UiT The Arctic University of Norway

Faculty of Science and Technology, Department of Chemistry

Exploring multipartite genomes using pangenome analysis

Gene distribution, gene expression and genome openness

Cecilie Bækkedal Sonnenberg

A dissertation for the degree of Philosophiae Doctor, May 2023

Exploring multipartite genomes using pangenome analysis

Gene distribution, gene expression and genome openness

Cecilie Bækkedal Sonnenberg

A dissertation for the degree of Philosophiae Doctor

Department of Chemistry, Faculty of Science and Technology,

May 2023

To my remarkable and greatly missed Onkel Jan

CONTENTS

ACKNOWLEDGEMENT

The work presented in this thesis has been carried out at the University of Tromsø – The Arctic University of Norway, Faculty of Science and Technology, at the research group Molecular Biosystems and Bioinformatics.

It all began in 2010, during my fourth year studying molecular biotechnology, which mostly consisted of lab courses that I didn't particularly enjoy. However, on the first day of the "Introduction to Bioinformatics" course, which was taught by Peik Haugen and the Molecular Biosystems and Bioinformatics group, I finally found a technique in biotechnology that I was passionate about. Even though I didn't understand much at the time, I knew that this was what I wanted to do. Later that year, when I had to find a research group to conduct my master thesis, I contacted Peik, and he has more or less been my supervisor since. Contacting him was an excellent decision because Peik is truly the best supervisor one could get. He has been there for me as an academical cliff, always available and always with a smile. He has a magical way of knowing precisely when to challenge me, when to give me responsibility and when to take the pressure off of me. I am extremely grateful for everything you have thought me, Peik, from writing skills, thinking outside the box, accepting disappointing results and being patient and precise.

Besides working on this thesis, my life has been very eventful these eight years that I have spent on this PhD degree. I am grateful for many amazing days out in the mountains, both with and without skis. Tromsø is truly an unique place on Earth. Two definite highlights over the years have been marrying my irreplicable husband, Markus, and giving birth to our daughter, Tiril Marie. To complete this PhD thesis would have been impossible without Markus, who has taken care of literally everything so that I could write, and I am endlessly grateful. I always come home to a delicious dinner, a tidy house and a very happy Tiril. You are really a superhuman and superdad! A special thanks to Tiril for bringing joy and happiness into our lives, for making me laugh and unwind and taking me on adventures in the Duplo universe. I also want to thank my supportive family and friends for always being there for me and for accepting that I have been in my own little PhD bubble lately.

A big thanks to Espen Robertsen, who has provided computer assistance whenever I needed it. Thank you to all of my colleagues, past and present, for our enjoyable fellowship.

Last but not least, I would like to express my gratitude to Onkel Jan, to whom this thesis is dedicated, for introducing me to science and critical thinking at a young age.

Cecilie Bækkedal Sonnenberg

May 2023, Tromsø, Norway

ABSTRACT

Bacteria are small single-celled organisms that are found in nearly every habitat on Earth. While bacterial genomes usually consist of one large circular replicon, about 10% of bacteria have organized their genes onto several large replicons. These multipartite bacteria are often found in symbiotic or pathogenic relationships with other higher organisms and are believed to have greater ability to adapt to new niches and to changing environments. However, much remains unknown about multipartite bacteria. In this study we aimed to gain a better understanding of why some bacteria have organized their genes on several large replicons. To do so, *Vibrionacaeae* and *Pseudoalteromonas,* which both consist of two large replicons, were used as model systems.

In **Paper 1**, pangenome analysis and transcriptomic data of *Vibrionaceae* revealed a highly organized distribution pattern of different gene types on the chromosome, and a strong correlation between gene expression and distance to the origin of replication. In **Paper 2**, *Pseudoalteromonas* showed a similar distribution pattern and correlation with gene expression on the chromosome as in *Vibrionaceae.* In **Paper 3**, pangenome analysis showed that *Vibrio* and *Pseudoalteromonas* have a larger repertoire of genes than genomes with one chromosome. Furthermore, horizontally transferred genes are inserted into specific regions on the replicons. In **Paper 4**, seven new complete genomes of *Vibrio anguillarum* genomes were presented.

Overall, results from these studies have increased our understanding of how multipartite genomes are organized with respect to their genes, how they are expressed and where newly acquired genes are retained on the replicons.

LIST OF PAPERS

Paper 1

Cecilie Bækkedal Sonnenberg, Tim Kahlke and Peik Haugen. 2020. ***Vibrionaceae* core, shell and cloud genes are non-randomly distributed on Chr 1: An hypothesis that links the genomic location of genes with their intracellular placement**. BMC Genomics 21 (1), 695. doi: 10.1186/s12864-020-07117-5

Paper 2

Cecilie Bækkedal Sonnenberg and Peik Haugen. 2021. **The *Pseudoalteromonas* multipartite genome: Distribution and expression of pangene categories, and a hypothesis for the origin and evolution of the chromid.** G3 Genes|Genomes|Genetics, 11 (9), jkab256. doi: 10.1093/g3journal/jkab256

Paper 3

Cecilie Bækkedal Sonnenberg and Peik Haugen. 2023. **Bipartite genomes in Enterobacterales: Independent origins of chromids, elevated openness and donors of horizontally transferred genes.** International journal of molecular sciences, 24 (5), 4292. doi: 10.3390/ijms24054292

Paper 4

Kåre Olav Holm, Cecilie Bækkedal, Jenny Johansson Söderberg AND Peik Haugen. 2015. **Complete Genome Sequences of Seven *Vibrio anguillarum* Strains as Derived from PacBio Sequencing.** Genome biology and evolution, 10 (4), 1127-1131. doi: 10.1093/gbe/evy074

ABBREVIATIONS

CID	chromosomal interaction domain
COG	Clusters of Orthologous Groups
GTDB	Genome Taxonomy database
HTG	horizontally transferred gene
kb	kilobases (1000 basepairs)
LCA	Last Common Anchestor
Mb	megabases (1,000,000 basepairs)
mRNA	messenger RNA
NAPs	nucleoid-associated proteins
NCBI	National Center for Biotechnology Information
ori	origin of replication
RNAP	RNA polymerase
RNA-seq	RNA sequencing
RP	ribosomal proteins
ter	replication terminus region

1 BACKGROUND

1.1 The bacterial cell

Bacteria are small single-celled organisms without a defined enveloped nuclei or organelles, and with a size of only a few micrometers in diameter (Westoby et al. 2021). They are found in nearly every habitat on Earth, from fresh and marine waters, terrestrial ecosystems and even in the clouds (Whitman et al. 1998; Dillon et al. 2021). Before the 1990s bacteria were viewed as primitive “bags of enzymes” with unstructured and randomly folded DNA (Levin and Angert 2015). However, during the last couple of decades it has become evident that bacteria contain a highly organized intracellular space (Surovtsev and Jacobs-Wagner 2018). For instance, the bacterial genome is organized into a confined space called nucleoid (Surovtsev and Jacobs-Wagner 2018), cytoskeletal filaments provide the cell with mechanical support (Ingerson-mahar and Gitai 2012) and processes such as replication and transcription are highly regulated (Yubero and Poyatos 2020). One important component of this organization is the packaging of chromosomal DNA into the nucleoid. Since the chromosomal DNA of bacteria is much longer than the length of the cell, it must be highly compacted to fit inside the bacterial cell (reviewed by Dame et al. 2020). In *Escherichia coli*, for example, the DNA must be compressed 1000 times to fit into the cytoplasm (Sherratt 2003), as the cell is 1.5—2.5 micrometers in length and the DNA measures 1.5 millimeters in length when fully extended (Surovtsev and Jacobs-Wagner, 2018). The nucleoid is compact yet flexible, allowing it to adjust to changes in the environment (Dame et al. 2020). This enables the appropriate chromosomal regions to be accessible at the right time, thus facilitating crucial cellular processes such as DNA replication and transcription.

1.1.1 Structural and spatial organization of the bacterial genome

Bacterial genomes usually consist of one large circular chromosome (i.e. one circular DNA ; Volff and Josef, 2000)**,** often accompanied by one or more auxiliary small circular DNAs, named plasmids. When subjected to torsional stress, the circular DNAs undergo supercoiling and form supercoiled domains (**Figure 1**). Supercoiled domains with an average size of 10 kilobases (kb) (Higgins et al. 1996; Postow et al. 2004) form chromosomal interaction domains (CIDs) (Le et al. 2013). CIDs typically span 100–200 kb and the majority of CIDs are separated by > 2 kb regions that contain highly expressed genes (Le et al. 2013). Exactly how chromosomes are organized into CIDs may vary during different growth phases. For instance, in the Gram-negative bacterium *Caulobacter crescentus*, the chromosome is structured into 23 CIDs during exponential growth in a nutrient-rich environment but organized into 29 CIDs during starvation (Le and Laub 2016). Several DNA binding proteins, including topoisomerases, structural maintenance of chromosomes (SMC) complexes, and nucleoid-associated proteins (NAPs), play crucial roles in maintaining the supercoiling of DNA and the compaction and dynamic nature of the nucleoid (reviewed by Badrinarayanan et al. 2015 and Le Berre et al. 2022). While topoisomerases help to maintain the supercoiled state of DNA by removing or inserting supercoils into the genome, NAPs influence chromosome organization by bending, wrapping or bridging DNA, and structural maintenance of chromosome complexes (SMC) encircles DNA. At a final level of organization, macrodomains of up to 1 Megabases (Mb) have been discovered (Valens et al. 2004). Macrodomains in *E. coli*, for example, include Ter (terminus), Ori (origin), Left, Right, and two unstructured domains (**Figure 2A**).

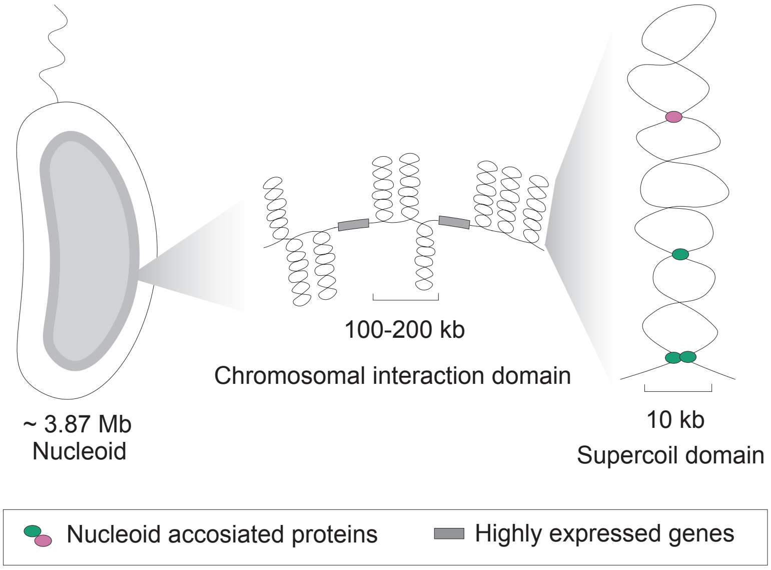

Figure 1: Packaging of DNA in the bacterial cell. The bacterial genome is highly compacted into the nucleoid. The nucleoid is composed of several chromosomal interaction domains, which each consist of supercoiled domains. Green and pink dots indicate nucleoid-associated proteins and grey boxes indicate regions of highly expressed genes.

The spatial arrangement, i. e. the three-dimensional arrangement, of replicons within a cell, has been established in some species. The *E. coli* chromosome has a transverse orientation, with the origin of replication and the terminus region in the middle of the cell and the two chromosomal arms extended on the opposite sides of the cell (Wang et al. 2006), as shown in **Figure 2A**. Both *Vibrio cholerae* (David et al. 2014; Fogel and Waldor 2005; Val et al. 2016;

Srivastava and Chattoraj 2007) and *C. crescentus* (Le et al. 2013; Yildirim and Feig 2018) have chromosomes that are longitudinally oriented, with the replication origin located at one pole and the replication terminus found at the opposite pole, as shown in **Figure 2B** and **C**. In addition, in *V. cholerae*, which contain a genome consisting of two replicons, the smaller replicon is placed in one half of the cell, with the origin of replication located at the middle of the cell and the terminus region placed in at the new pole. Both the smaller replicon of *V. cholerae* and the chromosome of *C. crescentus* have a helicoidal shape, while the largest replicon of *V. cholerae* has an open structure (Val et al. 2016; Yildirim and Feig 2018).

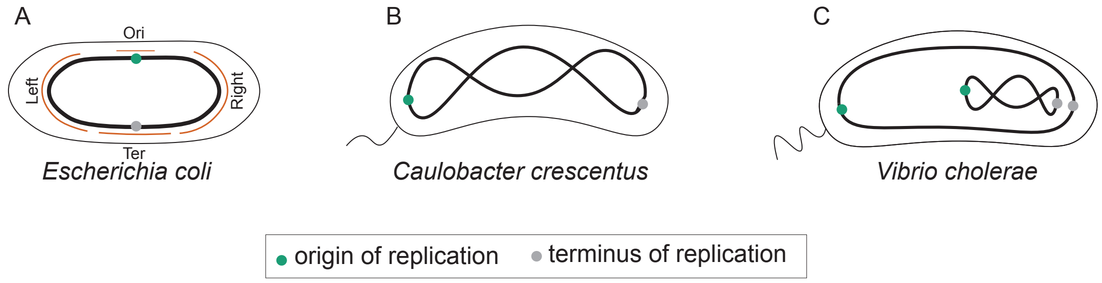

Figure 2: Subcellular placement of chromosomes in three model bacteria. The A) *E. coli* chromosome has a transverse orientation. The macrodomains Left, Right, Ori, and Ter are highlighted in orange. The chromosomes of B) *C. crescentus* and C) the largest replicon of *V. cholerae* are longitudinally oriented. The largest replicon of *V. cholerae* has an open structure, whereas the smaller replicon of *V. cholerae* and the chromosome of *C. crescentus* have a helicoidal shape. Green and grey dots, respectively, indicate the origin of replication and terminus of replication, respectively.

In summary, the chromosomal DNA is supercoiled and packed at multiple levels, so that it is both compact enough to fit inside the bacterial cell and flexible so that transcription and replication can occur. Also, the intracellular placement and orientation of chromosomes within cells varies among different organisms.

1.1.2 Transcription and gene expression

In order to function and survive, bacteria are dependent on effective and rapid response to changes in their environments (Boor 2006). When bacteria sense a change in their surroundings, this environmental stimulus can be converted into cellular signals, which may alter gene expression in order to produce RNA or proteins to respond to the environmental change (Boor 2006). How quickly and accurately the cellular signals can influence gene expression is dependent on, in part, how the gene content is organized. In bacteria, the gene content is in general very compact, with genes occupying 90% of the genome (Achaz et al. 2003). The genes are organized into operons, which typically consist of 3—4 consecutively arranged genes (Zheng et al. 2002). The total number of genes in the same operon may however be dozens (Zheng et al. 2002). Genes within an operon are often functionally related, for example, encoding physically interacting proteins or proteins involved in the same metabolic pathway (Daruvar et al. 2002). The operon is controlled by a common promoter (or in some cases several alternative promoters), where regulatory proteins can bind to promote or inhibit transcription (Bervoets and Charlier 2019).

For transcription to start, RNA polymerase (RNAP) binds to a promoter sequence at the front of the operon (reviewed by Bervoets and Charlier 2019). A transcription factor called sigma factor, binds to the RNAP and assists RNAP by helping it recognize the promoter site and initiate transcription, forming the RNAP holoenzyme. Bacterial cells typically have multiple sigma factors, and the sigma factor which is used is determined by environmental stimuli and what type of products the cell needs. As the RNAP proceeds down the DNA strand, ribonucleotides attach to form the messenger RNA (mRNA). When the RNAP reaches a termination signal, transcription is finished, and the mRNA molecule can serve as template for protein synthesis. Translation starts when ribosomes bind to the correct site on the mRNA, called ribosomal binding site, and tRNAs charged with amino acids are brought to the ribosome and added to what becomes a growing chain of amino acids until the ribosome reaches at the stop codon (see Ii and Ibba 2020)

Rate of transcription and translation of single genes and operons varies depending on growth conditions, with gene expression being activated or inhibited in response to environmental stimuli. Regulation of gene expression occurs at multiple levels, with some genes/operons being transcriptionally regulated by binding of regulators to promoter regions, and some genes/operons being regulated by post-transcriptional modification of mRNA, or regulated during translation (Ii and Ibba 2020). RNA sequencing (RNA-seq) is a powerful technique for studying gene expression levels at the RNA level. It allows researchers to investigate how genes are expressed, on a global scale (thus known as transcriptomics), under various conditions. RNA-seq quantifies all RNA molecules in a sample collected at a particular time-point and can therefore be used to understand how bacteria respond to environmental changes or different treatments. RNA-seq became popular with the advent of the next-generation sequencing technologies, such as pyrosequencing (Roche, 454 technology), sequencing-by-synthesis (e.g., Illumina), semiconductor sequencing (Thermo Fisher, Ion

Torrent) and sequencing-by-ligation (SOLiD) (see Hong *et al.* 2020). The so called third generation sequencing such as single-molecule real-time (SMRT) sequencing by Pacific Biosciences (Brown et al. 2014) and nanopore sequencing by Oxford Nanopore technologies (Quick et al. 2015) can produce longer reads, up to several thousand nucleotides.

1.1.3 Replication and multifork replication

For a bacterial population to grow, each bacterial cell must replicate its genome, so that each newly divided cell can inherit one copy of the genome (see Donnell et al. 2013). Replication begins at a specific position on the chromosome known as the origin of replication (**Figure 3**), which is often shortened as "*oriC*" in bacteria with one replicon and "*ori1*" and "*ori2*" in bacteria with several replicons. Here, the DNA helix is locally unwound after helicases bind and separate the DNA strands, which enable the replisome to assemble. The replisomes, which consist of multiple proteins, i.e., DNA polymerases, helicases and a clamp loader, form replication forks, before they move bidirectionally along the chromosome arms until they meet at the opposite side of the chromosome, i.e., the terminal region (*ter*). As replication proceeds, the duplicated chromosomes start to separate and one of the duplicated chromosomes travels across the cell to the opposite cell pole (reviewed by Gogou et al. 2021). After replication is finished, the cell divides into two daughter cells, each of which has a single copy of the chromosome.

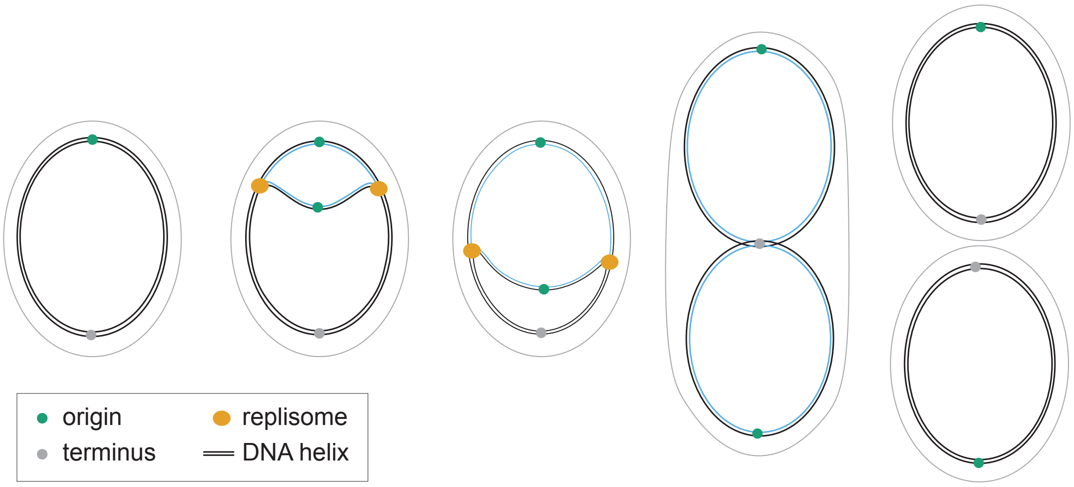

Figure 3: Replication begins at the origin of replication, and two replication forks form as the replisome progresses down each chromosome arm and ends at terminus of replication. During replication, the cell doubles, resulting in each cell possessing a single copy of the chromosome. The origin of replication and terminus of replication are indicated by green and grey dots, respectively, and the replisome is denoted as an orange dot. The newly replicated DNA strand is shown in blue.

During periods of rapid bacterial growth, some bacteria like *E. coli*, *Bacillus subtilis*, and *V. cholerae* can initiate multiple rounds of replication before completing the first (**Figure 4**). This is made possible because the time required for chromosome replication is longer than the time needed for the cell to divide (Cooper and Helmstetter 1968; Couturier and Rocha 2006; Stokke et al. 2011). In *E. coli*, for example, replication typically takes 45-60 minutes, while it only takes 20 minutes for the cell to double its mass and divide (Cooper and Helmstetter 1968). As a result, the new daughter cell inherits a chromosome that has already undergone partial division. This process, known as multifork replication, leads to an increase in the number of DNA copies near the *ori* compared to that of later-replicating regions. This phenomenon is referred to as the "gene dosage effect" (Couturier and Rocha 2006).

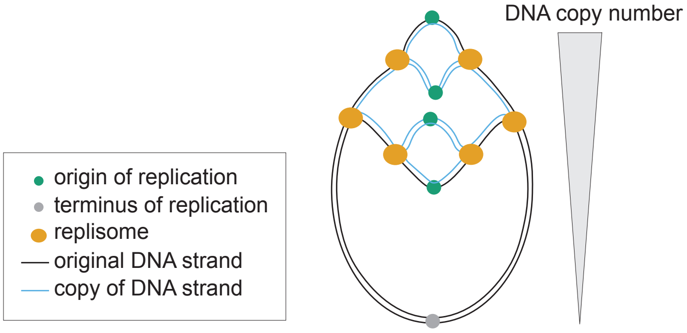

Figure 4: Replication with multiple replication forks. New replication rounds have begun on the chromosome that are already being replicated. This results in more copies of genes near the origin of replication compared to terminus of replication. The total gene copy numbers across the chromosome are illustrated by a gray cone. Origin and terminus of replication are represented by green and grey dots, respectively, while the replisome is denoted as an orange dot. The newly replicated DNA strands is shown in blue.

1.2 Multipartite genomes

The first bacterial genome consisting of two replicons was discovered in 1989 and belonged to *Rhodobacter sphaeriodes* (Suwanto and Kaplant 1989). Bacteria with their genes organized on several large replicons, i.e., multipartite genomes, have later been discovered in several phyla. These include *Actinomycetota* (formerly *Actinobacteria* (Oren and Garrity 2021)), *Deinococcota* (formerly *Deinococcus-Thermus*), *Bacillota* (formerly *Firmicutes*) and *Pseudomonadota* (formerly *Proteobacteria*), *Bacteroidota* (formerly *Bacteroidetes*) and *Spirochaetota* (formerly *Spirochetes*) (diCenzo and Finan 2017, all renaming of Oren and

Garrity 2021). Today approximately 10% of the known bacterial genomes are multipartite (Harrison et al. 2010; diCenzo and Finan 2017; Almalki et al. 2023). The majority of multipartite genomes are found within the phylum *Pseudomonadota*, with *Alphaproteobacteria* accounting for 25%, *Betaproteobacteria* for 46% and *Gammaproteobacteria* for 28% (Harrison et al. 2010; diCenzo and Finan 2017; Almalki et al. 2023). However, these numbers are based on currently available genome sequences, which may present a biased representation of existing bacterial populations. As the ability to cultivate a broader variety of bacteria improves, and their genome sequences become available, these numbers and the distribution of multipartite bacteria can be expected to change.

Multipartite genomes are often associate with animals, humans or plants as either pathogens or symbionts (Harrison et al. 2010; diCenzo and Finan 2017; Almalki et al. 2023). They are also associated with high abiotic stress tolerance (Misra et al. 2018). *Sinorhizobium meliloti* (phylum *Pseudomonadota*), for example, is a free-living bacterium containing a genome of two large replicons and a megaplasmid, lives in a symbiotic relationship with the root nodes of legume plants (Gibson et al. 2009). Another example is *Deinococcus radiodurans* (phylum *Deinoccocota*), which has two large chromosomes and a megaplasmid, and can withstand high levels of ionizing radiation (White et al. 1999). *D. radiodurans* even survived a year in space, orbiting Earth outside the International Space Station (Ott et al. 2020).

The smaller of the two replicons of *Rhodobacter sphaeriodes* was originally named a "secondary chromosome" (Suwanto and Kaplant 1989). Today, the terms "secondary chromosome", "chromid", and "Chr 2" are used interchangeably to describe additional replicons in multipartite bacteria. In this thesis, additional replicons will from here on be referred to as "chromids". In 2010, Harrison et al introduced the term "chromid", which is a combination of "*chrom*osome" and "plasm*id"*, to describe additional replicons in multipartite bacteria. The researchers proposed the following three criteria to define chromids; i) it should have a nucleotide composition that resembles that of the chromosome, ii) it should contain plasmid-type maintenance and replication systems, and finally, iii) it should have presence of core genes. Furthermore, diCenzo and Finan (2017) suggested that additional replicons larger than 350 kb in size should be regarded as chromids. However, the chromosome, which is the largest replicon in the genome, contains the majority of the core (essential) genes (Harrison et al. 2010). Chromids typically fall between the size range of plasmids and chromosomes, being larger than plasmids but smaller than chromosomes (Harrison et al. 2010). The median size of chromids is

1.26 (Mb), which is smaller than the corresponding number of chromosomes (i.e., 3.65 Mb) but larger than that of megaplasmids (i.e., the 46.2 kb) (diCenzo and Finan 2017).

1.2.1 Possible advantages of the multipartite genome structure

Over the years, multiple hypotheses have been proposed to explain the persistence of multipartite genomes in nature. In general terms, the hypotheses can be divided into two categories: Replication-related hypotheses, and niche specialization and adaptation hypotheses.

The latter are based on that the chromid is advantageous. For example, Cooper et al (2010) observed that genes on *Vibrio* and *Burkholderia* chromids evolve faster than on the chromosomes (Vaughn S Cooper et al. 2010) . They hypothesized that chromids act as evolutionary "test beds", as they are subjected to weaker selective pressure and higher mutation rates. These properties may enable neutral genes to resist being selected against before they have the chance to provide the host with some kind of advantage at particular conditions or niches. On the same note, the chromid of *Flammeovirgaceae,* a family of polysaccharide-degrading bacteria within the phylum *Bacteroidota*, show a higher degree of evolutionary plasticity, and also experience a more relaxed selection pressure compared to the corresponding chromosome (Feng et al. 2021). Furthermore, the genes on the *Burkholderiaceae* chromids evolve more rapidly and acquire genes faster than the chromosomes, thus leading to the idea that the primary benefit of having several replicons, is that the genetic malleability of the bacteria is increased, thus allowing for expansion of gene content through accumulation of horizontally transferred genes (diCenzo et al. 2019).

In terms of replication-related hypotheses, many of the models that explain of how replication in multipartite occurs are based on knowledge from the genus *Vibrio*. In brief, replication begins with initiation of replication of the chromosome, followed by initiation of chromid replication, i.e., when the replication of the chromosome is approximately two-thirds done. This synchronization enables both the chromosome and chromid to complete replication simultaneously (Rasmussen et al. 2007; Val et al. 2016). Also, in some bacteria, multifork replication occurs during fast growth (Couturier and Rocha 2006; Rasmussen et al. 2007; Srivastava and Chattoraj 2007). One potential advantage of having several replicons is that it might take less time to complete replication of the entire genome (Rasmussen et al. 2007; Choudhary et al 2012). Also, the necessary number of forks would be less (Srivastava and Chattoraj 2007), and a chromosome with fewer forks should be less vulnerable to damage. Furthermore, the delayed replication of the chromid may allow for reduced complexity of overlapping replication cycles in fast growing bacteria and less complicated chromosome segregation during fast growth (Rasmussen et al. 2007). Additionally, it has been suggested that difference in gene expression levels between the replicons is a way to regulate and fine-tune, in a replicon-specific manner, gene expression of certain genes (Balsiger et al. 2004; Srivastava and Chattoraj 2007; Couturier and Rocha 2006; Dryselius et al. 2008). According to a final theory (Slater et al. 2009), the existence of two or more replicons may enable the expansion of the genome size, and thus allow for the presence of more genes. This has been supported by observations showing that monopartite genomes (i.e., genomes with just one replicon) are typically smaller than multipartite genomes (Harrison et al. 2010; diCenzo and Finan 2017). It is essential to emphasize that these hypotheses are not mutually exclusive, and there may be several reasons that can explain the existence of multipartite genomes.

1.2.2 Hypotheses for the origin of chromids

Several hypotheses regarding the origin of multipartite genomes have been proposed, i.e., from where chromids originate and how they evolve. The most widely accepted is the plasmid hypothesis, which suggests that chromids originate and evolve from megaplasmids that are acquired by bacteria. Over time, core genes are transferred from the chromosome to the megaplasmid, transforming it into a chromid. Therefore, a hallmark of such chromids would be that they use plasmid-like replication machineries, which consist of a replication initiation protein and partitioning systems. Phylogenetic analyses of the partitioning system genes *parA* and *parB* within *Burkholderiaceae* (*Betaproteobacteria*) suggest that the chromids arose from two ancestral plasmids (diCenzo et al. 2019). Similarly, in *Rhizobiaceae* (*Alphaproteobacteria*), it is believed that chromids originated from a small number of plasmids (Harrison et al. 2010). However, although virtually all observations of chromid to date support the plasmid hypothesis, other scenarios cannot be excluded, e.g., that the chromid is a result of i) a split from a chromosome (schism hypothesis), ii) recombination between a plasmid and a chromosome (de novo), iii) recombination between a plasmid and a chromid (rebirth) (Harrison et al. 2010) and iii) uptake of an entire chromid from other bacteria through conjugation (Choudhary et al 2012).

1.3 *Vibrionaceae* and *Pseudoalteromonas* - model systems to study multipartite genomes

Vibrionaceae and *Pseudoalteromonas,* which both belong to *Enterobacterales*, carry the only known multipartite genomes within *Gammaproteobacteria* (**Figure 5**). *Pseudoalteromonas* represents one of 44 genera in *Alteromonadaceae.* In contrast, all bacteria in *Vibrionaceae* carry a chromid. They belong to one of eight genera, including *Vibrio*, *Aliivibrio*, *Paraphotobacterium*, *Photobacterium*, *Salinivibrio*, *Thaumasiovibrio* and *JCM-19050.* Both the chromid of *Vibrionaceae* (Fournes et al. 2018; Heidelberg et al. 2000) and *Pseudoalteromonas* (Médigue et al. 2005; Liao et al. 2019; Rong et al. 2016) are believed to have originated from a plasmid which over time acquired essential genes from the chromosome. Based on estimates of time since divergence, *Vibrionaceae* has existed for a much longer time than *Pseudoalteromonas* (Xie et al. 2021; Liao et al. 2019). The *Pseudoalteromonas* chromosome and chromid have coexisted for about 378–502 million years, whereas the chromosome and chromid of *Vibrionaceae* have coexisted for 900–110 million years. Moreover, while *Vibrionaceae* has been extensively studied for many years, research on *Pseudoalteromonas* genomes has only gained traction during recent years. This is reflected by the number of complete genome sequences in the National Center for Biotechnology Information (NCBI) RefSeq database (O'leary et al. 2016), which includes 74 complete *Pseudoalteromonas* genomes and 608 *Vibrionaceae* genomes as of April 2023. The fact that *Pseudoalteromonas* and *Vibrionaceae* are members of the same order, but originated at different times, makes them an interesting case for studying multipartite genomes.

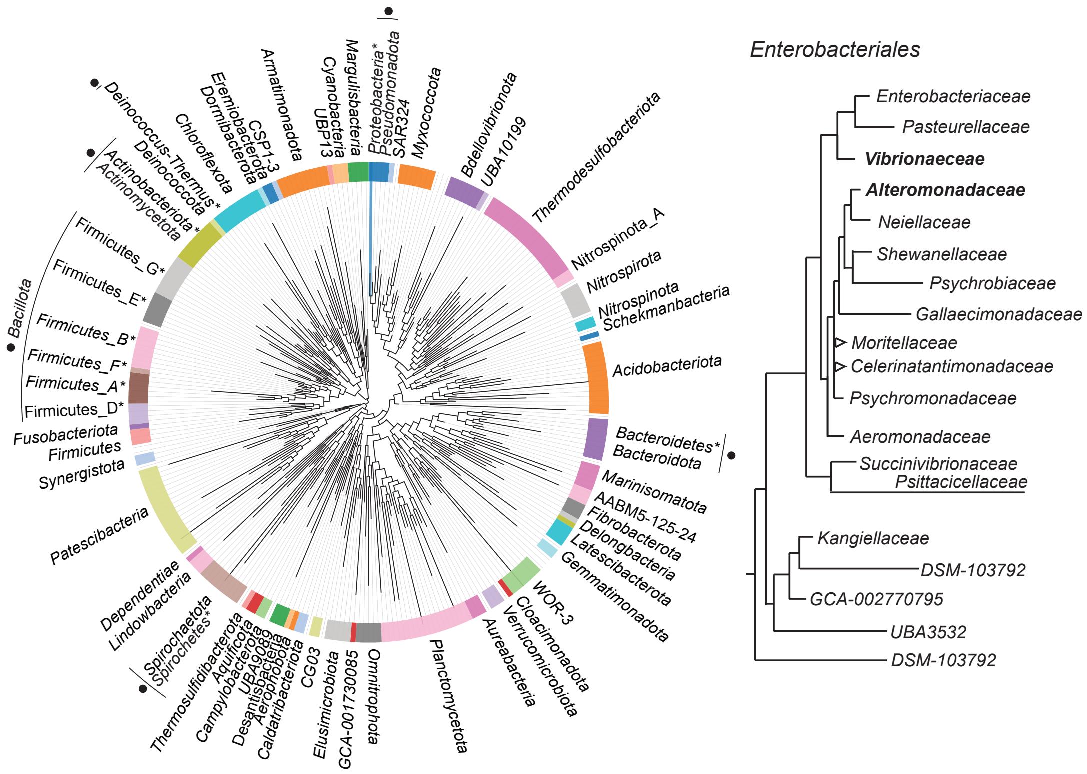

Figure 5: Evolutionary relationship between bacterial phyla, and between families within *Enterobacterales*. Phylogeny between bacterial phyla is shown in the left tree. The phylum containing *Gammaproteobacteria* is highlighted in blue. Phyla with multipartite bacteria are indicated with black dots. Phyla names marked with an asterisk, refer to their previous name before being renaming (Oren and Garrity 2021). Evolutionary relationships between *Enterobacterales* families are illustrated to the right. Families containing multipartite bacteria are displayed in bold. Both trees the are derived from the Genome Taxonomy database (GTDB).

1.3.1 *Vibrionaceae*

Vibrionaceae was first documented in 1854 as the causal agent of two independent cholera outbreaks in London and Florence (Carboni 2021). The first reports indicating that the *Vibrionaceae* genome comprise two replicons were published in the late 1990s (TRUCKSIS et al. 1998; Yamaichi et al. 1999). The first genome sequence of *V. cholerae* was published in 2000 by Heidelberg and colleagues and confirmed the presence of two large replicons (Heidelberg *et al*. 2000). *Vibrionaceae* is a family of gram-negative, rod-shaped bacteria with one or more polar flagella (Zhu et al. 2013). They are widely distributed in aquatic environments, including oceans, lakes and brackish waters, where they can either swim freely or interact with other organisms, including fish, corals, and humans, either as symbionts or pathogens (Takemura et al. 2014; Thompson et al. 2004). Perhaps the most famous species include the pathogens *V. cholerae*, *Vibrio parahaemolyticus* and *Vibrio vulnificus*, which cause cholera, foodborne illnesses, gastrointestinal infections and wound infections, respectively, thereby posing a significant threat to human health (Onohuean et al. 2022).

Replication of the chromosome and chromid, which are approximately 3 and 1 Mb in length respectively, are synchronized (Val *et al.* 2016), with replication of the chromosome and chromid terminating at the same time (Rasmussen *et al.* 2007). The chromosome is first replicated about two-thirds, before the chromid replication starts. This happens when a locus on the chromosome called *ctrS* (Chr 2 replication triggering site), a non-coding locus of 150 bp, is replicated (Val *et al.* 2016). While the replication initiator protein DnaA initiates replication on the chromosome, RctB initiates replication on the chromid (Duigou et al. 2006). RctB can bind to several regulatory sites, including the 39-mer and 12-mer regulatory sites in origin of replication on the chromid (*ori2*), which either promotes or inhibits replication initiation, respectively (Venkova-Canova and Chattoraj 2011). Interestingly, research recently revealed that after *crtS* is replicated on the chromosome, binding of RctB to *crtS* prevents RctB from inhibiting replication at the 39-mer site, which allows replication initiation of the chromid (Fournes et al. 2021).

Vibrios are known to be some of the fastest-growing bacteria on Earth, with *V. cholerae* being capable of dividing every 17 minutes (Soler-Bistu *et al.* 2015), and *Vibrio natriegens* holding the record for the fastest-growing bacteria with a doubling time of just 10 minutes (Weinstock et al. 2016). Studies on *V. cholerae* (Rasmussen *et al.* 2007; Srivastava and Chattoraj 2007), *V. vulnificus*, and *V. parahaemolyticus* (Dryselius *et al* 2008) have demonstrated that during rapid growth, multifork replication occurs, resulting in a higher copy number of DNA near origin of replication (*ori1*) compared to terminus (*ter1*) on the chromosome (known as the gene dosage effect). However, only a single round of replication occurs per cell cycle on the chromid.

1.3.2 *Pseudoalteromonas*

Pseudoalteromonas was first described as *Alteromonas* (Baumann et al. 1972), but was later reclassified into a separate genus and given its current name, which comes from the Greek word "pseudo" and means "false or similar" (Gauthier et al. 1995). *Pseudoalteromonas* is a genus of gram-negative and heterotrophic bacteria, which have rod-shaped cells with flagella (reviewed by Parrilli *et al.* 2021). The genus has been isolated from various marine habitats, such as coastal, open and deep-sea waters and sediments. *Pseudoalteromonas* is often associated with healthy animals, such as invertebrates and fishes and algae. *Pseudoalteromonas* is also known for its ability to live in extreme environments. For instance, *Pseudoalteromonas* sp. SM9913 has been isolated from deep sea sediments collected at 1855m depth (Chen et al. 2003), and *Pseudoalteromonas haloplanktis* TAC125 was isolated from Antarctic coastal seawaters (Médigue *et al.* 2005) and can survive temperatures ranging from -2.5 to 29 °C (Sannino et al. 2017; Piette et al. 2011). *Pseudoalteromonas* species can be divided into pigmented and non-pigmented, with pigmented species being linked to the production of natural products like antimicrobial, anti-fouling, algicidal, and compounds relevant to pharmaceuticals. Only a few strains of *Pseudoalteromonas* have been identified as pathogenic, including *Pseudoalteromonas agarivorans*, which is harmful to the sponge *Rhopaloeides odorabile*. *Pseudoalteromonas piscicida* is another interesting species, which has been demonstrated to attack and kill competing bacteria, e.g., *V. parahaemolyticus* (Richards et al. 2017). Using electron microscopy, it was discovered that *P. piscicida* transfers lytic vesicles to the outer surface of *V. parahaemolyticus*, resulting in holes in the cell wall, followed by gathering of *P. piscicida* cells around the dying vibrio to assimilate nutrients.

It is noteworthy that most *Pseudoalteromonas* chromids replicate unidirectionally, a phenomenon that has not previously been observed for other chromids (Xie *et al.* 2021; Médigue *et al.* 2005). In unidirectional replication, replication proceeds clockwise from origin of replication to terminus of replication. However, in some *Pseudoalteromonas*, such as the closely related *Pseudoalteromonas spongiae* and *Pseudoalteromonas piratica*, chromids are replicated bidirectionally. Bidirectional replication is suggested to have evolved from unidirectional replication through gene insertion events of, for example, prophage-like regions. Interestingly, despite replication mode, all chromids have a *tus* gene at the terminus region (Xie et al. 2021). In *E. coli*, *tus* encodes a replication fork trap system that enforces replication termination in the terminus region (Galli et al. 2019). The *tus* gene is thought to have originated from a plasmid and is present in the majority of *Enterobacteriaceae* and *Aeromonadaceae* genomes. Moreover, regardless of the replication mode, both the chromosome and the chromid replications terminate simultaneously.

1.3.3 Differences in taxonomic classification in NCBI and GTDB

The development of new methods for taxonomic classification of prokaryotes has resulted in recent changes in the taxonomic classification of Bacteria. In 2018, the Genome Taxonomy database (GTDB) was released, and it bases its taxonomy on a set of 120 single copy marker proteins (Parks et al. 2022). On the other hand, NCBI (O'leary et al. 2016) is a curated classification and nomenclature database based on, among others, current taxonomic literature

(Federhen 2012). According to NCBI´s classification, the genus *Pseudoalteromonas* belongs within the family *Pseudoalteromonadaceae* and order *Alteromondales*, while *Vibrionaceae* belongs to "*Vibrionales*" (**Table 1**). Both *Alteromonadales* and "*Vibrionales*" belong within the phylum *Gammaproteobacteria*. However, according to GTDB, both *Pseudoalteromonas* belong to the family *Alteromonadaceae*, which together with *Vibrionaceae* belong to the order *Enterobacterales*, *Gammaproteobacteria* (**Figure 5**). In **Paper 1** and **2**, we adhered to the taxonomic classification system used by NCBI. However, in **Paper 3**, we followed the GTDB classification. Furthermore, in 2022, the International Committee on Systematics of Prokaryotes proposed new names for phyla, and *Proteobacteria* was renamed *Pseudomonadota* (Oren and Garrity 2021).

Table 1: Comparison of GTDB and NCBI taxonomic assignments of families within the Enterobacterales order.

Enterobacterales order	NCBI family	NCBI order
Alteromonadaceae	*Alteromonadaceae, Colwelliaceae, Chromatiaceae, Idiomarinaceae, Pseudoalteromonadaceae*	*Alteromonadales, Chromatiales*
Aeromonadaceae	*Aeromonadaceae*	*Aeromonadales*
Celerinatantimonadaceae	*Alteromonadaceae*	*Alteromonadales*
Enterobacteriaceae	*Budviciaceae, Enterobacteriaceae, Erwiniaceae, Hafniaceae, Morganellaceae, Pectobacteriaceae, Yersiniaceae*	*Enterobacteriales*
Gallaecimonadaceae	No records	No records
Kangiellaceae	*Kangiellaceae*	*Oceanospirillales*
Neiellaceae	*Psychromonadaceae*	*Alteromonadales*
Moritellaceae	*Moritellaceae*	*Alteromonadales*
Pasteurellaceae	*Pasteurellaceae*	*Pasteurellales*
Pittacicellaceae	*Pasteurellaceae*	*Pasteurellales*
Psychrobiaceae	unclassified *Gammaproteobacteria*	No records
Psychromonadaceae	*Alteromonadaceae, Moritellaceae, Psychromonadaceae*	*Oceanospirillales*
Shewanellaceae	*Ferrimonadaceae, Shewanellaceae*	*Alteromonadales*
Vibrionaceae	*Vibrionaceae*	*Vibrionales*

1.4 Pangenome analysis and openness of pangenomes

Pangenome analysis is a branch of comparative genomics, and is a method used to describe the complete set of all genes (i. e. the pangenome) present in a set of genomes, often a taxonomic unit like a bacterial family, genus or species. The pangenome (from the Greek word “pan” which means whole) concept was first introduced by Tettelin et al. in (2005), soon after many new genome sequences became available as a result of new DNA sequencing technologies in mid-2000s. This resulted in the possibility to compare multiple bacterial genomes and genomic variations at the same time, instead of analyzing single genomes. The first pangenome studies were presented by Tettelin et al. (2005) and Hogg et al. (2007), based on eight *Streptococcus agalactiae,* or twelve *Heamophilus influenzae* genomes, respectively. These studies showed that the pangenome contained significantly more genes than genomes from any single strain. This finding contradicted the widely accepted idea that sequencing the genome of a single isolate from a particular species was enough to reflect the genomic makeup of that species.

Pangenome analysis was initially used to study the diversity of bacteria, but it soon expanded to include studies of fungi, plants, animals and even humans (Golicz et al. 2020). Today, pangenome studies are widely used in research for investigating for example pathogenic mechanisms, environmental adaptation and evolutionary relationships. Moreover, pangenome analysis has proven useful in identification of potential vaccine targets and as an epidemiological tool. It was, for example, used for identifying the origin of a cholera outbreak in Haiti in 2010. At first, it was unclear whether a local or an Asian *V. cholerae* strain was responsible for the epidemic, but pangenome analysis demonstrated that the outbreak was caused by strains originating from Southeast Asia (Chun et al. 2009). Pangenomics was also used during the COVID-19 pandemic to investigate the evolutionary relationships among various SARS-CoV-2 genomes (Parlikar et al. 2020).

1.4.1 Calculation of the pangenome

Pangenome analysis divides genes within a taxonomic unit into pangene categories based on homology. The main three pangene categories are: core genes that are present in all the genomes, shell genes that are present in two or more genomes but not in all, and cloud genes that are present in one or two genomes (**Figure 6**). Additionally, genes that are found in 95% of genomes are categorized as "softcore" genes. In general, core genes typically encode housekeeping functions related to replication, transcription and translation, as well as genes with regulatory functions (Tettelin et al. 2005; Tettelin et al. 2008). However, their functions may not necessarily be essential. Shell and cloud genes are not necessary for the basic survival of a species, but they typically encode genes that offer selective advantages, such as niche adaptation, colonization of new hosts, antibiotic resistance, and pathogenicity (Tettelin *et al.* 2008). Many cloud genes encode proteins of unknown function.

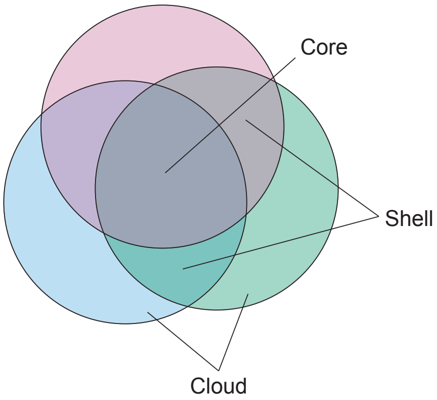

Figure 6: Pangenome of three genomes. Each large circle represents a genome. The genes shared by all genomes constitute the core genes, genes present in only two of the genomes make up shell genes and genes found in only one genome form the cloud gene category.

1.4.2 Calculation of pangenome openness

An analysis that is often used in conjunction with pangenome analysis, is the calculation of the "openness" of a pangenome. The openness of a pangenome reflects the ability of bacteria under study to acquire new genes, which is a major driving force in the evolution of bacterial genomes and is associated with the ability to survive in various niches. Bacterial species that can colonize multiple environments have more opportunities to exchange genetic material and thus have more open pangenomes. Interactions with other organisms can result in large pangenomes with a high number of shell and cloud genes (Rouli et al. 2015). In contrast, bacteria living in isolation tend to have closed pangenomes with limited opportunities to acquire external genes. Heap's law can be used to describe the openness of a pangenome, i. e. the pangenome size and the number of new genes added for each new genome sequence. Heap's law is formulated as: $n = kN^{\gamma}$, where n is the pangenome size, N is the number of genomes used, and k and γ (often referred to as "Heaps´s exponent") are the fitting parameters. A pangenome is considered open if $\gamma > 0$, indicating that for each new genome added, new genes are being added to the pangenome, as illustrated in **Figure 7**. In contrast, a closed pangenome is characterized by $\gamma < 0$, indicating that the majority of genetic information has already been revealed in previously sequenced genomes, and the pangenome size approaches a constant as more genomes are added.

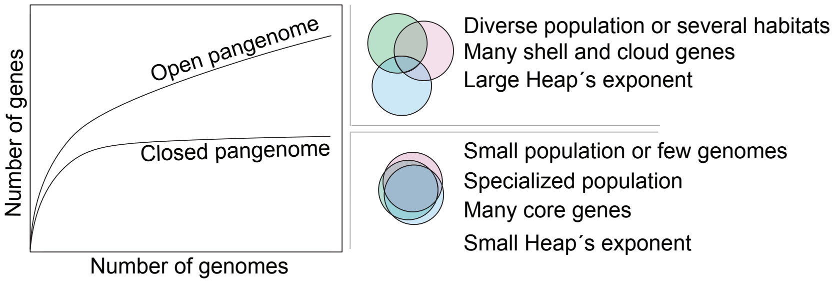

Figure 7: Estimation of pangenome openness using Heaps law. The openness is estimated though curve fitting of the pangenome size (number of genes in the pangenome) on the y-axis versus the number of genomes on the x- axis using Heap´s law. The pangenome sizes the median of pangenome size of random combinations of genomes. An open pangenome is characterized by a curve that increases and large Heap´s exponent, indicating a diverse population, that has the ability to live in several habitats has a pangenome consisting of many shell and cloud genes. Conversely, a closed pangenome has a flattened curve and is associated with a small population or few genomes, a high number of core genes, and bacteria that live in one niche. Large circles represent three genomes and illustrates open and closed pangenomes.

A pangenome analysis can be influenced by several factors, which can affect the results such as pangenome size and distribution of pangene categories. Some important factors to consider when conducting a pangenome analysis are: i) the parameters used during homology clustering, specifically sequence identity and sequence coverage (Tettelin et al. 2008), ii) the quality of the genome sequences, including the potential for sequencing errors or the use of draft genomes (Park et al. 2019), and iii) the choice of bacterial strains, as species diversity can affect the pangenome size (Costa et al. 2020; Tettelin et al. 2008).

1.4.3 Pangenome of *Pseudoalteromonas*

In 2017, Bosi et al. (2017) conducted a pangenome analysis of 38 *Pseudoalteromonas* genomes from various ecological niches, including 15 isolates from Antarctica (Bosi et al. 2017). They found that the *Pseudoalteromonas* genus has an open pangenome with a large proportion of unique genes (80%), while only 7% of the genes were core genes and 13% were accessory genes. The study also revealed that *Pseudoalteromonas* genomes have gained genes since their last common ancestor (LCA), with the average *Pseudoalteromonas* genome today containing 4245 genes compared to the estimated 2999 genes in their LCA. The authors suggested that horizontal gene transfer, particularly through transduction, has significantly influenced the genomic diversity and openness of the genus. Furthermore, the researchers calculated the panmobilome of *Pseudoalteromonas*, which includes all mobile genetic elements in each pangene category, and found that the majority of the mobile genetic elements originated from plasmids, with only a small portion originating from viruses.

1.4.4 Pangenome of *Vibrionacaeae*

Kahlke et al. (2012) calculated the pangenome of 64 *Vibrionaceae* genomes, which included a wide range of genetically diverse isolates, including pathogenic, non-pathogenic, environmental and clinical strains (Kahlke et al. 2012). The study found that the core genome accounted for 18%, while 78% of the genes were accessory genes and 3% were unique genes. Interestingly, the researchers also identified unique core genes, which are genes that are shared by any specific group of genomes in the dataset. They discovered 12 unique core genes in *V. cholerae*, some of which are related to aerotaxis and biosynthesis, suggesting that these genes play a role in adaptation to its specific niches. Another study, which included 20 *Vibrio* genomes, discovered that the *Vibrio* pangenome was open and contained a large number of shell genes (Lin et al. 2018). In addition, the researchers found that *Vibrio* had significantly more horizontally transferred genes (HTGs) than other marine bacteria and suggested that HTGs provide *Vibrio* genomes with a genetic diversity which is helpful in niche survival. Furthermore, analysis of gene gain/loss events indicated that *Vibrio* has experienced gene expansion throughout evolution, with an estimated increase from 2828 genes in the LCA to an average of 4547 genes in *Vibrio* genomes today. Based on a pangenome analysis of eleven *Vibrionaceae* strains, Lilburn et al. (2010) found that most of the genomic diversity, i. e. shell and cloud genes, on the chromosome was accounted for by genomic islands (Lilburn et al. 2010). On the chromid, the genomic diversity was more evenly distributed and significant genetic variation was observed on the superintegron. The superintegrons, which is 120 kb long, function as a gene capture system and acts as a reservoir for genes with functions related to adaption. It is important for the high genetic variability of, for example, *V. cholerae* (reviewed by Escudero and Mazel 2017).

2 AIMS OF STUDY

The main goal of this study was to gain a better understanding of why some bacteria have organized their genes onto several large replicons, i.e., on one chromosome and one or several chromids. The following questions were addressed using a pangenome approach and *Vibrionaceae* and *Pseudoalteromonas* as model systems:

1) Are different gene types distributed randomly between the replicons, or is it possible to find distinct patterns?
2) Given that genes are distributed in a non-random fashion, is there a correlation between how genes are distributed on the replicons and how the genes are expressed?
3) Do bacteria with several large replicons have a greater capacity to acquire genes from other organisms than bacteria with one replicon? If so, where are these new genes inserted on the replicons?

3 SUMMARY OF PAPERS

Paper 1

***Vibrionaceae* core, shell and cloud genes are non-randomly distributed on Chr 1: An hypothesis that links the genomic location of genes with their intracellular placement**

Cecilie Bækkedal Sonnenberg, Tim Kahlke and Peik Haugen.2020. BMC Genomics 21 (1), 695.

In this work we studied gene distribution on the chromosome and chromid of *Vibrionaceae*. Furthermore, we studied how gene expression levels correlates with genomic location and how the pangene categories contributes to the observed gene expression levels.

The pangenome analysis of 124 *Vibrionaceae* genomes and mapping of the pangene categories core, softcore, shell and cloud back to their genomic locations, showed that core and softcore genes were overrepresented around *ori1* on the chromosome, whereas shell and cloud genes were overrepresented the regions surrounding *ter1*. Publicly available RNA-Seq data from *Vibrio natriegens* and *Aliivibrio salmonicida* revealed that gene expression strongly correlated with the distance to *ori1*, with higher expression levels closer to *ori1*. Under fast-growing conditions all pangene categories contributed to high expression pattern around *ori1*, while softcore, shell and cloud contributed under slow growing conditions. The chromid showed no distribution bias, and the gene expression pattern did not correlate with distance to *ori2* or *ter2*. Furthermore, based on the subcellular organization of the chromosome and chromid in *V. cholerae*, we presented a hypothesis suggesting that core/softcore and shell/cloud are spatially separated into distinct intracellular regions in the cell.

Paper 2

The *Pseudoalteromonas* multipartite genome: Distribution and expression of pangene categories, and a hypothesis for the origin and evolution of the chromid.

Cecilie Bækkedal Sonnenberg and Peik Haugen.2021. G3 Genes|Genomes|Genetics 11 (9), jkab256.

In this work we studied the gene distribution on the chromosome and chromid of *Pseudoalteromonas*, and how gene expression levels correlates with genomic location and how the pangene categories contribution to the observed gene expression levels.

Based on pangenome analysis of 25 *Pseudoalteromonas* genomes, followed mapping of the pangene categories core, softcore, shell and cloud back to their genomic locations, we discovered that core genes were significantly overrepresented around terminus on the chromids. However, on the chromosome, all pangene categories were more evenly distributed on the chromosome, and core/softcore genes were weakly overrepresented around *ori1*, and shell/cloud genes were weakly overrepresented regions around *ter1*. Publicly available RNA-Seq data from *Pseudoalteromonas fuliginea* were used to analyze gene expression under optimal and sub-optimal growth conditions. Gene expression strongly correlated with the distance to *ori1*, with higher expression levels closer to *ori1* under both fast and slow growing conditions. Under fast growth all pangene categories contributed to high expression pattern around *ori1*, while only shell genes contributed under slow growing conditions. Furthermore, 78 chromid hallmark genes (i. e. genes located on the chromids in all the 25 genomes) were identified, and BLAST searches showed that the majority of the genes originated from *Alteromonadales*, indicating that this is where the chromid originated from. Finally, a large number of genes associated with iron-acquisition and homeostasis were identified on the chromids.

Paper 3

Bipartite genomes in *Enterobacterales*: Independent origins of chromids, elevated openness and donors of horizontally transferred genes.

Cecilie Bækkedal Sonnenberg and Peik Haugen. 2023. International journal of molecular sciences, 24 (5), 4292.

In this work we used *Vibrio* and *Pseudoalteromonas* to investigate the genome openness of bipartite genomes, and to determine which types of genes that are more likely to have been acquired horizontally, thus leading to open bipartite genomes.

Pangenome analysis and Heap´s law was used to calculate the openness of *Vibrio* and *Pseudoalteromonas* and monopartite genomes from the same order (*Enterobacterales*). Bipartite genomes were more open than monopartite genomes. The *Vibrio* chromosome and chromid were equally open, whereas the *Pseudoalteromonas* chromid was more open than the chromosome. Codon usage bias calculations and the HGTector software was used to detect horizontally transferred genes among the pangene categories, core, softcore, shell and cloud. This revealed that the pangene categories shell and cloud contribute most to the openness of the bipartite genomes. Based on our previous studies of gene distribution in *Vibrionaceae* and *Pseudoalteromonas,* we proposed a hypothesis suggesting that the chromids and the chromosome terminus region contributes to the genomic plasticity of bipartite genomes. Furthermore, the majority of the horizontally transferred genes in both *Vibrio* and *Pseudoalteromonas* were predicted to have originated from the genus *Shewanella.* Finally, based on phylogenetic analysis using parA and parB protein sequences, we suggested that chromids of *Vibrionaceae* and *Pseudoalteromonas* originated from two separate plasmid acquisition events and that both chromids were acquired from plasmids that belong to *Enterobacterales.*

Paper 4

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

Kåre Olav Holm, Cecilie Bækkedal, Jenny Johansson Söderberg, and Peik Haugen. 2015. Genome Biology and Evolution. 10 (4), 1127-1131.

In this work we presented seven new complete *Vibrio anguillarum* genomes and performed basic genome comparison and pangenome analysis of the in total eleven complete *V. anguillarum* genomes (as of March 2018). Furthermore, we described the structural features of superintegrons on the chromid of *V. anguillarum* and identified novel insertion sequences.

The seven *V. anguillarum* genomes were *de novo* assembled using long-sequence PacBio reads. Genome comparison of the 11 complete *V. anguillarum* strains using the global BLAST comparison tool, BRIG, revealed both previously described and undescribed genome gaps, which is genomic regions consisting of cloud genes only present in *V. anguillarum* NB10 and/or shell genes. The pangenome analysis indicated an open pangenome, with a large amount of shell and cloud genes. Moreover, 18 new insertion sequences were identified using the ISfinder database.

4 DISCUSSION

The initial goal of this project was to study gene distribution and gene expression in the multipartite genomes of *Vibrionaceae*. As the work progressed, it became clear that there was a distinct gene distribution pattern on the chromosome, and that the location of genes was closely linked to how genes were expressed. To see if this was unique to *Vibrionaceae*, we used the same methodology on *Pseudoalteromonas*, which together with *Vibrionaceae* represent the only bacteria within *Gammaproteobacteria* with genomes distributed on multiple replicons. Interestingly, we found that *Pseudoalteromonas* has similar gene distribution and gene expression patterns as described for *Vibrionaceae*. Intrigued by our findings we next set out to determine the openness of the *Vibrionaceae* and *Pseudoalteromonas* genomes, this to learn more about factors that may contribute to the diverse lifestyle of the bacteria, e.g., genomic flexibility of multipartite genomes. By comparing the openness of multipartite and monopartite genomes, we identified genes (and their location) that are responsible for elevated openness of multipartite genomes.

Although there are several intriguing findings that warrant discussion, I have chosen to go more in depth on the following topics; the holistic use of pangenome analysis, gene distribution pattern, gene expression, pangenome openness and spatial distribution of genes in the *V. cholerae* cell. These topics are discussed in separate sections below.

4.1 Holistic use of pangenome analysis

The introduction of first-generation sequencing in the 1970s was a game changer and facilitator to the field of molecular biology. Later, high-throughput sequencing in the mid-2000s, resulted in massive numbers of genome sequencing projects and therefore increased availability of genomes. The first bacterial genome to be fully sequenced was *Heamophilus influenzae* in 1995 (Fleischmann et al. 1995). This was soon followed by completion of the genomes of other bacteria such as *E. coli* in 1997 (Blattner et al. 1997), *B. subtilis* in 1997 (Kunst et al. 1997) and *V. cholerae* in 2000 (Heidelberg et al 2000). At the time, it was believed that the variability within a bacterial species was limited, and that a genome of a single isolate of a given species was sufficient to describe the genomic content of that species (Medini et al. 2020). However, the first pangenome analyses revealed that there was a considerable genomic variation between closely related strains, and that the total number of genes in all genomes (later called pangenome) contained many more genes than that found in just one strain, highlighting the need for multiple sequences (Tettelin et al. 2005, 2008; Medini et al. 2005; Hogg et al. 2007).

Since then, there have been further advances in sequencing technology, resulting in a tremendous increase in the number of complete genome sequences, which has reached 37,944 as of April 2023 in NCBI RefSeq database (O'leary *et al.* 2016). For instance, we increased the number of complete *V. anguillarum* genomes from four to eleven in 2015 by adding seven new strains using the PacBio long-reads sequencing (**Paper 4**).

We have taken advantage of the large number of complete genome sequences to investigate genomes in a pangenomic context. The pangene categories derived from pangenome analysis have been used as a foundation to study gene distribution, gene expression, horizontally transferred genes and codon usage. More specifically, each gene from the pangene categories core, softcore, shell and cloud were mapped to their position on its respective genome to study gene distribution patterns (**Paper 1** and **Paper 2**). Gene expression data was then mapped back to the genomic location of the pangenes, to investigate the relationship between gene expression and gene distribution, and to identify which pangene categories that contributed to the observed expression levels (**Paper 1** and **Paper 2**). Additionally, codon usage and the number of horizontally transferred genes (HTGs) were detected for each pangene category, resulting in a more detailed picture of horizontal gene transfer and genome plasticity (**Paper 3**). In all, this approach has allowed for a comprehensive study of why some bacteria carry their genes on multiple large replicons.

This study demonstrates the potential of combining pangenome analysis with other research areas and bioinformatic tools, and it would have been interesting to add additional research fields into our analysis. Other researchers have also combined pangenome analysis with various fields. One example is a study that combined pangenomics and metabolomics in *Pseudoalteromonas luteoviolaceae* to investigate the diversity of potential bioactive secondary metabolites between species (Maansson et al. 2016). This could be especially interesting in *Pseudoalteromonas*, which is known for its ability to produce bioactive molecules. Another study combined pangenome analysis and transcriptomics to identify the pan-regulon, which includes all genes that are regulated by a single transcription factor, in this case Ferric uptake regulator, in closely related *E. coli* strains (Gao et al. 2020). Identifying the pan-regulon in *Vibrionaceae* and *Pseudoalteromonas* could potentially provide a better understanding of the gene expression levels on the chromosome and chromid.

In summary, we have used pangenome analysis as a foundation to study gene distribution, gene expression, horizontally transferred genes and codon usage in *Vibrionaceae* and *Pseudoalteromonas*, and it has proven to be an effective approach to study multipartite genomes.

4.2 Is gene distribution on the chromosome of *Vibrionaceae* and *Pseudoalteromonas* organized or random?

The bipartite genome structure of *Vibrionaceae* and *Pseudoalteromonas,* combined with their ability to live in, and adapt to diverse environments, makes them intriguing research topics. The complex nature of bacteria with multipartite genomes has been discussed since the discovery of such genomes (Suwanto and Kaplant 1989). It is widely believed that multipartite genomes have high genomic plasticity, with the chromid playing a key role in creating this variability (Escudero and Mazel 2017; Dicenzo et al. 2019; Vaughn S. Cooper et al. 2010). To enhance our understanding of why some bacteria carry their genes on multiple large replicons, it is necessary to consider several factors, such as the overall genome organization, distribution of genes across replicons, gene expression levels, genomic regions that may aid in adaptability and horizontal gene transfer, mechanisms for niche adaptation, and pathogenicity. This is obviously a very complex task. Therefore, as a starting point, we determined how different gene categories are distributed among chromosomes and chromids in *Vibrionaceae* and *Pseudoalteromonas,* this to address the following question: Can we establish whether the distribution of genes in bipartite genomes is random or highly organized, when considering the different gene types?

Gene order in bacteria was first established by Rocha (2004), which showed that essential genes, particularly those that are highly expressed, tended to be located near the origin of replication (*oriC*) in the two fast-growing bacteria *E. coli* and *B. subtilis* (Rocha 2004). However, the slow-growing *Mycobacterium tuberculosis* and moderately fast-growing bacteria *C. crescentus* (doubling time of 90 minutes), showed no apparent bias in gene distribution. Furthermore, in 2006, Couturier and Rocha showed that genes related to transcription and translation in *E. coli* were located near *oriC* and suggested that this positioning was related to gene dosage resulting from multifork replication. They also showed that genes related to translation and transcription, i. e. genes encoding tRNA and ribosomal proteins and protein-coding genes with the 5% strongest codon usage bias, were overrepresented on the chromosome in multipartite bacteria. In contrast, the chromid of *Agrobacterium*, *Brucella*, *Burkholderia*, *Vibrio* and *Photobacterium* contained fewer translation and transcription associated genes than expected. Interestingly, no significant bias was found between the chromosome and chromid in *Leptospira interrogans,* which was the slowest growing bacteria studied. This finding was consistent with results from Heidelberg et al. (2000), that already in 2000 indicated that the *Vibrio* chromosome contained more genes involved in essential biosynthesis pathways than the chromid. Dryselius et al. (2008) found that genes contributing to growth, both essential and non-essential, were significantly overrepresented on the chromosome in *Vibrio* and were located in early replicating regions. Functional annotation of five *Vibrio* genomes showed overrepresentation of Clusters of Orthologous Groups (COG) categories crucial for proliferation on the chromosome. In contrast, categories of little importance for growth were overrepresented on the chromid. Recently, pangenome analysis has been used as basis to explore gene order and distance to *oriC*. Among 101 *Rhodobacteraceae* (*Alphaproteobacteria*) genomes, 83 showed a tendency for highly conserved genes to cluster in closer proximity to *oriC*. Some of the species showed extreme cases, where 22 had core genes that clustered significantly near *oriC*, while in eight species core genes clustered close to terminus (Kopejtka et al. 2019). In a study of 401 *Klebsiella pneumonia* strains, softcore genes (present in 95% of the genomes) were located near *oriC*, whereas shell and cloud genes (present in 5% of the genomes) were located farther away from *oriC* (Comandatore et al. 2019). Furthermore, a series of intriguing studies on the importance of the genomic location of a locus known as *s10-spc-α* (S10) in *V. cholerae* have been conducted by moving the S10 away from its position close to *ori1* (Soler-Bistué et al. 2015; Soler-Bistué et al. 2017; Soler-Bistué et al. 2020; Larotonda et al. 2023). In *Vibrionaceae*, half of the RP genes are found in S10 (Soler-Bistué et al. 2015). The importance of the genomic location of S10 in evolution and cellular physiology, was recently demonstrated by Larotonda et al. (2023). The researchers created strains with S10 located either near or far from *ori1* and subjecting them to 1000 generations of evolution. Even though strains with S10 distant from *ori1* were able to survive, they displayed decreased fitness and infectivity. This decrease in growth rate persisted for more than 1000 generations, which highlights the critical importance of genomic location, at least for this particular locus. To sum up, previous research on gene order has mainly focused on genomic location of specific gene types, typically those important during rapid growth, such as highly expressed genes and genes involved in translation and transcription. Our analysis differs from these prior studies because we consider all genes in the genomes, divided into pangene categories, rather than focusing on genes based on their function or expression.

In my research, I conducted studies on the distribution of core genes between the chromosome and chromid in *Vibrionaceae* (**Paper 1**) and *Pseudoalteromonas* (**Paper 2**). My findings are in agreement with previous studies where genes associated with growth were typically found on the chromosome, instead of on the chromid (Heidelberg et al. 2000; Rocha 2004; Couturier and Rocha 2006; Dryselius et al. 2008). Most core genes were found on the chromosome, while the majority of genes on the chromid are shell and cloud genes (**Paper 1** and **Paper 2**) (**Figure 8**). Notably, core genes are not necessarily essential, or growth-related, but the majority of the functionally annotated core genes on the chromosome correspond to the COG categories "Translation, ribosomal structure and biogenesis", followed by "Amino Acid metabolism and transport", "Transcription", "Replication and repair" (unpublished data based on functional annotation from **Paper 3**).

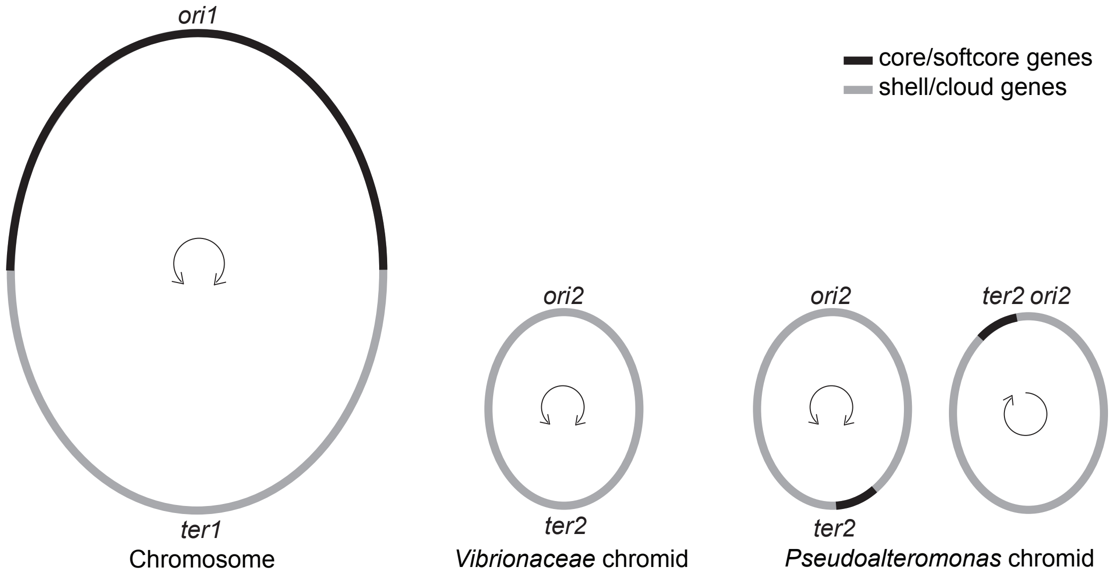

Figure 8: Model of gene distribution on the chromosome and chromid of *Vibrionaceae* and *Pseudoalteromonas*. On the chromosome of both *Pseudoalteromonas* and *Vibrionaceae*, the majority of core/softcore genes (black) are located on the upper half, close to *ori1*, while shell/cloud genes (grey) are overrepresented on the lower half, close to *ter1*. There is no distribution bias on the *Vibrionaceae* chromid, On the of *Pseudoalteromonas* chromid, core/softcore genes are overrepresented around *ter2*, regardless of the chromid is uni- or bidirectionally replicated. The majority of genes on the chromids are shell and cloud.

The distribution pattern of core/softcore genes on the chromosome in both *Vibrionaceae* and *Pseudoalteromonas* is similar to that observed in monopartite bacteria capable of fast growth, such as *E. coli*, *B. subtilis* (Rocha 2004; Couturier and Rocha 2006), *K. pneumoni,* and *Rhodobateriaea* (Kopejtka et al. 2019; Comandatore et al. 2019). Our findings revealed that the region proximal to *ori1*, specifically the upper half of the chromosomes, show a higher concentration of core and softcore genes (**Figure 8**). Conversely, the lower half, near *ter1*, show an overrepresentation of shell and cloud genes. However, the pattern is less clear in

Pseudoalteromonas than in *Vibrionaceae* (**Paper 2**), where there instead is a weak overrepresentation of core and softcore genes near *ori1* and of shell genes near *ter1*. One potential explanation for this less pronounced gene distribution pattern in *Pseudoalteromonas* could be that *Vibrionaceae* is able to grow faster than *Pseudoalteromonas*. Couturier and Rocha (2006) observed that in multipartite bacteria, the stronger the observed gene dosage effect, the higher the overrepresentation of highly expressed genes on the chromosome (Couturier and Rocha 2006). *V. cholerae* has a short doubling time of 16 min (Couturier and Rocha 2006), whereas *P. haloplanktis* has a doubling time of 31 minutes at 20 °C (Médigue et al. 2005).

The *Vibrionaceae* chromid show no gene distribution bias, whereas on the chromid of *Pseudoalteromonas*, core/softcore genes are overrepresented around *ter2* irrespective of replication mechanism (**Figure 8**). Most *Pseudoalteromonas* chromids replicate unidirectionally, and some exceptions, such as the *P. spongiae* and *P. piratica* replicate bidirectionally. Interestingly, in **Paper 2** we identified 71 "chromid hallmark genes", i. e. core genes that are present on all the *Pseudoalteromonas* chromids. Out of these, 31 are found in clustered around *ter2*, and they are involved in functions such as histidine biosynthesis, DNA binding protein, acetolactate synthase, biopolymer transport system and cell division.

In summary, we used a pangenome approach to investigate gene distribution patterns in *Vibrionaceae* and *Pseudoalteromonas* and found that genes are far from randomly distributed on the studied multipartite genomes. Instead, they are found highly organized on the chromosomes, but less organized on the chromids. The gene distribution on *Vibrionaceae* chromid was virtually random, whereas core/softcore genes were significantly overrepresented close to terminus on the *Pseudoalteromonas* chromid, irrespective of replication direction.

4.3 Identifying correlation patterns between the genomic placement of genes and their expression level

The gene order analysis described above revealed that genes are non-randomly distributed on the chromosomes of *Vibrionaceae* and *Pseudoalteromonas*, with core/softcore genes located near the origin of replication and shell/cloud genes close to the terminus. A similar organization of genes has been observed in other bacteria that exhibit fast growth rates, for example in *E. coli* and *B. subtilis*, where translational and transcriptional genes are located close to *oriC* (Couturier and Rocha 2004). During fast growth of these bacteria, genes located near *oriC* are typically expressed at higher rates compared to genes located at the opposite side of the circular replicon, with a genome-wide gene expression that decreases towards the terminus. This expression trend was attributed to the gene dosage effect, which occurs due to several rounds of replication during fast growth, resulting in multiple gene copies around *oriC*. Gene dosage is also in effect in *Vibrionaceae*. Studies of *V. cholerae* (Rasmussen et al. 2007, Srivastava and Chattoraj 2007), *V. vulnificus*, and *V. parahaemolyticus* (Dryselius et al 2008) showed that gene dosage is growth-dependent and only in effect on the chromosome, i.e., the chromid is unaffected and typically displays lower levels of gene expression compared to the chromosome.

The overrepresentation of core/softcore genes near *ori1* in *Vibrionaceae* (**Paper 1**) and *Pseudoalteromonas* (**Paper 2**), along with previous research indicating higher expression levels of *ori*-proximal genes in fast growing bacteria, served as the foundation for examining whether there is a correlation between gene location and expression, in *Vibrionaceae* and *Pseudoalteromonas*. To be more specific, our aim was to investigate whether global gene expression patterns are different between rapid and slow growing bacteria, and if so, to establish which gene types contribute to the difference. To address our aim, we used publicly available RNA-seq data from bacteria grown under fast- and slow growing conditions, and publicly available bioinformatic algorithms and software.

As anticipated, we found that gene expression decreases with increasing distance to *ori1* on the chromosomes during fast growth, whereas gene expression on the chromid is lower and relatively even (**Paper 1** and **Paper 2**). A similar pattern was observed during fast growth in *Vibrio splendidus* (Toffano-Nioche et al. 2012) and *V. parahaemolyticus* (Dryselius et al. 2008). Here, the expression level of the *V. splendidus* chromosome was on average 3.6 times higher than that of the chromid (when excluding rRNA operons), and the expression levels decreased gradually with increasing distance to *ori1* during exponential growth, in a rich medium. Similarly, in V. *parahaemolyticus,* both gene expression and DNA copy numbers decreased towards *ter1*, indicating that chromosomal gene dosage levels correspond with the gene expression levels. The chromidal gene expression in V. *parahaemolyticus* was lower, with no apparent link between gene dosage and expression.

There are several proposed explanations for the elevated expression close to *ori1*. For example, genes around *ori1* often have functions related to cell growth, and elevated expression of this region can therefore facilitate efficient and reliable responses to varying growth demands (Dryselius et al. 2008), as there is a substantial need for their gene products during growth (Slager and Veening 2016).

In contrast to the distinct patterns in gene expression and DNA copy numbers observed under fast growing conditions, as described above, multifork replication is expected to be absent, or at least minimal, under slow growing conditions. When analyzing data from *A. salmonicida*, *V. natriegens* and *P. fuliginea*, from samples collected under slow growth, we were therefore surprised to find gene expression patterns resembling that seen under fast-growing conditions. A similar observation was done by Dryselius (2008) when *V. parahaemolyticus* was grown under sub-optimal conditions. Although the *V. parahaemolyticus* chromosome maintained a higher numbers of DNA copies around *ori1* compared to *ter1*, the cell doubling time was longer compared to when grown under fast growth conditions. The authors explained their observation by suggesting that replication stalled or slowed down. Moreover, others have observed that change in DNA copy numbers throughout the genome can occur during slow growth and when multifork replication is absent. For instance, in *V. cholerae*, the relocation of the ribosomal protein gene cluster S10 under slow growing conditions led to a change in its DNA copy number depending on its genomic location (Soler-Bistué et al. 2017). The DNA copy numbers of S10 varied from 1.5 copies when S10 was located near *ori1,* to one when it was closer to terminus on either the chromosome or the chromid.

Based on the observation described above, gene copy numbers are only weakly elevated around *ori1* during slow growth. This suggests that multifork replication cannot alone explain the elevated expression levels on the upper half of the chromosome during such growth conditions. Our data show that softcore, shell and cloud genes contribute to increased expression around *ori1* during slow growth in *Vibrionaceae*, whereas only shell genes contribute to the same effect in *Pseudoalteromonas*. Clearly, there are likely several and complex reasons for this, that go beyond gene dosage effects. Contributing factors that potentially can influence gene expression include supercoiling of DNA (reviewed by Martis et al 2019), binding nucleoid-associated proteins (NAPs) to DNA (Le Berre et al. 2022) and variations in the transcriptionally regulatory regions in genes, e.g., proximal and distal promoters and ribosomal binding sites (Bervoets and Charlier 2019). Supercoiling of DNA is affected by alterations in the environment and regulates gene expression on a global scale (Peter et al. 2004; Martis B. et al. 2019). For instance, in a study of *E. coli,* the gradient of negative supercoiling was highest in the terminus region during exponential growth phase, whereas supercoiling was more uniform during stationary growth phase (Lal et al. 2016). Another interesting mechanism that can contribute to regulation of gene expression is xenogeneic silencing. The histone-like nucleoid structuring protein (H-NS) can function as a transcriptional silencer by binding to AT-rich sequences located inside bacterial promoters and thereby modulating RNA polymerase binding (Forrest et al. 2022). H-NS may act to repress expression of genes obtained through horizontal gene transfer, specifically those with a higher AT-content than the host genome, which could potentially lower the fitness of the bacterium (Navarre et al. 2006). Interestingly, shell and cloud genes in both *Vibrionaceae* and *Pseudoalteromonas* have a higher AT-content compared to the corresponding core and softcore genes (**Paper 3**), which are overrepresented in the terminus-region of the chromosome, and on the entire chromid. This could imply that they are more susceptible to binding by H-NS, which may result in silencing and lower gene expression. A study on H-NS binding in *V. cholerae* revealed that H-NS has a greater affinity for the chromid than the chromosome and a preference for suppressing genes located within genomic islands, which are believed to have been laterally acquired (Ayala, Wang, Benitez, et al. 2015; Ayala, Wang, Silva, et al. 2015). The majority of these genes were found within the superintegron on the chromid and a *Vibrio* pathogenicity island on the chromosome. This is intriguing because the superintegron spans a considerable portion, ranging from 7—29%, of the total chromid genomic sequence (**Paper 4**).

In summary, gene expression on the chromosome of *Pseudoalteromonas* and *Vibrionaceae* decreases with increasing distance from *ori1,* under fast-growing conditions. A main contributing factor is likely multifork replication, which results in high DNA copy numbers of genes near *ori1*. The elevated expression levels on the upper half of the chromosome were contributed to by all pangene categories in both *Pseudoalteromonas* and *Vibrionaceae*. During slow growing conditions, gene expression is elevated around *ori1*. This can, however, be attributed to shell genes in *Pseudoalteromonas,* and softcore, shell and cloud genes in *Vibrionaceae*. Gene dosage effects could not be detected for either of the two chromids. Instead, they both showed a generally lower level of expression, compared to the corresponding chromosome.

4.4 Exploring chromosome and chromid openness, and the contributing factors

To continue our quest to better understand why some bacteria contain multipartite genomes, we next turned our attention to pangenome openness. Pangenome openness refers to the ability of an organism to acquire new genes, which occur mostly by horizontal gene transfer (Treangen and Rocha 2011). Several factors can influence openness, such as the degree of interactions with competing and cooperating species, the number of niches that the bacteria inhabits and their lifestyle (Tettelin et al. 2008). Bacteria that live in a variety of environments or niches typically have more open genomes than those that live in isolation (McInerney et al. 2017). In fact, it is believed that bacteria with multipartite genomes often are identified as pathogens or symbionts of animals, humans and plants, because their genomes contain a high diversity of genes (i. e. large amount of shell and cloud genes) and that this trait plays a central role in their lifestyle. The chromid, in particular, has been believed to play a critical role in their successful spread and diversity (Heidelberg et al. 2000; Cooper et al. 2010a; Galardini et al. 2013; diCenzo et al. 2014; diCenzo et al. 2019). In our studies, we calculated the openness of *Vibrio* and *Pseudoalteromonas* genomes*,* and compared the result to that of monopartite bacteria from the same order. This was done using pangenome analysis and curve fitting using Heap´s law. The results show that *Vibrio* and *Pseudoalteromonas* have more open pangenomes than monopartite bacteria (**Paper 3**). The higher capacity of bacteria with multipartite genomes to acquire new genes, could imply that the size of their genomes grow larger over time. Earlier studies showed that multipartite genomes are indeed, on average, larger than monopartite genomes (Harrison et al. 2010; diCenzo and Finan 2017). This trait was attributed to the presence of chromids since they observed little difference in the size of the corresponding chromosomes (diCenzo and Finan 2017). To examine if the same is true for our dataset, we compared the sizes of *Vibrio* and *Pseudoalteromonas* genomes with genomes of closely related bacteria harboring only one large replicon (i.e., bacterial genera with monopartite genomes). The box plot in **Figure 9** shows that *Vibrio* and *Pseudoalteromonas* genomes are not larger in size than monopartite genomes, and we can therefore conclude that the increased openness of *Vibrio* and *Pseudoalteromonas* cannot be explained by extra voluminous genomes. Moreover, we calculated the openness of chromosomes and chromids separately, and found that the *Pseudoalteromonas* chromid is more open than the chromosome, whereas the *Vibrio* chromosome and chromid are equally open.

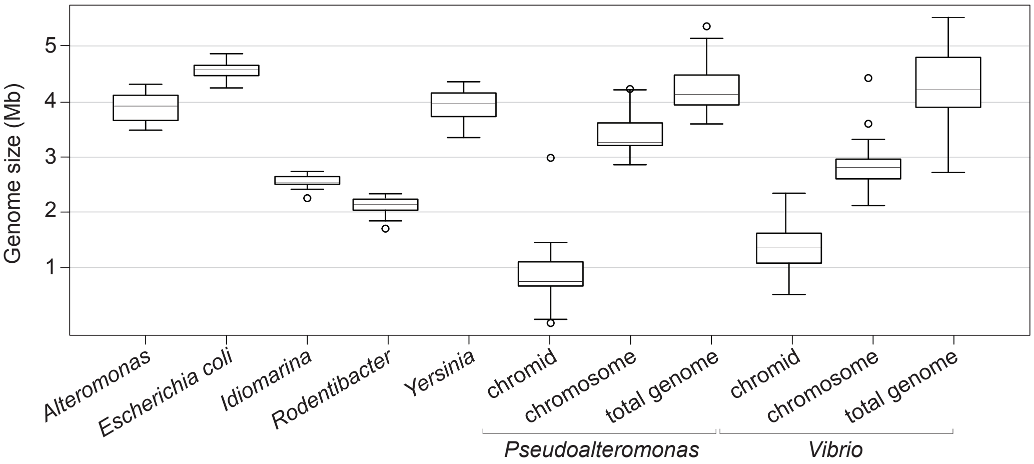

Figure 9: Boxplot of genome size of monopartite and multipartite genomes. Genome size of the monopartite genomes *Alteromonas*, *E. coli*, *Idiomarina*, *Rodentibacter* and *Yersinia* and the size of the chromid, chromosome and total genome of *Pseudoalteromonas* and *Vibrio* are measured in megabases.

Next, we identified the types of genes that are likely to have been acquired through horizontal gene transfer, and thus contribute to the open pangenomes of *Vibrio* and *Pseudoalteromonas*. This was done by using codon usage bias analysis and identification of putatively horizontally transferred genes. First, we found that shell and cloud genes exhibit an atypical codon usage compared to core and softcore genes. Atypical codon usage is typically seen in horizontally transferred genes where the codon usage of the donor deviates from that of the host (Tuller et al. 2011). Furthermore, we found that the vast majority of HTGs in *Vibrio* belong to the shell or cloud gene categories, whereas in *Pseudoalteromonas* HTGs are more evenly distributed across all pangene categories.

Interestingly, if the results described above is seen in the light of data showing that the majority of shell and cloud genes typically populates the chromosomal region surrounding *ter1*, and the entire chromid, then we can postulate that these regions can serve as safe "landing sites" for new genes and thus contribute to an increased genome openness. This model is intriguing since it suggests that a specific region on the chromosome and the entire chromid contribute to elevated openness, thus challenging the predominant idea that it is the chromid that constitutes the preferred landing site for incoming genes in bacteria with multipartite genomes. The chromid has been described as both an evolutionary test bed, where genes are weakly preserved and evolve more rapidly than the on the chromosome (Cooper *et al.* 2010), and as a niche-specialized replicon where new genes accumulate (diCenzo *et al.* 2019). Our results extend this theory to include the terminus proximate region of the chromosome. In fact, the number of shell and cloud genes in the chromosome terminus region and the chromid are roughly equal in

Vibrio and *Pseudoalteromonas*, thus demonstrating the significance of the terminus chromosome region. Interestingly, increased occurrence of mobile genetic elements near the terminus region on the chromosome has been observed in several other bacteria (Kopejtka et al. 2019; Oliveira et al. 2017; Touchon and Rocha 2016; Rocha 2004; Bobay et al. 2012; Esin et al. 2018). This has been proposed to be a way of minimizing disruption of genome organization, as it avoids affecting early replicating and highly expressed genes (Bobay et al. 2012; Oliveira et al. 2017). In Escherichia and Salmonella, for example, the frequency of prophages increases with distances to oriC and integration of phages are selected against in chromosomal regions with most highly expressed genes (Bobay et al. 2012).

Taken together, *Vibrio* and *Pseudoalteromonas* have more open genomes than monopartite bacteria from the same order. The elevated openness is due to incoming genes that today belong to the pangene categories shell and cloud, which are typically located in the lower region of the chromosome and on the chromid. The possibility to integrate foreign genes on both the chromid and the chromosome terminal region is likely an important advantage for bacteria with several replicons, in terms of niche specialization and diversification.

4.5 *V. cholerae* genome in the subcellular space

Accumulating experimental data supports that the intracellular space in bacterial cells is highly organized (reviewed by Surovtsev and Jacobs-Wagner 2018). Several factors contribute to the organization, such as chromatin compaction through nucleoid-associated proteins (reviewed by Dame et al. 2020), a complex network of structural protein fibers (i.e., a cytoskeleton) (Ingerson-mahar and Gitai 2012) and macromolecular crowding that influences diffusion and mobility of molecules in the cytoplasm (reviewed by Berg et al. 2017). Chromosomes are compacted into the nucleoid which is spatially organized in the cytoplasm (Wang and Rudner 2015; Dame et al. 2020). Studies of the spatial organization of the chromosome and chromid of *V. cholerae* provide an example of how the interior of a bacterium is organized (David et al. 2014; Fogel and Waldor 2005; Val et al. 2016; Srivastava and Chattoraj 2007). The chromosome stretches across the whole cytoplasm, with *ori1* spatially positioned at the old pole area and *ter1* at the new pole. The chromid spans from the mid cell, where *ori2* is positioned in the center of the cell, to the new pole, where *ter2* is located. This spatial positioning of the replicons leads to a partial separation of the pangene categories into separate intracellular regions, with core and softcore genes crowding the old pole, and shell and cloud genes heavily populating the new pole (**Paper 1**) (**Figure 10**).

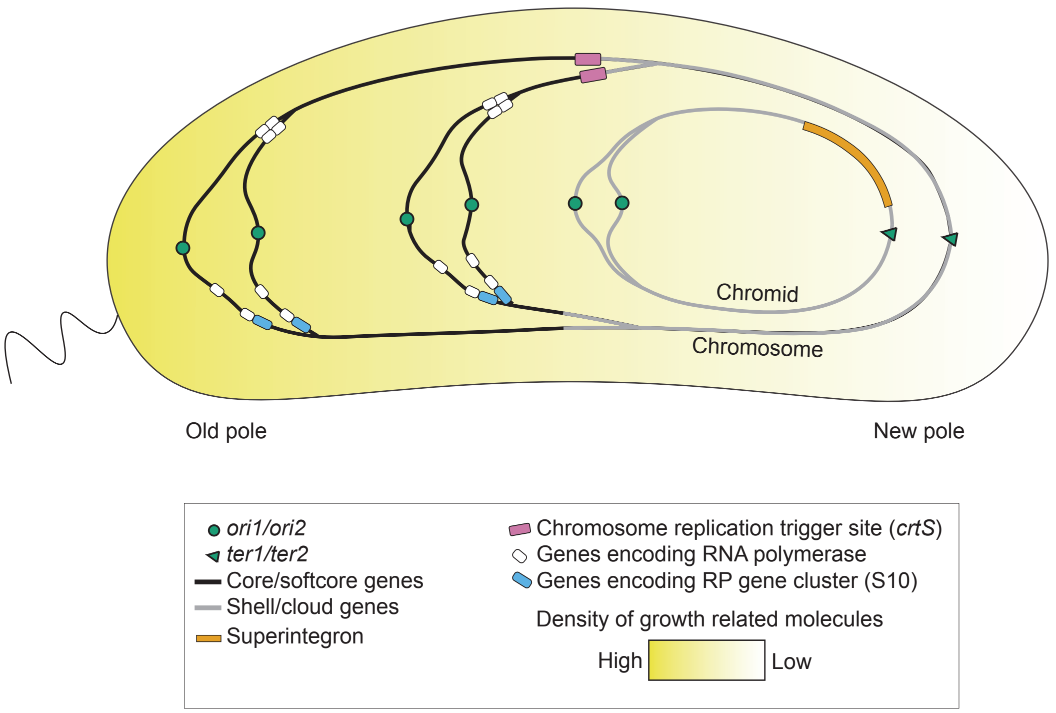
Chromid
Chromosome
Old pole
New pole
ori1/ori2
ter1/ter2
Core/softcore genes
Shell/cloud genes
Superintegron
Chromosome replication trigger site (crtS)
Genes encoding RNA polymerase
Genes encoding RP gene cluster (S10)
Density of growth related molecules
High
Low

Figure 10: Model of a *V. cholerae* cell with subcellular location chromosome and chromid. The core/softcore genes (black) are spatially enriched in the intracellular region near the old pole, while the shell/cloud genes (grey) are enriched in the new pole. Multiple rounds of replication (indicated by four green replication origins on the chromosome) results in several copies of RNAP encoding genes (white boxes) and RP encoding genes (blue boxes), all located in the old pole. The intensity of the yellow background color reflects the abundance of growth-related molecules in the cytoplasmic space. Replication of ctrS on the chromosome leads to the replication start of the chromid. A superintegron (orange) is present on the chromid. *Ori1*/*ori2* and *ter1*/*ter2* are indicated as green and grey dots, respectively.

The *Vibrionaceae* genome structure has been shaped over the course of millions of years of evolution (Xie et al. 2021) and it is therefore reasonable to assume that the partial separation of gene categories seen in today's genome and intracellular space, has been selected for and maintained because it is beneficial for survival and reproduction of the cells. The perhaps most plausible explanation is that physical separation of core/softcore genes from presumably less critical genes is favorable. For example, this can enable more specific and coordinated regulation of core genes (which often have growth related functions) at all cellular levels, including DNA packing, transcription, translation, post translational modification and transportation. Another example is that less critical genes can be acquired and stored without negatively impacting the regulation of core genes and their products. In *Vibrionaceae,* the new pole is enriched in shell and cloud genes (**Paper 1**), which also account for the majority of identified horizontally transferred genes in *Vibrio* (**Paper 3**). To be able to live in various environments and to adapt to changing conditions, bacteria need to acquire and maintain a diverse repertoire of genes (Bobay and Ochman 2017). Vibrios are known for its highly plastic genomes and the ability to acquire a diverse repertoire of genes (Lin et al. 2018; Escudero and Mazel 2017). Hence, the ability to insert new genes on both the chromid and the lower part of the chromosome may constitute a significant advantage, and a possible reason for the ecological success of vibrios. And, as a further consequence of the skewed gene distribution, the subcellular new pole area might serve as a gene reservoir where initially "useless" genes can be stored long enough until they serve a selective advantage and therefore are kept. The potential benefit of horizontally transferred genes will likely change depending on changes in the surrounding environment (Oliveira et al. 2017). This is supported by observations showing that less beneficial genes can be preserved in large populations (compared to smaller populations) (Bobay et al. 2012), and that HTGs with neutral or deleterious effects on the host can be retained in genomes as long as they do not disrupt essential gene functions {Formatting Citation}.

Other potential factors that might contribute to give cells with separated pangene categories a selective advantage in the battle for survival are: i) Difference in cytosolic chromatin density. During slow growth, the density of chromatin is likely to be higher in the new pole, whereas during fast growth, there may be a large amount of chromatin throughout the cell as multiple rounds of replication result in more DNA in the old pole. ii) The abundance of transcription and translation related proteins, such as RNAP and ribosomes, may vary with growth phases and in the intracellular space. For example, accumulation of RNAP, in so called transcription foci, has been observed in proximity to *ori* in several bacteria during fast growth (Jin et al. 2018). iii) Recent evidence suggest that molecular crowding are important for bacterial fitness (Soler-Bistué et al. 2020) and that molecular crowding can vary in different regions of the intracellular space (Berg et al. 2017). This will most likely affect diffusion of proteins and RNA, as well as their interactions and collisions with other molecules. All of the aforementioned factors, as well as others, may have an uneven impact on the two regions, resulting in different intracellular environments in the two poles and possibly evolving and maintaining this genome structure.

Based on the spatial placement of chromosome and chromid in *V. cholerae*, core/softcore genes and shell/cloud genes are separated into the old and the new pole respectively. This separation of gene types into different subcellular regions is likely advantageous, as it might allow for separate regulation of core/softcore genes. Furthermore, it might provide the bacteria with a reservoir of HTGs in the new pole, which can be used to adapt to changing conditions and niche specialization.

6 CONCLUDING REMARKS

In this thesis I have presented the main findings of the work I have done during my time as a PhD student. It has been a fascinating and rewarding journey, and a few moments stand out as particularly memorable. One of these moments was when we had just finished the analysis that showed that there was a significant gene distribution *Vibrionaceae*. When we then found the papers about the intracellular location of the chromosome and chromid in *V. cholerae*, it felt like we found the missing piece of the puzzle, which laid the groundwork for our hypotheses and further research. It was also exciting to see how, as I finished more and more of the gene distribution analysis, a similar gene distribution in *Pseudoalteromonas*, as in *Vibrionaceae*, gradually emerged.

Although we have gained new knowledge about the genome structure of multipartite bacteria, new questions have emerged. For example, I am curious whether the gene distribution patterns we have discovered can be found in other multipartite bacteria, or if their genes are organized differently. I am also eagerly awaiting the resolution of the spatial arrangement of the chromid and chromosome in *Pseudoalteromonas*, as I am curious of how the gene distribution corresponds to the spatial arrangement. Furthermore, although not discussed in this thesis, we did put forward a hypothesis in **Paper 1** suggesting that genes are spatially located close to the site of function of their products. This hypothesis would have been interesting to investigate further using both bioinformatics and laboratory experiments.

Overall, I hope that this study will serve as an inspiration to other researchers, both in terms of research on multipartite bacteria and the holistic use of pangenome analysis.

7 REFERENCES

Achaz G, Coissac E, Netter P, Rocha EPC. 2003. Associations between inverted repeats and the structural evolution of bacterial genomes. Genet. Soc. Am. Assoc. 164:1279–1289.

Almalki F, Choudhary M, Azad RK. 2023. Analysis of multipartite bacterial genomes using alignment free and alignment-based pipelines. Arch. Microbiol. 205:25.

Ayala JC, Wang H, Benitez JA, Silva AJ. 2015. RNA-Seq analysis and whole genome DNA-binding pro file of the *Vibrio cholerae* histone-like nucleoid structuring protein (H-NS). Genom. Data. 5:147–150.

Ayala JC, Wang H, Silva AJ, Benitez JA, Control CD. 2015. Repression by H-NS of genes required for the biosynthesis of the *Vibrio cholerae* biofilm matrix is modulated by the second messenger cyclic diguanylic acid. Mol. Microbiol. 97:630–645.

Badrinarayanan A, Le TBK, Laub MT. 2015. Bacterial chromosome organization and segregation. Annu. Rev. Cell. Dev. Biol. 31:171–199.

Balsiger S, Ragaz C, Baron C, Narberhaus F. 2004. Replicon-specific regulation of small heat shock genes in *Agrobacterium tumefaciens*. J. Bacteriol. 186:6824–6829.

Baumann L, Baumann P, Mandel M, Allen RD. 1972. Taxonomy of aerobic marine eubacteria. J. Bacteriol. 110:402–429.

Berg J Van Den, Boersma AJ, Poolman B. 2017. Microorganisms maintain crowding homeostasis. Nat. Publ. Gr. 15:309–318.

Le Berre D, Reverchon S, Muskhelishvili G, Nasser W. 2022. Relationship between the chromosome structural dynamics and gene expression; A chicken and egg dilemma? Microorg. 10:846.

Bervoets I, Charlier D. 2019. Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. FEMS Microbiol. Rev. 43:304–339.

Blattner FR et al. 1997. The complete genome sequence of *Escherichia coli* K-12. Science. 277:1453–1462.

Bobay L, Ochman H. 2017. The evolution of bacterial genome architecture. Front. Genet.

8:72.

Bobay L, Rocha EPC, Touchon M. 2012. The adaptation of temperate bacteriophages to their host genomes. Mol. Biol. Evol. 30:737–751.

Boor KJ. 2006. Bacterial stress responses: what doesn’t kill them can make then stronger. PLoS Biol. 4:e24.

Bosi E et al. 2017. The pangenome of (Antarctic) *Pseudoalteromonas* bacteria: Evolutionary and functional insights. BMC Genomics. 18:93.

Brown SD et al. 2014. Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia. Biotechnol. Biofuels. 7:40.

Carboni GP. 2021. The enigma of Pacini ’ s *Vibrio cholerae* discovery. J. Med. Microbiol. 70.

Chen X, Zhang Y-Z, Gao P-J, Luan X-W. 2003. Two different proteases produced by a deep-sea psychrotrophic bacterial strain, *Pseudoaltermonas* sp. SM9913. Mar. Biol. 143:989–993.

Choudhary, M., Cho, H., Bavishi, A., Trahan, C., Myagmarjav B. 2012. Evolution of multipartite genomes in prokaryotes. In: Evolutionary biology: Mechanisms and trends. Pontarotti, P. editor. Springer, Berlin, Heidelberg.

Chun J et al. 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae.* Proc. Natl. Acad. Sci. 106:15442–15447.

Comandatore F et al. 2019. Gene composition as a potential barrier to large recombinations in the bacterial pathogen *Klebsiella pneumoniae*. Genome Biol. Evol. 11:3240–3251.

Cooper S, Helmstetter CE. 1968. Chromosome replication and the division cycle of *Escherichia coli* B/r. J. Mol. Biol. 31:519–540.

Cooper Vaughn S., Vohr SH, Wrocklage SC, Hatcher PJ. 2010. Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput. Biol. 6:e1000732.

Costa SS, Guimarães LC, Silva A, Soares SC, Baraúna RA. 2020. First steps in the analysis of prokaryotic pan-genomes. Bioinform. Biol. Insights. 14:1177932220938064.

Couturier E, Rocha EPC. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol. Microbiol. 59:1506–1518.

Dame RT, Rashid FZM, Grainger DC. 2020. Chromosome organization in bacteria: mechanistic insights into genome structure and function. Nat. Rev. Genet. 21:227–242.

Daruvar A De, Collado-vides J, Valencia A. 2002. Analysis of the cellular functions of *Escherichia coli* operons and their conservation in *Bacillus subtilis*. J. Mol. Evol. 55:211–221.

David A et al. 2014. The two cis-acting sites, *parS1* and o*riC1*, contribute to the longitudinal organisation of *Vibrio cholerae* chromosome I. PLoS Genet. 10:e1004448.

DiCenzo GC, Finan TM. 2017. The divided bacterial genome. Microbiol. Mol. Biol. Rev. 81:e00019-17.

diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. 2014. Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLoS Genet. 10:e1004742.

Dicenzo GC, Mengoni A, Perrin E. 2019. Chromids aid genome expansion and functional diversification in the family *Burkholderiaceae*. Mol. Biol. Evol. 36:562–574.

Dillon KP, Correa F, Judon C, Sancelme M. 2021. Cyanobacteria and algae in clouds and rain in the area of puy de Dôme, Central France. Appl. Environ. Microbiol. 87:e01850-20.

Donnell MO, Langston L, Stillman B. 2013. Principles and concepts of DNA replication in Bacteria, Archaea, and Eukarya. Cold Spring Harb. Perspect. Biol. 5:a010108.

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.

Duigou S´ephane et al. 2006. Independent control of replication Initiation of the two *Vibrio cholerae* chromosomes by DnaA and RctB. J. Bacteriol. 188:6419–6424.

Escudero JA, Mazel D. 2017. Genomic plasticity of *Vibrio cholerae*. Int. Microbiol. 20:138–148.

Esin A, Ellis T, Warnecke T. 2018. Horizontal gene flow into *Geobacillus* is constrained by the chromosomal organization of growth and sporulation. bioRxiv. doi:

10.3389/fmicb.2016.00723

Federhen S. 2012. The NCBI Taxonomy database. 40:D136–D143.

Feng Z et al. 2021. The second chromosome promotes the adaptation of the genus *Flammeovirga* to complex environments. Microbiol. Spectr. 9:e00980-21.

Fleischmann RD et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 269:496–512.

Fogel MA, Waldor MK. 2005. Distinct segregation dynamics of the two *Vibrio cholerae* chromosomes. Mol. Microbiol. 55:125–136.

Forrest D, Warman EA, Erkelens AM, Dame RT, Grainger DC. 2022. Xenogeneic silencing strategies in bacteria are dictated by RNA polymerase promiscuity. Nat. Commun. 13:1149.

Fournes F et al. 2021. The coordinated replication of *Vibrio cholerae* ' s two chromosomes required the acquisition of a unique domain by the RctB initiator. Nucleic Acids Res. 49:11119–11133.

Fournes F, Val M-E, Skovgaard O, Mazel D. 2018. Replicate once per cell cycle: replication control of secondary chromosomes. Front. Microbiol. 9:1–1833.

Galardini M, Pini F, Bazzicalupo M, Biondi EG, Mengoni A. 2013. Replicon-dependent bacterial genome evolution: The case of *Sinorhizobium meliloti*. Genome Biol. Evol. 5:542–558.

Galli E et al. 2019. Replication termination without a replication fork trap. Sci. Rep. 9:8315.

Gao Y et al. 2020. Reconstruction of Fur pan-regulon uncovers the complexity and diversity of transcriptional regulation in *E. coli*. bioRxiv. doi: 10.1101/2020.05.21.109694.

Gauthier G, Gauthier M, Christen R. 1995. Phylogenetic analysis of the genera *Alteromonas*, *Shewanella*, and *Moritella* using genes coding for small-subunit rRNA sequences and division of the genus *Alteromonas* into two genera, *Alteromonas* (emended) and *Pseudoalteromonas* gen. nov., and proposal of twelve new species combinations Int. J. Syst. Bacteriol. 45:755–761.

Gibson KE, Kobayashi H, Walker GC. 2009. Molecular determinants of a symbiotic chronic infection. Annu. Rev. Genet. 42:413–441.

Gogou C, Japaridze A, Dekker C, Leonard A. 2021. Mechanisms for chromosome segregation in Bacteria. Front. Microbiol. 12:685687.

Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. 2020. Pangenomics comes of age: from bacteria to plant and animal applications. Trends Genet. 36:132–145.

Harrison PW, Lower RPJ, Kim NKD, Young JPW. 2010. Introducing the bacterial ‘chromid’: not a chromosome, not a plasmid. Trends Microbiol. 8:141–148.

Heidelberg JF et al. 2000. DNA sequence of both chromosomes of the cholera pathogen V*ibrio cholerae*. Nature. 406:477–483.

Higgins NP, Yang X, Fu Q, Roth JR. 1996. Surveying a supercoil domain by using the γδ resolution system in *Salmonella typhimurium*. J. Bacteriol. 178:2825–2835.

Hogg JS et al. 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:R103.

Hong M et al. 2020. RNA sequencing: new technologies and applications in cancer research. J. Hematol. Oncol. 13:166.

Ii RT, Ibba M. 2020. Translational regulation of environmental adaptation in bacteria. JCB Rev. 295:10434–10445.

Ingerson-mahar M, Gitai Z. 2012. A growing family: the expanding universe of the bacterial cytoskeleton. FEMS Microbiol. Rev. 36:256–266.

Jin DJ, Martin CM, Sun Z, Cagliero C, Zhou YN. 2018. Nucleolus-like compartmentalization of the transcription machinery in fast-growing bacterial cells. Crit. Rev. Biochem. Mol. Biol. 52:96–106.

Kahlke T, Goesmann A, Hjerde E, Willassen NP, Haugen P. 2012. Unique core genomes of the bacterial family *Vibrionaceae* : insights into niche adaptation and speciation. BMC Genomics. 13:1.

Kopejtka K et al. 2019. Clustered core- and pan-genome content on *Rhodobacteraceae* chromosomes. Genome Biol. Evol. 11:2208–2217.

Kunst F et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus* *subtilis*. Nature. 390:249–256.

Lal A et al. 2016. Genome scale patterns of supercoiling in a bacterial chromosome. Nat. Commun. 7:11055.

Larotonda L, Mornico D, Khanna V, Bernal-bayard J. 2023. Chromosomal position of ribosomal protein genes affects of V*ibrio cholerae*. mBio. 14:e0343222.

Le TBK, Imakaev M V., Mirny LA, Laub MT. 2013. High-resolution mapping of the spatial organization of a bacterial chromosome. Science. 342:731–734.

Le TBK, Laub MT. 2016. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. 35:1582–1595.

Levin PA, Angert ER. 2015. Small but mighty: Cell size and bacteria. Cold Spring Harb. Perspect. Biol. 7:a019216.

Liao L et al. 2019. Multipartite genomes and the sRNome in response to temperature stress of an Arctic *Pseudoalteromonas fuliginea* BSW20308. Environ. Microbiol. 21:272–285.

Lilburn TG, 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.

Lin H, Yu M, Wang X, Zhang XH. 2018. Comparative genomic analysis reveals the evolution and environmental adaptation strategies of vibrios. BMC Genomics. 19:135.

Maansson M et al. 2016. An integrated metabolomic and genomic mining workflow to uncover the biosynthetic potential of bacteria. mSystems. 1:e00028-15.

Martis B. S, Forquet R, Reverchon S, Nasser W, Meyer S. 2019. DNA supercoiling : an ancestral regulator of gene expression in pathogenic bacteria? Comput. Struct. Biotechnol. J. 17:1047–1055.

McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat. Microbiol. 2:17040.

Médigue C et al. 2005. Coping with cold: The genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. Genome Res. 15:1325–1335.

Medini D, Donati C, Rappuoli R, Tettelin H. 2020. The Pangenome: A data-driven discovery in biology. In: The pangenome: diversity, dynamics and evolution of genomes. Medini D, Tettelin H editors. Springer pp. 3–20.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr. Opin. Genet. Dev. 15:589–594.

Misra HS, Maurya GK, Kota S. 2018. Maintenance of multipartite genome system and its functional significance. J. Genet. 97:1013–1038.

Navarre WW et al. 2006. Selective silencing of foreign DNA protein in *Salmonella*. Science. 313:236–239.

O'leary NA et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733-745.

Oliveira PH, Touchon M, Cury J, Rocha EPC. 2017. The chromosomal organization of horizontal gene transfer in bacteria. Nat. Commun. 8:841.

Onohuean H, Agwu E, Nwodo UU. 2022. A global perspective of *Vibrio* species and associated diseases: Three-decade seta-Synthesis of aesearch Advancement. Environ. Health Insights. 16:1–14.

Oren A, Garrity GM. 2021. Valid publication of the names of forty-two phyla of prokaryotes. Int. J. Syst. Evol. Microbiol. 71:005056.

Ott E et al. 2020. Molecular repertoire of *Deinococcus radiodurans* after 1 year of exposure outside the International Space Station within the Tanpopo mission. Microbiome. 8:150.

Park C, Zhang J. 2012. High expression hampers horizontal gene transfer. In: Genome Biology and Evolution. 4:523–532.

Park S et al. 2019. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. Front. Microbiol. 10:834.

Parks DH et al. 2022. GTDB : an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 50:D785–D794.

Parlikar A, Kalia K, Sinha S, Patnaik S. 2020. Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2. PeerJ. 8:e9576.

Parrilli E, Tedesco P, Fondi M, Luisa M. 2021. The art of adapting to extreme environments: The model system *Pseudoalteromonas*. Phys. Life Rev. 36:137–161.

Peter BJ et al. 2004. Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. Genome Biol. 5:R87.

Piette F et al. 2011. Life in the cold : a proteomic study of cold-repressed proteins in the Antarctic bacterium P*seudoalteromonas haloplanktis* TAC125. Appl. Environ. Microbiol. 77:3881–3883.

Postow L, Hardy CD, Arsuaga J, Cozzarelli NR. 2004. Topological domain structure of the *Escherichia coli* chromosome. Genes Dev. 18:1766–1779.

Quick J, Quinlan AR, Loman NJ. 2015. Erratum: A reference bacterial genome dataset generated on the MinION TM portable single-molecule nanopore sequencer. Gigascience. 4:6.

Rasmussen T, Jensen RB, Skovgaard O. 2007. The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle. EMBO J. 26:3124–3131.

Richards GP et al. 2017. Mechanisms for *Pseudoalteromonas piscicida*-induced killing of Vibrios and other bacterial pathogens. Appl. Environ. Microbiol. 83:e00175–17.

Rocha EPC. 2004. The replication-related organization of bacterial genomes. Microbiology. 150:1609–1627.

Rong JC et al. 2016. Complete genome sequence of a marine bacterium with two chromosomes, *Pseudoalteromonas translucida* KMM 520T. Mar. Genomics. 26:17–20.

Rouli L, Merhej V, Fournier P, Raoult D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect. 7:72–85.

Sannino F et al. 2017. A novel synthetic medium and expression system for subzero growth and recombinant protein production in *Pseudoalteromonas haloplanktis* TAC125. Appl. Microbiol. Biotechnol. 101:725–734.

Sherratt DJ. 2003. Bacterial chromosome dynamics. Science. 301:780–785.

Slager J, Veening JW. 2016. Hard-wired control of bacterial processes by chromosomal gene location. Trends Microbiol. 24:788–800.

Slater SC et al. 2009. Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. J. Bacteriol. 191:2501–2511.

Soler-Bistué A et al. 2015. Genomic location of the major ribosomal protein gene locus determines *Vibrio cholerae* global growth and infectivity. PLoS Genet. 11:e1005156.

Soler-Bistué A et al. 2020. Macromolecular crowding links ribosomal protein gene dosage to growth rate in *Vibrio cholerae*. BMC Biol. 18:43.

Soler-Bistué A, Timmermans M, Mazel D. 2017. The proximity of ribosomal protein genes to oric enhances *Vibrio cholerae* fitness in the absence of multifork replication. mBio. 8:e00097-17.

Srivastava P, Chattoraj DK. 2007. Selective chromosome amplification in *Vibrio cholerae*. Mol. Microbiol. 66:1016–1028.

Stokke C, Waldminghaus T, Skarstad K. 2011. Replication patterns and organization of replication forks in *Vibrio cholerae*. Microbiology. 157:695–708.

Surovtsev I, Jacobs-Wagner C. 2018. Subcellular organization: A critical feature of bacterial cell replication. Cell. 172:1271–1293.

Suwanto A, Kaplant S. 1989. Physical and genetic mapping of the R*hodobacter sphaeroides* 2 .4 .1 genome: Presence of two unique circular chromosomes. J. Bacteriol. 171:5850–5859.

Takemura AF, Chien DM, Polz MF. 2014. Associations and dynamics of *Vibrionaceae* in the environment, from the genus to the population level. Front. Microbiol. 5:38.

Tettelin H et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome'. Proc. Natl. Acad. Sci. U. S. A. 102:13950–13955.

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

Thompson FL, Iida T, Swings J. 2004. Biodiversity of Vibrios. Microbiol. Mol. Biol. Rev. 68:403–431.

Toffano-Nioche C et al. 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. 18:2201–2219.

Touchon M, Rocha EPC. 2016. Coevolution of the organization and structure of prokaryotic genomes. Cold Spring Harb. Perspect. Biol. 8:a018168.

Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7:e1001284.

Trucksis M, Michalski J, Deng YK, Kaper JB. 1998. The *Vibrio cholerae* genome contains two unique chromosomes. Proc. Natl. Acad. Sci. U. S. A. 95:14464–14469.

Tuller T et al. 2011. Association between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acids Res. 39:4743–4755.

Val M-E et al. 2016. A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. Sci. Adv. 2:e1501914.

Valens M, Penaud S, Rossignol M, Cornet F, Boccard F. 2004. Macrodomain organization of the *Escherichia coli* chromosome. EMBO J. 23:4330–4341.

Venkova-Canova T, Chattoraj DK. 2011. Transition from a plasmid to a chromosomal mode of replication entails additional regulators. Proc. Natl. Acad. Sci. 108:6199–6204.

Volff J, Josef A. 2000. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. FEMS Microbiol. Lett. 186:143–150.

Wang X, Liu X, Possoz C, Sherratt DJ. 2006. The two *Escherichia coli* chromosome arms locate to separate cell halves. Genes Dev. 20:1727–1731.

Wang X, Rudner DZ. 2015. Spatial organization of bacterial chromosomes. Curr. Opin. Microbiol. 22:66–72.

Weinstock MT, Hesek ED, Wilson CM, Gibson DG. 2016. *Vibrio natriegens* as a fast-growing host for molecular biology. Nat. Methods. 13:849–851.

Westoby M et al. 2021. Cell size, genome size, and maximum growth rate are near-independent dimensions of ecological variation across bacteria and archaea. Ecol. Evol. 11:3956–3976.

White O et al. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science. 286:1571–7.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: The unseen majority. Proc. Natl. Acad. Sci. 95:6578–6583.

Xie B Bin et al. 2021. Evolutionary trajectory of the replication mode of bacterial replicons. mBio. 12:e02745-20.

Yamaichi Y, Iida T, Park K. 1999. Physical and genetic map of the genome of *Vibrio parahaemolyticus*: presence of two chromosomes in *Vibrio* species. Mol. Microbiol. 31:1513–1521.

Yildirim A, Feig M. 2018. High-resolution 3D models of *Caulobacter crescentus* chromosome reveal genome structural variability and organization. Nucleic Acids Res. 46:3937–3952.

Yubero P, Poyatos JF. 2020. The impact of global transcriptional regulation on bacterial gene order. iScience. 23:101029.

Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S. 2002. Computational identification of operons in microbial genomes. Genome Res. 12:1221–1230.

Zhu S, Kojima S, Homma M, Visick KL. 2013. Structure, gene regulation and environmental response of flagella in *Vibrio*. Front. Microbiol. 4:410.

PAPER 1

Sonnenberg *et al. BMC Genomics* (2020) 21:695
https://doi.org/10.1186/s12864-020-07117-5

BMC Genomics

RESEARCH ARTICLE Open Access

Vibrionaceae core, shell and cloud genes are non-randomly distributed on Chr 1: An hypothesis that links the genomic location of genes with their intracellular placement

Cecilie Bækkedal Sonnenberg[1], Tim Kahlke[2] and Peik Haugen[1*]

Abstract

Background: The genome of *Vibrionaceae* bacteria, which consists of two circular chromosomes, is replicated in a highly ordered fashion. In fast-growing bacteria, multifork replication results in higher gene copy numbers and increased expression of genes located close to the origin of replication of Chr 1 (*ori1*). This is believed to be a growth optimization strategy to satisfy the high demand of essential growth factors during fast growth. The relationship between *ori1*-proximate growth-related genes and gene expression during fast growth has been investigated by many researchers. However, it remains unclear which other gene categories that are present close to *ori1* and if expression of all *ori1*-proximate genes is increased during fast growth, or if expression is selectively elevated for certain gene categories.

Results: We calculated the pangenome of all complete genomes from the *Vibrionaceae* family and mapped the four pangene categories, core, softcore, shell and cloud, to their chromosomal positions. This revealed that core and softcore genes were found heavily biased towards *ori1*, while shell genes were overrepresented at the opposite part of Chr 1 (i.e., close to *ter1*). RNA-seq of *Aliivibrio salmonicida* and *Vibrio natriegens* showed global gene expression patterns that consistently correlated with chromosomal distance to *ori1*. Despite a biased gene distribution pattern, all pangene categories contributed to a skewed expression pattern at fast-growing conditions, whereas at slow-growing conditions, softcore, shell and cloud genes were responsible for elevated expression.

Conclusion: The pangene categories were non-randomly organized on Chr 1, with an overrepresentation of core and softcore genes around *ori1*, and overrepresentation of shell and cloud genes around *ter1*. Furthermore, we mapped our gene distribution data on to the intracellular positioning of chromatin described for *V. cholerae,* and found that core/softcore and shell/cloud genes appear enriched at two spatially separated intracellular regions. Based on these observations, we hypothesize that there is a link between the genomic location of genes and their cellular placement.

Keywords: Pangenome, Genome architecture, *Vibrionaceae*, *Aliivibrio salmonicida*, *Vibrio natriegens*, Gene dosage

* Correspondence: peik.haugen@uit.no
[1]Department of Chemistry and Center for Bioinformatics (SfB), Faculty of Science and Technology, UiT The Arctic University of Norway, N-9037 Tromsø, Norway
Full list of author information is available at the end of the article

Background

Bacteria that belong to the family *Vibrionaceae* are rich in most aqueous habitats, from the deep seas to fresh and brackish waters, and in temperature zones ranging from the polar to tropical areas. They exist as free-swimming cells or associated with other organisms, either in a symbiotic relationship or as pathogens of e.g. fish, corals and even humans [1, 2]. Despite the notorious reputation of some *Vibrionaceae* species, (e.g., *Vibrio cholerae* and *Vibrio vulnificus)* it is the diversity of non-pathogenic *Vibrionaceae* species that makes these bacteria so successful and ecologically important [3]. The facultative anaerobic bacterium *Vibrio natriegens,* for example, fixes atmospheric nitrogen (N_2) into ammonia (NH_3), and thus provides its surroundings with a critical nutrient [4].

As of April 2020, the RefSeq database contains 306 complete *Vibrionaceae* genomes (representing 57 species), with genomes from new species being added on a regular basis. One characteristic feature shared by almost all *Vibrionaceae* genomes is a highly unusual bipartite structure consisting of a large (Chr 1) and a smaller (Chr 2) chromosome [5, 6]. It is proposed that bacteria with bipartite genomes have a selective advantage for the adaptation to very different environmental conditions [7], and that division into multiple smaller replicons may reduce replication time, thus allowing for faster generation time and a competitive advantage [8, 9]. The unconventional genome constellation is expected to require tightly regulated and synchronized replication to ensure proliferation and control of gene expression during changes in the surrounding environment.

In *V. cholerae,* replication of Chr 1 and Chr 2 is highly coordinated [10]. When the replication fork approaches *crtS* in Chr 1 (Chr 2 replication triggering site), a hitherto unknown mechanism triggers replication of Chr 2 [11, 12]. Interestingly, there is a short pause (corresponding to replication of approx. 200 kbp) between the *crtS* replication and the initiation of Chr 2 replication. The exact function of this pause is yet unknown, but it is hypothesized to be needed for activation of the *rctB* (Chr 2's own replication initiator) and *ori2* initiation system [12]. In other words, the chromosomal position of *crtS* and the pause contribute to synchronize termination of Chr 1 and Chr 2 replication. Furthermore, the synchronized termination is likely linked to coordination of chromosome segregation and cell division [12].

Another intriguing phenomenon regarding replication of *Vibrio* genomes is that genes surrounding *ori* can be found in multiple copies during the replication process due to successive initiations of replication from *ori* (i.e., multifork replication) [13, 14]. This phenomenon is a hallmark of fast-growing bacteria, such as *V. cholerae* and *V. natriegens,* and is believed to be a growth optimization strategy to satisfy the high demand of essential growth factors during fast growth [15–17]. Using an elegant genetic approach, Soler-Bistué et al. (2015) showed that by relocating the major ribosomal protein gene locus (*s10-spec-α*) of *V. cholerae* further away from *ori1*, growth rate, the gene copy number and mRNA abundance of this cluster were reduced [18]. The authors concluded that there is a strong correlation between chromosomal gene position and effects on the bacterial physiology. Later, the same model system (i.e., *V. cholerae* with relocated *s10-spec-α* locus) was used to study effects on bacterial fitness under slow growth conditions (i.e., no multifork replication) [19]. One conclusion from this study was that bacterial fitness was reduced when the *s10-spec-α* locus was located distal to *ori1,* which demonstrates that genomic positioning of ribosomal protein genes not only affects growth, but also cell fitness across the whole life cycle. In a recent study, Soler-Bistué et al. (2020) showed that relocation of the *s10-spec-α* locus lead to higher cytoplasm fluidity and the authors suggested that changes in the macromolecular crowding of the cytoplasm impacts the cellular physiology of *V. cholerae.* Interestingly, the protein production capacity in *V. cholerae* was independent of the position of the *s10-spec-α* locus [20].

In an interesting approach, Dryselius et al. (2008) used qPCR and microarray to study how copy numbers of genes vary across the entire genome of several *Vibrio* species (*V. parahaemolyticus, V. cholerae* and *V. vulnificus*) under different growth conditions, and then monitored how the data correlated with gene expression levels (also using microarray) [21]. The authors found greatest differences in gene copy numbers across Chr 1 compared to Chr 2 when grown in a rich medium. In general, the trend is that gene copy numbers increase from the terminus towards the origin of replication, and that this increase is reflected by increasing gene expression levels. The same trend was not found for slow-growing bacteria (i.e., when grown in minimal medium). Also, for Chr 2 gene expression levels were low and apparently independent of gene copy number effect. Similar findings were later described in *V. splendidus* [22]. Here, genes located on Chr 1 were 3.6 × more expressed compared to those located on Chr 2, and the highest expression values were typically associated with genes surrounding the origin of replication on Chr 1.

In summary, the genome of *Vibrionaceae* bacteria, which consists of two circular chromosomes, is replicated in a highly ordered fashion. In fast-growing bacteria, replication results in higher gene copy numbers, and increased expression of genes located close to the origin of replication of Chr 1. That the expression of growth-related genes located close to *ori1* is elevated during fast growth is known, but a general picture of which gene types are found close to *ori1,* and how expression of each gene type is affected, is however not known. To address this knowledge gap we revisited the intriguing topic of genome architecture in *Vibrionaceae*. In a pangenome approach we used available genomes to calculate and divide clusters of orthologous genes into the main categories "core", "softcore", "shell" (accessory) and "cloud" (unique), and used this information to determine how the corresponding genes are distributed on Chr 1 and Chr 2 of selected *Vibrionaceae* genomes. Data from publicly available gene expression experiments was mapped back to the pangenes to determine gene expression profiles under different environmental conditions such as expression data from the fast-growing bacterium *V. natriegens* grown under optimal or minimal growth conditions, and data from the fish-pathogen *Aliivibrio salmonicida* grown under salt concentration and temperature that mimics the physiological conditions during infection. Our results show a non-random distribution of genes on the two chromosomes of *Vibrionaceae*. The gene distribution was then compared with global gene expression trends, and we find a strong correlation between expression levels and distance from *ori1.* Surprisingly, despite a biased gene distribution pattern, all pangene categories contribute to a skewed expression pattern at fast-growing conditions. Finally, based on our data we propose an hypothesis that describes how pangenes are spatially distributed inside *Vibrionaceae* bacterial cells, and we discuss possible implications of the proposed hypothesis.

Results

Pangenome calculations based on 124 complete *Vibrionaceae* genomes identifies 710 clusters of orthologous core genes

To categorize all genes associated with *Vibrionaceae* genomes into distinct classes, we downloaded all complete genomes from the NCBI RefSeq database (124 as of May 2018, see Additional file 1), and then used GET_HOMOLOGUES v3.1.0 [23] to cluster orthologous protein sequences based on the OrthoMCL algorithm. The pangenome calculations identified a total of 61,512 clusters, of which 710 were encoded by genes found in all 124 genomes (i.e., core genes). The remaining clusters are distributed among softcore (encoded by ≥117 genomes), shell (encoded by 116 ≤ and ≥ 3 genomes) and cloud (encoded by ≤2 genomes), and contain 1796, 14,642 and 45,074 clusters, which represents 3, 23 and 73% of the total clusters, respectively. In individual genomes, core gene clusters represent 1.2% of the pangenome, and comprise 10—17% of the total genes. Similarly, softcore constitutes 24—34% (1489—1796 genes per genome) of the total genes.

Core and softcore genes densely populate the upper half of Chr 1

The four gene categories core, softcore, shell and cloud, were next mapped to their chromosomal locations to investigate whether they are randomly or non-randomly distributed on each chromosome. First, genes of eleven selected *Vibrionaceae* representatives (see Additional file 2 for phylogeny of the 11 genomes) were classified as either upper or lower (i.e., upper or lower half of the chromosome) based on their chromosomal location on Chr 1 and Chr 2 in relation to their distance of the origin of replication. As presented in Fig. 1 (complete table of pangene distribution is available as Additional file 3 and chi-squared test is available as Additional file 4), core and softcore genes are significantly overrepresented (adjusted chi-square *P*-value ≤0.05) in the upper half of Chr 1 in all investigated genomes. Similarly, shell and cloud genes on Chr 1 are significantly overrepresented (adjusted chi-square P-value ≤0.05) in the lower half of Chr 1 in 8 genomes, thus supporting a non-random distribution of genes on Chr 1. In contrast to Chr 1, genes of all categories are much more evenly distributed on Chr 2. Although shell, cloud and softcore genes show non-random distribution on Chr 2 in some of the investigated genomes (softcore 3/11, shell 1/11, cloud 2/11), the majority of genomes show no significant bias (adjusted chi-square *P*-value ≤0.05). Furthermore, core genes were not significantly overrepresented in either lower or upper half of Chr 2 in any of the genomes.

To provide a more fine-grained picture of the core (710—721) and shell (749—2753) gene distributions, we plotted the distribution of core and shell genes on Chr 1 and Chr 2 of eleven *Vibrionaceae taxa* using the genome comparison tool Circos [24] (Fig. 2). Each plot was centered on *mioC* (Chr 1) and *rctB* (Chr 2). Our results show that although the exact distribution pattern varies between species, the biased distributions of core and shell, as described above, are striking and readily visible with the naked eye. Interestingly, although core genes densely populate the upper half of Chr 1, the region immediately surrounding *ori1* contains very few core genes. This region (denoted "i" in Fig. 2) is, in contrast, densely populated by softcore genes (at least in *V. natriegens* and *A. salmonicida,* see section below). Also, a region (denoted "ii" in Fig. 2) of approximately 500 kb surrounding *ter1* is densely populated with shell genes (and hence sparsely populated with core genes). For Chr 2, the chi-square test supported no significant bias in gene distribution (Additional file 4), and Fig. 2b supports this general picture although some local clustering of gene categories will occur. In summary, the results presented here reveal that core, softcore, shell and cloud genes are non-randomly distributed on Chr 1. Core and softcore

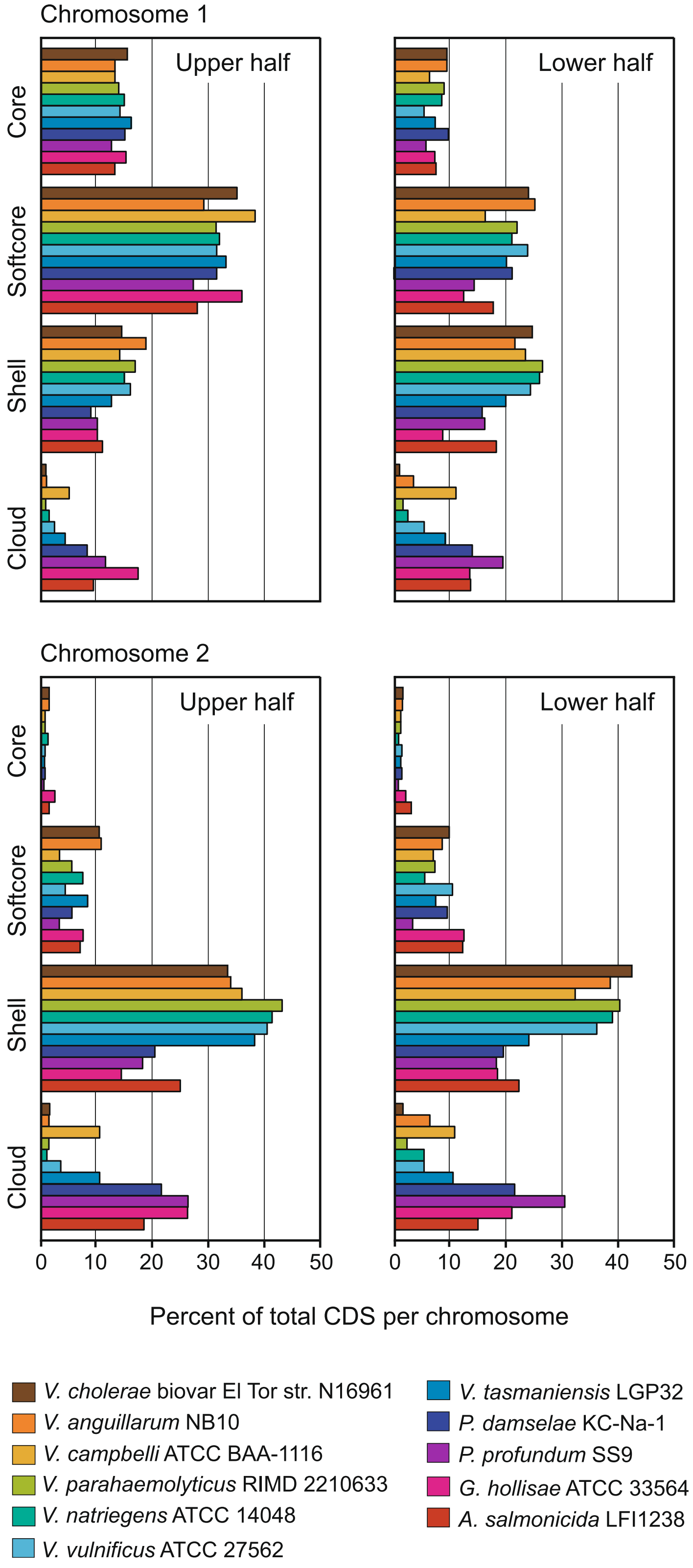

Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Distribution of the four pangene categories between upper and lower half of 11 *Vibrionaceae* genomes. Bars in the histogram show percent of total CDSs per chromosome for each pangene category. Core and softcore genes are overrepresented on the upper half of Chr 1, shell and cloud genes are overrepresented on the lower half. On Chr 2 the genes are more evenly distributed between the upper and lower halves of Chr 2

genes are more likely to be located on the upper half of Chr 1, whereas shell and cloud genes tend to be located closer to the replication terminator. For Chr 2, the distribution of the four pangene categories are in general randomly distributed showing locational bias only for a few genomes.

Expression levels of genes located on Chr 1 of *V. natriegens* and *A. salmonicida* generally correlate with distance to *ori1*

Figure 3 shows how core, softcore, shell and cloud pangenes are distributed on Chr 1 and Chr 2 of *V. natriegens* and *A. salmonicida.* The pattern is consistent with

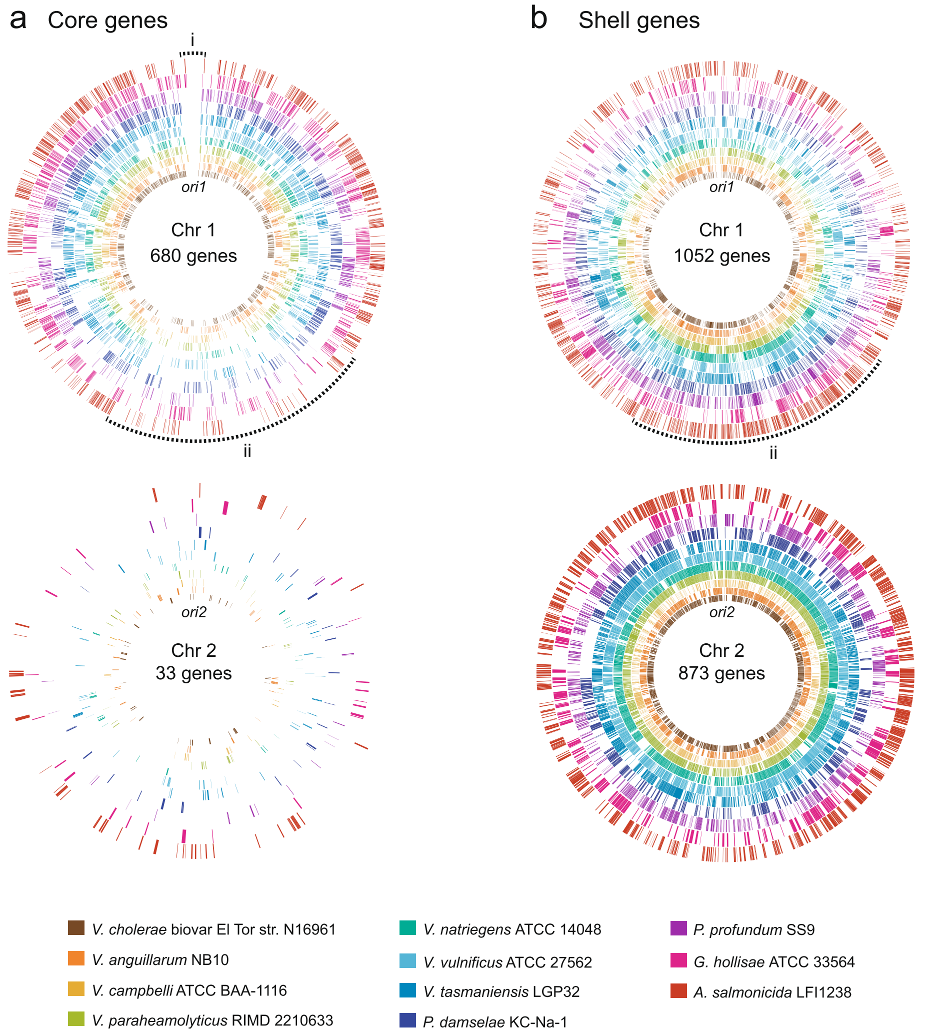

Fig. 2 Distribution of 710 core genes in 11 *Vibrio* genomes. Location of core (**a**) and shell (**a**) genes on Chr 1 and Chr 2 of 11 *Vibrionaceae* genomes. Circular plots are arranged regarding the phylogenetic relationship of the investigated isolates. Each plot is centered at a gene assumed to be close to the replication origin: *mioC* on Chr 1 and *rtcB* on Chr 2. As shown, a majority of core genes on Chr 1 is located closer to *ori1* than to ter. Shell genes show the opposite distribution pattern on Chr 1, where majority of shell genes accumulate closer to ter. On Chr 2 both core and shell genes are randomly distributed. The dashed line "i" indicates a region on Chr 1 surrounding *ori1* that contains very few core genes. The dashed line "ii" shows a region on Chr 1 of approximately 500 kb surrounding ter that is more sparsely populated with core genes than the rest of the chromosome

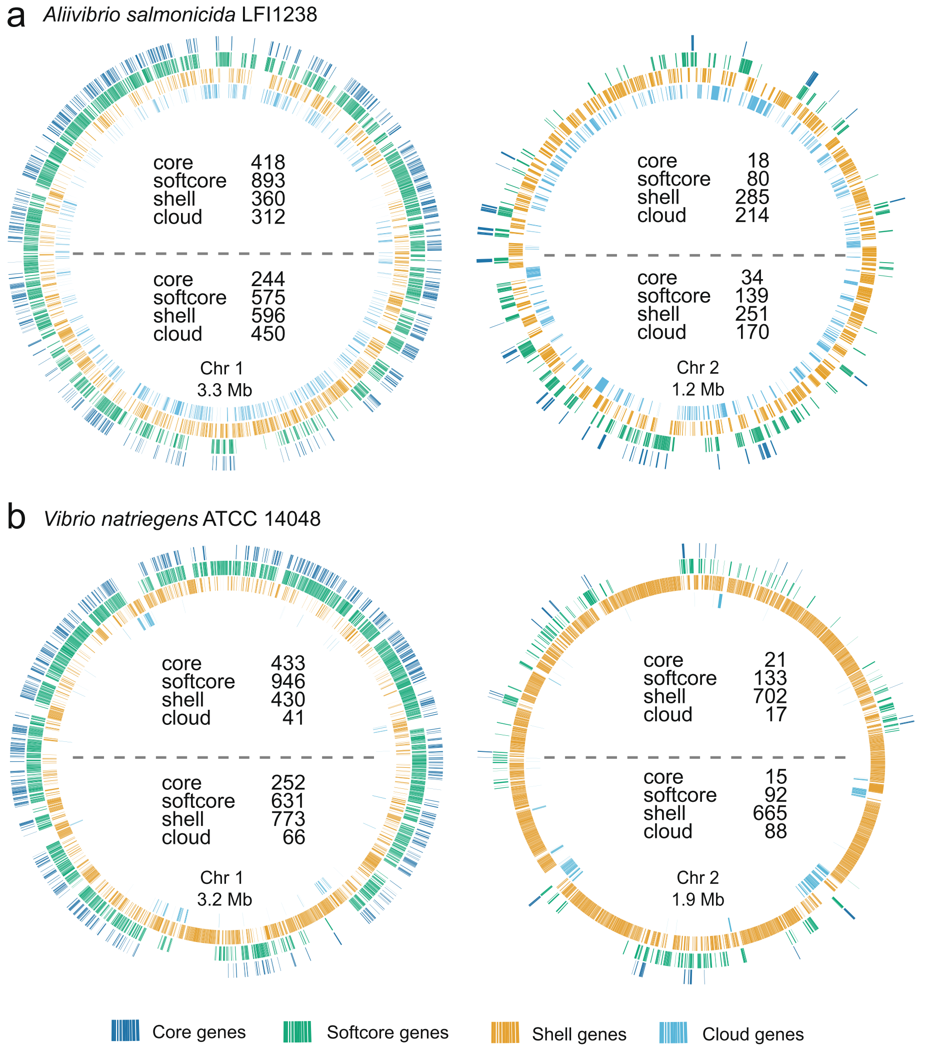

Fig. 3 Distribution of the four pangene categories on Chr 1 and Chr 2 for (**a**) *A. salmonicida* LFI1238 and (**b**) *V. natriegens* ATCC 14048. The number of genes in each pangene category in the upper and lower half is written inside each chromosome. A dashed line visualises the separation of the upper and lower half of the chromosomes

the biased gene distribution pattern described above, with core and softcore genes being overrepresented at the upper half of Chr 1, and shell and cloud genes being overrepresented at the lower half. The two species were chosen as models for comparison of gene expression data with pangene distribution patterns. Specifically, we were curious to examine if regions that are densely populated by core/softcore pangenes are expressed at high levels, compared to regions more sparsely populated by core/softcore pangenes. This expectation is based on previous data from *V. parahaemolyticus* and *V. cholerae,* which showed that growth rates have large impacts on the copy number (gene dosage) of genes located on Chr 1, as well as on gene expression levels [9, 10, 21]. Fast- and slow-growing bacterial representatives were therefore chosen for this particular comparative analysis. *V. natriegens* is a fast-growing bacterium commonly found in estuarine mud, with doubling times below 10 min at favourable conditions [25]. *A. salmonicida* is, in contrast, a slow growing *Vibrionaceae* bacterium, and the causative agent of cold-water vibriosis in e.g., Atlantic salmon and cod [26, 27]. To correlate gene distribution with gene expression data, publicly available RNA-seq data of *V.* natriegens and *A. salmonicida* were downloaded from the Sequence Read Archive [28] at NCBI. For *V. natriegens,* datasets from growth in minimal and optimal (rich) medium at 37 °C to mid log phase were chosen [29]. For *A. salmonicida*, a dataset originating from growth in LB medium containing 1% NaCl at 8 °C to mid log phase was used [30]. EDGE-PRO 1.3.1 [31] was used to align cDNA reads to the *V. natriegens* ATCC 14048 (NBRC 15636, DSM 759) (assembly no. GCA_001456255.1) or

A. salmonicida LFI1238 (assembly no. GCF_000196495.1) genome, and to calculate expression values as reads per kilobase per million (RPKM) for all protein coding sequences (CDS).

Figure 4 shows global expression maps of *V. natriegens* and *A. salmonicida* chromosomal genes centered around the median. Data points ($\log_2$ ratio RPKM CDS:RPKM median) for each CDS are shown, as well as a trend line averaged over a sliding window of 200 data points. For Chr 1 the general picture is similar in all three datasets, i.e., RPKM values are typically above the median value at the upper half (i.e., the region closest to the origin of replication), but lower at the region surrounding the terminus, independent of growth conditions. This is somewhat surprising since the observed expression patterns described above was expected for fast growing cultures (i.,e *V. natriegens* in rich medium), but not for slow growing cultures (i.e., *A. salmonicida* in LB 1% NaCl and 8 °C and *V. natriegens* in minimal medium, see Additional file 5). The rationale is that gene copy numbers (also known as "gene dosage"), and thus expression levels are expected to be correlated with growth rates/multifork replication [21].

A more detailed circular expression map is available in Additional file 6 and shows that region "i" (see Fig. 2), which encodes mostly softcore genes, contains a highly expressed proton-translocating ATP synthase (F_0F_1 class) gene cluster (atpIBEFHAGDC). The ATPase cluster is well described in *Escherichia coli* as an operon located 84 min on the chromosome (close to *oriC*), and with gene expression levels varying according to cell growth rate [32]. The ATP synthase cluster represents softcore genes, and are present in both bacteria. Moreover, the detailed map shows that region "ii", which is densely populated with shell genes, differs from the remaining lower half of Chr 1 by being expressed far below median in *V. natriegens* at both fast and slow growth conditions. For *A. salmonicida* the main picture is the same, but less pronounced, meaning that the majority of shell genes located in "ii" are expressed below median.

For Chr 2, the results are more ambiguous, although overall similar between minimal and rich growth. For *A. salmonicida,* expression around the terminus is, on average, higher compared to that of regions adjacent to *ori2*. For *V. natriegens*, expression is generally higher than median in regions surrounding the terminus, but varies across the remaining parts of Chr 2. Similar to Chr 1, little difference could be determined between the slow- and the fast-growing datasets of Chr 2.

In summary, we found that global expression levels for Chr 1, consistently correlate with the distance to the origin of replication. The log2 ratio of RPKM CDS:RPKM median decreases as the distance from origin of replication increases.

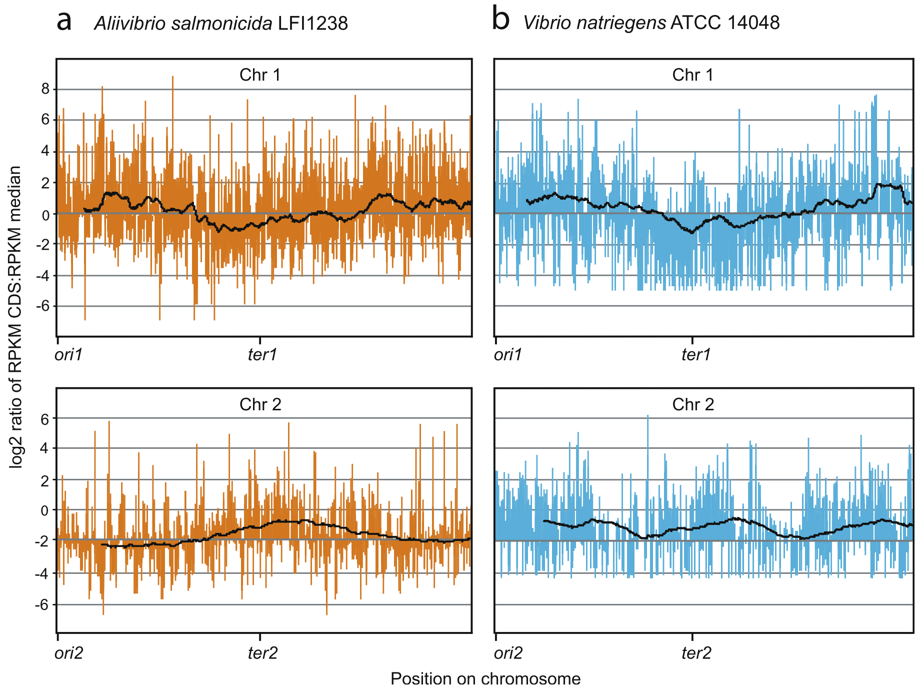

Fig. 4 Global expression maps of (**a**) *A. salmonicida* LFI1238 and (**b**) *V. natriegens* ATCC 14048 chromosomal genes centered around the median. Data points ($\log_2$ ratio RPKM CDS:RPKM median) for each CDS are shown, as well as a trend line averaged over a sliding window of 200 data points. *V. natriegens* ATCC 14048 is grown under fast-growing conditions and *A. salmonicida* LFI1238 is grown under suboptimal conditions

All pangene categories contribute to higher expression levels around *ori1* at fast-growth conditions, but not at slow-growth conditions

The global trend described above can be explained by generally higher expression levels of all pangene categories located close to *ori1*, or, higher expression of three or less of the four pangene categories. To discriminate between the two alternatives, we calculated the RPKM median value for each pangene category, and compared the median values for genes located on the upper or lower halves of Chr 1 (Table 1). The Wilcoxon signed-rank test strongly support ($P\text{-adj} \leq 0.05$) that median values for all four pangene categories are significantly higher for genes located on the the upper half, i.e., when *V. natriegens* is cultured at fast-growth ("optimal") conditions. Notably, when grown under slow-growing conditions, median values for softcore, shell and cloud genes located on the upper half are significantly higher. Core genes are in contrast, expressed at equal levels on both halves. This applies for both *V. natriegens* (RPKM median = 370 and 360, *P*-adj = 0.321) in minimal medium, and *A. salmonicida* (RPKM median = 301 and 309, *P*-adj = 0.717) at suboptimal conditions. Conversely, we can therefore state that genes from all pangene categories located on the lower half are generally expressed at lower levels compared to those on the upper half (except for core genes at slow growth conditions). To summarize, we conclude that gene expression levels correlate with distance to *ori1* (Fig. 4), and genes from all four pangene categories contribute to this trend when grown under fast-growing conditions, whereas softcore, shell and cloud genes contribute at slow-growing conditions.

Discussion

Inspired by the discovery of multifork replication and increased copy numbers of genes surrounding the origin of replication, researchers have for decades studied how different categories of genes are distributed on chromosomes and at which level these genes are expressed. Here, we revisited this topic and describe hitherto hidden/unrecognized global gene distribution and expression patterns in *Vibrionaceae*. First, we mapped pangenes to their chromosomal positions and revealed that core and softcore genes are found heavily biased towards the *ori1* of Chr 1. Shell genes are, in contrast, overrepresented at the opposite part of Chr 1 (i.e., close to *ter*). We next found that gene expression strongly correlates with chromosomal distance to *ori1*. This trend is caused by higher expression of all pangene categories at fast-growing conditions, whereas softcore, shell and cloud genes are responsible for biased (higher) expressing on the upper half of Chr 1 at slow-growing conditions.

Pangene categories are non-randomly distributed on Chr 1

In this work we report a clear pattern where core/softcore genes are overrepresented on the upper half of Chr 1 of *Vibrionaceae*, particularly at regions corresponding to 10–11 and 1–2 O'clock on Chr 1, and shell/cloud genes are overrepresented in the *ter1* region (Fig. 2). In comparison, no clear pattern was recorded for Chr 2, i.e., the distribution of pangenes appear generally independent of location. For Chr 1, the core/softcore gene distribution pattern resembles that described for genes involved in translation and transcription in *E. coli* [16, 17, 33] and in several *Vibrio* species [16, 17, 21]. More precisely, Couturier and

Table 1 Comparison of gene expression levels for pangenes located on the upper or lower halves of Chr 1

	A. salmonicida				***V. natriegens* slow-growth**				***V. natriegens* fast-growth**			
	core	**softcore**	**shell**	**cloud**	**core**	**softcore**	**shell**	**cloud**	**core**	**softcore**	**shell**	**cloud**
Upper half[a]												
Q_1	152	118	42	42	188	126	21	5	249	170	36	37
Q_2	301	245	89	67	370	288	71	147	447	341	93	269
Q_3	853	633	197	197	1101	760	190	426	1059	719	241	581
Max	34,254	34,254	6473	13,656	23,238	23,238	17,161	5533	35,274	35,274	28,737	4049
Lower half[a]												
Q_1	151	89	34	25	143	83	4	4	178	109	0	0
Q_2	309	207	65	47	360	192	28	18	328	232	26	17
Q_3	695	486	133	82	966	565	74	59	696	480	97	62
Max	53,501	8098	19,837	23,646	14,116	14,116	15,800	463	16,521	17,549	17,550	535
P-value Q_2 [b]	0.71	0.01	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00

[a] Q_1 is the RPKM value at the first quartile. Q_1 is defined as the middle number between the smallest number and the median (i.e., the second quartile Q_2), if the data numbers (in this case RPKM values) are ordered from smallest to largest. The third quartile (Q_3) is the middle value between the median (Q_2) and the maximum (Max) value

[b] Adjusted *P*-values from Wilcoxon signed-rank test, to test if Q_2 values (median) of genes located on the upper half of Chr 1 are significantly different from Q_2 values of genes located on the lower half. Values below 0.05 are considered significant

Rocha (2006) showed that genes involved in translation and transcription in four *Vibrio* species are typically found close to *ori1* of Chr 1. Chr 2 contained, in contrast, fewer genes related to translation and transcription than would be expected. Iida and coworkers [21] later found that genes related to growth (both essential and contributing) are located in close proximity to *ori1* in *V. cholerae*. Over-representation of core/softcore genes, many of which are important for growth, at the region proximate to *ori1* of *Vibrionaceae* Chr 1 can be explained by an increase in demand for *ori1*-proximate gene products during fast growth (i.e., multifork replication results in elevated gene copy numbers and increased transcription levels). For example, genes that encode ribosomal RNA and ribosomal proteins are found clustered in the upper half of Chr 1, and are expressed at extremely high levels, which support this hypothesis.

Moreover, we found that during fast growth of *V. natriegens,* core, softcore, shell and cloud genes are all expressed at higher levels on the upper half of the chromosome compared to the lower half. In slow-growing *V. natriegens* and *A. salmonicida*, only softcore, shell and cloud genes followed the same trend, which suggests that regulatory mechanisms other than “gene dosage” are in play, to ensure a relatively low and uniform expression of core genes independent of chromosomal position during slow growth.

Why are core and softcore genes clustered at the old pole area of cells?

It is well documented in the literature that the intracellular space of bacteria is highly organized, with defined structures at specific locations (reviewed by Surovtsev and Jacobs-wagner 2019) [34]. For example, Chr 1 and Chr 2 of *V. cholerae* are spatially organised in a longitudinal orientation inside the cell, with their chromatin stretching from one pole to the other. *ori1* and *ter1* of Chr 1 are located at the old and new poles, respectively, whereas *ori2* and *ter2* of Chr 2 stretches from the new pole towards the cell’s center, respectively (Fig. 5). The organization of Chr 1 and Chr 2 in *V. cholerae* has been established by both fluorescence tag microscopy [9, 35, 36] and chromosome conformation capture (3C) [11]. In the light of this knowledge, our data then suggest that core/softcore and shell/cloud genes are enriched at two spatially separated intracellular regions, i.e., at the two extreme poles of *Vibrionaceae* cells, given that the spatial positioning of chromatin described for *V. cholerae* applies to all representatives within the family. We emphasize, however, that this hypothesis is based on limited data and should be further tested in future experiments before any strong conclusion can be made. Below we further speculate on why core and softcore genes appear clustered at the old (flagellated) pole area.

Given a non-random structural organization of the genes (as hypothesized above), this then suggests to us that there is a link between gene placement and their function, and that the underlying reasons for the strong distribution pattern could be very complex. The full complexity of factors that affects gene expression can be illustrated by e.g., chromatin packing [37–41], nucleoid-associated proteins (NAPs) [42–44], Structural Maintenance of Chromosome complex (SMC) [45], RNA polymerase (RNAP) [46–50], transcription factors and promoter strength/chromosomal position [43, 51] and macromolecular crowding [20]. Perhaps the most fundamental factor is chromatin packing and organization. The density of chromatin is determined by a number of circumstances, including differential abundance/availability of macromolecular machineries [38, 41, 46–50, 52, 53]. In this respect the bipartite DNA organization of *Vibrionaceae* represents a special case because Chr 1 stretches from pole to pole, whereas Chr 2 prolongates from the new pole towards the cell center, thus suggesting that the chromatin density varies between the two halves of the cell. Higher chromatin density will presumably reduce the diffusion of

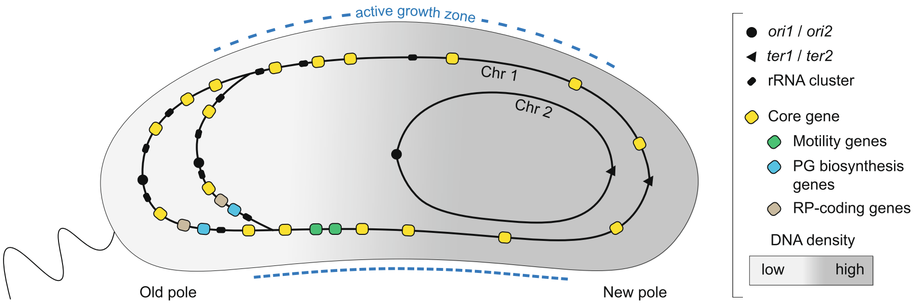

Fig. 5 Subcellular distribution of Chr 1 and Chr 2 in *V. cholerae*. Core genes are spatially enriched in the intracellular region near the old pole. Coloured core gene clusters (related to motility, peptidoglycan biosynthesis and ribosomal proteins) represent core gene products that co-localize with growth/survival-related reactions in the old pole of the cell. Two replication origins on Chr 1 indicate multifork replication. Active growth zones are indicated with blue dashed lines along the axis of the cell. Small dashes illustrate fast peptidoglycan growth and long dashes illustrate slower growth

macromolecular particles, such as proteins and ribosomes, in the nucleoid/DNA meshwork. Given that the DNA density is lower in the old pole area, the extra cytoplasmic space will presumably result in increased diffusion and transport of gene products, which provides a plausible explanation for the high abundance of core genes (many of which are growth related), and also the ribosomal protein clusters and rRNA clusters, in this subcellular region. Production of core gene products will therefore coincide and co-localize with the greatest number of growth/survival-related reactions and processes in the cell. A number of such cases can be mentioned, albeit we highlight two potential cases below.

The insertion of peptidoglycan (PG) in the cell wall happens in a dispersed manner, with the active growth zones along the axis [54]. To form the inner curvature of *Vibrio* cells, PG insertion is biased along the outer curve. Genes involved in cell wall synthesis are located in close proximity to *ori1* on *V. cholerae* Chr 1, with the main gene cluster related to nascent PG synthesis positioned approximately 0.38 Mb from *ori1.* This suggests that the first step of PG synthesis preferentially takes place in the old pole area. Similarly, motility related genes are found clustered 0.6 Mb from *ori1*, which is spatially close to the flagellum at the old pole.

Conclusions

Our results show a non-random organization of pangene categories on the two chromosomes of *Vibrionaceae*, with an overrepresentation of core and softcore genes around *ori1.* Gene distribution was compared with global gene expression trends and showed that during fast growth, all pangene categories contribute to a skewed expression pattern in respect to *ori1*. From our data and previous literature, we can deduce that core and softcore genes are overrepresented at the old pole area of *V. cholerae*. We hypothesize that this pattern can be beneficial due to spatial links between the structural organization of core genes and their cellular function, and that differences in intracellular DNA densities might further contribute to the biased gene distribution. These findings add to the growing list of examples of spatial order in bacteria, and scientists will surely continue to study the interplay between genome organization, gene activity and cellular function. We envision to explore how different pangene categories are distributed on chromosomes of other bacterial orders, and to search for similar spatial links to gene functions to investigate if our current findings are part of a general trend in Bacteria, or specific to *Vibrionaceae.*

Methods

Genome retrieval and gene annotation

As of May 2018 a total of 124 complete *Vibrionaceae* genomes were publicly available at the National Center for Biotechnology Information (NCBI) which were downloaded from the RefSeq database at NCBI [55] (see Additional file 1 for a complete list). All genome sequences were re-annotated using RAST (Rapid Annotation using Subsystem Technology) version 2.0 [56] with default settings. The annotation of the 124 genome sequences resulted in a total of 555,513 annotated protein sequences.

Pangenome approach to extract core, softcore, shell and cloud genes from large genome dataset

To categorize the annotated *Vibrionaceae* protein sequences into four categories (core, softcore, shell and cloud genes) we performed pangenome analysis using the software package GET_HOMOLOGUES (v3.1.0 (20180103)) [23]. The clustering algorithm OrthoMCL was used to cluster homolog protein sequences. The parameter “minimum percent sequence identity” was set to 50 and “minimum percent coverage in BLAST query/subj pairs” was set to 75 (default).

Comparison of core, softcore, shell and cloud genes from 11 species

We chose 11 representative species (based on phylogeny and scientific interest i. e. number of papers published in PubMed) to study the distribution of core, softcore, shell and cloud genes on Chr 1 and Chr 2. Chr 1 and Chr 2 were divided into “upper half” (close to *ori*) and “lower half” (close to *ter*) and the number of core, softcore, shell and cloud genes in each half were counted (see Additional file 3). The 11 species were used to study the exact chromosomal positions of core and shell genes on Chr 1 and Chr 2. The DoriC database [57] was used to locate *ori1* and *ori2* in Chr 1 and Chr 2 to subsequently center the plotted chromosomes at origin of replication, respectively at *mioC* on Chr 1 and *rtcB* on Chr 2. The software package Circos [24] was used to visualize the gene distributions on the chromosomes.

Analysing gene expression: mapping of read files on reference genomes

To study gene expression of core, softcore, shell and cloud genes in *A. salmonicida* LFI1238 and *V. natriegens* ATCC 14048 (NBRC 15636, DSM 759), the following datasets were downloaded from the Sequence Read Archive [28] at the NCBI: for *V. natriegens* ATCC 14048 datasets from growth in minimal (BioSample accession no. SAMN10926309, SAMN10926310 and SAMN10926313) and optimal (rich) medium (sample no. SAMN10926311, SAMN10926312 and SAMN10926329) at 37 °C to OD_{600nm} 0.3—0.5 [29]; for *A. salmonicida* LFI1238 one dataset (sample no. SAMEA4548122, SAMEA4548133, SAMEA4548134) originating from growth in LB medium containing 1% NaCl at 8 °C to mid log phase (OD_{600nm} ~

0.5) [30]. The salt concentration is expected to be similar to the concentration the bacterium would experience inside its natural host (Atlantic salmon), where the bacterium is known to cause cold water vibriosis at temperatures below 10 °C [26, 27]. Hence, 8 °C was used in the experiment. The quality of the reads was checked using FastQC [58]. EDGE-pro v1.0.1 (Estimated Degree of Gene Expression in Prokaryotes) [31] in Galaxy was used to align cDNA reads to *V. natriegens* ATCC 14048 (assembly no. GCA_001456255.1) and *A. salmonicida* LFI1238 (assembly no. GCF_000196495.1) and estimate gene expression as reads per kilobase per million (RPKM) for all protein coding sequences (CDS). The RPKM values were then used to calculate the $\log_2$ ratio RPKM CDS: RPKM median to make global expression maps for each of the three datasets.

Statistical analysis

Statistical analysis was performed using R in RStudio. Significance of gene distribution on either the upper or lower half of the chromosomes was performed using R's chisq.test() function for the non-parameteric chi-squared test (see Additional file 4). Significance of gene expression between gene classes located on the upper or lower half of the chromosomes was performed using R's wilcox.test() function for unpaired Wilcoxon signed-rank tests (see Additional file 4). For both analyses *P*-values were Bonferroni corrected for multiple comparisons using R's p.adjust() function.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-07117-5.

Additional file 1: Table S1. Complete list of the 124 *Vibrionaceae* genomes used in this study.

Additional file 2: Fig. S1. The 11 *Vibrionaceae* representatives mapped to a phylogeny.

Additional file 3: Table S2. Distribution and percent of total number of CDSs per chromosome of core, softcore, shell and cloud genes on «upper half» and «lower half» of Chr 1 and Chr 2 of 11 representative *Vibrionaceae* genomes.

Additional file 4: Table S3. Statistical analysis of gene distribution (chi-squared test) and gene expression (Wilcoxon signed-rank test) between "upper half" and "lower half" of Chr 1 and Chr 2.

Additional file 5: Fig. S2. Global expression maps of *V. natriegens* ATCC 14048 (grown under slow-growing conditions) chromosomal genes centered around the median. Data points ($\log_2$ ratio RPKM CDS:RPKM median) for each CDS are shown, as well as a trend line averaged over a sliding window of 200 data points.

Additional file 6: Fig. S3. Circular visualization of pangene distribution and gene expression ($\log_2$ ratio RPKM CDS:RPKM median) of (a) *A. salmonicida* LFI1238 and *V. natriegens* ATCC 14048 grown under (b) fast- and (c) slow-growing conditions.

Acknowledgements

We thank Professor Tomoo Sawabe, Hokkaido University, for providing multiple sequence alignments.

Authors' contributions

PH and CBS designed the study and wrote the manuscript. CBS performed all bioinformatics analysis. TK did statistical analyses and contributed to the writing of the manuscript. All authors contributed to proofreading and approved on the final manuscript.

Funding

The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. The funder had no role in study design, data collection and analysis or preparation of the manuscript.

Availability of data and materials

All data analysed during this study are included in this published article, its additional files and publicly available repositories. The RNA-seq datasets used in this study are available at Sequence Read Archive at Bioproject Accession PRJNA522293 [29] and PRJEB17700 [30].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Chemistry and Center for Bioinformatics (SfB), Faculty of Science and Technology, UiT The Arctic University of Norway, N-9037 Tromsø, Norway. [2]Climate Change Cluster, University of Technology Sydney, Sydney, NSW, Australia.

Received: 9 June 2020 Accepted: 29 September 2020
Published online: 06 October 2020

References

1. Thompson FL, Iida T, Swings J. Biodiversity of Vibrios. Microbiol Mol Biol Rev. 2004;68:403–31.
2. Takemura AF, Chien DM, Polz MF. Associations and dynamics of *Vibrionaceae* in the environment, from the genus to the population level. Front Microbiol. 2014;5:38.
3. Montánchez I, Kaberdin VR. *Vibrio harveyi*: a brief survey of general characteristics and recent epidemiological traits associated with climate change. Mar Environ Res. 2020;154:104850.
4. Hoff J, Daniel B, Stukenberg D, Thuronyi BW, Waldminghaus T, Fritz G. *Vibrio natriegens*: an ultrafast-growing marine bacterium as emerging synthetic biology chassis. Environ Microbiol. 2020;10.
5. Okada K, Iida T, Kita-Tsukamoto K, Honda T. Vibrios commonly possess two chromosomes. J Bacteriol. 2005;187:752–7.
6. di Cenzo GC, Finan TM. The Divided Bacterial Genome: Structure, Function, and Evolution. Microbiol Mol Biol Rev. 2017;81:e00019–7.
7. Val ME, Kennedy SP, El Karoui M, Bonné L, Chevalier F, Barre FX. FtsK-dependent dimer resolution on multiple chromosomes in the pathogen *Vibrio cholerae*. PLoS Genet. 2008;4:e1000201.
8. Egan ES, Fogel MA, Waldor MK. MicroReview: divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol. 2005;56:1129–38.
9. Srivastava P, Chattoraj DK. Selective chromosome amplification in *Vibrio cholerae*. Mol Microbiol. 2007;66:1016–10289.
10. Rasmussen T, Jensen RB, Skovgaard O. The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle. EMBO J. 2007;26:3124–31.
11. Val ME, Marbouty M, de Lemos MF, Kennedy SP, Kemble H, Bland MJ, et al. A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. Sci Adv. 2016;2:e1501914.
12. Kemter FS, Messerschmidt SJ, Schallopp N, Sobetzko P, Lang E, Bunk B, et al. Synchronous termination of replication of the two chromosomes is an evolutionary selected feature in *Vibrionaceae*. PLoS Genet. 2018;14:e1007251.
13. Cooper S, Helmstetter CE. Chromosome replication and the division cycle of *Escherichia coli* B/r. J Mol Biol. 1968;31:519–40.

14. Stokke C, Waldminghaus T, Skarstad K. Replication patterns and organization of replication forks in *Vibrio cholerae*. Microbiology. 2011;157:695–708.
15. Slager J, Veening JW. Hard-wired control of bacterial processes by chromosomal gene location. Trends Microbiol. 2016;24:788–800.
16. Rocha EPC. The replication-related organization of bacterial genomes. Microbiology. 2004;150:1609–27.
17. Couturier E, Rocha EPC. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol. 2006;59:1506–18.
18. Soler-Bistué A, Mondotte JA, Bland MJ, Val ME, Saleh MC, Mazel D. Genomic location of the major ribosomal protein gene locus determines *Vibrio cholerae* global growth and infectivity. PLoS Genet. 2015;11:e1005156.
19. Soler-Bistue A, Timmermans M, Mazel D. The proximity of ribosomal protein genes to oric enhances *Vibrio cholerae* fitness in the absence of multifork replication. MBio. 2017;8.
20. Soler-Bistué A, Aguilar-Pierlé S, Garcia-Garcerá M, Val M-E, Sismeiro O, Varet H, et al. Macromolecular crowding links ribosomal protein gene dosage to growth rate in *Vibrio cholerae*. BMC Biol. 2020;18:43.
21. Dryselius R, Izutsu K, Honda T, Iida T. Differential replication dynamics for large and small *Vibrio* chromosomes affect gene dosage, expression and location. BMC Genomics. 2008;9:559.
22. Toffano-Nioche C, Nguyen AN, Kuchly C, Ott A, Gautheret D, Bouloc P, et al. 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. 2012;18:2201–19.
23. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013;79:7696–701.
24. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
25. Eagon RG. *Pseudomonas natriegens*, a marine bacterium with a generation time of less than 10 minutes. J Bacteriol. 1962;83:736–7.
26. Colquhoun DJ, Sørum H. Temperature dependent siderophore production in *Vibrio salmonicida*. Microb Pathog. 2001;31:213–9.
27. Enger O, Husevåg B, Goksøyr J. Seasonal variation in presence of *Vibrio salmonicida* and total bacterial counts in Norwegian fish-farm water. Can J Microbiol. 1991;37:618–23.
28. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39:D19–21.
29. Lee HH, Ostrov N, Wong BG, Gold MA, Khalil AS, Church GM. Functional genomics of the rapidly replicating bacterium *Vibrio natriegens* by CRISPRi. Nat Microbiol. 2019;4:1105–13.
30. Thode SK, Bækkedal C, Söderberg JJ, Hjerde E, Hansen H, Haugen P. Construction of a fur null mutant and RNA-sequencing provide deeper global understanding of the *Aliivibrio salmonicida* Fur regulon. PeerJ. 2017;5:e3461.
31. Magoc T, Wood D, Salzberg SL. EDGE-pro: estimated degree of gene expression in prokaryotic genomes. Evol Bioinformatics Online. 2013;9:127–36.
32. Kasimoglu E, Park SJ, Malek J, Tseng CP, Gunsalus RP. Transcriptional regulation of the proton-translocating ATPase (atpIBEFHAGDC) operon of *Escherichia coli*: control by cell growth rate. J Bacteriol. 1996;178:5563–7.
33. Ardell DH, Kirsebom LA. The genomic pattern of tDNA operon expression in *E. coli*. PLoS Comput Biol. 2005;1:0086–99.
34. Surovtsev I, Jacobs-Wagner C. Subcellular organization: a critical feature of bacterial cell replication. Cell. 2018;172:1271–93.
35. Fogel MA, Waldor MK. Distinct segregation dynamics of the two *Vibrio cholerae* chromosomes. Mol Microbiol. 2005;55:125–36.
36. David A, Demarre G, Muresan L, Paly E, Barre FX, Possoz C. The two cis-acting sites, parS1 and oriC1, contribute to the longitudinal organisation of *Vibrio cholerae* chromosome I. PLoS Genet. 2014;10:e1004448.
37. Martis BS, Forquet R, Reverchon S, Nasser W, Meyer S. DNA supercoiling: an ancestral regulator of gene expression in pathogenic bacteria? Comput Struct Biotechnol J. 2019;17:1047–55.
38. Dorman CJ. DNA supercoiling and transcription in bacteria: a two-way street. BMC Mol Cell Biol. 2019;20:26.
39. Dorman CJ, Dorman MJ. DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. Biophys Rev. 2016;8:209–20.
40. Yildirim A, Feig M. High-resolution 3D models of *Caulobacter crescentus* chromosome reveal genome structural variability and organization. Nucleic Acids Res. 2018;46:3937–52.
41. Brocken DJW, Tark-Dame M, Dame RT. The organization of bacterial genomes: towards understanding the interplay between structure and function. Curr Opin Syst Biol. 2018;8:137–43.
42. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol. 2010;8:185–95.
43. Sobetzko P, Travers A, Muskhelishvili G. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. Proc Natl Acad Sci U S A. 2012;109:E42–50.
44. Dame RT, Tark-Dame M. Bacterial chromatin: converging views at different scales. Curr Opin Cell Biol. 2016;40:60–5.
45. Brandão HB, Paul P, Van Den Berg AA, Rudner DZ, Wang X, Mirny LA. RNA polymerases as moving barriers to condensin loop extrusion. Proc Natl Acad Sci U S A. 2019;116:20489–99.
46. Jin DJ, Cabrera JE. Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in *Escherichia coli*. J Struct Biol. 2006;156:284–91.
47. Jin DJ, Mata Martin C, Sun Z, Cagliero C, Zhou YN. Nucleolus-like compartmentalization of the transcription machinery in fast-growing bacterial cells. Crit Rev Biochem Mol Biol. 2017;52:96–106.
48. Yang S, Kim S, Kim DK, Jeon An H, Bae Son J, Hedén Gynnå A, et al. Transcription and translation contribute to gene locus relocation to the nucleoid periphery in *E coli*. Nat Commun. 2019;10:5131.
49. Weng X, Bohrer CH, Bettridge K, Lagda AC, Cagliero C, Jin DJ, et al. Spatial organization of RNA polymerase and its relationship with transcription in *Escherichia coli*. Proc Natl Acad Sci U S A. 2019;116:20115–23.
50. Martin CM, Sun Z, Zhou YN, Jin DJ. Extrachromosomal nucleolus-like compartmentalization by a plasmid-borne ribosomal RNA operon and its role in nucleoid compaction. Front Microbiol. 2018;9:1115.
51. Engstrom MD, Pfleger BF. Transcription control engineering and applications in synthetic biology. Synth Syst Biotechnol. 2017;2:176–91.
52. Bremer H, Dennis PP. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. EcoSal Plus. 2008;3.
53. Le TBK, Imakaev MV, Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. Science. 2013;342:731–4.
54. Bartlett TM, Bratton BP, Duvshani A, Zhu J, Shaevitz JW, Gitai Z, et al. A periplasmic polymer curves *Vibrio cholerae* and promotes pathogenesis. Cell. 2017;168:172–85 e15.
55. O'leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, Mcveigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2015;44:D733–45.
56. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.
57. Luo H, Gao F. DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. Nucleic Acids Res. 2019;47:D74–7.
58. Andrews S. FastQC. Babraham bioinformatics. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Last accessed 27 Apr 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PAPER 2

OXFORD
G3
Genes | Genomes | Genetics

G3, 2021, 11(9), jkab256
DOI: 10.1093/g3journal/jkab256
Advance Access Publication Date: 21 July 2021
Investigation
Featured

The *Pseudoalteromonas* multipartite genome: distribution and expression of pangene categories, and a hypothesis for the origin and evolution of the chromid

Cecilie Bækkedal Sonnenberg and Peik Haugen *

Department of Chemistry and Center for Bioinformatics (SfB), Faculty of Science and Technology, UiT The Arctic University of Norway, Tromsø N-9037, Norway

*Corresponding author: Department of Chemistry and Center for Bioinformatics (SfB), Faculty of Science and Technology, UiT The Arctic University of Norway, Tromsø N-9037, Norway. Email: peik.haugen@uit.no

Abstract

Bacterial genomes typically consist of one large chromosome, but can also include secondary replicons. These so-called multipartite genomes are scattered on the bacterial tree of life with the majority of cases belonging to Proteobacteria. Within the class gammaproteobacteria, multipartite genomes are restricted to the two families *Vibrionaceae* and *Pseudoalteromonadaceae*. Whereas the genome of vibrios is well studied, information on the *Pseudoalteromonadaceae* genome is much scarcer. We have studied *Pseudoalteromonadaceae* with respect to the origin of the chromid, how pangene categories are distributed, how genes are expressed relative to their genomic location, and identified chromid hallmark genes. We calculated the *Pseudoalteromonadaceae* pangenome based on 25 complete genomes and found that core/softcore are significantly overrepresented in late replicating sectors of the chromid, regardless of how the chromid is replicated. On the chromosome, core/softcore and shell/cloud genes are only weakly overrepresented at the chromosomal replication origin and termination sequences, respectively. Gene expression is trending downwards with increasing distance from the chromosomal *oriC*, whereas the chromidal expression pattern is more complex. Moreover, we identified 78 chromid hallmark genes, and BLASTp searches suggest that the majority of them were acquired from the ancestral gene pool of Alteromonadales. Finally, our data strongly suggest that the chromid originates from a plasmid that was acquired in a relatively recent event. In summary, this study extends our knowledge on multipartite genomes, and helps us understand how and why secondary replicons are acquired, why they are maintained, and how they are shaped by evolution.

Keywords: *Pseudoaltermonas*; pangenome; multipartite; chromid; Alteromonadales; secondary replicons

Introduction

Multipartite genomes are recognized by the concurrent presence of multiple replicons, *i.e.*, cells contain one or more large replicons in addition to the chromosome (Harrison *et al.* 2010). The majority of bacteria with multipartite genomes are associated with high tolerance to abiotic stresses, or are associated with animals, human, or plants as pathogens or symbionts (Misra *et al.* 2018). This observation, in addition to other data, has prompted scientists to hypothesize that multipartite genomes play crucial roles in the successful spread and establishment of bacteria into a broad range of ecological niches (Heidelberg *et al.* 2000). A striking and well-studied example is the bacterium *Vibrio fischeri*. Some strains colonize the light-emitting organ of squid (*e.g.*, the Hawaiian bobtail squid *Euprymna scolopes*), and produce bioluminescent light that enables the host to evade predators by counter-illumination (Soto and Nishiguchi 2014). Other strains are in contrast pathogens, which apparently is made possible due to the presence of a gene capture system, a superintegron, and pathogenicity islands located on the chromid (Soto and Nishiguchi 2014; diCenzo *et al.* 2019). Therefore, a link between the two-replicon genome architecture and the bacteria's lifestyle has been suggested. The reality is, however, that although carrying one or more extra replicon may promote new opportunities for a bacterium to move into new niches, other bacteria thrive in the same environment without additional large replicons, thus demonstrating that multipartite genomes are probably not necessary to succeed in that environment. Our general understanding of the origin, evolution, and functional roles of multipartite genomes remains fragmented, and multiple other equally likely hypotheses have been proposed to explain their existence. For example, carrying genes on more than one large replicon allows for replicon-specific gene dosage, and consequently replicon-specific gene expression regulation (*e.g.*, Couturier and Rocha 2006; Dryselius *et al.* 2008). Also, it has been suggested that the presence of multiple replicons allows bacteria to contain larger genomes, and may reduce the time required to complete replication thus allowing for rapid cell growth and division (diCenzo *et al.* 2014). Finally, we recently proposed a hypothesis that the presence of two large replicons allows for intracellular spatial separation of different categories of genes, and that there is a link

Received: April 08, 2021. **Accepted:** July 06, 2021

between the skewed gene placement and their function (Sonnenberg *et al.* 2020). The underlying reason for the separated distribution of gene categories in 3D is however likely very complex, and not easy to dissect.

Given that one or more of the hypotheses above are correct, then it is not surprising that bacteria with multipartite genomes are indeed widely distributed. They are as of today found scattered across the bacterial kingdom, into 6 of 81 phyla listed in the NCBI taxonomy system (*i.e.*, Proteobacteria, Bacteriodetes, Actinobacteria, Firmicutes, Deinococcus-Thermus, and Spirochaetes) (diCenzo and Finan 2017). It is highly likely however that more examples of multipartite genomes remain to be discovered, especially when considering that the number of complete genomes in the databases is still relatively low (∼23,000), and dominated by Proteobacteria (57%) and Terrabacteria (34%). One hundred one of 127 multipartite genomes group within the phylum Proteobacteria (diCenzo and Finan 2017), which means that the remaining 26 genomes are distributed among five other phyla. Inside the class gamma-proteobacteria, multipartite genomes are restricted to *Vibrionaceae* and *Pseudoalteromadaceae*. The multipartite genome of *Vibrionaceae* consists of one large circular chromosome (4.1–2.7 Mb) known as Chromosome 1 (Chr 1), and one smaller circular replicon (2.3–0.7 Mb) knows as Chromosome 2 (Chr 2), hereafter referred to as the *Vibrionaceae* chromid, in accordance with the nomenclature by Harrison *et al.* (2010). Replication of the *Vibrionaceae* chromosome and chromid is precisely coordinated by a mechanism that is partly understood. Briefly, when the replication fork approaches *crtS* (chromid replication triggering site) in Chr 1, a hitherto unknown mechanism triggers replication of the chromid (Val *et al.* 2016; Kemter *et al.* 2018), there is a brief pause in Chr 1 replication before the cycle ends in a synchronized termination of replication. In fast-growing bacteria, such as *Vibrio cholerae* and *Vibrio natriegens*, replication results in higher gene copy numbers of genes surrounding the origin of replication of Chr 1 (*ori1*) (this is known as "gene dosage effect"). Consequently, expression of genes typically decreases with increasing distance from *ori1* (Dryselius *et al.* 2008; Toffano-Nioche *et al.* 2012). This correlation does not necessarily apply to slow-growing bacteria, or fast-dividing bacteria grown under poor (sub-optimal) conditions. We recently published a study where we calculated the *Vibrionaceae* pangenome based on 124 genomes to study how the four pangene categories (core, softcore, shell, and cloud) are distributed on the genome (Sonnenberg *et al.* 2020). The analysis showed that core/softcore genes are typically found clustered around *ori1*, whereas shell and cloud genes densely populate terminus-proximate regions on Chr 1. On the chromid, genes are more randomly distributed, with no strong distribution pattern. On Chr 1, gene expression levels strongly correlate with distance to *ori1*, with higher expression levels around *ori1*. Interestingly, under slow-growing conditions all categories, except core genes, contribute to this pattern. This prompted us to question whether the observed gene distribution and expression patterns are specific to *Vibrionaceae*, or represent a general trend among bacteria with multipartite genomes.

The family *Pseudoalteromonadaceae* represents an excellent opportunity to study multipartite genomes, *e.g.*, how the genes are distributed and expressed, and its origin and evolution. As of March 2021, the Refseq database contains 53 complete *Pseudoalteromonadaceae* genomes. All genomes are bipartite and consist of one chromosome (3.1–4.9 Mb) and one chromid (0.6 – 1.8 Mb). Bosi *et al* (2017) calculated the *Pseudoalteromonas* pangenome based on 38 genomes (Bosi *et al.* 2017). Briefly, they described the pangenome as open and with a large percentage (80%) of unique genes. Furthermore, they estimated the last common ancestor (LCA) of *Pseudoalteromonas* to contain an estimated 2999 genes, compared to an average of 4245 genes in the present-day genomes, which supports that the genome has undergone a considerable expansion. More recently, Liao *et al* (2019) studied the evolution of the *Pseudoalteromonas* genome (Liao *et al.* 2019). Using a phylogenetic approach and timescale analysis, they showed that the chromosome and chromid have coexisted, probably since *Pseudoalteromonas* diverged from the putative LCA 500 million years ago. The chromid apparently originates from a megaplasmid that over time obtained essential genes (Médigue *et al.* 2005; Rong *et al.* 2016; Liao *et al.* 2019; Xie *et al.* 2021).

In summary, *Vibrionaceae* and *Pseudoalteromadaceae* represent the only two families from the gamma-proteobacteria class with multipartite genomes. Whereas the *Vibrionaceae* genome is well studied, the information on *Pseudoalteromadaceae* is scarce. In this study, we set out to gain insight into how pangene categories are distributed on *Pseudoalteromonadaceae* chromosomes and chromids, how genes are expressed relative to their genomic location, which genes can be regarded as hallmark genes of the chromid, and the origin and evolution of the chromid. We present data that support observations on gene distribution and global expression from other bacterial chromosomes, as well as data showing chromid-specific patterns that suggest specific roles of secondary replicons. Several pieces of evidence suggest the likely source of the chromid and its hallmark genes.

Materials and methods

Genome retrieval and gene annotation

A total of 25 *Pseudoalteromonas* genomes that were available at the onset of this project (mid 2019) at the National Center for Biotechnology Information (NCBI), were downloaded from the RefSeq database at NCBI (O'Leary *et al.* 2016) (see Supplementary File S1 for a complete list). The following genomes were excluded from the analysis: *Pseudoalteromonas atlantica* T6 (GCF_000014225.1) misplaced into *Pseudoalteromonas*, and later reclassified and renamed to *Paraglaciecola atlantica* T6. *P. atlantica* ECSMB14104 and *Pseudoalteromonas marina* ECSMB141043 are assembled into one contig, and the nature of their chromids could not be reliably resolved using Mauve. All genome sequences were re-annotated using RAST (Rapid Annotation using Subsystem Technology) version 2.0 (Aziz *et al.* 2008) with default settings. Mauve (Darling *et al.* 2004) was used to align genomes that were annotated with only one replicon.

Phylogenetic analysis

Phylogenetic relationships between Alteromonadales genomes were inferred using the nucleotide sequences *gyrB, recA, rpoD, recN*, and *topA* as described earlier (Busch *et al.* 2019), and included the seven families *Alteromonadaceae, Colwelliaceae, Idiomarinaceae, Moritellaceae, Pseudoalteromonadaceae, Psychromonadaceae*, and *Shewanellaceae* (see Supplementary Figure S1 for complete phylogeny). The nucleotide sequences were aligned using MUSCLE (Edgar 2004). Only unambiguously aligned positions were kept using BioEdit (Hall 1999), which resulted in a 9216 nt sequence alignment. MEGAX was used to generate a Maximum Likelihood (ML) tree, with the settings GTR (General Time Reversible) model, Gamma Distributed with Invariant (G + I), and Bootstrap with 1,000 pseudoreplicates (Kumar *et al.* 2018; Stecher *et al.* 2020). An ML-phylogeny of *Pseudoalteromonas* was based on 469 single-copy marker genes identified by EzTree

(Wu 2018). The robustness of nodes was tested with a bootstrap analysis inferred from ML−GTR+G + I.

Pangenome calculations

To classify the annotated *Pseudoalteromonas* protein sequences into four categories (core, softcore, shell, and cloud genes), we performed pangenome analysis using the software package GET_HOMOLOGUES (v3.1.0 (20180103) (Contreras-Moreira and Vinuesa 2013). The clustering algorithm MCL was used to cluster homologous protein sequences. The parameter "minimum percent sequence identity" was set to 50 and "minimum percent coverage in BLAST query/subj pairs" was set to 75 (default).

Mapping of core, softcore, shell, and cloud genes on the *Pseudoalteromonas* genome

To study the distribution of core, softcore, shell, and cloud genes of *Pseudoalteromonas*, the chromosome and chromid sequences were divided into 4, 6, 8, 10, and 12 equally sized sections, with sector one starting at origin of replication (*gidA* on the chromosome and *parA* on the chromid). For each sector, the number of core, softcore, shell, and cloud genes were counted. The number of genes in each sector was then divided by the total gene number per replicon (probability of a gene belonging to a sector). The probability of a gene belonging to a sector given equal distribution between sectors was calculated for each of the 4, 6, 8, 10, and 12 sized sectors (1 divided on numbers of sectors). Then, the log10 ratio was calculated of the probability of a gene belonging to a sector divided by the probability given an equal distribution of genes between all sectors. Only a summary of the results when chromosomes and chromids are divided into six sectors are presented in the paper itself. The summary was made by calculating log10 ratio of: The probability of a gene belonging to a sector on average (Average #genes in a sector/Average total #genes)/The probability given an equal distribution of genes between all sectors (1/#sectors). See Supplementary File S2 for data. Kruskal–Wallis test and Dunńs test were used to perform pairwise comparisons of number of genes between all sections (see Supplementary File S3 for data).

Gene expression analyses

RNA-seq datasets from *P. fuliginea* BSW20308 grown at three different temperatures, *i.e.*, 32° (BioSample accession no. SAMN06226833, SRR11593421, SRR11593421, and SRR11593422), 15° (sample no. SRR11593423, SRR11593424, and SRR11593425), and 4° (sample no. SRR11593426, SRR11593427, and SRR11593428) (Liao et al. 2019) were downloaded from the NCBI Sequence Read Archive (Leinonen *et al.* 2011) and analyzed. The quality of the reads was checked using FastQC (Andrews 2010). EDGE-pro v1.0.1 (Estimated Degree of Gene Expression in Prokaryotes) (Magoc et al. 2013) in Galaxy was used to align cDNA reads to the genome assembly (no. GCF_000310105.2) and estimate gene expression as reads per kilobase per million (RPKM) for all protein-coding sequences (CDS). The RPKM values were then used to calculate the $\log_2$ ratio RPKM CDS: RPKM median to make global expression maps for each of the three datasets. To identify which pangene categories contribute to the gene expression pattern, the chromosome was divided into "upper" and "lower" halves, and the chromid was divided into "upper" and "lower" halves, as well as "right" and "left" halves, and the RPKM median value for each pangene category was calculated (Supplementary File S4).

BLASTp searches

Homologs of chromid hallmark genes were identified by BLASTp when using the nonredundant database, and excluding the *Pseudoalteromonadaceae* family (taxid: 267888), with the thresholds: *e*-value $< 1e^{-15}$, sequence identity >30% and sequence coverage >70% (see Supplementary File S6).

Statistical analysis

Statistical analysis was performed using R in RStudio (RStudio Team 2021). Kruskal–Wallis test and Dunn' s test were used to perform pairwise comparisons of number of genes in replicon sections. The tests were chosen because the data did not follow a normal distribution, and sample sizes were low. R's Kruskal.test() function for the rank-based nonparameteric Kruskal–Wallis test and the dunn.test() function for *post hoc* Dunn's test was used (see Supplementary Files S2 and S3 for data). Significant difference of gene expression between replicon halves and replicons was performed using R's wilcox.test() function for unpaired Wilcoxon signed-rank tests (see Supplementary File S4 for data). For all analyses, *P*-values were Bonferroni corrected for multiple comparisons using R's p.adjust() function.

Results

Pseudoalteromonadaceae branches off from families with monopartite genomes

Figure 1 shows the overall phylogenetic relationships between bacterial families and genera that form the order Alteromonadales (see Supplementary Figure S1 for complete phylogeny). The ML-tree (GTR+G + I model) was based on the concatenated nucleotide sequences of *gyrB*, *recA*, *rpoD*, *recN*, and *topA* from selected bacteria from the seven families *Alteromonadaceae*, *Colwelliaceae*, *Idiomarinaceae*, *Moritellaceae*, *Pseudoalteromonadaceae*, *Psychromonadaceae*, and *Shewanellaceae*. The analysis shows that each genera and family forms well-supported monophyletic groups, similar to previous studies (Williams et al. 2010; Martin *et al.* 2015). Notably, the family *Pseudoalteromonadaceae* (includes only the genus *Pseudoalteromonas*), which exclusively contains bacteria with multipartite genomes, branches off from the monopartite genome-containing clades as a crown group together with its sister *Alteromonadaceae*. None of the bacteria outside of *Pseudoalteromonadaceae* contain multipartite genomes. These two observations strongly support that the chromid was acquired by the most recent LCA of *Pseudoalteromonadaceae*, likely in a single event (indicated with an arrow in Figure 1). A single origin of the chromid is supported by a phylogenetic analysis that showed congruent phylogenies between the chromosome and chromid (Liao *et al.* 2019). Finally, two separate estimates of time since divergence suggest that *Pseudoalteromonadaceae* branched off approximately 500, and 502–378 million years ago (Liao *et al.* 2019; Xie et al. 2021). Compared to *Vibrionaceae*, which also exclusively contains multipartite genomes, the birth of *Pseudoalteromonadaceae* is relatively recent. The time since divergence of *Vibrionaceae* is estimated to approximately 1100–900 million years ago (Xie et al. 2021).

The *Pseudoalteromonadaceae* pangenome contains 1399 core genes

The definite point of origin of the *Pseudoalteromonadaceae* chromid prompted us to study the multipartite genome in more detail, *e.g.*, to identify where the chromid replicon was acquired from,

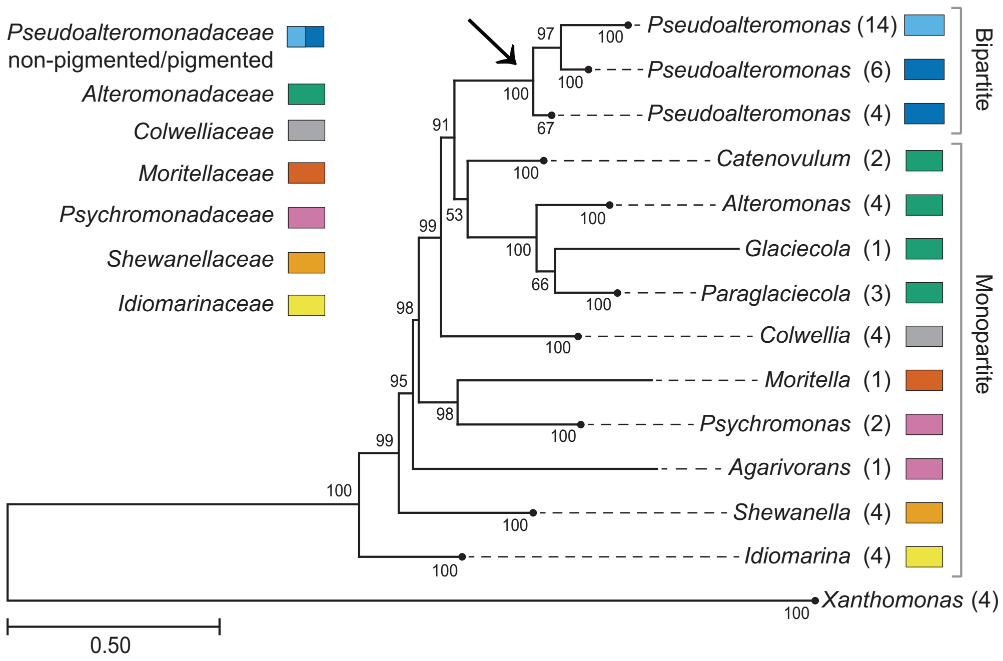

Figure 1 Summary of an ML-phylogenetic tree showing evolutionary relationships between Alteromonadales families. See Supplementary Figure S1 for complete tree. Multipartite genomes are restricted to *Pseudoalteromonadaceae*, which is placed at the crown of the tree and branches off from families containing monopartite genomes. The arrow highlights the LCA of *Pseudoalteromonadaceae*, and the likely point of origin of the *Pseudoalteromonadaceae* chromid. The color scheme shows family affiliation of genera. Numbers of strains included in collapsed nodes are shown in parentheses. Bootstrap values at the nodes were calculated using the ML method, and the GTR+G+I model, with 1000 replicates.

and how it has evolved after its acquisition. We used a pangenome approach as previously described (Sonnenberg *et al.* 2020). Briefly, available complete genomes were downloaded and re-annotated using RAST (Aziz *et al.* 2008). Genome datasets were then used to cluster orthologous groups of protein sequences based on the MCL algorithm (Van Dongen 2000) in GET_HOMOLOGUES (Contreras-Moreira and Vinuesa 2013). By using 25 available complete *Pseudoalteromonadaceae* genomes, which were available by the onset of the calculations (see Supplementary File S1 for complete list), we found a total of 24,991 clusters. The clusters were sub-categorized into 1399 core (encoded by all 25 genomes), 1606 softcore (encoded by ≥23 genomes), 7688 shell (encoded by ≤22 and ≥3 genomes), and finally 15,697 cloud (encoded by ≤2 genomes). This result is comparable to the calculations reported by Bosi *et al* (2017), based on 38 *Pseudoalteromonas* genomes (mostly draft genomes), which identified a total of 22,530 clusters, sub-divided into 1571 core (encoded by all 38 genomes), 2901 shell (encoded ≤37 and ≥2 genomes), and 18,058 cloud (encoded by one strain) (Bosi *et al.* 2017).

The distribution of core/softcore genes on the *Pseudoalteromonas* chromid strongly correlates with the direction of replication

To establish the distribution pattern of *Pseudoalteromonas* pangenes, we mapped all representative genes from the four pangene categories core, softcore, shell, and cloud to their chromosomal or chromidal locations. First, chromosome and chromid sequences were divided into 4, 6, 8, 10, or 12 equally sized sectors (or bins), with sector one starting from the origin of replication and proceeding clockwise. For each sector, the number of genes from each category were counted. At least for primary replicons, previous data from other bacterial families (Comandatore *et al.* 2019; Kopejtka *et al.* 2019; Sonnenberg *et al.* 2020), have shown that core/softcore genes densely populate regions that are replicated early in the replication cycle, and we hypothesized that *Pseudoalteromonas* would generate a similar distribution pattern. Notably, a recent study showed that most *Pseudoalteromonas* chromids are replicated unidirectionally, except for *Pseudoalteromonas spongiae* and *Pseudoalteromonas piratica* chromids, which are replicated bidirectional (Xie et al. 2021).

Figure 2A shows the result mapped onto a *Pseudoalteromonas* ML-phylogeny based on 469 single-copy marker genes identified by ezTree (Wu 2018). Heatmaps summarize the result for Clade 1 (unidirectional replication of chromid), and for Clade 2 (bidirectional replication of chromid). The heatmaps are based on average values from the 25 analyzed genomes (see Materials and Methods). Only data for chromids divided into 6 sectors are shown (see Supplementary File S2 for all datasets). The Kruskal–Wallis and the Dunńs *post hoc* tests were used to identifying significant over- or under-representation of gene numbers between all pairs of sectors (see Supplementary File S3). The main finding is that core/softcore genes densely populates late replicating chromidal sectors, regardless of if chromids are replicated uni- or bi-directionally, which is surprising and opposite of what we expected. Specifically, for unidirectionally replicated chromids (Clade 1), core/softcore genes are strongly overrepresented in sector 6, and underrepresented in sectors 2 and 3. For bidirectionally replicated chromids (Clade 2), core/softcore genes are strongly overrepresented in sectors 3 and 4, and underrepresented in sectors 2 and 6.

None of the pangene categories are, in contrast, significantly over- or under-represented in specific regions of the chromosome. Instead, core/softcore genes are only weakly overrepresented in sectors 1, 5, and 6 (near origin of replication), shell genes are weakly overrepresented in sectors 3 and 4 (near

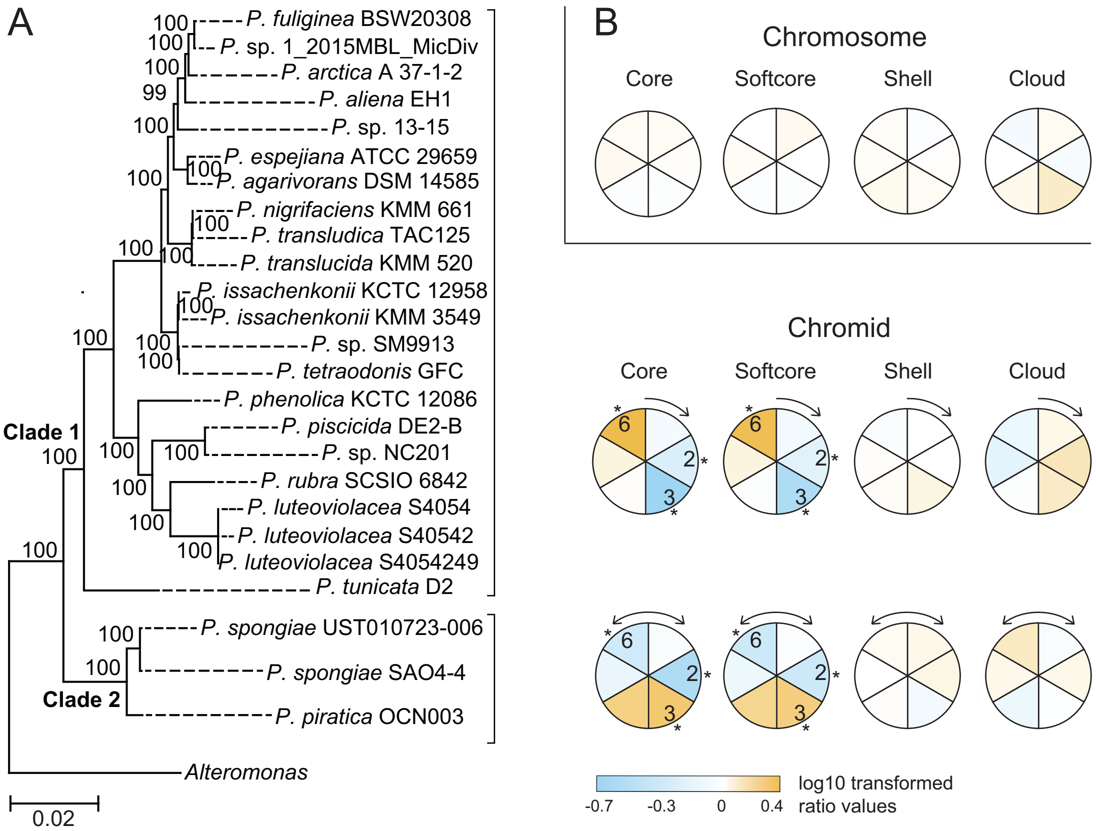

Figure 2 Heatmaps of distribution of core, softcore, shell, and cloud genes in *Pseudoalteromonas* genomes. Genes were placed into one of six equally sized sectors of chromids (A) or chromosomes (B), with sector 1 starting at the origin of replication (12 o'clock). Unidirectionally replicated chromids are found in species that belong to Clade 1, as shown in the ML-phylogeny (GTR+G+I model, 1000 replicates), whereas bidirectionally replicated chromids belong to representatives of Clade 2. Heatmaps are based on log $_{10}$ ratio values of the probability of a gene belonging to a sector on average divided by the probability given an equal distribution of genes among all sectors. Positive values (shades of orange) suggest that gene categories are overrepresented, whereas negative values (shades of blue) suggest underrepresentation. Asterisks indicate significant over- or under-representation of gene categories using Dunńs test *P*-value ≤ 0.05 (see Supplementary Files S2 and S3 for more details). The phylogenomic tree is based on 469 single marker genes identified by EzTree (Wu 2018). Bootstrap values at the internal nodes were inferred from a ML−G+I analysis.

terminus of replication), and cloud genes are weakly overrepresented in sector 3. The general pattern is therefore similar to, but less pronounced than what has been reported for *e.g.*, *Vibrionaceae* (Sonnenberg *et al.* 2020), *Klebsiella pneumonia* (Comandatore *et al.* 2019), and *Rhodobacteraceae* (Kopejtka *et al.* 2019).

In summary, by dividing the two *Pseudoalteromonas* replicons into 4–12 sectors and calculating the $\log_{10}$ ratio of the probability of a gene belonging to a sector divided by the probability given an equal distribution, we showed that core/softcore genes are significantly overrepresented in late replicating sectors of the chromid, regardless of how the chromid is replicated, *i.e.*, unidirectionally (Clade 1 strains) or bidirectionally (Clade 2 strains). Chromosomal genes are in contrast more evenly distributed into each sector of the replicon.

Gene dosage is in effect on the *Pseudoalteromonas fuliginea* BSW20308 chromosome, but not on the chromid

It is well established that the copy number of *ori*-proximate genes can increase during rapid growth due to the formation of multiple concurrent replication forks, which in turn result in multiple copies of the replicon (*e.g.*, a chromosome), and increased gene expression. This is known as the "gene dosage effect." To date, this has been described for the *Vibrionaceae* Chr 1 (Rasmussen *et al.* 2007; Srivastava and Chattoraj 2007; Dryselius *et al.* 2008; Toffano-Nioche *et al.* 2012), *Escherichia coli*, *Bacillus subtilis*, and *Streptomyces* (Couturier and Rocha 2006; Lato and Golding 2020). To establish if a gene dosage effect is in play in *Pseudoalteromonas* (for the chromosome, the chromid or both), we downloaded data from one of two available RNA-seq experiments stored at the NCBI Sequence Read Archive (Leinonen *et al.* 2011). In the selected experiment, *P. fuliginea* BSW20308 was grown in Difco marine broth 2216 and harvested at 4° (lowest temperature with growth), 15° (optimal growth), or 32° (maximum temperatures with growth) (Liao *et al.* 2019). These datasets, therefore, provide an excellent chance to test gene dosage effects at fast and slow growth, which is highly relevant because gene dosage has been reported to be particularly strong at rapid growth (and therefore rapid replication). The three RNA-seq datasets (each in triplicates) were analyzed as previously described (Sonnenberg *et al.* 2020). Briefly, cDNA reads were mapped onto the *P. fuliginea* BSW20308 genome (assembly no. GCF_000310105.2) and reads RPKM was calculated for all protein CDS.

Figure 3 shows global expression maps of the chromosome and chromid when *P. fuliginea* BSW20308 was grown at 4°, 15°, or 32°. Data points ($\log_2$ ratio RPKM CDS: RPKM median) are centered around the RPKM median. Moreover, for each plot a trend line averaged over a sliding window of 100 data points was added to show

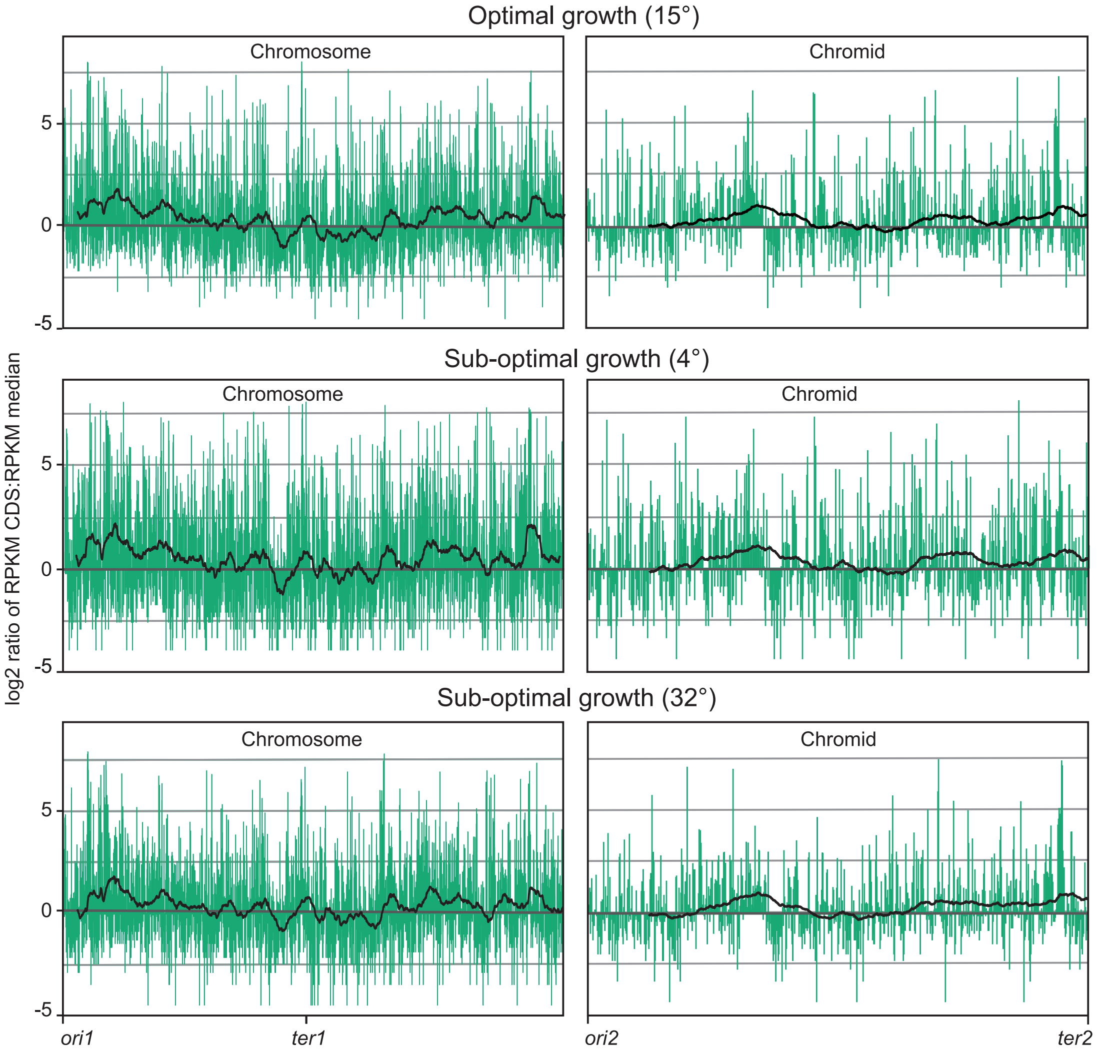

Figure 3 Global expression maps of *P. fuliginea* BSW20306 chromosomal and chromid genes centered on the median. Data points ($\log_2$ ratio RPKM CDS: RPKM median) for each CDS are shown, as well as a trend line averaged over a sliding window of 100 data points. The temperatures 4° and 32° corresponds to sub-optimal growth conditions and 15° corresponds to optimal growth conditions.

the overall direction of the data. Expression from the primary replicon (*i.e.*, the chromosome) is trending downwards starting from *ori1* and ending at *ter1*, with a low point at position 1,734,472. In other words, RPKM values are typically higher on the upper half compared to the lower half, which is expected if gene dosage is in effect on a bidirectionally replicated chromosome. This finding is strongly supported by the Wilcoxon signed-rank test (P-adj $\leq$ 0.05) (see Supplementary File S4). Expression from the chromid is elevated in the intermediate and late replicating regions. This expression pattern is opposite of what is expected if gene dosage is in effect on a unidirectionally replicated chromid, as in this case. If gene dosage was in effect then overlapping replication cycles would increase the number of DNA copies on the chromidal half (*i.e.*, the "right" half) which is replicated first. The Wilcoxon test does not, however, support significant differences in gene expression neither between upper and lower halves, or left and right halves (see Supplementary File S4).

In summary, we found that gene dosage appears to be in effect on the *Pseudoalteromonas* chromosome, but not on the chromid. This applies for all three tested temperatures, 4°, 15°, or 32°, which can be regarded as the minimum, optimum or maximum growth temperatures, respectively.

All pangene categories contribute to higher gene expression on the upper half of the *Pseudoalteromonas* chromosome under optimal growth temperature

To establish which pangene categories contribute to the gene dosage effect on the *P. fuliginea* BSW20308 chromosome, we calculated the RPKM median value for each pangene category (Table 1). The Wilcoxon signed-rank test strongly support (P-adj $\leq$ 0.05) that all four pangene categories contribute, when *P. fuliginea* BSW20308 is cultured at optimal conditions (15°). Interestingly, when grown at sub-optimal conditions (4° and 32°), the same test identifies only shell genes as significant contributors (see Table 1). The RNA-seq data further shows that RPKM median values of core and softcore genes are generally higher than that of shell and cloud genes (see Supplementary File S4), and this is valid for all three datasets except for the chromosome when grown under 32°. As expected, RPKM values are generally highest when grown at optimal temperature (15°), slightly lower at 4° and lowest at 32°. Overall, expression from chromosomal genes is higher compared to chromidal genes at 15° (RPKM median = 45 and 31, P-adj = 0.0), 4° (RPKM median = 30 and 20, P-adj = 0.0), and 32° (RPKM median = 22 and 20.5, P-adj = 0.043).

To summarize, under optimal growth conditions all four pangene categories contribute to higher gene expression on the upper part of the chromosome, whereas only shell genes contribute under sub-optimal conditions. As expected, absolute RPKM values are generally highest for core and softcore genes, and the median RPKM value for the chromosome is significantly higher than that of the chromid.

The *Pseudoalteromonas* chromid originates from an ancestral plasmid similar to those found in extant species of Alteromonadales

To investigate where the *Pseudoalteromonas* chromid originates from, we performed BLASTp searches using the chromid ParA, ParB, and RepA proteins as queries against the nr. database. The tripartite ParA-ParB-*parS* system consists of a ParA ATPase, a ParB CTPase and DNA-binding protein, and a centromere-like *parS* site, and is responsible for faithful segregation of replicons during cell growth and division in approximately three quarters of bacteria (Jalal *et al.* 2020). RepA is the replication initiator protein in *Pseudoalteromonas* chromids (Xie et al. 2021). The fundamental function of the partitioning system and the replication initiator protein, together with their widespread distribution in Bacteria and Archaea, make ParA, ParB, and RepA excellent candidates for finding clues to the origin of the chromid.

ParA and ParB BLASTp searches both identified homologs from draft genomes of *Rheinheimera* and *Catenovulum* as best hits (*e*-values = 0/0, Identities = 39%/41%; see top 20 list in Supplementary File S5). We believe that these hits represent auxiliary chromosomal ParA and ParB sequences originating from integrated plasmids. Following the top hits are a number of high-scoring matches against plasmids from *Pseudoalteromonas*, *Shewanella*, *Vibrio*, *Catenovulum*, *Alteromonas*, and *Glaciecola*. RepA BLASTp identified *Rheinheimera* and *Shewanella* draft genomes and *Catenovolum sediminis* plasmid as best hits, followed by *Aeromonas* plasmids. The BLASTp results therefore strongly suggest that the *Pseudoalteromonas* chromid originates from an ancestral plasmid, or possibly a megaplasmid, similar to those found in extant Alteromonadales species. Moreover, the relatively large size of today's *Pseudoalteromonas* chromids suggests that the acquired plasmid or megaplasmid has accumulated a vast number of genes that over time evolved into an in-dispensable replicon. A similar origin has been suggested for the *Vibrionaceae* chromid (*i.e.*, Chr 2) (Heidelberg *et al.* 2000; Fournes *et al.* 2018) and other chromids (Harrison *et al.* 2010; diCenzo *et al.* 2019).

More than half of the chromid hallmark genes in *Pseudoalteromonas* originates from the ancestral gene pool of Alteromonadales

Given that the *Pseudoalteromonas* chromid originates from an ancestral plasmid, then new questions emerge. For example, which type of genes are associated with chromids? Potential genetic sources could be genes from the *Pseudoalteromonas* chromosome, and/or genes from chromosome, chromid or plasmid DNA from closely or distantly related bacteria. To address this, we used the results from our pangenome analysis of *Pseudoalteromonas* genomes, and calculated the number of genes from each pangene category that are associated with the chromid. Any gene that was found at least once on a chromid was included. We found 164 core, 746 softcore, 2097 shell, and 4790 cloud genes, in total 7633 genes.

To find the genetic source of chromid genes we carefully selected a set of proteins and used them as queries in BLASTp searches. Of the 164 core genes only 78 are always located on the chromid (see Supplementary File S6 for complete list of genes). These are hereafter referred to as "chromid hallmark genes." Their ubiquitous presence on chromids support that they were acquired by the LCA, before diversification of *Pseudoalteromonas* took place (see arrow in Figure 1). Interestingly, about half (31) of the chromid hallmark genes are found clustered close to the replication terminus, and include genes and operons involved in histidine biosynthesis (*hisIEFAHBCDG*), DNA binding protein (*hupB*), acetolactate synthase (*ilvBH*), biopolymer transport system (*tonB-exbB-exbD*), and cell division (*minCDE*) (see Supplementary File S6 for more information). For the ancestral plasmid to be maintained and become part of the stable genome we see today, the chromid hallmark genes probably provided a selective advantage. We, therefore, regard these genes as great candidates for studying the origin of early chromid genes. All chromid hallmark proteins were used as queries in BLASTp searches (Supplementary File S6). In total, 42 (58%) of the proteins produced the highest scoring matches to sequences from Alteromonadales (after excluding matches from *Pseudoalteromonas*), followed by Chromatiales (11%), Vibrionales (10%), and Oceanospirillales (8%). This suggests that more than half of the current hallmark

Table 1 Comparison of gene expression levels for pangenes located on the upper or lower halves of the *P. fuliginea* BSW20308 chromosome

	Optimal growth conditions (15°)				Sub-optimal growth conditions (4°)				Sub-optimal growth conditions (32°)			
	Core	**Softcore**	**Shell**	**Cloud**	**Core**	**Softcore**	**Shell**	**Cloud**	**Core**	**Softcore**	**Shell**	**Cloud**
Upper half*a*												
Q_1	41	39	26	20	30	29	11	7	20	19	10	8
Q_2	84	82	43	37	77	74	24	15	41	39	18	15
Q_3	253	245	94	70	290	267	67	30	115	107	43	34
Max	11,063	11,063	134,285	37,786	48,320	48,320	169,723	53,846	5,208	5,208	35,549	8,083
Lower half*a*												
Q_1	29	28	16	13	24	21	7	5	18	17	8	6
Q_2	65	64	28	26	66	65	16	12	40	38	14	13
Q_3	185	181	59	55	267	232	46	32	97	94	32	27
Max	11,172	11,172	19,840	1,635	6,927	6,927	17,176	578	2,047	2,047	4,884	523
P-value Q_2*b*	0.00	0.00	0.00	0.00	0.08	0.07	0.00	0.15	0.40	0.40	0.00	0.20

[a] Q_1 is the RPKM value at the first quartile. Q_1 is defined as the middle number between the smallest number and the median (*i.e.*, the second quartile Q_2), if the data numbers (in this case RPKM values) are ordered from smallest to largest. The third quartile (Q_3) is the middle value between the median (Q_2) and the maximum (Max) value.

[b] Adjusted *P*-values from Wilcoxon signed-rank test, to test if Q_2 values (median) of genes located on the upper half of the chromosome are significantly different from Q_2 values of genes located on the lower half. Values below 0.05 are considered significant.

genes originates from the ancestral gene pool of Alteromonadales, whereas the remaining genes were acquired from diverse sources, mostly other gamma-proteobacteria.

The *Pseudoalteromonas* chromid contains a large number of genes with roles in iron uptake and homeostasis

A surprisingly large number of genes associated with iron-acquisition and homeostasis are located on the *Pseudoalteromonas* chromid. For example, in all 25 genomes, two *bfr* genes that encode bacterioferritin are located on the chromid, often flanked by *bdf* (encodes bacterioferritin-associated ferredoxin) and *iutA* (aerobactin siderophore receptor gene). Moreover, two complete *tonB-exbB-exbD* systems are found in all genomes, one on the chromid and one on the chromosome. And, in addition to *iutA*, six other TonB-dependent siderophore receptor genes are associated with the chromid, including *fhuA* (ferrichrome), *fhuE* (coprogen, rhodoturulate), *viuA* (vibriobactin), *fepA* (enterobactin), *desA* (deferoxamine B), and *vctA* (enterobactin). As previously reported for *Pseudoalteromonas tunicata* D2 (Thomas *et al.* 2008), a complete siderophore biosynthesis gene cluster is yet to be found in any of the *Pseudoalteromonas* genomes, even though they carry a relatively large number of siderophore receptor genes. This suggests that *Pseudoalteromonas* are "cheaters" *i.e.*, they have siderophore receptors on their surface with affinity to compounds produced by other bacteria (Payne *et al.* 2016). This mechanism is used as a strategy to avoid being discriminated against by other bacteria in the constant struggle between microorganisms to acquire iron.

Discussion

The number of studies on multipartite bacterial genomes has steadily increased along with the number of available finished genomes in public databases. As of March 16th 2021, there are 306,881 bacterial genome assemblies listed in the NCBI genome database, of which 22,910 are denoted as "complete" (7.5%). However, whereas some phyla are well represented, with *e.g.*, 57% of complete genomes belonging to Proteobacteria, and 34% belonging to Terrabacteria, most groups of bacteria are poorly represented, or not represented at all. Opportunities for doing studies on many gap-free multipartite genomes from single families are therefore rare. *Pseudoalteromonas* represents one of these rare cases. The chromid appears to originate from a relatively recent event that can be placed at a specific branch on the evolutionary tree with high confidence. We have therefore taken the opportunity to study the *Pseudoalteromonas* genome, and mapped how different gene categories are distributed and expressed in order to shed light on possible mechanisms that have shaped the chromosome and chromid.

We found that the *Pseudoalteromonas* genome partly confirms observations from other families, *e.g.*, that core/softcore genes appear more frequent around *ori1*, and shell (accessory) genes occur more frequent around *ter1* (Figure 2). We recently reported a similar strong correlation for *Vibrionaceae* Chr 1 (Sonnenberg *et al.* 2020). Using a slightly different pangenomic approach Comandatore et al. (2019) found a similar distribution pattern in *K. pneumonia*, whereas Kopejtka et al. (2019) reported a more complex picture with 22 species from *Rhodobacteraceae* showing clustering of core genes close to *oriC*, and eight species showing clustering around *ter* (Comandatore *et al.* 2019; Kopejtka *et al.* 2019). One plausible hypothesis is that core/softcore genes, which are associated with essential cell processes, are typically overrepresented around *oriC* because their gene products are of high demand during fast growth (Slager and Veening 2016). The rationale is that several concurrent initiations of replication from *oriC* results, on average, in higher "doses" of *oriC*-proximate genes. In turn, this leads to increased gene expression (the "gene dosage" effect) (Couturier and Rocha 2006). Analyses of *V. natriegens* and *A. salmonicida* (Sonnenberg *et al.* 2020), *Salmonella enterica* (Garmendia *et al.* 2018), and eleven bacterial data sets of diverse origin (Lato and Golding 2020) all confirmed that overall expression decreases with increasing distance to *oriC*. Our current analysis of a *P. fuliginea* BSW20308 RNA-seq data replicates a similar pattern (Figure 3).

The distribution pattern for chromid genes is in contrast very different, and perhaps more difficult to explain. For hitherto unknown reasons, the presence of core genes strongly correlates with distance to *ter2*. Interestingly, a recent study concluded that chromids belonging to the *P. spongiae* group are replicated bidirectionally, whereas chromids in all other *Pseudoalteromonas* are replicated unidirectionally (Xie et al. 2021). Accordingly, in bidirectionally replicated chromids *ter2* is located at 6 o'clock, and in unidirectionally replicated chromids *ter2* is located at 12 o'clock. The fact that core genes are overrepresented at *ter2* suggests that the genes are typically found in chromid sections that are replicated in the final part of the replication cycle, a situation that is opposite to that of *e.g.*, the *Pseudoalteromonas* and *Vibrionaceae* chromosomes where the gene dosage effect is in play. Gene dosage is apparently not in effect in *Pseudoalteromonas* chromids which suggests that we need another explanation for the clustering of core genes.

We can only speculate on why core/softcore genes tend to be located at *ter2*, but an intriguing possibility that we recently introduced for *Vibrionaceae* (Sonnenberg *et al.* 2020), is that the genomic distribution of gene categories is directly linked with how genes are organized into subcellular locations. In *V. cholerae*, Chrs 1 and 2 are longitudinally organized, with *ori1* located at the old pole, *ter1* and *ter2* located at the new pole, and *ori2* placed at the cell center (Fogel and Waldor 2005; Srivastava and Chattoraj 2007; David *et al.* 2014). Together, the data from *V. cholera* suggests to us that core/softcore and shell/cloud genes are enriched at two separate cellular locations, *i.e.*, at the old and new poles, respectively. Given that this hypothesis is correct, then it is plausible that a similar pattern/mechanism is in play in *Pseudoalteromonas*. It should be stressed that this remains per today a hypothesis, although the cytoplasmic position of individual gene loci have previously been successfully predicted based on the spatially organization of chromosomes (reviewed in Surovtsev and Jacobs-Wagner 2018). Moreover, for the hypothesis to be valid for *Pseudoalteromonas* there is an additional prerequisite that must be fulfilled: *ter2* is located at 6 or 12 o'clock (relative to *ori2*) depending on the replication mechanism that is in play (uni- or bidirectional). If *ter2* is deciding the subcellular destination of *ter2*-proximate core genes then they should, in principle, be located at the same subcellular compartment regardless of the replication mechanism. If, however, *ori2* is the decisive genetic component for intracellular positioning of the chromid, then *ter2*/associated core genes will be located at different spatial places depending on the replication mechanism (and positioning of *ter2* relative to *ori2*). Unfortunately, there is currently no evidence to suggest how *Pseudoalteromonas* cells are spatially organized intracellularly with regards to their chromatin. We note that the Min system (*minCDE*), which represents one of the best-studied proteins involved in cellular self-organization (reviewed by Wettmann and Kruse 2018), is located in the vicinity of *ter2*, but the significance of this is currently unclear to us.

Our results suggest that today's chromid in *Pseudoalteromonas* originated from a plasmid that was acquired in a single event in the LCA of this family. By comparing the chromid ParAB with database sequences we found that the best hits belong to plasmids found within today's representatives of Alteromonadales (Supplementary File S5). An early acquisition of chromid is further supported by congruent phylogenies of the chromosome and chromid, which support that the two replicons have coexisted since the LCA of *Pseudoalteromonas* (Liao *et al.* 2019; Xie et al. 2021). Given an early acquisition of a plasmid or megaplasmid, what then were the main driving forces for retaining and expanding the replicon size into a relatively large chromid? diCenzo *et al.* (2019) recently proposed that the main advantage with secondary replicons, is that they enable increased genetic flexibility and potential to acquire new genetic material (diCenzo *et al.* 2019). As a result, the bacterium is better suited to take advantage of new niche opportunities. It is an appealing concept, and several pieces of evidence from our study support the hypothesis. Perhaps the most compelling evidence comes from our pangenome calculations that identify the *Pseudoalteromonas* chromid as open and extremely flexible. A total of 7633 genes are associated with the chromid, which is approximately 10x greater that the number of genes encoded by individual chromids (553–1567 genes; median = 781). Moreover, chromid genes are generally expressed at a lower level, which have been suggested to increase the likelihood of newly acquired genes to be retained in the genome (Park and Zhang 2012; diCenzo *et al.* 2019). This is likely because highly expressed and mostly more critical genes on the chromosome are not disrupted, which then leads to less fitness cost for the bacterium. As a final piece of the puzzle, the vast majority of chromid genes in *Pseudoalteromonas* belong to the categories shell or cloud (see Supplementary File S2), which provides further support for the hypothesis that new genes are preferentially maintained on the chromid and thus increases the genetic plasticity of the *Pseudoalteromonas* genome.

To summarize, we provide data showing that *Pseudoalteromonas* core/softcore genes are weakly overrepresented at *oriC*-proximate regions, whereas shell/unique genes are weakly overrepresented around *ter1*. This distribution fits with patterns reported earlier for other bacteria (Comandatore *et al.* 2019; Kopejtka *et al.* 2019; Sonnenberg *et al.* 2020). Similarly, we found that gene expression is trending downwards with increasing distance to *oriC*, which also fits a general pattern among many bacteria (Garmendia *et al.* 2018; Lato and Golding 2020; Sonnenberg *et al.* 2020). For secondary replicons, the situation appears more complex. Here, the distribution pattern for pangene categories, as well as global expression maps, vary greatly among the studied bacteria. Perhaps the reason for the apparent lack of general trends is a direct result of the specialized roles of chromids, which have been shaped by the acquired and retained set of (mostly shell/unique) genes. Finally, we hypothesize that the gene distribution patterns reported by us and others are directly linked to how the DNAs are organized intracellularly, such that different pangene categories are enriched at separate subcellular locations based on their specialized biological functions.

Data availability

Supplemental data are available in Supplementary Figure S1 and Supplementary Files S1–S6. Supplementary Figure S1 shows phylogenetic relationships between Alteromonadales families. A list of all 25 *Pseudoalteromonas* genomes used in this study are available in Supplementary File S1. Supplementary File S2 contains distribution data of core, softcore, shell, and cloud on *Pseudoalteromonas* chromosome and chromid divided into 4, 6, 8, 10, and 12 sectors. Supplementary File S3 contains statistical analysis of pairwise comparisons of number of genes between sectors (Kruskal–Wallis and Dunńs test). Statistical analysis of gene expression of *P. fuliginea* BSW20306 chromosome and chromid using Wilcoxon signed-rank test is available in Supplementary File S4. Supplementary File S5 contains BLASTp results with chromid ParA, ParB, and RepA as queries. A list of chromid hallmark genes and BLASTp results is available in Supplementary File S6. Supplemental material is available at figshare: https://doi.org/10.25387/g3.14900463.

Funding

This work was supported by the UiT The Arctic University of Norway. The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. The funder had no role in study design, data collection, and interpretation, or the decision to submit the work for publication.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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

Bosi E, Fondi M, Orlandini V, Perrin E, Maida I, *et al.* 2017. The pangenome of (Antarctic) *Pseudoalteromonas* bacteria: evolutionary and functional insights. BMC Genomics. 18:93.

Busch J, Agarwal V, Schorn M, Machado H, Bradley S, *et al.* 2019. Polybrominated products in the genus *Pseudoalteromonas*. Environ Microbiol. 21:1575–1585.

Comandatore F, Sassera D, Bayliss SC, Scaltriti E, Gaiarsa S, *et al.* 2019. Gene composition as a potential barrier to large recombinations in the bacterial pathogen *Klebsiella pneumoniae*. Genome Biol Evol. 11:3240–3251.

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

Couturier E, Rocha EPC. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol. 59:1506–1518.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–1403.

David A, Demarre G, Muresan L, Paly E, Barre FX, *et al.* 2014. The two cis-acting sites, *parS1* and *oriC1*, contribute to the longitudinal organisation of *Vibrio cholerae* chromosome I. PLoS Genet. 10:e1004448.

diCenzo GC, Finan TM. 2017. The divided bacterial genome. Microbiol Mol Biol Rev. 81:e00019-17.

diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. 2014. Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLoS Genet. 10: e1004742.

diCenzo GC, Mengoni A, Perrin E. 2019. Chromids aid genome expansion and functional diversification in the family *Burkholderiaceae*. Mol Biol Evol. 36:562–574.

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.

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

Fogel MA, Waldor MK. 2005. Distinct segregation dynamics of the two *Vibrio cholerae* chromosomes. Mol Microbiol. 55:125–136.

Fournes F, Val ME, Skovgaard O, Mazel D. 2018. Replicate once per cell cycle: replication control of secondary chromosomes. Front Microbiol. 9:1833.

Garmendia E, Brandis G, Hughes D. 2018. Transcriptional regulation buffers gene dosage effects on a highly expressed operon in *Salmonella*. MBio. 9:e01446–18.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. – ScienceOpen. Nucleic Acids Symp. Ser. 41:95–98.

Harrison PW, Lower RPJ, Kim NKD, Young JPW. 2010. Introducing the bacterial "chromid": Not a chromosome, not a plasmid. Trends Microbiol. 18:141–148.

Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, *et al.* 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature. 406:477–483.

Jalal AS, Tran NT, Le TB. 2020. *ParB* spreading on DNA requires cytidine triphosphate *in vitro*. Elife. 20:e53515.

Kemter FS, Messerschmidt SJ, Schallopp N, Sobetzko P, Lang E, *et al.* 2018. Synchronous termination of replication of the two chromosomes is an evolutionary selected feature in *Vibrionaceae*. PLoS Genet. 14:e1007251.

Kopejtka K, Lin Y, Jakubovičová M, Koblízek M, Tomasch J. 2019. Clustered core- and pan-genome content on *Rhodobacteraceae* chromosomes. Genome Biol Evol. 11:2208–2217.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35:1547–1549.

Lato DF, Golding GB. 2020. Spatial patterns of gene expression in bacterial genomes. J Mol Evol. 88:510–520.

Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database Collaboration 2011. The sequence read archive. Nucleic Acids Res. 39:D19–D21.

Liao L, Liu C, Zeng Y, Zhao B, Zhang J, *et al.* 2019. Multipartite genomes and the sRNome in response to temperature stress of an Arctic *Pseudoalteromonas fuliginea* BSW20308. Environ Microbiol. 21:272–285.

Magoc T, Wood D, Salzberg SL. 2013. EDGE-pro: estimated degree of gene expression in prokaryotic genomes. Evol Bioinform Online. 9:127–136.

Martin M, Barbeyron T, Martin R, Portetelle D, Michel G, *et al.* 2015. The cultivable surface microbiota of the brown alga *Ascophyllum nodosum* is enriched in macroalgal-polysaccharide-degrading bacteria. Front Microbiol. 6:1487.

Médigue C, Krin E, Pascal G, Barbe V, Bernsel A, *et al.* 2005. Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis TAC125*. Genome Res. 15: 1325–1335.

Misra HS, Maurya GK, Kota S, Charaka VK. 2018. Maintenance of multipartite genome system and its functional significance in bacteria. J Genet. 97:1013–1038.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, *et al.* 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733–D745.

Park C, Zhang J. 2012. High expression hampers horizontal gene transfer. Genome Biol Evol. 4:523–532.

Payne SM, Mey AR, Wyckoff EE. 2016. Vibrio iron transport: evolutionary adaptation to life in multiple environments. Microbiol Mol Biol Rev. 80:69–90.

Rasmussen T, Jensen RB, Skovgaard O. 2007. The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle. EMBO J. 26:3124–3131.

Rong JC, Liu M, Li Y, Sun TY, Pang XH, *et al.* 2016. Complete genome sequence of a marine bacterium with two chromosomes, *Pseudoalteromonas translucida* KMM 520T. Mar Genomics. 26:17–20.

RStudio Team, 2021. RStudio: Integrated Development for R. Boston, MA: RStudio Team.

Slager J, Veening JW. 2016. Hard-wired control of bacterial processes by chromosomal gene location. Trends Microbiol. 24:788–800.

Sonnenberg CB, Kahlke T, Haugen P. 2020. *Vibrionaceae* core, shell and cloud genes are non-randomly distributed on Chr 1: an hypothesis that links the genomic location of genes with their intracellular placement. BMC Genomics. 21:695.

Soto W, Nishiguchi MK. 2014. Microbial experimental evolution as a novel research approach in the *Vibrionaceae* and squid-Vibrio symbiosis. Front Microbiol. 5:593.

Srivastava P, Chattoraj DK. 2007. Selective chromosome amplification in *Vibrio cholerae*. Mol Microbiol. 66:1016–1028.

Stecher G, Tamura K, Kumar S. 2020. Molecular evolutionary genetics analysis (MEGA) for macOS. Mol Biol Evol. 37: 1237–1239.

Surovtsev I, Jacobs-Wagner C. 2018. Subcellular organization: a critical feature of bacterial cell replication. Cell. 172:1271–1293.

Thomas T, Evans FF, Schleheck D, Mai-Prochnow A, Burke C, *et al.* 2008. Analysis of the *Pseudoalteromonas tunicata* genome reveals properties of a surface-associated life style in the marine environment. PLoS One. 3:e3252.

Toffano-Nioche C, Nguyen AN, Kuchly C, Ott A, Gautheret D, *et al.* 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. 18:2201–2219.

Val M-E, Marbouty M, de Lemos Martins F, Kennedy SP, Kemble H, *et al.* 2016. A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. Sci Adv. 2:e1501914.

Van Dongen SM. 2000. Graph clustering by flow simulation [Doctoral dissertation]. Utrecht: University of Utrecht.

Wettmann L, Kruse K. 2018. The min-protein oscillations in *Escherichia coli*: an example of self-organized cellular protein waves. Philos Trans R Soc Lond B Biol Sci. 373:20170111.

Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, *et al.* 2010. Phylogeny of gammaproteobacteria. J Bacteriol. 192: 2305–2314.

Wu YW. 2018. ezTree: an automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among uncultivated prokaryotic draft genomes. BMC Genomics. 19:921.

Xie BB, Rong JC, Tang BL, Wang S, Liu G, *et al.* 2021. Evolutionary trajectory of the replication mode of bacterial replicons. MBio. 12: e02745–20.

Communicating editor: D. Baltrus

PAPER 3

International Journal of
Molecular Sciences

MDPI

Article

Bipartite Genomes in *Enterobacterales*: Independent Origins of Chromids, Elevated Openness and Donors of Horizontally Transferred Genes

Cecilie Bækkedal Sonnenberg and Peik Haugen *

Department of Chemistry, The Center for Bioinformatics (SfB), Faculty of Science and Technology,
UiT The Arctic University of Norway, N-9037 Tromsø, Norway
* Correspondence: peik.haugen@uit.no

Abstract: Multipartite bacteria have one chromosome and one or more chromid. Chromids are believed to have properties that enhance genomic flexibility, making them a favored integration site for new genes. However, the mechanism by which chromosomes and chromids jointly contribute to this flexibility is not clear. To shed light on this, we analyzed the openness of chromosomes and chromids of the two bacteria, *Vibrio* and *Pseudoalteromonas*, both which belong to the *Enterobacterales* order of *Gammaproteobacteria*, and compared the genomic openness with that of monopartite genomes in the same order. We applied pangenome analysis, codon usage analysis and the HGTector software to detect horizontally transferred genes. Our findings suggest that the chromids of *Vibrio* and *Pseudoalteromonas* originated from two separate plasmid acquisition events. Bipartite genomes were found to be more open compared to monopartite. We found that the shell and cloud pangene categories drive the openness of bipartite genomes in *Vibrio* and *Pseudoalteromonas*. Based on this and our two recent studies, we propose a hypothesis that explains how chromids and the chromosome terminus region contribute to the genomic plasticity of bipartite genomes.

Keywords: *Vibrionaceae*; Pseudoalteromonas; multipartite; bipartite; pangenome; horizontal gene transfer; codon usage bias; chromid

check for updates

Citation: Sonnenberg, C.B.; Haugen, P. Bipartite Genomes in *Enterobacterales*: Independent Origins of Chromids, Elevated Openness and Donors of Horizontally Transferred Genes. *Int. J. Mol. Sci.* **2023**, *24*, 4292. https://doi.org/10.3390/ijms24054292

Academic Editor: Grzegorz Wegrzyn

Received: 31 January 2023
Revised: 14 February 2023
Accepted: 18 February 2023
Published: 21 February 2023

1. Introduction

Multipartite genomes refer to the presence of multiple replicons in a single bacterial cell and include one large chromosome, as well as one or more replicons (typically average size of 1.5 Mb), called chromids [1,2]. Bacteria with multipartite genomes are commonly found as pathogens or symbionts in animals, humans, and plants, as well as free-living bacteria [3,4] Although multipartite genomes are found throughout bacteria, 92% of those currently known are found in Proteobacteria or, using the validated name of this phylum, *Pseudomonadota* [5]). They are distributed among *Alphaproteobacteria*, *Betaproteobateria* and *Gammaproteobateria*, with 25%, 46% and 28% of multipartite bacteria found in each group, respectively [4]. Out of all multipartite bacteria, the majority (88%) are bipartite, i.e., they consist of one chromosome and one chromid.

The prevailing theory for the origin of bipartite genomes is that chromids have their origin from plasmids or megaplasmids that have been captured and domesticated by the ancestral host (the plasmid hypothesis) [1]. However, alternative hypotheses exist, such as that chromids can arise from a split of the chromosome (the schism hypothesis) [6], that the entire chromid is acquired through conjugation from another bacterium [7], or that the chromid arises through recombination between a chromid and a plasmid (chromid "rebirth") [1]. The majority of known chromids have originated from a plasmid or megaplasmid and have plasmid-like replication machineries. For example, in *Betaproteobacteria* the majority of chromids are found within the *Burkholderiaceae* family [8] and are thought to have originated from two ancestral plasmids. Similarly, in *Alphaproteobacteria*, most chromids are found within *Rhizobiaceae* and are believed to originate from a relatively small number of plasmids [1].

Exactly why 10% of the currently available bacterial genomes are multipartite, and which purpose the extra replicons may serve is still unclear. Several hypotheses have been suggested [1,2]. One hypothesis is that chromids acquire and loose genes more rapidly, thus providing bacteria with an increased genetic plasticity. This can be advantageous in terms of environmental specialization and niche-specificity [8–10]. For example, studies have suggested that the gene content of chromids varies more than in chromosomes [7,11], and thus evolve more rapidly and acquire new genes at a faster rate [8], and finally, experience more relaxed selection pressure (i.e., greater evolutionary plasticity) [12]. This hypothesis is also known as the test bed hypothesis [11]. Other suggested hypotheses are that chromids can contribute with replicon-specific gene regulation and expression [13–15], reduce the number of overlapping replication cycles required during fast growth [16] and that extra replicons are responsible for larger genomes and increased genome content [17].

Several different calculations can be performed to provide new insights into the plasticity of multipartite genomes, and potentially differentiate between the alternative hypotheses of their existence. One commonly used approach is to estimate the rate of growth of the so-called pangenome of a species (or genus or a family), also known as the "openness" of a genome [18]. The open or closed state of a pangenome depends on the ability of the bacteria to acquire new genes, for example, through horizontal gene transfer. In an open pangenome, new genes are added to the pangenome as more genomes are sequenced or added to the analysis. In contrast, a closed pangenome approaches a constant size as more genomes are added. Heap's law can be used to describe the pangenome size and number of new genes added for each new genome sequences and is formulated as: $n = kN^{\Upsilon}$, where n is the pangenome size, N is the number of genomes used and k and γ are the fitting parameters. If $\gamma < 0$, the pangenome is closed, and if $\gamma > 0$, the pangenome is open [19].

Another frequently used method to study the flexibility of genomes and horizontal gene transfer, is through calculation of codon usage. Codon usage can differ between organisms, as well as between genes of the same genome [20,21]. The typical codon usage of an organism, i.e., the preferential use of certain synonymous codons in typical genes, can be distinguished from the codon usage of highly expressed genes (optimal codon usage), and codon usage of horizontally transferred genes (HTGs) (atypical codon usage) [22,23]. Optimal codon usage corresponds to the use of the most abundant tRNAs in the organism, thus leading to faster translation (protein synthesis) [20]. HTGs on the other hand have a codon usage similar to its donor organism. To what extent the codon usage of an HTG deviates from the recipient genomes depends on how distantly related the donor and recipient genomes are. Variations in relatedness between the donor and recipient, as well as amelioration (that codon usage evolves towards that of the typical genome over time) are limitations that can lead to underestimation of HTGs [24].

Within *Gammaproteobacteria*, bipartite genomes are exclusively found in *Vibrionaceae* and *Pseudoalteromonas,* both of which belong to the *Enterobacterales* order (according to the Genome Taxonomy database (GTDB)) [25]. *Vibrionaceae* consists of eight genera, all of which have bipartite genomes, whereas *Pseudoalteromonas* is the only bipartite genus among the 44 genera within *Alteromonadaceae*. According to estimates of time since divergence, *Pseudoalteromonas* is much younger than *Vibrionaceae* [26,27]. Both the *Vibrionaceae* and the *Pseudoalteromonas* chromids are believed to have originated from plasmids from the same order [26,28–32]. The replication of chromosomes and chromids of *Vibrionaceae* have been heavily studied, with research showing that both replicons are bidirectionally replicated, and the replication is highly coordinated with synchronized termination of the replicons [16,33,34]. Replication of most *Pseudoalteromonas* chromids occur in an unidirectionally manner, while some are replicated bidirectionally. Additionally, the replication termination has been proposed to be synchronized [27]. We recently studied the global gene distribution and gene expression in *Vibrionaceae* [35] and *Pseudoalteromonas* [32]. Briefly, we calculated the pangenomes of 124 *Vibrionaceae* and 25 *Pseudoalteromonas* genomes, mapped the pangene categories on the genomes and compared the gene distribution with gene expression under fast and slow growth conditions. In both cases, core and softcore genes were overrepresented around the origin of replication (*ori1*), whereas shell and unique genes densely populated the regions surrounding the replication terminus (*ter1*). Gene expression strongly correlated with the distance to *ori1*, with higher expression levels closer to *ori1*. The *Vibrionaceae* chromids did not display any distinct gene distribution pattern. In contrast, the core genes of *Pseudoalteromonas* chromids were found to have a strong correlation with *ter2*, regardless if the chromid was replicated bi- or uni-directionally. Gene expression in chromids did not correlate with distance to *ori* or *ter*. Based on the subcellular organization of chromosome and chromid in *Vibrio cholerae* [15,16,36,37] we found that core/softcore and shell/cloud was spatially separated into separated intracellular regions (the poles of *V. cholerae*). This led us to propose a hypothesis that the bipartite genome structure enables intracellular spatial separation of different pangene categories and that there is a connection between gene placement and gene function.

Extensive research has been conducted on the maintenance and advantages provided by chromids in multipartite bacteria. Some hypotheses propose that chromids provide advantages such as replication specific gene regulation, increased gene content and reduced replication cycles during fast growth [13–17]. Other hypotheses suggest that chromids offer increased genomic plasticity and that they are a preferred location for horizontally transferred genes [8,9,11,12]. However, the extent to which chromosomes and chromids contribute to the overall plasticity and openness of bipartite genomes is not well understood. Our study aims to address this knowledge gap by calculating the openness of chromids and chromosomes of the bipartite bacteria *Vibrio* and *Pseudoalteromonas*, as well as monopartite genomes, and use codon usage and horizontal gene transfer analysis to determine which genes that contribute to the openness. Based on our data and two recent studies, we propose a hypothesis that describes how chromids and a specific region of the chromosomes appear to contribute to the genomic plasticity of bipartite genomes. Additionally, we establish the origin of *Vibrionaceae* and *Pseudoalteromonas* chromids.

2. Results

2.1. Vibrio and Pseudoalteromonas Belong to the Same Bacterial Order

The only known cases of bacteria with bipartite genomes within the class of *Gammaproteabacteria* are *Pseudoalteromonas* and *Vibrionaceae*. The overall phylogenetic relationship between bacterial families and their respective genera that form the order *Enterobacterales* are presented in Figure 1. The phylogenetic tree is based on information derived from GTDB release 89 [25], and lineages with bipartite genomes are highlighted.

The fact that *Vibrionaceae* and *Pseudoalteromonas* belong to the same order, raises the possibility, although unlikely, that their chromids originate from a single acquisition event in a common ancestor. Such a scenario would invoke a common origin followed by long-term retainment of the chromid, and then massive losses in all representatives of *Enterobacterales*, except *Vibrionaceae* and *Pseudoalteromonas*. A more likely explanation is that the chromids originate from two separate acquisition events.

2.2. Separate Origin of Chromids in Vibrionaceae and Pseudoalteromonas

We used ParA and ParB as phylogenetic markers to discriminate between the two hypotheses, i.e., a common or separate origin of the *Vibrionaceae* and *Pseudoalteromonas* chromids. ParA and ParB have fundamental roles in partitioning of replicons [38], and their conserved function and widespread distribution in bacteria and archaea make them suitable for establishing the origin of the chromids. A concatenated ParA–ParB alignment was created from sequences identified by BLASTp when using ParA and ParB sequences from *Pseudoalteromonas* and *Vibrionaceae* chromids as queries against the nr. protein database. The final dataset included a total of 376 residues from ParA and 313 residues from ParB (few residues were kept due to highly divergent regions that could not be reliably aligned).

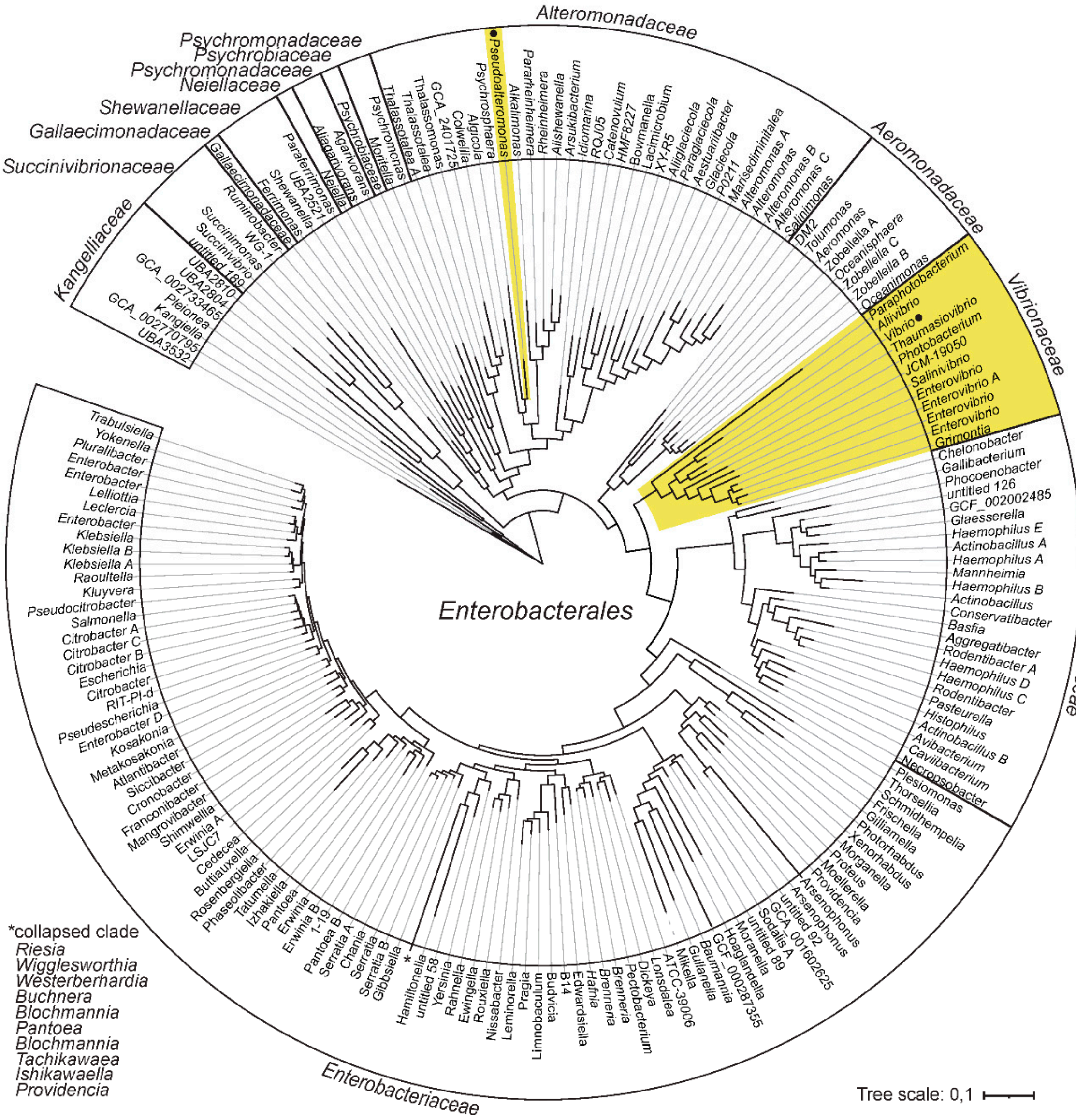

Figure 1. Phylogeny and distribution of bipartite genomes within *Enterobacterales*. Phylogenetic relationship between bacterial families and their respective genera are derived from the Genome Taxonomy database (GTDB). Lineages with bipartite genomes are highlighted in yellow, and genera investigated in this study are indicated with black dots.

The resulting maximum likelihood tree, based on the concatenated protein sequences of ParA and ParB and the WAG + G+I model, shows the evolutionary relationships between chromidal sequences from *Vibrio* and *Pseudoalteromonas* (Figure 2). Chromosomal sequences were used as the outgroup. Here, chromidal ParA–ParB from *Vibrionaceae* branches together with plasmid sequences from *Alteromonas*, *Pseudoalteromonas* and *Paraglaciecola* (Plasmid group 2), whereas chromidal *Pseudoalteromonas* ParA–ParB form a sister group with another set of plasmids, i.e., from *Shewanella*, *Vibrio* and *Pseudoalteromonas* (Plasmid group 1). These relationships are supported by bootstrap values of 90% and 75%, respectively. In summary, our result agrees with separate origins of the *Vibrionaceae* and *Pseudoalteromonas* chromids and suggests that both chromids were acquired from plasmids belonging to the *Enterobacterales* gene pool.

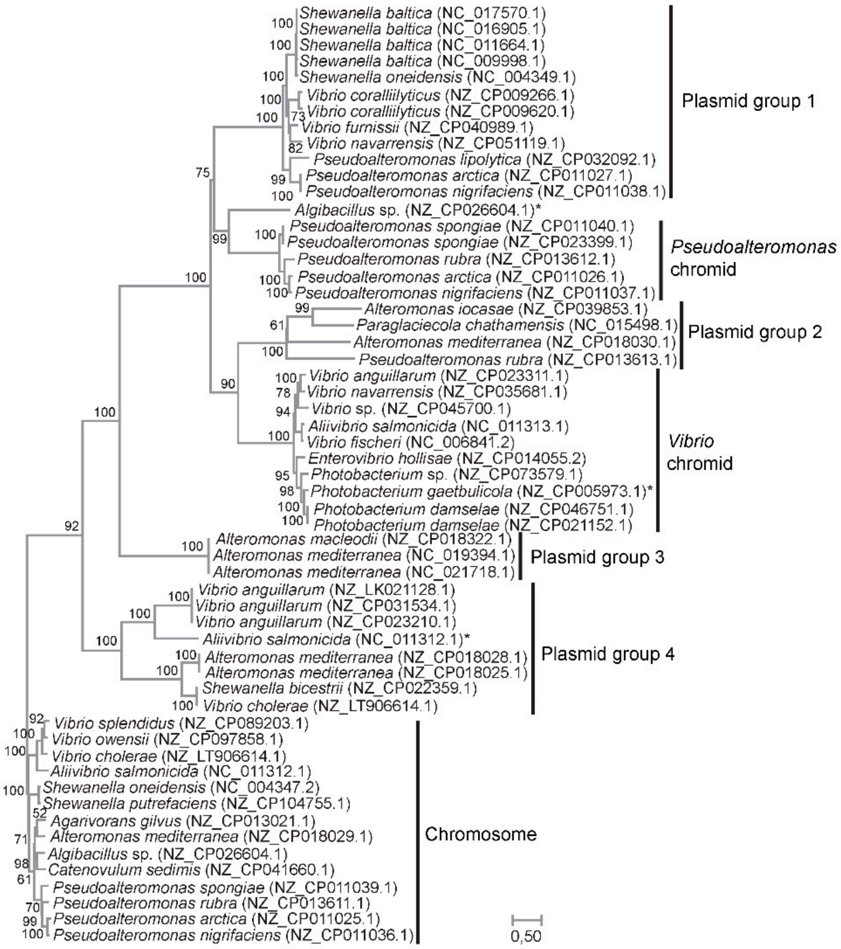
100
Shewanella baltica (NC_017570.1)
Shewanella baltica (NC_016905.1)
100
Shewanella baltica (NC_011664.1)
Shewanella baltica (NC_009998.1)
100
Shewanella oneidensis (NC_004349.1)
100
Vibrio coralliilyticus (NZ_CP009266.1)
100
73
Vibrio coralliilyticus (NZ_CP009620.1)
Vibrio furnissii (NZ_CP040989.1)
82
Vibrio navarrensis (NZ_CP051119.1)
Pseudoalteromonas lipolytica (NZ_CP032092.1)
75
99
Pseudoalteromonas arctica (NZ_CP011027.1)
100
Pseudoalteromonas nigrifaciens (NZ_CP011038.1)
Plasmid group 1
Algibacillus sp. (NZ_CP026604.1)*
Pseudoalteromonas spongiae (NZ_CP011040.1)
99
100
Pseudoalteromonas spongiae (NZ_CP023399.1)
Pseudoalteromonas rubra (NZ_CP013612.1)
100
100
Pseudoalteromonas arctica (NZ_CP011026.1)
100
Pseudoalteromonas nigrifaciens (NZ_CP011037.1)
Pseudoalteromonas chromid
99
Alteromonas iocasae (NZ_CP039853.1)
61
Paraglaciecola chathamensis (NC_015498.1)
Alteromonas mediterranea (NZ_CP018030.1)
100
Pseudoalteromonas rubra (NZ_CP013613.1)
Plasmid group 2
90
100
Vibrio anguillarum (NZ_CP023311.1)
78
Vibrio navarrensis (NZ_CP035681.1)
94
Vibrio sp. (NZ_CP045700.1)
100
Aliivibrio salmonicida (NC_011313.1)
100
Vibrio fischeri (NC_006841.2)
Enterovibrio hollisae (NZ_CP014055.2)
95
Photobacterium sp. (NZ_CP073579.1)
98
Photobacterium gaetbulicola (NZ_CP005973.1)*
100
Photobacterium damselae (NZ_CP046751.1)
100
Photobacterium damselae (NZ_CP021152.1)
Vibrio chromid
92
100
Alteromonas macleodii (NZ_CP018322.1)
Alteromonas mediterranea (NC_019394.1)
Alteromonas mediterranea (NC_021718.1)
Plasmid group 3
100
Vibrio anguillarum (NZ_LK021128.1)
100
Vibrio anguillarum (NZ_CP031534.1)
Vibrio anguillarum (NZ_CP023210.1)
100
Aliivibrio salmonicida (NC_011312.1)*
100
Alteromonas mediterranea (NZ_CP018028.1)
100
Alteromonas mediterranea (NZ_CP018025.1)
Shewanella bicestrii (NZ_CP022359.1)
100
Vibrio cholerae (NZ_LT906614.1)
Plasmid group 4
92
Vibrio splendidus (NZ_CP089203.1)
100
Vibrio owensii (NZ_CP097858.1)
100
Vibrio cholerae (NZ_LT906614.1)
Aliivibrio salmonicida (NC_011312.1)
100
Shewanella oneidensis (NC_004347.2)
Shewanella putrefaciens (NZ_CP104755.1)
52
Agarivorans gilvus (NZ_CP013021.1)
71
Alteromonas mediterranea (NZ_CP018029.1)
Algibacillus sp. (NZ_CP026604.1)
98
Catenovulum sedimis (NZ_CP041660.1)
61
Pseudoalteromonas spongiae (NZ_CP011039.1)
70
Pseudoalteromonas rubra (NZ_CP013611.1)
99
Pseudoalteromonas arctica (NZ_CP011025.1)
100
Pseudoalteromonas nigrifaciens (NZ_CP011036.1)
Chromosome
0,50

Figure 2. ML-tree based on the concatenated protein sequences of ParA and ParB and the WAG + G+I model. The tree shows the evolutionary relationships between chromidal sequences from *Vibrio* and *Pseudoalteromonas*, and sequences from plasmids carrying related ParA and ParB pairs. Chromosomal sequences were used as the outgroup. Clades containing plasmid sequences were designated Plasmid group 1–4 for clarity. Asterix denotes chromosomal sequences with an auxiliary pair of ParA and ParB. Bootstrap values (ML method, WAG + G+I model, 1000 pseudoreplicates) are associated with the nodes. Branch lengths are proportional to the number of substitutions per site (see scale).

2.3. The Chromids in Pseudoalteromonas and Vibrio Play a Significant Role in the Openness of the Two Genomes

It has been proposed that the main advantage of keeping multiple replicons is increased genetic flexibility, often termed "openness" (e.g., [8,11,12,32]). A commonly used method to estimate the openness of a pangenome, is to perform curve fitting of the pangenome size versus number of genomes using Heaps' law [18,19]. Heaps' law is formulated as $n = kN^{\gamma}$, where an exponent $\gamma > 0$ indicates an open pangenome, i.e., the pangenome will grow/gain genes as new genomes are sequenced and added to the analysis. An exponent $\gamma < 0$ indicates a closed pangenome that will not grow in size as new genomes are added. To estimate to what extent the chromosome and the chromid contribute to the pangenome openness we made two separate datasets consisting of 50 complete *Vibrio* and 26 complete *Pseudoalteromonas* genomes. The datasets are non-redundant, meaning that only one complete genome per available species was included (see Table S1 for complete list of bipartite genomes). We then calculated the pangenome size and Heaps' exponent for the chromosome, chromid and total genome (see Table S3). The pangenome of *Vibrio* consists of 822 core (encoded by all 50 genomes), 1505 softcore (encoded by $\geq$47 genomes), 8463 shell (encoded by $\leq$46 and $\geq$3 genomes), and 37,177 cloud (encoded by $\leq$2 genomes). The *Pseudoalteromonas* pangenome consists of 1386 core (encoded by all 26 genomes), 1787 softcore (encoded by ≥24 genomes), 5096 shell (encoded by ≤23 and ≥3 genomes), and finally 20,635 cloud (encoded by ≤2 genomes).

The calculated pangenome sizes are presented (Figure 3), with the sizes being relative to the number of genomes added (median of 100 randomly generated combinations of genome datasets). For both *Vibrio* and *Pseudoalteromonas*, the size of the chromosomal, chromidal and total genomes increase as more genomes are added to the analysis, more in the beginning of the curve and less after 10 genomes are added. The Heaps' exponent associated with the *Vibrio* chromid (0.668 ± 0.001) and the chromosome (0.660 ± 0.003) are virtually identical. This means that the two replicons are equally "open", but because of its bigger size, the chromosome hosts the majority of new genes. For *Pseudoalteromonas*, the chromid exponent (0.685 ± 0.007) is considerably larger than that of the chromosome (0.594 ± 0.002) and total genome (0.601 ± 0.003). With the highest Heaps' exponent, the chromid contributes considerably to the openness of the *Pseudoalteromonas* genome. In summary, we have used Heaps' law to evaluate the openness of the chromosome and chromid of *Vibrio* and *Pseudoalteromonas* by calculating the pangenome sizes and Heaps' exponents. The *Vibrio* chromosome and chromid are equally open, whereas the *Pseudoalteromonas* chromid is more open than the chromosome.

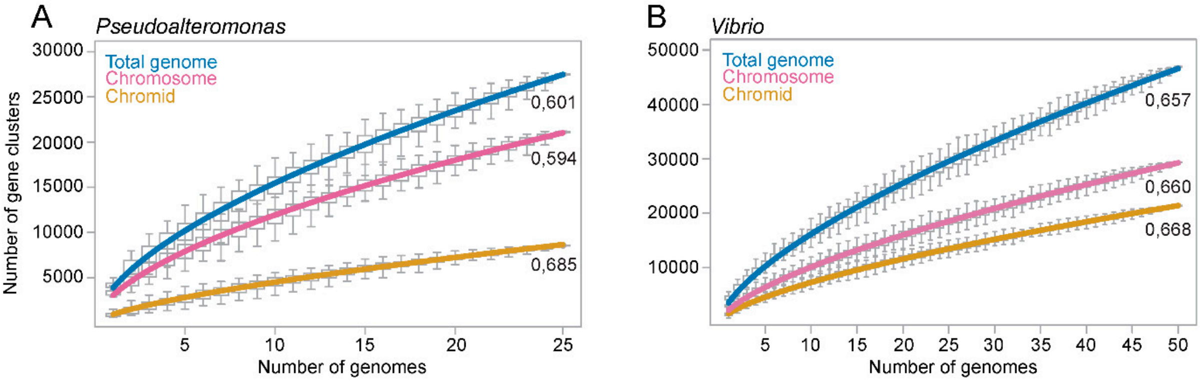

Figure 3. Graphs showing the calculated pangenome sizes of *Pseudoalteromonas* and *Vibrio* relative to the number of added genomes. For *Pseudoalteromonas* (**A**) and *Vibrio* (**B**), the number of gene clusters continues to grow as more genomes are added to the analysis, which shows that the chromids, chromosomes and total genomes are open. Each data point in the graph is based on the median of pangenome size of 100 randomly generated datasets (strain orders). The Heaps' exponents are shown associated with each graph and are used to evaluate the openness of the genomes.

2.4. Bipartite Genomes Are More Open Compared to Monopartite Genomes

Next, we compared the openness of the *Pseudoalteromonas* and *Vibrio* genomes to that of monopartite genomes of closely related genera. Hypothetically, the structural organization of genomes into one or multiple replicons can have a major impact on the flexibility of the genomes. The four relatively closely related genera *Alteromonas, Idiomarina, Rodentibacter* and *Yersinia* (all from *Enterobacterales*) with monopartite genomes were chosen for the analysis, for comparison to bipartite genomes (see Table S2 for complete list of monopartite genomes). For each genera, the Heaps' exponent was calculated from a random combination of an increasing number of genomes (using seven permutations) (see Table S3). This was conducted to test what effect the number of genomes and genome combinations have on the resulting Heaps' exponent. A dataset consisting of 27 *Escherichia coli* (species level) genomes was added as a control.

Plots with Heaps' exponent relative to the number of genomes for monopartite genomes are presented in Figure 4A. When the number of genomes is small, the distribution of Heaps' exponent is wide for *Yersinia, Alteromonas* and *Rodentibacter*, whereas for *Idiomarina*, the distribution is smaller. The corresponding plots for *Vibrio* and *Pseudoalteromonas,* show that the Heaps' exponent is widely distributed when only a few numbers of genomes are included in the datasets (Figure 4B). As the number of genomes increases, the exponents are less distributed (see Table S3 for complete list of Heaps' exponents). Similarly, the calculations for *Pseudoalteromonas* chromids vary greatly for small datasets but become more stable as the number of included genomes increases. These results show, as expected, that larger dataset (>10 genomes) result in more stable Heaps' values.

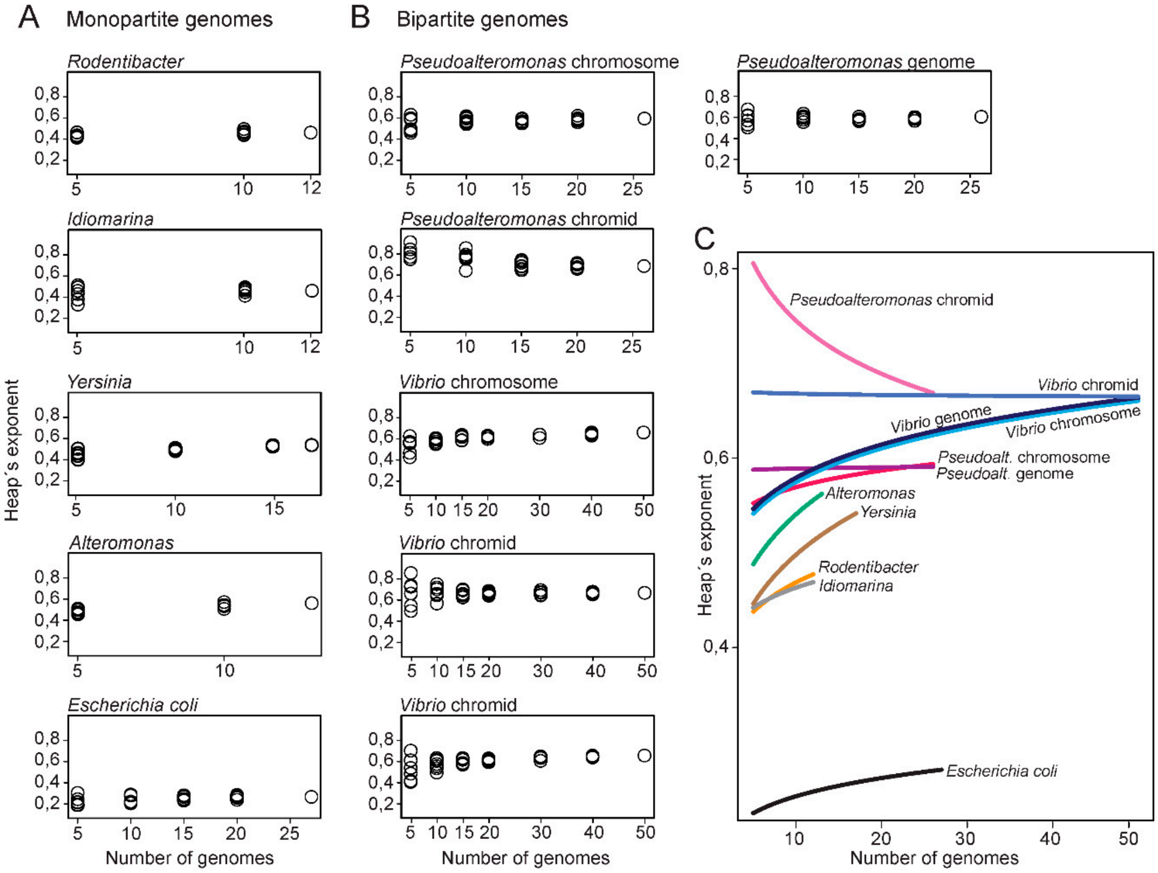

Figure 4. Plots of Heaps' exponents against the number of genomes. The analysis was carried out for datasets with monopartite (**A**) or (**B**) bipartite genomes. Each of the Heaps' exponents are made from the median number of pangenome sizes from 100 randomly generated strain orders. (**C**) Rarefaction curves of Heaps' exponents plotted against number of genomes. The curves can be regarded as a summary and of the results from (**A**,**B**) through curve fitting of the Heaps' exponents.

A summary of the results from Figure 4A,B through curve fitting of the Heaps' exponents, show that all bipartite replicons have larger Heaps' exponents compared to the monopartite genomes (Figure 4C). For example, at 10 genome datasets the lowest Heaps' value for bipartite are 0.618, whereas the highest Heaps' value for monopartite are 0.572. These results show that, with the currently available genomes, bipartite genomes have more open pangenomes, and thus appear more genetically flexible than monopartite counterparts. Chromids have the most open state of all replicons compared. Notably, how the exponent will change when more genomes become available is however unclear.

In summary, we plotted the Heaps' exponent relative to the size of genome datasets to compare openness of monopartite versus bipartite genomes. With the currently available datasets, bipartite genomes appear more open than that of closely related monopartite bacteria.

2.5. Codon Usage Is Specific for Each Pangene Category Rather Than for Each Replicon Type

Next, we used codon usage bias calculations to further explore the plasticity of bipartite genomes. Newly acquired genes are expected, in general, to have different codon usage profiles compared to those of most genes, especially genes with essential cellular roles (e.g., for cellular growth). Codon bias analyses are therefore used for exploring evolutionary aspects, including lateral transfer of genes.

Therefore, we first measured the relative synonymous codon usage (RSCU) for all individual genes in each of the 50 *Vibrio* and 26 *Pseudoalteromonas* genomes and performed a correspondence analysis of the RSCU values. Variations in codon usage among different pangene categories were explored by dividing the gene datasets into core, softcore, shell and cloud genes, and visualize the gene categories in different colors. Axis1 and Axis2 correlate with the two main influencing factors of codon usage bias. They represent 10.98% and 8.07% of the total variation for *Vibrio* and 10.97% and 7.52% of the total variation for *Pseudoalteromonas*, respectively.

Both *Vibrio* and *Pseudoalteromonas* have a broad distribution of codon usage, that are to a great extent specific for each pangene category (Figure 5A,B). In *Vibrio*, core and softcore genes are densely clustered toward the upper and lower right quadrants, whereas the shell and especially cloud genes are distributed towards upper left quadrant. In *Pseudoalteromonas*, core and softcore genes are distributed densely in upper left quadrant, shell genes toward the lower quadrants and in upper left quadrant.

PCA plots of the RSCU data described above (from Figure 5A,B) show that codon usage clusters based on pangene categories and not on the type of replicon (Figure 4C). This result is supported by correlation analysis of the RSCU values for each pangene category and analysis of median effective number of codons (ENC) for each pangene category (see Table S4 for global RSCU values and Table S5 for correlation plot and ENC values).

In summary, we performed COA and PCA on RSCU values to identify major trends of codon usage patterns in *Vibrio* and *Pseudoalteromonas*. Both type of plots show that codon usage is specific for each pangene category rather than type of replicon. This is valid for both *Pseudoalteromonas* and *Vibrio.* Similar codon usage for each pangene category indicates that they also have different evolutionary trajectories, which we explore further (see below).

2.6. Shewanella Represents the Top Donor of HTGs to Vibrio and Pseudoalteromonas

To identify putatively horizontally transferred genes (HTGs) in *Vibrio* and *Pseudoalteromonas*, we used HGTector [39], which is a software for genome-wide detection of horizontal gene transfer events based on homology searches. For *Pseudoalteromonas*, we defined horizontally transferred genes as all genes that originate from a donor outside of *Alteromonadaceae*, whereas for *Vibrio* horizontally transferred genes come from outside *Vibrionaceae*.

The number of HTGs detected for each pangene category on each replicon is presented in Figure 6A,B. HTGs comprise 11% and 23% of the total number of genes in the pangenomes in *Vibrio* [24,529 genes/7308 gene clusters (12 core, 32 softcore, 1496 shell, 4765 cloud)] and *Pseudoalteromonas* [19,970 genes/4310 gene clusters (309 core, 424 softcore, 2510 shell, 2389 cloud)], respectively. In *Vibrio,* the majority of HTGs (98%) are shell or cloud genes. These are distributed on the chromosome, where they make up 15% of shell and 13% of cloud genes, and on the chromid where they make up 20% (shell) and 16% (cloud). Notably, the *Vibrio* dataset contains 35 plasmids (from 19 genomes), of which 27% of shell genes and 13% of cloud genes are HTGs. For *Pseudoalteromonas*, about half of the HTGs are core and softcore genes. Of these, 15% and 18% of softcore genes are distributed on chromosomes and chromids, respectively. The other half of HTGs corresponds to chromosomal genes where they make up 24% of shell and 12% of cloud genes, respectively, and the corresponding numbers for chromidal genes are 30% (shell) and 13% (cloud). Six genomes contain one plasmid each. Here, 30% of HTGs represent shell and 14% represent cloud genes.

To summarize, in *Vibrio*, the identified horizontally transferred genes are typically shell and cloud genes located on both the chromosomes and chromids. In *Pseudoalteromonas,* the HTGs are more evenly distributed among all pangene categories from both chromosomes and chromids.

Phylogenetic distribution of the bacterial gene donors, i.e., the bacterial families from where the predicted HTGs originated from, show that in both *Vibrio* and *Pseudoalteromonas* the main contributors are families within the *Gammaproteobacteria* orders *Enterobacterales* and *Pseudomonadales* (Figure 6C,D). *Enterobacterales* and *Pseudomonadales* accounts for 66% and 22% of the total HTGs in *Vibrio* and 61% and 21% in *Pseudoalteromonas*, respectively. For *Pseudoalteromonas*, the top three donor genera are *Shewanella* (17%; *Shewanellaceae*), followed by *Vibrio* (11%; *Vibrionaceae*) and *Photobacterium* (5%; *Vibrionaceae*). Similarly, for *Vibrio* the top three donors are *Shewanella* (13%; *Shewanellaceae*), *Marimonas* (6%; *Marinomonadaceae*), and *Psychromonas* (6%; *Psychromonadaceae*).

In summary, we found that the majority of HTGs in *Vibrio* and *Pseudoalteromonas* originates from *Enterobacterales* and *Pseudomonadales*, with *Shewanella* representing the top donor of all genera.

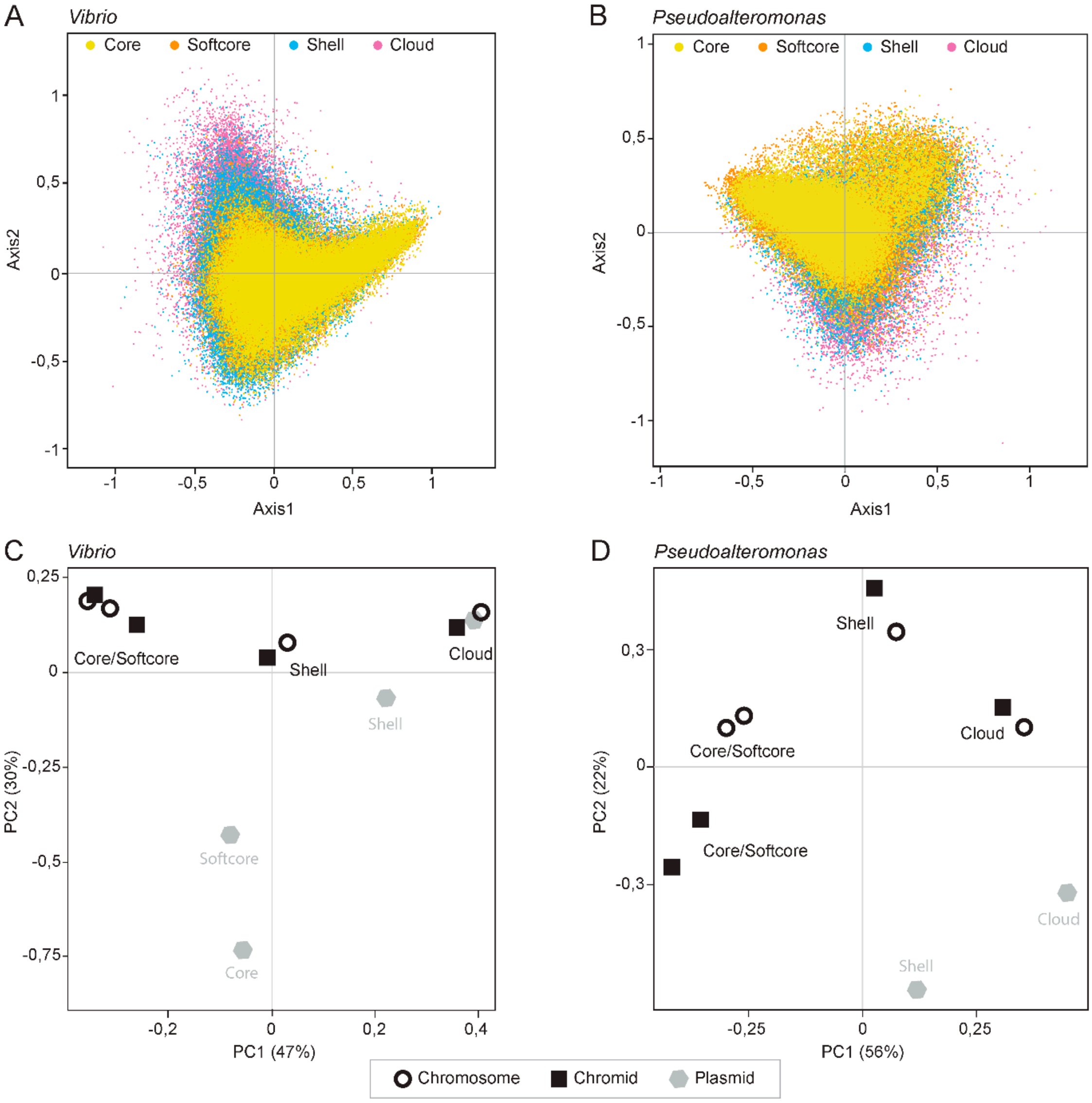

Figure 5. Correspondence analysis of relative synonymous codon usage (RSCU). The analyses are based on 50 *Vibrio* (**A**) and 26 *Pseudoalteromonas* (**B**) genomes. Core, softcore, shell and cloud genes are indicated with yellow, orange, blue and pink colors, respectively. The genes are distributed on primary and secondary axes which account for 10.98% and 8.07% in *Vibrio* and 10.97% and 7.52% *Pseudoalteromonas* of the total variation. Principal component analysis PCA) plots of the RSCU data from *Vibrio* (**C**) and *Pseudoalteromonas* (**D**) are shown. Both type of plots show that codon usage is specific for each pangene category rather than type of replicon.

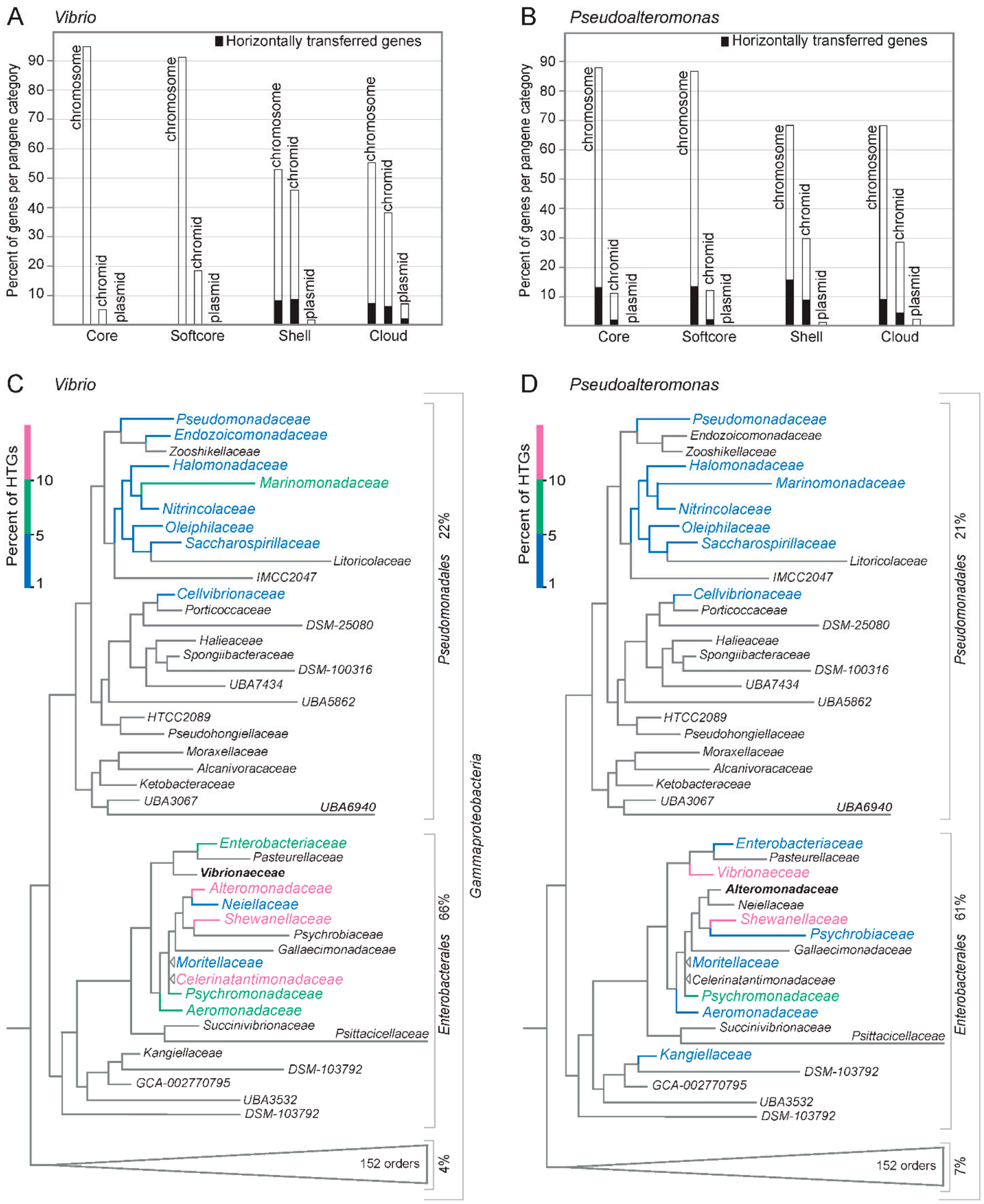

Figure 6. Horizontally transferred genes in *Vibrio* and *Pseudoalteromonas*, and the phylogenetic distribution of their donors. The number of HTGs in *Vibrio* (**A**) and *Pseudoalteromonas* (**B**) were predicted using the HGTector software. The data is shown as percentage of HTGs in each pangene category (core, softcore, shell and cloud), and also they are distributed among the three types of replicons (chromosomes, chromids and plasmids). HTGs were defined as genes with closest BLASTp hits outside of its family (i.e., *Vibrionaceae* and *Alteromonadaceae*, respectively). Next, the predicted bacterial donors of HTGs that reside in *Vibrio* (**C**) and *Pseudoalteromonas* (**D**) are shown mapped onto a phylogeny of *Gammaproteobacteria*. The top donors are shown in colorblindness-friendly color codes, from 1–5% (blue), 5–10% (green) and 10–15% (reddish purple). The majority of HTGs originates from other families within *Enterobacterales*, with *Shewanella* (at genus level) as the top donor to both *Vibrio* and *Pseudoalteromonas*.

3. Discussion

Here, we continue our studies on the bipartite genomes of *Vibrionaceae* and *Pseudoalteromonas*. According to GTDB, *Vibrionaceae* and *Pseudoalteromonas* both belong to *Enterobacterales* [25]. Based on an inferred ParAB phylogeny, we first established that the *Vibrio* and *Pseudoalteromonas* chromids do not share the same last common ancestor. The chromids originate from two separate plasmid acquisition events from plasmids within the *Enterobacterales* gene pool. We then calculated the pangenome and openness of the *Vibrio* and *Pseudoalteromonas* genomes and found that the *Vibrio* chromosome and chromid are equally open (i.e., the chromosome and chromid pangenome size increase at a similar rate as more genomes are added to the analysis), whereas the *Pseudoalteromonas* chromid is more open than the chromosome. Compared with monopartite genomes, bipartite are more open, at least based on today's available genome datasets. We next used codon usage bias calculations to elucidate which type of genes are more likely to have been acquired horizontally, thus leading to open bipartite genomes in *Vibrio* and *Pseudoalteromonas.* The data support that codon usage is specific to each pangene category regardless of which replicon they reside in. The vast majority of HTGs in *Vibrio* are shell or cloud genes, whereas HTGs in *Pseudoalteromonas* are more evenly distributed among all pangene categories.

By comparing the bipartite genomes of *Vibrio* and *Pseudoalteromonas* with monopartite genomes of related bacterial families, we showed that bipartite genomes appear more open than monopartite. The increased openness suggests that bipartite genomes have a higher capacity to acquire genes [40]. Using codon usage bias calculations and the HGTector tool we, therefore, set out to identify which type of genes are typically horizontally acquired by vibrios and pseudoalteromonases. We found that the codon usage in both *Vibrio* and *Pseudoalteromonas* group based on which pangene category genes belong to, and not based on which replicon genes reside on (chromidal or chromosomal placement). Notably, codon usage of cloud genes differs most from that of core genes (compared to shell genes), which are typically more highly expressed and therefore assumed to use codons better adapted to the translation machinery (adaption) [18,21]. This supports that cloud genes include a higher portion of more recently acquired genes. A similar pattern was reported for the multipartite bacterium *Sinorhizobium meliloti*, where codon usage of core genes on the chromosome and chromid were more similar than when compared to unique genes on the same replicons [41]. To conclude, less optimal codon usage of shell and cloud genes agree with data from our HGTector analysis, which suggests that as much as 98% of the detected HTGs in vibrios are either cloud or shell genes.

For *Pseudoalteromonas*, the general picture is similar, but here the HGTector result suggests that about half of the HTGs are core/softcore genes, whereas the other half corresponds to shell and cloud genes. The high proposition of HTGs among core/softcore is somewhat puzzling to us. To be detected as HTG, BLAST searches must identify the closest hit outside of *Alteromonadaceae*. We speculate that this result can be explained by the fact that *Pseudoalteromonas* is relatively young compared to *Vibrio* [502–378 vs. 1100–900 million years ago [26,27], respectively], and more genes will thus potentially be identified as HTG among core/softcore. The rationale is that HTGs in the last common ancestor (LCA) of extant *Pseudoalteromonas* bacteria have had approx. 500 million fewer years to adapt to the translation machinery than the corresponding genes in *Vibrio*. Moreover, *Pseudoalteromonas* have had less time to diverge from the LCA into different species, which subsequently can occupy various biological niches (such as *Vibrio*, that comprises at least 140 species). Consequently, our pangenome analyses identified 1386/1787 and 822/1505 core/softcore genes in *Pseudoalteromonas* and *Vibrio*, respectively. To summarize, HTGs in *Vibrio* are almost exclusively from the shell and cloud categories, whereas about half of HTGs in *Pseudoalteromonas* are shell and cloud genes.

Based on the results presented above, a new question arises: if a significant portion (>98% and >50%) of HTGs belong to the shell and cloud categories, where in the genomes are they typically located, and could their location explain why bipartite genomes are more flexible than monopartite genomes? In the light of this and previous studies, we suggest that the chromid and the lower half of the chromosome are particularly available for integration of new genes, and thus contribute to the elevated flexibility/openness of bipartite genomes (Figure 7). We recently mapped the pangene categories on the genomes of *Vibrionaceae* [35] and *Pseudoalteromonas* [32] and discovered distinct distribution patterns. On the chromosomes, core and softcore genes are overrepresented around the origin of replication (*ori1*), whereas shell and unique genes densely populate the regions surrounding the replication terminus (*ter1*). The *Vibrionaceae* chromids showed no clear gene distribution pattern, but for *Pseudoalteromonas*, the distribution of core genes strongly correlates with *ter2*, regardless of its position [i.e., *Pseudoalteromonas* chromids are replicated bi- or uni-directional, hence the position of *ter2* varies [27]]. Other studies have also found a correlation between density of mobile genetic elements and proximity to the *ter* region. Kopetja et al., discovered that in *Rhodobacterales*, core genes are located near *oriC*, whereas phages are located near the terminus [42]. A similar finding was reported by Oliviera et al. [43]. Using a diverse genome dataset, they found a higher frequency of "hot-spots" for horizontal gene transfer that contained prophages near *terC*. The evolutionary process responsible for this distribution pattern is discussed elsewhere [32,35], but from the current results we conclude that chromids and the lower halves of chromosomes appear to be favored "landing sites" for gene acquisition in bipartite genomes.

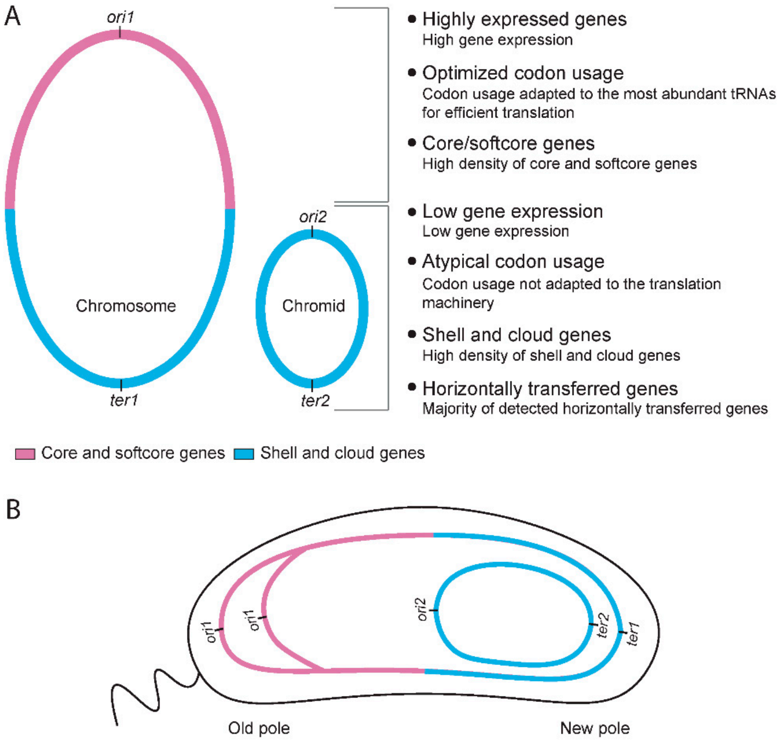

Figure 7. Summary of key characteristics of bipartite genomes in *Vibrio* and *Pseudoalteromonas*, and a putative model for accepted landing sites of HTGs. (**A**) Genes on the upper half of the chromosome are statistically more highly expressed, more likely to be core or softcore genes, and the codon usage is well adapted to the translational machinery. Genes located on the lower half of the chromosome, or the chromid, are statistically lower expressed, more likely to be shell or cloud genes, and have atypical codon usage less adapted to the translational machinery (compared to core/softcore). (**B**) Sketch of a hypothetical cell with a bipartite genome, and depicting the subcellular location of a chromosome and a chromid. The model is based on our pangenome calculations and genomic mapping of pangene types [32,35], and data from *V. cholerae* where the subcellular position of replicons have been determined [15,16,36,37]. Based on the genomic characteristics described in A, we hypothesize that chromids and the lower halves of the chromosomes are favored "landing sites" for gene acquisition in bipartite genomes.

4. Material and methods

4.1. Enterobacterales Reference Tree

The phylogenetic tree of *Enterobacterales* was made using Annotree [44], which is based on phylogeny and taxonomic nomenclature from the Genome Taxonomy database (GTDB) [25]. According to GTDB, *Pseudoalteromonas* and *Vibrionaceae* both group within the order *Enterobacterales*. Whereas following the NCBI taxonomy classification, *Vibrionaceae* and *Pseudoalteromonas* belong to separate orders (i.e., "*Vibrionales*" and *Pseudoalteromonadales*). Notably, in addition to multipartite genomes in *Vibrionaceae* and *Pseudoalteromonas*, there are reports of single strains with chromids in *Alteromonas mediterranea* [45] and in *Plesiomonas shigella* [46].

4.2. ParAB phylogenetic tree

BLASTp was used to compile ParA and ParB protein sequences from the databases using ParA and ParB from *Vibrionaceae* and *Pseudoalteromonas* as queries. The protein sequences were aligned using MUSCLE [47]. The alignment was manually adjusted using BioEdit [48], and only unambiguously aligned positions were kept for phylogenetic inference. A total of 689 aa positions were kept. MEGA11 was used to generate a Maximum Likelihood (ML) tree using the WAG model, Gamma distribution of evolutionary rates among sites, with invariant sites allowed (WAG + G + I) [49,50]. Bootstrap analysis with the same parameters as described above was performed with 1000 pseudoreplicates.

4.3. Genome Retrieval and Gene Annotation

One dataset for each of the genera *Pseudoalteromonas, Vibrio, Alteromonas, Yersinia, Idiomarina* and *Rodentobacter* and *E. coli* was made based on taxonomy of Genome Taxonomy database [25]. The genomes were downloaded from the RefSeq database at National Center for Biotechnology Information (NCBI) [51]. All *Vibrio* and *Pseudoalteromonas* genomes were complete (see Table S1 for complete lists of bipartite genomes). For a bipartite genome to be included in the study, its chromid had to meet the following criteria: it must possess a plasmid-type replication system, have a nucleotide composition close to that of the chromosome and contain core genes [1]. Direct evidence of the physical presence of chromids exist for *V. cholerae* [15,16,52]. and in *Pseudoalteromonas tunicata* and *Pseudoalteromonas spongiae* [27], all of which are included in the study. We allowed draft genomes with up to 200 contigs to be included for datasets of monopartite genomes (*Alteromonas, Yersinia, Idiomarina* and *Rodentobacter* and *E. coli*) (see Table S2 for complete list of monopartite genomes). All genomes were re-annotated using RAST (Rapid Annotation using Subsystem Technology) version 2.0 [53]. To make the datasets non-redundant, FastANI [54] was used to calculate average nucleotide identity values for all genomes against all genomes to select one genome per species.

4.4. Pangenome Calculation

To classify the annotated protein sequences of each of the seven datasets from *Pseudoalteromonas, Vibrio, Alteromonas, Yersinia, Idiomarina, Rodentobacter* and *E. coli* into four pangenome categories, we performed pangenome analysis using the clustering algorithm MCL in the software package GET_HOMOLOGUES (https://github.com/eead-csic-compbio/get_homologues, accessed on 15 August 2022)) [55]. The parameter "minimum percent sequence identity" was set to 50 and "minimum percent coverage in BLAST query/subj pairs" was set to 75 (default) [56]. To calculate the openness of pangenomes, pangenome analysis was performed using 100 permutations (for each datapoint). The median values of the combinations was used to perform curve fitting and calculate Heaps' exponent using power-law regression in the "aomisc package" in R v.4.0.3 [57] (see Table S3).

4.5. Calculation of Codon Usage

To investigate codon usage bias, codonW [58] was used to calculate relative synonymous codon usage (RSCU) and perform correspondence analysis of all genes in *Pseudoal-*

teromonas and *Vibrio.* Correspondence analysis (COA) was used to identify the major trends of codon usage among the four pangene categories. Each gene is described by a vector of 59 variables (codons) that correspond to the RSCU value of each synonymous codon. Codons without synonymous alternatives were excluded from the analysis (methionine, tryptophane and stop codons UAA, UAG, UGA). CodonW was also used to calculate global RSCU values of the pangenome categories separated based on their respective replicon (either chromosome, chromid or plasmid). The RSCU values were then plotted on a principal component analysis (PCA) (see Table S4 for global RSCU values). Effective number of codons was calculated using the R package "vhcub" [59] (see Table S5). ENC is used to estimate the overall codon bias for each gene in a dataset. ENC values range from 20 to 61, where all synonymous codons are used equally at 61 and only one codon used at 20 [60].

4.6. Prediction of Horizontally Transferred Genes

HGTector v2.0b3 [39] was used to identify putatively horizontally transferred genes in *Vibrio* and *Pseudoalteromonas*. A database consisting of 25,859 bacterial RefSeq proteins was downloaded from NCBI [51] and compiled using DIAMOND [61]. DIAMOND BLASTP searches with *Vibrio* pangenes and *Pseudoalteromonas* pangenes as queries was performed with the parameters e-value $< 1 \times 10^{-5}$, sequence identity > 30%, and sequence coverage > 50%. To search for horizontally transferred genes in *Pseudoalteromonas*, the parameter "self group" was set to *Pseudoalteromonas* (TaxID: 53246) and "close group" to *Alteromonadaceae* (TaxID: 226, 2848171, 135575, 28228, 1621534, 2071980, 336830, 2800384, 67575, 89404, 1249554, 111142, 2800384, 907197, 1518149, 366580, 1751872, 249523, 265980, 1407056, 2834759, 2125985, 296014, 1406885, 1172191, 137583, 2848177, 2661818, 2798470, 2851088). To search for horizontally transferred genes in *Vibrio*, the parameter "self group" was set to *Vibrio* (TaxID: 662) and "close group" was set to *Vibrionaceae* (TaxID: 641).

4.7. Statistical Analysis

Statistical analysis was performed using R in RStudio [62]. Correlation analysis was performed using the cor() function with Pearsons correlation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24054292/s1.

Author Contributions: P.H. and C.B.S. designed the study and wrote the manuscript. C.B.S. performed all bioinformatics analysis. Both authors contributed to proofreading and approved on the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the UiT The Arctic University of Norway with a PhD fellowship to C.B.S. The publication charges for this article have been funded by UiT The Arctic University of Norway. The funder had no role in study design, data collection, and interpretation, or the decision to submit the work for publication.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in Tables S1 and S2.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Harrison, P.W.; Lower, R.P.J.; Kim, N.K.D.; Young, J.P.W. Introducing the Bacterial "Chromid": Not a Chromosome, not a Plasmid. *Trends Microbiol.* **2010**, *18*, 141–148. [CrossRef]
2. DiCenzo, G.C.; Finan, T.M. The Divided Bacterial Genome. *Microbiol. Mol. Biol. Rev.* **2017**, *81*, e00019-17. [CrossRef] [PubMed]
3. Misra, H.S.; Maurya, G.K.; Kota, S.; Charaka, V.K. Maintenance of Multipartite Genome System and Its Functional Significance in Bacteria. *J. Genet.* **2018**, *97*, 1013–1038. [CrossRef] [PubMed]
4. Almalki, F.; Choudhary, M.; Azad, R.K. Analysis of Multipartite Bacterial Genomes Using Alignment Free and Alignment-Based Pipelines. *Arch. Microbiol.* **2023**, *205*, 25. [CrossRef] [PubMed]

5. Oren, A.; Garrity, G. Valid publication of the names of forty-two phyla of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **2021**, *71*, 004851. [CrossRef]
6. Egan, E.S.; Fogel, M.A.; Waldor, M.K. MicroReview: Divided Genomes: Negotiating the Cell Cycle in Prokaryotes with Multiple Chromosomes. *Mol. Microbiol.* **2005**, *56*, 1129–1138. [CrossRef]
7. Choudhary, M.; Cho, H.; Bavishi, A.; Trahan, C.; Myagmarjav, B. Evolution of Multipartite Genomes in Prokaryotes. In *Evolutionary Biology: Mechanisms and Trends*; Pontarotti, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 301–323.
8. Dicenzo, G.C.; Mengoni, A.; Perrin, E. Chromids Aid Genome Expansion and Functional Diversification in the Family *Burkholderiaceae*. *Mol. Biol. Evol.* **2019**, *36*, 562–574. [CrossRef] [PubMed]
9. Galardini, M.; Pini, F.; Bazzicalupo, M.; Biondi, E.G.; Mengoni, A. Replicon-Dependent Bacterial Genome Evolution: The Case of *Sinorhizobium Meliloti*. *Genome Biol. Evol.* **2013**, *5*, 542–558. [CrossRef]
10. diCenzo, G.C.; MacLean, A.M.; Milunovic, B.; Golding, G.B.; Finan, T.M. Examination of Prokaryotic Multipartite Genome Evolution through Experimental Genome Reduction. *PLoS Genet.* **2014**, *10*, e1004742. [CrossRef]
11. Cooper, V.S.; Vohr, S.H.; Wrocklage, S.C.; Hatcher, P.J. Why Genes Evolve Faster on Secondary Chromosomes in Bacteria. *PLoS Comput. Biol.* **2010**, *6*, e1000732. [CrossRef]
12. Feng, Z.; Zhang, Z.; Liu, Y.; Gu, J.; Cheng, Y.; Hu, W.; Li, Y.; Han, W. The Second Chromosome Promotes the Adaptation of the Genus *Flammeovirga* to Complex Environments. *Microbiol. Spectr.* **2021**, *9*, e00980-21. [CrossRef]
13. Dryselius, R.; Izutsu, K.; Honda, T.; Iida, T. Differential Replication Dynamics for Large and Small *Vibrio* Chromosomes Affect Gene Dosage, Expression and Location. *BMC Genom.* **2008**, *9*, 559. [CrossRef]
14. Couturier, E.; Rocha, E.P.C. Replication-Associated Gene Dosage Effects Shape the Genomes of Fast-Growing Bacteria but Only for Transcription and Translation Genes. *Mol. Microbiol.* **2006**, *59*, 1506–1518. [CrossRef] [PubMed]
15. Srivastava, P.; Chattoraj, D.K. Selective Chromosome Amplification in *Vibrio Cholerae*. *Mol. Microbiol.* **2007**, *66*, 1016–1028. [CrossRef] [PubMed]
16. Rasmussen, T.; Jensen, R.B.; Skovgaard, O. The Two Chromosomes of *Vibrio Cholerae* Are Initiated at Different Time Points in the Cell Cycle. *EMBO J.* **2007**, *26*, 3124–3131. [CrossRef]
17. Slater, S.C.; Goldman, B.S.; Goodner, B.; Setubal, J.C.; Farrand, S.K.; Nester, E.W.; Burr, T.J.; Banta, L.; Dickerman, A.W.; Paulsen, I.; et al. Genome Sequences of Three *Agrobacterium* Biovars Help Elucidate the Evolution of Multichromosome Genomes in Bacteria. *J. Bacteriol.* **2009**, *191*, 2501–2511. [CrossRef]
18. 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.; et al. Genome Analysis of Multiple Pathogenic Isolates of *Streptococcus Agalactiae*: Implications for the Microbial "Pan-Genome. " *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13950–13955. [CrossRef]
19. Tettelin, H.; Riley, D.; Cattuto, C.; Medini, D. Comparative Genomics: The Bacterial Pan-Genome. *Curr. Opin. Microbiol.* **2008**, *11*, 472–477. [CrossRef]
20. Ikemura, T. Codon Usage and TRNA Content in Unicellular and Multicellular Organisms. *Mol. Biol. Evol.* **1985**, *2*, 13–34. [CrossRef]
21. Plotkin, J.B.; Kudla, G. Synonymous but Not the Same. *Natl. Rev. Genet.* **2011**, *12*, 32–42. [CrossRef]
22. Tuller, T.; Girshovich, Y.; Sella, Y.; Kreimer, A.; Freilich, S.; Kupiec, M.; Gophna, U.; Ruppin, E. Association between Translation Efficiency and Horizontal Gene Transfer within Microbial Communities. *Nucleic Acids Res.* **2011**, *39*, 4743–4755. [CrossRef]
23. Komar, A.A. The Yin and Yang of Codon Usage. *Hum. Mol. Genet.* **2016**, *25*, R77–R85. [CrossRef]
24. Tuller, T. Codon Bias, TRNA Pools, and Horizontal Gene Transfer. *Mob. Genet. Elem.* **2011**, *1*, 75–77. [CrossRef]
25. Parks, D.H.; Chuvochina, M.; Rinke, C.; Mussig, A.J.; Chaumeil, P.A.; Hugenholtz, P. GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **2022**, *50*, D785–D794. [CrossRef]
26. Liao, L.; Liu, C.; Zeng, Y.; Zhao, B.; Zhang, J.; Chen, B. Multipartite Genomes and the SRNome in Response to Temperature Stress of an Arctic *Pseudoalteromonas Fuliginea* BSW20308. *Environ. Microbiol.* **2019**, *21*, 272–285. [CrossRef]
27. Xie, B.B.; Rong, J.C.; Tang, B.L.; Wang, S.; Liu, G.; Qin, Q.L.; Zhang, X.Y.; Zhang, W.; She, Q.; Chen, Y.; et al. Evolutionary Trajectory of the Replication Mode of Bacterial Replicons. *MBio* **2021**, *12*, e02745-20. [CrossRef]
28. Fournes, F.; Val, M.E.; Skovgaard, O.; Mazel, D. Replicate Once per Cell Cycle: Replication Control of Secondary Chromosomes. *Front. Microbiol.* **2018**, *9*, 1833. [CrossRef]
29. Heidelberg, J.F.; Elsen, 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.; et al. DNA Sequence of Both Chromosomes of the Cholera Pathogen *Vibrio Cholerae*. *Nature* **2000**, *406*, 477–483. [CrossRef]
30. Médigue, C.; Krin, E.; Pascal, G.; Barbe, V.; Bernsel, A.; Bertin, P.N.; Cheung, F.; Cruveiller, S.; D'Amico, S.; Duilio, A.; et al. Coping with Cold: The Genome of the Versatile Marine Antarctica Bacterium *Pseudoalteromonas Haloplanktis TAC125*. *Genome Res.* **2005**, *15*, 1325–1335. [CrossRef]
31. Rong, J.C.; Liu, M.; Li, Y.; Sun, T.Y.; Pang, X.H.; Qin, Q.L.; Chen, X.L.; Xie, B. Bin Complete Genome Sequence of a Marine Bacterium with Two Chromosomes, *Pseudoalteromonas Translucida KMM 520T*. *Mar. Genom.* **2016**, *26*, 17–20. [CrossRef]
32. Sonnenberg, C.B.; Haugen, P. The *Pseudoalteromonas* Multipartite Genome: Distribution and Expression of Pangene Categories, and a Hypothesis for the Origin and Evolution of the Chromid. *G3* **2021**, *11*, jkab256. [CrossRef]

33. Kemter, F.S.; Messerschmidt, S.J.; Schallopp, N.; Sobetzko, P.; Lang, E.; Bunk, B.; Spröer, C.; Teschler, J.K.; Yildiz, F.H.; Overmann, J.; et al. Synchronous Termination of Replication of the Two Chromosomes Is an Evolutionary Selected Feature in *Vibrionaceae*. *PLoS Genet.* **2018**, *14*, e1007251. [CrossRef]
34. Val, M.-E.; Marbouty, M.; de Lemos Martins, F.; Kennedy, S.P.; Kemble, H.; Bland, M.J.; Possoz, C.; Koszul, R.; Skovgaard, O.; Mazel, D. A Checkpoint Control Orchestrates the Replication of the Two Chromosomes of *Vibrio Cholerae*. *Sci. Adv.* **2016**, *2*, e1501914. [CrossRef]
35. Sonnenberg, C.B.; Kahlke, T.; Haugen, P. *Vibrionaceae* Core, Shell and Cloud Genes Are Non-Randomly Distributed on Chr 1: An Hypothesis That Links the Genomic Location of Genes with Their Intracellular Placement. *BMC Genom.* **2020**, *21*, 695. [CrossRef]
36. David, A.; Demarre, G.; Muresan, L.; Paly, E.; Barre, F.X.; Possoz, C. The Two Cis-Acting Sites, ParS1 and OriC1, Contribute to the Longitudinal Organisation of *Vibrio Cholerae* Chromosome I. *PLoS Genet.* **2014**, *10*, e1004448. [CrossRef]
37. Fogel, M.A.; Waldor, M.K. Distinct Segregation Dynamics of the Two *Vibrio Cholerae* Chromosomes. *Mol. Microbiol.* **2005**, *55*, 125–136. [CrossRef]
38. Jalal, A.S.; Tran, N.T.; Le, T.B. ParB Spreading on DNA Requires Cytidine Triphosphate in Vitro. *Elife* **2020**, *20*, e53515. [CrossRef]
39. Zhu, Q.; Kosoy, M.; Dittmar, K. HGTector: An Automated Method Facilitating Genome-Wide Discovery of Putative Horizontal Gene Transfers. *BMC Genom.* **2014**, *15*, 717. [CrossRef]
40. Medini, D.; Donati, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The Microbial Pan-Genome. *Curr. Opin. Genet. Dev.* **2005**, *15*, 589–594. [CrossRef]
41. López, J.L.; Lozano, M.J.; Lagares, J.A.; Fabre, M.L.; Draghi, W.O.; Del Papa, M.F.; Pistorio, M.; Becker, A.; Wibberg, D.; Schlüter, A.; et al. Codon Usage Heterogeneity in the Multipartite Prokaryote Genome: Selection-Based Coding Bias Associated with Gene Location, Expression Level, and Ancestry. *MBio* **2019**, *10*, e00505-19. [CrossRef]
42. Kopejtka, K.; Lin, Y.; Jakubovičová, M.; Koblízek, M.; Tomasch, J.; Moran, N. Clustered Core- And Pan-Genome Content on *Rhodobacteraceae* Chromosomes. *Genome Biol. Evol.* **2019**, *11*, 2208–2217. [CrossRef]
43. Oliveira, P.H.; Touchon, M.; Cury, J.; Rocha, E.P.C. The Chromosomal Organization of Horizontal Gene Transfer in Bacteria. *Nat. Commun.* **2017**, *8*, 841. [CrossRef] [PubMed]
44. Mendler, K.; Chen, H.; Parks, D.H.; Lobb, B.; Hug, L.A.; Doxey, A.C. AnnoTree: Visualization and Exploration of a Functionally Annotated Microbial Tree of Life. *Nucleic Acids Res.* **2019**, *47*, 4442–4448. [CrossRef]
45. López-Pérez, M.; Ramon-Marco, N.; Rodriguez-Valera, F. Networking in Microbes: Conjugative Elements and Plasmids in the Genus *Alteromonas*. *BMC Genom.* **2017**, *18*, 36. [CrossRef] [PubMed]
46. Adam, Y.; Brezellec, P.; Espinosa, E.; Besombes, A.; Naquin, D.; Paly, E.; Possoz, C.; van Dijk, E.; Barre, F.X.; Ferat, J.L. *Plesiomonas Shigelloides*, an Atypical *Enterobacterales* with a *Vibrio*-Related Secondary Chromosome. *Genome Biol. Evol.* **2022**, *14*, evac011. [CrossRef]
47. Edgar, R.C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
48. Hall, T.A. BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT.–ScienceOpen. *Nucleic Acids Symp. Ser.* **1999**, *41*, 95–98.
49. Stecher, G.; Tamura, K.; Kumar, S. Molecular Evolutionary Genetics Analysis (MEGA) for MacOS. *Mol. Biol. Evol.* **2020**, *37*, 1237–1239. [CrossRef] [PubMed]
50. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef] [PubMed]
51. O'leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; Mcveigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2015**, *44*, D733–D745. [CrossRef]
52. Val, M.-E.; Soler-Bistué, A.; Bland, M.J.; Mazel, D. Management of Multipartite Genomes: The *Vibrio Cholerae* Model. *Curr. Opin. Microbiol.* **2014**, *22*, 120–126. [CrossRef] [PubMed]
53. Aziz, R.K.; Bartels, D.; Best, A.A.; DeJongh, M.; Disz, T.; Edwards, R.A.; Formsma, K.; Gerdes, S.; Glass, E.M.; Kubal, M.; et al. The RAST Server: Rapid Annotations Using Subsystems Technology. *BMC Genom.* **2008**, *9*, 75. [CrossRef] [PubMed]
54. Jain, C.; Rodriguez-R, L.M.; Phillippy, A.M.; Konstantinidis, K.T.; Aluru, S. High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *Nat. Commun.* **2018**, *9*, 5114. [CrossRef]
55. Contreras-Moreira, B.; Vinuesa, P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Appl. Environ. Microbiol.* **2013**, *79*, 7696–7701. [CrossRef]
56. Costa, S.S.; Guimarães, L.C.; Silva, A.; Soares, S.C.; Baraúna, R.A. First Steps in the Analysis of Prokaryotic Pan-Genomes. *Bioinform Biol Insights* **2020**, *14*, 1177932220938064. [CrossRef]
57. Onofri, A. The Broken Bridge between Biologists and Statisticians: A Blog and R Package. 2020. Available online: https://www.statforbiology.com (accessed on 1 September 2022).
58. Peden, J.F. Analysis of Codon Usage. Ph.D. Thesis, University of Nottingham, Nottingham, UK, 2000.
59. Anwar, A.M.; Soudy, M. vhcub: Virus-Host Codon Usage Co-Adaptation Analysis. 2019. Available online: https://CRAN.r-project.org/package=vhcub (accessed on 1 September 2022).
60. Wright, F. The "effective Number of Codons" Used in a Gene. *Gene* **1990**, *87*, 23–29. [CrossRef] [PubMed]

61. Buchfink, B.; Xie, C.; Huson, D.H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* **2014**, *12*, 59–60. [CrossRef] [PubMed]
62. RStudio Team. RStudio: Integrated Development for R. 2021. Available online: http://www.rstudio.com/ (accessed on 6 September 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

PAPER 4

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

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

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

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

Accepted: April 6, 2018

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

Abstract

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

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

Introduction

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

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

Materials and Methods

Bacterial Isolates

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

Table 1
Complete *Vibrio anguillarum* Genomes

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

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

DNA Isolation and DNA Sequencing

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

Genome Analysis

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

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

Results and Discussion

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

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

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

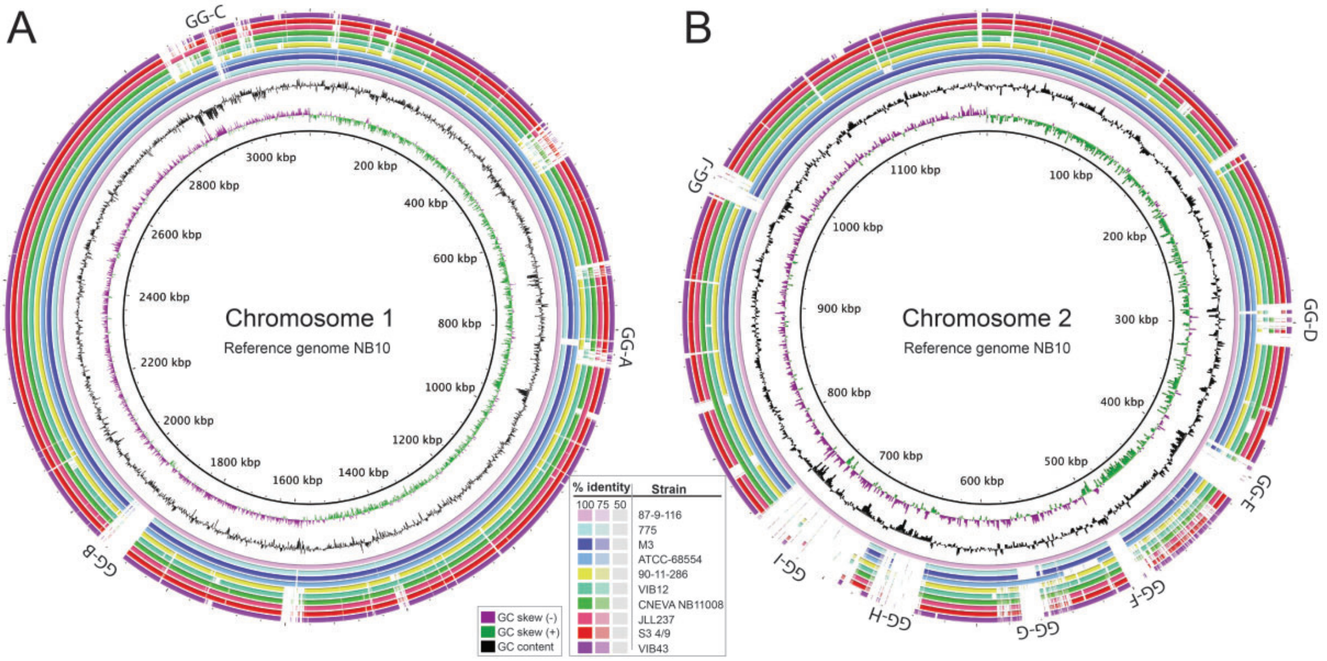

FIG. 1.—Comparison of 11 complete *Vibrio anguillarum* genomes. The figure was generated using the BLAST Ring Image Generator (BRIG) tool, and Chromosome 1 (*A*) and Chromosome 2 (*B*) of strain NB10 as central references (black ring in center). Genomic gaps (GG-A–GG-J) are the same as previously described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

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

Notably, according to the ATCC Bacteriology Collection, the strain names ATCC-68554 and 775 are synonymous. To verify this relationship ATCC-68554 was acquired directly from the ATCC Bacteriology Collection, cultured under standard conditions, and finally sequenced. A global pairwise genome comparison of the two genomes was done using the Artemis Comparison Tool (ACT) and the DOUBLE ACT v.2 server (Carver et al. 2005; see supplementary fig. S2, Supplementary Material online). This shows that both Chromosome 1 and Chromosome 2 sequences are highly similar, except for a region located approximately between positions 470,000–597,000 on Chromosome 2 (according to the ATCC-68554 sequence). This region represents a so-called "superintegron" (SI), which normally begins with an integron integrase, but in ATCC-68554 it is truncated and thus nonfunctional (locus tag CLI14_17245). *AttC*-containing regions (i.e., the SIs) are marked in yellow in supplementary figure S2C, Supplementary Material online. In ATCC-68554 this sequence is 54 kb longer than that of 775. In 775, the SI is followed by a 26-kb region, which is not present in ATCC-68554. The latter 775-specific sequence contains CDSs of various functions, and one tRNA-Gly gene. These discrepancies may be explained by technical artefacts during sequencing and assembly, or by real differences in the genomes, perhaps as a result of subculturing of the bacterium in different laboratories. The SIs and associated insertion sequences (ISs) are described in more detail below.

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

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

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

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

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

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

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

Supplementary Material

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

Acknowledgments

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

Literature Cited

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Associate editor: Howard Ochman