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
UiT Norges arktiske universitet
Tim Kahlke
University of Technology Sydney
Peik Haugen (peik.haugen@uit.no)
UiT Norges arktiske universitet https://orcid.org/0000-0002-4711-5513

Research article

Keywords: pangenome, genome architecture, Vibrionaceae, Aliivibrio salmonicida, Vibrio natriegens, gene dosage

Posted Date: September 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34687/v3

Version of Record: A version of this preprint was published at BMC Genomics on October 6th, 2020. See the published version at https://doi.org/10.1186/s12864-020-07117-5.

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.

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 1,796, 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% (1,489—1,796 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 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*

Fig. 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 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 [10, 21, 25]. 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 minutes at favourable conditions [26]. *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 [27, 28]. 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 [29] at NCBI. For *V. natriegens,* datasets from growth in minimal and optimal (rich) medium at 37 °C to mid log phase were chosen [30]. For *A. salmonicida*, a dataset originating from growth in LB medium containing 1% NaCl at 8 °C to mid log phase was used [31]. EDGE-PRO 1.3.1 [32] 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).

Fig. 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 [33]. 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.

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-adj ≤ 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, 34] and in several *Vibrio* species [16, 17, 21]. More precisely, Couturier and 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*. Overrepresentation 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) [35]. 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, 36, 37] and chromosome conformation capture (3C) [38]. 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 [39–43], nucleoid-associated proteins (NAPs) [44–46], Structural Maintenance of Chromosome complex (SMC) [47], RNA polymerase (RNAP) [48–52], transcription factors and promoter strength/chromosomal position [45, 53] 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 [40, 43, 48–52, 54, 55]. 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 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 [56]. 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 [57] (see Additional file 1 for a complete list). All genome sequences were re-annotated using RAST (Rapid Annotation using Subsystem Technology) version 2.0 [58] 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 [59] 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 [29] 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 [30]; 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) [31]. 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 [27, 28]. Hence, 8 °C was used in the experiment. The quality of the reads was checked using FastQC [60]. EDGE-pro v1.0.1 (Estimated Degree of Gene Expression in Prokaryotes) [32] 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.

Declarations

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 [30] and PRJEB17700 [31].

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Competing interests. The authors declare that they have no competing interests.

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.

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

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.

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. 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. diCenzo GC, Finan TM. The Divided Bacterial Genome: Structure, Function, and Evolution. Microbiol Mol Biol Rev. 2017;81:e00019-17.
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 Martins F, 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. Srivastava P, Chattoraj DK. Selective chromosome amplification in *Vibrio cholerae*. Mol Microbiol. 2007;66:1016–28.
26. Eagon RG. *Pseudomonas natriegens*, a marine bacterium with a generation time of less than 10 minutes. J Bacteriol. 1962;83:736–737.
27. Colquhoun DJ, Sørum H. Temperature dependent siderophore production in *Vibrio salmonicida*. Microb Pathog. 2001;31:213–9.
28. 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.
29. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39:D19–D21.
30. 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.
31. 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.
32. Magoc T, Wood D, Salzberg SL. EDGE-pro: estimated degree of gene expression in prokaryotic genomes. Evol Bioinform Online. 2013;9:127–36.
33. 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.
34. Ardell DH, Kirsebom LA. The genomic pattern of tDNA operon expression in *E. coli*. PLoS Comput Biol. 2005;1:0086–99.
35. Surovtsev I, Jacobs-Wagner C. Subcellular organization: a critical feature of bacterial cell replication. Cell. 2018;172:1271–93.

36. Fogel MA, Waldor MK. Distinct segregation dynamics of the two *Vibrio cholerae* chromosomes. Mol Microbiol. 2005;55:125–36.
37. 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.
38. Val ME, Marbouty M, de Lemos Martins F, 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.
39. Martis B. S, 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.
40. Dorman CJ. DNA supercoiling and transcription in bacteria: a two-way street. BMC Mol Cell Biol. 2019;20:26.
41. Dorman CJ, Dorman MJ. DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. Biophys Rev. 2016;8:209–20.
42. 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.
43. 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.
44. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nature Reviews Microbiology. 2010;8:185–95.
45. 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.
46. Dame RT, Tark-Dame M. Bacterial chromatin: Converging views at different scales. Curr Opin Cell Biol. 2016;40:60–5.
47. Brandão HB, Paul P, Berg AA Van Den, 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. Engstrom MD, Pfleger BF. Transcription control engineering and applications in synthetic biology. Synth Syst Biotechnol. 2017;2:176–91.
54. Bremer H, Dennis PP. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. EcoSal Plus. 2008;3.
55. Le TBK, Imakaev M V., Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. Science. 2013;342:731–4.
56. 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;168:172-185.e15.
57. 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:D733-745.
58. 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.

59. 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.

60. Andrews S. FastQC. Babraham Bioinformatics. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Last accessed 27 April 2020.

Table

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	6 473	13 656	23 238	23 238	17 161	5 533	35 274	35 274	28 737	4 049
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	8 098	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.

Figures

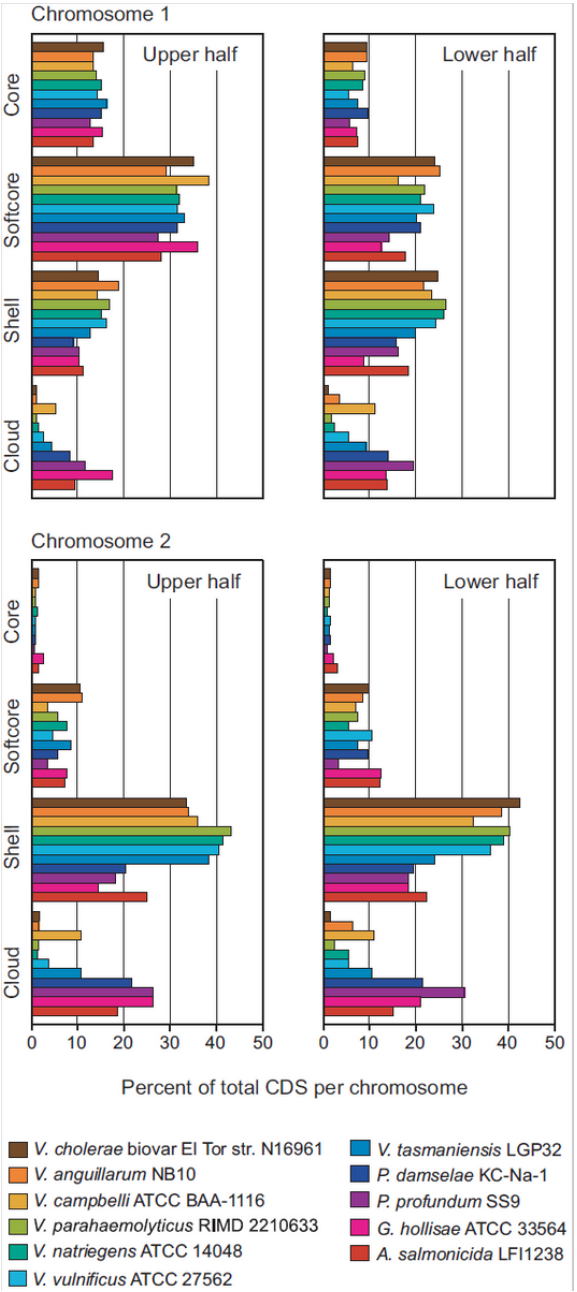

Figure 1

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.

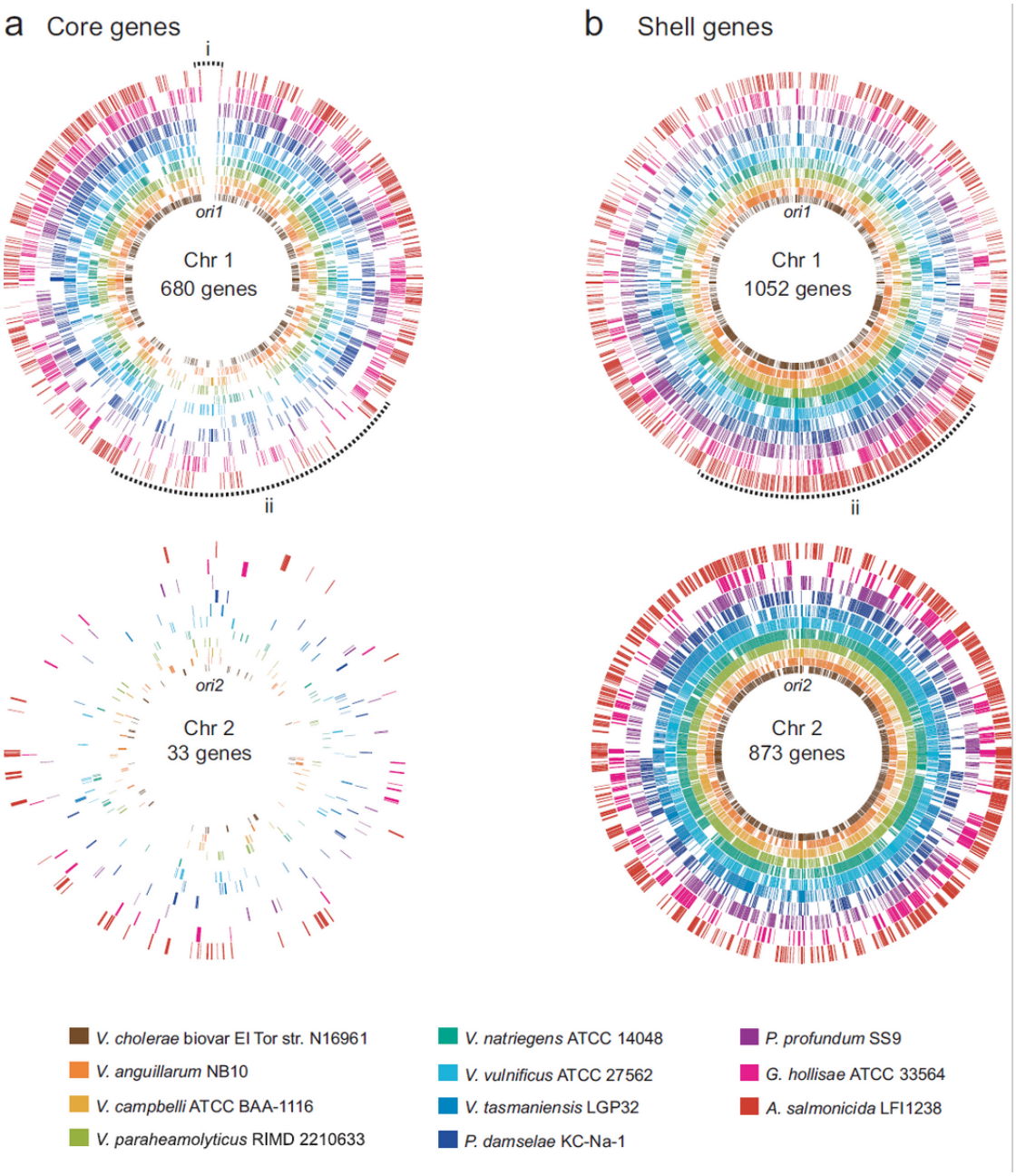

Figure 2

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.

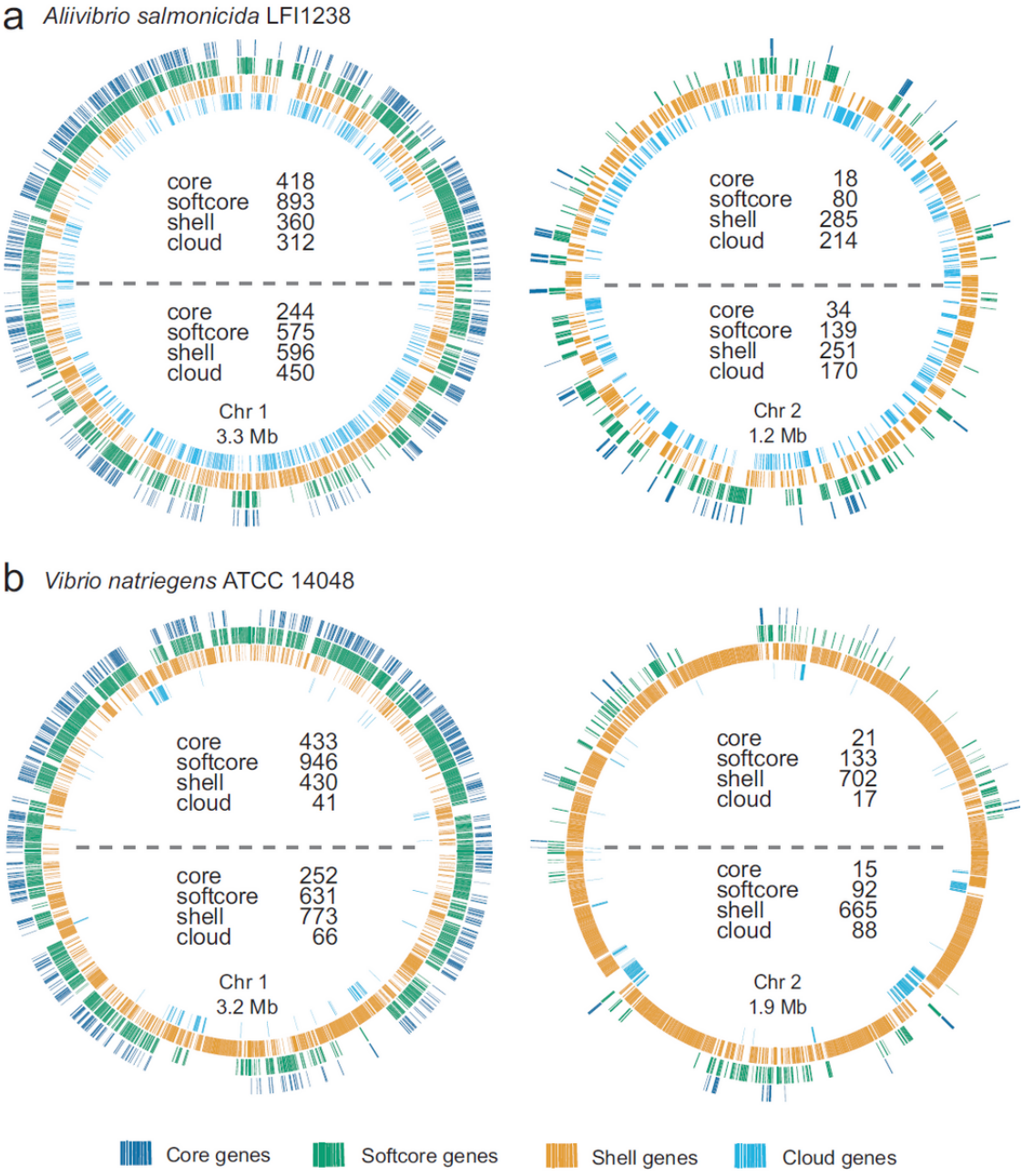

Figure 3

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.

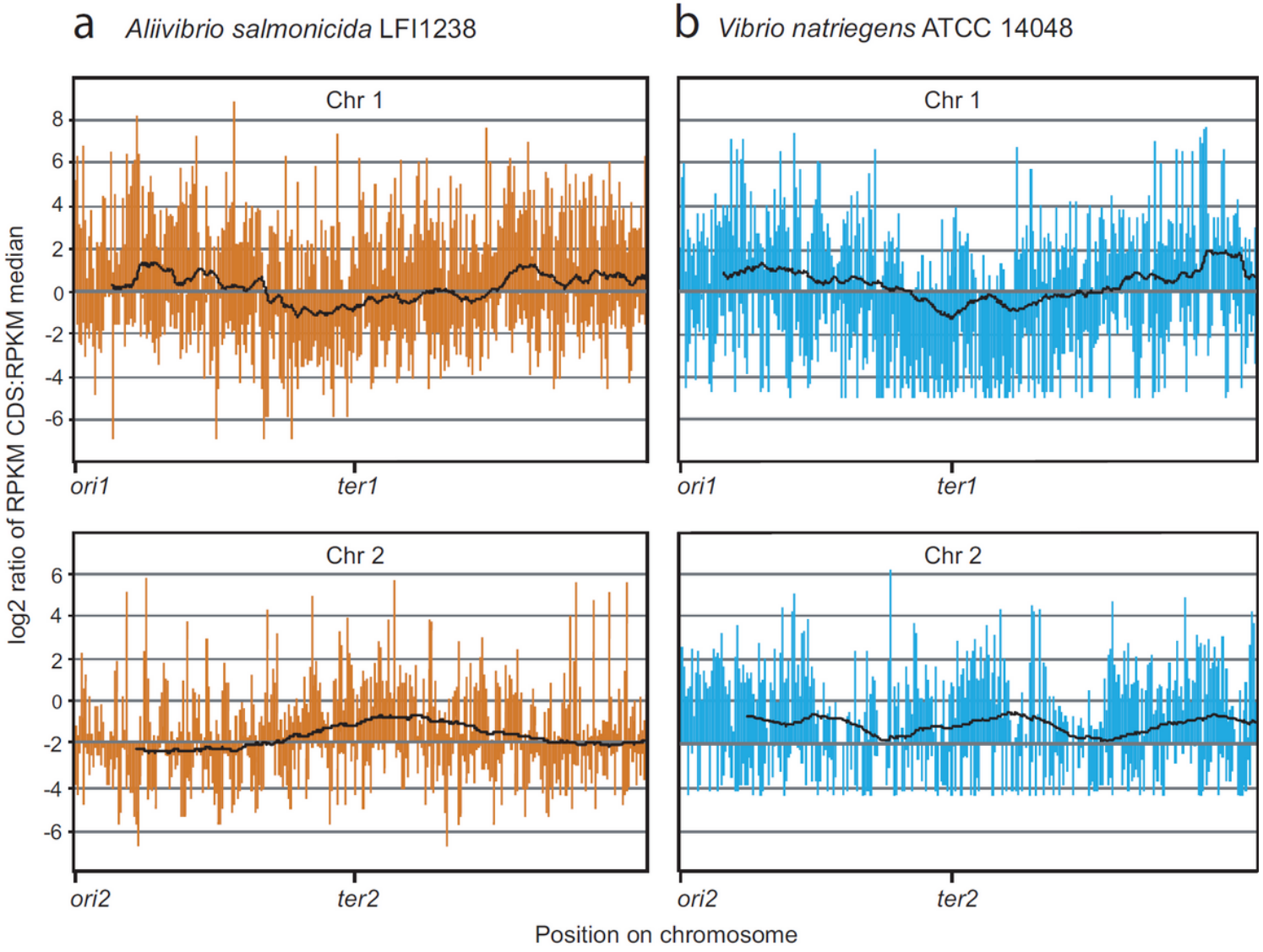

Figure 4

Global expression maps of (a) A. salmonicida LFI1238 and (b) V. natriegens ATCC 14048 chromosomal genes centered around the median. Data points (log2 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.

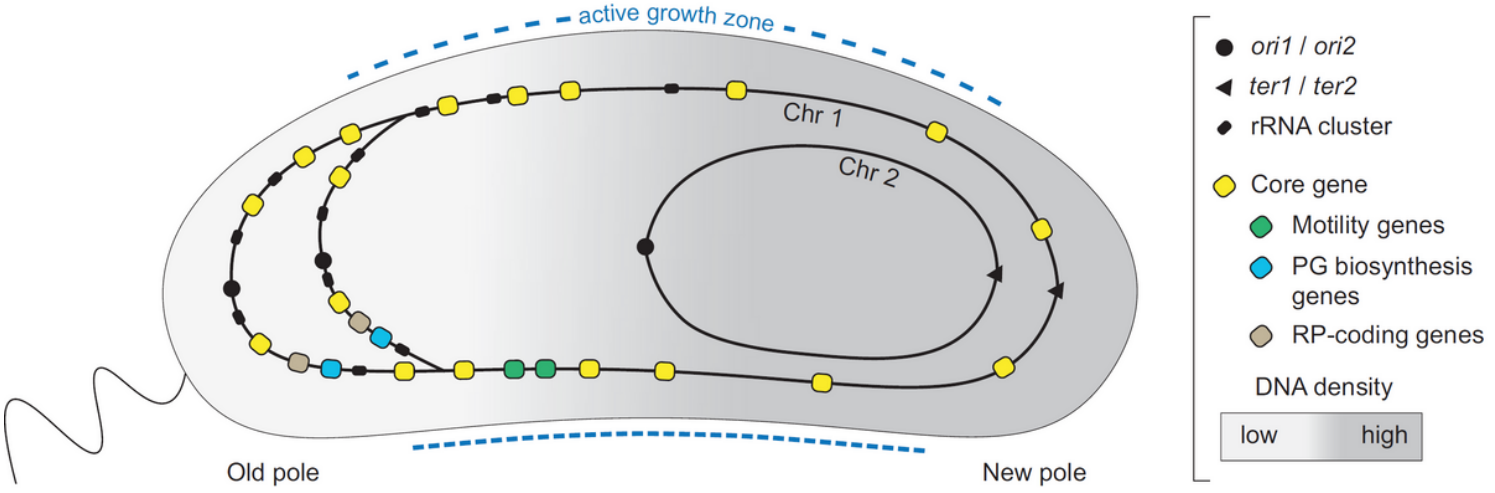

Figure 5

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1TableS1.xlsx
- Additionalfile2.Fig.S1.pdf
- Additionalfile3TableS2.xlsx
- Additionalfile4TableS3.xlsx
- Additionalfile5Fig.S2.pdf
- Additionalfile6Fig.S3.pdf