Switching on the light: using metagenomic shotgun sequencing to characterize the intestinal microbiome of Atlantic cod

Even Sannes Riiser,1* Thomas H.A. Haverkamp,1* Srinidhi Varadharajan,1 Ørnulf Borgan,2 Kjetill S. Jakobsen,1 Sissel Jentoft1 and Bastiaan Star1*
1Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, PO Box 1066, Blindern, N-0316 Oslo, Norway.
2Department of Mathematics, University of Oslo, PO Box 1053, Blindern, N-0316 Oslo, Norway.

Summary
Atlantic cod (Gadus morhua) is an ecologically important species with a wide-spread distribution in the North Atlantic Ocean, yet little is known about the diversity of its intestinal microbiome in its natural habitat. No geographical differentiation in this microbiome was observed based on 16S rRNA amplicon analyses, yet such finding may result from an inherent lack of power of this method to resolve fine-scaled biological complexity. Here, we use metagenomic shotgun sequencing to investigate the intestinal microbiome of 19 adult Atlantic cod individuals from two coastal populations in Norway—located 470 km apart. Resolving the species community to unprecedented resolution, we identify two abundant species, Photobacterium iliopiscarium and Photobacterium kishitanii, which comprise over 50% of the classified reads. Interestingly, the intestinal P. kishitanii strains have functionally intact lux genes, and its high abundance suggests that fish intestines form an important part of its ecological niche. These observations support a hypothesis that bioluminescence plays an ecological role in the marine food web. Despite our improved taxonomical resolution, we identify no geographical differences in bacterial community structure, indicating that the intestinal microbiome of these coastal cod is colonized by a limited number of closely related bacterial species with a broad geographical distribution.

Introduction
The fish intestinal microbiome comprises a complex and specialized gut bacterial community providing a multitude of biological functions in the host, including metabolism, growth, development, and immunity (reviewed in Wang et al., 2017; Ghanbari et al., 2015; Sullam et al., 2012; Izvekova et al., 2007). For instance, studies of laboratory-reared zebrafish have demonstrated that the intestinal microbiome regulates 212 genes stimulating gut epithelial proliferation, promotion of nutrient metabolism, and innate immune responses (Rawls et al., 2004). Moreover, several studies of aquaculture freshwater fish have shown that gut bacterial communities produce a wide range of digestive enzymes (Sugita et al., 1997; Bairagi et al., 2002) and is involved in synthesis of vitamins (Sugita et al., 1991). Despite this known biological importance, the composition of the intestinal microbiome in wild fish populations remains poorly understood. To date, studies of the fish intestinal microbiome have revealed a limited phylogenetic diversity, with genera from Proteobacteria, Firmicutes, and Bacteroidetes constituting up to 90% of the sequence reads across different species (Verner-Jeffreys et al., 2003; Ward et al., 2009; Ghanbari et al., 2015; Givens et al., 2015; Riiser et al., 2018; Talwar et al., 2018). Apart from this relatively low bacterial diversity, several studies have reported a limited geographical differentiation between intestinal bacterial communities, indicating a strong influence of host-associated factors on the composition of the gut microbiome (Ye et al., 2014; Llewellyn et al., 2016; Riiser et al., 2018). Nevertheless, most studies have been limited either because of their focus on cultured fish species (Desai et al., 2012; Wu et al., 2013; Zarkasi et al., 2014, 2016; Schmidt et al., 2016; Dehler et al., 2017) or because of methodological approaches that offer limited taxonomical resolution (e.g. 16S rRNA amplicon sequencing (Star et al., 2013; Ye et al., 2014; Llewellyn et al., 2016; Riiser et al., 2018) or dependence on bacterial cultivation (Kim et al., 2007; Martin-Antonio et al., 2007; Valdenegro-Vega et al., 2013). Therefore, there remains a
lack of detailed, baseline compositional data comparing healthy wild fish from the same species that live in different habitats with a variety of environmental conditions (Uchii et al., 2006; Egerton et al., 2018).

Atlantic cod (Gadus morhua) is an economically, ecologically, and culturally important species of the North Atlantic Ocean and represents a unique study system of the fish gut microbiome for fundamental as well as applied purposes. First, Atlantic cod, as well as the whole gadiform lineage, has lost the Major Histocompatibility Complex (MHC) II of the adaptive immune system (Star et al., 2011; Malmstrøm et al., 2016). This species also has an altered set of Toll-like receptors (TLRs), with a lack of TLR 1, 2, 3, and 4, and gene expansions of the intracellular TLR 7, 8, and 9 (Star et al., 2011; Malmstrøm et al., 2016; Solbakken et al., 2016). These components of the adaptive and innate immune system are specifically involved in bacterial and viral recognition, hence likely affect the interaction between Atlantic cod and its intestinal microbiome (Star et al., 2011; Star and Jentoft, 2012; Malmstrøm et al., 2016; Solbakken et al., 2016). Second, Atlantic cod is exposed to a variety of environmental conditions (e.g., salinity and temperature) due to its ability to exploit a wide range of ecological niches (Righton et al., 2010), which in turn may influence the composition of the host microbiome. It has a large geographical distribution, which comprises various subpopulations with divergent migratory and feeding behaviour (Cohen et al., 1990; Gode and Michalsen, 2000; Michalsen et al., 2008; Link et al., 2009), and hence possibly distinctive gut microbiomes. Finally, there have been significant investments to domesticate Atlantic cod for aquaculture purposes. Various factors have prevented this industry to be profitable, for instance through difficulties in immunization of juvenile cod (Samuelsen et al., 2006; Froese and Pauly, 2012), but also through to an inefficient digestion of formulated food of larvae in the prestomach stage (Hamre, 2006; Lie et al., 2018). Providing baseline data of the natural composition of intestinal microbiome in Atlantic cod may help efforts to improve the profitability of this industry.

The intestinal microbiome of Atlantic cod has so far been studied using both culture-based methods (Ringo et al., 2006; Dhanasiri et al., 2011) and culture-independent methods based on 16S rRNA amplicon sequencing (Star et al., 2013; Riiser et al., 2018). These methods show an abundance of Bacteroidales, Erysipelotrichales, Clostridiales and especially Vibrionales (Ringo et al., 2006; Dhanasiri et al., 2011; Star et al., 2013; Riiser et al., 2018). A single Vibrionales oligotype was found to numerically dominate the Atlantic cod intestinal microbiome, comprising more than 50% of all the sequence data (Riiser et al., 2018), suggesting that these microbiomes are not particularly complex. It is well known however, that 16S rRNA-based analyses can be confounded by amplification bias, 16S rRNA gene copy number variation and a lack of taxonomic resolution (Konstantinidis et al., 2006; Liu et al., 2008; Youssef et al., 2009; Vasileiadis et al., 2012; Shaya et al., 2013; Birtel et al., 2015; Amore et al., 2016; Noecker et al., 2016; Zhang et al., 2018). It has been found that 16S rRNA has an especially low power in distinguishing various Vibrionales species (Sawabe et al., 2007; Machado and Gram, 2015), and therefore substantial species differentiation may exist in these communities in the absence of 16S rRNA divergence (Konstantinidis et al., 2006; Noecker et al., 2016). These limitations can be mitigated by the use of shotgun metagenomics, which offers enhanced detection of bacterial species, a better estimation of diversity, and a more in-depth insight into the functional composition of microbiomes (Lewellyn et al., 2014; Romero et al., 2014; Ghanbari et al., 2015; Merrifield and Rodlides, 2015; Colston and Jackson, 2016; Ranjan et al., 2016; Tamecki et al., 2017). Despite these advantages, however, only a handful of studies has used metagenomics approaches to investigate the intestinal microbiome in fish, and the existing studies are all limited in their number of samples investigated, their community characterization at the lower taxonomical levels (i.e., species) or geographical sampling range, with a focus on Pacific aquaculture species (Xing et al., 2013; Xia et al., 2014; Hennersdorf et al., 2016; Tyagi et al., 2019). Nevertheless, there exist no studies that use metagenomic shotgun sequencing to characterize the geographical structure and community complexity in the intestinal microbiome of wild fish.

Here, we investigate the intestinal microbial community structure of 19 adult individuals of coastal Atlantic cod from different habitats in Norway, located 470 km apart (Fig. 1A) using metagenomic shotgun sequencing. No geographical differentiation of the intestinal microbiome between these locations was previously observed based on 16S rRNA amplicon sequencing (Riiser et al., 2018), providing an opportunity to test the enhanced resolution of shotgun metagenomics in a spatial and environmental context. First, we compare the genome-wide taxonomic composition and diversity based on metagenomic shotgun sequencing to that of the 16S rRNA marker-gene analysis. Second, we assess strain-level variation of the most abundant bacterial members of the intestinal community by using reference-based read mapping and comparing genome-wide single nucleotide variation. Finally, we explore the genome-wide coverage of the two most abundant bacterial strains in the Atlantic cod intestines to infer the functionality of specific genes and loss of genes.

Results

The Atlantic cod intestinal microbiome order-level composition

We obtained a data set of 198 million paired-end reads from 19 specimens caught in the coastal waters of Lofoten...
After quality trimming, the number of reads for each specimen varied from 835,000 to 7,000,000 reads (average 3 million sequences), comprised between 18.2% and 91.5% (mean: 66.7%) of host (Atlantic cod) DNA and between 8.5% and 81.8% (mean: 33.3%) bacterial DNA (Table 1, Supporting Information Table S1). 80% of the paired-end reads were classified, of which 96% at the order level (Supporting Information Table S2). The community profiles, based on non-normalized read counts, show a large overlap when clustering individuals from Lofoten and Sørøya using multivariate nonmetric multidimensional scaling (NMDS, Fig. 1B). The Atlantic cod intestinal microbiome is numerically dominated by bacteria of the order Vibrionales, which has a mean relative abundance of 81.8% and represents >76% of the reads in all except four individuals (Fig. 2A, Table 2). In relative abundance, this order is followed by Alteromonadales (3.6%), Fusobacteriales (3.1%), Clostridiales (2.9%) and Bacteroidales (1.7%). In total, the five orders with highest relative abundance constitute 94% of all classified sequences. A 16S rRNA-based analysis from the same locations shows that Vibrionales are the most abundant, followed by Fusobacteriales, Clostridiales, Bacteroidales and Alteromonadales (Fig. 2B, Table 2, reproduced from Riiser et al., 2018). A statistical comparison detects significant differences in the classification of the Atlantic cod intestinal microbiome (ANOVA for compositional data, \( p = 10^{-10} \)) for either the metagenomic shotgun sequencing or 16S rRNA-based classification.

The individual samples vary in diversity estimated by Shannon (H), Simpson (D) and Inverse Simpson (1/D) indices based on nonnormalized order-level read counts (Supporting Information Fig. S1 and Table S3). The variation in alpha diversity is reflected in the abundance profile in Fig. 2A, where in particular, four Lofoten samples (01, 04, 05, 09) and one Sørøya sample (09) contain higher relative abundances of orders other than Vibrionales. Top-down reduction of linear regression models based on the alpha diversity indices ends up with models containing no significant covariates (Supporting Information Table S4), indicating that neither location, length or sex have an impact on the within-sample diversity. Similarly, PERMANOVA analysis based on the beta diversity measures Bray-Curtis and Jaccard reveals no statistically significant differences in community structure at the order level between Lofoten and Sørøya (Table 3).

The species-level composition within Vibrionales

Overall, 55.3% of the reads are classified to the species level (Supporting Information Table S2). Of these, Photobacterium, Aliivibrio and Vibrio species are consistently found in all individuals, and constitute between 39% and 94% (mean: 77.3%) of all species-level reads (Fig. 3A). The Vibrionales community is dominated by Photobacterium illopuscarium (mean relative abundance: 40.3%) and Photobacterium kishitani (MRA: 26.6%) (Fig. 3B), while specific samples also have a high relative abundance of Aliivibrio logei (maximum relative abundance (MRA): 19.4%).
P. piscicola (MRA: 38.8%), Aliivibrio wodanis (MRA: 18.5%), Aliivibrio fischeri (MRA: 10.1%) and Aeromonas salmonicida (MRA: 10.0%). We detect no significant difference in the intestinal Vibrionales species community structure between Lofoten and Sørøya (Table 3). Metagenomic shotgun sequencing identifies a set of clearly separated, highly abundant Photobacterium, Aliivibrio and Vibrio species in the Atlantic cod intestines (Fig. 3B). We retrospectively assessed whether 16S rRNA-based taxonomic profiling is able to provide an equally detailed description of the bacterial community by analysing the 16S V4 sequences of these Vibrionales species (Supporting Information Table S5, Fig. S2, File S1). Several of the species share identical V4 sequences

---

**Table 1.** Metagenomic sequences before and after trimming, quality filtering and host DNA removal. The table shows per sample the number of original (raw) reads, the percentage of reads remaining after trimming and filtering, percentage of host DNA, percentage of bacterial DNA and the final number of reads used in the microbiome analysis. PhiX- and human DNA sequences represent a negligible proportion and are therefore excluded from the table. The bottom two rows show total and mean values per column. On average, 33.3% of the quality filtered reads per sample are used for microbiome analysis. For details, see Supporting Information Table S1.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads</th>
<th>After quality trimming/filtering (%)</th>
<th>Host DNA (%)</th>
<th>Bacterial DNA (%)</th>
<th>Final reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_01</td>
<td>10,883,740</td>
<td>85.9</td>
<td>87.3</td>
<td>12.7</td>
<td>1,187,649</td>
</tr>
<tr>
<td>L_02</td>
<td>11,140,950</td>
<td>87.9</td>
<td>62.2</td>
<td>37.8</td>
<td>3,699,538</td>
</tr>
<tr>
<td>L_03</td>
<td>9,891,322</td>
<td>90.2</td>
<td>41.2</td>
<td>58.8</td>
<td>5,249,515</td>
</tr>
<tr>
<td>L_04</td>
<td>10,587,865</td>
<td>86.9</td>
<td>85.2</td>
<td>14.8</td>
<td>1,364,663</td>
</tr>
<tr>
<td>L_05</td>
<td>8,423,091</td>
<td>89.1</td>
<td>57.7</td>
<td>42.3</td>
<td>3,171,737</td>
</tr>
<tr>
<td>L_06</td>
<td>10,879,319</td>
<td>89.6</td>
<td>30.5</td>
<td>69.5</td>
<td>6,772,948</td>
</tr>
<tr>
<td>L_07</td>
<td>10,082,237</td>
<td>91.8</td>
<td>31.3</td>
<td>68.7</td>
<td>6,361,506</td>
</tr>
<tr>
<td>L_08</td>
<td>9,114,703</td>
<td>87.3</td>
<td>80.5</td>
<td>19.5</td>
<td>1,549,210</td>
</tr>
<tr>
<td>L_09</td>
<td>11,105,189</td>
<td>89.1</td>
<td>62.2</td>
<td>37.8</td>
<td>3,733,846</td>
</tr>
<tr>
<td>L_10</td>
<td>11,140,743</td>
<td>84.7</td>
<td>86.0</td>
<td>14.0</td>
<td>1,320,875</td>
</tr>
<tr>
<td>S_01</td>
<td>10,631,475</td>
<td>84.2</td>
<td>91.5</td>
<td>8.5</td>
<td>834,564</td>
</tr>
<tr>
<td>S_02</td>
<td>11,527,589</td>
<td>92.6</td>
<td>18.2</td>
<td>81.8</td>
<td>7,018,741</td>
</tr>
<tr>
<td>S_03</td>
<td>9,855,514</td>
<td>84.2</td>
<td>83.7</td>
<td>16.3</td>
<td>1,353,671</td>
</tr>
<tr>
<td>S_04</td>
<td>9,259,707</td>
<td>92.6</td>
<td>79.3</td>
<td>20.7</td>
<td>1,959,505</td>
</tr>
<tr>
<td>S_05</td>
<td>11,539,193</td>
<td>82.2</td>
<td>85.7</td>
<td>14.3</td>
<td>1,247,744</td>
</tr>
<tr>
<td>S_06</td>
<td>10,272,359</td>
<td>86.5</td>
<td>77.7</td>
<td>22.3</td>
<td>2,779,436</td>
</tr>
<tr>
<td>S_07</td>
<td>14,395,326</td>
<td>87.4</td>
<td>69.7</td>
<td>30.3</td>
<td>2,431,383</td>
</tr>
<tr>
<td>S_08</td>
<td>9,186,209</td>
<td>88.6</td>
<td>50.4</td>
<td>49.6</td>
<td>3,727,749</td>
</tr>
<tr>
<td>S_09</td>
<td>8,476,896</td>
<td>88.6</td>
<td>66.7</td>
<td>33.3</td>
<td>56,885,711</td>
</tr>
<tr>
<td>Total</td>
<td>198,396,427</td>
<td>87.4</td>
<td>66.7</td>
<td>33.3</td>
<td>2,993,985</td>
</tr>
<tr>
<td>Mean</td>
<td>10,441,917</td>
<td>87.4</td>
<td>66.7</td>
<td>33.3</td>
<td>2,993,985</td>
</tr>
</tbody>
</table>

---

![Fig. 2. Taxonomic composition of the intestinal microbiome in Atlantic cod specimens from Lofoten and Sørøya. A. Relative abundance of metagenomic shotgun sequences classified to bacterial orders. Colours represent the 30 orders with highest relative abundance, including reads that could not be assigned to the order level (yellow). Numbers 1–10 and 1–9 represent individual specimens. Bars below the stacked bar plot show the square root transformed counts of paired-end reads classified to order level per individual. B. Relative abundance of 16S rRNA V4 sequences from Riiser et al., 2018 classified to bacterial orders. Numbers 1–12 and 1–10 represent individual specimens. The category 16S/Both indicates taxa identified by 16S analysis alone (normal font) and by 16S as well as metagenomic sequence analysis (bold). [Color figure can be viewed at wileyonlinelibrary.com]© 2019 The Authors. *Environmental Microbiology* published by Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*, 21, 2576–2594
Table 2. The 10 most abundant orders in the Atlantic cod intestinal microbiome. The table shows the 10 most highly abundant orders in the metagenomic shotgun sequencing analysis, their mean, minimum and maximum relative abundance. The corresponding values from the 16S rRNA-based analysis are reproduced from Riiser et al. (2018). NP: Not present.

<table>
<thead>
<tr>
<th>Order</th>
<th>Metagenomic (n = 19)</th>
<th>16S rRNA (n = 22)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (%)</td>
<td>Mean (%)</td>
</tr>
<tr>
<td></td>
<td>Minimum (%)</td>
<td>Minimum (%)</td>
</tr>
<tr>
<td></td>
<td>Maximum (%)</td>
<td>Maximum (%)</td>
</tr>
<tr>
<td>Vibionales</td>
<td>81.8</td>
<td>64.3</td>
</tr>
<tr>
<td></td>
<td>46.7</td>
<td>27.4</td>
</tr>
<tr>
<td></td>
<td>95.4</td>
<td>97.3</td>
</tr>
<tr>
<td>Alteromonadales</td>
<td>3.6</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>11.3</td>
<td>5.5</td>
</tr>
<tr>
<td>Fusobacteriales</td>
<td>3.1</td>
<td>17.1</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>25.4</td>
<td>39.9</td>
</tr>
<tr>
<td>Clostridiales</td>
<td>2.9</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>17.4</td>
<td>22.6</td>
</tr>
<tr>
<td>Bacteroidales</td>
<td>1.7</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>15.4</td>
<td>22.9</td>
</tr>
<tr>
<td>Enterobacteriales</td>
<td>1.2</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td>-</td>
</tr>
<tr>
<td>Oceanospirillales</td>
<td>0.9</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>14.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Bacillales</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
<td>-</td>
</tr>
<tr>
<td>Mycoplasmatales</td>
<td>0.2</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>1.1</td>
<td>12.1</td>
</tr>
<tr>
<td>Pseudomonadales</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 3. PERMANOVA analysis of diversity differences between bacterial communities from Lofoten and Sørøya (beta diversity). The table shows $R^2$ and p-values from multivariate statistical analyses to test for community composition differences based on reads classified at order and species level. The results are based on read counts normalized by common scaling. Degrees of freedom ($df$): 18.

<table>
<thead>
<tr>
<th></th>
<th>Bray-Curtis</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order-level classification</td>
<td>$R^2$</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.409</td>
</tr>
<tr>
<td>Species-level classification</td>
<td>$R^2$</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.617</td>
</tr>
</tbody>
</table>

Within-Vibionales levels of single nucleotide variant heterogeneity

We assessed the heterogeneity of the reads mapping to each of the 15 most abundant Vibionales bacterial reference genomes (Supporting Information Table S6). These 15 genomes all obtained sufficient coverage across the majority of samples to confidently identify SNVs with a greater than 10-fold coverage. Sequence similarity estimates based on the average nucleotide identity (ANI) and mash distance among these 15 genomes reveal a clear separation between the Aliivibrio-, Photobacterium- and Vibrio species (Supporting Information Fig. S3, Table S8). The Aliivibrio species are more similar to each other than the Photobacterium species, and Vibrio renipiscarium has a higher sequence divergence compared to the other genomes. The overall differences in sequence diversity among the species (Supporting Information Fig. S3) are reflected in the results of SNV analysis (Supporting Information Fig. S4), for example, species from the Aliivibrio cluster all have a lower SNV density than most Photobacterium species. Based on sequence similarity (%ANI) of these genomes, the results of six reference genomes that represent different species clusters are reported here (Fig. 4 and Supporting Information Fig. S3).

Overall, the different reference assemblies vary in the mean fold coverage, the density of variable sites within each individual sample and in the total number of SNVs observed in all samples. For instance, almost 5000 SNVs are detected in the Photobacterium angustum S14 genome, but the average density within specimens is low (max. 4.7/Kbp). In contrast, P. iliopiscarium yields less SNVs (1299) overall, yet a higher average density (max. 43.4/Kbp). The density of variable sites varies across specimens for several of the reference genomes, reflecting varying levels of heterogeneity in the bacterial populations within specimens. This pattern is particularly strong for P. iliopiscarium, varying from 0.1 to 43.4 variant positions per Kbp per individual specimen (Fig. 4). Likewise, the variation analysis of the two Aliivibrio genomes (A. salmonicida and A. sp. 1S128) indicate that sample L_03 consists of a complex mix of Aliivibrio strains. Despite the overall differences in SNV abundance between reference strains, we observe no statistically significant differences (based on Tracy-Widom and Chi-squared statistics) in SNV profiles between Lofoten and Sørøya among any of the 15 Vibionales strains (Supporting Information Fig. S5, Table S9).

Genome-wide discrepancies between abundant Photobacterium strains and their closest relatives

Per individual, 85% of the P. iliopiscarium genome and 45% of the P. kishitanii genome are sequenced to a depth of minimum 5-fold coverage, respectively (Supporting Information Table S10). Whereas reads aligned to P. iliopiscarium provide near complete coverage of the

© 2019 The Authors. Environmental Microbiology published by Society for Applied Microbiology and John Wiley & Sons Ltd., Environmental Microbiology, 21, 2576–2594
entire assembly in all individuals, reads aligned to *P. kishitanii* show consistent lack of alignments in a specific genomic region between 60 and 80 kbp (Fig. 5). This region in the Mediterranean *P. kishitanii* reference genome (Supporting Information Table S7) contains a prophage (Machado and Gram, 2017), and the deletion found here suggests that the North Atlantic population of this species lacks this particular prophage (30–50 kbp), as well as other host DNA. The difference in observed coverage between the two species translates directly to the number of genes lost; while only seven genes are absent in the Atlantic cod *P. iliopiscarium* strains, 698 genes are absent (with zero

© 2019 The Authors. *Environmental Microbiology* published by Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*, 21, 2576–2594
We obtained gene ontology data for 400 of the 698 genes that are absent from the P. kishitanii strain (Supporting Information Table S12 and Fig. S6). A striking number of sequences encodes membrane or membrane-associated cellular components (GO CC classification: membrane, membrane part, Supporting Information Fig. S6). Independent of functional annotation, a blast search indicates that the majority of the 698 gene sequence reads matches P. kishitanii (Supporting Information Fig. S6), confirming the coverage (in the Atlantic cod P. kishitanii strains compared to their reference assemblies (Fig. 5, Supporting Information Table S11)).

Fig. 4. Sequenced Vibrionales genome assemblies relative to reference genomes. For six Vibrionales reference genomes, the figure displays (from top to bottom) (top) read coverage per single nucleotide variant (SNV) position in each sample from Lofoten (red numbers) and Sørøya (blue numbers), (middle) variation density (number of variable positions per kbp. Reported in each individual sample, independent of coverage in the other samples) per sample and (lower) heatmap of a randomly chosen subset of 400 SNVs. In the heatmap, each row represents a unique variable nucleotide position, where the colour of each tile represents the two most frequent competing nucleotides in that position. The shade of each tile represents the square root-normalized ratio of the most frequent two bases at that position (i.e., the more variation in a nucleotide position, the less pale the tile is). The y-axis of the coverage- and variation density plots are scaled across the reference genomes. For each genome, the density plot (on top) is annotated with the maximum variation density value (grey number). [Color figure can be viewed at wileyonlinelibrary.com]
presence of this species (and not \textit{P. phosphoreum} ANT-2200, as the reference is classified) in the Atlantic cod intestines. In contrast to \textit{P. kishitanii}, only seven genes are absent in the Atlantic cod-associated \textit{P. iliopiscarium} strain compared to its closest relative. Only one of these is successfully annotated and is assigned a function in \textquote{chromosome partitioning}'.

\textbf{The \textit{P. kishitanii} \textit{lux operon}}

\textit{P. kishitanii} is known to contain the \textit{lux operon} (i.e., encoding luciferase activity) necessary for bioluminescence. Due to the high relative abundance of this bacterium (or a closely related strain) in the Atlantic cod gut, and the unclear role of such a bioluminescence feature in the intestinal compartment, we

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image}
\caption{Representation of the two most abundant \textit{Photobacterium} species among the individual samples. For each of the two most abundant \textit{Photobacterium} species in the Atlantic cod samples, \textit{Photobacterium iliopiscarium} (A) and \textit{P. kishitanii} (B), the figure gives an overview of the sequence coverage distribution at the assembly level (circle) and gene level (upper right square). In the assembly overview, each bar represents 20,000 bp of a contig. Starting from the center, the concentric rings display the GC %, $\log_{10}$ coverage of the 19 samples, and predicted ribosomal RNAs. The coverage scale is identical for all samples, and the maximum value is given in the extracted selection above the assembly overview. The assembly overview metadata shows information on the total number of reads mapped per sample, and physical parameters associated with each individual fish. In the gene overview, each bar represents an individual gene, and the genes are ordered by differential coverage across samples. Maximum $\log_{10}$ coverage is given below the figure. The metadata gives information of the total number of reads and the number and percentage of reads mapped. In contrast to the incomplete \textit{P. iliopiscarium} assembly (289 contigs) (A), the \textit{P. kishitanii} assembly (B) consists of only three scaffolds, annotated in the outer layer. For \textit{P. kishitanii}, the presence of the \textit{lux operon} is annotated with a black square. The (*) denotes sections of the reference genome completely missing in \textit{P. kishitanii} in the Atlantic cod intestines, while (**) denotes genes in the reference genome absent in the \textit{P. kishitanii} in the cod samples. [Color figure can be viewed at wileyonlinelibrary.com]}
\end{figure}
investigated the putative loss of the *lux* operon in the *P. kishitanii* strain associated with the Atlantic cod intestinal samples. All *lux* genes (*lux C, D, A, B, F, E, G*) of the operon are identified in the *P. kishitanii* strain in Atlantic cod (Supporting Information Fig. S7, Table S13). Their mean coverage across all samples ranges from 12.5 to 21X, and the coverage of each gene per sample correlates with the total number of mapped paired reads per sample (Supporting Information Table S13). No insertions or deletions are observed in the *lux* operon gene sequences (Supporting Information File S2). We find between 5 and 23 nonsynonymous substitutions in Atlantic cod *P. kishitanii* *lux* genes compared to the reference sequence. None of these substitutions results in a stop-codon, and there is no indication that the translation of the complete *lux* operon is disabled in the *P. kishitanii* strains in the Atlantic cod intestine.

**Discussion**

Here, we have used metagenomic shotgun sequencing to provide a first in-depth characterization of the Atlantic cod intestinal microbiome determined down to species- and strain-level resolutions. In contrast to previous 16S rRNA data, which yielded a single numerically dominant OTU belonging to genus *Photobacterium* (Riiser et al., 2018), we find at least nine bacterial *Photobacterium* species that occur in varying abundances in the Atlantic cod gut. Based on their 16S V4 sequences, eight of these species cluster into a single OTU (at 97% sequence identity), demonstrating the increased taxonomical resolution provided by metagenomic shotgun sequencing.

Two related species (*P. iliopiscarium* and *P. kishitanii*) are particularly abundant, comprising 67% of reads classified to genus *Photobacterium* and more than 50% of all reads classified. Both have previously been isolated from the intestines of Atlantic cod (e.g., Dhanasiri et al., 2011), although these species differ in their perceived ecological niches. *P. kishitanii* is a cosmopolitan, wide-spread facultative psychrophilic bacterium (Urbanczyk et al., 2011; Machado and Gram, 2017). It is most known for containing the *lux-rib* operon, which is essential for quorum sensing and generating bioluminescence in the light organs in—among others—*Gadiform* deep-water fish (Ast and Dunlap, 2005). In contrast, *P. iliopiscarium* is a nonluminous bacterium that has been isolated from the intestines of several cold-water species, including Atlantic cod, yet the ecological distribution of this bacterium is still poorly known (Onarheim and Raa, 1990; Onarheim et al., 1994; Urakawa et al., 1999; Ast and Dunlap, 2005; Smith et al., 2007). Based on phylogenetic analyses, *P. iliopiscarium* has lost the *lux-rib* operon, presumably due to niche specialization (Machado and Gram, 2017).

The high abundance of *P. kishitanii*—with full repertoire of *lux* genes—in the Atlantic cod intestinal samples is particularly interesting. Zooplankton feeding on luminescent bacteria has been found to glow, which makes them more vulnerable to predation (Zarubin et al., 2012). Bioluminescence has therefore been suggested to be an adaptation encouraging fish ingestion, allowing efficient dispersal of the bacteria through their fish hosts (Takemura et al., 2014). Although luminescent bacteria have been long known from excrement pellets (Andrews et al., 1984) and a wide range of fish taxa (Ruby and Morin, 1979), for instance captive Atlantic halibut (*Hippoglossus hippoglossus*) (Verner-Jeffreys et al., 2003), their relative abundances in wild fish intestines have never been reported. Here, we observe that such luminescent bacteria comprise an abundant component (26.6% of reads for *P. kishitanii*) of the intestinal microbiota in Atlantic cod. This observation suggests that fish intestines form a particularly rich niche for bioluminescent bacteria.

We compared the genomic organization of the two most abundant *Photobacterium* species in the Atlantic cod intestinal microbiome to their closest relatives by investigating genome-wide alignments. A near complete read coverage across the reference genome of *P. iliopiscarium* was observed, indicative of limited large-scale genomic rearrangements in those strains sampled from the Atlantic cod intestines. In contrast, a consistent lack of read coverage in distinct genomic regions across the *P. kishitanii* reference genome demonstrates the absence of specific regions in all Atlantic cod-associated strains. This lack results in the absence of nearly 700 of the 4300 genes annotated on the *P. kishitanii* reference assembly. Such an observation is not uncommon among *Photobacterium* species, and as little as 25% of genes is expected to be conserved between different strains of this genus (Machado and Gram, 2017). Nevertheless, the consistent absence of the same genomic region in all individuals indicates that these intestines have been colonized by a closely related *P. kishitanii* strain in both geographical locations. Interestingly, the missing genes predominantly encode components of the cell membrane. Given that the bacterial cell membrane plays a central role in host-microbiome interaction, and the fact that Atlantic cod has lost the MHC II pathway and possesses a special TLR repertoire (Star et al., 2011; Solbakken et al., 2016), it is possible that the loss of these genes represents a functional adaptation to the peculiar immune host environment.

The functional role of *P. iliopiscarium*, *P. kishitanii*, and other members of the genus *Photobacterium* in the Atlantic cod intestines and the reason for their high abundance (host-selection or environmental exposure) remains unclear. In a genome-based functional profiling analysis of metagenome-assembled genomes (MAGs) closely related to *P. iliopiscarium* and *P. kishitanii*—derived from...
the intestines of Northeast Atlantic cod—Le Doujet et al., 2019 finds a high relative abundance of the valine-glycine repeat protein G (VgrG) protein of the Type VI secretion system (T6SS) (Le Doujet et al., 2019). Several T6SSs are involved in antibacterial processes, including in the human gut (Russell et al., 2014). This suggests that *P. iliopiscarium* and *P. kishitanii* may provide antagonistic activity toward bacterial pathogens in the Atlantic cod gut, as has previously been observed for several members of the *Photobacterium* genus (MacDonald et al., 1986; Caipang et al., 2010; Ray et al., 2012; Egerton et al., 2018). Le Doujet et al., 2019 also observed a high relative abundance of genes involved in chitin and N-acetylglucosamine utilization, indicating that the *Photobacterium* species may aid in the breakdown of the exoskeleton of crustacean prey, of which chitin is the major component. Further, histamine biosynthesis by *P. kishitanii* associated with marine fish has been observed in several studies (Bjornsdottir-Butler et al., 2016; Machado and Gram, 2017), suggesting that this bacterium may be involved in immune responses in the Atlantic cod gut. Members of the *Photobacterium* genus have also been shown to aid in the digestive process of Dover sole (*Solea solea*), i.e., by degrading chitin (MacDonald et al., 1986), while others show antagonistic activity towards common bacterial pathogens in Atlantic cod (MacDonald et al., 1986; Caipang et al., 2010; Ray et al., 2012; Egerton et al., 2018). Such roles in protective immunity or digestion suggest an evolutionary benefit of host selection for the colonization by *Photobacterium*. Host selection for certain taxa (classified based on 16S rRNA) has been observed in zebrafish and Atlantic salmon parr (Roeselers et al., 2011; Dehler et al., 2017). It may be assumed that bacteria more intimately associated with their host (i.e. through a strong association with the mucosal layer relative to the general gut content) are actively selected for. Based on such an assumption, host selection for *Photobacterium* in Atlantic cod is implied by a significantly higher abundance of this genus associated with the intestinal mucosal layer relative to the gut content based on 16S RNA classification (Risser et al., 2018). Nonetheless, it is currently not clear if this higher abundance of the genus in the mucosal layer is due to the increased selection for specific *Photobacterium* strains, for example, *P. iliopiscarium* or *P. kishitanii*. Hence, more elaborate functional studies are required to investigate the roles of *P. iliopiscarium, P. kishitanii* and the other members of *Photobacterium* in the Atlantic cod intestines, and whether their high abundance are due to its unique immune system or by external, ecological factors (Star et al., 2011; Star and Jentoft, 2012).

Our results shed light on the order-level classification based on 16S rRNA amplicon sequencing versus metagenomic shotgun sequencing. There are significant differences in the order-level bacterial community composition detected by the two analysis methods. For instance, *Fusobacteriales* has an average relative abundance of 17.1% based on 16S rRNA, yet comprises 3.1% of the metagenomic shotgun data. Interestingly, a member of the *Fusobacteriales* (*Cetobacterium somerae*) that has been isolated from the intestinal tract of fish (Tsuchiya et al., 2008) has a particularly low GC content (28.5%) (ecogenicomic.org, 2013). A bias against such low GC content has been observed during library preparation (for instance due to the enzymatic fragmentation applied in our protocol), amplification and sequencing (Benjamini and Speed, 2012), and could explain lower *Fusobacteriales* relative abundance in the metagenomic data. This lower proportion of *Fusobacteriales* may contribute to the increased relative abundance of *Vibrionales* in the metagenomic shotgun data. Despite such differences, however, both methods do identify a similar set of abundant microbial taxa and show a dominant presence of *Vibrionales* in the intestines of Atlantic cod.

Several 16S rRNA-based studies have reported limited effects of geographic location on the composition and diversity of the fish intestinal microbiome. In Atlantic salmon (*Salmo salar*), little differentiation was observed in populations from both sides of the Atlantic Ocean, and the intestinal microbiome composition was rather associated with life stage (Llewellyn et al., 2016). Similarly, the gut microbiome of invasive Silver carp (*Hypophthalmichthys molitrix*) collected at highly separated sampling spots in the Mississippi river basin was affected by sampling time rather than location (Ye et al., 2014). Finally, no significant differences in intestinal microbiome composition were detected in Atlantic cod from Lofoten and Sørøya, separated by 470 km (same locations as in this study) using 16S rRNA analyses (Risser et al., 2018). Our in-depth characterization of the Atlantic cod intestinal microbiome using metagenomic shotgun sequencing allowed us to re-address if significant geographical population structure could be demonstrated at the species or within-species level based on genome-wide data. First, at the level of species, we observe no significant geographical differences using genome-wide protein-based analyses. This lack of differentiation is partly due to the presence of the two *Photobacterium* species (*P. iliopiscarium* and *P. kishitanii*), which are abundant in all specimens. Second, based on SNV variation across the genome of the 15 most abundant *Vibrionales* species, we find that the gut of each fish specimen contains a unique and diverse set of strains of each species, nonetheless, no significant geographical differences are observed. Both the protein-based and strain-level approaches assessing the diversity of *Vibrionales* indicate that the microbial community composition of the gut is not related to the geographic location where the cod specimens were caught. This absence of geographical substructure, even based on genome-wide data, suggests that the intestinal microbiome of Atlantic cod...
Wild coastal Atlantic cod (*G. morhua*) specimens were collected in Lofoten (N68.0619167, W13.5921667) (10 individuals, August 2014) and Sørøya (N70.760418, W21.782716) (9 individuals, September 2013) (Fig. 1A, Supporting Information Table S14). A 3 cm long part of the hindgut (immediately above the short, wider rectal chamber) was aseptically removed post-mortem by scalpel and stored on 70% ethanol. The samples were frozen (−20°C) for long-term storage. Relevant metadata such as length, weight, sex, and maturity were registered. Age was determined by studying otoliths. Although different individuals were used here, these were collected on the same time and location as the fish used in a previous 16S rRNA-based study (Riiser et al., 2018). We always strive to reduce the impact of our sampling needs on populations and individuals. Therefore, samples were obtained as a byproduct of conventional business practice. Specimens were caught by commercial vessels, euthanized by local fishermen and were intended for human consumption. Samples were taken post-mortem and no scientific experiments have been performed on live animals. This sampling follows the guidelines set by the ‘Norwegian consensus platform for replacement, reduction and refinement of animal experiments’ (Norecopa) and does not fall under any specific legislation in Norway, requiring no formal ethics approval.

**Sample preparation and DNA extraction**

Intestinal samples were split open lengthwise, before the combined gut content and mucosa was gently removed using a sterile disposable spatula. Each individual sample was washed in 500 μl 100% ETOH and centrifuged before the ethanol was allowed to evaporate, after which dry weight was measured before proceeding to DNA extraction. DNA was extracted from between <10 and 300 mg dry weight of gut content using the MoBio Powersoil HTP 96 Soil DNA Isolation Kit (Qiagen, Valencia, CA) according to the DNA extraction protocol (v. 4.13) utilized by the Earth Microbiome Project (Gilbert et al., 2010). DNA was eluted in 100 μl Elution buffer and stored at −20°C. Due to high methodological consistency between biological replicates in previous experiments, only one sample was collected per fish (Riiser et al., 2018).

**Sequence data generation and filtering**

Quality and quantity of the DNA was measured using a Qubit fluorometer (Life Technologies, Carlsbad, CA), and normalized by dilution. DNA libraries were prepared using the Kapa HyperPlus kit (Roche Sequencing, Pleasanton, CA) and paired-end sequenced (2 × 125 base pairs) on an Illumina HiSeq2500 using the HiSeq SBS V4 chemistry with dual-indexing in two independent sequencing runs. Read qualities were assessed using FastQC (Andrews, 2010), before adapter removal, singleton read identification, de-duplication and further read quality trimming were performed using Trimomatic (ver. 0.36) (Bolger et al., 2014) and PRINSEQ-lite (ver. 0.20.4) (Schmieder and Edwards, 2011) (Supporting Information Table S15). PhiX, host and human sequences were removed by mapping reads to the phiX reference genome [GenBank: J02482.1], the Atlantic cod genome assembly (gadMor 2), [P. iliopiscarium](https://www.ncbi.nlm.nih.gov/nuccore/J02482.1] and a masked version of the human genome (HG19) (Genome Reference Consortium, 2009) using BWA (ver. 0.7.13) (Li and Durbin, 2009) or BBMap (ver. 37.53) (JGI) with default parameters and discarding matching sequences using seqtk (ver. 2012.11) (Li, 2012). All sequence data have been deposited in the EMBL database European Nucleotide Archive (ENA) under study accession number PRJEB29346.

**Taxonomic profiling**

Taxonomic classification of quality-trimmed and filtered metagenomic paired-end reads was performed using Kaiju (ver. 1.5.0) (Menzel and Krogh, 2016) (‘greedy’ heuristic approach, −e 5), with the NCBI nr database (rel. 84) (incl. Proteins from fungal and microbial eukaryotes) as reference (O’Leary et al., 2016). Counts of reads successfully assigned to orders and species were imported into RStudio (ver. 1.1.383) (Racine, 2010) based on R (ver. 3.4.2) (R Core Team, 2017) for further processing. Final results were visualized using the R package ggplot (ver. 2.2.1) (Wickham, 2009). Note: Based on a recent recategorization (Machado and Gram, 2017), we refer to
the reference strain *Photobacterium phosphoreum* ANT-2200 (acc. nr. GCF_000613045.2) as *P. kishitanii* (Supporting Information Table S7).

**Assessment of Vibrionales species resolution based on 16S rRNA V4 region**

RNA sequences of the most highly abundant Vibrionales species were downloaded from RefSeq (accessed 12.12.18) (Supporting Information Table S5), before 16S rRNA sequences were extracted using a custom script. Next, the 16S rRNA sequences were imported into Geneious (ver. 10.2.2) (Geneious), where the V4 regions (one or multiple from the same assembly) were identified and extracted. Finally, the V4 regions of the different Vibrionales species were aligned (Supporting Information File S1) using the MAFFT algorithm with default parameters, generating a sequence similarity matrix (Supporting Information Table S6).

**Sequence variation analysis and genome similarity estimations**

In order to assess the heterogeneity of Vibrionales species in our bacterial populations, we analysed the sequence variation in Vibrionales genomes present in the intestinal metagenome of each fish. Initially, paired-end reads from each sample were mapped against 109 complete or scaffold-level *Photobacterium, Aliivibrio* or *Vibrio* genomes downloaded from NCBI RefSeq (rel. 84) (O’Leary et al., 2016) (Supporting Information Table S16). The relatedness between the 15 reference genomes recruiting the highest portion of reads (Supporting Information Table S7) was then estimated based on Average Nucleotide Identity (ANI) and Mash genome distances using FastANI (ver. 1.1) (Jain et al., 2018) and Mash (ver. 2.1) (On dov et al., 2016) (Supporting Information Fig. S3, Table S8). For the sequence variation analysis, paired-end reads from each individual were mapped to the 15 reference genomes using the Snakemake workflow (Köster and Rahmann, 2012) of anvi’o (ver. 5.1) (Eren et al., 2015a) with default parameters in the ‘all-against-all’ mode (with anvi-profile --min-coverage-for-variability 10). In anvi’o, contigs are divided into ‘splits’ of maximum 20,000 bp. Splits with outlier mean coverage values (above the 98-percentile, 4–7 splits per sample), potentially containing repetitive sequences, were removed, and samples of low coverage were filtered (0–2 samples per reference genome). For each individual sample, variable sites (with min. 10X coverage) were identified, and the mean number of these per 1000 bp calculated (variation density). Next, variable sites with a minimum of 10X coverage in all samples were defined as single nucleotide variants (SNVs, anvi-gen-variability-profile --min-occurrence 1 --min-coverage-in-each-sample 10). Coverage, variation density and SNV profiles were plotted in RStudio following the R script provided by anvi’o (Eren et al., 2015b). The anvi’o SNV output was converted to .vcf format using a custom-developed script (https://github.com/srinidhi202/AnvioSNV_to_vcf), and the resulting .vcf files were used for principal component analysis (PCA) to test for geographical differences as implemented in smartpca (ver. 6.1.4) (EIGENSOFT) (Patterson et al., 2006). The variant analysis results of six reference genomes that represent different species clusters (based on average nucleotide identity) are reported in the results section.

**Statistical analysis**

Differences in order-level classification between metagenomic shotgun sequencing and 16S rRNA amplicon sequencing (Fig. 2) were tested using ANOVA for compositional data (van den Boogaart and Tolosana-Delgado, 2013, section 5.3.3.2) using the R package compositions (ver. 1.40–2) (van den Boogaart and Tolosana-Delgado, 2008). Six orders common to both approaches (Fig. 2 legend, bold) and an ‘others’ category (which contained the remaining orders) were used for the ANOVA test. Model assumptions were verified as described in section 5.3.8 of van den Boogaart and Tolosana-Delgado, 2013. Within-sample diversity (alpha diversity) was calculated using the diversity function in the R package vegan (ver. 2.4–1) (Oksanen et al., 2017) based on Shannon, Simpson and Inverse Simpson indices calculated from non-normalized order-level read counts. Differences in alpha diversity were studied using linear regression. The optimal model (i.e., the model that best describes the individual diversity) was identified through a ‘top-down’ strategy including all covariates (Supporting Information Table S4), except age and weight, which highly correlated with length (r = 0.78 and 0.94), and selected through t-tests. Model assumptions were verified through plotting of residuals. Differences in bacterial community structure (beta diversity) between Lofoten and Sørøya were visualized using nonmetric multidimensional scaling (NMDS) plots based on the Bray–Curtis dissimilarity index and tested using permutation-based multivariate analysis of variance (PERMANOVA) using the metaMDS and adonis functions in vegan (ver. 2.4–1) with both Bray–Curtis dissimilarity and Jaccard index. *Adonis* was run with 20,000 permutations. PERMANOVA assumes the multivariate dispersion in the compared groups to be homogeneous; this was verified (p > 0.05) using the betadisper function (vegan) (Supporting Information Table S17). All beta diversity analyses were based on sequence counts normalized using a common scaling procedure, following McMurdie and Holmes, 2014 (McMurdie and Holmes, 2014). This
method multiplies the sequence count of every unit (e.g., species) in a given library with a factor corresponding to the ratio of the smallest library size in the data set to the library size of the sample in question, replacing rarefying (i.e., random subsampling to the lowest number of reads). Normalizing using this procedure effectively results in the library scaling by averaging an infinite number of repeated subsamplings. PERMANOVA analysis was performed on normalized counts of reads classified at the order and species level (Kajiu). We used Tracy-Widom and Chi-squared statistics, as implemented in smartpca (Patterson et al., 2006), to test for significant geographical differences in the distribution of SNVs per Vibrionales reference genome, while correcting for multiple testing using sequential Bonferroni (Holm, 1979).

**Genome-wide characterization of Photobacterium**

Genome-wide coverage of the two most abundant bacterial strains (P. kishitanii, ANT-2200 and P. iliopiscarium, ATCC 51761) was obtained by mapping all paired-end reads from each cod specimen towards the respective reference genomes (GCF_000613045.2, GCF_000949935.1), and visualized using the anvi’o command ‘anvi-interactive’. We selected the ANT-2200 strain as genomic reference for its improved assembly characteristics (200X coverage and a chromosome resolved assembly) compared to the P. kishitanii reference strain (80X coverage, 117 contigs). Next, ‘anvi-export-gene-coverage-and-detection’ (Eren et al., 2015a) together with the predicted gene loci identified in the automated NCBI prokaryotic genome annotation process (PGAP) were used to detect genes with zero coverage in all specimens and that are therefore consistently absent in these bacterial strains. To obtain updated GO information, we manually annotated the sequences of genes from those missing regions through a blastx search (using blast + (ver. 2.6.0) (Altschul et al., 1990; Camacho et al., 2009)) to the nr database (accessed 10. 12. 18) using default parameters, keeping the top 5 hits. The .xml results file and gene sequences was imported into Blast2GO (ver. 5.2.5) (Conesa et al., 2005; Conesa and Götz, 2008; Götz et al., 2008, 2011), where an InterPro search (Jones et al., 2014; Mitchell et al., 2019), GO mapping, functional annotation and visualization was conducted with default parameters. PHASTER (Zhou et al., 2011; Amdt et al., 2016) was used to screen the P. kishitanii genome for the presence of prophages. The regions around the identified prophage sequences were manually inspected for the presence of other phage-associated genes (e.g. capsid heads, terminases, integrases) that could have been missed by the PHASTER algorithm.

The lux operon was not annotated in the original P. kishitanii RefSeq assembly (GCF_000613045.2). Therefore, the P. kishitanii lux genes (Supporting Information Table S13) were identified using the lux sequences of Photobacterium phosphoreum (AB367391.1) in a local blast search against the P. kishitanii reference genome with blast+ (ver. 2.6.0) (Altschul et al., 1990; Camacho et al., 2009), and manually annotated. Paired-end reads from each sample were then mapped against the annotated reference genome, and reads (.bam files) mapping to the lux genes were combined per location using samtools (ver. 1.3.1) (Li et al., 2009) (‘samtools merge’) to yield a consensus sequence for each location per lux gene. The coverage distribution and possible loss of function (due to insertions, deletions, stop codons, etc.) of these lux gene consensus sequences was inspected using Geneious (ver. 10.2.2) (Geneious), Integrative Genomics Viewer (ver. 2.4.16) (Robinson et al., 2011; Thorvaldsdóttir et al., 2013) and the ExPASy Translate online tool (Artimo et al., 2012).

**Acknowledgements**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. This work was funded by a grant from the Research Council of Norway (project no. 222378) and University of Oslo (Strategic Research Initiative)–both to KSJ. We thank Borge Iversen and Helle Tessand Baalsrud for their kind help in sampling cod specimens in Lofoten, and Martin Malmstrøm, Paul Ragnar Berg and Monica Hongro Solbakken for sampling in Sørøya.

**References**


© 2019 The Authors. Environmental Microbiology published by Society for Applied Microbiology and John Wiley & Sons Ltd., Environmental Microbiology, 21, 2576–2594
The Atlantic cod intestinal microbiome


Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

Figure S1. Within-sample diversity of Atlantic cod. Boxplots of Shannon (A), Simpson (B) and Inverse Simpson (C) diversity in samples from Lofoten and Sørøya. The samples are grouped by location, and each of the 19 individuals is represented by a point. The middle band represents the median, while the upper and lower band shows the 75th and 25th percentile. The boxplots also show the minimum and maximum alpha diversity values.

Figure S2. Multiple alignment of Vibrionales 16S V4 sequences. The figure shows a multiple alignment of the 16S V4 region of the most highly abundant Photobacterium, Vibrio and Alterivibrio species (Supporting Information Table S5). The full alignment is supplied as a fasta file in Supporting Information File S1, and the corresponding similarity matrix is shown in Supporting Information Table S6.

Figure S3. Similarity of 15 Vibrionales reference genomes. Heatmap showing the similarity of the reference genomes of the 15 most abundant Vibrionales species present in the Atlantic cod samples, based on (A) % Average Nucleotide Identity and (B) MASH distance. Squares with a black border in panel (A) represent the selection of reference genomes presented in Fig. .

Figure S4. Variation analysis of 15 Vibrionales reference genomes. For each of the 15 most abundant Vibrionales genomes, the figure displays (from top to bottom), (1) read coverage per single nucleotide variant (SNV) position in each sample, (2) variation density (number of variable positions per Kbp. reported in each individual sample, independent of coverage in the other samples) per sample and (3) heatmap of a randomly chosen subset of 400 SNVs. In the heatmap, each row represents a unique variable nucleotide position, where the colour of each tile represents the two most frequent competing nucleotides in that position. The shade of each tile represents the square root-normalized ratio of the most frequent two bases at that position (i.e., the more variation in a nucleotide position, the less pale the tile is). The y-axis of the coverage- and variation density plots are scaled across the reference genomes, and the mean SNV coverage across all samples is noted to the right of each coverage plot. For each genome, the density plot is annotated with the maximum variation density value. The total number of SNVs identified per reference genome is noted in parentheses after the strain name. Samples from Lofoten are numbered in red, while samples from Sørøya are numbered in blue.

Figure S5. Principal component analysis (PCA) of SNVs in Lofoten and Sørøya. Principal component analysis of single nucleotide variant (SNV) distribution in individuals from Lofoten (red) and Sørøya (blue). Each plot represents one of the 15 Vibrionales reference genomes with highest mean abundance, and the ordering is similar to the 15 SNV plots in Fig. S4. Overlapping clusters indicate no spatial separation of the Atlantic cod intestinal microbiome. The p-value from a Chi-squared
test of geographical differences is included in each plot. Detailed statistics for each plot are given in Supporting Information Table S9.

**Figure S6.** Functional analysis of the 698 *P. kishitanii* genes missing in our closely related *Photobacterium* strain. (A) Overview of the complexes or compartments where the gene products of the missing genes are potentially active. The x-axis represents the number of gene sequences assigned to each GO (Gene ontology) category. (B) Overview of the taxonomy affiliated with hits from a blast search with the 698 gene sequences. The top five hits were kept for each individual sequence search. The x-axis represents the number of hits associated with each taxonomical category.

**Figure S7.** Coverage of the lux operon in *Photobacterium kishitanii*. The figure shows the coverage of the *Photobacterium kishitanii* lux operon, as displayed in the Integrative Genomics Viewer. The coverage of individual samples has been summed for both Lofoten (max. Coverage: 401X) and Sørøya (max. Coverage 228X). The horizontal red bars represent the individual lux genes in the operon (luxC, luxD, luxA, luxB, luxF, luxE and luxG).

**File S1.** Multiple alignment of *Vibrio* 16S V4 sequences in .fasta format. The file contains a .fasta-formatted multiple alignment of the 16S V4 region of the most highly abundant *Photobacterium*, *Vibrio* and *Aliivibrio* species (Supporting Information Table S5).

**File S2.** Lux gene sequences of *Photobacterium kishitanii* in .fasta format. The file contains, for each gene in the lux operon, the reference genome sequence and the consensus sequence from both Lofoten and Sørøya.

**Table S1.** Sample sizes
Number of reads per sample before, during and after the trimming and filtering steps. The final two columns show the number of classified paired-end reads and its percentage of all paired-end reads. The lower table shows a summary of the data per location.

**Table S2.** Classification of sequences by Kaiju (ver. 1.5.0)
The table shows the numbers of total classified reads and number of reads classified at the species and order level per sample. A detailed overview of reads per order-level taxon per sample starts at column P. The lower table shows the first part of the same data, but at a relative scale. Sums and mean values are given in the green and yellow rows.

**Table S3.** Alpha diversity values
Alpha diversity estimates of the Atlantic cod intestinal microbial samples, calculated from non-normalized counts of reads classified at order level. See also Fig. S1.

**Table S4.** Alpha diversity differences - Linear regression model
Results from linear regression analysis used in testing for significant effects of location, sex or length on alpha diversity. The beyond optimal model including all covariates is presented here. The 'top-down' strategy, selecting suitable covariates through t-tests, results in an ‘optimal’ model with no covariates, indicating that neither location, sex or length have a significant effect on alpha diversity.

**Table S5.** Accessions used for 16S V4 sequence analysis
The table shows all genomes (assemblies) used for the retrieval of 16S sequences ("*.rna_from_genomic.fna.gz") used in the multiple alignment of *Vibionales* V4 sequences.

**Table S6.** Similarity (% identity) of V4 sequences
The matrix shows the % identity between the 16S V4 region from the different *Vibionales* species. Green cells indicate an identity of 100%, yellow an identity > 97%.

**Table S7.** *Vibionales* genomes used for ANI- and variation analysis
Overview of the 15 *Vibionales* reference genomes used for genome similarity analysis measured by Average Nucleotide Identity (ANI) and variation analysis for the identification of single nucleotide variants (SNVs) within each genome. These genomes had the highest mean abundance among our samples after reference mapping.

**Table S8.** Reference genomes similarity
Results from genome similarity analyses based on average nucleotide identity (ANI) and mash distance. From the top: Table A) Jspecies website - ANIb, Table B) Jspecies website - Tetra, Table C) mash and Table D) fastANI. Plots in Fig. S3 are based on data from fastANI and mash.

**Table S9.** Significance tests of SNV distributions.
For each of the 15 *Vibionales* reference genomes, the tables show statistics for PCA of SNV distribution (Fig. S4, Fig. S5), including significance of two first PC axes (Tracy-Widom Statistic) and significance of between-group testing (Chi-square test). No Tracy-Widom p-values are significant after sequential Bonferroni correction (right table).

**Table S10.** Coverage breadth of *P. iliopiscarium* and *P. kishitanii*
The tables show per sample the portion of each reference genome that is covered at a sequencing depth of at least 5X or 10X.

**Table S11.** Genome- and gene level features of *P. iliopiscarium* and *P. kishitanii*
For each of the two most highly abundant *Vibionales* species in the Atlantic cod gut, the table shows information on assembly, total number of genes, total number of annotated genes and the number of genes with zero coverage in the Atlantic cod samples.

**Table S12.** Functional annotation of the 698 missing *Photobacterium kishitanii* genes
The table shows functional annotation data for each of the 698 genes absent in the *Photobacterium kishitanii* strain associated with Atlantic cod compared to its most closely related reference genome (GCF_000613045.2).

**Table S13.** *Photobacterium kishitanii* lux operon
The table shows accession number, length, position, and mean coverage per sample for each of the genes in the *Photobacterium kishitanii* lux operon. Gene sequence are given in the rightmost column.

**Table S14.** Metadata
Metadata collected for all specimens used in the study. Red and blue bars are applied to visualize associations between weight, length and age.
Table S15. Filtering parameters
Parameters used for quality filtering and trimming of metagenomic shotgun sequences in Trimmomatic (ver. 0.36) and PRINSEQ-lite (ver. 0.20.4).

Table S16. Vibrionales reference genomes
The table lists all reference genomes used for mapping of paired-end reads from the 19 intestinal microbiome samples.

Table S17. Homogeneity tests and PERMANOVA results for three datasets
Results from homogeneity and PERMANOVA tests on the datasets based on order- and species-level read counts. All tests were performed on normalized data using the beta diversity measures Bray–Curtis dissimilarity and Jaccard index. P-values are marked in bold; values <0.05 indicate statistical significance.