Metatranscriptomics reveals a shift in microbial community composition and function during summer months in a coastal marine environment

Ben J. G. Sutherland1,2,3*, Jan F. Finke2,4*, Robert Saunders5, Snehal Warne5, Angela D. Schulze1, Jeff H. T. Strohm1, Amy M. Chan2, Curtis A. Suttle2,6,7,8 and Kristina M. Miller1

1 Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, BC, Canada V9T 6N7
2 Department of Earth, Ocean and Atmospheric Sciences, The University of British Columbia, Vancouver, BC, Canada V6T 1Z4
3 Sutherland Bioinformatics, Lantzville, BC, Canada V0R 2H0
4 Hakai Institute, PO Box 309, Heriot Bay, BC, Canada, V0P 1H0
5 RKS Labs, Parksville, BC, Canada V9P 2A2
6 Department of Microbiology and Immunology, The University of British Columbia, Vancouver, BC, Canada V6T 1Z3
7 Department of Botany, The University of British Columbia, Vancouver, BC, Canada V6T 1Z4
8 Institute for the Oceans and Fisheries, The University of British Columbia, Vancouver, BC, Canada, V6T1Z4

* these authors contributed equally to this work.

*Author for correspondence: BJGS
Current Address:
Sutherland Bioinformatics, Lantzville BC, Canada, V0R 2H0
Sutherland.Bioinformatics@protonmail.com

*Author for correspondence: JFF
Department of Earth, Ocean and Atmospheric Sciences, The University of British Columbia, Vancouver, BC, Canada, V6T 1Z4
Email: jfinke@eoas.ubc.ca

Data Deposition: Raw metatranscriptome sequence data has been uploaded to SRA under BioProject PRJNA738385, BioSample SAMN19728638-SAMN19728657.
Abstract

Temperate coastal marine waters are often thermally stratified from spring through fall but can be dynamic and disrupted by tidal currents and wind-driven upwelling. These mixing events introduce deeper, cooler water with a higher partial pressure of CO$_2$ (pCO$_2$), and its associated microbial communities to the surface. Anecdotally, these events impact shellfish hatcheries and farms, warranting improved understanding of changes in composition and activity of marine microbial communities in relation to environmental processes. To characterize both compositional and functional changes associated with abiotic factors, here we generate a reference metatranscriptome from the Strait of Georgia over representative seasons and analyze metatranscriptomic profiles of the microorganisms present within intake water containing different pCO$_2$ levels at a shellfish hatchery in British Columbia from June through October. Abiotic factors studied include pH, temperature, alkalinity, aragonite, calcite and pCO$_2$. Community composition changes were observed to occur at broad taxonomic levels, and most notably to vary with temperature and pCO$_2$. Functional gene expression profiles indicated a strong difference between early (June-July) and late summer (August-October) associated with viral activity. The taxonomic data suggests this could be due to the termination of cyanobacteria and phytoplankton blooms by viral lysis in the late season. Functional analysis indicated fewer differentially expressed transcripts associated with abiotic variables (e.g., pCO$_2$) than with the temporal effect. Microbial composition and activity in these waters varies with both short-term effects observed alongside abiotic variation as well as long-term effects observed across seasons. The analysis of both taxonomy and functional gene expression simultaneously in the same samples by environmental RNA (eRNA metatranscriptomics) provided a more comprehensive view for monitoring water bodies than either would in isolation.

Keywords: metatranscriptome assembly; metatranscriptomics; microbial community; oyster; pCO2; seasonality
Introduction

Metatranscriptomics characterizes expressed genes (i.e., RNA transcripts) that are present in an environmental sample. These transcripts may be within the cell, or in the extracellular space. While metagenomics profiles the taxonomy of a sample, metatranscriptomics can profile the biological functions that are present in the sample and are active or differentially accumulated in particular environments (Aguiar-Pulido et al. 2016). Metatranscriptomics can also be used to deduce taxonomic information for dominant taxa in communities (Shi et al. 2011; Neves et al. 2017; Marcelino et al. 2019), in particular with longer transcriptome contigs, which are expected to produce correct taxonomic assignments (Shakya et al. 2019). Metatranscriptomics is typically conducted through next-generation sequencing, which has the benefit of identifying novel genes and functions not known to be used by the identified taxa (Gilbert et al. 2008). Millions of marine microorganisms and viruses occur within a millilitre of seawater and their abundance and community composition varies with regions and environmental variables (Gilbert et al. 2011; Wigington et al. 2016; Finke et al. 2017; Yu et al. 2018; Gregory et al. 2019). A large scale global study found indications that the functional gene content in marine microbial samples is largely shaped by taxonomic composition (Salazar et al. 2019). However, the detailed distribution of generated community functions and their link to environmental variables remains to be studied (Moran 2015). Altogether, the environmental metatranscriptomic approach has strong potential for profiling active processes and community composition under changing conditions in marine molecular ecology (Moran et al. 2013; Cristescu 2019).

Coastal water microbial communities are shaped by widely varying environmental conditions, including temperature, salinity, and especially pCO$_2$ and pH. (Salisbury et al. 2008; Joint et al. 2011; Ray et al. 2012; Lv et al. 2016; Lee et al. 2017). These are factors that are impacted by seasons, tides, or even biological activity. Microbes tolerate pH fluctuations under regular conditions and co-vary with water bodies in the short-term (Joint et al. 2011), but generally adjust in the long term. Microbial, bacterial and protist communities show significant responses to high pCO$_2$ concentrations (Ray et al. 2012; Zhang et al. 2015; Thomson et al. 2016). In the past, average pCO$_2$ concentrations have increased from 280 parts per million (ppm) (Friedlingstein et al. 2019) to currently 400 ppm (Blunden and Arndt 2020) and are deemed to increase to an average of 1000 ppm by the end of this century, with an associated rise of sea surface temperature by 1ºC and drop of pH by 0.29 (Kirtman 2013; Pachauri 2014; Bindoff 2019).

Many microorganisms are expected to vary with temperature and/or pCO$_2$, and this may include shellfish pathogens such as *Vibrio* (Gomez-Leon et al. 2005; Green et al., 2019), *Roseovarius* (Maloy et al. 2007), and *Mikrocytos* (Carnegie et al. 2003), as well as harmful algae (Landsberg 2002) and viruses (Renault et al. 2010). By the nature of their intertidal habitat and filter feeding lifestyle, in particular oysters are affected by varying environmental conditions and microbiomes (Wegner et al. 2013; Lokmer et al. 2016;
As a consequence, changes in microbial communities are expected to affect oyster hatcheries. Interestingly, for detecting an oyster parasite, environmental RNA (eRNA) was found to be superior to detect the species than when using environmental DNA (eDNA) (Merou et al. 2019). The impact of temperature increases and pCO$_2$ on overall microbial communities and microbial activity is difficult to predict, but is expected to be significant with wide-ranging consequences. This serves as a motivation to develop comprehensive ways to improve our understanding of changes in marine microbial community composition and activity in response to environmental processes.

Here we compare the microbial composition, active genes, and enriched functional categories of differentially expressed genes within samples of intake water at a shellfish hatchery over a range of pCO$_2$ levels, temperatures, and months. We apply comparative metatranscriptomics by first developing a de novo metatranscriptome assembly using the samples in the study. We take a taxonomic approach to view the community composition of each sample. In parallel, we quantify the relative expression levels of each transcript among the samples and perform differential gene expression analysis in relation to environmental metadata. Collectively, this work profiles hatchery intake water across 20 dates over a two-year period. Using environmental metadata, multivariate clustering, differential expression analysis, and functional enrichment analysis, we characterize these different sampling events in relation to date of collection and environmental parameters.

Materials and Methods

Water collection, RNA extraction and environmental variables

Water samples were collected from the ocean water intake at a Pacific oyster hatchery in Qualicum Bay, central East Coast Vancouver Island, British Columbia over a two year period, with the majority of samples collected in 2015 (Supplemental File S1). On each collection day, six litres of water were taken and filtered through sterile 0.22 µm cellulose acetate filters (Millipore, Burlington, MA). Filters were stored at -80ºC. Temperature was measured by a mercury thermometer, salinity with a refractometer and pH with a glass probe pH meter (Jenco, San Diego, CA). Alkalinity was measured with a HI901 titrator (Hanna Instruments, Smithfield, RI). pCO$_2$ was determined with a LI840A infrared gas analyzer (Li-Cor, Lincoln, NB). Aragonite and calcite were calculated with the CO2SYS.BAS program (https://github.com/jamesorr/CO2SYS-Excel/blob/master/CO2sys_mod.bas). The filters were then used for RNA extraction by the Power Water RNA isolation kit (MoBio, Carlsbad, CA) following manufacturer’s instructions, including the alternate lysis step. Output total RNA was depleted for ribosomal RNA and prepped for RNA-seq using the Scriptseq Complete Gold (Epidemiology) kit (Illumina, San Diego, CA). A total of ~50 ng of ribosomal depleted RNA was used as an input to transcriptome libraries. Individual barcoded libraries were randomly multiplexed into groups of four samples using equimolar
quantities and sequenced on a MiSeq using a v3 600 reagent kit (Illumina, San Diego, CA) to generate paired-end 250 bp reads. Environmental variables were analyzed in a principal component analysis (PCA) using the Vegan package (Oksanen et al. 2016) in the R environment (R Core Team, 2022), where missing values were imputed with unit averages.

Bioinformatics and metatranscriptome assembly

Raw and quality trimmed sequence data were inspected with FastQC (Andrews 2010) and MultiQC (Ewels et al. 2016). Quality trimming was conducted to remove low quality reads and adapters with Cutadapt (Martin 2011) using flags -q 20 to remove < Q20 data from the 3'-end of the read and -m 50 to remove reads shorter than 50 bp. Results were output as an interleaved fastq. Putative ribosomal RNA (rRNA) reads were removed using SortMeRNA (Kopylova, Noé, and Touzet 2012) using all suggested rRNA databases, and therefore enriching for putative messenger RNA (mRNA) in silico.

Two different approaches were taken to assemble the reference metatranscriptome. First, a reference metatranscriptome was assembled using the metatranscriptome-optimized assembler IDBA-tran (Peng 2012) with a majority of the samples (16 / 20 samples; 191,303,324 reads), not including all samples due to computational constraints. Samples used for this approach comprised an equal number of samples from both ends of the pCO₂ range present in the collections. This assembly was conducted with default settings and run on 56 threads. Second, a reference metatranscriptome was generated by first assembling each of the 20 libraries individually using IDBA-tran, and then merging these 20 assemblies into a single assembly using CD-HIT-EST (Li and Godzik 2006). CD-HIT-EST merged contigs at 95% similarity, and dedupe.sh of BBTools (Bushnell, Rood, and Singer 2017) was used to de-duplicate with default parameters (i.e., the ‘merged assembly approach’). These assemblies were compared based on the total number of contigs, total length, multi-mapping proportions, and mapping percentages to select the best assembly.

Reads for each sample were aligned against the reference metatranscriptome using Bowtie2 (Langmead and Salzberg 2012) in end-to-end mode allowing for multi-mappings. A maximum of 40 alignments were retained for each read. Alignments were then filtered to remove low quality mappings (i.e., retain mapq ≥ 2). Retained alignments were quantified using eXpress (Roberts et al. 2013). Effective counts from eXpress were output into a table in R, and imported into edgeR (Robinson and Oshlack 2010). Filtering was conducted to only retain contigs against which at least five reads mapped in the sample with the fewest reads (i.e., 3.86 counts per million; CPM), and requiring that the contig was represented at this CPM level or higher in at least five samples. Retained transcripts were normalized for library size using the TMM normalization method of edgeR v.3.28.1.
Taxonomic community analysis
The expressed transcripts of the metatranscriptome were annotated for taxonomic identity using BLASTn (Camacho et al. 2009) against the nt database of NCBI, retaining a maximum of 100 alignments and descriptions per record. Best annotations were selected based on the e-value and a minimum cut-off at $E < 10^{-5}$. Best match phylogenetic lineages of annotated transcripts were extracted with a custom python tool (see Data Accessibility) based on the subject sequence id using the ranked lineage database (NCBI), and exporting different levels of the taxonomy. The taxonomic overview and characterization used all expressed genes, at the kingdom, phylum and order level, and further analysis was conducted at the genus level. A canonical correlation analysis (CCA) of genus abundances and environmental variables, as well as ANOVA tests for significance of regressions were performed with the vegan package in R. To compare community composition at different lineage levels, pairwise distances were calculated using the Bray-Curtis dissimilarity measure in vegan, and Mantel tests were performed using the ade4 package (Dray and Dufour 2007). Linear models of genus abundance versus pCO$_2$ concentration and t-tests for genus abundance between seasons were conducted in R. Significant regressions were defined by $p < 0.05$, adjusted p-values were determined using the Benjamin-Hochberg procedure.

Differential expression analysis
To view detailed gene expression trends among samples, log2 CPM values were used for multidimensional scaling (MDS) plots. The plotMDS function of limma v.3.42.2 (Ritchie et al. 2015) was used to generate MDS plots using both the leading log2-fold-change as well as a PCA (gene.selection = common). Samples were grouped into binary groups for pCO$_2$ (low/normal versus high) and season for differential expression analysis. High pCO$_2$ was considered when the value was greater than 700 ppm, and low/normal was considered less than 700 ppm. Early summer was considered as June through July, and late summer was considered for August through October (see Table 1). Expression levels for each transcript were analyzed in a generalized linear model (i.e., glmFit and glmLRT) in edgeR to analyze the effect of pCO$_2$ and the effect of early vs. late summer, and their interaction. Genes with pairwise $p \leq 0.05$ after Benjamini-Hochberg multiple test correction were considered differentially expressed.

To annotate transcripts with functions, expressed transcripts were assigned UniProt descriptions and identifiers by using BLASTx (Altschul 1997) to align contigs against the Swiss-Prot database (UniProt 2017) using the pipeline go_enrichment (Eric Normandeau, see Data Accessibility), with flags --max_target_seqs 1 in outfmt 6 format, and only retaining hits with $E < 10^{-5}$. The UniProt identifier was used as an input for Gene Ontology (GO) enrichment analysis in DAVID bioinformatics (Huang, Sherman, and Lempicki 2009), using differentially expressed lists compared against all expressed genes in the metatranscriptome for those transcripts annotated with UniProt identifiers.
Results

Sampling and environmental conditions

Intake water at the commercial oyster hatchery on east coast Vancouver Island, BC was sampled on four separate days in 2014 (June-August) and on 16 days in 2015 (June-November), with the total samples collected being seasonally balanced between the early summer (i.e., June-July, n = 11) and late summer (i.e., August-October, n = 9). Environmental variables were measured from the intake water, including pH, temperature, salinity, alkalinity, aragonite, calcite and pCO₂, as shown in Table 1 (for complete data see Supplemental File S1).

The focal variable of this study, pCO₂, ranged from 81-1060 ppm (average = 588 +/- 290 ppm). Following IPCC assessments (Gattuso 2014) we classify samples into low pCO₂ < 400 ppm (n = 6), medium pCO₂ 400-700 ppm (n = 4), and high pCO₂ > 700 ppm (n = 10) samples. A two dimensional PCA of environmental conditions shows variability among samples, separating them by low, medium and high pCO₂ conditions (Figure 1). Samples S25 and S35, having concentrations of 397 and 386 ppm pCO₂ cluster close to the medium pCO₂ samples. Samples S22, S26 and S36 have almost identical environmental conditions. The PCA describes 77% of total variation in the first two principal components, with 52% explained by PC1 and 25% by PC2. The most influential variables in the PCA are pCO₂ and pH, separating the samples by their pCO₂ classes. The contributions of pCO₂ and pH are inverse, as are salinity and temperature. Calcium carbonate (CaCO₃), aragonite and calcite differ in their variation from the main axes described by pH/ pCO₂ and salinity/ temperature. Notably, the grouping of samples by environmental variables does not display clustering by sampling months and the associated early vs. late summer classification (Figure 1).

Sequencing and assembly

The metatranscriptome libraries each yielded on average 14.7 M (s.d. = 2.8 M) paired-end reads. Depletion of rRNA in library preparation removed most of the rRNA from a majority of the samples, although for six of the 20 samples, 30-65% of sequenced library remained as rRNA (Figure S1). Residual rRNA was removed in silico, and is primarily comprised of bacterial 23s and eukaryotic 28s rRNA, but also archaeal rRNA.

The reference metatranscriptome was assembled from a total of 191,303,324 mRNA reads (33,050,961,278 bp) originating from pooling the non-rRNA data from 16 of the samples (n = 8 from each of low pCO₂ and high pCO₂). This input produced a final assembly of 8,003,896 contigs (total length = 2,468,090,451 bp; longest contig: 103,407 bp; N50 = 272 bp; number contigs > 500 bp = 600,691). This assembly was compared to other assemblies that used fewer libraries, or those that were individually assembled by sample then subsequently merged (i.e., ‘merged assembly’; see Methods). The collectively-
assembled contigs with the highest number of input samples show fewer multi-mapping reads than did the merged assembly. The percentage of reads aligning a single time increases substantially until four libraries were added, then tapers off to not increase notably with eight, 12, or 16 libraries. The addition of more libraries after four libraries did not substantially increase the percentage of reads mapping, but it also did not increase redundancy as evaluated by multi-mapping. The 16 library collectively-assembled assembly also has a similar number of contigs and total length to the other assemblies (see Supplemental Results; Figure S2). Therefore, this collectively-assembled assembly was chosen to be used for all downstream functional analyses (i.e., the ‘final reference metatranscriptome assembly’).

Aligning reads from individual samples against the final reference metatranscriptome assembly resulted in an average alignment rate per sample of 56% (median = 57%; min. = 37%; max = 68%), with an average of 40% of the total alignments per sample with both read pairs aligning concordantly a single time. On average per sample, 12% of reads align concordantly more than once, which may indicate remaining redundancy in the assembly. Applying low expression filters removed the majority of contigs from the metatranscriptome, retaining 32,866 contigs with CPM ≥ 3.86 in at least 5 of 20 samples.

**Taxonomic analysis**

To infer the community compositions of samples, the taxonomic lineages for expressed transcripts were analyzed. Taxonomic assignment was successful for 19,315 of the 32,866 expressed transcripts (59%), but the taxonomic resolution varied among contigs. For example, 89% of the annotated transcripts resolve to at least the phylum level, 80% to class, 79% to genus and only 30% to the species level. The data includes some expected non-microbial taxa such as fish or molluscs and some suspected false taxonomic assignments, but is vastly dominated by microbes. A total of 708 genera in 104 classes and 58 phyla are annotated, but only 10 microbial phyla are present at an abundance level of greater than 0.5% of the total reads. When combined, these 10 phyla account for over 95% of all reads, yet several other phyla are represented in the data as well (Supplemental File S2). These 10 dominant phyla show an abundance of reads assigned to bacteria (~74%), specifically *Proteobacteria, Bacteroidetes, Firmicutes, Actinobacteria* and *Cyanobacteria*. Overall, fewer reads are assigned to archaea (~11%), viruses (~9%) and eukaryotes (~1.6%). *Archaea* are represented by *Thaumarcheota* and *Euryarcheota*. Viruses are largely in the class *Caudovirales (Uroviricota)* which includes, for example, bacteriophages and cyanophages, and the *Nucleocytoviricota* that represent the nucleocytoplasmic large DNA viruses (NCLDV). Eukaryotes are in the phylum *Bacillariophyta*. Figure 2 shows the variation in community composition of these ten phyla among samples, *Proteobacteria* and *Bacteroidetes* are clearly dominating in most samples, especially in the early season. *Cyanobacteria* and *Bacillariophyta* are also mostly present in early season samples. In
late season there is an increase in relative abundance for the *Thaumarchaeota* and *Euryarchaeota*, but especially the *Uroviricota* and *Nucleocytoviricota* viruses.

The overall variation in community composition among samples was evaluated based on Bray-Curtis similarity in a pairwise distance analysis. When compared at different taxonomic levels, variations in community compositions are congruent, and this is true for comparisons of genus to class (R=0.97), class to phylum (R=0.91) and genus to phylum (R=0.90), all showing significant congruence (Mantel test p ≤ 0.01). A CCA of taxa composition for the 708 genera and environmental variables produces a significant model (P=0.013), explaining 56% of the total variation in taxon composition through environmental variables (Figure 3). Of that variation the first dimension (CCA1) describes 39% of the variation and the second dimension (CCA2) describes 27%. The relationship between sample composition as indicated by the sample labels and the genera (dots) are shown, where taxa are coloured according to the corresponding four kingdoms. Generally, kingdoms are distributed across the CCA, but eukaryotes show some grouping with samples S21, S15 and S13. Viruses show grouping with late summer samples S31, S34, S36 and S37. The effect and strength of environmental variables on the sample composition is indicated by vectors. Temperature and salinity, and pH and pCO$_2$ describe the predominant dimensions of variation. A stepwise regression determined that temperature (p=0.001) and pCO$_2$ (p=0.028) are significant environmental variables affecting the composition of samples. Additionally, early vs. late summer is significant (p=0.023) in separating samples by composition. Correlating all 708 genera to pCO$_2$, our main environmental variable of interest, revealed 67 genera showing a linear correlation of their log$_{10}$ transformed abundances to pCO$_2$ concentrations (p < 0.05; Supplemental File S3), although of these, only two are significant based on multiple test corrected values (i.e., *Methyloceanibacter* and *Deferrribacter*). The genera with the highest linear correlations are predominantly in the *Proteobacteria*, *Cyanobacteria*, and *Picornavirales* and *Caudovirales*. Similarly, 93 genera show significant variation between the early and late season samples, especially *Caudovirales* and *Algavirales* (Supplemental File S4). Of these, 23 are significant based on multiple test corrected values, including *Caudovirales*. The top ten genera for both models are summarized in Table 2.

Gene expression analysis

After the low expression filter was applied, in total 22,121 (67%) of the expressed contigs (n = 32,866) were functionally annotated with UniProt identifiers (BLASTx E < 10-5). All gene expression analysis was conducted using both the annotated and unknown transcripts except for functional enrichment analysis, which depends on the UniProt identifier.

Using the filtered expression data, an unsupervised multidimensional scaling plot (MDS plot) was used to group samples by similarity in gene expression (Figure 4). Based on gene expression signatures this...
plot indicates similarity among samples from a common season, irrespective of sampling year, where samples from June-July (early summer) are separated from August-October (late summer/fall). Early season samples (blue) and late season samples (red) separated on PC1 (Figure 4), with one sample being an exception, S19, which groups outside of its season. Notably, S19 was the only sample from 2014 that was collected in August or later (collected Aug. 6th, 2014; Supplemental File S1). No grouping within the MDS plot was observed for an effect of year nor of technical aspects of sample handling.

Based on the observations of the effect of pCO$_2$ as a key environmental variable on taxonomic composition and a focal variable of the study, a differential expression analysis was conducted using pCO$_2$ separated into either high or medium/low levels, and early summer vs. late summer as binary explanatory variables, as well as their interaction. There is a larger effect of early vs. late summer than of pCO$_2$, where 2,765 transcripts are found differentially expressed based on early vs. late, and 720 based on pCO$_2$. Of these, 45 are differentially expressed in both contrasts (Supplemental File S5). Of the transcripts affected by season, 318 are over-expressed in early summer, and 2447 are over-expressed in late summer/early fall. Of the transcripts affected by pCO$_2$, 553 are over-expressed by high pCO$_2$ and 167 are under-expressed. There are no transcripts showing a significant interaction effect of sampling period and pCO$_2$.

Differentially expressed transcripts were used for Gene Ontology (GO) enrichment analysis (Table 3). Annotated transcripts overexpressed in the early summer (n = 318) are enriched for metabolic processes and biosynthesis (e.g., cellular macromolecule metabolic process; n = 32 genes; p = 0.01), and those overexpressed in the late summer are most notably enriched with viral process (biological process; n = 53; p << 0.0001) and virion (cellular component; n = 30; p << 0.0001). The viral process category is mainly enriched with transcripts annotated from phage taxa (n = 40 of 53 transcripts; Supplemental File S6).

Genes overexpressed in high pCO$_2$ include cobalamin biosynthetic process (n = 5, p = 0.003), organic substance biosynthetic process (n = 81, p = 0.004), and DNA replication (n = 17, p =0.005). Genes with lower expression at high pCO$_2$ include protein maturation (n = 6, p << 0.001) and response to abiotic stimulus (n = 6, p < 0.001).

**Discussion**

*Profiled characteristic conditions indicate short-term environmental fluctuations*

The collected samples reflect typical intake water of oyster hatcheries from standard aquaculture practices on East Coast Vancouver Island (Helm 2004), spanning over several months and two years. The variables were analyzed in a PCA to understand the covariation of environmental variables and how they shape the sampling conditions. The variation in environmental data describes the characteristic water flow and carbon chemistry of intertidal oyster habitats and the region of sampling (Strait of Georgia) (Dickinson et al. 2012; Ianson et al. 2016); salinity and temperature are inversely affected by influxes of cold seawater with
relatively high salinity and warmer fresh water with lower salinities. Water with higher pCO$_2$ concentrations
is characterized by lower pH values and the combined effect of pH, pCO$_2$, and salinity is reflected in the
CaCO$_3$ concentrations, with calcium being one of the salts in seawater. This relationship shows in the
sample separation by low, medium and high pCO$_2$ values and the force loadings of environmental variables.
Calcite and aragonite are both derivatives of CaCO$_3$, varying based on the pCO$_2$ and the pH of water (Doney
2010), and their influence is thus mediated between these two variables. The clustering of samples observed
when considering environmental data highlights the impact of pCO$_2$ and salinity and suggests that
variations in environmental conditions observed in the scale of the present study (two years) are driven by
short term influxes of fresh or marine waters through tidal cycles and upwelling of deep waters.

*Taxonomic composition varies with abiotic factors and early and late summer*

The microbial community analysis here is based on sequences with taxonomic assignment. However, we
expect this to be a representative description, with the possible exception of largely uncharacterized taxa in
reference databases. Derived from taxonomic assignments of contigs and their respective expression levels,
community similarity among samples proved to be congruent across different taxonomic levels. This
validates the approach of using taxonomic assignments down to the genus level. Importantly, it also
indicates that shifts in community composition happen at higher taxonomic levels and not just at the genus
or species level. Contigs assigned to non-microbial taxa or terrestrial taxa are expected to be either the
product of dispersed tissue or cells in the water column or alternatively incorrect taxonomic assignment.

Across samples the microbial communities are clearly dominated by *Proteobacteria*, *Bacteriodetes*,
*Firmicutes*, *Actinobacteria* and *Cyanobacteria*. Both eukaryote and archaea microorganisms are notably
less represented, and even matched in abundance by viral sequences. The described bacterial phyla are
commonly found to be dominant in marine (Sunagawa et al. 2015) and coastal waters (Yung et al. 2016;
Yu et al. 2018), and constitute the microbiome of oysters themselves (Trabal et al. 2012; Lokmer et al. 2015;
Lokmer et al. 2016; Dubé et al. 2019; Stevick et al. 2019). In particular, *Proteobacteria* (*Vibrionales*) and
*Bacteriodetes* (*Flavobacteria*) represent common marine microbes and pathogens (Gomez-Leon et al. 2005;
Schulze et al. 2006; Paillard et al. 2008; Chen et al. 2017). Observations by Stevick and co-workers (Stevick
et al. 2019) found that *Cyanobacteria* were among the dominant phyla in the oyster rearing water
communities and *Synechococcus* is a ubiquitous cyanobacterial genus in coastal environments (Partensky
et al. 1999; Tai and Palenik 2009). Similarly, the abundant eukaryote *Bacillariophyta* and *Chlorophyta,*
both present in the samples taken here, are common phytoplankton in coastal waters (Worden, Nolan, and
Palenik 2004; Armbrust 2009).

Matching the microbial community, the dominant *Caudovirales* includes general bacteriophages
and specifically cyanophages (Weinbauer and Rassoulzadegan 2004). The present results are thus mirroring
the general ubiquity of heterotrophic bacteria, but also indicating their lysis. The Nucleocytoviricota include giant viruses commonly infecting protists (Fischer et al. 2010; VanEtten et al. 2010), but also phycodnaviruses infecting eukaryote phytoplankton (e.g., Chlorophyta), both being abundantly present in coastal waters (VanEtten et al. 1982, 2002). Taken together, the observation of a low presence of cyanobacteria sequences and Bacillariophyta sequences alongside high abundance of plankton viruses and phages in some late season samples, may indicate the lysis of cyanobacteria and algae blooms later in the season. This observation would match phytoplankton bloom patterns described for the northern Strait of Georgia where diatoms (Bacillariophyta) and prasinophytes (Chlorophyta) are mostly dominant with only periodic cyanobacteria blooms (Del Bel Belluz et al. 2021). Additionally, the general decrease in Proteobacteria in the late season samples is observed alongside increased presence of phage (Uroviricota) activity, which also fits with the understanding that bacterial blooms are terminated by viral lysis. As well, viruses show a correlation to late summer samples in the canonical correspondence analysis and the t-test following multiple test correction (i.e., Caudovirales).

In the linear models of taxa to pCO$_2$, our main environmental variable of interest, 67 correlations are significant based on individual tests ($p < 0.05$), but only two are significant based on multiple test corrected p-values, and so the 65 taxa should be considered as putatively characteristic of changes in pCO$_2$ due to water influx. Future studies may indicate the broader reliability of the finding in these taxa. Several of these 67 taxa, including Proteobacteria, Bacteroidetes, Cyanobacteria, Firmicutes and Deferribacteria match bacterial taxa with abundances that are associated with tidal cycles and salinity (Lee et al. 2017; Chen et al. 2019). There remains a large portion of unexplained variation in the taxonomic data when all taxa are considered. The CCA shows that only about 50% of the variation among microbial communities could be explained by the combined environmental variables. The effect of environmental variables on the taxonomic community data in the CCA matches their interplay in the PCA based on environmental data (Figure 1). The CCA also confirms that early vs. late summer sampling time has a significant effect on the community composition, most notably for viruses, although overall to a lesser effect than observed in the functional gene expression results (see below). We could not identify the abiotic variables responsible for the early vs. late summer effect, which indicates other variables that may be involved in community differences are missing (e.g., nutrients or irradiance levels, among others).

In any case, the observed effect and explanatory levels of environmental variables matches previous studies, where temperature has been shown to be a main variable correlated to community composition of coastal marine bacterioplankton (El-Swais et al. 2015; Yung et al. 2016; Yu et al. 2018). Further, Yu and co-workers also established the effect of pH that corresponds to our observations on the influence of pCO$_2$ on community composition. Therefore, although some abiotic or biotic variables may be missing from the study, the results fit with other studies on abiotic factors influencing microbial communities.
Overall, the data show that the microbial communities in coastal waters vary at high taxonomic levels with season. The communities further vary alongside with environmental variables that are driven by tidal cycles and upwelling. As demonstrated, metatranscriptome data can be used to monitor the presence of a wide range of microbes and putative pathogens in seawater, expanding on the use of specific microbe probing (e.g., Merou et al. 2019).

**Differential RNA transcript abundance reveals functional variation between early and late summer**

Environmental variables are not only expected to alter microbial community composition, but also to influence the function of cells on the transcriptomic level. Further, a change in community composition does not necessarily indicate a change in functions within the community if one taxon is replaced by a similar taxon with similar functions. A study comparing taxonomic and functional composition in marine microbial samples revealed a disconnect between taxonomy and functions (Louca et al. 2016). In the functional transcriptome analysis of the present study, early and late summer sampling shows a clear effect on overall transcriptome expression, but environmental variables including pCO2 do not show clear groupings of samples in the gene expression data. Consequently, similar to the community composition results, here functional variation is expected to have been influenced by factors outside of the measured environmental variables (e.g., nutrients and irradiance levels, among others). Observing this pattern in gene expression despite having samples spread out over two years suggests some consistency to this seasonal trend. However, the grouping of the only August sample (S19) from 2014 not with the late summer but rather with the early summer unlike the August samples in 2015 may indicate a difference in the exact timing of the putatively seasonal shift in expression. Without more samples from 2014 after the August sample (S19), it is not possible to tell if the late season signature would appear again, but the early samples from both years consistently group. To determine if this late season signature persists in multiple years, and when the switch from early to late occurs each year, would require future work.

The early vs. late summer effect is most notably enriched for transcripts involved in phage viral activity, matching the findings in the taxonomic analysis. Viruses are recognized as major components in marine ecosystems (Suttle, 2007; Brussaard et al. 2008), but their replication dynamics remain elusive. However, in the North Pacific Ocean it has been observed that viral productivity and abundance is higher in summer (July) than in winter (Jan-Feb) (Gainer et al. 2017), leading these researchers to conclude that seasonality is an important consideration to understand viral dynamics (Jiang and Paul 1994; Tsai et al. 2013). A study within a Korean bay over different seasons found that the number of reads and unique species of viruses identified differed depending on month; the most viral reads were found in March and December and the fewest reads in June and September (Hwang 2017). It is recognized that viral composition changes depend seasonally on a range of factors including temperature, salinity, dissolved
oxygen, primary production and nutrient concentrations (Brum et al. 2015; Fuhrman et al. 2015). Generally, viral abundance, specifically for dsDNA viruses, has been found to correlate with nutrient concentration, as well as heterotrophic bacteria abundance (Wigington et al. 2016; Finke et al. 2017). Viruses infecting bacteria (i.e., bacteriophages) are typically considered among the dominant group of viruses in marine environments (Breitbart et al. 2002; Steward et al. 2013). In the Korean bay study, 73% of the viral reads were from bacteriophage, and 26% were from algal viruses, with only 1% involving other viruses (Hwang et al. 2017). The effect of pCO$_2$ was also investigated in the differential expression analysis, but this had a lesser effect than that the temporal effect observed. Interestingly, in the present study, overexpression of protein chaperones was observed at low pCO$_2$, which is the opposite to previous observations where metatranscriptomic responses of overexpressed chaperonin transcripts was observed in high pCO$_2$ mesocosms (Gilbert et al. 2008).

Marine microbial communities are often comprised of a few dominant species and many rare ones, and results from the Tara Oceans Expedition found approximately 37 k bacterial and archaeal species, 100 k protist groups and 5.5 k double-stranded bacterial and archaeal virus populations (summarized in Moran 2015). The present study is successful in assessing this high degree of diversity in gene expression profiles among samples. As is suggested from the community composition and transcript activity of the present study, previous metatranscriptome studies have also found a substantial increase in viral transcripts after a phytoplankton bloom, presumably due to infected cells in lytic stage (Gilbert et al. 2008).

Overall, taxonomic and transcriptomic profiling complemented each other and when combined provided a comprehensive view of the changes observed in this study. This matches the observation by Salazar et al. (2019) that the correlation between taxonomic composition and functional composition is variable. The community composition analysis provided more information on taxa associated with abiotic factors, and the functional analysis highlighted the large effect of viral activity that differed over time during the summer months. Being able to assess viral activity alongside microbial data is a major advantage of the metatranscriptomic (eRNA) approach applied here to assess marine microbial communities.

Conclusions

In the present study, microbial communities were derived from metatranscriptome data, which avoided primer bias that can occur in amplicon-based approaches, and captures data from all superkingdoms. The observed variability among microbial communities was found to be associated with both temperature and pCO$_2$, as well as other changes that occurred between early and late summer. Changing environmental conditions may occur over short time scales through upwelling or currents, or may have large, structured changes that occur temporally. These temporal and abiotic factors appeared to be disconnected and were two different trends occurring in the marine environment. The functional gene expression analysis pointed
to a strong difference in viral activity moving into the late season, and a much lesser effect of abiotic factors such as pCO₂, temperature and salinity. Together, these analyses provide community composition and functional gene differences associated with abiotic factors and time, and likely captured viral termination of bacterial, cyanobacterial and algal blooms later in the season. Being able to indiscriminately assess eukaryote, prokaryote and viral taxa, describing a combined bloom and viral lysis across several host virus systems proved to be the advantage of using metatranscriptome analysis. The improved understanding of changes in marine microbial communities with abiotic factors helps to monitor water bodies. Metatranscriptomics of eRNA allowed the characterization of both community and gene expression activity changes simultaneously, providing a comprehensive view of these dynamic water bodies.

Acknowledgements

This project would not have been possible without the involvement, contributions, and sample provision from the industry partner, Island Scallops (ISL; CEO Robert Saunders). J. Finke and B. Sutherland were supported during this work by a grant from the Gordon and Betty Moore Foundation (GBMF, Grant number 5600) awarded to C. Suttle and K. Miller. This project was supported by an Aquaculture Collaborative Research and Development Program (ACRDP) grant from Fisheries and Oceans, Canada, awarded to K. Miller (Grant number P-14-02-001). The authors declare that they do not have any conflicts of interest in this work.

Data Accessibility

Supplemental files including the reference metatranscriptome (fasta), quantified transcript expression levels (.csv), and annotation for the functional and taxonomic analyses (.txt) are available on FigShare: https://doi.org/10.6084/m9.figshare.18128966.v1 Pipeline for analyzing metatranscriptome data: https://github.com/bensutherland/eRNA_taxo Pipeline used for annotating transcripts: https://github.com/enormandeau/go_enrichment Script to extract taxonomic lineages: https://github.com/janfelix/bioinformatic_tools/blob/master/taxid2lineage.py
References


https://doi.org/10.3354/dao054219


https://doi.org/10.1242/jeb.061481.


595
596 Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. “Systematic and Integrative Analysis
597 of Large Gene Lists Using DAVID Bioinformatics Resources.” 
599
600 Hwang, Jinik, So Yun Park, Miryak Park, Sukchan Lee, and Taek Kyun Lee. 2017. “Seasonal Dynamics
601 and Metagenomic Characterization of Marine Viruses in Goseong Bay, Korea.” 
603
604 Iason, Debby, Susan E. Allen, Benjamin L. Moore-Maley, Sophia C. Johannessen, and Robie W.
605 Macdonald. 2016. “Vulnerability of a Semiclosed Estuarine Sea to Ocean Acidification in
606 Contrast with Hypoxia.” 
608
609 Jiang, S C, and J H Paul. 1994. “Seasonal and Diel Abundance of Viruses and Occurrence of
610 Lysogeny/Bacteriocinogeny in the Marine Environment.” Marine Ecology Progress Series 104:
612
613 Kim, Bo Mi, Kyobum Kim, Ik Young Choi, and Jae Sung Rhee. 2017. “Transcriptome Response of the
614 Pacific Oyster, Crassostrea Gigas Susceptible to Thermal Stress: A Comparison with the Response
616 https://doi.org/10.1007/s13273-017-0011-z.
617
620 of the Intergovernmental Panel on Climate Change.”
621
623 of Ribosomal RNAs in Metatranscriptomic Data.” 
625
628
631
632 Lee, Eunhee, Doyun Shin, Sung Pil Hyun, Kyung Seok Ko, Hee Sun Moon, Dong Chan Koh, Kyoochul
633 Ha, and Byung Yong Kim. 2017. “Periodic Change in Coastal Microbial Community Structure
634 Associated with Submarine Groundwater Discharge and Tidal Fluctuation.” Limnology and
636
639 https://doi.org/10.1093/bioinformatics/btl158.
640
641 Lokmer, Ana, M Anouk Goedknegt, David W Thielges, Dario Fiorentino, Sven Kuenzel, John F Baines,
643 Microbiota across Multiple Scales” Front. Microbiol. 7 (August): 1–18.
645
648 https://doi.org/10.1038/ismej.2014.160.
649
652 https://doi.org/10.1126/science.aaf4507
653
654 Lv, Xiaofei, Bin Ma, Junbao Yu, Scott X. Chang, Jianming Xu, Yunzhao Li, Guangmei Wang,
655 Guangxuan Han, Guan Bo, and Xiaojing Chu. 2016. “Bacterial Community Structure and Function
656 Shift along a Successional Series of Tidal Flats in the Yellow River Delta.” 
658
660 Crassostreae , the Etiological Agent of Juvenile Oyster Disease (Now to Be Known as Roseovarius
https://doi.org/10.1016/j.aquaculture.2007.04.008.


Salazar, Guillem, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Christopher M. Field, Luis Pedro Coelho, Corinne Cruaud, Stefan Engelen,
Van Etten, J L, L C Lane, and D D Dunigan. 2010. “DNA Viruses: The Really Big Ones (Giruses).”


Tables

Table 1. Summary statistics of ocean chemistry variables. pCO₂ was the target abiotic variable of the study, but the effect of the other variables was also considered in the context of metatranscriptome profiles. For calcite and aragonite, the saturation states are shown.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Median</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.56</td>
<td>8.35</td>
<td>7.8</td>
<td>7.78</td>
<td>0.22</td>
</tr>
<tr>
<td>Temperature (ºC)</td>
<td>11.4</td>
<td>16.7</td>
<td>14.4</td>
<td>14</td>
<td>1.5</td>
</tr>
<tr>
<td>Salinity (%)</td>
<td>28</td>
<td>29</td>
<td>28.2</td>
<td>28</td>
<td>0.41</td>
</tr>
<tr>
<td>Alkalinity (µmol/kg)</td>
<td>795.1</td>
<td>2634.2</td>
<td>1501.1</td>
<td>1400</td>
<td>465.1</td>
</tr>
<tr>
<td>Aragonite (Ω)</td>
<td>0.34</td>
<td>2.72</td>
<td>0.93</td>
<td>0.65</td>
<td>0.69</td>
</tr>
<tr>
<td>Calcite (Ω)</td>
<td>0.54</td>
<td>3.15</td>
<td>1.39</td>
<td>1.04</td>
<td>0.86</td>
</tr>
<tr>
<td>pCO₂ (ppm)</td>
<td>81.2</td>
<td>1059.9</td>
<td>588.3</td>
<td>606.6</td>
<td>289.9</td>
</tr>
</tbody>
</table>
Table 2. Top ten significant genera of linear models against low, medium, and high pCO$_2$ concentration and T-test of early vs. late summer. P-values express the model significance and P-adj. is adjusted p-value with Benjamini-Hochberg. Statistic values indicate positive and negative correlations.

**Linear model genus vs. pCO$_2$**

<table>
<thead>
<tr>
<th>R2</th>
<th>P-value</th>
<th>P-adj</th>
<th>Slope</th>
<th>Genus</th>
<th>Order</th>
<th>Phylum</th>
<th>Superkingdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.718</td>
<td>&lt;0.001</td>
<td>0.038</td>
<td>0.001</td>
<td>Methyloceanibacter</td>
<td>Rhizobiales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>0.597</td>
<td>&lt;0.001</td>
<td>0.038</td>
<td>-0.002</td>
<td>Deferrribacter</td>
<td>Deferribacterales</td>
<td>Deferribacteres</td>
<td>Bacteria</td>
</tr>
<tr>
<td>0.552</td>
<td>&lt;0.001</td>
<td>0.063</td>
<td>-0.001</td>
<td>Gossypium</td>
<td>Malvales</td>
<td>Streptophyta</td>
<td>Eukaryota</td>
</tr>
<tr>
<td>0.614</td>
<td>0.001</td>
<td>0.098</td>
<td>-0.002</td>
<td>Thioalkalivibrio</td>
<td>Chromatiales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>0.694</td>
<td>0.001</td>
<td>0.102</td>
<td>-0.002</td>
<td>Mitrocomella</td>
<td>Leptotheccata</td>
<td>Cnidaria</td>
<td>Eukaryota</td>
</tr>
<tr>
<td>0.466</td>
<td>0.001</td>
<td>0.102</td>
<td>-0.002</td>
<td>Sogarnavirus</td>
<td>Picornavirales</td>
<td>Pisuviricota</td>
<td>Viruses</td>
</tr>
<tr>
<td>0.460</td>
<td>0.001</td>
<td>0.102</td>
<td>-0.001</td>
<td>Methylovorus</td>
<td>Nitrosomonadales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>0.447</td>
<td>0.001</td>
<td>0.108</td>
<td>-0.001</td>
<td>Methylophilus</td>
<td>Nitrosomonadales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>0.462</td>
<td>0.001</td>
<td>0.108</td>
<td>-0.001</td>
<td>Aeromonas</td>
<td>Aeromonadales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>0.470</td>
<td>0.002</td>
<td>0.155</td>
<td>0.001</td>
<td>Steinhofvirus</td>
<td>Caudovirales</td>
<td>Uroviricota</td>
<td>Viruses</td>
</tr>
</tbody>
</table>

**T-test genus vs. early and late summer**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>P-value</th>
<th>P-adj</th>
<th>Stderr</th>
<th>Genus</th>
<th>Order</th>
<th>Phylum</th>
<th>Superkingdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.053</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.121</td>
<td>Coregonus</td>
<td>Salmoniformes</td>
<td>Chordata</td>
<td>Eukaryota</td>
</tr>
<tr>
<td>6.884</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>0.118</td>
<td>Biomphalaria</td>
<td></td>
<td>Mollusca</td>
<td>Eukaryota</td>
</tr>
<tr>
<td>4.129</td>
<td>0.001</td>
<td>0.030</td>
<td>0.169</td>
<td>Leucotheavirus</td>
<td>Caudovirales</td>
<td>Uroviricota</td>
<td>Viruses</td>
</tr>
<tr>
<td>3.759</td>
<td>0.004</td>
<td>0.088</td>
<td>0.207</td>
<td>Nesterenkonia</td>
<td>Micrococcales</td>
<td>Actinobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>3.713</td>
<td>0.002</td>
<td>0.051</td>
<td>0.217</td>
<td>Glaciecola</td>
<td>Alteromonadales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>3.693</td>
<td>0.002</td>
<td>0.062</td>
<td>0.164</td>
<td>Emcibacter</td>
<td>Emcibacterales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>3.614</td>
<td>0.002</td>
<td>0.062</td>
<td>0.125</td>
<td>Glaesserella</td>
<td>Pasteurellales</td>
<td>Proteobacteria</td>
<td>Bacteria</td>
</tr>
<tr>
<td>3.566</td>
<td>0.004</td>
<td>0.095</td>
<td>0.217</td>
<td>Nephromyces</td>
<td>Nephromycida</td>
<td>Apicomplexa</td>
<td>Eukaryota</td>
</tr>
<tr>
<td>3.560</td>
<td>0.006</td>
<td>0.110</td>
<td>0.194</td>
<td>Tannerella</td>
<td>Bacteroidales</td>
<td>Bacteroidetes</td>
<td>Bacteria</td>
</tr>
<tr>
<td>3.221</td>
<td>0.006</td>
<td>0.110</td>
<td>0.112</td>
<td>Borrelia</td>
<td>Spirochaetales</td>
<td>Spirochaetes</td>
<td>Bacteria</td>
</tr>
</tbody>
</table>
Table 3. Gene Ontology (GO) enrichment for pCO$_2$ and season. GO enrichment analysis indicates the effect of pCO$_2$ and season on several biological processes, most notably a highly significant enrichment for viral processes in the late season. Columns shown are the GO Term, count in gene list, p-value for enrichment test, count in background list, and fold enrichment. Full GO enrichment results are presented in Supplemental Materials.

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Count</th>
<th>P-value</th>
<th>Count (background)</th>
<th>Fold Enrich.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overexpr. in high pCO$_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009236−cobalamin biosynthetic process</td>
<td>5</td>
<td>0.0030</td>
<td>18</td>
<td>7.9</td>
</tr>
<tr>
<td>GO:0006260−DNA replication</td>
<td>17</td>
<td>0.0048</td>
<td>226</td>
<td>2.1</td>
</tr>
<tr>
<td>GO:0019058−viral life cycle</td>
<td>8</td>
<td>0.015</td>
<td>75</td>
<td>3.0</td>
</tr>
<tr>
<td>GO:0039693−viral DNA genome replication</td>
<td>4</td>
<td>0.020</td>
<td>17</td>
<td>6.7</td>
</tr>
<tr>
<td>Overexpr. in low pCO$_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0051604−protein maturation</td>
<td>6</td>
<td>6.1E-06</td>
<td>30</td>
<td>21.0</td>
</tr>
<tr>
<td>GO:0006950−response to stress</td>
<td>10</td>
<td>0.0032</td>
<td>344</td>
<td>3.1</td>
</tr>
<tr>
<td>GO:0006457−protein folding</td>
<td>5</td>
<td>0.0067</td>
<td>83</td>
<td>6.3</td>
</tr>
<tr>
<td>Overexpr. in early season</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0003735−structural constituent of ribosome</td>
<td>8</td>
<td>0.0059</td>
<td>134</td>
<td>3.6</td>
</tr>
<tr>
<td>Overexpr. in late season</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0016032−viral process</td>
<td>53</td>
<td>7.3E-42</td>
<td>91</td>
<td>9.7</td>
</tr>
<tr>
<td>GO:0006260−DNA replication</td>
<td>55</td>
<td>1.8E-20</td>
<td>226</td>
<td>4.1</td>
</tr>
<tr>
<td>GO:0051701−interaction with host</td>
<td>23</td>
<td>5.1E-16</td>
<td>45</td>
<td>8.5</td>
</tr>
<tr>
<td>GO:0046718−viral entry into host cell</td>
<td>17</td>
<td>1.2E-13</td>
<td>27</td>
<td>10.5</td>
</tr>
</tbody>
</table>
Figures

Figure 1. Principal Components Analysis (PCA) of samples based on environmental conditions. Samples are indicated as labels, their corresponding pCO$_2$ classification is indicated by colour (blue=low, violet=medium, red=high). Arrow direction and length indicate the relative effect and strength of environmental variables. Here the positions of samples are entirely from environmental conditions, not based on taxonomic or gene expression data.
**Figure 2.** Community composition of the top 10 phyla across samples. Stacked bar-plot of the relative abundance (total number of alignments to taxa) across samples for microbial phyla above 0.5% total abundance, phyla are indicated by color coding in legend, samples are arranged by early season vs. late season.
Figure 3. Canonical correspondence analysis (CCA) showing the effect of environmental variables on the variation in relative abundance of genera among samples. Samples are shown as labels and the relative association of genera is shown as crosses. Crosses for genera are color coded by their corresponding superkingdom. The effect of environmental variables on the composition of genera in the samples is indicated by grey arrows, the arrow length corresponds to scaled effect strength. The effect of early and late summer is indicated in grey labels.
Figure 4. Unsupervised multidimensional scaling (MDS) plot on samples based on gene expression. Dimension 1 explains the most variation, separating the late and early season samples. Samples are labeled by sample number and colors indicate early vs. late season. Samples S13, S15, S16, and S19 are from 2014, and the rest are from 2015. Full details on samples can be viewed in Supplemental File S1.
Supplementary Materials

Supplemental Figure 1. Proportion of non-rRNA and rRNA reads per sample. Results indicate that for a majority of the samples, the rRNA depletion was successful, enriching the amount of messenger RNA for analysis.
Supplemental Figure 2. Assessment of reference metatranscriptome assemblies. (A) Total contig numbers (red) increase with input data size linearly, and so does total length of output assembly (black). The individual assembly of eight libraries that were subsequently merged together (shown to the right of the hatched line) has a similar number of contigs and length as the eight libraries assembled simultaneously. (B) The percentage of unaligned reads decreases as more libraries are added, initially with a rapid decrease until four libraries are added, and then with a more gentle slope as additional libraries are added. The benefit of increasing from eight to 16 libraries is not as evident in the percentage of reads aligned, suggesting that once four or eight libraries are assembled together, not much benefit is added by increasing the number of libraries in this dataset. Interestingly, the merged assembly has slightly fewer unaligned reads, but double the number of multi-mapping reads (align > 1, grey), indicating a strong amount of redundancy is still present in the merged assembly that is not present in the 16 library simultaneous assembly.
Supplemental Data

Supplemental File S1. Complete environmental and metadata for all samples.

Supplemental File S2. Overview of lineage data assigned to contigs.

Supplemental File S3. Full list of genera with significant linear models to pCO$_2$ concentrations.

Supplemental File S4. Full list of genera with significant different abundances between early and late summer samples.

Supplemental File S5. Full differential gene expression analysis including genes differentially expressed between pCO$_2$ levels and between seasons, as well as genes found in both comparisons.

Supplemental File S6. Full Gene Ontology analysis including GO enrichment for each differentially expressed gene list for Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). Viral transcripts identified in BP in the season differential analysis are also included.