MarMic Laboratory Rotation I: March 6 – April 28, 2017
Department of Molecular Ecology

Metatranscriptome Analysis of Helgoland Marine Sediments during a spring phytoplankton bloom

M. Sc. student:
Matthew Schechter mschecht@mpi-bremen.de

Department head:
Prof. Dr. Rudolf Amann ramann@mpi-bremen.de

Supervisors:
David Probandt dproband@mpi-bremen.de
Dr. Katrin Knittel kknittel@mpi-bremen.de

Submitted: May 12, 2017

Abstract:

Phytoplankton blooms over continental shelves contribute significantly to total ocean primary production and the resulting phytodetritus is exported to the sublittoral marine sediments. Due to this, it is important to understand how microbial communities respond to these carbon inputs for global carbon cycling implications. Metatranscriptomic analysis was used to measure gene expression of the upper sediment and the lower sediment off the coast of Helgoland, Germany with an emphasis on glycoside hydrolase expression. Analysis revealed there was no significant gene expression difference between the upper and lower sediment. This may have been caused by hydrological forces such as advection and sediment resuspension.

Table of Contents

Table of Contents

1. Introduction

Continental shelf marine sediments are dynamic environments that are actively involved in the global carbon cycle. In fact, 20% of all ocean primary production occurs in the water column above continental shelves (Liu 2010). This is due to high nutrient input from rivers and aeolian dust which produces seasonal phytoplankton blooms. At water depths ≤ 100 meters, 15%-50% of all primary production carbon products are exported to the sediment (Bauer, Cai et al. 2013, Woulds, Bouillon et al. 2016). Continental shelf sediments tend to be permeable, course sands (Huettel, Berg et al. 2014). This allows for advection forces to transport oxygen and water column-derived phytodetritus into shelf sediments at fast rates (Ehrenhauss, Witte et al. 2004). With greater carbon and O_2 input, microbial respiration rates increase the flux of carbon remineralization.

Organic carbon substrate induced bacterioplankton successions have been described in depth for the waters off Helgoland, Germany (Teeling, Fuchs et al. 2016). There is vast evidence that members of bacterial classes *Flavobacteriia*, *Gammaproteobacteria*, and *Alphaproteobacteria* arise based on availability of carbon substrates (i.e. polysaccharides) from phytodetritus after phytoplankton spring blooms. This has been analyzed on a population level via CARD-FISH, proteomics, and metagenomics.

During the progression of an algal bloom, different polysaccharides become available. For example, exopolysaccharides are abundant at the start, while intracellular carbon substrates (i.e. laminarin) become available at the die-off of the bloom (Teeling, Fuchs et al. 2016). The enzymes investigated by Teeling et. al. via metagenomics that degrade algal polysaccharides are referred to as Carbohydrate Active enZymes (CAZymes)(Cantarel, Coutinho et al. 2009). The abundance of CAZymes can infer which polysaccharides are available during a spring phytoplankton bloom (Teeling, Fuchs et al. 2016). CAZymes are divided into classes that reflect their function and amino acid composition: Glycoside Hydrolases (GHs), Glycosyl Transferases (GTs), Polysaccharide lyases (PLs), and Carbohydrate esterases (CEs). GHs are particularly important for marine bacteria heterotrophic metabolism of polysaccharides because they cleave glycosidic linkages yielding monomers that can be further respired or fermented.

Since advection pulls oxygenated water into continental shelf sediment, respiration occurs to metabolize phytodetritus. When O_2 becomes limited deeper in the sediment, microbial communities change metabolic strategies to fermentation and other carbon oxidation methods with terminal electron acceptors such as nitrate, iron, manganese, and sulfate. Benthic microbial community responses to phytodetritus are interlinked to biogeochemical rates and fluxes of organic carbon in this environment. Studies via 16S rRNA gene tag sequencing and stable isotope labeled carbon have shown that benthic microbial communities are poised to quickly remineralize when there are spikes on incoming carbon (Gihring, Humphrys et al. 2009). Yet, no studies have emphasized gene expression levels of the benthic microbial community in response to phytodetritus input. The aim of this investigation is to analyze continental shelf marine sediments from Helgoland, Germany to uncover differential gene expression between the upper sediment (US) and the lower sediment (LS) layers with a focus on CAZymes.

2. Methods

2.1. Sampling site description

Samples were taken on May 24, 2016 from a *marine sub-littoral zone* (ENVO:01000126) 200 meters east of the Kabeltonne German ecological research station 40 km off the German North Sea coast between the Islands of Helgoland an Düne. The water depth was on average 8 m, the water temperature was between 11-13 °C, and chlorophyll a concentration was 4 mg m^{-3}. The samples consisted of *sandy sediment* (ENVO:01000118) with a median diameter of 34X micrometers and permeability of 8.48 x 10^{-11} m^2 (calculated from Eq. 1).

2.2. Sampling procedure

The samples were taken from sublittoral sediments in the southern North Sea at site Helgoland roads (HelRoads) on May 25, 2016 verb missing. Three sediment push cores were retrieved by scientific divers from a water depth of 8 m. Subsequently, the cores were transported to the lab within 30 minutes at in situ temperature and immediately sectioned into 0-1.5 cm upper sediment (US) and 5-6 cm lower sediment (LS) frozen using liquid nitrogen (-196 °C).

2.3. RNA isolation, cDNA synthesis, and sequencing

Total RNA was extracted from each sediment core (three replicates per depth layer) using the RNeasy isolation kit (Qiagen, Helden, Germany) by the Vertis Biotechnologie AG (Freuising, Germany) including the DNase treatment. Clear 16S and 23S rRNA absorption peaks of the isolated RNA for all 6 RNA samples using a Shimadzu MultiNA microchip. The Illumina stranded RNA library preparation protocol was followed according to the manual in the TruSeq stranded RNA kit, with the following modifications: (1) rRNA depletion step was omitted; (2) RNA fragmentation was shortened from 8 to 2 minutes to achieve longer RNA fragments. cDNA fragment length ranged from 150nt to 450nt. After library preparation cDNA lengths increased, ranging from 230 to 450nt. The total RNA library fragments were extracted from the preparative gel and sequenced. All 6 libraries were combined equally and sequenced in three consecutive Illumina HiSeq Rapid v2 runs in single read mode (250nt).

2.4. Sequence processing

All 6 samples were sequenced on three separate Illumina HiSeq single end rapid mode runs (250nt). Each sequence file was error corrected and quality trimmed using bbmap tools (ecc.sh; bbduk, forcetrimleft = 15 forcetrimright = 237), and remaining PhiX sequencing control was removed using bbduk as well. Sequence file quality was checked using FASTQC for before and after quality control comparison. Next, rRNA reads were removed from each sequence file using SortMeRNA using the following reference databases: prokaryotic 5S, 23S and 16S SILVA REF rRNA; eukaryotic 5S, 8S, 28S, and 18S SILVA REF rRNA (v. 128), rfam 5S and 5.8s. Text depicting the exact commands and tools with syntax and flags used for sequences processing can be found in the appendix.

2.5. Metatranscriptome Comparison and taxonomic identification

ORF prediction was done by FragGeneScan (illumina training set), annotation of ORF-predicted reads was done by Ultrafast Protein Classification (UProC) Pfam (Meinicke, 2015), and GH Hidden Markov Models were used for GH annotation from dbCAN (Yin et. al., 2012).

2.6. Gene Ontology (GO) annotation and Environment Ontology (EnvO)

After Pfam annotation, Pfam module hits were annotated with Gene Ontology terms using a custom R-script (Christiane Hassenrück; https://github.com/chassenr/NGS/tree/master/META_G_T). Pfam annotations were accompanied with Gene Ontology Biological Process (BP) and Molecular Function (MF). BP and MF were clustered into higher order family terms to extract summary conclusions from the overall gene expression in the two depths. Additionally, total raw BP and MF annotations were analyzed for percent abundance. For visualization of clustered GO terms, REVIGO (reduce + visualize Gene Ontology) was used to summarize GO terms that were annotated to the samples metatranscriptomes Pfam annotation (Supek, Bosnjak et al. 2011).

Environmental Ontology (EnvO) terms were used to describe the sampling location. Terms in this document have this syntax: *term* (ENVO:########). Additionally, all EnvO terms are listed in the appendix (Table 11) with hyperlinks (Buttigieg, Morrison et al. 2013, Buttigieg, Pafilis et al. 2016).

2.7. Statistical analysis

For detection of transcripts that differentiate US and LS, the R-package "Analysis of Differential Abundance Taking Sample Variation into Account" (ALDEx2) was used (Fernandes et. al., 2013). Welch's t-test and 128 Monte Carlo simulations were used to estimate underlying distributions.

2.8. Permeability Calculation

Sample sediment permeability was calculated via median grain size (d_g^2) with empirical relation (Gangi 1985):

Eq. 1: $k = Dar \times 735 \times 10^6 \times d_g^2$

where k is permeability in m^2 and Dar (9.869×10^{-13}) is conversion of the unit Darcy into m^2.

3. Results

3.1. RNAseq processing outcome

Raw transcript data files were pre-processed before transcriptome annotation steps. The bbmap sequencing error correction tool suite ecc.sh yielded marginal differences for all samples when analyzed via FASTQC for "per sequence base quality." After ribosomal RNA separation via SortMeRNA, only 6.11%-7.42% US and 6.62%-8.05% LS of the total RNA pool were sorted as mRNA (Table 1). Next,

leftover PhiX sequencing control RNA was removed using bbduk. This lowered the samples triplicates average percent of retained mRNA to 6.69% US and 7.01% LS (Table 1).

Table 1: Sequence processing statistics. Given are triplicated averages of upper sediment (US) and lower sediment (LS)

Sequence processing step	US (0-1.5 cm) read count	LS (5-6 cm) read count
Pre-processed read count	84,043,645	85,014,035
After rRNA separation (SortMeRNA)	5,722,094 (6.8%)	6,069,072 (7.1%)
After PhiX removal (bbduk)	5,626,046 (6.7%)	5,960,376 (7.0%)

84.43% US and 85.11% LS were predicted to have open reading frames of the processed mRNA reads (Table 2). Interestingly, only 16.28% US and 16.27% mRNA reads were annotated using Pfam reference database (v. 28).

Table 2: mRNA annotation statistics, counts and percentages are triplicated averages of US and LS.

Annotation	US (0-1.5 cm) read count	LS (5-6 cm) read count
mRNA	5,626,046 reads	5,960,376 reads
ORF predicted mRNA (% of mRNA reads)	4,749,990 reads (84%)	5,072,952 reads (85%)
Pfam annotated mRNA (% of mRNA reads)	915,972 (16%)	969,860 (16%)

3.2. Gene expression in upper and lower sediment depth layer

Gene expression analysis of the Pfam annotation counts matrix (ALDEx2) did not reveal any transcripts that contributed significantly to US and LS metatranscriptome differentiation (Figure 1). In fact, most annotated transcripts (relative gene expression) were expressed at highly similar levels in both depths. In figure 1, transcript expression levels are visualized on a color scale from black to grey to red. Black dots symbolize Pfam modules that are similarly expressed in both sediment layers, while grey dots symbolize moderate differential expression and red dots represent differentially expressed genes, respectively. Pfam annotations that contributed most to differentiating the US and LS are listed in the appendix (Table 9).

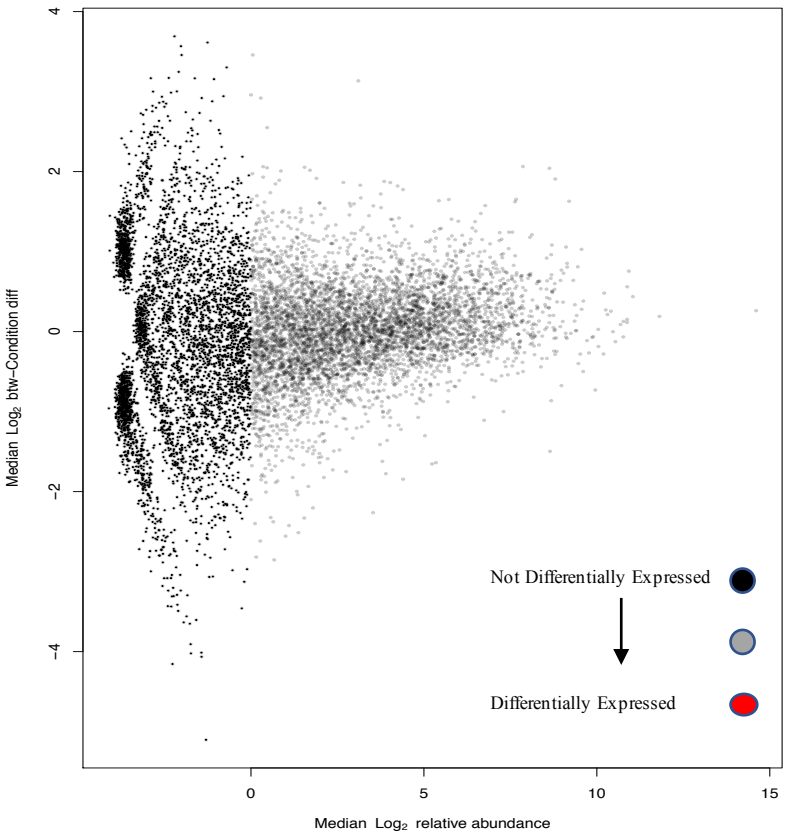

Fig. 1: The graph above represents relative gene expression plotted against gene expression variance between sediment depths. The x-axis the is log transformed expression data while the y-axis is the variance in expression between US and LS.

3.3. Abundant functions detected in upper and lower sediment depth layers

The top two expressed genes (averaged triplicates of Pfam counts) in US and LS were 'photosynthetic reaction center protein' (PF00124) at 11.25% and 10.25% of total transcripts in US and LS, respectively, and 'Ribulose bisphosphate carboxylase large chain, catalytic domain' (PF00016) at 1.67% and 1.37% in US and LS, respectively (Table 3). Furthermore, both depths shared the top 11 highest relative abundance expressed genes, but in different order of relative contribution to total annotated Pfams. Three of the top expressed genes were photosynthesis related: 'photosynthetic reaction center protein' (PF00124) at 11.25% in US and 10.25% in LS; 'photosystem II protein' (PF00421) at 1.03% in US and 0.80% in LS; 'Photosystem I psaA/psaB protein' (PF00223) at 0.90% in US and 0.71% in LS.

Table 3: Pfam annotations of top 11 mRNA transcripts from upper sediment and lower sediment depth layer

	US (0-1.5 cm)			LS (5-6 cm)		
	Count Avg.	Pfam	Description	Count Avg.	Pfam	Description
1	11.25%	PF00124	Photosynthetic reaction centre protein	10.25%	PF00124	Photosynthetic reaction centre protein
2	1.67%	PF00016	Ribulose bisphosphate carboxylase large chain, catalytic domain	1.37%	PF00016	Ribulose bisphosphate carboxylase large chain, catalytic domain
3	1.28%	PF00118	TCP-1/cpn60 chaperonin family	0.88%	PF07690	Major Facilitator Superfamily
4	1.03%	PF00421	Photosystem II protein	0.87%	PF00069	Protein kinase domain
5	0.99%	PF00009	Elongation factor Tu GTP	0.87%	PF00005	ABC transporter

			binding protein			
6	0.81%	PF07690	Major Facilitator Superfamily	0.80%	PF00421	Photosystem II protein
7	0.90%	PF00223	Photosystem I psaA/psaB protein	0.74%	PF00012	Hsp70 protein
8	0.75%	PF00005	ABC transporter	0.71%	PF00009	Elongation factor Tu GTP binding protein
9	0.70%	PF00069	Protein kinase domain	0.71%	PF00223	Photosystem I psaA/psaB protein
10	0.84%	PF00012	Hsp70 protein	0.68%	PF00115	Cytochrome C and Quinol oxidase polypeptide I
11	0.62%	PF00115	Cytochrome C and Quinol oxidase polypeptide I	0.66%	PF00118	TCP-1/cpn60 chaperonin family

Pfam annotations were joined with their respective GO terms and were analyzed in low resolution "super clusters" as well as high resolution raw terms to uncover conclusions about functionality of the metatranscriptomes (Table 4, Table 5). BP and MF super clustering revealed photosynthesis and housekeeping genes to be the most upregulated functions in both sediment layers. Of the clustered molecular functions, the second most clustered MF term was "Mg 2^+ binding".

Table 4: Gene ontology biological process REVIGO super clusters

Super Cluster	US (0-1.5 cm)	LS (5-6 cm)
Photosynthesis, light reaction	38%	35%
Oxidation-reduction process	22%	24%
Metabolism	9%	11%
Carbohydrate metabolism	2%	2%

Table 5: Gene ontology molecular function REVIGO super clusters

Super Cluster	US (0-1.5 cm)	LS (5-6 cm)
Electron transport within photosynthesis	25%	25%
Mg 2^+ binding	17%	15%
Oxidoreductase activity	9%	9%
DNA binding	8%	6%

In both US and LS, the most abundant raw BP terms were: (1) 'photosynthesis, light reaction' (GO:0019684); (2) 'photosynthetic electron transport in photosystem II' (GO:0009772); (3) 'oxidation-reduction process' (GO:0055114) (Table 6). Concurrently, the most abundant GOMF terms were: (1) 'electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity' (GO:0045156); (2) 'ATP binding' (GO:0005524); (3) 'oxidoreductase activity' (GO:0016491); (4) 'DNA binding' (GO:0003677) (Table 7).

Table 6: Metatranscriptome annotations with Gene Ontology Biological Processes (BP). Ordered by percent abundance of BP found in metatranscriptome.

	US (0-1.5 cm)		LS (5-6 cm)	
	Avg. Count	BP	Avg. Count	GBP
1	18.14%	photosynthesis, light reaction	16.77%	photosynthesis, light reaction
2	16.61%	photosynthetic electron transport in photosystem II	15.55%	photosynthetic electron transport in photosystem II

3	9.91%	oxidation-reduction process	11.56%	oxidation-reduction process
4	5.23%	translation	5.74%	metabolic process
5	4.58%	metabolic process	3.72%	translation
6	3.28%	transport	3.37%	transport
7	2.23%	regulation of transcription, DNA-templated	2.47%	transmembrane transport
8	2.28%	photosynthesis	2.07%	regulation of transcription, DNA-templated
9	2.11%	Transcription, DNA-templated	1.99%	proteolysis
10	2.00%	Transmembrane transport	1.65%	photosynthesis

Table 7: Metatranscriptome annotations with Gene Ontology Molecular Functions (MF)

	US (0-1.5 cm)		LS (5-6 cm)	
	Avg. Count	MF	Avg. Count	MF
1	14.97%	electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity	13.67%	electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity
2	9.22%	ATP binding	8.75%	ATP binding
3	4.91%	Oxidoreductase activity	5.98%	Oxidoreductase activity
4	4.77%	Structural constituent of ribosome	3.93%	catalytic activity
5	4.57%	DNA binding	3.30%	structural constituent of ribosome
6	2.85%	Catalytic activity	2.42%	protein binding
7	2.56%	Magnesium ion binding	2.22%	magnesium ion binding
8	2.02%	protein binding	2.01%	electron carrier activity
9	1.98%	electron carrier activity	1.89%	Heme binding
10	1.66%	heme binding	1.31%	ATPase activity

3.4. Expression of glycoside hydrolases

To identify potential candidate enzymes for the degradation of carbohydrates, dbCAN Hidden Markov Models were used to look at the expression of glycoside hydrolases. Of total transcripts, 8.98×10^{-5}% (of total reads) and 8.34×10^{-5}% (of total reads) were assigned to glycoside hydrolases for the upper and lower sediment depth layer, respectively. These reads represented 24 different glycoside hydrolases. There were not clear differences in the expression levels of glycoside hydrolases between sediment layers. The top expressed was GH family 109 (37.64% US, 32.62% LS of total expressed GH), followed by 23, 3, 16 (Table 8). The latter 21 GH were similar but found in different orders of relative expression.

Table 8: Metatranscriptome annotations with Gene Ontology Molecular Functions (MF)

	US (0-1.5 cm)			LS (5-6 cm)		
	GH HMM	% of GH	Description	GH HMM	% of GH	Description
1	GH109	37.64%	α-N-acetylgalactosaminidase, removes antigens from red blood cells, could be cell wall related	GH109	32.62%	α-N-acetylgalactosaminidase, removes antigens from red blood cells, could be cell wall related
2	GH23	24.40%	lysozyme/peptidoglycan/chitinase	GH23	26.03%	lysozyme/peptidoglycan/chitinase
3	GH3	12.16%	beta-glucan; cellulosic biomass degradation, plant and bacterial cell wall remodeling, energy metabolism and pathogen defense	GH3	16.03%	beta-glucan; cellulosic biomass degradation, plant and bacterial cell wall remodeling, energy metabolism and pathogen defense
4	GH16	8.78%	beta-glucan; cellulosic biomass degradation, plant and bacterial cell wall remodeling, energy metabolism and pathogen defense	GH16	7.62%	beta-glucan; cellulosic biomass degradation, plant and bacterial cell wall remodeling, energy metabolism and pathogen defense
5	GH73	3.39%	bacterial peptidoglycan; cell wall	GH102	3.41%	peptidoglycan lytic transglycosylase

6	GH4	2.85%	broad glucose degradation, not common in phytoplankton bloom	GH73	2.94%	bacterial peptidoglycan; cell wall
7	GH74	2.62%	beta-1,4-linkages of glucans	GH4	2.78%	broad glucose degradation, not common in phytoplankton bloom
8	GH102	2.08%	peptidoglycan lytic transglycosylase	GH74	1.83%	beta-1,4-linkages of glucans
9	GH114	1.54%	degradation of polygalactosamine	GH114	1.43%	degradation of polygalactosamine
10	GH117	1.31%	Coencode with GH2 in polysaccharide utilization loci	GH24	1.11%	found in GH2 PULS

4. Discussion

Understanding how microbial communities respond to phytodetritus inputs in coastal marine sediments has important implications for global carbon cycling. To investigate the active microbial community at a functional level, mRNA was extracted for metatranscriptomic analysis. A specific interest in this study was to explore difference in gene expression between US and LS and investigate expression of glycoside hydrolases.

4.1. US and LS metatranscriptome comparison

No significant differences were detected between the US and LS when comparing two independent annotations of total differential gene expression (Pfam) and glycoside hydrolase gene expression. Unlike pelagic marine sediments, upper shelf marine sediments tend to be permeable, subjected to strong hydrodynamic forces (waves and tides), and be below shallow water columns (Huettel, Berg et al. 2014). These phenomena provide consistent input of bottom water throughout upper sediment layers. Concurrently, mechanical energy from tides, waves, and bottom currents mix the upper sediment, particularly in the upper 20 cm (Bennett, Hulbert et al. 2002).

Based on the sampling site's sediment permeability (8.48×10^{-11} m^2) and sediment depth (0-6 cm) in light of figure 2 from Huettel and colleagues (2014), the main bottom water transport mechanism is advection. Furthermore, the sampling site's location adjacent to the German Island Helgoland with an 8-meter water column depth warrants strong wave and tidal energy on the upper sediment layers. Due to this, sediment layers here are likely resuspended regularly.

Similar gene expression in US and LS may be a result of similar environmental conditions. If sampling site permeability was comparable between the US and LS before the samples were taken, then advection could saturate both sampling depths with compositionally consistent bottom water. Similar compositional water column input would warrant similar gene expression. Although some substrate would be assimilated or remineralized in the upper sediment layer, the lack of resistance by the sediment on the pore water flow may not effectively biofilter substrates in the space between US and LS. Additionally, sampling location in the *marine sub-littoral zone* (ENVO:01000126) warrants strong sediment disturbances due to localized hydrological forces resulting in resuspension. This may homogenize upper sediment microbial communities regularly. Overall, similar sample metatranscriptomes may have been caused by the combination of similar advective bottom water flow and sediment mixing.

Pfam annotation showed high expression of photosynthesis in both US and LS. Additionally, the top two most abundant BP annotations were "photosynthesis, light reaction" (18.14% US, 16.77 % LS) and "photosynthetic electron transport in photosystem II" (16.61% US, 15.55% LS). A requirement for the synthesis of chlorophyll is the co-factor Mg^{2+}. This is reflected in the second largest MF term cluster being "Mg^{2+} binding". Photosynthesis gene expression has context in the US (0-1.5cm) but little light can reach the LS (5-6 cm). At such shallow water depths, micro algae and cyanobacteria contribute to the benthic microbial community (Huettel, Berg et al. 2014). Sufficient light reaches the sediment surface, where phototrophs profit from both the sun light and nutrients released from the sediment. Although, light penetration into the sediment is observed, it usually does not reach deeper than few millimeter (Kuhl, Lassen et al. 1994). One possible explanation for expression levels of photosynthesis genes in the LS may be the migration of benthic diatoms (Kuhl, Lassen et al. 1994). Upon stress, they can migrate down to 12 cm (Kingston 1999). Detected mRNA levels related to photosynthesis could therefore be a historical artifact.

4.2. Expressed glycoside hydrolases

Glycoside hydrolases (GHs) are CAZymes used in the initiation of microbial remineralization of complex phytodetritus carbon polymers that are transported to sediments. GHs are either endo- or exo-acting enzymes that cleave glycosidic bonds in carbohydrate polymers. To resolve differences in GH gene expression between US and LS, the metatranscriptomic data sets were annotated against the dbCAN reference database. GH counts did not differ between US and LS. In fact, the top 4 GH counts were the same for both US and LS (GH109, GH 23, GH 3, GH 16). This is interesting because GHs were expected to play more of a functional role in phytodetritus degradation in the US but not the LS, and in turn resolve gene expression between the depths. Yet, the results further demonstrate continued similarity in gene expression.

Of all expressed GH families in US and LS, 29% were related to bacterial peptidoglycan degradation with the top expressed being GH23 (24.40% US, 26.03% LS) (www.cazy.com). The next top expressed included GH families 73, 102, 25. A possible explanation for this is the up-regulation of genes related to cell growth and cell division driven by high organic carbon input. GH3 was found in high relative abundance in both sediment layers (12.16%, 16.03% of total GHs respectively). The GH3 family contains alpha and ß-glucosidase that could represent ß-glucosidase activities measured in permeable surface sediments (Böer, Arnosti et al. 2009).

GH annotations were compared to GH gene frequencies in bacterioplankton metagenomes from the same sampling site (Teeling et. al, 2016) to explore similarities of GH content between the water column and sediments at Helgoland Roads. Although the Teeling and colleagues (2016) investigation was based on metatranscriptomics and proteomics, comparisons can still be drawn regarding genetic content between sites. One similarity between the two studies was some of the most abundant GHs found in water column metagenomes were also highly expressed in both sediment depth layers (GH23, GH73, GH102, GH109). One explanation for this could be the shallow water column. In 8 meters of depth, primary production products are not being entirely degraded in the microbial loop because it is so shallow. Thus, microbial communities in the upper sediment could have access to the same carbohydrates as water column microbes. GH23 and GH73 are both related to hydrolyses of bacterial peptidoglycan. This is an important function cell wall repair and cell division. A possible explanation for them being found in both water column and benthic microbial communities is that they both have access

to the same high concentrations of phytodetritus polysaccharides; thus, they may up regulate these GHs for cell wall recycling and division due to r-strategist lifestyles. Seeding of water column microbes into surface sediments can be excluded as an explanation, since the microbial community in North Sea surface sediments have been shown to be significantly different to the water column (Probandt, Knittel et al. 2017).

5. Future directions

There are many future directions that may be taken in this investigation. One issue that should be addressed is that less than 20% of all ORF predicted mRNA reads were annotated. One way to tackle this problem would be to use different sequencing technologies to increase sequence length (i.e. Illumina MiSeq). The samples in this investigation were sequenced on Illumina HiSeq platform which resulted in shorter reads, only spanning genes partially thus decreasing the chance of successful annotation. Annotation tools using models or patterns to identify genes are more sensitive, precise and accurate with longer reads. Additionally, using different sequencing technology could allow for deeper sequencing. Functional genes that differential express metagenomes are in relatively low abundance. Housekeeping genes can easily overshadow their expression due to upregulation.

Moreover, having metagenomic data from the same site would allow mapping of annotated reads back to bins revealing more accurate expression levels. This will confirm sample gene expression taxonomically and proportionalize expression levels with gene counts. If metagenomic data is not available, reads could be mapped against annotated mRNA reads, analyzed for depth and coverage, and compared to original annotation counts to make conclusions.

Another way to increase read annotation is to assemble reads via a *de novo* metatranscriptome assembler. This can create mRNA contigs which increase the sensitivity of annotation tools due to more genetic information available to annotate (Grabherr, Haas et al. 2011). Lastly, another way to increase percent of reads annotated is use bioinformatic tools with improved sensitivity and precision such as HMM-GRASPx (Zhong, Edlund et al. 2016).

Despite experimental result limitations based on limited differential expression and functional genes being dominated by photosynthesis, cell cycle, and housekeeping genes, it was shown that gene expression at this sampling site is very similar between US and LS. This shows that advective flow and sediment resuspension cause the upper *sandy sediment* (ENVO:01000118) levels of *marine sub-littoral zone* (ENVO:01000126) to harbor microbial communities with similar metatranscriptomes.

Appendix

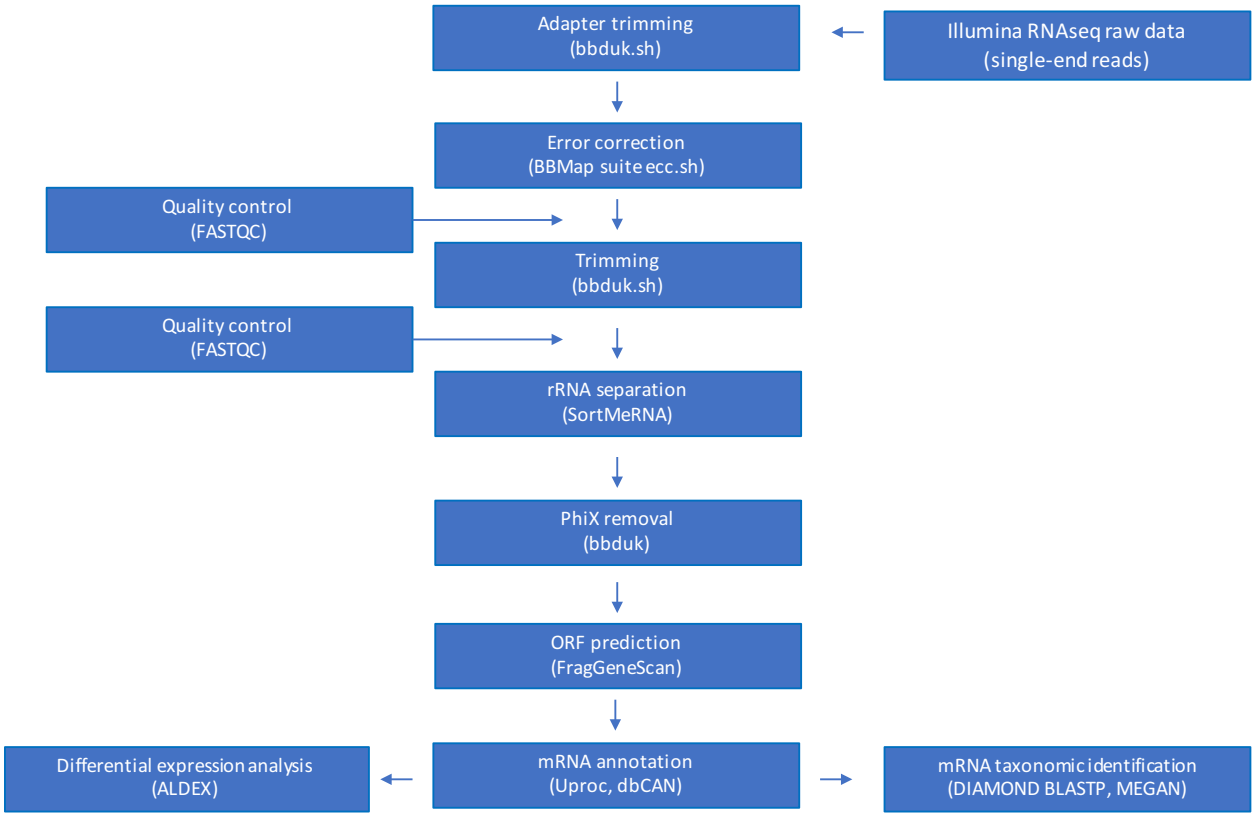

Fig. 2: Sequence processing pipeline from raw reads to functional annotation and statistical analysis.

Table 9: Pfam annotations that contributed most to differentially expressing US and LS from ALDEx2.

Pfam module	Description	Expected p value*	US % abundance	LS % abundance
Pfam11645	PD-(D/E)XK endonuclease	9.09E-03	0.015%	0.005%
Pfam04565	RNA polymerase Rpb2, domain 3	9.75E-03	0.026%	0.014%
Pfam00623	RNA polymerase Rpb1, domain 2	1.32E-02	0.047%	0.026%
Pfam13710	ACT domain	1.84E-02	0.008%	0.004%
Pfam04561	RNA polymerase Rpb2, domain 2	2.44E-02	0.047%	0.022%
Pfam14720	NiFe/NiFeSe hydrogenase small subunit C-terminal	2.48E-02	0.003%	0.003%
Pfam02170	PAZ domain	2.58E-02	0.002%	0.007%
Pfam10116	Protein required for attachment to host cells	3.18E-02	0.002%	0.005%
Pfam10116	Nickel-dependent hydrogenase	3.59E-02	0.044%	0.100%

Table 11: ENVO terms

Term	Number	Purl
marine sub-littoral zone	ENVO:01000126	http://purl.obolibrary.org/obo/ENVO_01000126
sandy sediment	ENVO:01000118	http://purl.obolibrary.org/obo/ENVO_01000118

Metatrancriptome.sh Pipline:

#----------

```
# Metatranscriptome pipeline for labrotation1_David
#
# wd: bigmem-* /scratch/mschecht
#
# Pipeline:
# 1. tar unzip cDNA files
# 2. gunzip cDNA files
# 3. FASTQC then analyze *.fastqc.html output
# 4. bbmap ecc.sh
# 5. FASTQC then analyze *.corrected.fastqc.html output
# 6. OPTIONAL trimming with bbduk.sh
# 7. FASTQC then analyze *.corrected.trimed.fastqc.html output
# 8. sortmeRNA
# 9. mRNA ORF prediction with fraggenescan
# 10. ORF annotation with UProcc
#*save ALL files to /bioinf/home/mschecht/labrotation1_David EVERYDAY*
#-------------------------------------------------------------------------------

#----------------------------
#step1: tar unzip cDNA files
#----------------------------

# *.fastq.tar.gz -> *.fastq.gz

tar -zxvf David_RNA-5-7_Run3.tar.gz

#-------------------------
#step2: gunzip cDNA files
#-------------------------

# *.fastq.gz -> *.fastq

#make directory "files_fastq_gz"
mkdir files_fastq_gz
#copy *.fastq.gz to files_fastq_gz directory
cp *.fastq.gz files_fastq_gz
#change directory to files_fastq_gz
cd files_fastq_gz
#run parallel "--dry-run" to see what the program will execute
time parallel --dry-run --eta -j 4 gunzip ::: *.fastq.gz

#if the dryrun went well then run w/o "--dry-run"
parallel --eta -j 4 gunzip ::: *.fastq.gz

#-------------------------------------------------
#step3: FASTQC then analyze *.fastqc.html output
#-------------------------------------------------

#.fastq -> .fastq.html

#run parallel "--dry-run" to see what the program will execute
time parallel --dry-run --eta -j 4 fastqc ::: *.fastq

#if the dryrun went well then run w/o "--dry-run"
time parallel --eta -j 4 fastqc ::: *.fastq

#Analyze fastqc *.html output

firefox *.html

#---------------------------------
#step4: Run bbmap suite ecc_mod.sh
#---------------------------------

#ecc_mod.sh has an up to date file path for the "java" command
#-Xmx is the max memory
#t=#of threads

#run parallel "--dry-run" to see what the program will execute
#time prints time
#progress = computers/CPU cores/max jobs to run
#-j 4 = number of processes in parallel
#eta is info on time
time parallel --dry-run --progress -eta -j 4 /bioinf/home/mschecht/programs/bbmap/ecc_mod.sh -Xmx100g t=5  in={.}.fastq out={.}.corrected.fastq ::: *.fastq

#if the dryrun went well then run w/o "--dry-run"
time parallel --progress -eta -j 4 /bioinf/home/mschecht/programs/bbmap/ecc_mod.sh -Xmx100g t=5  in={.}.fastq out={.}.corrected.fastq ::: *.fastq

#-------------------------------------------------------------
#step5: FASTQC then analyze *.corrected.fastqc.html output
#-------------------------------------------------------------

#repeat step3 and step4 on *.corrected.fastq then compare *.fastq to *.corrected.fastq before and after ecc_mod.sh error correction in step5
#pay attention to "per base sequence content"
#decided to trim or not to trim
```

```
#run parallel "--dry-run" to see what the program will execute
time parallel --dry-run --eta -j 4 fastqc ::: *corrected.fastq

#run parallel w/lo "--dry-run"
time parallel --eta -j 4 fastqc ::: *corrected.fastq

#----------------------------------------
#step6: "Optional" trimming with bbduk.sh
#----------------------------------------

#usage: bbduk.sh in=<input file> out=<output file> ref=<contaminant files>
#Main input. in=stdin.fq will pipe from stdin.
# -forcetrimleft=x : cut from x to beginning of sequence
# -forcetrimright=x : cut from x to end of sequence

#run parallel "--dry-run" to see what the program will execute
#time prints time
#progress = computers/CPU cores/max jobs to run
#-j 4 = number of processes in parallel
#eta is info on time

time parallel --dry-run --progress -eta -j 4 bbduk.sh forcetrimleft=15 forcetrimright=237 -Xmx100g t=5 in={.}.fastq out={.}.trimmed.fastq ::: *.corrected.fastq

#if the dryrun went well then run w/o "--dry-run"
time parallel --progress -eta -j 4 bbduk.sh forcetrimleft=15 forcetrimright=237 -Xmx100g t=5 in={.}.fastq out={.}.trimmed.fastq ::: *.corrected.fastq

#----------------------------------------------------
#step7: FASTQC then analyze *.corrected.trimmed.fastq
#----------------------------------------------------

#run parallel "--dry-run" to see what the program will execute
time parallel --dry-run --eta -j 4 fastqc ::: *corrected.trimmed.fastq

#run parallel w/o "--dry-run"
time parallel --eta -j 4 fastqc ::: *corrected.trimmed.fastq

#------------------------------------------------------------
#step8: use SortMeRna to filter out the rRNA from the mRNA
#------------------------------------------------------------

#usage:   ./sortmerna --ref db.fasta,db.idx --reads file.fa --aligned base_name_output`

#run parallel "--dry-run" to how program will execute
time parallel --dry-run --eta -j 2 /bioinf/software/sortmerna/sortmerna-2.0/bin/sortmerna --ref /bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-arc-16s-
id95.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-arc-16s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-arc-23s-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-arc-23s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-bac-16s-
id90.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-bac-16s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-bac-23s-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-bac-23s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-euk-18s-
id95.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-euk-18s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-euk-28s-
id98.fasta,/scratch/mschecht/databases/indexes/silva-euk-28s-id98.fasta-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/rfam-5s-database-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/rfam-5s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/rfam-5.8s-database-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/rfam-5.8s-db --reads /scratch/mschecht/files_fastq_gz/sortedrnafiles/{.}.fastq --aligned {.}.sorted_rRNA.fastq --sam --other DP_RNA-
07_S4_L002_R1_001.corrected.trimmed.sorted_mRNA.fastq --fastx --log -a 20 -m 24132 -v ::: *.fastq

#run parallel w/o "--dry-run"
time parallel --eta -j 2 /bioinf/software/sortmerna/sortmerna-2.0/bin/sortmerna --ref /bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-arc-16s-
id95.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-arc-16s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-arc-23s-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-arc-23s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-bac-16s-
id90.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-bac-16s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-bac-23s-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-bac-23s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-euk-18s-
id95.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/silva-euk-18s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-euk-28s-
id98.fasta,/scratch/mschecht/databases/indexes/silva-euk-28s-id98.fasta-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/rfam-5s-database-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/rfam-5s-db:/bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/rfam-5.8s-database-
id98.fasta,/bioinf/software/sortmerna/sortmerna-2.0/index/rfam-5.8s-db --reads /scratch/mschecht/files_fastq_gz/sortedrnafiles/{.}.fastq --aligned {.}.sorted_rRNA.fastq --sam --other DP_RNA-
07_S4_L002_R1_001.corrected.trimmed.sorted_mRNA.fastq --fastx --log -a 20 -m 24132 -v ::: *.fastq


#to index a database
#usage:   ./indexdb_rna --ref db.fasta,db.idx
#output is 3 files: db.stats; db.bursttrue_0.dat; db.kmer_0.dat; db.pos_0.data

/bioinf/software/sortmerna/sortmerna-2.0/bin/indexdb_rna --ref /bioinf/software/sortmerna/sortmerna-2.0/rRNA_databases/silva-euk-28s-id98.fasta,silva-euk-28s-id98.fasta-db -v

#----------------------------------------
#step9: ORF prediction with fraggenescan
#----------------------------------------

# usage: /bioinf/software/fraggenescan/fraggenescan-1.19/run_FragGeneScan.pl -genome=[seq_file_name] -out=[output_file_name] -complete=[1 or 0] -train=[train_file_name] (-
thread=[number of thread; default 1])
# [seq_file_name]: full path
# [output_file_name]: output file name including the full path
# [1 or 0]: 1 if complete genomic sequences; 0 if short sequence reads
# [train_file_name]: file name that contains model parameters; files found in the "train" directory; choose training file based on sequencing procedure
# [complete] for complete genomic sequences or short sequence reads without sequencing error
# [num_thread]: # of threads. Default 1.
```

```
/bioinf/software/fraggenescan/fraggenescan-1.19/run_FragGeneScan.pl
    -genome=/scratch/mschecht/RNA_18_trim_cor_mRNA.fastq.fastq
    -out=/scratch/mschecht/RNA_18_trim_cor_mRNA_ORFpredicted.fastq.fastq
    -complete=0 -train=/bioinf/software/fraggenescan/fraggenescan-1.19/train/illumina_10
```

Work Cited

Bauer, J. E., W. J. Cai, P. A. Raymond, T. S. Bianchi, C. S. Hopkinson and P. A. Regnier (2013). "The changing carbon cycle of the coastal ocean." Nature **504**(7478): 61-70.

Böer, S., C. Arnosti, J. Van Beusekom and A. Boetius (2009). "Temporal variations in microbial activities and carbon turnover in subtidal sandy sediments." Biogeosciences **6**(7): 1149-1165.

Buttigieg, P. L., N. Morrison, B. Smith, C. J. Mungall and S. E. Lewis (2013). "The environment ontology: contextualising biological and biomedical entities." Journal of biomedical semantics **4**(1): 43.

Buttigieg, P. L., E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls and C. J. Mungall (2016). "The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation." J Biomed Semantics **7**(1): 57.

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard and B. Henrissat (2009). "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics." Nucleic acids research **37**(suppl 1): D233-D238.

Ehrenhauss, S., U. Witte, F. Janssen and M. Huettel (2004). "Decomposition of diatoms and nutrient dynamics in permeable North Sea sediments." Continental Shelf Research **24**(6): 721-737.

Gangi, A. F. (1985). "Permeability of unconsolidated sands and porous rocks." Journal of Geophysical Research: Solid Earth **90**(B4): 3099-3104.

Gihring, T. M., M. Humphrys, H. J. Mills, M. Huette and J. E. Kostka (2009). "Identification of phytodetritus-degrading microbial communities in sublittoral Gulf of Mexico sands." Limnology and Oceanography **54**(4): 1073-1083.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol **29**(7): 644-652.

Huettel, M., P. Berg and J. E. Kostka (2014). "Benthic exchange and biogeochemical cycling in permeable sediments." Ann Rev Mar Sci **6**: 23-51.

Kingston, M. B. (1999). "Wave effects on the vertical migration of two benthic microalgae: Hantzschia virgata var. intermedia and Euglena proxima." Estuaries and Coasts **22**(1): 81-91.

Kuhl, M., C. Lassen and B. Jørgensen (1994). "Light penetration and light-intensity in sandy marine-sediments measured with irradiance and scalar irradiance fiberoptic microprobes Rid A-1977-2009." Marine Ecology-Progress Series.

Liu, K.-K. (2010). "Carbon and nutrient fluxes in continental margins : a global synthesis."

Probandt, D., K. Knittel, H. E. Tegetmeyer, S. Ahmerkamp, M. Holtappels and R. Amann (2017). "Permeability shapes bacterial communities in sublittoral surface sediments." Environ Microbiol.

Supek, F., M. Bosnjak, N. Skunca and T. Smuc (2011). "REVIGO summarizes and visualizes long lists of gene ontology terms." PLoS One **6**(7): e21800.

Teeling, H., B. M. Fuchs, C. M. Bennke, K. Kruger, M. Chafee, L. Kappelmann, G. Reintjes, J. Waldmann, C. Quast, F. O. Glockner, J. Lucas, A. Wichels, G. Gerdts, K. H. Wiltshire and R. I. Amann (2016). "Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms." Elife **5**: e11888.

Woulds, C., S. Bouillon, G. L. Cowie, E. Drake, J. J. Middelburg and U. Witte (2016). "Patterns of carbon processing at the seafloor: the role of faunal and microbial communities in moderating carbon flows." Biogeosciences **13**(15): 4343-4357.

Zhong, C., A. Edlund, Y. Yang, J. S. McLean and S. Yooseph (2016). "Metagenome and Metatranscriptome Analyses Using Protein Family Profiles." PLoS Comput Biol **12**(7): e1004991.