Predicted microbial secretomes and their target substrates in marine sediment

Orsi, William D.; Richards, Thomas A.; Francis, Warren

Published in:
Nature Microbiology

DOI:
10.1038/s41564-017-0047-9

Publication date:
2018

Document version
Peer reviewed version

Citation for published version (APA):
Predicted microbial secretomes and their target substrates in marine sediment

William D. Orsi 1,2*, Thomas A. Richards3, Warren R. Francis1,4

1. Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, 80333 Munich, Germany.
2. GeoBio-Center LMU, Ludwig-Maximilians-Universität München, 80333 Munich, Germany.
3. Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, EX4 4QD, UK.
4. Present address: Department of Biology, University of Southern Denmark, Odense, Denmark.

Running title: Subseafloor secretomes

Keywords: deep biosphere, metatranscriptome, secretome, extracellular enzymes, carbon cycling

*To whom correspondence should be addressed

Scientific drilling has identified a biosphere in marine sediments1, which contain many uncultivated microbial groups known only by their DNA sequences2-4. Recycling of organic matter in sediments is an important component of biogeochemical cycles because marine sediments are critical for long term carbon storage5. Turnover of carbon is hypothesized to be driven by the secretion of enzymes by microbial organisms5-7, which act to break down macro-molecules into constitutive monomers that can be transported into the cell for use. As such, the nature of the microbial secretome often influences the function of a community6. However, the microbial groups involved in this process and the biochemistry they perform is poorly understood. Here, we show that expressed genes from 5 – 159 meters below the seafloor (mbsf) encode numerous candidate peptidases and carbohydrate-active enzymes (“CAZymes”)9 targeted for secretion. The majority (90-99%) were assigned to Bacteria, of which 12% shared the highest sequence similarity with candidate phyla10,11. The remaining putatively secreted proteins shared highest sequence similarity with archael and fungal enzymes, which peak in two redox transition zones12. In the shallower redox zone at 30 mbsf, 20% of the transcripts encoding putative secreted peptidases were assigned to lineages7,13,14 of uncultivated archaea. The target compounds of the predicted secreted proteome show a preference for necromass in the form of microbial cell envelopes as well as plankton and algal detritus. The predicted fungal secreted proteome encodes CAZymes not present in the predicted bacterial or archael secreted proteomes, indicating that fungi putatively play a minimal but specialized role in subseafloor carbohydrate recycling.

Many microbes acquire nutrients and fixed carbon by secreting digestive enzymes (here defined for simplicity as the “secretome”) and transporting the digested material back to the cell. A major form of protein-mediated transfer across cell membranes is determined by the N-terminal signal peptides of the encoded protein, which signals the export of the protein through the cellular secretion (Sec) pathway. Signal
peptides have conserved amino acid sequence motifs that can be identified from predicted proteins using bioinformatic algorithms.

We detected 290,810 open reading frames (ORFs) present in previously sequenced metatranscriptomes from subseafloor sediments and scanned them for the presence of signal peptides using SignalP. These samples derive from anoxic subseafloor sediments of the Peru Margin (Ocean Drilling Program Leg 201, Site 1229D, water depth 151 m) and display two sulfate methane transition zones (SMTZs) exhibited by contrasting down core profiles of pore water sulfate and methane concentrations at 30 and 90 mbsf. Published metatranscriptomic sequences from these environments identified core metabolic activities from Bacteria, Archaea, and Fungi, (amino acid metabolism, lipid and carbohydrate metabolism) but did not focus on putatively secreted proteins or the predicted substrates of candidate secreted enzymes. We reanalyzed these data using all three variants of SignalP (version 4.1) which preferentially identifies Gram positive, Gram negative and eukaryotic secreted proteins and collectively identified 7,802 non-redundant metatranscriptome ORFs (3% of total ORFs) that contain a putative signal peptide (Table S1), indicating active microbial secreted proteins are present in the environments sampled. Searching these ORFs against the CAZy and SEED databases, updated with all predicted proteins from recently described “microbial dark matter” genomes and fungal genomes (see Methods), indicated that expressed genes from bacteria, archaea, and fungi encoded putatively secreted enzymes. To our knowledge this is the first analysis of predicted signal peptides in a marine sediment metatranscriptome, and provides insights into microbial mechanisms of organic matter recycling in anoxic subseafloor sediments.

The expressed ORFs putatively encoding secreted proteins branch throughout diverse clades across the tree of life (Figure 1). CAZymes, peptidases, and substrate transporters all putatively encoding a N terminal secretion signal peptide were found to be expressed by bacteria typically found in anoxic marine sediments such as: Alphaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, Bacteroidetes, Actinobacteria, Chloroflexi, and Firmicutes. Many expressed ORFs encoding candidate secreted proteins demonstrated significant levels of sequence similarity to CAZymes and also had highest level of sequence similarity to predicted proteins encoded by organisms corresponding to phylum level groups: Omnitrophica (OP3), Atribacteria (OP9 and JS1), Parcubacteria, and Microgenomates. Such ORFs had highest relative expression in the shallow SMTZ at 30 mbsf, reaching 20% of total transcripts of the ORFs encoding proteins putatively targeted for secretion (Figure 2A). Whilst the exact function of these proteins needs to biochemically determined, the presence of a N-terminal signal peptide and shared sequence similarity with CAZy enzyme classes indicates that many of the ORFs identified encode secreted proteins that function to bind and/or degrade assorted carbohydrates.
The secreted proteome detected with our metatranscriptomic approach is likely to be an underestimate for several reasons; *de novo* assembled transcriptomes are inherently incomplete, whereupon low-coverage transcripts may be fragmented, allowing for only prediction of partial ORFs. Thus, true N-terminal signal peptides may be missing from some fragmented transcripts. We note that comparison of read mapping to the metatranscriptome contigs (Figure S1) demonstrated that the 5’ ends of ORFs demonstrated consistently low levels of read coverage, likely because reads at the end of transcripts are curtailed because they cannot map outside the contig boundaries. This means that such metatranscriptome analysis systematically underestimates N [and 3’/C] terminal signal peptides. Additionally, SignalP does not have specific optimised settings for the Archaea, thus our analysis may be underestimating the true extent of the secreted proteome with specific biases against certain taxonomic groups. Furthermore, some proteins may be secreted via cellular mechanism that do not require N-terminal peptide. Collectively, these points suggest such approaches consistently under-estimate secreted protein diversity and function. We also note that all signal peptides identified with SignalP may not necessarily be secreted out of the cell\textsuperscript{15}, and thus we refer through the manuscript as transcripts encoding ORFs with signal peptides as putatively secreted.

The total relative abundance of expressed bacterial ORFs encoding proteins putatively targeted for secretion increases with sediment depth (Figure 2A). From the same set of bacterial ORFs, sequences with similarity to CAZy protein classes display a significant linear decrease with sediment depth ($r = -0.8$; least squares regression: $P = 0.02$; Figure 2B). This likely reflects hydrolysis rates of extracellular enzymes that are highest in the surface or near surface sediments where there is an influx of ‘fresh’ organic matter, and then decrease with sediment depth as diagenesis leads to an increased proportion of recalcitrant organic matter over time\textsuperscript{5,6}. Similarly, the relative abundance of expressed ORFs with highest sequence similarity to bacterial peptidases putatively targeted for secretion also decreased with depth, but their relative abundance was 8 - 20 times lower than CAZymes at all depths (Figure 2B).

Consistent with the significant decrease in CAZyme expression (Figure 2B), the relative abundance of transcripts that encode putatively secreted bacterial ABC transport systems for sugar substrates decreases exponentially with depth (Figure 2C), which follows a power law function ($R^2 = 0.66$, Figure 3). An exponential decrease in the relative expression of transporters for amino acids and peptides was also observed, which also follows a power law function with a near perfect fit ($R^2 = 0.95$; Figure 3). This is likely due to the decreasing availability of bioavailable amino acids with sediment depth at this site\textsuperscript{20}. Transcripts encoding signal peptides and which are annotated as ABC transporters for benzoate increased markedly in the deepest sample (Figure 2C). Benzoate is an important substrate resulting from the anaerobic decomposition of aromatic hydrocarbons in marine sediments\textsuperscript{21} and degradation of the benzoate could be associated with methane production\textsuperscript{22}. 
The most prevalent CAZy protein classes putatively targeted for secretion were glycoside hydrolase (GH) group 5 (cellulose family A, responsible for degrading cellulose⁹), carbohydrate esterase (CE) group 6 (carboxylesterase type B), CE1 (corresponding to beta-lactamases), and carbohydrate binding module (CBM) 50 (Figure 4). Most of these were assigned to bacterial groups (Figure 4). CBM50 is a protein domain of ca. 50 amino acid residues found attached to enzymes from CAZy enzyme classes GH18, GH19 (which were also expressed and targeted for secretion in the sediment samples: Figure 4) and are known to cleave chitin and peptidoglycan⁹. These results suggest that bacteria actively secrete CAZymes for targeted extracellular binding and/or degradation of carbohydrates, in particular cellulose, chitin and peptidoglycan, in deep anoxic sediments. Indeed, ‘plant’ or algal detritus and cell wall or cell surface layer components are the majority of the organic matter sources targeted by the CAZymes in our samples (Figure 4). In deep subseafloor sediments, adsorption of biopolymers to sediment particles and mineral grains would likely reduce their accessibility⁵. Rather, the abundance of CAZymes targeting cell wall necromass including peptidoglycan (Figure 4) suggests that partially intact cells and their cell envelopes, many of which consist of peptidoglycan²³, are preferred targets of the secreted proteins detected.

Expressed ORFs encoding proteins putatively targeted for secretion and assigned to the Archaea exhibited highest relative abundance at 30 mbsf (Figure 2A), recruiting 20% of reads mapping to ORFs putatively encoding secreted peptidases (Figure 2B). These putative peptidases all had highest sequence similarity to predicted peptidases from Bathyarchaeota¹³ and Thermoplasmatales (MBG-D)⁷ genomes (Figure 1). This supports prior genomic and biogeochemical observations suggesting that these Archaea influence protein cycling in sediments through secretion of extracellular peptidases⁷. Archaeal ORFs encoding proteins putatively targeted for secretion with sequence similarity to a cellulosomal glycoside hydrolase precursor GH48⁹ in the Hadesarchaea¹⁴ and Bathyarchaeota¹³ draft genomes were also detected at 30 mbsf (Figure 4). This CAZy class (GH48) was not detected among the ORFs assigned to Bacteria or Fungi. Bathyarchaeota genomes indicate a capacity for degradation of extracellular carbohydrates in sediments²⁴, which is consistent with their expression of ORFs with high sequence similarity to extracellular cellulosomes: complexes of multicellulolytic enzymes that break down cellulose and hemicellulose polysaccharides²⁵. Expressed ORFs encoding proteins putatively targeted for secretion were also assigned to Hadesarchaea, these included a putative polysaccharide lyase 1 (PL1), a gene family that includes pectate and pectin lyases (Figure 4), which can degrade complex pectin polysaccharides⁹. ORFs expressed by Hadesarchaea also shared sequence similarity to multiple CAZy carbohydrate binding modules (CBM5, 32, 50, 6, and 9), whereas Bathyarchaeota ORFs encoding proteins putatively targeted for secretion were classified as CAZy enzyme class CBM6.

The relative abundance of expressed fungal ORFs encoding enzymes targeted for secretion peak within the two SMTZs (Figure 2A), which suggests that active fungal populations may play a role in
recycling of organic matter at these geochemical interfaces. Indeed, the fungal transcripts encoding CAZymes targeted for secretion at these depths, and other depths (Figure 2B), indicate that subseafloor fungi are active in the extracellular degradation of carbohydrates. Interestingly, two putatively fungal proteins (contig_65210 & contig_2308 - Table S2) demonstrated high levels of protein similarity to Hypocreales database sequences, consistent with the detection of ribosomal RNA of this group in deep sediment samples26. Our analysis indicates that fungi are involved in the recycling of extracellular organic matter down to 91 mbsf through the production of proteins with sequence similarity to putatively secreted carbohydrate binding modules (CAZy carbohydrate binding modules 43, 18, and 48), and glycoside hydrolases (CAZy enzyme classes GH47, and 17). None of these CAZy enzyme classes were found to be expressed by bacteria or archaea (Figure 4), and thus by secreting these enzymes fungi may play a specialized role in the degradation of specific classes of carbohydrates. For example, enzymes of the CAZy class GH17 are important for the breakdown of β-glucans through the hydrolysis of β-D-glucose that releases α-glucose9. The glucose resulting from this extracellular breakdown of polysaccharides by fungi is likely utilized by bacteria, as putative bacterial transmembrane transport systems for sugars were found to be expressed throughout the sampling core (Figure 2C). Moreover, the sugar residues in archaeal pseudopeptidoglycan are β-(1,3) linked N-acetylglucosamine and N-acetyltalosaminuronic acid27. Thus, it may be possible for fungal CBM43 enzymes to attach to archaeal pseudopeptidoglycan cell envelopes in marine sediments.

Bacterial secretion of glycoside hydrolase group 71 (Figure 4) was detected whose sole representative is α-1,3-glucanase, and enzyme responsible for endohydrolysis of (1,3)-α-D-glucosidic linkages in isolichenin and pseudonigeran, both components of fungal cell walls28. This indicates that some bacteria putatively recycle fungal cell necromass in the anoxic sediments. Our results directly support hypotheses put forth by genome centric studies suggesting a role for candidate phyla in the turnover of recalcitrant sedimentary carbon10,17 through their secretion of carbohydrate active enzymes, and show that fungi should also be considered as participants in this process.

The marine subsurface exhibits a diversity of habitats and provinces, and thus our results from a single drilling site are not sufficient to generalize about universal principles of the subsurface secreted proteome. Nevertheless, the methodological approach and the results described here allow us to identify which microbes are performing which aspects of carbon recycling in deep marine subsurface sediment. Comparisons with additional subsurface metatranscriptomes will help to establish general patterns and processes and identify key microbial players within these important and cryptic environments.
Methods

Assembled contigs deriving from metatranscriptome datasets from Peru Margin sediments' 5, 30, 50, 70, 91, and 159 meters below the seafloor (Ocean Drilling Program Leg 201, Site 1229D; 77° 57.4590’ W, 10° 58.5721’ S) were searched for ORFs using FragGeneScan\textsuperscript{29}. ORFs were searched using contigs resulting from the original assemblies performed by Orsi \textit{et al.}\textsuperscript{8}. For additional details regarding sampling, contamination controls, metatranscriptome sequencing, and \textit{de novo} assembly please see Orsi \textit{et al.}\textsuperscript{8}. Relative expression of ORFs was calculated by mapping read counts to ORFs after normalizing against ORF length (reads per kilobase mapped: RPKM). Only those ORFs with a minimum average coverage >5 were considered for downstream analysis. To account for uneven sequencing depth between the six metatranscriptomes, RPKM values were normalized (rarified) to the sample with the lowest number of annotated sequences (n = 48 million reads).

ORFs were then sequentially searched for signal peptides using SignalP 4.1\textsuperscript{15} with the default settings for eukaryotes, gram negative bacteria, and gram positive bacteria in turn. Most ORFs containing signal peptides were detected with all settings (Table S1), indicating applying SignalP to mixed communities of microbes is in itself not able to differentiate between transcripts deriving from gram negative bacteria, gram positive bacteria, eukaryotes, or potentially Archaea. Prior to downstream analysis, duplicate signal peptide containing ORFs detected with multiple settings were removed. To further annotate the taxonomic affiliation and putative function of the non-redundant ORFs found to contain a signal peptide (n=7,802), we searched these ORFs for related proteins using BLASTp (using BLOSUM62 matrix) against an updated SEED database (www.theseed.org) containing all predicted proteins from the high quality draft genomes and single cell genomes from recently described “microbial dark matter”\textsuperscript{7,10,11,13,14,17-19} that were downloaded from the NCBI protein database. Our database also included all fungal genomes from the NCBI RefSeq database. The total number of predicted proteins in the updated database was 37.8 million. Our criteria for assigning homologous ORFs with BLASTp were bit score greater than 50, minimum alignment length of 30 amino acids, minimum sequence similarity of 30%.

ORFs were assigned to high level taxa (phylum level, or class level for Proteobacteria) using the phylum or class level affiliation of the top hit in the genome database. ORFs encoding enzymes targeted for secretion with sequence similarity to CAZYmes were identified through BLASTp searches against the CAZY database (www.cazy.org) using the same parameters and homology criteria described above. Annotation of peptidases come from those available in the SEED database, and for “dark matter” genomes we used the NCBI annotation provided in the publicly available data. Signal peptide containing ORFs annotated as chaperone proteases (\textit{e.g.}, HtrA proteases\textsuperscript{30}) that assist in membrane insertion of other proteins were not here regarded as part of the “secretome” that would be directly acting on extracellular detrital compounds, and were excluded from further analysis. On the basis of BLASTp identified sequence
similarities we cannot assign highly resolved (e.g., genus level) phylogenetic affiliations to our ORFs. Rather, the phyla or class (in the case of Proteobacteria) affiliation of the top hits were used to group ORFs into higher level taxonomic groups (Figure 4).

A new python script was written, to create a matrix summarizing the two BLASTp searches (e.g., against the updated genome database, and the CAZY database). This matrix contained summed RPKM counts per gene, per taxon, per sample for the BLASTp search against the genome database, which was then amended with CAZYme class homology acquired from the BLASTp against the CAZY database. The python script is publicly available at bitbucket.org/wrf. All statistical analyses were conducted in R (www.r-project.org). The phylogenetic tree displayed in Figure 1 was acquired from the Supplemental Data of Hug et al¹⁹, with modified coloring. Colors in the tree indicate higher level taxonomic groups where predicted proteins in representative genomes were found to be homologous with expressed ORFs encoding enzymes targeted for secretion in the metatranscriptomes.

Data availability: Data is available as raw sequence reads through the NCBI Short Read Archive (SRA) under accession number SRA058813 and as assembled contigs in MG RAST (metagenomics.anl.gov) under accession numbers 4515478.3, 4515477.3, 4515476.3, 4510337.3, 4515336.3, 4515335.3. Fasta files containing the expressed ORFs with signal peptides are available from the corresponding author upon request.

References:


Figure legends:

Figure 1. Classes of expressed ORFs encoding putatively secreted enzymes in subseafloor metatranscriptomes across the tree of life. Tree is modified from that of Hug et al and is colored to show the microbial groups where subseafloor transcripts were assigned, that were found to encode ORFs (n=890) containing putative signal peptides from various enzyme and protein classes (see legend). The colors of the branches are only used to indicate which portion of the tree corresponds to the different phylum and class level groups (e.g., all highlighted branches do not have putative sequences in the secreted protein dataset), colors of the text in the taxonomic name match the colors in the tree.

Figure 2. Relative abundance of expressed ORFs encoding putatively secreted proteins. (A) The percent of reads per sample (after rarefaction) mapping to bacterial (n=3,695 transcripts), archaeal (n=98 transcripts), or fungal (n=94 transcripts) transcripts in subseafloor metatranscriptomes that encode ORFs with a putative secretion signal peptide. (B) The percent of reads per sample targeted for secretion (e.g., the “secretome”) mapping to transcripts encoding ORFs (n=613) with sequence similarity to either CAZymes or peptidases. (C) The percent of reads per sample targeted for secretion (e.g., the “secretome”) mapping to transcripts encoding ORFs with sequence similarity to transmembrane transport proteins (n=424 ORFs). Note the different scale in the axes for Bacteria, Archaea, and Fungi. In the bacterial panel, the black portion of the histogram corresponds to reads assigned to recently discovered microbial “dark matter” groups (Parcubacteria, Microgenomates, Atribacteria, and Omnitrophica candidate phyla) whose ecological functions are not fully understood. TRAP: tripartite ATP-independent periplasmic transporters, TonB: TonB-dependent transporters. SMTZ: sulfate methane transition zone.
Figure 3. Distributions of expressed ABC-transporters for sugars and amino acids/peptides. The top x-axis represents the percentage of reads mapping to transcripts containing a putatively secreted sugar (n=55 ORFs) or amino acid/peptide transporter (n=130 ORFs), out of all reads mapping to signal peptide containing ORFs (n=7,802 ORFs). The gray shading indicates the sulfate methane transition zones was observed, with higher cell counts. Cell count data is re-plotted from D’Hondt et al., 2004.

Figure 4. Distribution of CAZy protein classes with sequence similarity to putatively secreted enzymes. The colors in the heatmap correspond to the number of reads (after rarefaction) mapping to transcripts encoding ORFs (n=548) with sequence similarity to proteins belonging to various CAZy protein classes (right hand side of panel). Colored circles represent potential organic matter sources targeted by different CAZy enzyme classes, ‘animal detritus’ refers to compounds other than chitin (e.g., keratan sulfate). GH: glycoside hydrolase, CBM: carbohydrate binding module, PL: polysaccharide lyase, GT: glycosyl transferase, CE: carbohydrate esterase.

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through Project OR 417/1-1 (W.D.O.), and also by Ludwig-Maximilians Universität München Junior Researcher Fund (W.D.O.). T.A.R. is supported by a Royal Society University Research Fellowship. We thank Dr. Oliver Voigt (LMU Munich) for his help in installing the stand alone (command line) implementation of SignalP.

Author contributions W.D.O and T.A.R conceived the idea for the study, W.D.O. wrote the paper, W.D.O and W.F. analyzed data, W.F. developed analytical tools. All authors participated in data interpretation and provided editorial comments on the manuscript.
Gammaproteobacteria
CAZYmes
Proteases, peptidases
Substrate transporters
Glycosyltransferases
Polysaccharide lyases
Carbohydrate esterases
Carbohydrate binding modules
Glycoside hydrolases

Alphaproteobacteria
Epsilonproteobacteria
Deltaproteobacteria
Omnitrophica (OP3)
Planctomycetes
Bacteroidetes
Actinobacteria
Chloroflexi
Firmicutes

Encoded protein groups targeted for secretion in metatranscriptomes
Hadesarchaea
Nanohaloarchaea
Bathyarchaeota (MCG)
Thermoplasmata (MBG-D)
Atribacteria (OP9, JS1)
Latescibacteria
Bacteroidetes
Omnitrophica (OP3)
Epsilontrobia
Deltaproteobacteria
Alphaproteobacteria

Microgenomates

Eukaryotes
Ascomycota
Basidiomycota
Bathyarchaeota (MCG)
Thermoplasmata (MBG-D)
Atribacteria (OP9, JS1)
Latescibacteria
Bacteroidetes
Omnitrophica (OP3)
Epsilontrobia
Deltaproteobacteria
Alphaproteobacteria

Archaea

Hadesarchaea
Nanohaloarchaea
Bathyarchaeota (MCG)
Thermoplasmata (MBG-D)
Atribacteria (OP9, JS1)
Latescibacteria
Bacteroidetes
Omnitrophica (OP3)
Epsilontrobia
Deltaproteobacteria
Alphaproteobacteria

Microgenomates