Michael Hoffert

Bioinformatics and Genomics Final Paper

3.14.17

Metagenomic SNV profile analysis of selective pressures on a North Pacific marine microbial population

Abstract

The composition and evolution of marine microbial communities is due in part to interactions between populations with specific metabolic and adaptive requirements and environmental features of the oceanic region. The TARA Oceans project represents a collection of metagenomic and physiochemical data from marine environments around the world. Analysis of the SNV profiles of specific populations in the TARA datasets allows us to quantify the presence, strength, and source of selective pressures on these microbial populations. We hypothesized that the combination of nitrate and oxygen deficiency in the North Pacific deep chlorophyll maximum layer would allow sulfur-oxidizing bacteria to flourish and diversify alongside photoautotrophs. We assembled SNV profiles of genomes extracted from metagenomic data and attempted to classify sources of variation in the profiles according to site-specific metadata. Despite observed lack of sulfur and nitrogen oxidizing microbes, we discovered the presence of genetically variable Prochlorococcus populations coexisting as the result of a clonal bloom. The source of this bloom remains unidentified and merits further investigation into this region.

Metagenomic SNV profile analysis of selective pressures on marine microbial populations

Introduction

Oceanic microbial communities are collections of microbes that represent a large number of metabolic and taxonomic groups. A combination of environmental features, nutrient availability and depth create unique conditions for particular microbial groups to flourish or decline in specific marine environments. As a result, the genetic composition and diversity of any marine microbial community develops specific to its oceanic region of origin. The TARA Oceans program seeks to catalogue the morphological, ecological, and molecular diversity of the world's oceans by collecting and extracting metagenomic DNA sequence data for communities of bacteria, archaea, plankton, viruses and eukaryotic organisms at a variety of depths and locations. The program also collected physiochemical data to characterize the environmental contexts of specific marine microbiomes. These data allow for the analysis of the evolutionary and ecological developments of marine communities in a genetic context through the coupling of metadata and metagenomic analysis. We sought to determine if the selective pressures exerted by environmental factors on marine microbe populations could be detected and quantified based on metagenomic analysis of the populations.

A common method of analyzing the evolution movement of genomes are single nucleotide variants (SNVs) [1]. SNVs are DNA polymorphisms introduced to DNA by mutation or replication errors. The number and abundance of specific SNVs in a sequence are used to quantify the selective pressure on the sequence: lower numbers of SNVs or high ratios of a particular SNV indicate that selective pressures prevent the introduction of further variance to the organism's DNA. Analysis of SNVs provides a method of determining the presence and nature of selective pressures on oceanic microbe genomes from the TARA dataset when paired with physiochemical metadata. In order to generate SNV profiles, we mapped sequenced reads from available TARA datasets to a DNA assembly from the deep chlorophyll maximum (DCM) layer of the anoxic, nitrate-limited North Pacific sample site and constructed SNV profiles for extracted genomes in order quantify selective pressures on the extracted genomes. Assignment of KEGG Orthologies to the contigs in the binned genomes was conducted to determine the identity of contigs that experience differential selective pressure and correlate the functional definitions of orthologues and associated contigs with ecological conditions in the DCM. We predict nitrate and oxygen deficiency in the North Pacific DCM will increase selective pressure on microbes that use nitrogen and oxygen as electron receptors, creating profiles with few, highly conserved SNVs. We also expect an increase in the abundance and variation of SNVs in contigs related to sulfur oxidizing pathways, as selective pressure is decreased on sulfur-oxidizing microbes with the overall decrease in competitive ability of other forms of metabolism in this zone. Because the layer is a DCM, we expect abundance of photoautotrophic organisms.

Methods

Assembly and mapping of metagenomic data – Sample collection and DNA extraction/sequencing were performed according to the procedure in Sunagawa et. al 2015 [9]. Metadata for the TARA datasets was found on the EBI metagenomics website [3]. 12 million unpaired reads were randomly recruited from the North Pacific deep chlorophyll maximum (DCM) layer TARA dataset (ERR598995) and assembled with IDBA-UD [7]. Assembly quality was analyzed quast.py and the completed assembly. The reads used for the assembly were mapped against the resulting assembly using bowtie2 and samtools [5][6]. Additional mappings of 1 million reads from mesopelagic, surface and deep chlorophyll maximum layers for Southern Ocean (Antarctic), South Pacific, North Atlantic (Portuguese coast), Arabian Sea, and North

Atlantic (East Coast) datasets were mapped to the large North Pacific assembly using bowtie2 and samtools.

Binning metagenomic mappings on North Pacific assembly - A contig database was constructed from the large North Pacific DCM assembly for binning. Gene calls for the contig database were generated using centrifuge. The resulting taxonomy and previously generated mappings of other sample sites to the North Pacific DCM assembly were merged by anvi'o with a minimum contig length of 1000. Binning of the mapped datasets was conducted with anvi'o [2] based on tetranucleotide frequency, GC content, and completeness of universal single-copy gene sets. SNV profiles for identified bins were generated with anvi'o. Bin contigs were annotated according to KEGG Orthology using the KEGG Automatic Annotation Server (KAAS) with BLAST search and nucleotide entries selected. The representative prokaryotic KO set was assigned with single-directional best hits (SBHs). Open reading frame (ORF) calls on the large North Pacific assembly were performed with Prodigal [4]. Interproscan [8] was used for subsequent annotation of the ORFs.

Results

Characterization of collected metadata - The North Pacific deep chlorophyll maximum layer lies 115 meters below the surface of the ocean, with water temperatures of 15.3°C and salinity at 34.4 psu. Despite being somewhat colder (15.3°C vs. ~25°C) than comparably nitrogen-limited DCMs in the Arabian Sea and South Pacific, the North Pacific DCM contains similar amounts of oxygen and chlorophyll to these layers [3]. Although the North Pacific is characterized as an anoxic zone, the North Pacific DCM has the highest dissolved oxygen content of any North Pacific layer at 225 µmol per kilogram. This value is comparable to 211.6 µmol/Kg and 179.9µmol/Kg for the Arabian Sea and South Pacific DCMs, which come from sample locations in anoxic zones. It is characterized by limited concentrations of nitrate at 0.732507 µmol per liter.

Assembly and mapping – The North Pacific DCM was assembled from 12 million randomly sampled unpaired reads from the dataset with IDBA-UD. The resulting assembly was composed of 8496 contigs, with a largest contig length of 96,742 base pairs and an N50 of 1380 bp. The GC content of the assembly was 42.11%. Mappings of other datasets to the North Pacific DCM assembly recruited between 0.5% and 5% of the total reads for each dataset. Based on visualizations generated post-mapping by anvi'o, mapping of the datasets to the North Pacific DCM resulted in low coverage over the majority of the assembly. The lowest read recruitment during mapping occurred from the Arabian Sea mesopelagic layer, which is characterized as extremely oxygen and nitrate limited. Mapping of the North Pacific DCM to itself resulted in average coverage between 50 and 5000 for contigs greater than 1000 base pairs in length.

Binning and phylogeny assignment – Although mappings generally recruited low percentages of reads, binning was conducted with anvi'o using all of the mapped data. Phylogeny assignment with centrifuge resulted in lower numbers of assigned phylogenies for the contigs in the North Pacific DCM database. Binning of the mapped datasets was conducted using the anvi-interactive visual interface and based on the merged and sorted mappings by anvi'o. Coverage appeared low in most regions of the assembly. 6 bins were extracted from the dataset. Bin 1 was taxonomically categorized as the genus Prochlorococcus, a group of extremely small abundant cyanobacteria. Although some taxonomic assignments in Bin 6 were also Prochlorococcus, a majority taxonomic assignment was not found. The average length of the bins was 476,649 base pairs and average N50 for the bins was 5960. Bin 6 was notably longer, larger, and more complete than the other bins, with a length of 1,358,095 bp, an N50 of 7035 bp, and percent completeness of 79.6% and redundancy of 4%. Only two other bins, Bin 1 and Bin

2, had a completeness greater than 1% at 15.1% and 5% respectively. Bin 6 also had a high GC concentration at 56% compared to 35-40% for the other bins in the dataset.

SNV profiles and variability analysis – The compiled SNV profiles for the bins were composed entirely or nearly entirely of reads from the North Pacific DCM dataset. Variability profiles between the bins were not consistent, but construction of an average SNV conservation vs. abundance plot from the ratio of most common nucleotide to second most common nucleotide and number of SNVs showed that bins 1 and 6 – both notable for high completeness – had SNV profiles that demonstrated differing conservation and generation of SNVs. Bin 6 had a high average n2n1 at approximately 0.7 for all SNVs, and 0.4 for SNVs with a length greater than 50. Bin 1 was among the lowest in terms of average n2n1, but contained a high number of SNVs at 42 per kilobase of contig for all SNVs in the profile.

KEGG Orthology assignment – After noting the differences between bins 1 and 6 in the previous analysis of the SNV profiles, we attempted to characterize the contents of the bins in order to determine the potential genetic origins of the SNVs and relate them to an ecological context. The KEGG Automatic Annotation Server was used to assign KEGG Orthologies to the contigs in each bin. The results of the annotation showed similarities in the top ten assigned orthologues between Bins 1, 2 and 6. The most frequently assigned orthologues were generally metabolic in nature. Large numbers of amino acid metabolism, cellular respiration and oxidative phosphorylation orthologues were observed. Few photosynthetic or chemolithotrophic orthologues were assigned in the bins. These observations agreed with simple searches of the Interproscan results for ORFs in the North Pacific DCM assembly, which had no matches for proteins related to sulfur or nitrogen-related metabolisms.

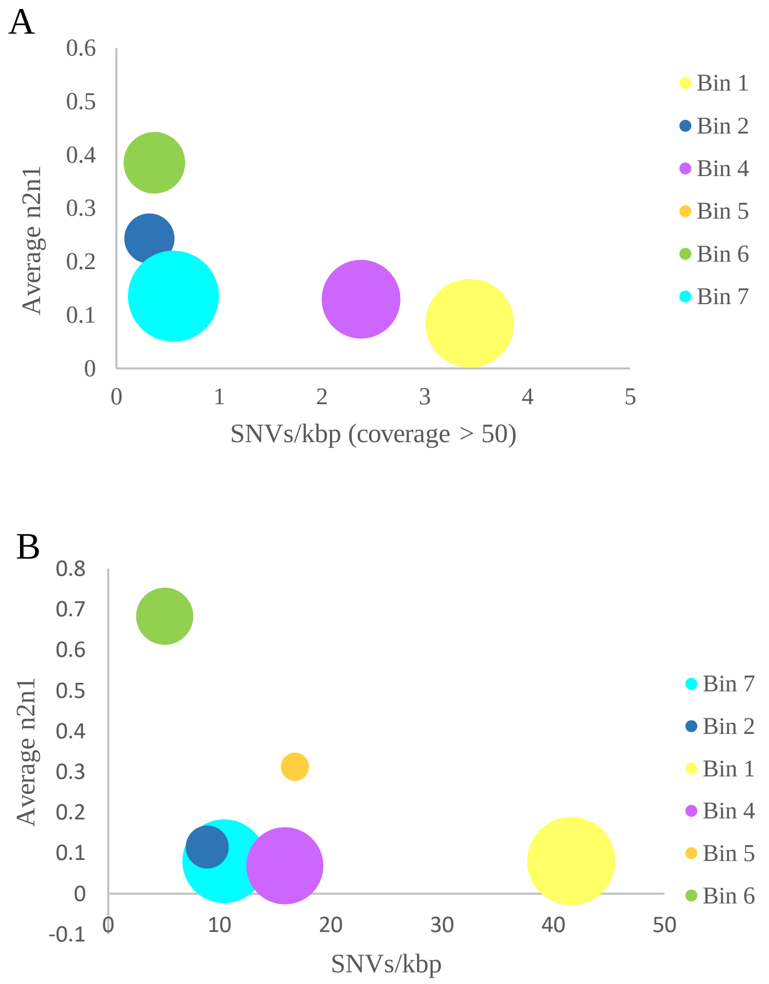

Figure 1 – SNV conservation and abundance in bins Bubble diameter corresponds to average coverage for considered SNVs for each bin. Bin 5 contained no SNVs with coverage greater than 50. (A) Number of SNVs per kilobase vs. average n2n1 for each bin extracted from metagenomic data mapped to North Pacific deep chlorophyll maximum layer assembly, excluding SNVs with coverage less than 50. (B) Number of SNVs per kilobase vs. average n2n1 for each bin extracted from metagenomic data mapped to North Pacific deep chlorophyll maximum layer assembly.

Table 1 – Summary of metagenomic bin data for 6 bins extracted from mappings of TARA sample data to North Pacific deep chlorophyll maximum assembly. Colors correspond to bubble colors from Figure 1.

Bin Name	Assigned taxonomy	Length (bp)	Number of contigs	N50	GC content	Percent completeness	Percent redundancy
Bin 1	Prochlorococcus	217772	148	1428	40.48755758	15.10791367	0.71942446
Bin 2	None	196188	151	1251	35.18957903	5.035971223	0
Bin 4	None	396029	82	14617	38.86879288	0	0
Bin 5	None	43314	28	1423	38.10620291	0.617283951	0
Bin 6	None	1358095	298	7035	53.91809224	79.62962963	4.32098765
Bin 7	None	648500	148	10006	37.59358465	0.71942446	0
Average	-	476649.67	142.50	5960.00	40.69	16.85	0.84

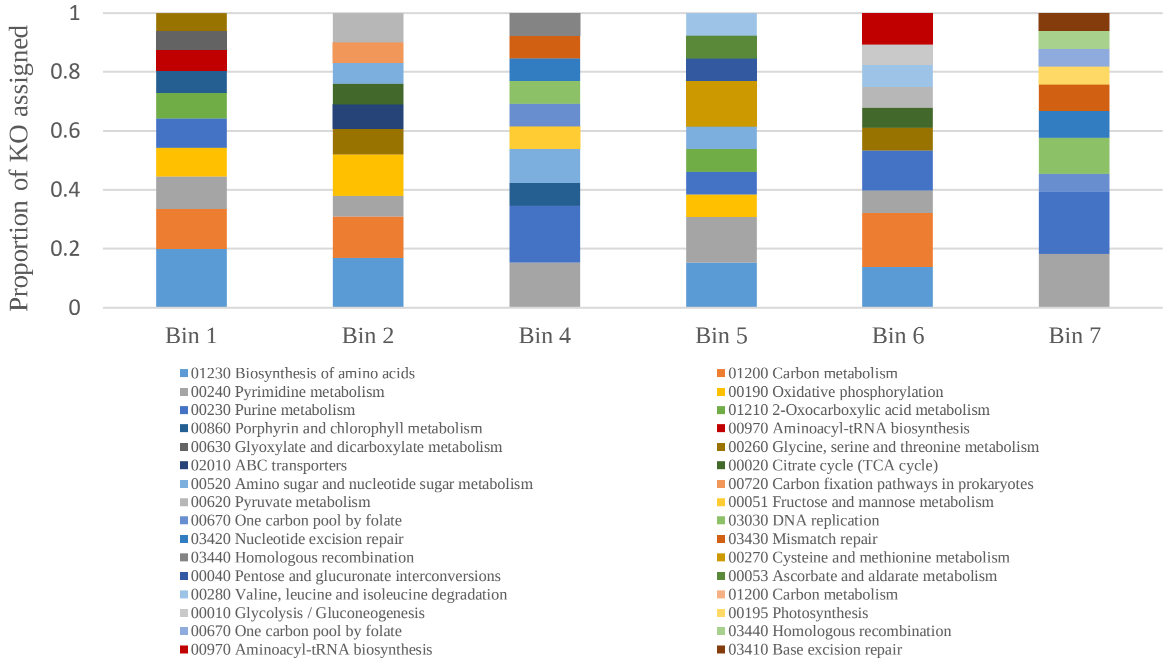

Figure 2 – Top 10 KEGG Orthology assignments for each bin based on the number of assigned contigs to each ortholog. Orthologs are colored identically across plots.

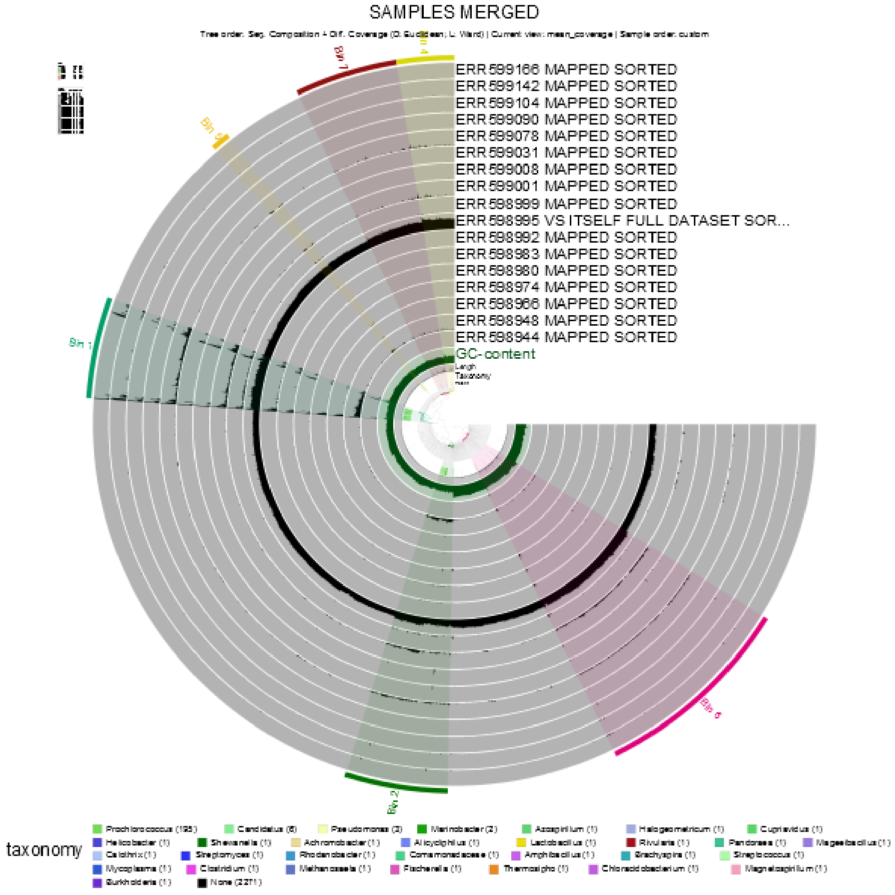

Figure 3 – Anvi'o [2] visualization of datasets mapped to North Pacific DCM for assembly. Bins are labeled and colored to show included contigs.

Discussion

Because the North Pacific DCM is both oxygen and nitrogen limited, we expected an increase in the proportion of sulfur-oxidizing microbes, as this form of metabolism is competitive in environments that lack both nitrate and oxygen as sources of potential electron acceptors. We did not observe any abundance of sulfur or nitrogen oxidizing organisms in the data. Although the identified presence of Prochlorococcus cyanobacteria supports our prediction, the hypothesis regarding the dominance of sulfur-oxidizing bacteria in the nutrient-limited zone is not supported. Both Interproscan and KAAS did not identify assembled sequences that represented sulfur or nitrogen oxidizing genes. Instead, there was an abundance of photosynthetic proteins in the Interproscan analysis and a large number of oxygen-based metabolic sequences in the KAAS results (Figure 2). This maybe be explained by the abilities of photoautotrophs to produce and use oxygen as an electron acceptor in the presence of light, which is present in the deep chlorophyll maximum. Additionally, amino acid metabolism was a commonly observed KO assignment. Because the North Pacific DCM is proximal to a nitrogen-rich mesopelagic zone [3], it is possible that organisms in the DCM use amino acids that diffuse upward from the nitrogen-rich mesopelagic layer as a source of nitrogen, explaining the lack of nitrogen-oxidizing genes in the data.

The taxonomic results of the binning were generally inconclusive with the exception of Bin 1, which was assigned the genus Prochlorococcus, a common oceanic cyanobacteria (Table 1). Although Bin 6 was not assigned a taxonomy, Prochlorococcus was identified in some of the contigs in the bin (Figure 3). Bin 2, the third most complete bin, was also observed to have a majority of Prochlorococcus (Figure 3). These results agree with metadata for the site [3], which indicates medium to high concentrations of chlorophyll, as well as Interproscan and KO assignments that identified photosynthetic and cellular respiration-related genes in the dataset (Figure 2).

The SNV profiles of the bins displayed low numbers of SNVs and low abundance of specific SNVs for bins 1, 2, 4, and 5, although Bin 4 appeared somewhat variable when all SNVs were considered. Because these bins had very low completeness and no assigned taxonomy, it is unlikely that they accurately represent microbial populations in the DCM. However, bins 1 and 6 had higher completeness and stronger assignment of phylogenies. Despite the homogeneity of the contigs, bins 1 and 6 displayed distinctly different patterns in their SNV profiles. Bin 6 exhibits high abundance of specific SNVs, but lower number of SNVs overall. Bin 1 is characterized by nearly 42 SNVs/kbp, an extremely high rate of variation compared to the rest of the bins in the set. Bin 1 also exhibits high variation within SNVs. These data indicate that the Prochlorococcus population in Bin 1 is under weak selective pressure, as variation is not eliminated from the population. Because Bin 6 is so complete relative to other bins and has a very high average n2n1, we conclude that this bin arose from a clonal bloom of organisms with a specific SNV profile. The resulting abundance of identical genomes led to high coverage and excellent assemblies of this region as well as a mostly complete bin, displayed by the extreme length of Bin 6 contigs and its very high completeness of 79.6% (Table 1). Additionally, the clonal bloom maintained the specific SNVs present in the source of the bloom, leading to the observed conservation of the SNV profile.

Although the difference between Bin 1 and Bin 6 may be explained by the presence of a clonal bloom, the possible taxonomic assignment to the Prochlorococcus genus in Bin 6 raises questions regarding the selective conditions under which a specific sub-set of the variable Prochlorococcus population from Bin 1 would be selectively allowed to explosively reproduce. Unfortunately, the current binning homogeneity of the genes represented in the assembly prevent us from specifically identifying unique features of the Bin 6 population, but its presence alongside the Prochlorococcus in Bin 1 implies more complex underlying interactions between the environmental features of the North Pacific deep chlorophyll maximum and its constituent microbial populations.

The results gathered in this experiment were largely hindered by poor mapping of other sample sites to the North Pacific DCM. Although 12 million reads were used for the assembly of the DCM, it is possible that mapping more than 1 million reads from other datasets to the assembly would create better coverage and population distinction while binning. Additionally, investigation into the diversity and composition of the North Pacific DCM populations could inform us as to why mappings onto this dataset are poorer than expected. In general, this would improve the reliability of the data, which was low in this study, and imply future directions for determination of selective environmental features. A method for contrasting the composition of various bins would also have been in precisely determining the sources of variation between the bins.

In conclusion, the North Pacific deep chlorophyll maximum layer did not display expected population abundance or SNV variation in sulfur-oxidizing microbe populations because none were identified in metagenomic data. However, binned genomes from populations of Prochlorococcus cyanobacteria displayed SNV profiles that indicated the coexistence of populations with differentially conserved SNV profiles, contradictory to our understandings of the effects of selective pressure. Poor mapping and homogenously represented genes hindered identification of the driving force behind the selective pressure, but it is possible that observed SNV patterns are due to complex interactions between the DCM and proximal layers, or unknown features of the oceanic region.

References

1. Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. Isme Journal. 2016;10(7):1589-601. doi: 10.1038/ismej.2015.241. PubMed PMID: WOS:000378292100005.

2. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. PeerJ Inc.; 2015;3: e1319. doi:10.7717/peerj.1319

3. EBI metagenomics
InterPro EMBL-EBI -
https://www.ebi.ac.uk/metagenomics/projects/ERP001736/samples/ERS494208/runs/ERR599166/results/versions/2.0

4. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–9. doi:10.1038/nmeth.1923

6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–9. doi:10.1093/bioinformatics/btp352

7. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. Oxford University Press; 2012;28: 1420–1428. doi:10.1093/bioinformatics/bts174

8. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res. 2005;33: W116- 20. doi:10.1093/nar/gki442

9. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348(6237). doi: ARTN 126135910.1126/science.1261359. PubMed PMID: WOS:000354877900031.