Deep Sequencing of a Coastal Bacterioplankton Metatranscriptome: How sequence coverage affects biogeochemical interpretation.

Scott Gifford, Shalabh Sharma, and Mary Ann Moran

Department of Marine Sciences, University of Georgia

Metatranscriptomics, the direct retrieval and sequencing of environmental RNA, has emerged as a powerful tool to identify genes actively being expressed by microbial communities. The technique has now been applied to a variety of systems, including coastal, open ocean, and terrestrial microbial communities. The dominant transcripts recovered from these previous analyses were involved with basic cellular machinery, such as protein and ribosomal synthesis. Due to the high transcript diversity of these communities, it is quite possible that we are missing important biogeochemically relevant transcripts due to a low sequence coverage which only catches the most abundant transcripts. In this study, we examined how deeply a metatranscriptome must be sequenced to catch important biogeochemical functions. Transcripts isolated from two replicate samples of a marine microbial community off the coast of Georgia were sequenced with four 454 FLX pyrosequencing runs, producing over 2 million reads (one million per replicate sample). After removal of rRNA, the sequences were BLASTed against NCBI's refseq and COG databases to identify the transcript's best hit to a taxon, gene, and function. We found that the majority (60%) of refseq hits were covered by only one transcript, and that the two replicate samples shared only ~35% of their refseq hits. Rarefaction analysis of the refseq hits showed that four plates were far from reaching saturation, and that 3 plates (~1.5 million reads) were needed to capture at least 75% of the total refseq diversity. On the other hand, both the taxa and COG function rarefaction curves indicated that the discovery of new taxa and functions had leveled off, and that 75% of their total diversity could be covered with just one plate (~500,000 reads). An examination of transcripts involved in the phosphorus cycle showed that the majority of phosphorus pathways would be discovered with a quarter of our full coverage. These initial results indicate that while our current sequencing levels do not come close to covering the entire transcript pool or identifying the majority of taxa specific functions, they are able to identify the majority of non-redundant functions and active taxa, as well as major components of biogeochemically important pathways.