Consiglio Nazionale delle Ricerche
ISMAR - Istituto di Scienze Marine
Sede Arsenale Tesa 104 - Castello 2737/F – 30122 Venezia, Italy
Tel +39 041 2407927 Fax +39 0412407940
direttore@ismar.cnr.it
C.F. 80054330586 - P.IVA 02118311006

PROGRAMMA DI RICERCA STM 2015

"*Metabolism and functioning of Mediterranean Sea microbial communities through the use of metatranscriptomics*"

Relazione sull'attività svolta

Struttura ospitante: Instituto Milenio de Oceanografía (IMO), Universidad de Concepción, Concepción, Cile

In riferimento a quanto indicato nella domanda di partecipazione al programma STM2015, gli obiettivi del presente progetto riguardavano l'analisi di dati metagenomici e metatrascrittomici di campioni di acqua raccolti nel Mar Mediterraneo Occidentale profondo, al fine di comprendere il significato ecologico delle comunità microbiche e le relazioni fra funzioni microbiche e le principali variabili ambientali.

Durante le tre settimane di permanenza presso l'IMO, sono state apprese ed applicate le fasi fondamentali ed i *tool* bioinformatici necessari per l'analisi di sequenze metagenomiche (MG) e metatrascrittomiche (MT). Brevemente, tali fasi sono state: il *prefiltering* delle sequenze per l'identificazione e la rimozione delle sequenze di scarsa qualità, l'identificazione tassonomica delle sequenze codificanti per l'RNA ribosomale, la ricerca e l'analisi dei geni codificanti per proteine, l'*assembly* ed i principali punti di analisi dei campioni metatrascrittomici.

La concentrazione di RNA estratto da alcuni campioni selezionati è risultata non sufficiente a garantire un livello di profondità e di qualità di sequenziamento necessarie per le successive analisi in programma. Pertanto, durante la permanenza presso i laboratori dell'*Instituto Milenio de Oceanografia*, le attività e le analisi previste dal presente progetto sono state realizzate su campioni diversi da quelli inizialmente previsti.

Il primo step delle analisi ha riguardato il *pre-filtering* delle sequenze (Figura 1). Tale step consiste nella ricerca e rimozione di sequenze o porzioni di sequenze che non rispondono ai criteri di qualità necessari per le fasi successive dell'analisi. I software utilizzati in tale fase sono indicati in Figura 1. Successivamente, è stata effettuata la *prediction* delle sequenze di RNA ribosomale attraverso il tool *metaxa* (Bengtsson et al., 2011, *Antonie van Leeuwenhoek Journal of Microbiology,* 100:471-475). Circa il 30% delle sequenze totali di RNA del MT sono risultate essere di tipo ribosomale e pertanto sono state separate dal resto delle sequenze attraverso l'utilizzo di uno script PERL, realizzato dal team di bioinformatici del gruppo di ricerca del Prof. Ulloa. Le sequenze di rRNA provenienti dal MG sono state identificate utilizzando il tool di allineamento *blastn* "contro" il database di RNA ribosomale SILVA (http://www.arb-silva.de/). Per le sequenze MT, è stato effettuato il cosiddetto *fragment recruitment* con il software FR-HIT (Beifang et al., 2011, *Bioinformatics*, 27:1704-1705) utilizzando il genoma di un *Thaumarchaeota*, precedentemente sequenziato, come genoma di riferimento.

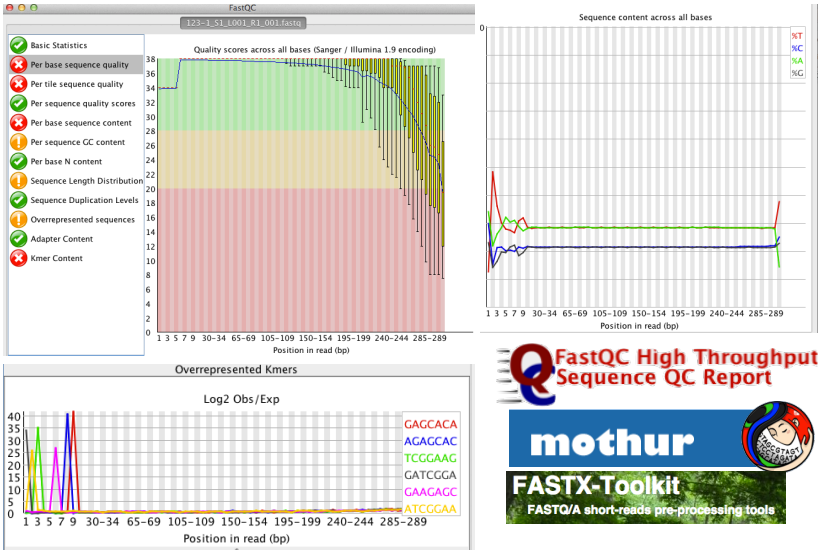

Figura 1. Analisi di qualità delle sequenze analizzate con FastQC. La figura indica inoltre alcuni degli altri software utilizzati per il *trimming* e il *filtering* delle sequenze (o porzioni di sequenza) di bassa qualità (i.e., Mothur (http://www.mothur.org/), FastX (http://hannonlab.cshl.edu/fastx_toolkit/)).

Per l'analisi dei campioni MG, le sequenze di rRNA sono state rimosse con uno script incluso nella pipeline QIIME (Caporaso et al., 2010, *Nature Methods*. 7:335–336). Prima di procedere con la ricerca dei *protein coding genes*, è stata creata una versione "*blastable*" del database KEGG (http://www.genome.jp/kegg/) con il software USEARCH (http://www.drive5.com/usearch/). La ricerca per le sequenze riconosciute come codificanti per proteine è stata effettuata con USEARCH.

Uno step importante dell'analisi di dati MG è il cosiddetto *assembly* ("assemblaggio") di genomi multipli a partire da sequenze corte (*reads*) provenienti dai singoli genomi di una comunità microbica. L'*assembly* delle sequenze metagenomiche è stato realizzato con il software Metavelvet (http://metavelvet.dna.bio.keio.ac.jp/) e, per i campioni analizzati, non ha comportato un significativo miglioramento della lunghezza dei frammenti. Infine, sebbene in maniera preliminare, è stato possibile conoscere ed utilizzare un potente strumento online di condivisione di dati metagenomici e metatrascrittomici, il IMG-ER (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi), realizzato dal Joint Genome Insitute (JGI), che consente di esaminare e confrontare il propri dati con un elevato numero di metagenomi depositati in precedenza e pubblici.

Sebbene il lavoro di analisi pianificato ed iniziato grazie al soggiorno di studio presso l'IMO necessiterà di ulteriori approfondimenti, l'obiettivo principale di questo progetto (l'apprendimento dei principali *tools* bioinformatici e *step* per l'analisi di dati MG e MT, e la loro applicazione ai fini della comprensione del funzionamento di comunità microbiche marine) è stato raggiunto con successo. Le competenze acquisite, decisamente innovative e "di frontiera" nell'ambito dell'oceanografica microbica, costituiranno un investimento per l'Istituto, anche in considerazione della disponibilità della Struttura Ospitante a stabilire collaborazioni future con l'ISMAR CNR.

Venezia, 22 Dicembre 2015