ANÁLISI DE IRREGULARIDADES EN NIVELES DE EXPRESIÓN GÉNICA UTILIZANDO SECUENCIACIÓN DE NUEVA GENERACIÓN ILLUMINA

ALEJANDRO SEPÚLVEDA GUTIÉRREZ
INGENIERO EN BIOINFORMÁTICA

RESUMEN

A partir del proyecto Fondecyt N11090234, se secuenció el Transcriptoma de Nannochloropsis salina para encontrar genes diferencialmente expresados en dos condiciones biotecnológicamente importantes: alta concentración de CO2 y alta intensidad de luz. Para esto, primero se generó una librería de expresión normalizada con ambas condiciones, que fue secuenciada por Roche-454 creando un Transcriptoma de Referencia. Luego, dos librerías sin normalizar fueron secuenciadas con Illumina y estas secuencias se utilizaron para mejorar lo obtenido desde 454 y para alinearlas sobre el Transcriptoma de Referencia, midiendo así la expresión génica. De esta forma, se observa la cantidad de genes diferencialmente expresados en las condiciones mencionadas previamente.

Sin embargo, al momento de alinear las lecturas IIlumina sobre la referencia, no se observó una distribución homogénea de las lecturas. Se observó que muchas secuencias, mapeaban de forma no homogénea y no sólo en la zona codificante de los contigs. Además, el mapeo mostró pics particulares en distintas partes del contig. Para evaluar las razones e implicancias de este comportamiento anómalo, en este trabajo se utilizaron diferentes herramientas bioinformáticas, informáticas y matemáticas, las cuales se enmarcaron en Biológicas y Artefactos del Secuenciador utilizado, de forma mutuamente excluyente. Una vez encontradas las causas de las mencionadas anomalías, se procedió a determinar en qué medida afectan a la Expresión Génica y Diferencial de N. salina. En consecuencia, se logra determinar que las causas de los pics son lecturas Illumina duplicadas durante el proceso de amplificación de las mismas, dichas anomalías, además, no afectan de forma significativa los niveles de expresión del organismo en estudio.

ABSTRACT

Fondecyt project N 11090234 sequenced the transcriptome of Nannochloropsis salina in order to find differentially expressed genes in two biotechnologically important conditions: high concentration of CO2 and high light intensity. With this aim, we generated a normalized expression library with both conditions, which was then sequenced using Roche-454 technology to create a Reference Transcriptome. Then, two un-normalized libraries were sequenced using Illumina technology and these sequences were used to improve those obtained from 454 and also were aligned on the reference transcriptome, in this manner measuring gene expression. As a result, we find the number of genes differentially expressed in the above conditions.

However, when Illumina reads were aligned to the reference, not observed a homogeneous distribution of reads, many of them mapped not only into the coding region of unigenes. In addition, the mapping showed specific peaks in different parts of the unigene. To evaluate the reasons and implications of this anomalous behavior, in this study we used different bioinformatics, computing and mathematics tools, which are framed in Biological and Sequencer Artifacts used, mutually exclusive. Once found the causes of the above anomalies, we proceeded to determine the extent affect Differential and Gene Expression for N. salina.

Accordingly, it is able to determine the causes of the peaks are Illumina duplicate readings during the amplification process thereof, such anomalies also not significantly affect expression levels of the organism under study.