Pipeline development for analysis of RNAseq data from single cells and cell populations from in vitro differentiation of human pluripotent stem cells

Robert Morey
University of California San Diego
Department of Reproductive Medicine
La Jolla, CA, 92037

Rathi D Thiagarajan
University of California San Diego
Department of Reproductive Medicine
La Jolla, CA, 92037

Louise Laurent
University of California San Diego
Department of Reproductive Medicine
La Jolla, CA, 92037

ABSTRACT

High-throughput next-generation sequencing of mRNA transcripts (RNAseq) has become the method of choice for determining the genome-wide mRNA expression levels in populations of cells and at the single cell level. RNAseq allows for the simultaneous discovery of novel transcripts and splice variants, allele specific expression analysis, and post-transcriptional base modifications at high resolution with low background noise. Moreover, recent developments in sequencing library construction have made it possible to interrogate the transcriptomes of single cells, allowing researchers to discern expression differences between seemingly homogenous cell populations. The capability of RNAseq to provide a comprehensive gene-expression profile has quickly made this technology a ubiquitous and crucial research tool in almost every field in the life sciences and is now being implemented for clinical use.

In spite of the utility and popularity of RNAseq, processing and analyzing the large datasets generated by this technology remains time consuming and unwieldy for those who wish to extend their analysis beyond a small set of standard tools, or to have the flexibility to select among several options at each step in the analysis process. Given the size of the datasets, large compute clusters such as "Trestles" available from XSEDE's San Diego Supercomputer Center (SDSC), are required to perform quality control, pre-processing, mapping, and secondary analysis of large datasets. Multiple software components from different sources are often needed to transform raw sequencing data into biologically relevant data that can be used to compare the transcriptomes of multiple samples. Unlike microarray gene-expression analysis, for which there are several data analysis packages that are accessible to most biologists, the unfamiliarity of the Linux environment to many in the life sciences has required biology-focused laboratories interested in using this technology to depend on core facilities or dedicate one or more lab members to performing these compute intensive tasks and subsequent tertiary data analyses.

We are developing an RNAseq data processing pipeline to be used by our lab's informatics personnel for analysis of RNAseq from single cells and cell populations from in vitro differentiation of human pluripotent stem cells (see Dr. Thiagarajan's poster for a description of the results of this analysis). We have recognized that this division of labor between the informatics-focused personnel who process and analyze the sequencing data and the "wet-lab" personnel who generate the biological samples and perform the downstream functional validation experiments can create both inefficiencies and gaps in communication. To bridge this gap, we hope to collaborate with XSEDE to help create a platform that integrates and centralizes the many different programs and scripts that are currently available on Trestles. This tool would allow life science researchers, without Linux experience, to leverage XSEDE's compute resources to process and analyze RNAseq data.