June 2015

Hands-on: De-novo transcriptome building

Esti Feldmesser

Introduction

In this workshop we will inspect how a transcriptome rebuilding looks like, how to evaluate its quality and how it is mapped to a reference genome. The data set in this workshop is a collection of mRNA sequences extracted from leukemia cell lines.

The RNA-seq reads have been mapped to the genome using Tophat. In a separate pipeline, reads were *de-novo* assembled to create transcripts using Trinity alone. Overall the Trinity assembly started from ~316 million paired-end reads from 4 samples and generated 35,568 contigs (transcripts) in 34,166 loci.

The obtained transcripts were mapped to the genome using Blat. The results are in a text file in psl format. For an explanation on the PSL format go to http://genome.ucsc.edu/FAQ/FAQformat.html and look for PSL.

Instructions

1. Visualize all the data together using IGV

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

We will need to download several files to the PC in order to see them in the browser.

a. The data is found on a public disk. To access the required file press the Windows button and type the following \\ngs001\Open_Data\Course2015- Assembly exercise (you can also follow the link).
b. Copy to your PC the file called Trinity.hg19.psl
c. Copy to your PC also the four files with the tdf suffx.
   A tiled data file (TDF) file (.tdf) is a binary file that contains data that has been preprocessed for faster display in IGV. In this case the files contain the read coverage of the RNA-seq samples

d. Open the IGV tool found on your desktop.

e. Allow the application IGV and/or Java to run when prompted
f. After the IGV window opens, look at the pool down menu at the top, left side and be sure that the browser shows the Human hg19 genome since the reads and transcripts were mapped to this genome version.
g. Right click on the left side at "RefSeq Genes", and then on "Expanded"
h. Go to File, Load from File… and load the psl file. Click on "Go" when prompted
i. Be sure the psl display is "Expanded"
j. Go to File, Load from File… and load the tdf files
k. You can change the range of display for the tdf files according to your needs by right clicking on their name and then on "Set Data Range…"
l. We will look at several differentially expressed genes in the IGV browser. Let's start with the AEBP1 gene. Write the name of the gene in the upper part of the IGV browser, to the left of the "Go" and click on "Go".
   Was the gene built correctly by Trinity? Can you understand why? Look for alternative isoforms built by Trinity, look at their nomenclature and try to understand how it works. Transcripts in introns of other genes are not alternative spliced isoforms.
m. Look for the ADCY6 gene. How was it built?
n. How do you think that Trinity results can be improved?