Exercise 7

Dena Leshkowitz & Gil Stelzer

Analyzing transcript assembly of RNA-Seq data

Introduction

In this exercise section we will run and evaluate assembly of transcripts using RNA-Seq data and a model genome (mouse). We will use RNA-Seq data from the article Using Synthetic Mouse Spike-In Transcripts to Evaluate RNA-Seq Analysis Tools. PLoS One. 2016 21;11(4). In this study, in vitro synthesized mouse spike-in control transcripts were added to the total RNA of differentiating mouse embryonic bodies, and their expression patterns were measured. This approach enables assessing the accuracy of the assembly tools.

Instructions

We will analyze one sample from the experiment explained above, namely sample C2 which is ES cells from day 0 to which spike-in transcripts were added. The library was created from total RNA (including spike-in mixes) and was processed using the Illumina TruSeq Strand Specific total RNA with RiboZero Gold (for rRNA removal) protocol (Illumina). The source for spike-in was in-vitro transcription (IVT) products of plasmid made from cDNA clones from the Fantom2 mouse cDNABook collection (RIKEN, Japan). The spike-ins were selected based on previous knowledge that they are not expressed in the ES biological system studied.
A list of the spike-ins and the quantities added to this sample are found in course_2019/transcript_analysis/spike_in/journal.pone.0153782.s009.docx (download to open).
The reads were ~50M paired-end 100-bp reads sequences. Were mapped with Tophat and transcripts were assembled using Stringtie and cufflinks programs.

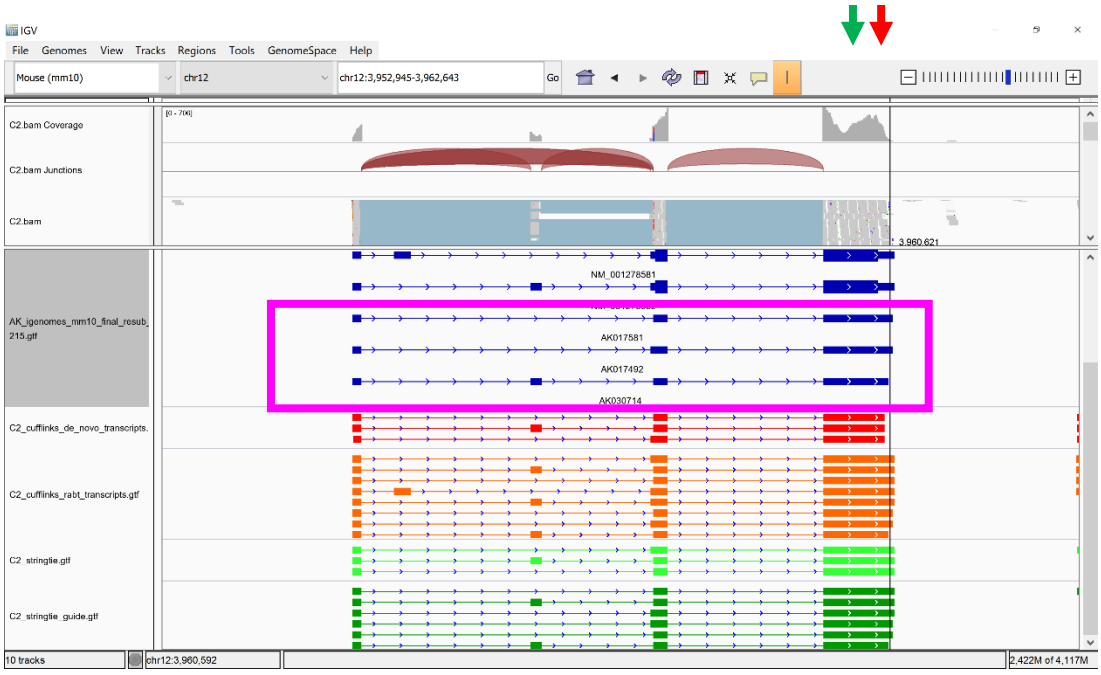
The files you will use in this section -
C2.bam - mapped reads
C2.bam.bai - index of mapped reads
C2_cufflinks_de_novo_transcripts.gtf - transcripts assembled with cufflinks
C2_cufflinks_rabt_transcripts.gtf - transcripts assembled with cufflinks with an annotation file (known transcripts)
C2_stringtie.gtf - transcripts assembled with stringtie
C2_stringtie_guide.gtf - transcripts assembled with stringtie with an annotation file (known transcripts)
AK_igenomes_mm10_final_resub_011215.gtf - refSeq annotation that includes the spike-ins transcripts.

Note - gtf files have special format used to hold information about a gene structure. Look here for details: http://mblab.wustl.edu/GTF22.html)

1. Loading built Cufflinks and Stringtie transcripts
   a. Open IGV viewer as instructed in Exercise 3.
   b. Assuming you mounted your class folder, go to the following folder - course_2019/transcript_analysis/spike_in
   c. Change the genome to mouse mm10
   d. Load all the gtf files and the bam file. If you are asked to index a gtf file, click on the "GO" and on the next popup click "OK".
   e. In order to view all isoforms for the Pomc gene we'll expand the gtf tracks (after typing the Pomc gene in the top selection field). Stand on the gtf file name in the IGV viewer, right click it and select the option "Squished" or "expanded".
   f. Under tracks (top panel) select "fit data to window" this helps to view all the tracks. This will cause all tracks to be displayed in "Squished" mode. In order to see transcript id's select "expanded" for the desired track.
   g. You may change the mouseover behaviour (yellow description pop-up) by clicking the yellow cartoon style speech bubble in the top icon bar (see green arrow in the figure below).
2. Inspecting Cufflinks and Stringtie transcripts
   a. The only Pomc transcripts in this loci are the ones spiked-in, since the endogenous gene is not expressed in this RNA sample. This locus contains two spike-ins that were added - **Table 1** – Concentration of Pomc spike-in transcripts added to the sequenced sample

	Concentration
AK030714	1000.0 attomoles/ul
AK017581	150.0 attomoles/ul
AK017492	150.0 attomoles/ul

   b. In order to view the exon borders and the aligned reads you will need to zoom-in.
   c. Notice that if you stand on with the mouse on a transcript from the cufflinks or stringtie tracks you will see quantification information (coverage, fraction, FPKM, TPM). We will observe these values to evaluate if the spike-ins proportions was kept.
   d. It can be easier to explore the track information if you color the four tracks in different colors. Do this by right click on the track and select "change track color"
   e. The line button (see red arrow) can help you evaluate the exons borders
   f. The spiked-in are shown in the pink box below.



g. In order to evaluate the accuracy of the transcript assemblies by the various tools please use the following criteria -
   I. The number of transcripts assembled
   II. How well the exons agree with the spike-in exons. Take into account differences in exons borders between the spike-ins and the assembled transcripts.

Question 1

Try to evaluate the accuracy of Pomc transcript assemblies by the various tools. Summarize your findings in the following table –

	Spike-in count detected accurately	Additional transcript count detected (FPKM > 0)	Is the proportion between the spiked-in correct (compared to table 1)
C2_cufflinks_de_novo_transcripts.gtf			
C2_cufflinks_rabt_transcripts.gtf			
C2_stringtie.gtf			
C2_stringtie_guide.gtf			

Question 2

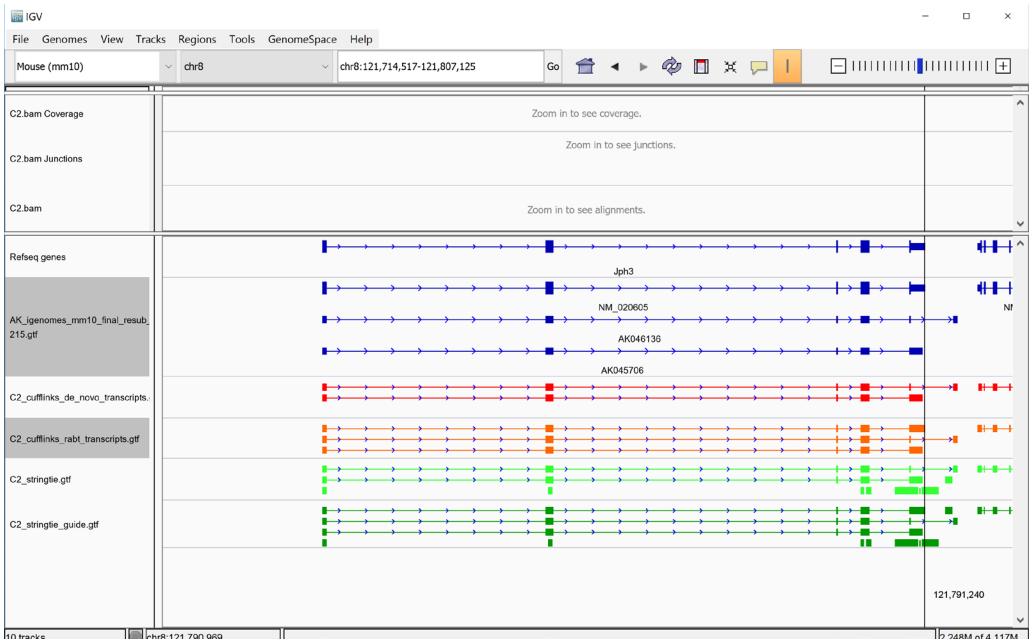
Summarize which program is more accurate?

Question 3

Repeat the above but for Jph3 transcripts –

Table 2 - Concentration of Jph3 spike-in transcripts added to the sequenced sample

	Concentration
AK045706	100.0 attomoles/ul
AK046136	450.0 attomoles/ul



	Spike-in count detected accurately	Additional transcript count detected (FPKM > 0)	Is the proportion between the spiked-in correct (compared to table 2)
C2_cufflinks_de_novo_transcripts.gtf			
C2_cufflinks_rabt_transcripts.gtf			
C2_stringtie.gtf			
C2_stringtie_guide.gtf			

Question 4

Summarize which program is more accurate?

Comparing short and long reads

In this section of the exercise, we will compare the information we get when we align short sequences and transcripts built from the short reads, to long reads produced from cDNA and direct sequencing using the MinION Oxford Nanopore technology.
The RNA was extracted from ES mouse cells – named undiff.
The files may be found in the following folder -
course_2019/transcript_analysis/short_vs_long

The files you will use in this section -

Undiff_cDNA.bam - BAM file of cDNA sequenced with MinION nanopore reads aligned with minimap2
Undiff_cDNA.bam.bai - index file of above
Undiff_dRNA.bam - BAM file of RNA sequenced with MinION nanopore reads aligned with minimap2
Undiff_dRNA.bam.bai - index of above
undiff_TruSeq_1.bam - BAM file of cDNA sequenced by Illumina using TruSeq library
undiff_TruSeq_1.bam.bai - index of above
undiff_TruSeq_1_cufflinks.gtf - Cufflinks built transcripts using BAM -
undiff_TruSeq_1_stringtie.gtf - Stringtie built transcripts using BAM -
gencode.vM15.annotation.gtf - GTF of gencode annotation file

Instructions

1. Remove all the tracks you previously loaded, right click and select “Remove Track”
2. Load the above bam and gtf files
3. Go to gene Ptp4a1

Question 5

Which sequencing technology captures the reads starting from the 5’ of the transcript?

Question 6

Is it possible to determine the exact full structure of transcripts that are expressed by any of the sequencing technologies?

Question 7

In which sequencing technology is there a smooth (even) expression throughout an exon?

Question 8

Which sequencing technology has more errors and indels?

Question 9

What difference can you see in the Illumina TruSeq sequencing when comparing Aplp1 to Ptp4a1?