June 2015

RNA-Seq Analysis with Chipster

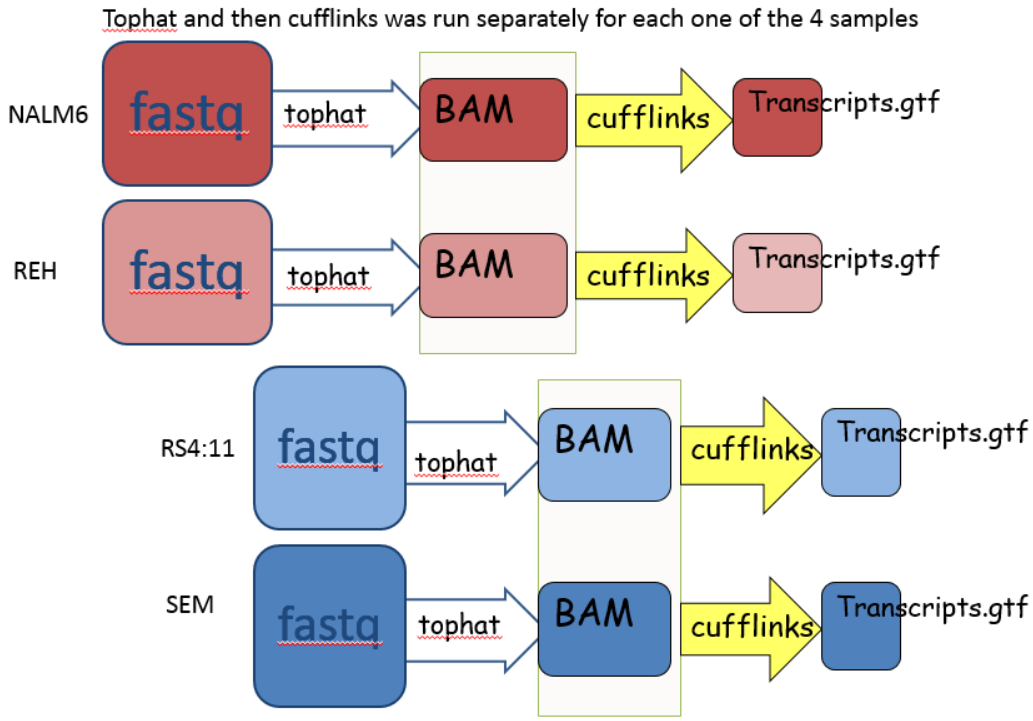
Dena Leshkowitz

Introduction

In this exercise we will learn how to analyse RNA-Seq data using the Tuxedo Suite tools: Tophat, Cuffmerge, Cufflinks and Cuffdiff. We will use again the RNA-Seq sequences derived from acute lymphoblastic leukemia (ALL) precursor B cell lines carrying a chromosome translocation (4:11), cells- RS4;11 and SEM, and compare to two precursor B cell lines that lack this translocation NALM6 and REH. This data is private do not distribute.

Following is flow of the pipeline:

1. Tophat & Cufflinks

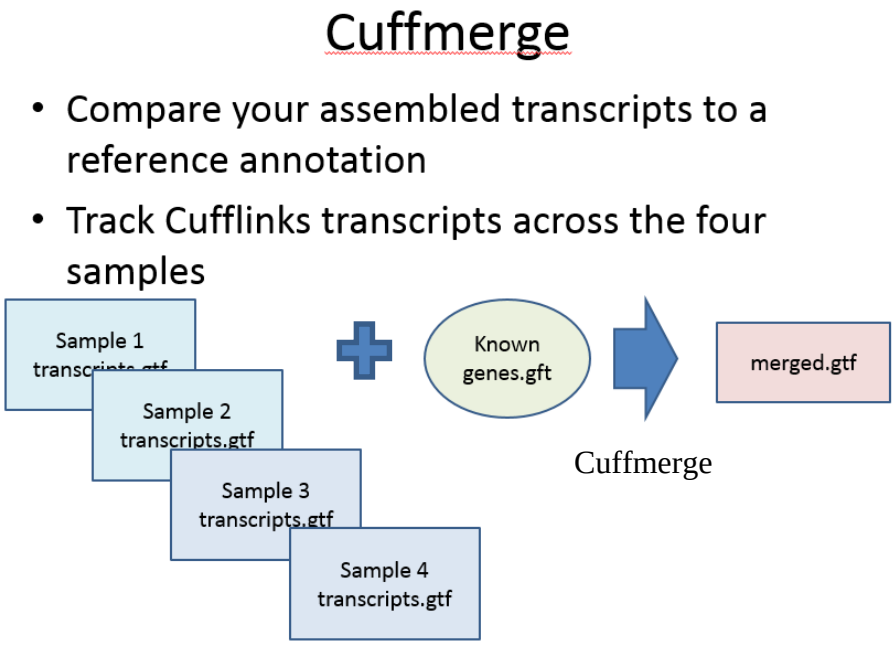


Unfortunately, do to time limitations and technical problems we have already run this analysis for you. For this exercise we used paired-end sequences of 100 base length. We had >40M reads per sample, yet the bam files you have contain only the sequences that mapped to chromosome 6. There are around ~3M fragments per aligned file.

The assembled transcripts: transcripts.gtf are in a format named **Gene transfer format** (**GTF**) used to hold information about gene structure (see explanation in http://asia.ensembl.org/info/website/upload/gff.html?redirect=no).

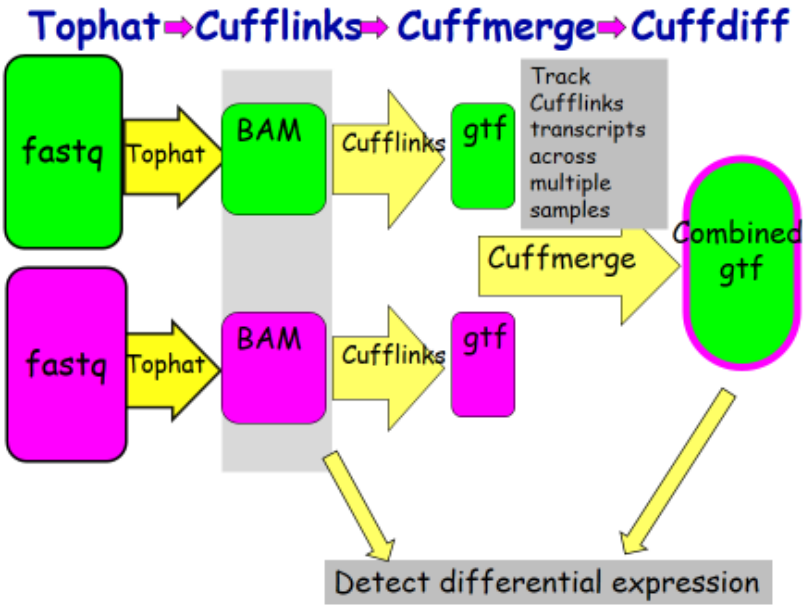
2. Cuffmerge

The four cufflinks outputs were merged to a single file using cuffmerge.



3. Cuffdiff

Finding differentially expressed transcripts is done using Cuffdiff which uses as input the assembled transcripts and the mapped reads.



Instructions:

1. Accessing the Data

Under D disk select folder “Course2015” and under it “Course2015-exercise4” you have the tophat bam file, cufflinks merged gtf file and the cuffdiff output.

2. Analysing Cufflinks output

Open the cufflinks merged file (only for transcripts on chromosome 6) merged_chr6.gtf with Excel.
Cufflinks was run using a refSeq annotation file (RABT mode). Make sure you understand the meaning of this term (you can look at http://cole-trapnell-lab.github.io/cufflinks/cufflinks/index.html). The transcripts and exons are coded to notify us whether this is a novel exon or one that is equivalent to a known exon.

Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

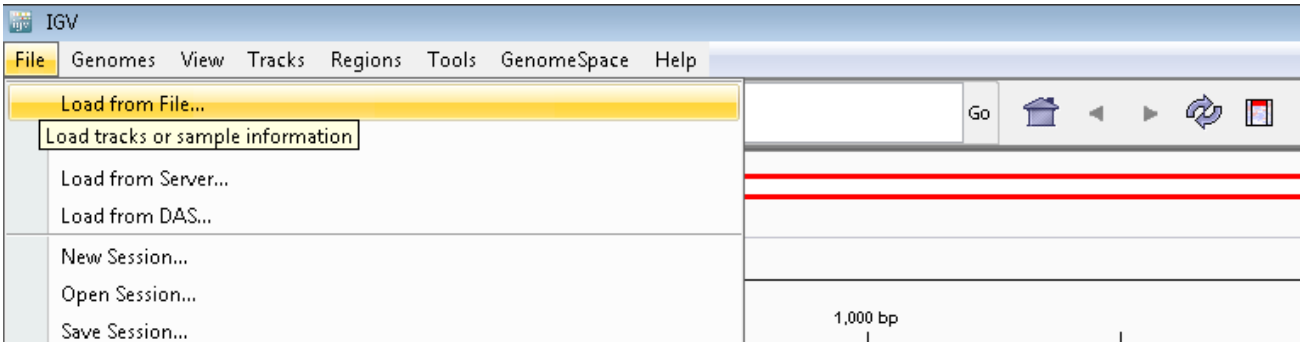
Open the merged_chr6.gtf file with excel and detect an exon which is a potential novel exon. Note the coordinates or the gene name. We will now analyse this with a genome browser.

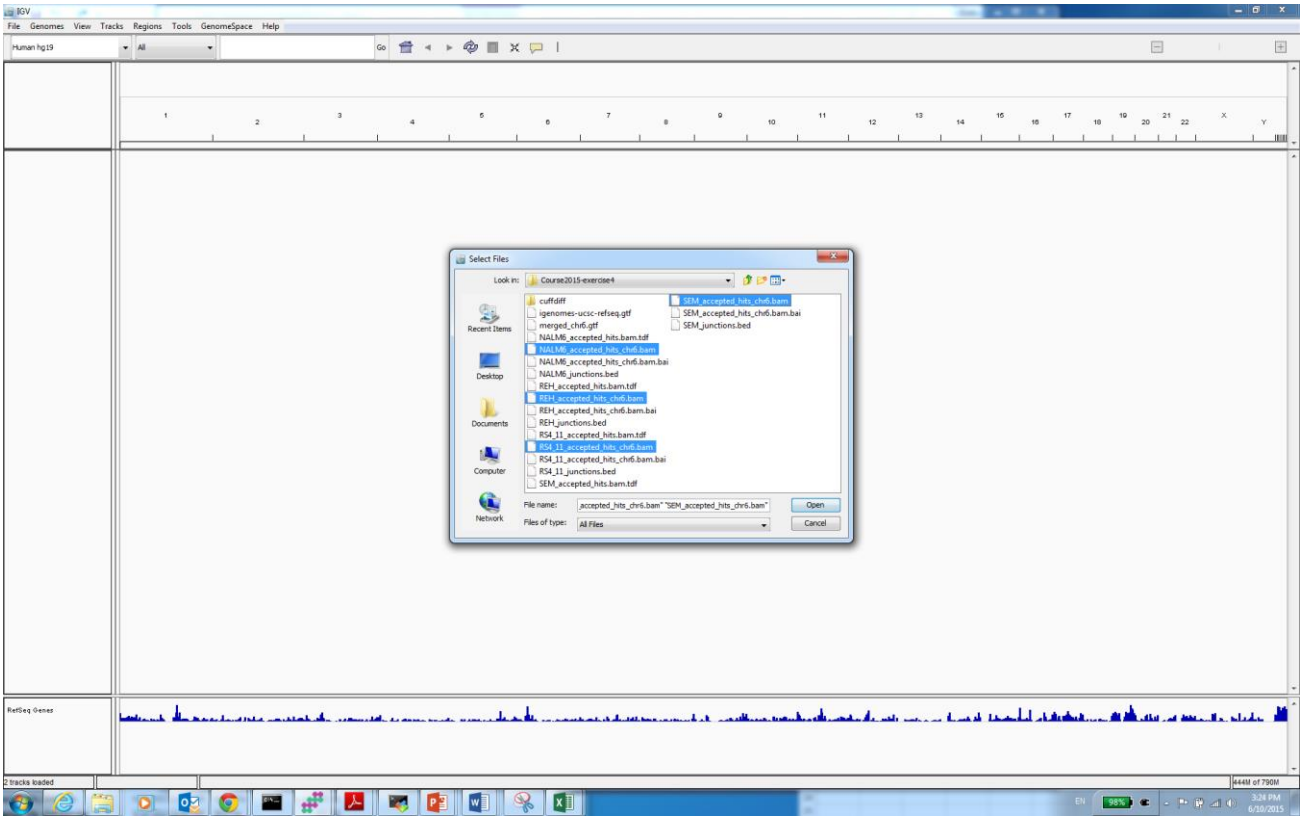
3. Browsing the alignments and assembled transcripts with a genome browser

a. Opening IGV and loading the files

We will use the Integrative genomics browser - IGV to view the mapped reads and the built transcripts. Open the IGV tool found on your desktop. Select run on the pop-up window. Once the application opened load the hg19 genome.

Load the files ending with bam (aligned read).

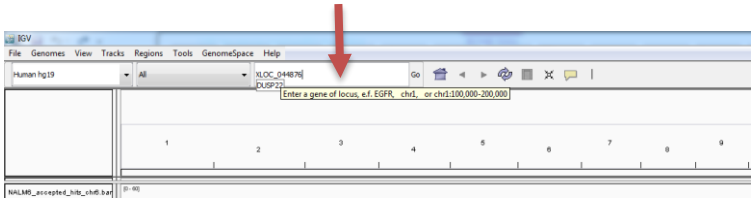




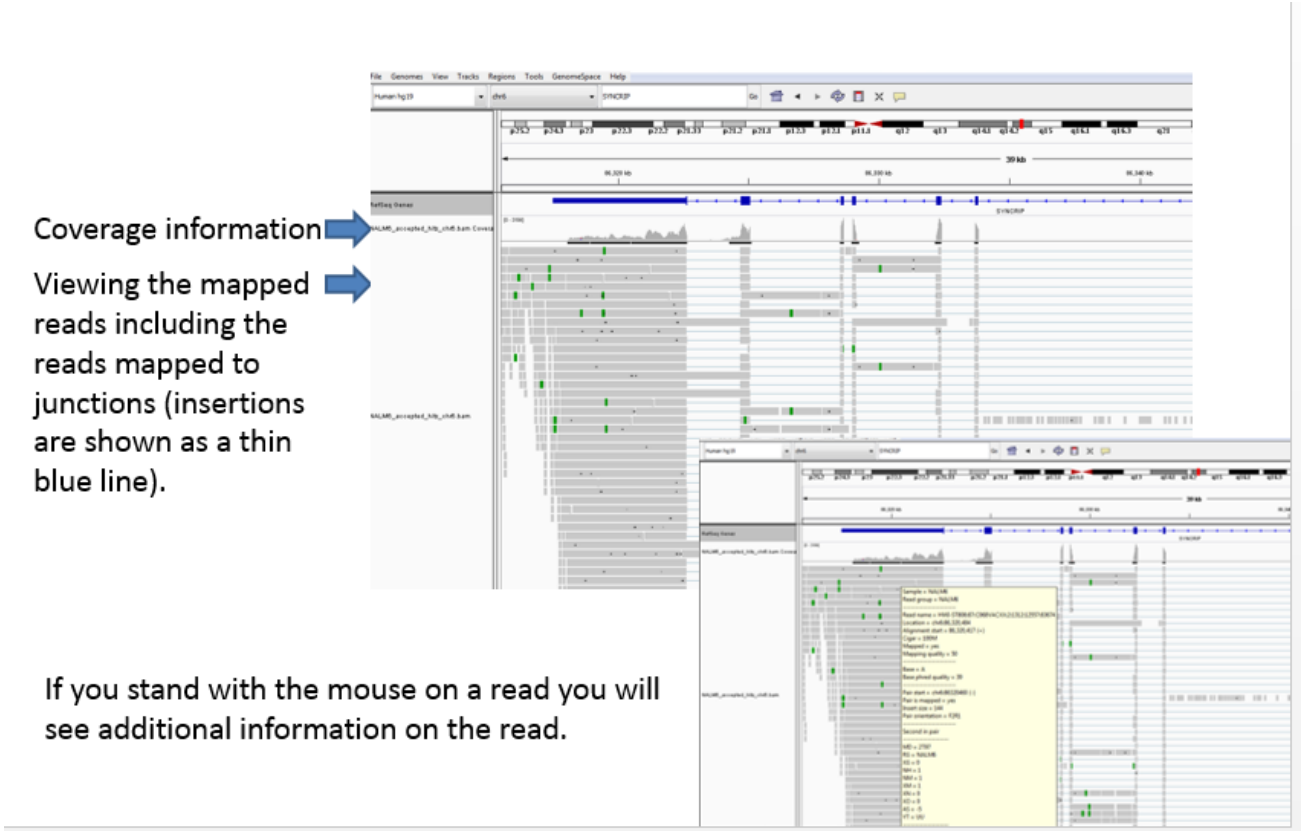
In order to view the reads we need to zoom in a certain region of the genome

Load the merged_chr6.gtf file (import it) as well, the program will notify you that it needs to build an index file select “Go” to permit it.

Now let’s zoom to the novel exon you identified previously. Type it in the window indicated by a red arrow below.

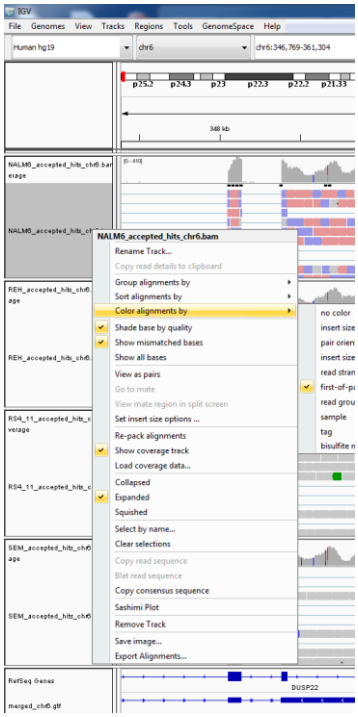


Now you can see the bam files at the read and base resolution.

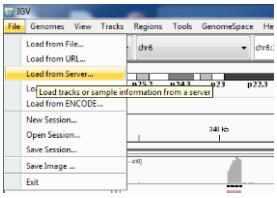


b. Trying different views and databases

There are various ways to color the reads- select a certain bam file and right click to change the coloring method.



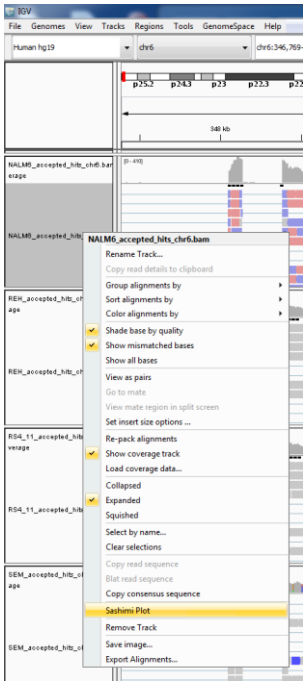
According to the refSeq annotation are you convinced the exon you selected is unknown i.e. novel?
RefSeq is conserved database that has relatively to other databases a small number of genes annotated. Therefore, let's load other databases and examine if this exon is also novel relatively to them. We can load Ensemble and Gencode, under File "Load from Server". A window will open, expand the information under "Gene" and select the above databases.

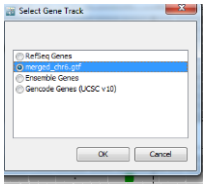


Is your exon novel also in regards to these databases?

c. Coverage on exons and junction – Shashimi plots

Shashimi plots have a very nice presentation of coverage in both exons and junction. Select a bam file and right click on it, then select the Shashimi plots. A window will appear - select the merged_ch6.gtf option and in the next window select all bam files to view the plot.





4. Analysing Cuffdiff output

Cuffdiff folder contains all the outputs from running cuffdiff with all the reads (>40M 100PE per sample) for the whole genome (not only chromosome 6). The Cuffdiff produces outputs for four different categories:

isoforms.fpkm_tracking	Transcript FPKMs
genes.fpkm_tracking	Gene FPKMs. Tracks the summed FPKM of transcripts sharing each gene_id
cds.fpkm_tracking	Coding sequence FPKMs. Tracks the summed FPKM of transcripts sharing each p_id, independent of tss_id
tss_groups.fpkm_tracking	Primary transcript FPKMs. Tracks the summed FPKM of transcripts sharing each tss_id

Open the isoforms_exp.diff file with Excel.

How many transcripts are quantified? Hint - filter on Status “OK” see the manual for explanation http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html

How many are differentially expressed? Hint - filter significant “yes”.

How would you compare this analysis to that performed on the gene level with HTSeq and DESeq, in regards to the number of differentially expressed genes?

Select a transcript that is differentially expressed and view the amount of reads from the various samples. Loading the files ending with tdf (found in the folder above cuffdiff) will allow you to see the coverage in non-zoomed mode and outside of chromosome 6.

Congratulations you reached the end.