BISC-577: Project # 4

Due on Tuesday, Mayl 05, 2015

Saket Choudhary
2170058637

Contents

Question # 1

(A): mRNA are a family of RNA that upon translation result into a sequence of amino acids as spedicifed by the corresponding codons as a result of gene expression
(B): transfer RNAs(tRNA) serves as a carrier of the amino acids trranspoting them to the ribosomes. Amino-acid-codon matching happens via the presence of an anticodon and is specific.
(C): Introns are 'inter-genic' regions that do not code for proteins and hence are absent in the mature RNA as they are removed via splicing. Exons on the other hand are the 'coding' regions of DNA. Mature RNA consists primarily of exons.
(D): Alternative splicing which involves removal of non-coding regions also gives rise to the possibility of multiple proteins being translated from the same gene depending on which exons are included and which ones are excluded. RNA silencing is another such process that increases RNA variability.
(E): Coding region of RNA consists of exons that for a protein. 5' UTRs and 3' UTRs which are also part of exon are upstream of initiation codon and downstream of the termination codon and both act as post transcriptional regulators. UTRs are not translated into proteins.

Question # 2

Gene ID: 3569 [IL6 interleukin 6]
HUGO: HGNC:HGNC:6018

B) $http://www.ncbi.nlm.nih.gov/gene/3569geneGenomicregions, transcripts, andproducts$

Exon count: 6 **From RefSeq:**

Transcript	Length(nucleotides bp)	No. of exons
$XM\_011515390.1$	2555	-
$XM\_005249745.3$	1969	-
$XM\_011515391.1$	978	-

I was not able to locate the number of exons on NCBI. So the number of exons is not indicated.

From Ensembl:

Transcript	Length(nucleotides bp)	No. of exons
$XM\_005249745.2$	1412	3
$NM\_000600.3$	1184	5

(C) None of the transcripts have the same number of exons as the original gene(6). This is expected, since the mature mRNA is a result of alternative splicing resulting in few exons being assembled while the introns are chunked off.

(D) Ensembl. The transcipts do not match. The XM_* comes from NCBI's automated eukaryotic genome annotation pipeline and are 'predicted' transcripts while the NM_* are the curated ones. This likely seems to differ, because the XM_* predicted transcripts arerefresshed periodically and the change might not reflect on Ensembl at the same time.

continued on next page...

Question # 3

A Splicing leads to removal of introns resulting in joining of exons. This process van suffer a lot of variation and hence mapping to a single reference as in the case of DNA is often not possible.
B Project Accession: $http://www.ncbi.nlm.nih.gov/bioproject/PRJNA257207$ There are 2 types of samples, investigating the reorganization of nuclear architecture ofhuman fobroblasts and MSCs. One type of samples come from early passage of replicative senescence while the other set are in late passage. r There are three biological replicates for the 2 type of conditions but no technical replicates(single run in each experiment)

C SRR1533801.fastq: 34507899
SRR1533801.fastq: 31246550

D Tophat is essentially an aligner that internally uses bowtie2. Reads from RNA-seq experiment will involve spliced regions. Hence a single read could have originally originated from two regions(exons) that are far apart on the genome(the reference sequence). Tophat first aligns the reads to the references, some of the reads will remain unmapped, possibly due to the alternative splicing(other reason might be contamination, mutations etc) in which case Tophat takes these unmapped reads and then infers the splice site regions. Bowtie2 cannot handle aligniing reads by splitting(allowing very large gaps)

E $sort\ -k5,5\ junctions.bed\ ||tail$
SP1: "chr2 216243994 216245583 JUNC00073269 9995 - 216243994 216245583 255,0,0 2 46,50 0,1539"
SP2: "chr19 55897756 55897987 JUNC00054766 999 + 55897756 55897987 255,0,0 2 50,50 0,181"
SRR1533801:
Number of splice sites: 125761
Splice junction with max reads: SP1(above)
Number of reads at SP1: 9995

SRR1533804:
Number of splice sites: 95823
Splice junction with max reads: SP2(above)
Number of reads at SP1: 999

F $awk'printf\ "\%s\%d\%d\%d\%d", \$1, \$3 - \$2, \$5, \$2, \$3'\ junctions.bed\ |\ sort\ -k2,2\ |\ tail$ SRR1533801 longest junction site(SP1): "chr4 9999 7 186231930 186241929"
SRR1533805 longest junction site(SP2): "chr5 9999 40 168139310 168149309"
$[chr][length][numerbfreads][start][end]$

continued on next page...

Question # 4

A Cufflinks takes an alignment file, assembles the transcript and estimates their abundance testing for differential expression.
Command used: $grep -r\ 'IL6'\ genes.gtf \| grepNM$ The 'genes.gtf' was downloaded from tophat's website and came bundled with other indices(iGenome bundle)
C FPKM measures the abundance of transcripts(RPKM for single end reads) is a normalized count to measure the abundance. Normalization is essential to adjust for the number of sequenced and mapped reads.
SRR1533804.fastq gene: "CUFF.8926 - - CUFF.8926 - - chr14:24702148-24702409 - - 9.99986 4.20454 15.7952 OK"
SRR1533804.fastq transcript: CUFF.8926.1 - - CUFF.8926 - - chr14:24702148-24702409 261 9.13859 9.99986 4.20454 15.7952 OK"

The 10^{th} column indicates the abundance which are 9.99 and 9.12 respectively for gene and transcript of SRR1533804

Question # 5

A "cuffdiff transcripts.gtf SRR1533801.bam SRR1533804.bam"
B "SERPINA9 SERPINA9 - chr14:94929057-94942670 q1 q2 OK 0 0.555802 inf -nan 0.00015 0.0408021 y" The logFC tunrs out to be inf, probably indicating a novel transcript not found in the original gtf file.
Command: "sort -k10,10 genediff — head"
C Given a control and treatement experiment, the differences can arise due to multiple attributes. As the number of attributes increase the probability of difference between control and experiment groups will tend to increase. Correcting for multiple testing adjusts this probability for those multiple attributes to reflect the corrected probability. **D**

Question # 6

A I did not get any output for differential splicing, splicingdiff was empty. I think I did the 'cuffdiff' part incorrect. I tried merging the gtf for control and treatment, but it seemed to fail with an invalid transcript id error. **B** cuffdiff finds significant changes in transcript levels. Given two samples and the number of reads mapping to each transcript, cuffduff's performs a hypothesis test, of how likely the change is due to the difference in two groups rather than just by chance

continued on next page...