SYSB 3036 W11: Gene Expression and RNA-Seq

Frank Aylward[1]

[1]Virginia Tech

dx.doi.org/10.17504/protocols.io.v4qe8vw

Frank Aylward
Virginia Tech

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

Getting started

1

Last tutorial we went over basic read mapping protocols using bowtie and SAMtools. We used data from a infection experiment that used Mycobacterium smegmatis and a bacteriophage called D29. We examined only one time-point that took place 15 minutes after infection. Today we will finish analyzing that dataset and then look at another time-point 60-minutes after infection and compare our results.

This will be easiest if you work in the same folder as last week, with your D29.fna file, the indexed bowtie2 files, the D29.bed file, etc.

If you did this in the same folder as last week, you should also have a nice T1.sort.bam file that we will also use later in this tutorial.

Look at BED file

2 Now we have the genome .FNA file and sorted bam files for two time-points. We can map the reads to the genome, as have already done, but that just tells us how many reads map to the genome as a whole. We still need information about where genes and other features are encoded in the mycobacteriophage D29 genome, so that we can find out what genes are expressed at different times. After all, knowing that a large number of reads map between coordinates 10377 and 11421 in the phage genome is not necessarily useful unless we also have information about what genes are encoded in that region.

For this next step we will use the mysterious D29.bed file that we downloaded last week but did not use. BED files are tab-delimited tables with information about the different features of a genome, their coordinates, which strand they are encoded in, and some annotation information. The tables are a bit large and somewhat unweildy to look at in the command line, but just to get an idea of a format let's look at the first two lines:

head -n 2 D29.bed

This should give you the general idea. For details on BED format see here: https://genome.ucsc.edu/FAQ/FAQformat.html#format1

Identify highly expressed genes with BEDtools

3 Now we have all of the necessary files to identify which genes were highly expressed in our transcriptome. To do this we will use a tool called BEDTools.

First let's take a look at the tool BEDtools and what utilities it has by typing 'bedtools'. You will see a large number of options- this is a very comprehensive tool!

Here we will be using the 'intersect' utility, which will tell us which reads in our sorted bam file intersect with regions in the BED file.

bedtools intersect -a D29.bed -c -bed -f 0.2 -b T1.sort.bam > intersect_T1.txt

A quick note on the flags used here.
-a and -b denote the BED file and sorted bam file, respectively.
-c indicates that we want to count the number of reads that overlap with BED regions. The default output is to provide each read individually with the feature it mapped to.
-f 0.2 denotes that the read has to overlap by at least 20% of it's length to be considered mapping to a feature. One can imagine a long polycistronic mRNA with many genes encoded in it; some reads may overlap just slightly with a specific gene. Here we require at least 20% overlap, which will help ensure reads are mapping properly.

The output will be a BED file, the same as we used for the input, but with an additional column appended to the end. This column will have the number of reads tha map to each feature.

Now a rather basic question that it would be nice to answer is: What are the highest expressed genes in our transcriptome?
To answer this we should be able to sort the intersect file by the last column with the Unix sort command.

sort -k 11,11 -rn intersect_T1.txt | cut -f 10,11 | head -n 5

So of the top 5 most highly expressed genes in the mycobacteriophage D29 genome, we have a DNA primase, a helicase, two hypothetical proteins, and one glutaredoxin. The first two are noteworthy since this particular transcriptome sample was taken after only 15 minutes of infection, and at this early stage we would expect to find genes associated with viral DNA replication ("early genes"). The hypothetical proteins are also not surprising, since many genes in bacteriophage genomes are uncharacterized and have no known function.

Repeat read mapping for another time-point

4 Now let's repeat some steps from last week using a different time-point.
We will use the same steps again in a protocol identical to last week, only this time we will use a different FASTQ read file that corresponds to the 60-minute time-point.

Since a lot of these commands are almost identical to those we used before, I will just list them here, with brief descriptions:

Get the new fastq file:
fastq-dump -X 10000 SRR5585002

Map the reads with bowtie2:
bowtie2 -U SRR5585002.fastq -x D29 -S T2.sam

Filter out the reads that did not map and convert to BAM format:
samtools view -bS -F 4 T2.sam > T2.bam

Sort the BAM file:
samtools sort T2.bam T2.sort

Index the BAM file:
samtools index T2.sort.bam

Get highly expressed genes for the new time-point

5 And now we can use the BED file in the same way to find out what genes are highly expressed in the 60-minute time-point:

bedtools intersect -a D29.bed -c -bed -f 0.2 -b T2.sort.bam > intersect_T2.txt

and

sort -k 11,11 -rn intersect_T2.txt | cut -f 10,11 | head -n 5

Note that now we have a different set of top-5-most-expressed genes. Now instead of primase or helicase genes we are finding structural genes like tail fibers, scaffold proteins, etc. This is interesting since these are typical "late genes" that are expressed later in the phage infection cycle. For phage, the first thing they have to do after infection is replicate their genomes inside of a host, and after that they have to package this DNA into new viral particles. Our results indicate that by 60 minutes the phage has moved onto this next phase.

Combine output files

6 We may also want to combine our T1 and T2 results into one file. Since the genes are ordered in the same way we can use a simple "paste" command.

paste intersect_T1.txt intersect_T2.txt > combined_mapping.txt

And to get the read counts for T1 and T2 we can specify columns 11 and 22:

cut -f 10,11,22 combined_mapping.txt | head

You can sort this in another pipe command if you want to investigate further and see the results side-by-side. For example, to see the most highly expressed genes in the T1 sample we can use:

cut -f 10,11,22 combined_mapping.txt | sort -rn -k 2,2 | head

You can use -k 3,3 to do the same with T2.