

C C A T T 0 1 0 0 0
G A G G A 0 1 1 0 1
G A A T T 0 0 1 1 0
A C A A G 0 0 1 0 0
T A C C A 0 0 1 1 0
T T A C A 0 1 0 0 0
A C C T C 0 0 0 1 0
A A G G A 0 0 0 0 0
G A T G A 0 1 1 0 0
T A G A T 0 0 1 0 0
G A T G A 1 0 1 0 0
T G T A G 1 0 0 0 0
T A G T A 0 0 0 0 0
G A T A T 1 0 0 0 0
G A G T G 1 0 0 0 0
A G A T T 1 0 0 0 0
G A G T A 1 0 0 0 0
T G A T G 1 0 0 0 0
A T T A G 1 0 0 0 0
T A G A T 1 0 0 0 0
G A G A 1 0 0 0 0
G T A 1 0 0 0 0
G A T 1 0 0 0 0
T A G 1 0 0 0 0
A G 1 0 0 0 0
G A 1 0 0 0 0
A 1 0 0 0 0

Tutorial

Tutorial: RNA-Seq Analysis Part II: Non-Specific Matches and Expression Measures

March 15, 2013



Tutorial: RNA-Seq Analysis Part II: Non-Specific Matches and Expression Measures

This tutorial is the second part of a series of tutorials about RNA-Seq analysis. We continue working with the data set introduced in the first tutorial.

Here, we will first explain how non-specific matches are treated, and then we will consider the effect of using different expression measures.

Running the same data set with and without non-specific matches

The term non-specific matches refers to reads that can be mapped equally well to more than one location on your reference. Since it is not possible to tell which transcript such reads actually came from, the Workbench has to decide where to place them. In such cases, the Workbench first estimates the expression of each gene based only on reads that map *uniquely* to that gene. It then uses this information to weight the distribution of the reads that can be mapped equally well to more than one location.

For example, imagine a situation where you have nine reads that match equally to two genes. Let's say one of these genes has twice the number of unique matches compared to the other gene. The first gene will, on average, have six of the nine reads counted towards its expression. The other one would get, on average, three of the reads.

Here, we focus on the effect of including these non-specific matches in the RNA-seq analysis. In the analysis done during the first tutorial in this series, we used the default mapping parameters. This means that the **Maximum number of hits for a read** was set to 10. This means that all reads that matched in 10 or fewer places were included in the mapping, with reads that could map to multiple locations being distributed as described above.

Run a **RNA-Seq Analysis** (🇺🇸) on the *Brain spike* sample. For this analysis, set the **Maximum number of hits for a read** to 1. You find this setting under **Mapping Settings** section of the wizard. Leave the rest of the settings as the defaults. If you have changed the parameters during earlier work, please set the parameters to default by clicking the button (↶) at the bottom of the dialog. If you do this, you will have to define the reference sequence again.

You will now have an RNA-Seq sample where all reads matching in more than one position are excluded. If you go to the **History** (📖) view of this new sample, you can see how many reads were not mapped. If you compare these numbers, you can see the first sample has 214631 unmapped reads whereas the second run without multihit reads has 226761 unmapped reads.

Now that we have more than one RNA-seq sample, it would make sense to save them with meaningful names. The histories of the two samples generated so far are shown in figure 1. Here, we have named our earlier sample **only specific**, and the new sample is called **including nonspecific**. To re-name saved samples, just click on the name in the navigation area on the left, and then click again. You should then be able to edit the object name. If you have chosen to open (rather than saving) your samples earlier, please click on the (💾) icon to **save** each sample. Now **re-name** them, preferably with the same names we have chosen.

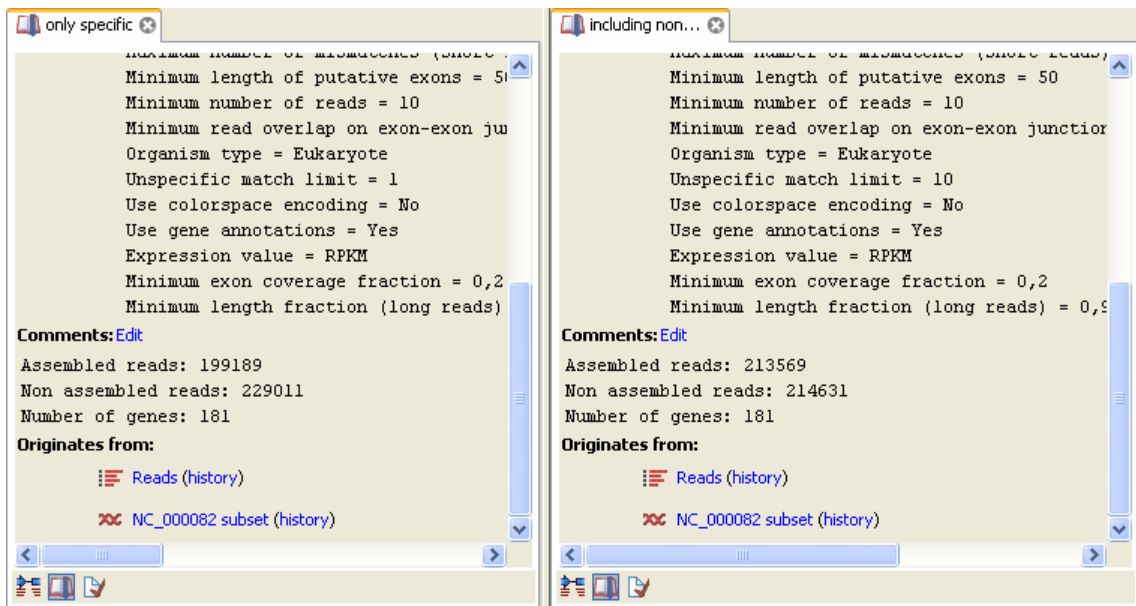


Figure 1: Comparing the history entries for the two samples.

Comparing the data in a scatter plot

Now, we want to see what this difference means in terms of the expression values. In order to compare the two samples, we set up an experiment:

Toolbox | Transcriptomics Analysis (🇺🇸) | Set Up Experiment (📊)

Select the two RNA-Seq samples (🇺🇸) that you have just saved and click **Next**. Choose an un-paired, two-group experiment, Choose to **Set new expression value** and choose **Genes: Total exon reads**. Click **Next**. Name the groups *including non-specific* and *only specific* and click **Next**. Right-click each of the samples and assign it to the appropriate group. Choose to open the results. Click on the **Finish** button.

You should now see an experiment based on the two samples. We will go into more details with the experiment later - for now we are interested in looking at the scatter plot. Click the **Scatter plot** (📊) icon at the bottom of the view.

At the bottom of the **Side Panel** you select the values to plot. Select *including non-specific Total exon reads* versus *only specific Total exon reads*, and you will see a view as shown in figure 2.

The scatter plot now shows the expression levels of the two samples. Since the RNA-Seq analysis was run on the same data set with the only difference being the treatment of non-specific matches, you can now directly see the effect of using and distributing the non-specific matches in this way.

Many of the genes have close to identical expression measures, as you can see by the number that lie along the $x=y$ line in the plot. This is to be expected, as we are working with the same underlying dataset. Some genes in this example do show higher expression in the sample that includes non-specific matches. To see the outliers more clearly, set the **Dot type** under **Dot properties** in the **Side Panel** to **Dot**.

The most outlying gene is *Sept5*. If you place your mouse on the dot as shown in figure 3, you can see the feature ID (gene name) and the x and y values of the dot.

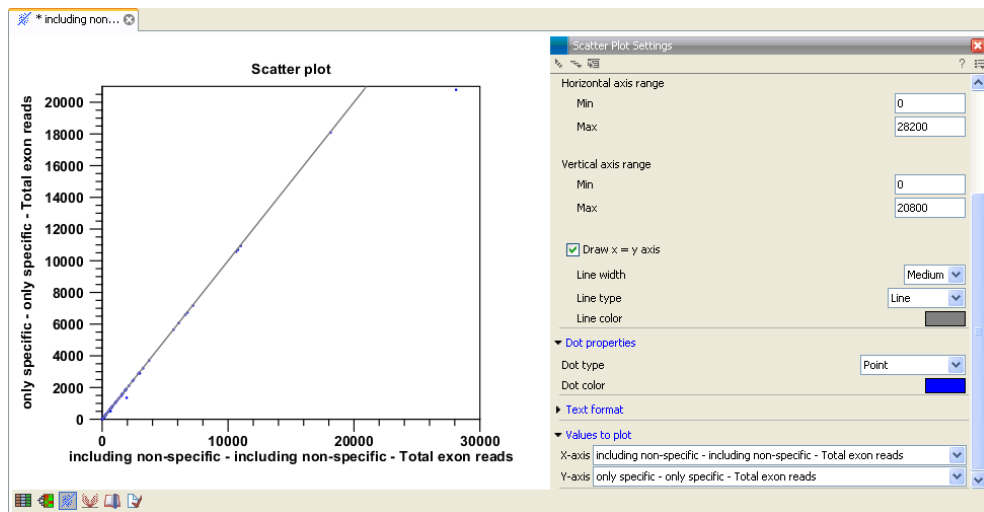


Figure 2: A scatter plot showing the effect of including non-specific matches in the expression measure.

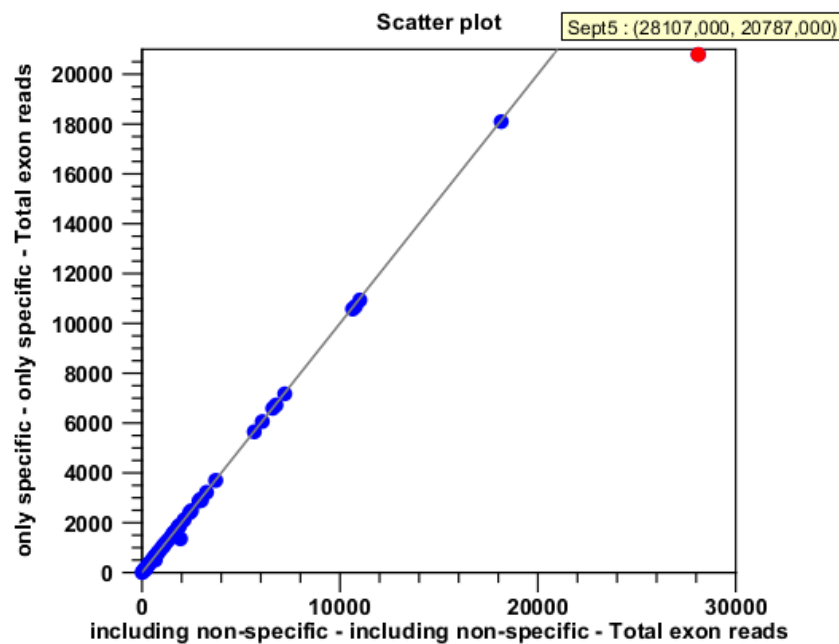



Figure 3: Gene *Sept5* is one of the genes showing a notable difference in expression.

Open the *including non-specific* RNA-Seq sample and locate this gene by typing *Sept5* in the filter at the top. Double-click to open the mapping. Double click on the tab at the top of this view so only the mapping data is visible. Zoom out all the way, e.g. by clicking on the () button in the top toolbar. Look at the end of the gene. You should see a lot of reads that are yellow, which is the color used for non-specific reads. In this case, these yellow reads are the ones contributing to a higher expression measure for *Sept5* in the sample where we included non-specific matches.

By looking at the gene annotations, you can also see the reason why there are so many non-specific matches. As shown in figure 4, there is an overlapping gene near the end. This means that all the reads that map to this part of the *Sept5* gene also map equally well to the beginning of the *Gp1bb* gene. These reads are then treated as non-specific matches. You may notice differences in the detail of how your mapping looks compared to the one in figure 4. Try playing

with the view settings in the right hand pane and see how this affects the information displayed to you.

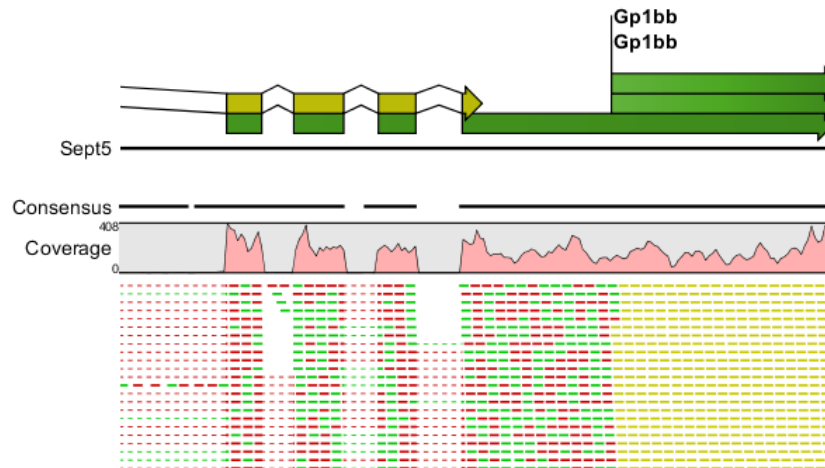


Figure 4: *Gp1bb* is overlapping the end of the *Sept5* gene.

If you opened the *Gp1bb* mapping you would also see the non-specific reads at the beginning where it overlaps with **Sept5**. Because we can see the overlap, we know why we have non-specific matches, but it could be that these reads would also match other places on the reference. It's easy to check if the same region is present other places in the reference by conducting a BLAST search.

To set up a BLAST search for this, select the relevant part of the gene, which here is the part that the yellow colored reads are mapping against. Ensure you have clicked on the Selection Tool button in the top tool bar. Then you can click and drag along the reference sequence to select a region. Once selected:

right-click the selection|BLAST against Local Data (🖱️)

This is illustrated in figure 5

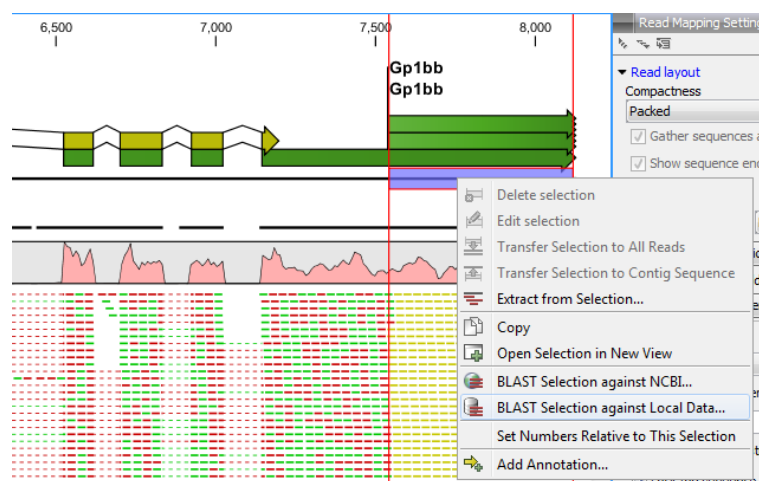


Figure 5: BLAST against the reference.

The idea with BLASTing this selection would be to use the reference sequence as target and see how many hits you would find. In this case there is only one good hit, but if you have a region of non-specific matches that are not due to overlapping genes, you can use this approach to try



to identify which other gene is "competing" for these reads. For now, please close the BLAST dialog. We do not include BLAST searching in this tutorial, but recommend that you instead read the BLAST tutorial to learn more about BLAST in the *CLC Main Workbench*.

Close the mapping view by clicking on the (X) in the top right hand side of the viewing tab. Go back to the experiment object and switch to the table view (table icon).

Enter *Gp1bb* in the filter and click with your mouse on the *Gp1bb* gene. Switch back to the scatter plot (scatter icon) and *Gp1bb* will now be high-lighted with a red color. Click **Zoom in** (zoom icon) and click a couple of times on the gene to zoom in (see figure 6).

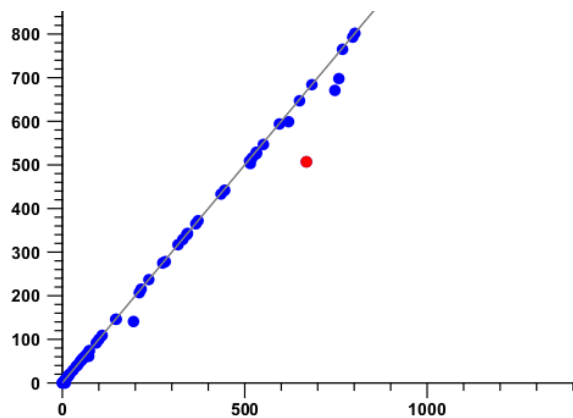


Figure 6: Zooming in on the *Gp1bb* gene in the scatter plot.

You can now see that this gene also exhibits differential expression between these two samples, but to a lesser degree than the *Sept5* gene. Open the *Gp1bb* mapping from the *including non-specific* sample, and you can see that there are fewer yellow reads than in the *Sept5* mapping. As explained above, the non-specific reads are distributed according to the number of unique reads, and when you compare the two results, it is evident that there are many more unique reads in the *Sept5* gene, so more of the non-uniquely mapped reads were counted towards its expression measure than were counted towards the *Gp1bb* gene. You can easily see the difference in the RNA-Seq result table as shown in figure 7.

including non...

Rows: 2 / 181 Filter: Match any Match all

| | | | | |
|------------|----------|-------|----------------------------------|----------------------------------|
| Feature ID | contains | sept5 | <input type="button" value="+"/> | <input type="button" value="X"/> |
| Feature ID | contains | Gp1bb | <input type="button" value="+"/> | <input type="button" value="X"/> |

| Feature ID | Unique exon reads | Total exon reads |
|------------|-------------------|------------------|
| Gp1bb | 421 | 567 |
| Sept5 | 15770 | 22117 |

 Expression value:

Figure 7: Comparing the number of unique reads between *Gp1bb* and *Sept5*.

From the scatter plot in figure 2, it is obvious that the decision on whether to include non-specific

matches or not is very important. For some genes, the difference in expression is substantial. The importance of being aware of the potential effects of non-specific matches becomes even more evident when looking at the full data set where the proportion of non-specific matches is higher. Consider that with the full reference transcriptome, there is a greater chance of finding sequences that are represented more times, e.g. arising through gene duplications.

It is hard to make general recommendations on how to treat non-specific matches. One of the pitfalls when including non-specific matches is that the number of unique matches can be too low to ensure a reliable distribution of the non-specific matches. One way of approaching this problem would be to run the same data set with different settings as shown in this tutorial. That will enable you to perform random checks of the genes whose expression is significantly altered, and you will be able to identify this kind of pattern. On the other hand, if you completely disregard non-specific reads, you may underestimate the expression levels of genes in gene families.

We refer to [Mortazavi et al., 2008] for an in-depth discussion of this topic.


The RPKM expression measure

Normalizing for sample size

The observations made from figure 2 lead to another important consideration when dealing with RNA-Seq analysis: you have to decide which expression measure you want to use. When you have several samples (as in this example with four different samples), these will have different numbers and qualities of reads. You will often see that there is quite a big difference between the samples in the number of reads that can be matched. This means that it can be hard to compare the expression of the same gene in different samples simply by looking at the number of reads matched, which is what we have done so far in this tutorial by examining total exon reads. When comparing the groups *including non-specific* versus *only specific total*, you can see this effect somewhat, since these two samples have 213,569 and 201,439 mapped reads, respectively. This means that you have an asymmetry in the scatter plot when using total exon reads as the expression measure (see we could see in figure 2).

There is another expression measure, RPKM (Reads Per Kilobase of exon model per Million mapped reads), which seeks to normalize for the different number of mapped reads between samples. We will now investigate RPKM in greater detail. Go back to the scatter plot in figure 2. Change the values to be plotted from total exon reads to RPKM for the two samples. You should now see a scatter plot as shown in figure 8.

Where figure 2 showed either dots falling on the $x = y$ axis or below, you now see dots falling primarily slightly above $x = y$ axis or below. This is because the RPKM takes into account that the total number of mapped reads is higher in the *including non-specific* sample than in the *only specific* sample. RPKM is defined as $RPKM = \frac{\text{total exon reads}}{\text{mapped reads(millions)} \times \text{exon length (KB)}}$.

Let's investigate two of the genes in the scatter plot. First, identify the two genes at the top of the scatter plot – one above the $x = y$ axis, one below. One of them is the *Sept5* gene that we have previously investigated. This still shows higher expression in the *including non-specific sample* because of the high number of non-specific matches. The other gene is *Sst*. Switch back to the experiment table  and compare the total exon reads for both samples (you can deselect sample columns under **Sample level** in the **Side Panel**, that will ease the overview). The value is 6600 and 6514, respectively, so the number of total exon reads is almost identical. What is then the reason that the PRKM value is higher for the *only specific* sample? This is because this

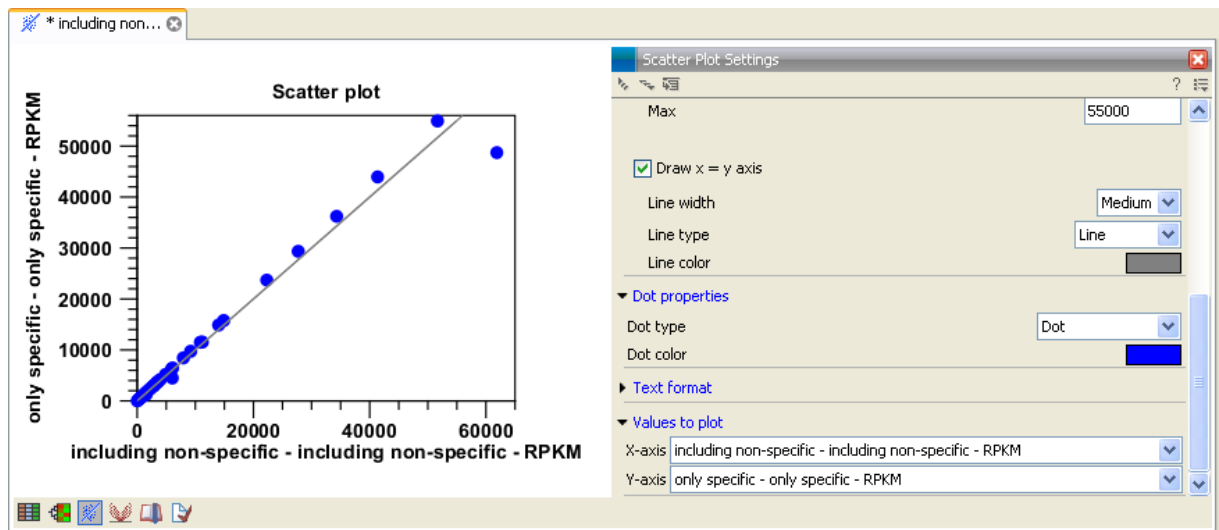


Figure 8: The effect of including non-specific reads compared using RPKM.

sample has a lower number of mapped reads, and the RPKM will thus be higher (see definition of RPKM above).

Normalizing for transcript length

In a sample, if two transcripts are present in the same number of copies, and the sequencing is unbiased, you would expect the same number of reads from each transcript. But if one transcript is short and the other is long, you would expect the long transcript to yield more reads. So if you wish to compare the expression of transcripts within the same sample, you need to take the transcript length into account.

If you look at the definition of RPKM above, you can see that besides number of mapped reads, the *exon length* is also considered. The idea behind this is to make it possible to compare expression levels of different transcripts.

Open the *only specific* sample and sort the table on total exon reads, you can see that the genes *Abcc5* and *Comt* (number 15 and 16 from the top) have almost the same number of reads (2,916 and 2,892). However, their expression value measured in RPKM is 2,333.78 and 11,456.38, respectively (see figure 9).

This is simply due to the difference in transcript length which you can also see in the table under **Exon length** (which sums the lengths of all the annotated exons).

Close all open views, save the experiment, and you are ready for part III.

only specific

Rows: 181 Filter:

| Feature ID | Expression values | Exon length | Total exon reads |
|-------------|-------------------|-------------|------------------|
| Camk2n2 | 23.571,09 | 1277 | 6069 |
| Eif4a2 | 14.687,97 | 1895 | 5612 |
| Zdhc8 | 3.758,47 | 4868 | 3689 |
| Etv5 | 4.165,51 | 3822 | 3210 |
| Abcc5 | 2.333,78 | 6197 | 2916 |
| Comt | 11.456,38 | 1252 | 2892 |
| Ppp1r2 | 3.462,28 | 4074 | 2844 |
| D16H22S680E | 8.263,88 | 1471 | 2451 |
| Rtn4r | 6.428,53 | 1874 | 2429 |
| Cldn5 | 8.432,15 | 1414 | 2404 |
| Abcf3 | 3.577,97 | 3288 | 2372 |
| Hira | 2.294,97 | 4534 | 2098 |
| Dgkg | 2.960,93 | 3201 | 1911 |

Figure 9: Nearly the same number of total exon reads for two genes leads to widely different RPKM values because of the difference in transcript lengths.



Bibliography

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628.