



Tutorial:
RNA-Seq Analysis
Part II (Tracks):
**Non-Specific Matches,
Mapping Modes and
Expression measures**

February 24, 2014

Tutorial: RNA-Seq Analysis Part II (Tracks): Non-Specific Matches, Mapping Modes and Expression measures

This tutorial is the second in 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 discuss the **Map to gene regions only (fast)** and the **Also map to inter-genic regions** option. Finally we will discuss different types of 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 random 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 well 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 set the **Maximum number of hits for a read** to 1. This means that only reads that had a unique match were mapped.

1. Now run an **RNA-Seq Analysis** (🔧) on the *ESC-1* sample again, with the same parameters as for the sample analyzed in the first tutorial, *except* for the **Maximum number of hits for a read** which you now set to 10. You will find this parameter in the "Mapping options" wizard step (figure 1).

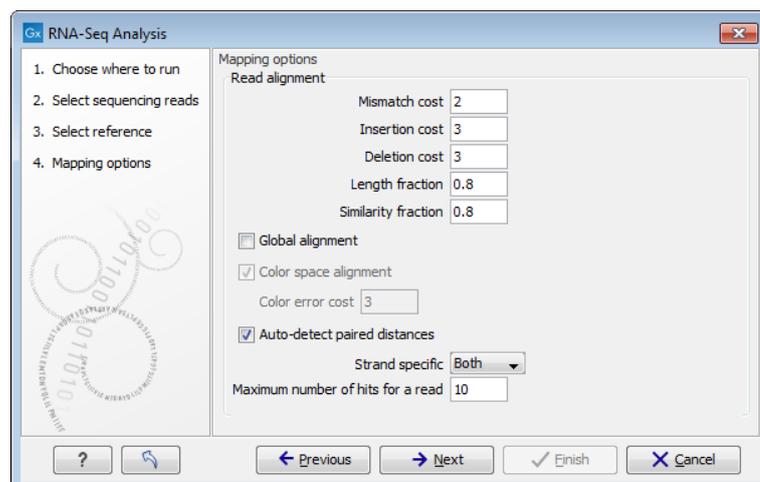


Figure 1: Change the "Maximum number of hits for a read" to 10.

If you have forgotten which parameters you used in your first analysis of the *ESC-1* sample, open one of the outputs you had for that analysis (e.g. the "ESC-1 (GE)" track) and click on

the (📄) at the bottom of the editor to see the history of the track (see Figure 2).

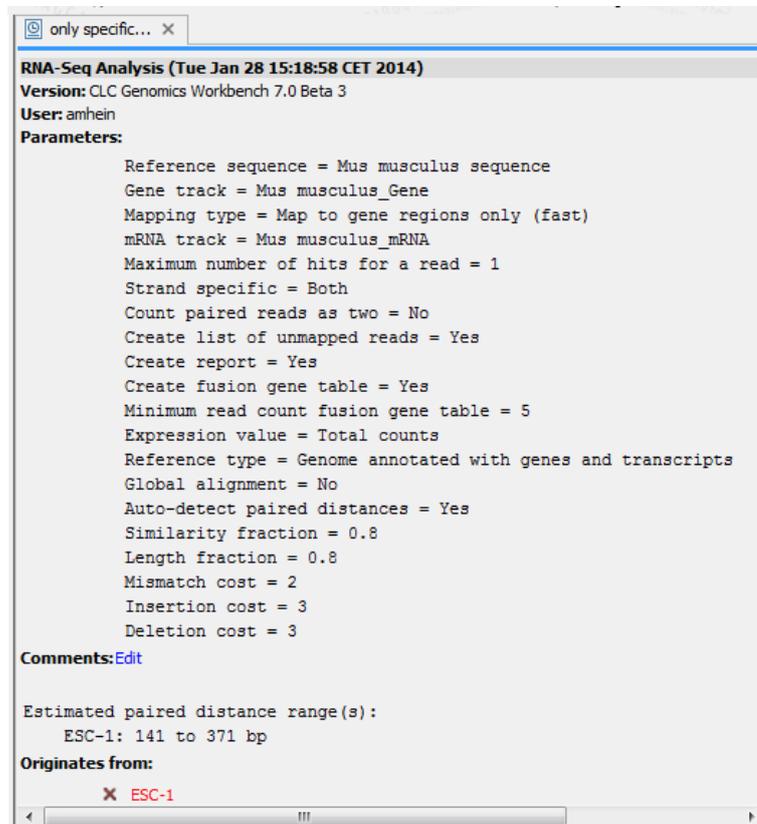


Figure 2: The history of the results shows the parameter values of the algorithm that were used to generate the results.

Setting the **Maximum number of hits for a read** parameter to 10 means that the algorithm will consider reads that have up to 10 equally good best matches, and distribute those among their match locations proportionally to the numbers of reads that map uniquely to the genes annotated in the mapping location regions.

Make sure that you choose "Save" in the final wizard step and create a folder in which to save your results.

2. Now that we have more than one RNA-seq sample, it would make sense to save them with meaningful names. Rename the results of your first analysis so that you can distinguish them from those from the second (e.g. give them the prefix "only specific"). To re-name saved samples, click on the name in the **Navigation Area** on the left, wait a second, and then click again. You should then be able to edit the object name. Alternatively, right-click on the name in the **Navigation Area** and select the option "Rename". If you have chosen to open (rather than save) your samples earlier, please click on the save (📄) icon in the Workbench toolbar to **save** each sample.
3. Open the RNA-seq reports generated in your two analyses. We wish to compare them side by side. To do that, right-click at the top pane of one of the reports and chose "view" and then "split vertically" (alternatively you can grab one of the reports by its pane and drag it to the side of the view). Hide the side-panels by clicking the side-panel icon (▶) and scroll down to the section "3 Mapping statistics" in the reports (figure 3).

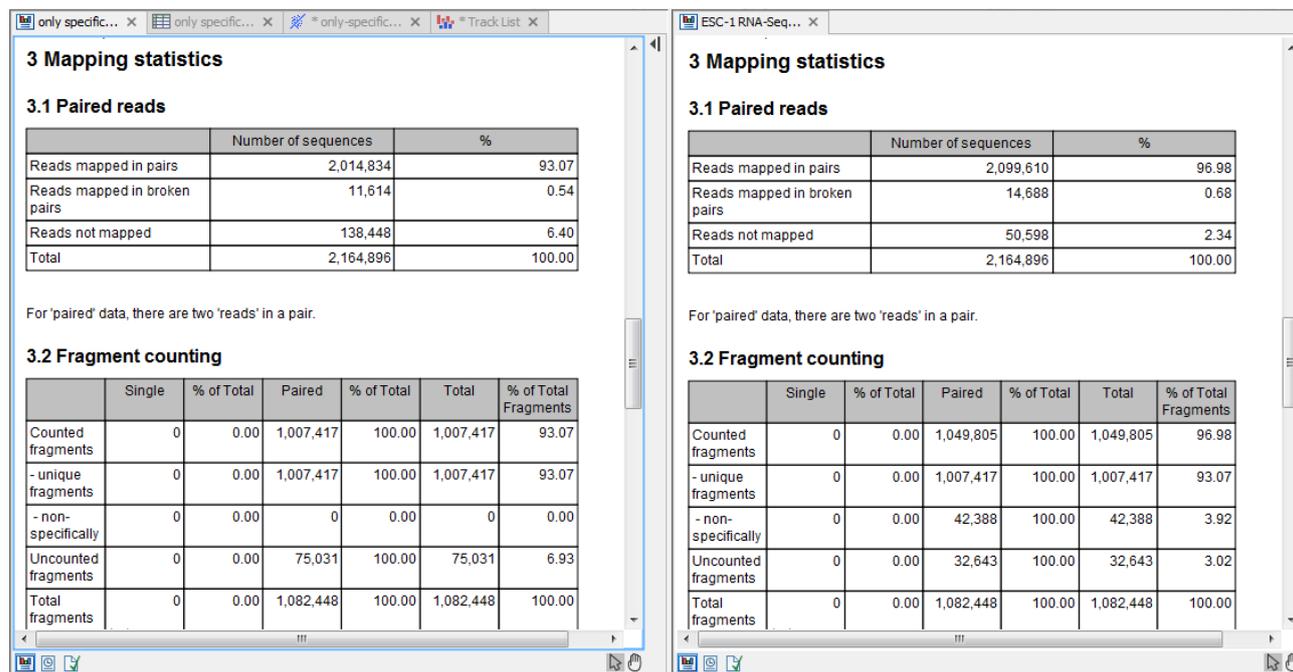


Figure 3: The RNA-seq report: 3 Mapping statistics.

The "3.1 Paired reads" tables in the reports tell you about the numbers of *reads* mapped in the analysis. We can see that when we only allowed specific matches we had 138,448 unmapped reads. When we allow non-specific matches this number is reduced by approximately 88,000 to 50,598. Most of the additionally mapped reads are mapped in pairs (around 85,000), whereas around 3000 of the additionally mapped were mapped in broken pairs. Table "3.2 Fragment counting" in the report talks about *fragments* as opposed to *reads*, so for paired data there are two reads for each fragment.

We cannot see from these numbers whether the additionally mapped reads were mapped to exonic or intronic regions. To see that, consider the table "3.3 Counted fragments by type". From these tables you can see that of the 42,388 non-specifically mapped *fragments* a number in the range of 24,748 were mapped to exonic regions, and about 17,640 to regions that were (at least partly) intronic. The numbers you get in your analysis may deviate a little from these numbers in this example for reasons that are discussed in the first tutorial in the series.

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 RNA-seq gene expression level tracks from your two analyses (figure 4) and click on the button labeled **Next**.
- Choose an un-paired, two-group experiment, choose to **Set new expression value** and choose **Genes: Total exon reads** (figure 5).

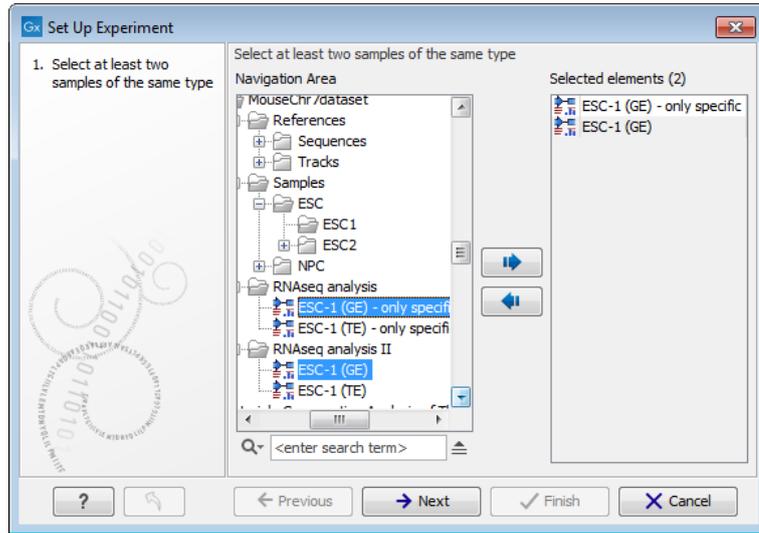


Figure 4: Select the RNA-seq expression level tracks.

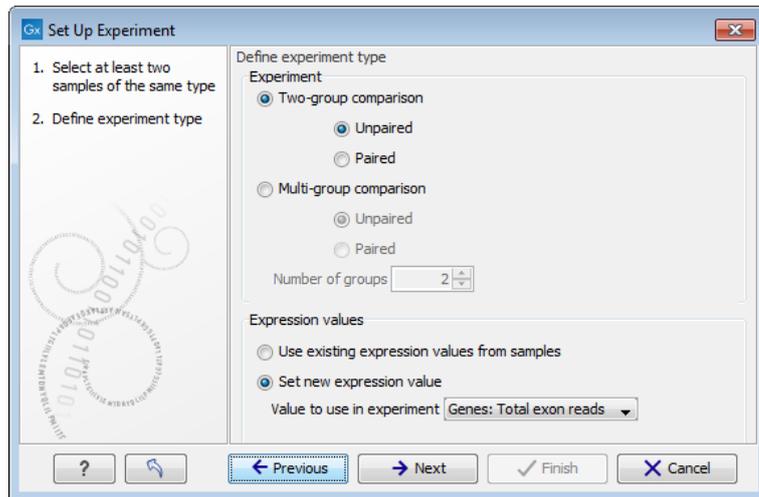


Figure 5: Adjust the parameters as shown in this figure.

7. Click on the button labeled **Next**. Name the groups *including non-specific* and *only specific* (figure 6) and click on the button labeled **Next**.



Figure 6: Rename the groups to be able to easily keep track of the two different analyses.

8. Right-click on each of the samples and assign them to the appropriate group (figure 7).

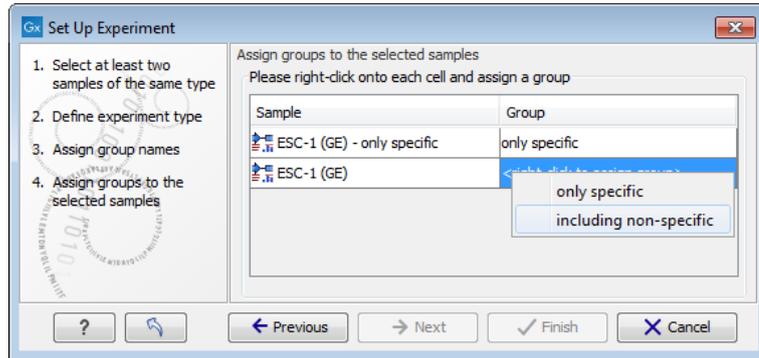


Figure 7: Assign the right groups to the samples by right-clicking on the group name to select the appropriate group.

9. Choose to open the results. Click on the button labeled **Finish**.

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. You will see a view similar to what is shown in figure 8. In figure 8 the diagonal line has been removed by unticking the "Draw x = y axis" box in the **Side Panel** (see the red arrow in the figure).

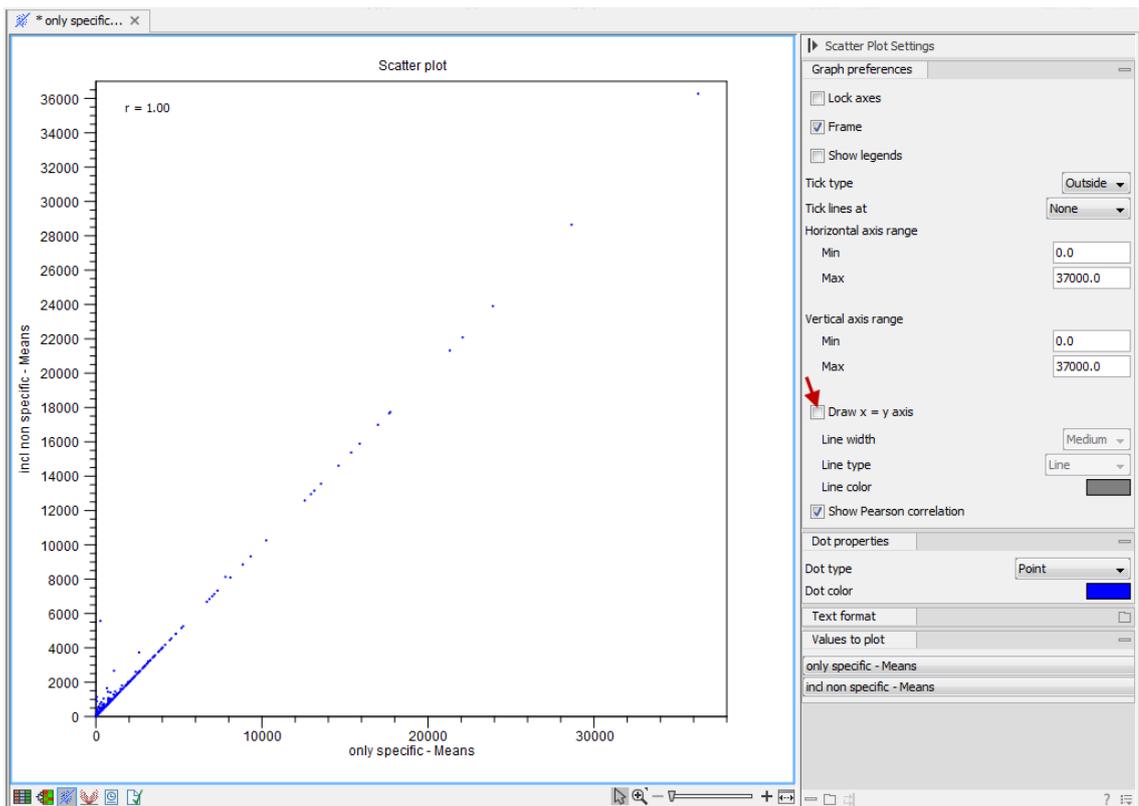


Figure 8: Comparing the "Total exon reads" counts for an analysis with only specific matches and one where non-specific matches are accepted.

The scatter plot now shows the expression levels in terms of "Total exon reads" count for the two samples. You can change the values in the "values to plot" settings of the scatter plot side panel. 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**.

If you place your mouse on the most outlying dot, the name of the gene for this point is shown and you can see that the dot is for the gene *Rps13*. We want to examine why we might get this huge difference in the expression level for this gene when we use different settings for the **Maximum number of hits** parameter. For this, create a track list that contains the gene expression level tracks (the tracks with extension "(GE)") and the reads tracks for your two analyses. (Remember that you can do that e.g. by selecting the tracks, right-clicking, choosing "New", and then "Track list"). Create a split view as shown in figure 9, e.g. by grabbing the track list by its pane and dragging it to the side of the viewer.

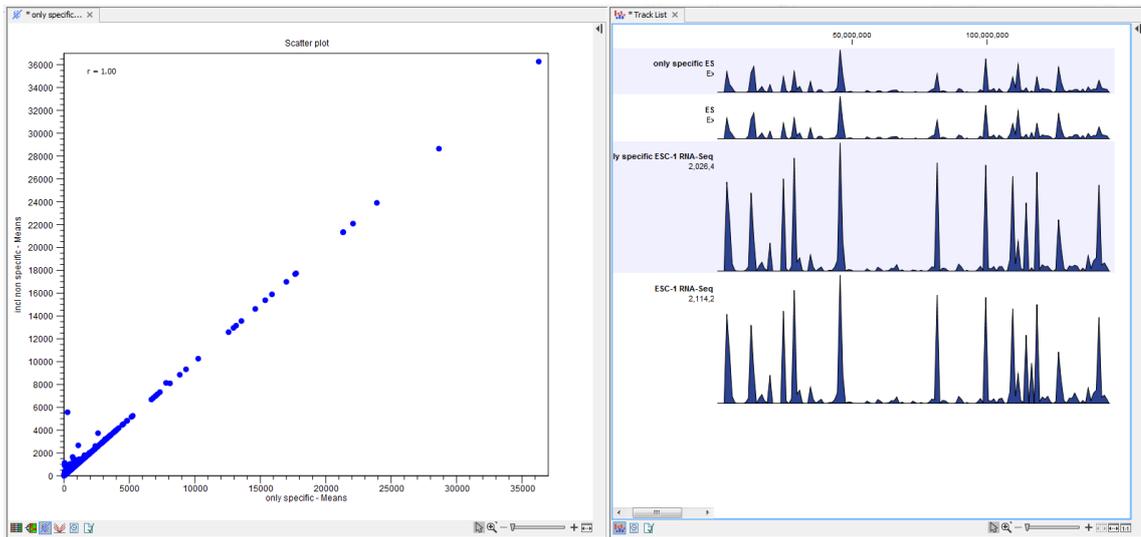


Figure 9: The experiment in scatter plot view shown next to the tracks list containing the gene expression level and RNA-seq reads tracks.

10. Now click on the dot for the gene *Rps13*. The dot will be "selected" (and become red) in the scatter plot. Moreover, because the gene expression level samples used in the experiment are also included in your track list, the region corresponding to the gene *Rps13* will become selected in the track list as well. In the bottom right corner of the track list editor click on the "zoom-to-selection" button (). You should get the view as seen in figure 10.

There are two striking differences between the reads tracks for the two analyses: (1) All reads in the reads track for the analysis that only allowed specific matches are blue. Since uniquely mapped reads are depicted in blue, this is expected. (2) Many more reads are mapped to this gene in the analysis that allowed non-specific matches (you can see that from the scale at the left of the reads track, which shows the maximum coverage in the region in the view. For the "only specific" sample this is around 300, whereas for the sample that allowed non-specific matches it is around 3,000). By far most of the reads that are mapped to the gene in the analysis that allowed non-specific matches are yellow. These are the non-uniquely mapped reads. Note that there are also blue reads in the second analysis. If you want to examine all the reads in the reads track you can expand the

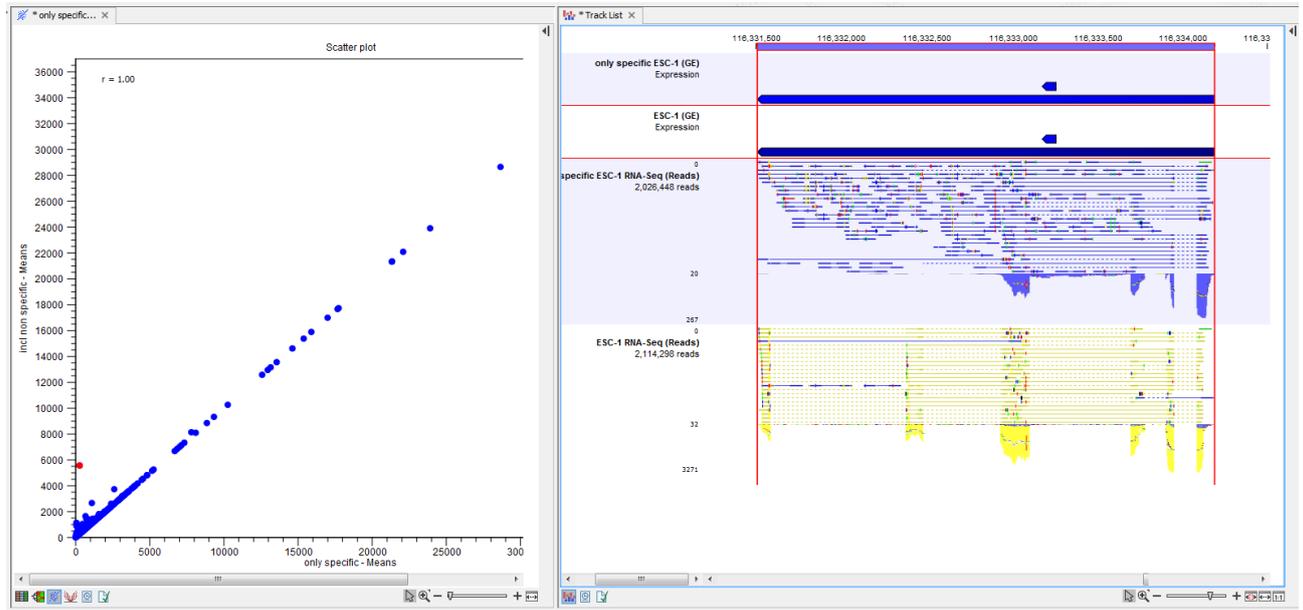


Figure 10: The dot in the scatter plot corresponding to the gene "Rps13" and the corresponding region in the trackm list "zoomed-to-selection".

area showing the visible reads by grabbing at the bottom of the track's editor and dragging it downwards. You will get a view similar to that in figure 11.

11. You may wonder why there are so many non specific matches for this gene. To understand this, go to the table view of your experiment by clicking on the table icon at the bottom of the experiment editor (see the red arrow in figure 12). Type the name of the gene, "Rps13" in the "Filter" box and press "Enter". You will see that two rows are displayed (figure 12); the row for the "Rps13" gene that we have been studying, but also a row for the gene "Rps13-ps2". This gene is a pseudo gene version of the "Rps13" gene.
12. Add the "Mus musculus_Gene" and "Mus musculus_mRNA" tracks to your track list by dragging them from the navigation area to the top of your track list.
13. Now click on the row for the "Rps13-ps2" gene in your experiment table. The displayed region in the track list will move to the area of this gene. You can see that the gene has a number of non specific matches (yellow reads) but no unique matches. As you can see there are no mRNA annotations for the pseudogene. So all reads that map across exon-exon junction for the "Rps13" gene are most likely uniquely assignable to the non-pseudogene version of the gene. The reads that map within a single exon which is also represented in the pseudogene will be non-specific. As non-specific matches are assigned proportionally to the number of unique match, most of them are mapped to the "Rps13" gene and not the "Rps13-ps2" gene (note that for genes with no unique matches we add a pseudo read count of 1 before distributing the non-specific matches, so that there is a chance of assigning non-specific matches to them).
14. Go back to the scatter plot view of your experiment. There is another gene whose "Total exon reads" count is different in the two analysis, and for a different reason. The count is around 3,700 when you allow non-specific matches and only 2,600 when you only consider unique matches (figure 13).

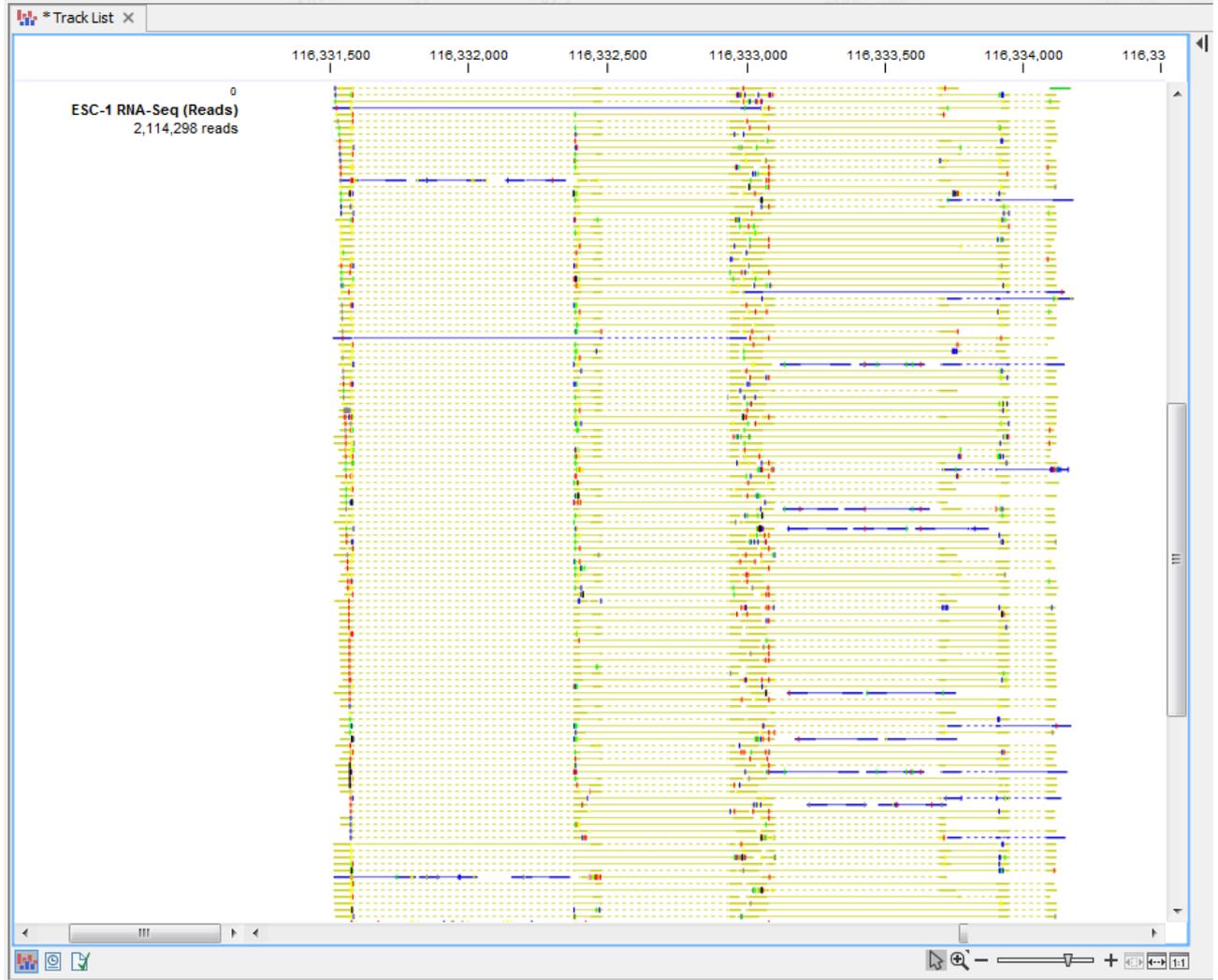


Figure 11: The reads track for the sample that allowed non-specific matches.

This is the "Ctbp2" gene. Click on the dot for this gene so that it gets selected in the experiment and track list. Double-click on the track list tab to maximize the track list view. As you can see, there are two overlapping genes in the area of this gene (figure 14). If you mouse-over the other gene you will see that the overlapping gene is the "Fgfr2" gene. Reads that map within a region that is exonic to both genes can not be assigned uniquely to one of the genes, so they are considered non-specific matches at the gene level, although they may map uniquely to this region in the reference genome. There is an exception to this: if the genes are on different strands and you have used a strand-specific protocol to generate your reads, you can specify which strand that you have been using in your protocol in the **strand specific** option of the third wizard step. When this is specified, reads will only be attempted mapped to the specified strand.

The default parameter settings when you run an RNA-seq analysis is to allow for non-specific matches. The problem, if you do not allow for them, is that you will get biased values for genes that belong to gene families that have pseudogenes or overlapping genes. You can use the "Ratio of unique to total exon reads" column in the expression track to help you see if the expression levels reported for your gene should be taken with some care.

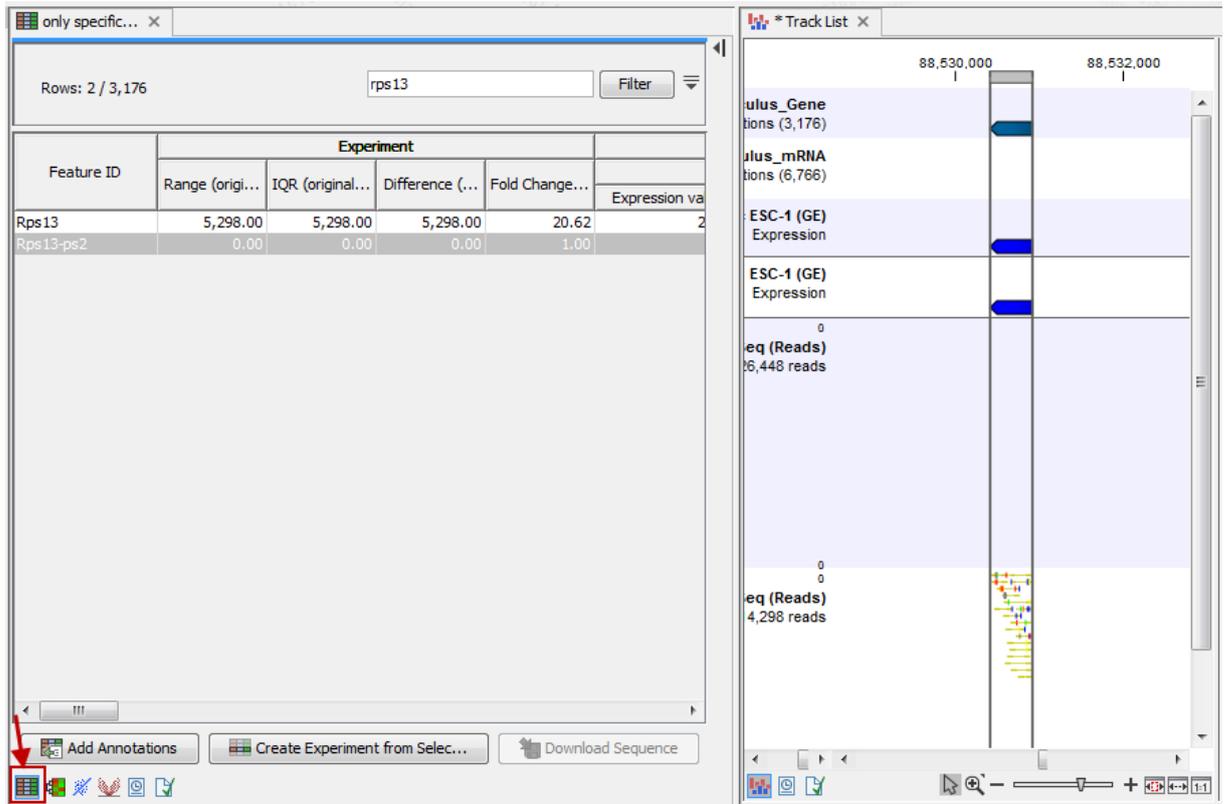


Figure 12: The "Rps13-ps2" pseudogene.

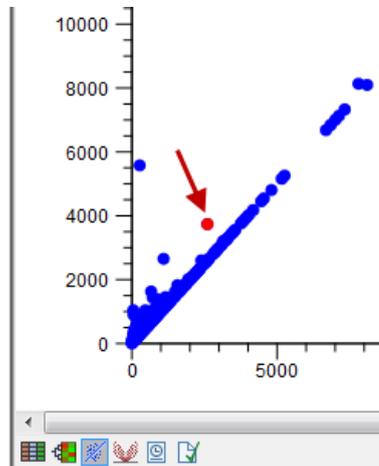


Figure 13: Find the "Ctbp2" gene (red dot).

Running the same data set with the "Also map to inter-genic regions" option

15. We will now explore the differences between results obtained when running with the **Map to gene regions only (fast)** option and the **Also map to inter-genic regions** option: Again run an RNA-seq analysis of the "ESC-1" sample, with the same parameters as in your last analysis (that is, with the **Maximum number of hits for a read** parameter set to 10) and this time choose **Also map to inter-genic regions** in wizard step 2. Save the results and rename them with the prefix "inter-genic". Also, give the names of the results from your last analysis the prefix "genes only".

16. Go to the toolbox and use the "Set Up Experiment" tool to create an experiment with

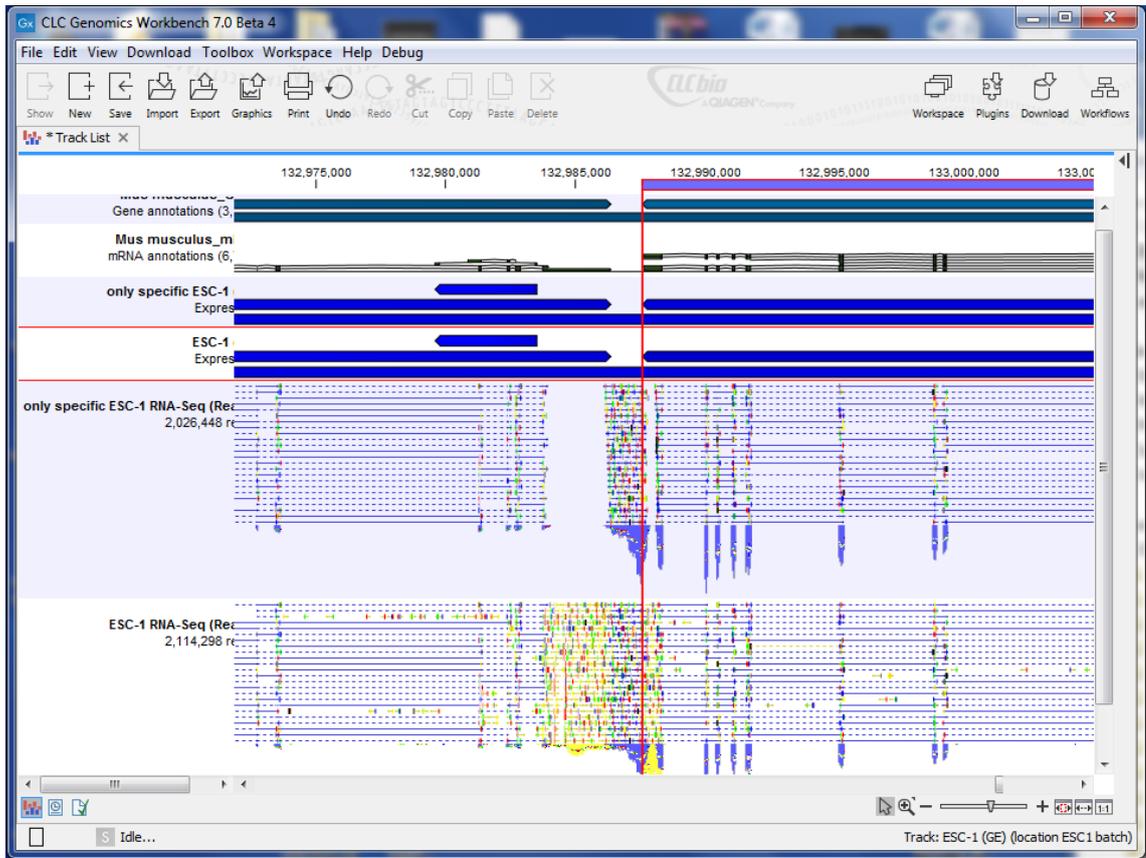


Figure 14: The "Ctbp2" gene overlaps with the "Fgfr2" gene.

the two groups called "Genes only" and "Inter-genic" from your two gene expression level tracks (the "genes only ESC-1 (GE)" track and the "inter-genic ESC-1 (GE)" track) as shown in figure 15.

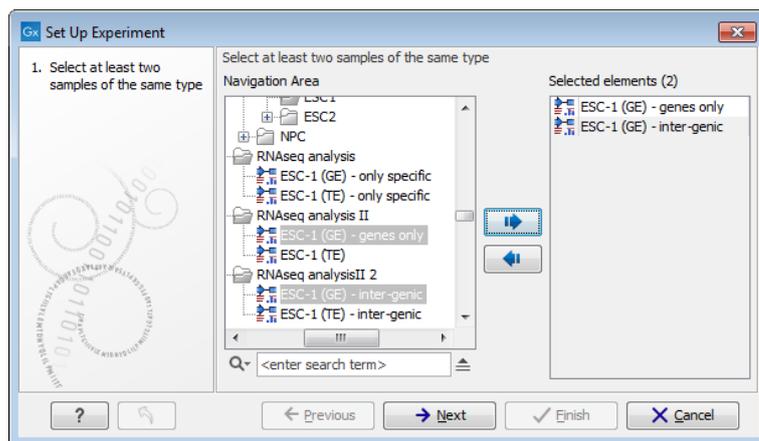


Figure 15: Create an experiment with the two gene expression tracks as shown in the figure.

Go to the scatter plot view of the experiment. In the side panel, choose "Dot". As you can see the expression levels are highly similar. However, if you zoom in at the low level values you will see that there are some differences (figure 16). Let your mouse hover over the outliers to let the names of the genes be displayed. One of the outlying dots is for the "Zscan4f" gene.

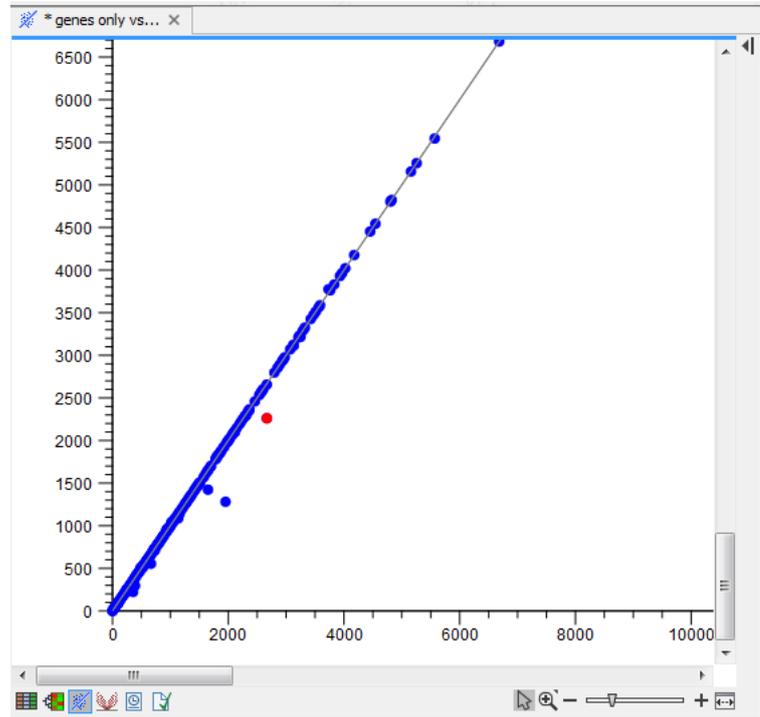


Figure 16: The expression level for the "Zscan4f" gene is higher when the "Map to gene regions only" option is used than with the "Also map to inter-genic regions" option.

17. Add your gene level expression and reads track from the inter-genic analysis ("inter-genic ESC-1 (GE)" and "inter-genic ESC-1 RNA-seq (reads)") to your track list by dragging them in.
18. Remove the "only specific ESC-1 (GE)" and "only specific ESC-1 RNA-seq (reads)" tracks from the track list by right-clicking on their names in the track list and choosing "Remove track".
19. Now click on the dot for the "Zscan4f" gene and the "zoom-to-selection" button (🔍). The track list will move to and zoom in on the area of this gene (Figure 17):
20. Now open the gene expression level tracks "genes only ESC-1 (GE)" and "Inter-genic ESC-1 (GE)" from the track list one at a time by double-clicking where their names are displayed in the track list. You want to arrange them as shown in figure 18. You do that by first dragging them in place so that the two tracks are shown one above the other (you can close all other tracks). "Filter" in the two track tables to see the rows for the "Zscan4f" gene. In the side panel of the gene expression level tracks, "Deselect all" columns. Then select the columns for unique and total gene reads, and the unique and total exon reads.

As you can see the numbers of unique and total gene and exon reads are higher when the inter-genic regions are ignored in the mapping than when they are included. This will happen if an acceptable match can be found for a read to a genic region, although there is a better match to an inter-genic region, when the inter-genic regions are ignored. It is our experience that, although there is this difference for a few genes, the difference is quite minor.



Figure 17: The "Zscan4f" gene. More reads are mapped to this gene when inter-genic regions are ignored than when they are included.

The RPKM expression measure

So far in these tutorials we have been working exclusively with the "Total exon reads" expression measure. You will often be in a situation where you have different sampling depths (different total numbers of reads) in your samples. In this case it might be problematic to just consider total counts. Similarly, since genes have different lengths, it is problematic to compare expression levels between different genes by looking at total counts: with the same level of expression a gene that is twice the length would also have twice the number of reads mapping to it.

21. For these reasons a different measure of expression, the RPKM value (see the first tutorial on RNA-seq analysis for a precise definition) has been proposed. The RPKM value normalizes the total exon reads counts by the total number of reads mapped for the sample and the total exon length of the gene. This gives a quite different expression measure. To see how different, close all other results and open your "genes only vs inter-genic" experiment. Go to the scatter-plot view and in the side panels "Values to plot" options choose "inter-genic ESC-1 (GE) Total exon reads" and "inter-genic ESC-1 (GE) RPKM". You will get a plot as shown in figure 19.

Which expression measure to use has been much debated and there is no simple answer; it depends upon your application. Some consensus seems to have been reached when

The image displays two screenshots of a software interface showing read counts for the Zscan4f gene. The top screenshot is titled 'genes-only ES...' and the bottom is 'inter-genic E...'. Both show a table with columns for Name, Unique gene reads, Total gene reads, Unique exon reads, and Total exon reads. The 'inter-genic E...' track shows higher counts for unique and total gene reads compared to the 'genes-only ES...' track.

Name	Unique gene reads	Total gene reads	Unique exon reads	Total exon reads
Zscan4f	1189	2834	1092	2667

Name	Unique gene reads	Total gene reads	Unique exon reads	Total exon reads
Zscan4f	1042	2389	959	2261

Figure 18: Counts for the "Zscan4f" gene when run with the Map to gene regions only (fast) and the Also map to inter-genic regions option.

it comes to statistical analysis of differential expression between two or more biological conditions: for such applications using the "Total exon reads" rather than the RPKM value, *together with* statistical tools particularly designed for count based analysis, seems to be generally most reliable. We will work with statistical analysis of differential expression in the third tutorial in this series of tutorials on RNA-seq analysis. However, when it comes to analyses that require that we compare expression levels for *different genes against each other* (rather than the *same gene across samples*), normalization for gene length and sampling depths may be required. For such applications the RPKM value may be more suitable than e.g. the "Total exon reads" count.

This concludes the second tutorial on RNA-seq analysis.

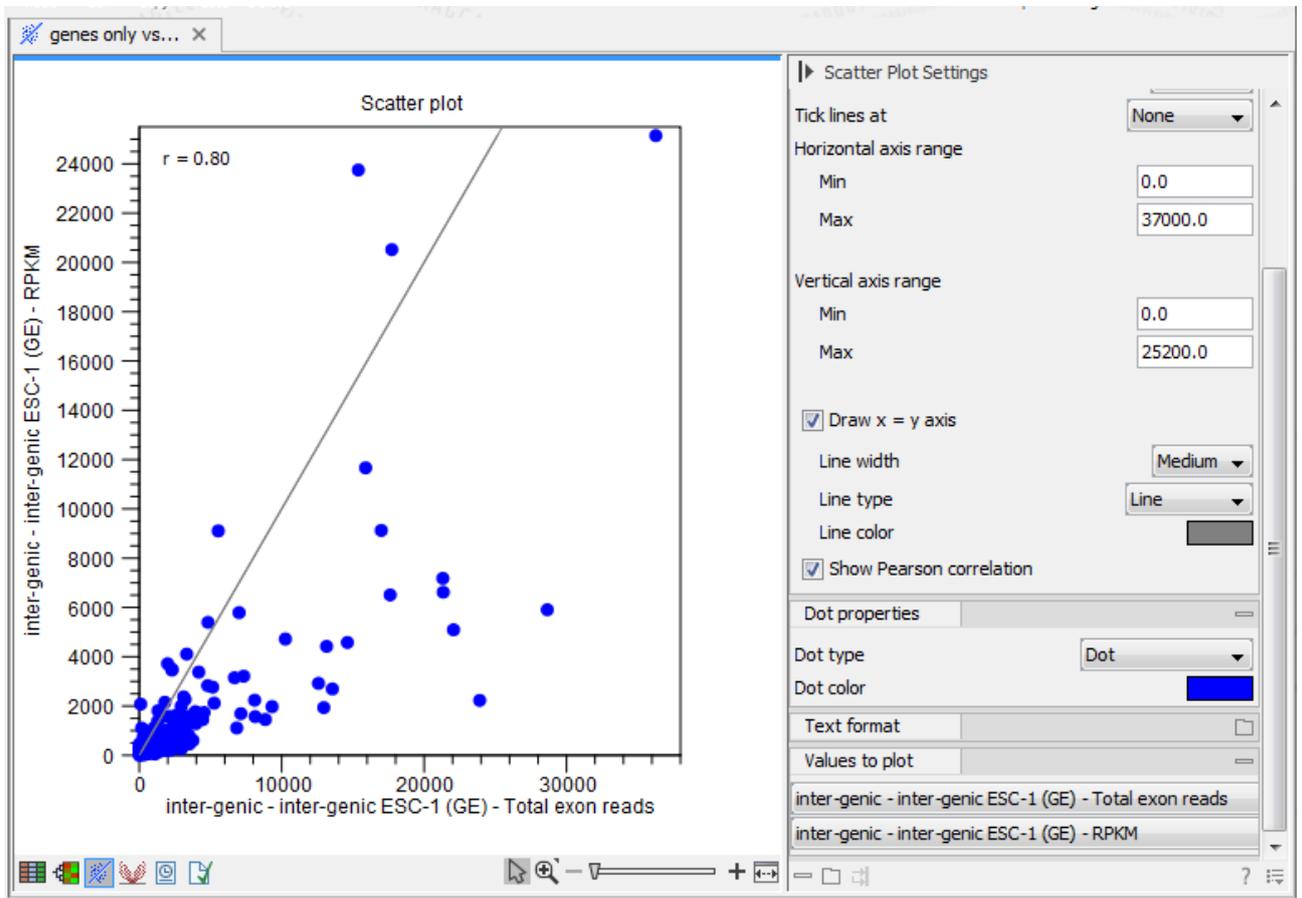


Figure 19: The "Total exon reads" versus the "RPKM" expression measure.