



# Tutorial

## RNA-Seq analysis with four tissues and six time-points

December 3, 2024

— Sample to Insight —

## RNA-Seq analysis with four tissues and six timepoints

### Introduction

The purpose of this tutorial is to illustrate how to analyze RNA-Seq data for multiple groups of samples and timepoints using *CLC Genomics Workbench*. We focus on the following types of analysis:

- Import and set up of data.
- Investigate outliers and structure in the data using a PCA plot and heat map.
- Identify differentially expressed genes and visualize them using a Venn diagram and heat map.
- Cluster differentially expressed genes according to the expression pattern across the timepoints.
- Detailed investigation of expression patterns of genes of interest.

### Data used in this tutorial

This tutorial uses the data set from "An Examination of Dynamic Gene Expression Changes in the Mouse Brain During Pregnancy and the Postpartum Period" by Ray et al 2016 ([GSE70732](#), [doi:10.1534/g3.115.020982](#)). The authors investigated developmental transformation of the female brain during pregnancy, parturition and postpartum, looking at four regions (cerebellum, hippocampus, hypothalamus and neocortex). The experimental setup included six timepoints corresponding to the following developmental stages:

- **Virgin:** Female mouse - unmated.
- **PC14** and **PC16:** Dam - 14 and 16 days post-conception, respectively.
- **PP1, PP3** and **PP10:** Dam - 1, 3, and 10 days postpartum, respectively.

Two or 3 samples were collected from each brain region at each stage, yielding a total of 71 samples.

We provide gene expression data for this tutorial, generated with single-end RNA-Seq reads downloaded from SRA and then analyzed using *RNA-Seq Analysis* using a mouse reference genome. Descriptive information about each sample is provided in a metadata table.

## Prerequisites



For this tutorial, you must be working with *CLC Genomics Workbench* 25.0 or higher. Note that versions higher than 25.0 may produce slightly different results than those shown here.

## General tips

- You can access the in-built manual by clicking the **Help** button on wizards or by selecting the **Help** option under the **Help** menu.
- Within wizards you can use the **Reset** button to change settings to their default values.
- Colors in plots can be changed by clicking the colored rectangles found in the [Side Panel](#).
- Columns in tables can be hidden by unticking their name in the [Side Panel](#).
- Columns in tables can be used to sort the rows, by successively clicking on the column name until the desired order (indicated by an arrow next to the column name) is achieved.
- Most of the tools of *CLC Genomics Workbench* require multiple inputs. When many data elements need to be selected, as it is the case in this tutorial, all elements located under a folder can be added by using the options **Add folder contents** or **Add folder contents (recursively)** found in the right-click menu.
- Many data elements produced by *CLC Genomics Workbench* tools have multiple views, indicated as icons in the lower left corner in the open files in the [View Area](#). Clicking on one of the view icons while pressing Ctrl (Cmd on Mac) key will open in split view such that both views are visible at the same time. Often, if viewing a table and a graphical representation in split view, selecting entries in the table will highlight them in the graphical representation. The order of the views can be changed using drag and drop, see [Arrange views in View Area](#).

## Import gene expressions and metadata

We start by downloading and importing the tutorial data.

1. Download the [data archive](#) containing gene expression data for each sample and a metadata table containing descriptive information about each sample.
2. Start the QIAGEN CLC Genomics Workbench.
3. Import the downloaded archive using [Standard Import](#).
  - (a) Start **Standard Import** from the top toolbar:  
**Import**  | **Standard Import** 
  - (b) Locate the archive using the **Add files** button and select **Automatic import** (figure 1).

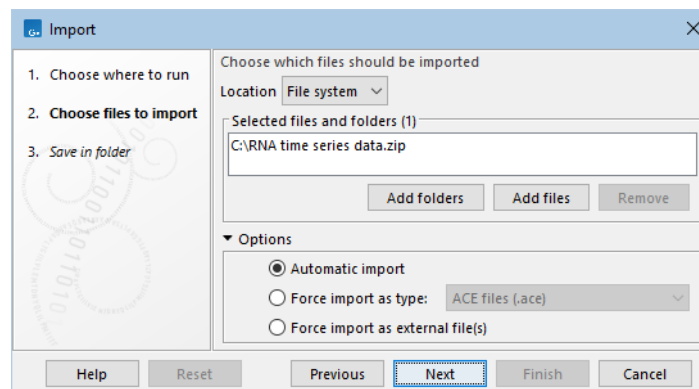


Figure 1: *Standard Import* configured to automatically import the zipped data.

- (c) Save the files in a folder of your choice.

The data needed for this tutorial is now imported: 71 gene expression tracks and the metadata table. Each track is associated with the corresponding row in the metadata table. See [Associating data elements with metadata](#) for information about how to create such associations.

The tools of *CLC Genomics Workbench* produce deterministic results, but changing the inputs order can lead to slightly different results. In this tutorial, the inputs are sorted alphabetically. This can be achieved by using the "Sort Folder" option from the right-click menu on folders in the [Navigation Area](#).

## Investigate outliers and structure in the data

In this section, we create a PCA plot and a heat map, which we use to visually inspect the full data set, to help identify any outliers among the samples and any interesting structure in the data.

### PCA plot

Using [PCA for RNA-Seq](#), we will first produce a PCA plot.

1. Start **PCA for RNA-Seq** by going to:

**Tools | RNA-Seq and Small RNA Analysis** (📁) | **Expression Plots** (📊) | **PCA for RNA-Seq** (📊)

2. Select the 71 gene expression tracks and choose to save the result in a folder named Plots.
3. Open the resulting PCA plot.
4. Configure the labels and legends from the Metadata tab in the **Side Panel** (figure 2):

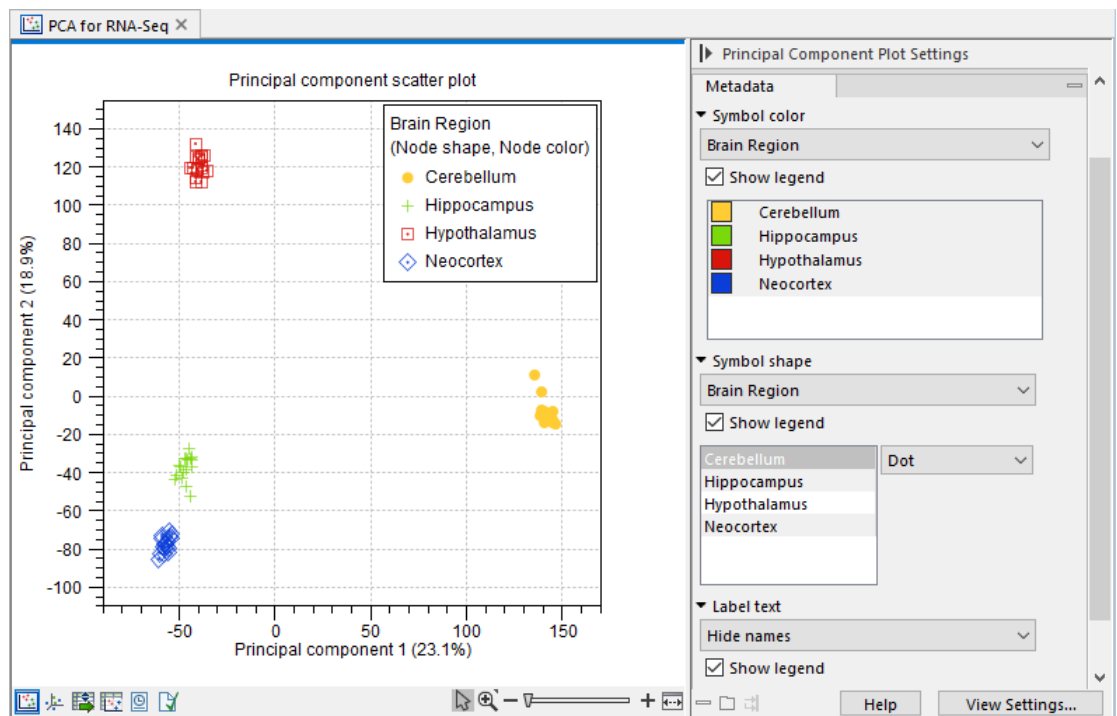


Figure 2: PCA plot shows the samples cluster into groups defined by the brain region.

- Under "Metadata", set "Symbol color" and "Symbol shape" to "Brain Region" from the drop-down menus.
- Under "Label text", choose "Hide names" in the drop-down menu.

As it can be seen in figure 2, each of the brain regions clusters together. There are no outliers. We therefore will use all the samples in the downstream analyses.

## Heat map

Using **Create Feature Level Heat Map for RNA-Seq**, we will produce a heat map of gene expressions for the genes with the most varied expression levels, that is, the genes with highest coefficients of variation (the ratio of the standard deviation to the mean).

1. Start **Create Feature Level Heat Map for RNA-Seq** by going to:

**Tools | RNA-Seq and Small RNA Analysis** (📁) | **Expression Plots** (📊) | **Create Feature Level Heat Map for RNA-Seq** (📊)

2. Select the 71 gene expression tracks as input.
3. Keep the default values in the "Distances" wizard step.
4. In the "Feature filters" wizard step, specify 50 for the "Fixed number of features". Keep the default values for the remaining options (figure 3).

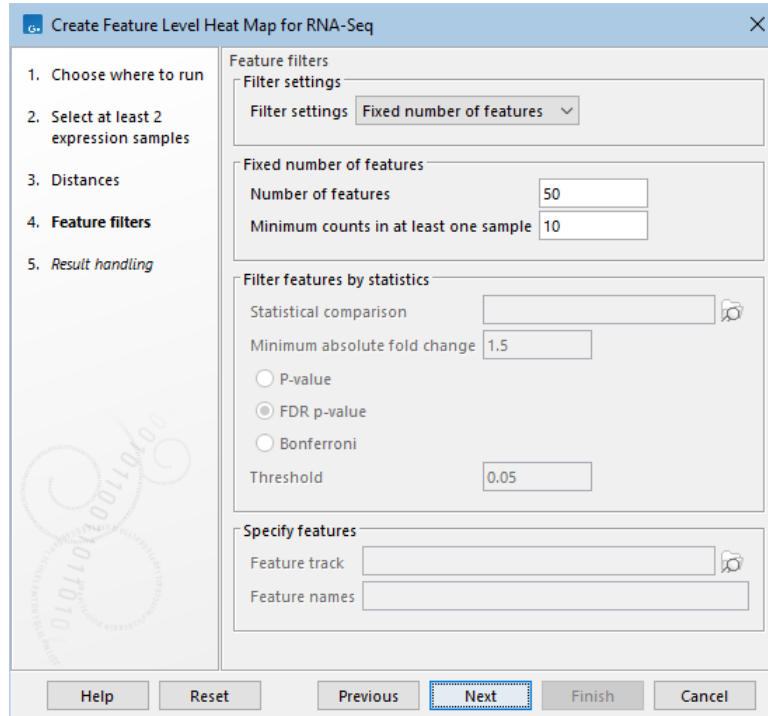


Figure 3: Set the "Fixed number of features" value to 50 in the "Feature filters" wizard of "Create Feature Level Heat Map for RNA-Seq".

5. Choose to save the results to the Plots folder.
6. Open the resulting heat map.

In the following steps, we use settings in the **Side Panel** to group, sort and label samples in the heat map. We first group the samples based on brain region, and afterwards, we group them based on both brain region and stage.

7. To group the samples by brain region (figure 4):
  - Under "Samples", untick "Show names above" and choose "Brain Region" from the "Order by" drop-down menu.
  - Under "Features", untick "Show names left".
  - Under "Metadata", choose "Brain Region" in the "Metadata layer #1" drop-down menu.
  - Drag and drop the brain regions listed under "Metadata layer #1" such that they are in the following order: Cerebellum, Hippocampus, Neocortex and Hypothalamus. This ordering results in the genes with high expression (red) lying along the diagonal, which can make interpretation easier.

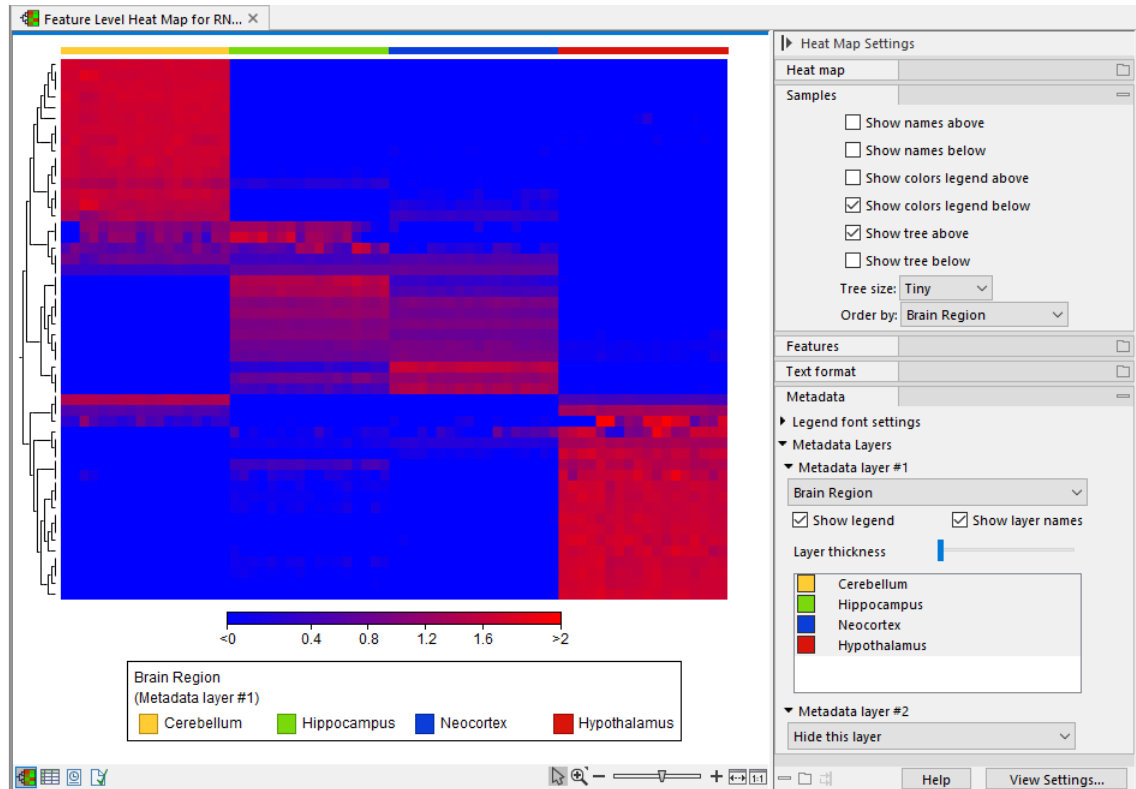


Figure 4: Heat map showing the 50 genes with the highest coefficients of variation. Samples are grouped by brain region, and brain regions are sorted such that genes with high expression are placed along the diagonal.

- To have the updated view settings applied each time this heat map is opened:
  - (a) Click on the **View Settings...** button in the bottom, right corner.
  - (b) Click **Save View Settings....**
  - (c) Provide a **View settings name**.
  - (d) Select **Save for this element only**, and click **OK**.
  - (e) Save the heat map (shortcut Ctrl+S or Cmd+S on Mac).

The view settings are now saved and used for this heat map only. See [View settings for the Side Panel](#) for more details.

8. To group the samples by both stage and brain region (figure 5):

- Select "Stage" in the "Order by" and "Metadata layer #1" drop-down menus, and add the "Brain Region" as a "Metadata layer #2".
- Drag and drop the stages listed under "Metadata layer #1" such that they are in chronological order: Virgin, PC14, PC16, PP1, PP3 and PP10.

## Identify differentially expressed genes

Figures 4 and 5 show that the brain region has a stronger effect on expression patterns than the developmental stage. To identify the genes differentially expressed in at least one

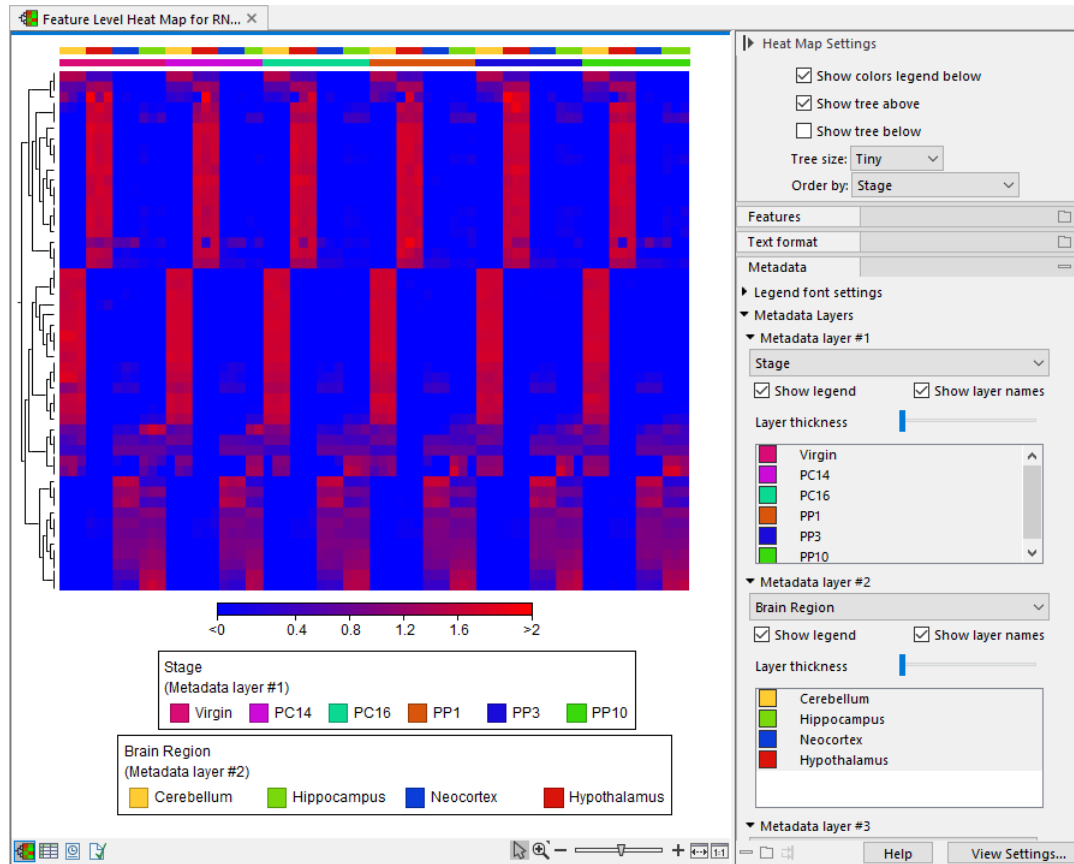


Figure 5: Heat map showing the 50 genes with the highest coefficients of variation. Samples are grouped by stage and brain region, and stages are sorted to preserve their chronological order.

developmental stage, we use **Differential Expression for RNA-Seq**, analyzing each brain region separately.

As seen earlier, data elements can be selected directly in the launch wizard of a tool. However, elements already selected elsewhere in the *CLC Genomics Workbench* are preselected as input in launch wizards. Here, we illustrate this feature by selecting gene expression tracks for hippocampus samples via their associations to the "All samples in project" **CLC Metadata Table**. These preselected elements are then selected for us in the the **Differential Expression for RNA-Seq** launch wizard.

1. Open the "All samples in project" metadata table.
2. Right-click on the "Brain Region" column of any row from the hippocampus, choose **Table filters** and click on **Brain Region = Hippocampus** (figure 6).

Only rows where the value in the "Brain Region" column is set to "Hippocampus" are now shown in the table.

Data originating from hippocampus samples have associations with these rows. To find these data elements:

3. Select all the rows in the table (shortcut Ctrl+A or Cmd+A on Mac) and click on the **Find Associated Data** button.



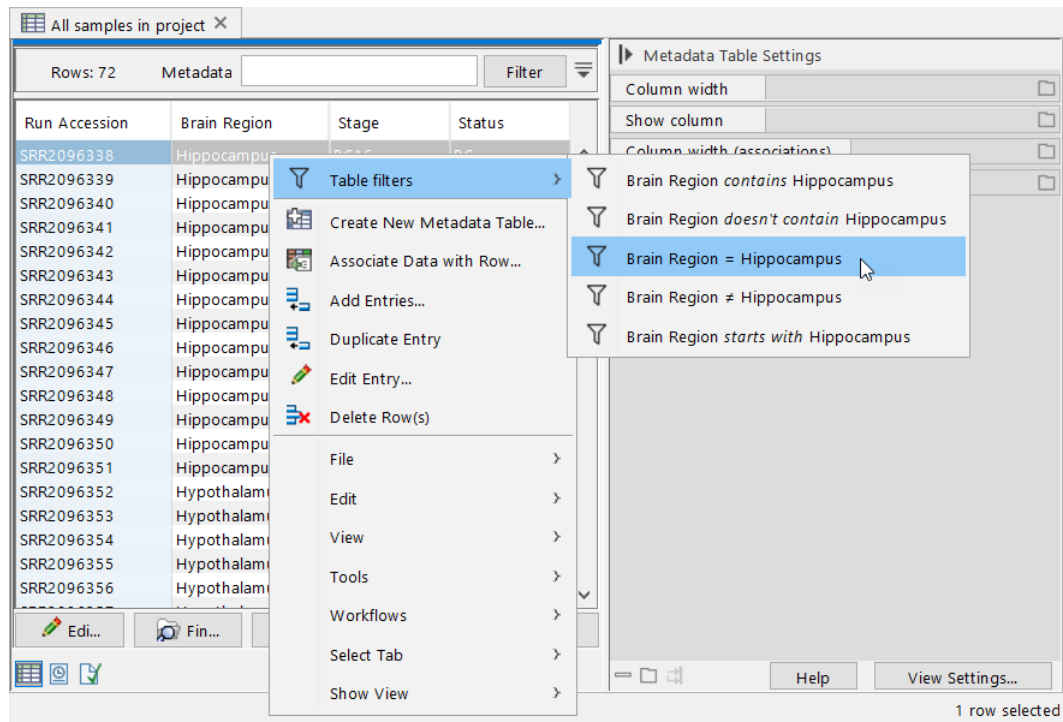
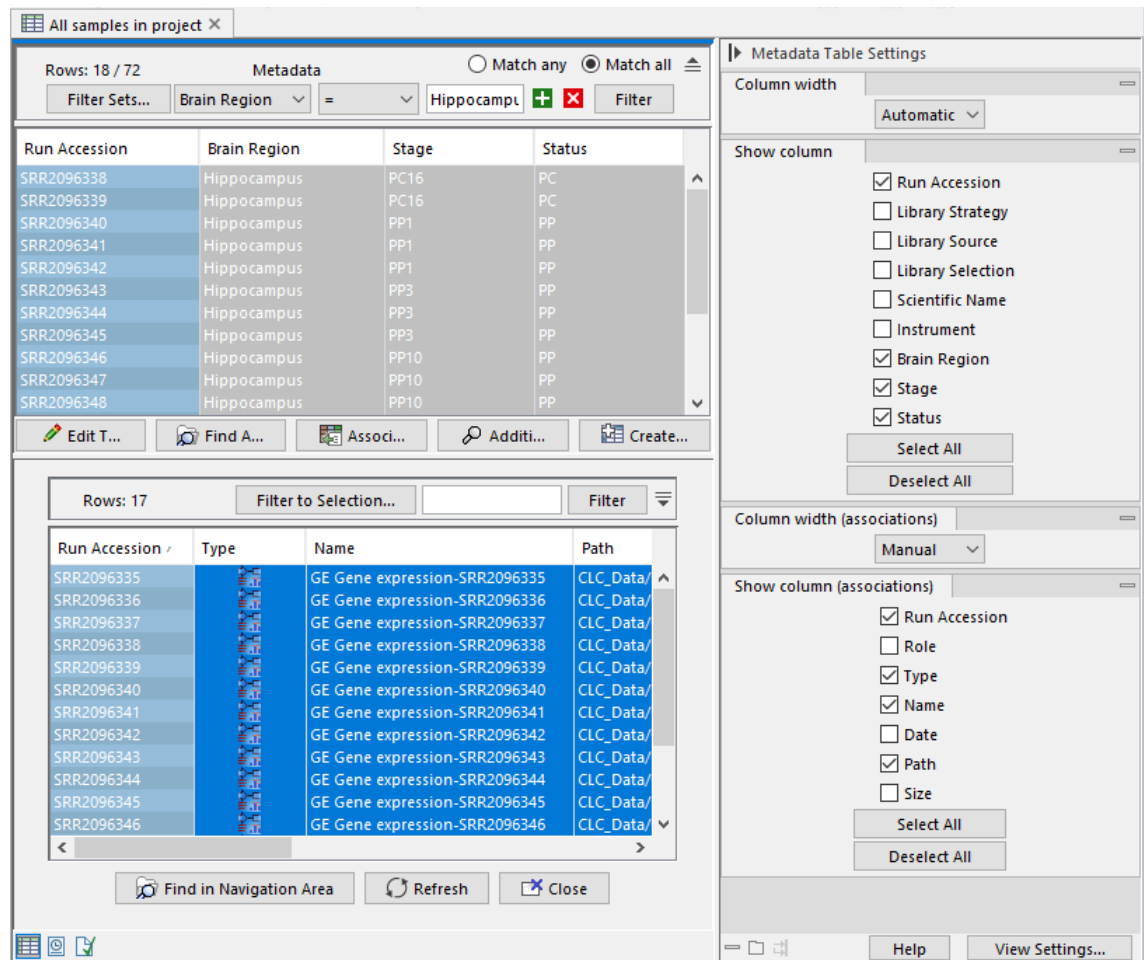


Figure 6: Filtering the table using the "Table filters" options.

4. In the resulting table, sort the samples by "Run Accession" by clicking on the column header.
5. Select the sorted samples (figure 7).
6. Start **Differential Expression for RNA-Seq** by going to:  
**Tools | RNA-Seq and Small RNA Analysis (🇺🇸) | Differential Expression (🇩🇪) | Differential Expression for RNA-Seq (🇩🇪)**
7. The gene expression tracks for hippocampus samples, selected earlier, are preselected for use when launching the tool.
8. In the "Configure normalization method" wizard step, keep the default settings.
9. In the "Experimental design and comparisons" wizard step, select the "All samples in project" metadata table, choose "Stage" in "Test differential expression due to", and select "Across groups (ANOVA-like)" under "Comparisons" (figure 8).
10. Keep the default values for the remaining options.
11. Choose to save the results in a folder named DE. Create a subfolder to hold results for each brain region.
12. Rename the results to reflect the brain region (e.g., "Hippocampus across all stages") by right-clicking on the results in the **Navigation Area** and using the **Rename** option (shortcut F2).

These element names are used in plots generated in downstream analysis.



The screenshot displays the CLC Metadata Table interface. The main table shows 18 rows of data with columns: Run Accession, Brain Region, Stage, and Status. The 'Brain Region' column is filtered to 'Hippocampus'. The 'Metadata Table Settings' panel on the right shows the 'Show column' list with 'Run Accession', 'Brain Region', 'Stage', and 'Status' selected. The 'Column width' is set to 'Automatic'.

Figure 7: Selecting data elements based on information in a CLC Metadata Table.

- Repeat steps 2-12 for the three remaining brain regions, cerebellum, hypothalamus and neocortex.

You should now have four statistical comparison tracks, one for each brain region, where differential expression was tested across all developmental stages.

### Investigating the differential expression results

We will use a Venn diagram to find genes that are differentially expressed in all brain regions, in at least one developmental stage.

- Start **Create Venn Diagram for RNA-Seq** by going to:

**Tools | RNA-Seq and Small RNA Analysis (📁) | Differential Expression (🔍) | Create Venn Diagram for RNA-Seq (📊)**

- Select the four statistical comparison tracks generated in the previous section (figure 9).
- Choose to save the result in the Plots folder.

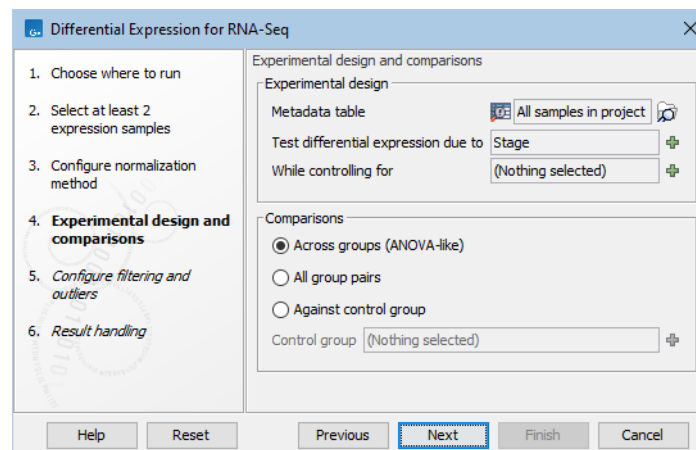


Figure 8: Differential expression will be tested across groups based on developmental stage.

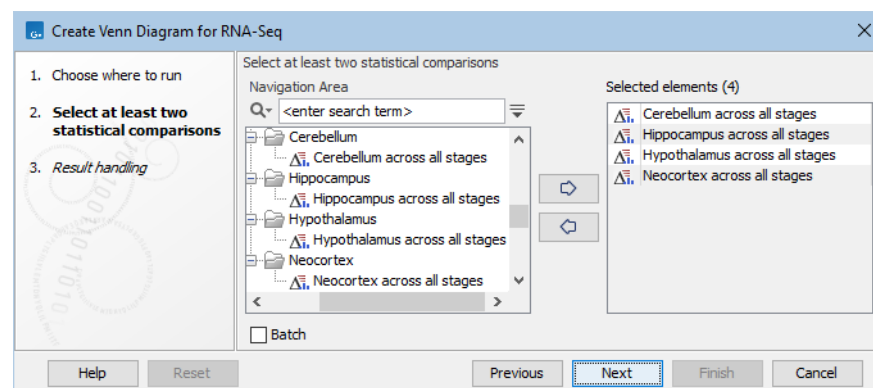



Figure 9: Creating a Venn diagram using the four statistical comparison tracks.

4. Open the resulting Venn diagram.
5. Set the "Min. absolute fold change" in the **Side Panel** to 1.  
The plot now shows that 40 genes are differentially expressed in all four brain regions (figure 10).
6. To investigate these 40 genes, open the Table view in a split view by clicking on the table icon (  ) at the bottom of the view while pressing Ctrl (Cmd on Mac) key.
7. In the Venn diagram, click on the intersection of all four brain regions.  
These 40 genes are now highlighted in the table.
8. To see just the highlighted rows in the table, click on **Filter to Selection** and choose **Filter to selected rows** (figure 10).
9. Select all 40 genes (shortcut Ctrl+A or Cmd+A on Mac).
10. Click on **Copy Gene Names to Clipboard**.
11. Paste the gene names to a text file to easily recover them later.  
We will use these 40 genes multiple times in this tutorial.

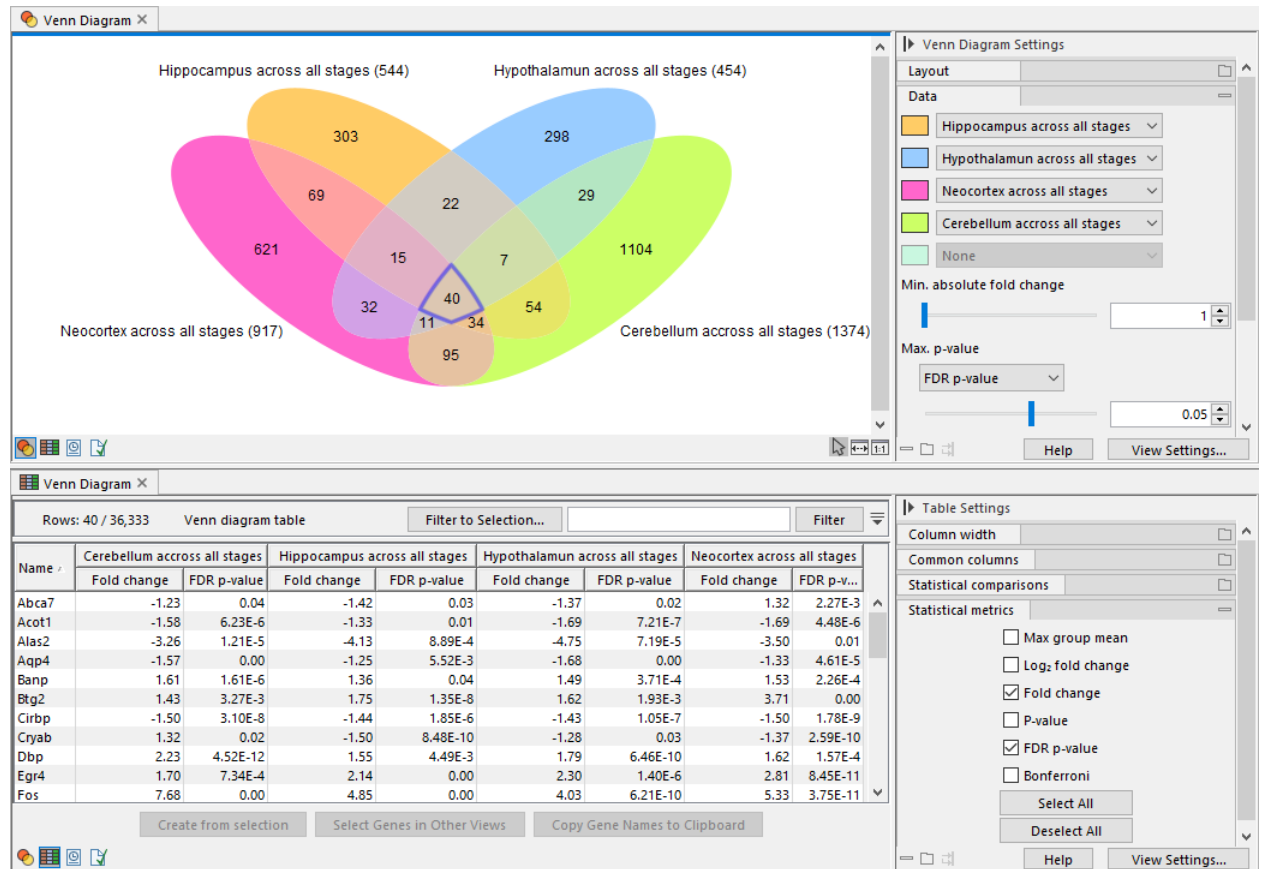


Figure 10: Top: Venn diagram showing the overlap of genes identified as differentially expressed in each brain region. The intersection of all regions, containing 40 genes, is selected. Bottom: Table view of the Venn diagram, filtered to show just the rows for only those 40 genes.

12. Start **Create Feature Level Heat Map for RNA-Seq** by going to:

**Tools | RNA-Seq and Small RNA Analysis** | **Expression Plots** | **Create Feature Level Heat Map for RNA-Seq**

13. Select the 71 gene expression tracks as input.

14. In the "Feature filters" wizard step, set "Filter settings" to "Specify features" and paste the copied gene names to the "Feature names" text field. This will create a heat map for the 40 selected genes from the Venn diagram.

15. Choose to save the result in the Plots folder.

16. Rename the resulting plot to "Heat Map for expression across all stages".

17. Open the heat map.

18. Under "Samples", untick "Show names above" and choose "Active metadata layers" from the "Order by" drop-down menu.

19. Under "Metadata", choose "Brain Region" and "Stage" for layers #1 and #2. Reorder the brain regions and stages as shown in figure 11.

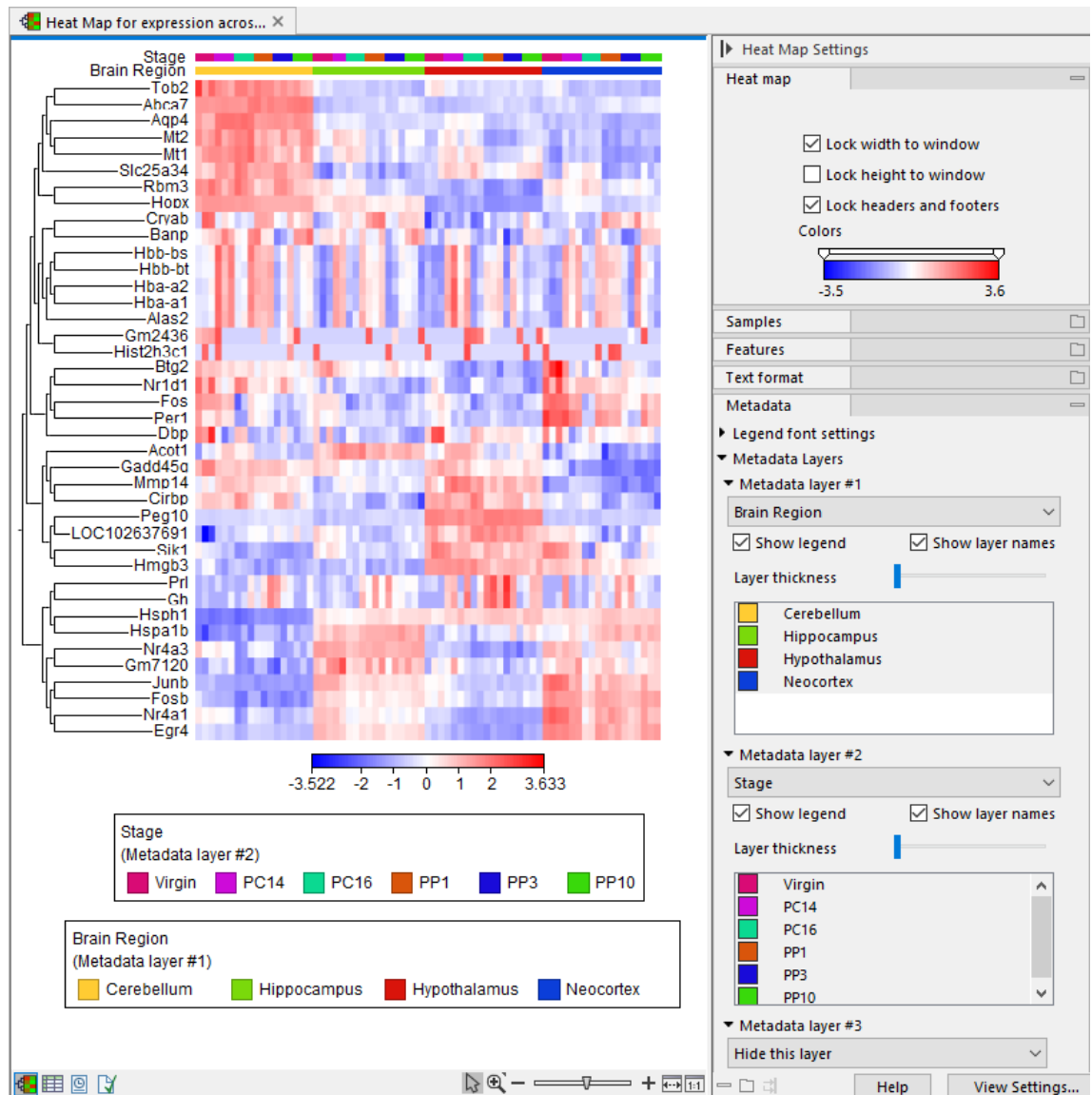


Figure 11: A heat map of the expression values for the 40 genes differentially expressed in at least one developmental stage in all four brain regions.

20. In the **Side Panel**, under "Heat map", click on the gradient and choose a gradient that starts at blue, goes through white and ends with red. Adjust the ranges of the gradient to -3.5 and 3.6, using the slider above the gradient.

These settings allow us to easily spot the over- and under-expressed genes. The chosen gradient has white for 0 values and the heat map normalizes gene expression such that the average expression is 0.

Figure 11 shows that the differentially expressed genes do not have the same expression pattern in all brain regions. For example, the first group of genes at the top of the heat map have high expression in cerebellum, but low expression in the other brain regions, while the last group of genes at the bottom of the heat map have high expression in hippocampus and necortex.

## Cluster differentially expressed genes in all brain regions

We are interested in identifying clusters that contain genes with expression patterns that are more similar to each other than they are to genes in other clusters.

Using **Create K-medoids Clustering for RNA-Seq**, we will analyze the 40 genes identified earlier as being differentially expressed in at least one stage in all four brain regions. The aim is to identify clusters of the genes with similar expression patterns across the developmental stages. We will order the samples chronologically, i.e. early developmental stages → late developmental stages, to help when interpreting the results in terms of expression across time.

A key output of this tool is a Sankey plot, providing an intuitive and interactive view of the clustering and flow of gene expression patterns across brain regions and developmental stages.

1. Copy the names of the 40 differentially expressed genes from step 11 on page 11.
2. Start **Create K-medoids Clustering for RNA-Seq** by going to:  
**Tools | RNA-Seq and Small RNA Analysis (🇺🇸) | Expression Plots (📊) | Create K-medoids Clustering for RNA-Seq (📊)**
3. Select the 71 gene expression tracks as input.

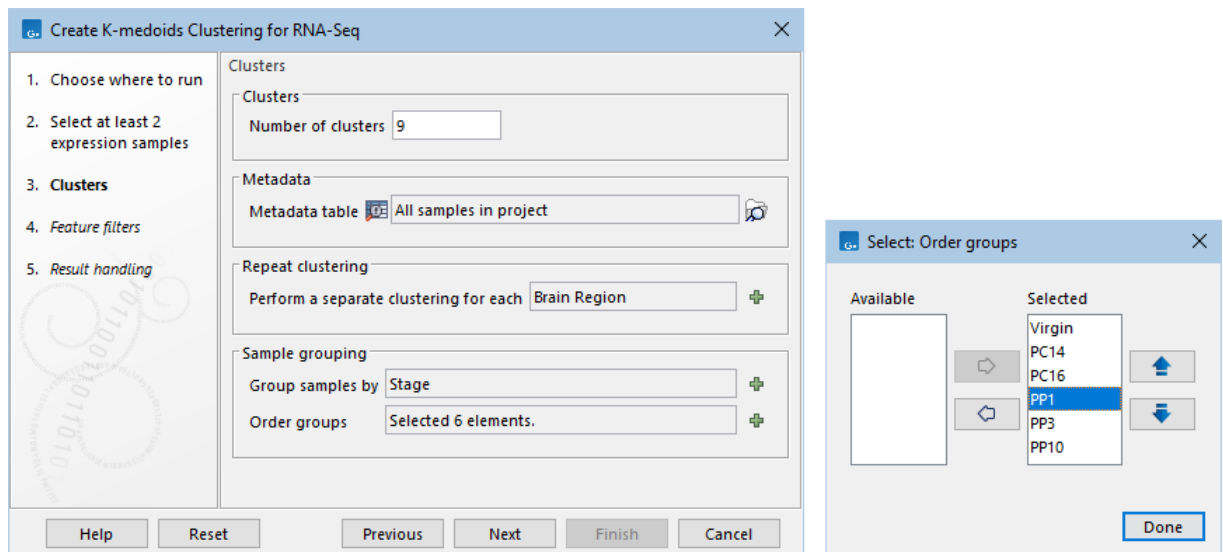


Figure 12: Creating 9 K-medoids clusters for each brain region, where samples are grouped by Stage. The order of the developmental stages is shown to the right. The blue arrows to the right can be used to reorder the time points.

4. In the "Clusters" wizard step, set the "Number of clusters" to 9 and configure the remaining options as shown in figure 12. In the "Sample grouping" section, click on (+) to the right of "Order groups" and order the stages to preserve their chronological order (Virgin, PC14, PC16, PP1, PP3, PP10).
5. In the "Feature filters" wizard step, set "Filter settings" to "Specify features" and paste the 40 gene names into the "Feature names" text field.
6. Choose to save the result in the Plots folder.

7. When opening the result, a split view is shown with a Sankey plot at the top and a line graph at the bottom (figure 13).
8. In the **Side Panel** under "Grouping", click the (+) button next to "Show stacks for:" and use the blue arrow buttons to reorder the brain regions, as shown in figure 13.

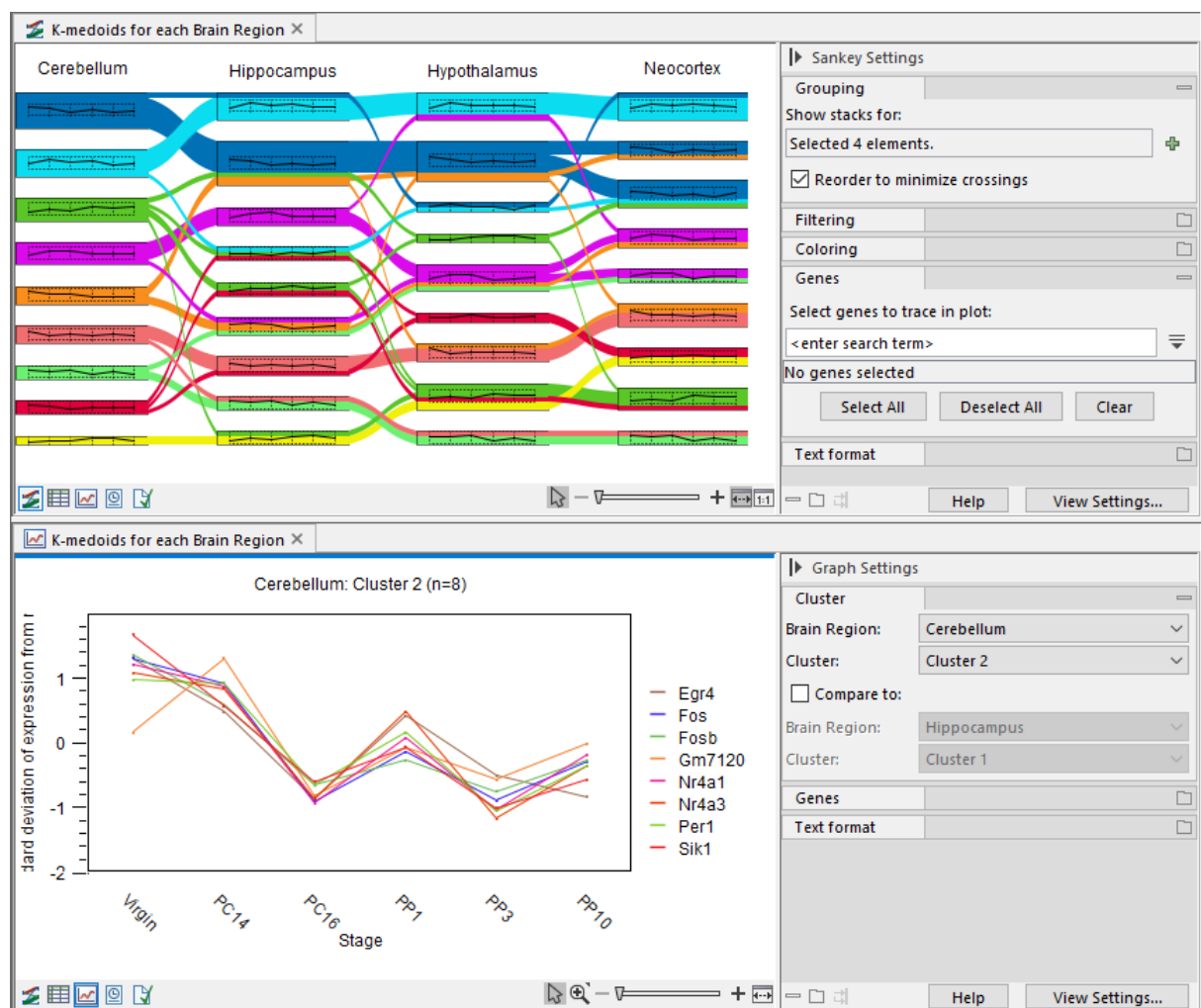


Figure 13: Top: Sankey plot showing 9 clusters for each of the four brain regions, and the flow of genes through the clusters. Each cluster shows a line graph with the general trend for the gene expression as a function of the six developmental stages. Bottom: The leftmost top cluster line graph plot.

The Sankey plot (figure 13) shows the 9 identified clusters for each of the brain regions and how many genes they share. A general trend for the gene expression as a function of stage is shown for each cluster as a line graph. Hover the mouse cursor over a cluster to reveal a tooltip containing the names of the genes in that cluster.

The line graph shows gene expressions across the developmental stages. A single cluster (figure 13) or two clusters (figure 14) can be specified in the **Side Panel**. When two clusters are specified, only genes common to both clusters are displayed. The expression values from the first cluster are shown as solid lines, values from the second cluster are shown as dotted lines.



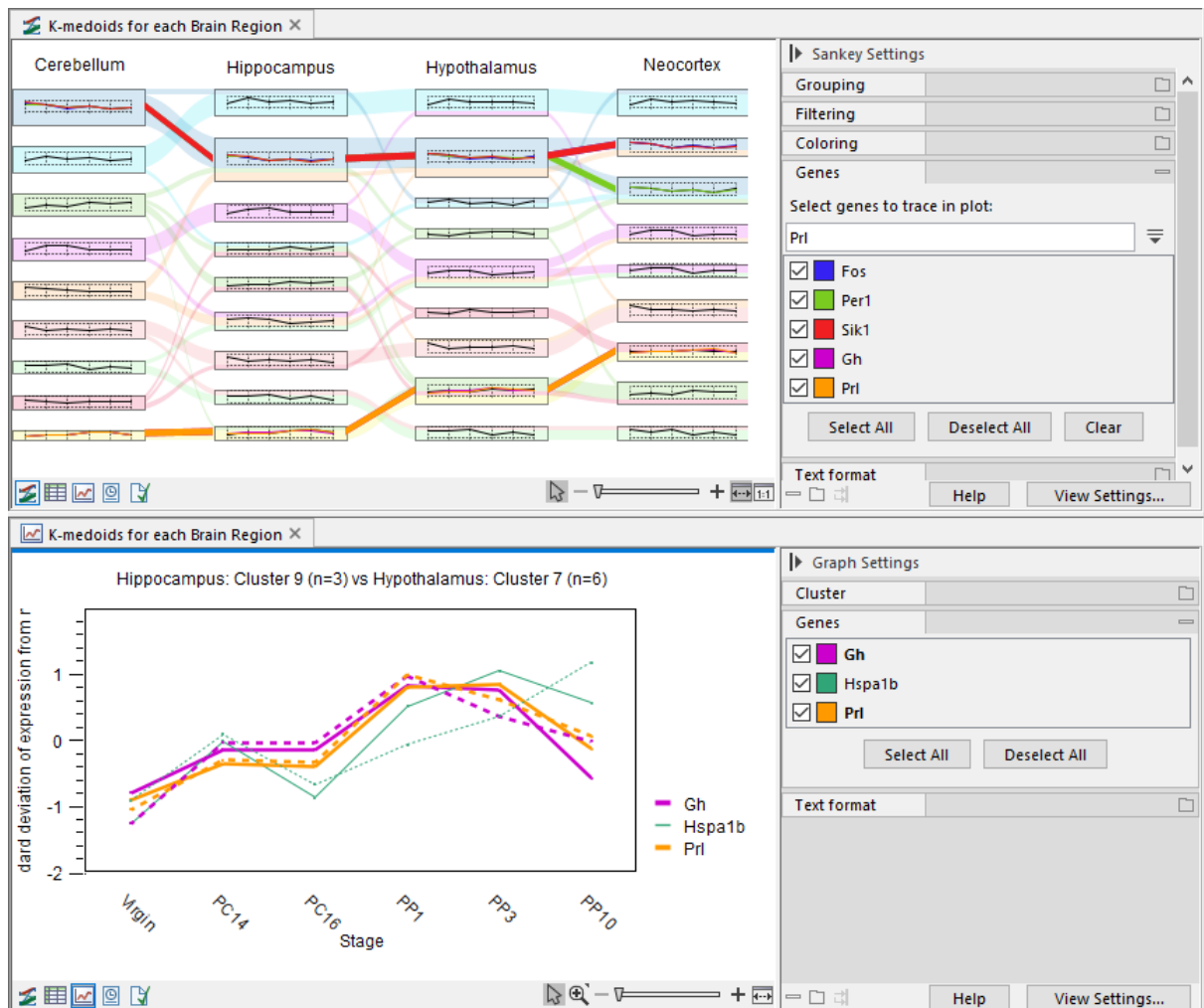


Figure 14: Top: Fos, Per1, Sik1, Gh and Prl are highlighted in the Sankey plot. The colored lines indicate the clusters the genes are located in for the different brain regions. When the lines are overlapping, as they do on the left side of the plot for Fos, Per1, Sik1 and throughout the plot for Gh and Prl, only the color of the last gene, here red and orange, is visible. The line graphs on the clusters now also show the expression of the selected genes, in addition to the general trend. Hippocampus cluster 9 and neocortex cluster 6 are selected. Flows opacity has been decreased to more easily trace the genes. Bottom: Gh, Prl and Hspa1b are found in both hippocampus cluster 9 and neocortex cluster 6. As Gh and Prl are highlighted in the Sankey plot, their lines are thicker and they appear in bold in the Side Panel. The line graph shows their expression in hippocampus cluster 9 as solid lines and neocortex cluster 6.

The clusters can be selected directly from the Sankey plot or through the options in the **Side Panel**:

- Clicking clusters in the Sankey plot selects them. Press the Ctrl (Cmd on Mac) key while clicking the clusters to select two clusters from two different brain regions.
- Under "Cluster" in the **Side Panel**, choose the brain region and the cluster to be shown in the plot. Tick "Compare to:" to add a second cluster to the plot.



It can often be useful to highlight specific genes in the plot and see in which clusters they are present (figure 14):

1. In the **Side Panel** under "Genes", start typing the desired gene name in the "Select genes to trace in plot" text field.  
A list of genes matching the text appears below.
2. Double click on the desired gene.  
The gene is now selected and highlighted in the plot.
3. Repeat the steps to add more genes to the selection.  
To more easily see the location of the selected genes, the opacity of the flows can be reduced under "Coloring" in the **Side Panel**.

### Further notes about the *Fos*, *Per1*, *Sik1*, *Gh* and *Prl* genes

In this tutorial, all these five genes are identified as differentially expressed in all brain regions.

*Fos*, *Per1* and *Sik1* are clustered together for three of the four brain regions, where they show similar expression patterns across the developmental stages (figure 14). In neocortex, *Per1* belongs to neocortex cluster 5, while *Fos* and *Sik1* belong to neocortex cluster 8. Clusters 5 and 8 exhibit similar expression trends and this could indicate that only 8 clusters instead of 9 are needed for neocortex.

*Gh* (Growth hormone) and *Prl* (Prolactin) cluster together in all brain regions. The clusters *Gh* and *Prl* are part of contain genes for which the expression increases after the mice give birth (figure 14).

In the original **paper**, the authors performed a differential expression analysis for all brain regions and clustered the differentially expressed genes. They identified *Fos*, *Per1* and *Sik1*, but not *Gh* and *Prl*. *Fos*, *Per1* and *Sik1* are all part of the same cluster (figure 5 in the paper). The authors discuss the importance of these *Gh* and *Prl* in postpartum and that they have been identified in previous studies.

### Further investigation of gene expression patterns

In this section, we explore expression levels of particular genes in more detail.

Using **Create Expression Browser**, we will create an expression browser that can generate a variety of bar charts. This helps exploring the expression level in more details and better understand the results of the differential analysis and clustering.

1. Start **Create Expression Browser** by going to:  
**Tools | RNA-Seq and Small RNA Analysis** (📄) | **Create Expression Browser** (🔍)
2. Select the 71 gene expression tracks.
3. In the "Select additional data" wizard, add the four statistical comparison tracks generated previously (figure 15).

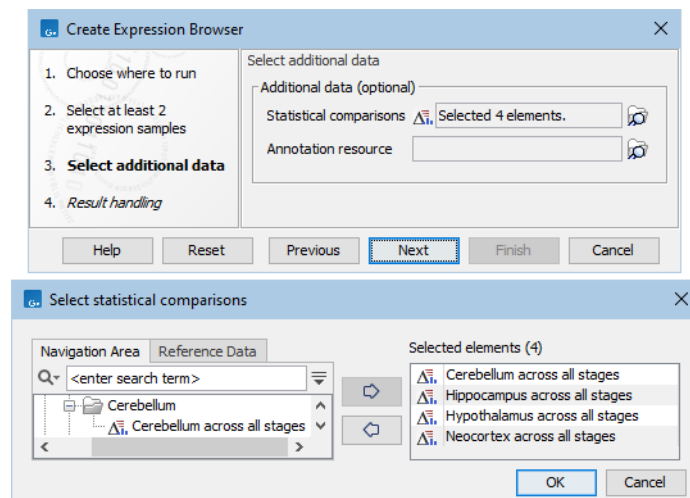




Figure 15: Creating an expression browser using the four statistical comparison tracks.

4. Choose to save the result in the DE folder.
5. Open the expression browser.
6. To visualize the bar chart together with the table in split view, press the Ctrl (Cmd for Mac) key and click on the  button below the displayed table.  
No genes are selected at this point.  
Let us view the expression for three of the genes discussed in the previous section on page 17: *Fos*, *Per1* and *Sik1*.
7. Click on the  button to open the advanced filtering options.
8. Filter the table by setting that the "Name" "is in list", write "Fos Per1 Sik1" in the text field and click **Filter** (figure 16).
9. Select all the rows in the table (shortcut Ctrl+A or Cmd+A on Mac).  
The bar chart updates to show the three genes.
10. In the bar chart **Side Panel**, under "Groups", select "Stage".  
The bar chart now shows each sample as a bar, and samples are grouped according to the developmental stage.
11. In the bar chart **Side Panel** under "Grouping", click "Collapse groups" to represent each group by one bar.  
This makes it easier to see that the genes have the same expression pattern across the developmental stages (figure 16).

## Statistical significance

When statistical comparison tracks containing pair-wise comparisons are used as input to the **Create Expression Browser** tool, the bar chart can highlight which comparisons are significant.

1. Using steps 1 to 8 on page 8, start **Differential Expression for RNA-Seq** for neocortex.

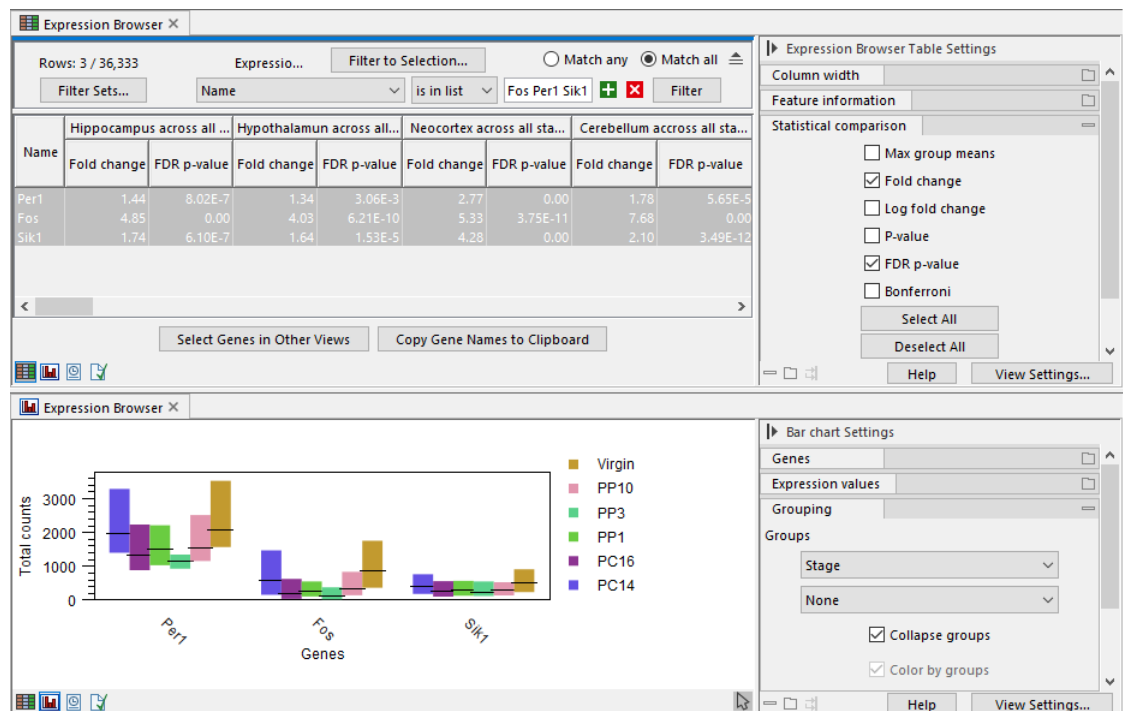
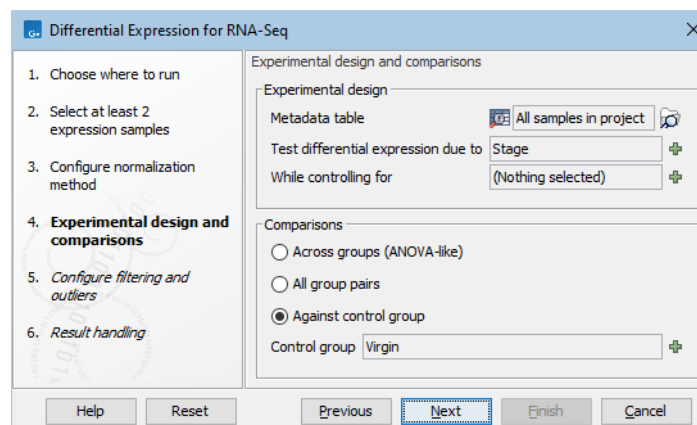


Figure 16: Expression Browser with split view. Top: Table with multiple genes selected. Bottom: Bar chart showing the genes selected in the table. Samples are grouped by developmental stage and groups are collapsed.


- In the "Experimental design and comparisons" wizard step, select the "All samples in project" metadata, choose "Stage" in "Test differential expression due to", set "Comparisons" to "Against control group" and choose "Virgin" under "Control group" (figure 17).



The figure shows the "Differential Expression for RNA-Seq" wizard step. The "Experimental design and comparisons" section is active. The "Experimental design" sub-section shows "Metadata table" set to "All samples in project", "Test differential expression due to" set to "Stage", and "While controlling for" set to "(Nothing selected)". The "Comparisons" sub-section shows "Across groups (ANOVA-like)" selected, "All group pairs" unselected, "Against control group" selected, and "Control group" set to "Virgin". The "Next" button is highlighted.

Figure 17: Differential expression will be tested for developmental stage against Virgin control group.

- Keep the default settings for the remaining options.
- Choose to save the results in the DE / Neocortex folder.
- Using the same approach (steps 1 to 6 on page 8), start **Create Expression Browser** for neocortex.

6. In the "Select additional data" wizard, add the five results from step 4 to the "Statistical comparisons".
7. Choose to save the result in the DE / Neocortex folder.
8. Open the resulting expression browser.
9. To visualize the bar chart together with the table in split view, press the Ctrl (Cmd for Mac) key and click on the  button below the displayed table.
10. Customize the expression browser to show the expression of *Gh* and *Prl* (figure 18):

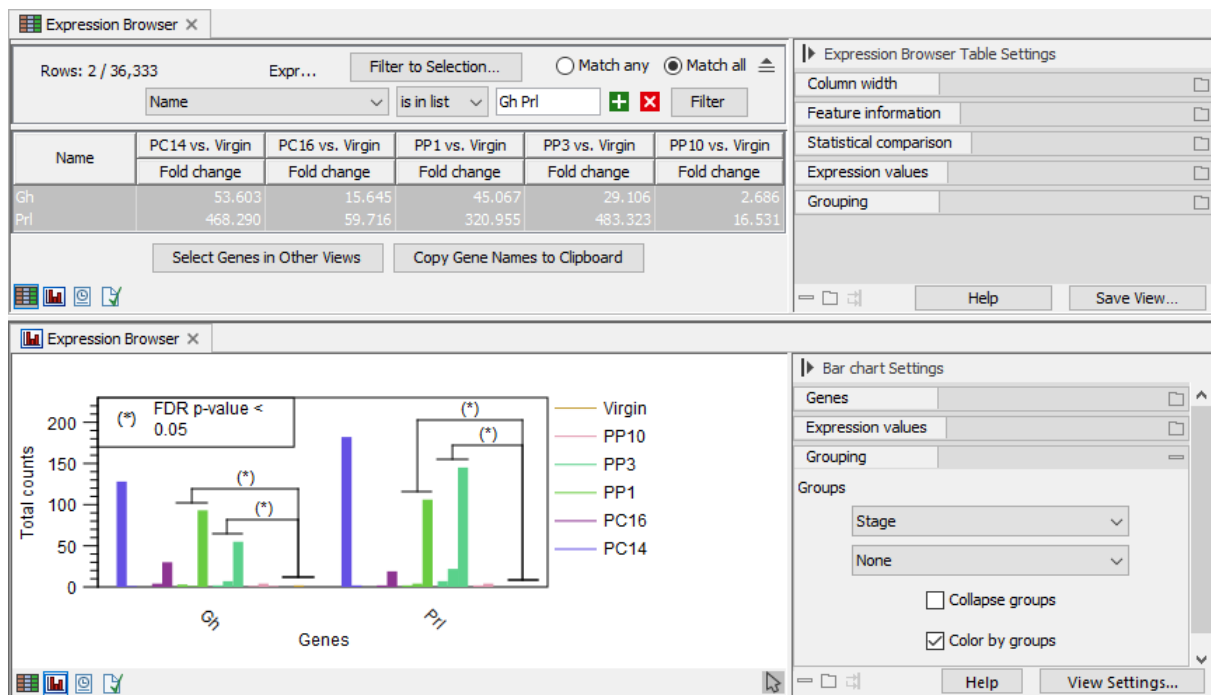



Figure 18: Expression Browser shows *Gh* and *Prl* in neocortex, and statistical significance is marked on the bar chart.

- Click on the  button to open the advanced filtering options.
- Filter the table by setting that the "Name" "is in list", write "Gh Prl" in the text field and click **Filter**.
- In the table view, select the table entries for *Gh* and *Prl* while pressing the Ctrl (Cmd for Mac) key.
- In the bar chart Side Panel, under "Groups", select "Stage".

Because the statistical comparison tracks included in the expression browser are pair-wise, the box chart now highlights the statistically significant comparisons.

*Gh* and *Prl*, discussed previously on page 17, increase in expression after the mice gave birth. Significant changes are observed for PP1 and PP3 in both genes, showing that this expression increase is indeed significant (figure 18).

**Pathway Analysis**

To interpret e.g. differentially expressed genes in their biological context, we recommend [Ingenuity Pathway Analysis \(QIAGEN IPA\)](#). If you do not have an IPA account, a free trial can be requested via [Trial request](#). Your IPA account credentials can then be used to upload the data from *CLC Genomics Workbench* with the [Upload to IPA](#) tool from the [Biomedical Genomics Analysis](#) plugin. Refer to the [CLC Genomics Workbench manual](#) for how to install plugins.

---