Tutorial

Advanced RNA-Seq analysis with upload to IPA

February 4, 2020

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Advanced RNA-Seq analysis with upload to IPA

The purpose of this tutorial is to illustrate how workflows can be used to easily run RNA-seq and downstream analyses for multiple samples and multiple groups of samples. In addition, it demonstrates the collaborative power of *CLC Genomics Workbench* and Ingenuity Pathway Analysis (IPA) for the analysis and interpretation of RNA-Seq expression data.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 20.0 or higher.

To use workflows that include upload to IPA, you must have the Ingenuity Pathway Analysis plugin installed. Installing plugins is described in the CLC Genomics Workbench manual. In addition, you must have access to IPA services. You can request a free trial by clicking on Request a trial.

General tips

- Within wizard windows you can use the **Reset** button to change settings to their default values.
- You can access the in-built manual by clicking on **Help** buttons or by selecting the "Help" option under the "Help" menu.

Download and import the data

The data we will analyze in this tutorial was sourced from the PRJNA328591 study, which included 12 mRNA profiles of Dengue virus 2 and mock infected cells at 24 and 36 hours post infection. To allow the tutorial to be completed in a reasonable amount of time, only a subset of the reads that map to chromosome 17 are used here.

The zip file distributed for use with this tutorial contains:

- A subset of the reads from the 12 samples described above.
- A metadata table.
- A reference sequence track for chromosome 17 of the human hg38 genome and corresponding gene and mRNA tracks.
- Four workflows:
    1. RNA-Seq and IPA analysis workflow
    2. RNA-Seq and IPA advanced analysis workflow
    3. RNA-Seq analysis workflow
    4. RNA-Seq analysis advanced workflow

The rest of this tutorial refers to worklows that include uploading results to IPA. If you do not have access to IPA, then please use the third and fourth workflows listed above instead of the ones named in the instructions.

To download and import the tutorial data:

1. Download the following zip file: `RNA_Seq_IPA_Dengue.zip`.
2. Start up the *CLC Genomics Workbench*.
3. Import the data by clicking the () **Import** button and selecting () **Standard Import**.

   Choose the zip file, leave the import type set to Automatic and save the imported data in your Navigation Area.

   You can alternatively drag-and-drop the zip file into the Navigation Area.

   Once the import is completed, you should see the folders and data elements in the Navigation Area as shown in figure 1.

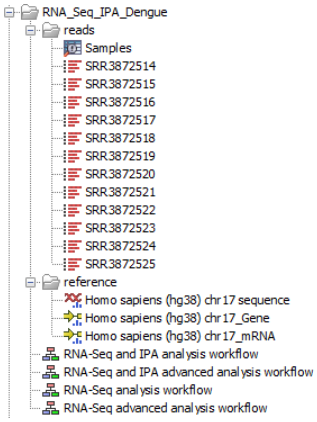

Figure 1: *The Navigation Area after the tutorial data has been imported.*

4. Associate the metadata rows with the imported reads. To do this,

   Open the "Samples" metadata table found in the "reads" folder.

   Click on **Associate Data...** | **Associate Data Automatically...**.

   Right-click on the "reads" folder and choose **Add folder contents** (figure 2).

   Keep clicking **Next** on each wizard step, leaving the default settings. On the final step, click **Finish**.

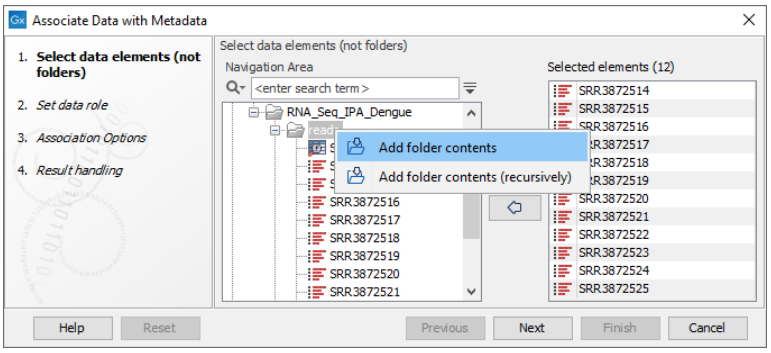

Figure 2: *The elements have been added using the option "Add folder contents".*

5. Select all the rows in the metadata table, and click **Find Associated Data**.

   This shows if all rows are associated correctly with the corresponding reads (figure 3).

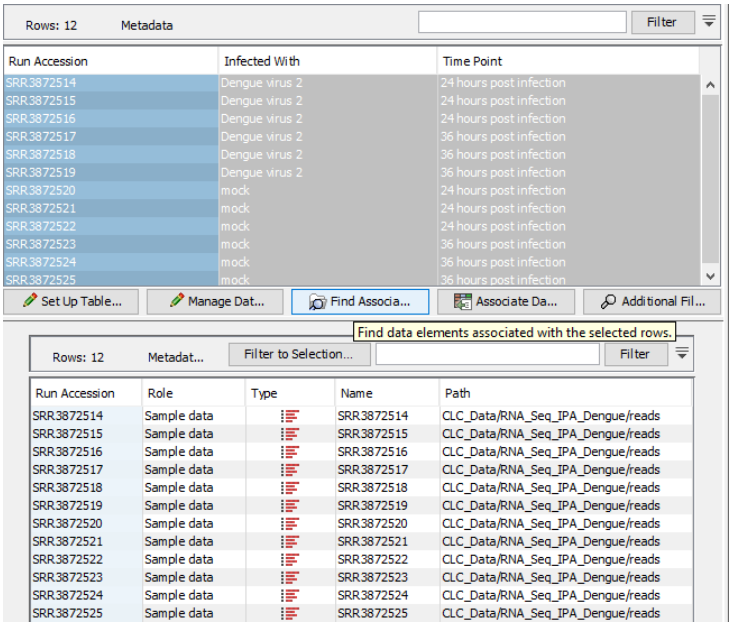

Figure 3: *The reads have been successfully associated with the metadata.*

Running the RNA-Seq analysis and uploading to IPA

The "RNA-Seq and IPA analysis workflow", provided with the tutorial data, shown in figure 4, will be used for the analysis.

In this workflow, the first steps **Trim Reads**, **RNA-Seq Analysis** and **Combine Reports Per Sample** are run once for each sample provided for that workflow run. Results from the **RNA-Seq Analysis** and **Combine Reports Per Sample** steps are then collected. These collected results are passed to the downstream steps, which are run just once. The downstream steps are where the expression results are further processed and reports for all the samples are combined into a single report.

The ability to run parts of a workflow a different number of times is achieved using the control flow elements, **Iterate** and **Collect and Distribute**. These are described further in the `Batching part of a workflow` section of the manual, with the section on `Advanced workflow batching` providing examples of their use.

Additionally, we will use the batch functionality to run the workflow twice, once for each time point post infection. For this, we will use `metadata to define the batch units`.

To run the workflow:

1. Open the "RNA-Seq and IPA analysis workflow" by double clicking on its name in the Navigation Area.
2. Start the workflow by clicking on the (▷) **Run** button near the bottom, on the right hand side.

   You will now step through the workflow wizard to specify input data and configure options before launching the workflow to run.
3. Specify the data to be analyzed by right-click on the "reads" folder and choosing the **Add folder contents** menu option.

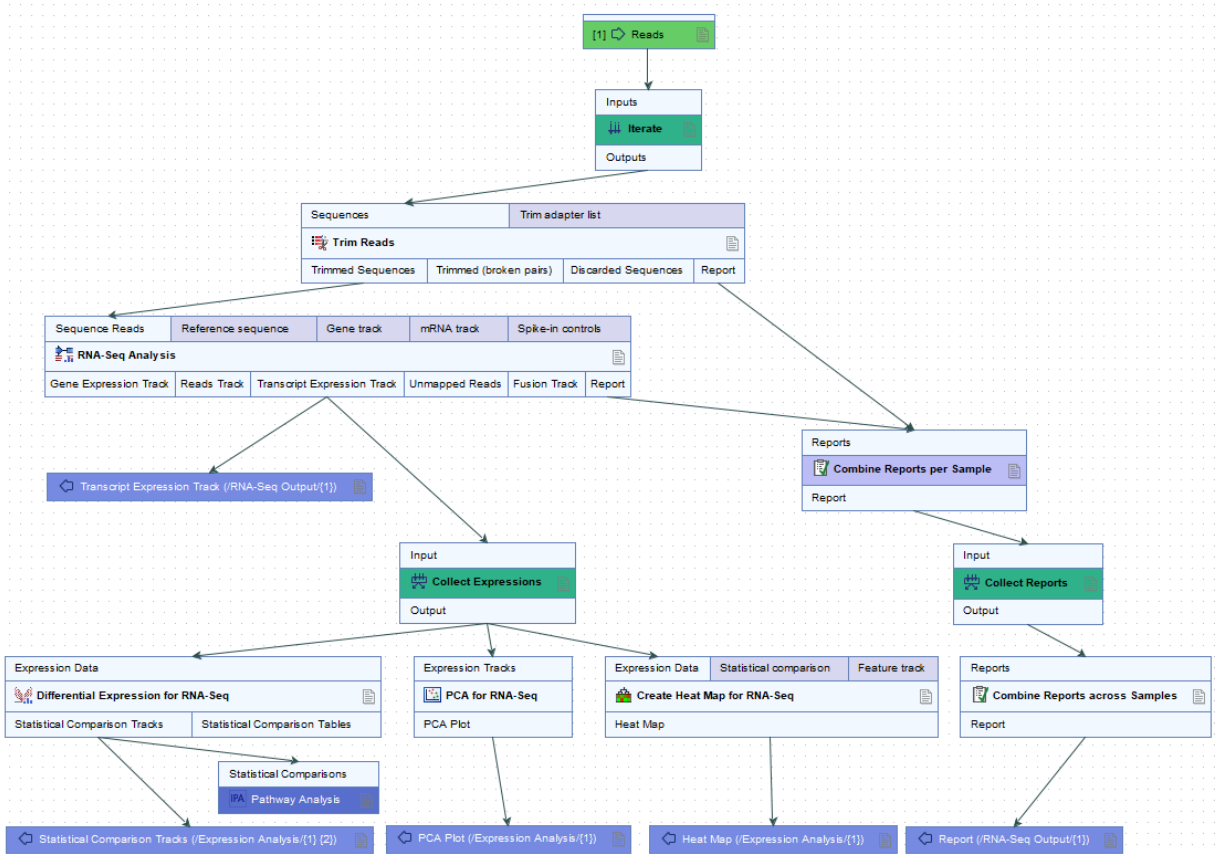

Figure 4: *A view of the "RNA-Seq and IPA analysis workflow".*

Check the "Batch" option below the data selection area (figure 5).

With the "Batch" option enabled, a warning text appears indicating that the workflow design itself will lead to at least part of the workflow being run multiple times with subsets of the inputs. This warning is to help us decide if we also intend to run the whole workflow multiple times using subsets of inputs. Here, this is indeed our intention.

Click **Next**.

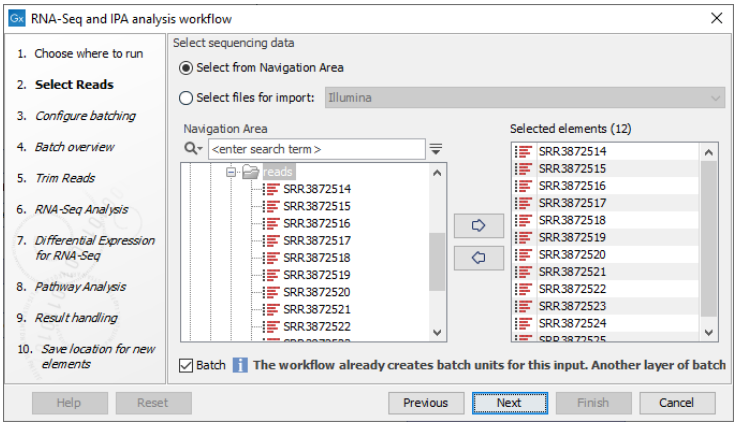

Figure 5: *The "Batch" box is checked and a warning is displayed in the wizard.*

In this tutorial, we imported the reads prior to launching the workflow. When analyzing your own data, you may prefer to use "Select files for import" to import data on the fly.

4. Configure how the workflow will run using the provided metadata table (figure 6).

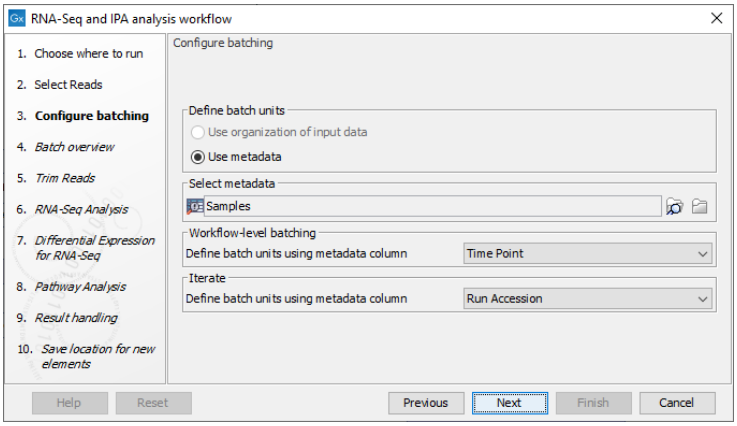

Figure 6: *The workflow execution is configured using the "Samples" metadata table.*

Click on the ( ) button at the right hand side of the "Selected metadata" field.

Choose the "Samples" metadata table.

Click **OK**.

In the "Workflow-level batching" area, click on the down arrow and select the option "Time Point".

This specifies that we wish to run the workflow once for each value in that column of the "Samples" metadata. Here, there are 2 values in that column, so the workflow will be run twice. One time, the input data elements associated with metadata table rows containing "24 hours post infection" will be used. The next time, the input data elements associated with metadata table rows with "36 hours post infection" are used.

In the "Iterate" area, click on the down arrow and select the option "Run Accession".

This specifies that during each workflow run, the data elements with the same value in the "Run Accession" column will be treated as part of a single sample. For this workflow, this means that such data elements will be analyzed together in the **Trim Reads**, **RNA-Seq Analysis** and **Combine Reports Per Sample** steps.

Click **Next**.

In this tutorial, we imported the metadata prior to launching the workflow. When analyzing your own data, it may be more convenient to use an excel format file containing metadata when launching workflows.

5. Review the organization of the input data in the Batch overview step (figure 7).

   If the organization of the input data is not as expected, batch unit configuration can be adjusted in the previous step by clicking on **Previous**.

   Click **Next**.

6. Leave the **Trim Reads** settings at their default values.

   Click **Next**.

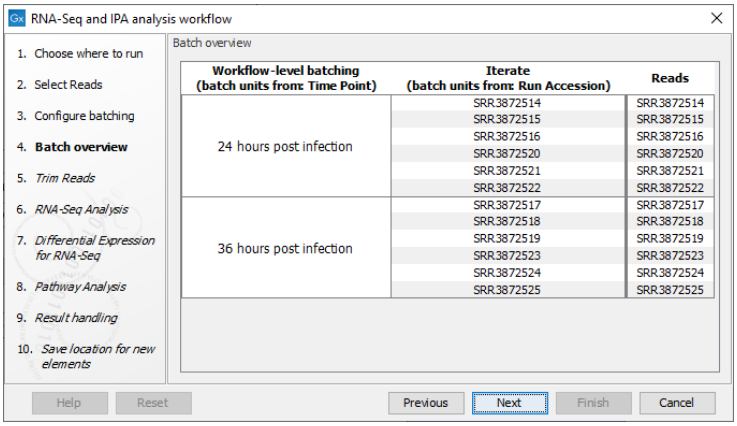

Figure 7: *The Batch overview step shows how the data is grouped for the analysis.*

7. Configure the **RNA-Seq Analysis** options:

   Specify the reference data to use so it looks like that shown in figure 8. Use the reference data provided with the tutorial data for this.

   Set the Strand specific option to "Reverse" as a KAPA stranded RNA-seq kit was used to generate this data.

   Click **Next**.

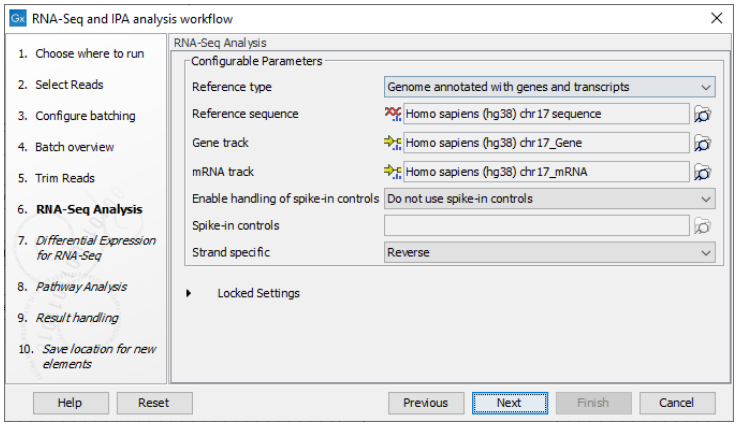

Figure 8: *RNA-Seq Analysis is configured with the provided references.*

8. Configure the **Differential Expression for RNA-Seq** options to look like those shown in figure 9). Specifically, set:

   - "Test differential expression due to" to the option "Infected With", and
   - Set "Control group" to "mock".

   Click **Next**.

9. Enter your Ingenuity username and password, leaving the remaining **Pathway Analysis** options at their default values.

   You will not see this step if you are running the "RNA-Seq analysis workflow".

   Click **Next**.

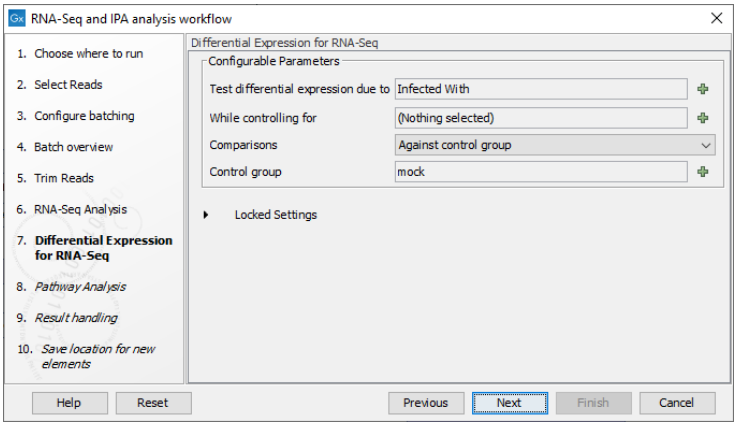

Figure 9: *Differential expression is tested due to "Infected With", using the "mock" group as control.*

10. Check the "Create subfolders per batch unit" and "Create workflow result metadata" boxes (figure 10). This will create two subfolders for each of the time points post infection, and a `Workflow Result Metadata` with information about the results.

    Click **Next**.

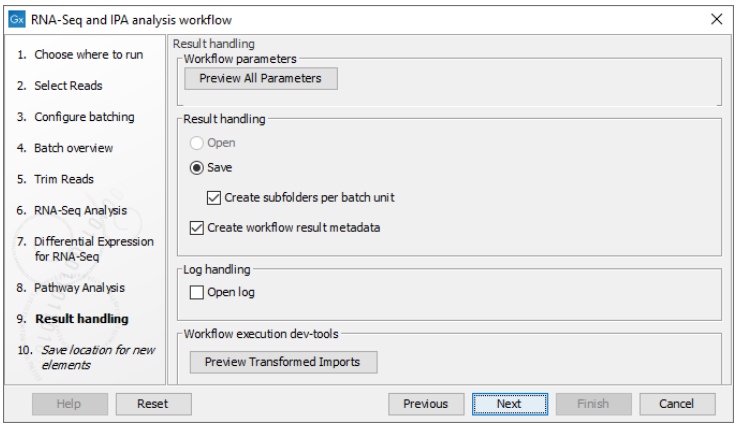

Figure 10: *"Create subfolders per batch unit" and "Create workflow result metadata" boxes are checked.*

11. Choose the location to save the results (for example a new "results" subfolder).

    Click **Finish**.

    The workflow will now execute. You can monitor the progress of the workflow in the "Processes" bar (figure 11). It will take some time for this workflow to run to completion.

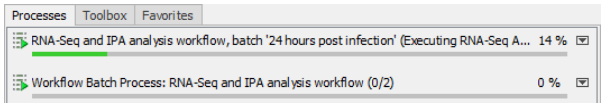

Figure 11: *The "Workflow Batch Process" indicates how many batches have been completed.*

Results interpretation

Results from the analyses carried out by the workflow will be placed in the "results folder", as shown in figure 12.

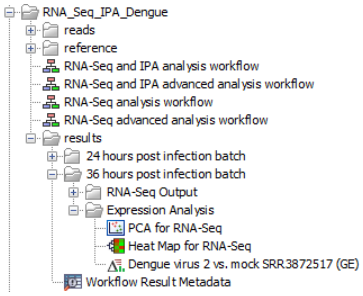

Figure 12: *Results from the analysis are saved into folders, visible in the Navigation Area.*

The statistical comparisons were automatically uploaded to IPA as part of the workflow run. You will receive an email when the IPA analysis is complete.

The results from each batch unit are saved into a subfolder, named after the data grouping indicated in the "Batch overview" (figure 7). Within each of these folders, there are subfolders for particular types of result data:

- "RNA-Seq Output" contains the outputs of **RNA-Seq Analysis** and the combined report,
- "Expressions Analysis" contains the differential expression, PCA plot and heat map.

A single `Workflow Result Metadata table` is generated for the workflow run, and is saved within the "results" folder. This can be particularly useful for finding all the output elements of a workflow when there are many batch runs involving many outputs.

Combined report

Open one of the reports found in the "RNA-Seq Output" folders.

The `combined report` summarizes information from the samples in that batch unit. It can be used to quickly review the results of the trimming and RNA-Seq analyses. In this report, samples highlighted in yellow are outliers, making it easy to spot any problematic samples.

PCA plot

Open one of the "PCA for RNA-Seq" plots found in the "Expression Analysis" folders.

The `PCA` plot is colored by "Infected With". Note that you can:

- change the display settings in the "Metadata" side panel section,
- click-and-drag to move the legend for optimal visibility.

We can see that, as expected, the samples cluster by the infection type (figure 13).

You can also visualize the plot in 3D by clicking on the ( ) icon (figure 13). You can click-and-drag to change the orientation of the axis.

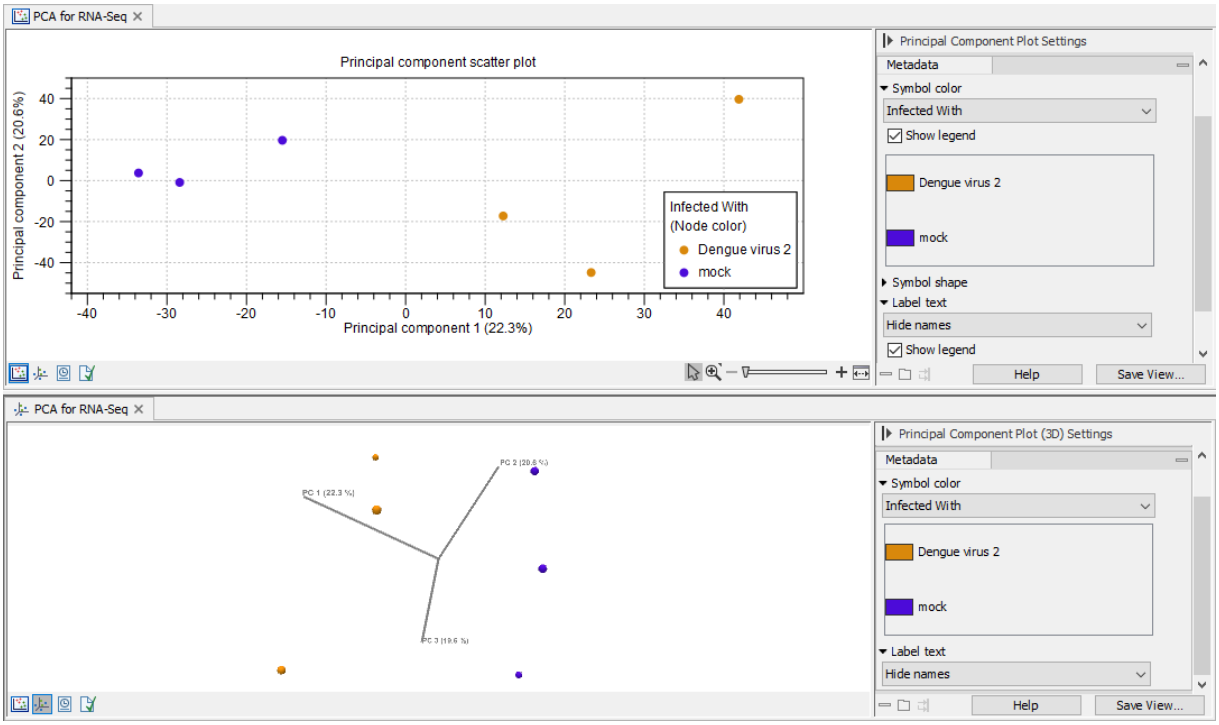

Figure 13: *Top: The PCA plot when samples are colored by infection type. Bottom: The 3D view of the same PCA plot. Label text is hidden in both views.*

Heat map

Open one of the "Heat Map for RNA-Seq" plots found in the "Expression Analysis" folders.

The `hierarchical clustering algorithm` is pre-configured in the workflow to select the 25 "most interesting" transcripts in the heat map, based on the coefficient of variation (relative standard deviation). The samples are also clustered on the horizontal axis, by unsupervised clustering.

To see that the samples cluster by the infection type, add "Infected With" as a metadata layer in the "Metadata" side panel section (figure 14).

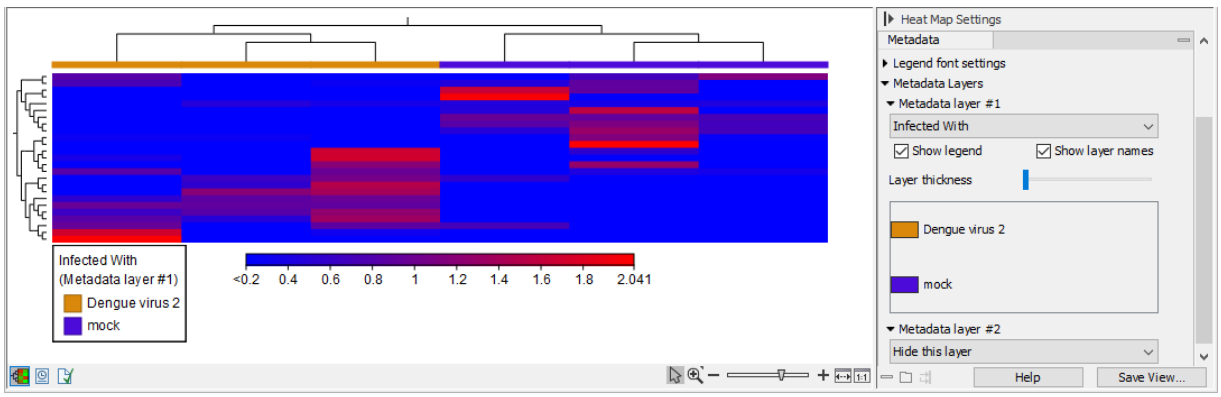

Figure 14: *The "Metadata" side panel section options for heat maps. The names of the samples (in the "Samples" section) and features (in the "Features" section) are hidden for increased visibility.*

Statistical comparison

Open one of the "Dengue virus 2 vs. mock" comparisons: it will automatically display the table view ( ) of the track.

We will first investigate the volcano plot of the results.

Click the volcano plot icon ( ) at the bottom of the view to open this plot.

This dataset is small and has very few significant transcripts under FDR-correction, therefore the volcano plot is not very dense. We can improve the visibility in this particular case by changing the settings in the "Values" side panel section (figure 15).

- Choose "P-value" in the "P-value type" drop down menu.
- Change the "Lower limit on p-values" to 1E-5.

You can select transcripts by simply clicking on the point representing them. This will turn those points red and their names will be displayed next to the point.

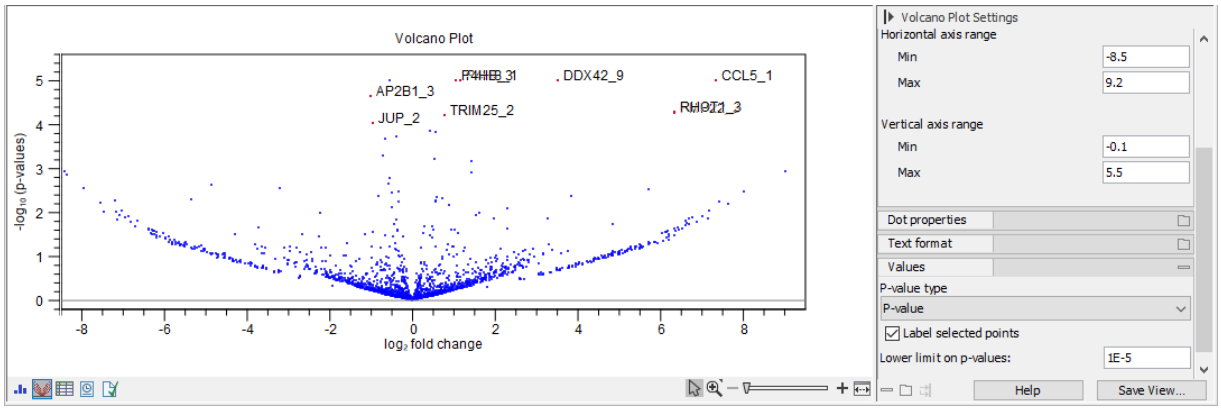

Figure 15: *Volcano plot and the corresponding side panel options.*

Transcripts selected in one view are also selected in other views. For example, rows in the table corresponding to points selected in the volcano plot will also be selected, as will the corresponding positions in the track view. Using this functionality, points of relevance can be highlighted in graphical views, like the volcano plot, by filtering for particular characteristics in the table and then selecting the visible table rows.

For example, to show only points in the volcano plot representing samples with an FDR value less than 0.05 and a fold change greater than 1.5:

1. Open the table view by clicking on the ( ) at the bottom of the window.
2. Click on ( ) next to the **Filter** button. If you cannot see the **Filter** button, expand the width of your viewing area.
3. Select
   - **FDR p-value** in the first field,
   - **abs value <** in the second field,

- type **0.05** in the third field.

4. Add a new filtering criterion by clicking on ().
5. Select
   - **Fold change** in the first field,
   - **abs value >** in the second field,
   - type **1.5** in the third field.
6. Click **Filter**.
7. Select all remaining rows in the table.
8. Go to the volcano plot view ().
9. The selected transcripts are now highlighted in red and have their names displayed.
10. Axes ranges can be altered in the "Graph preferences" side panel section (figure 15).

You can read more about statistical comparisons in the CLC Genomics Workbench manual.

Finally, you can launch IPA to look for genes and associated pathways that are differentially expressed between the Dengue virus 2 and mock infected cells.

Remember that we are only analyzing genes present on chromosome 17 in this tutorial. IPA requires whole genome analysis to output a comprehensive picture for the whole genome.

Comparison across time points

We ran the workflow in batch mode with two batch units, one for each time point post infection. This leads to a rather limited view of the data where the two time points cannot be compared.

Venn diagram

Make sure you do not close the statistical comparison from before.

We will first create a Venn diagram from the two statistical comparisons to see the overlap between the differentially expressed genes at the two time points:

1. Go to:
   Toolbox | () **RNA-Seq and Small RNA Analysis** | () **Create Venn Diagram for RNA-Seq**
2. Choose the two "Dengue virus 2 vs. mock" comparisons as input.
3. Click **Next**.
4. Choose to "Open" the result.
5. Click **Finish**.

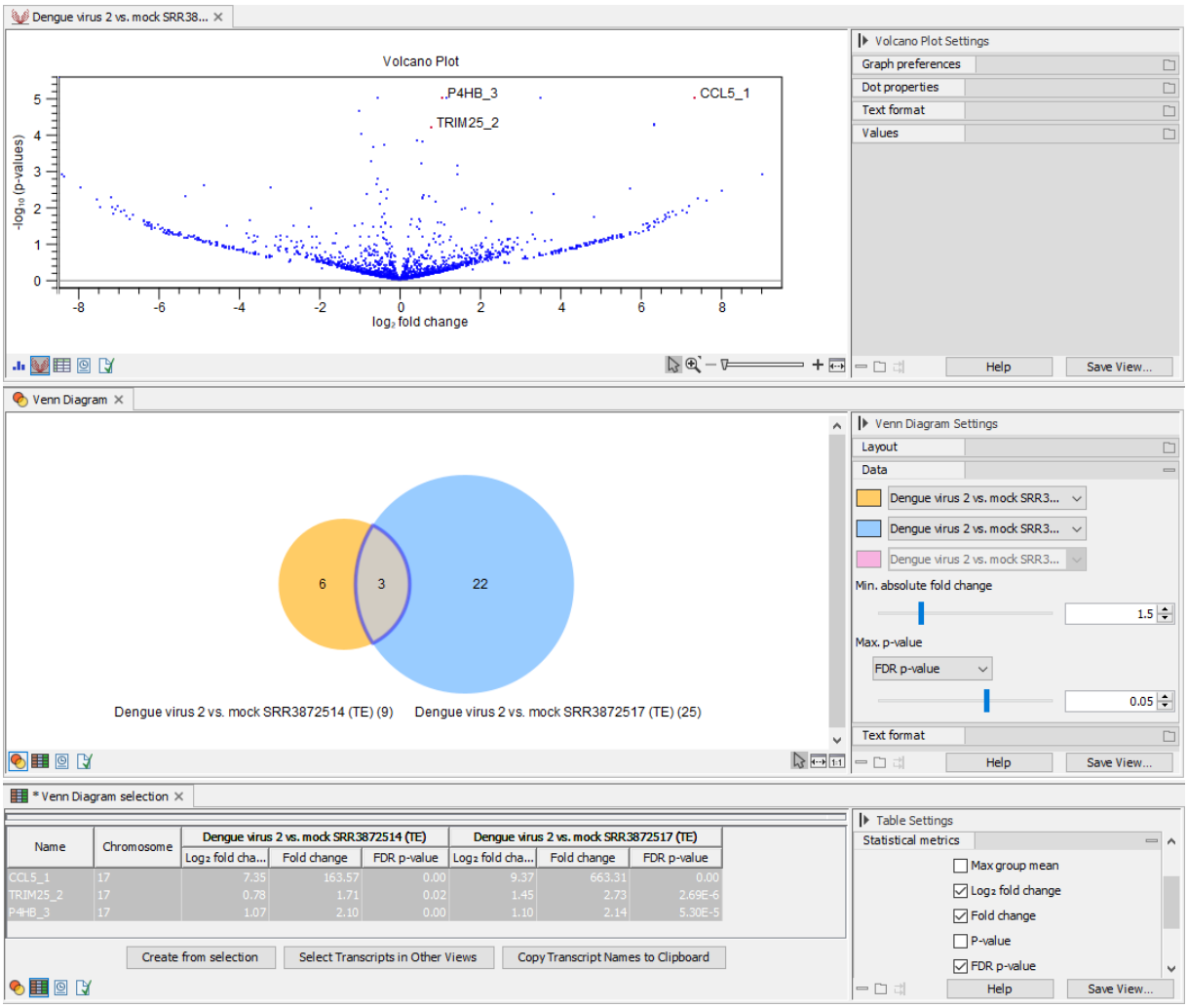

Figure 16: *From top to bottom: Statistical comparison, Venn diagram, and selection from the diagram. The same transcripts are selected in all views.*

The diagram shows how many transcripts were detected to be differentially expressed in the two comparisons.

Select Transcripts in Other Views allows you to do synchronized selections between expression tracks, statistical comparison tracks, and the table view of Venn diagrams (figure 16):

1. Select the middle intersection with the transcripts that were differentially expressed in both comparisons.
2. Go to the table view ( ).
3. The same transcripts will still be selected in the table. It can be hard to find them.

   Click on **Create from selection** to open a new table in split view with the selected transcripts.
4. Click **Select Transcripts in Other Views** to highlight the transcripts in the opened statistical comparison.
5. You can arrange the windows in different `split views`.

PCA and heat map across time points

We can create additional PCA plots and heat maps from all 12 expression tracks:

1. Go to:

   Toolbox | () **RNA-Seq and Small RNA Analysis** | () **PCA for RNA-Seq**

2. Right-click on the "results" folder and choose **Add folder contents (recursively)** to add all 12 expression tracks.

3. Click **Next**.

4. Choose to "Open" the result.

5. Click **Finish**.

6. In the "Metadata" side panel selection, set

   - "Symbol color" to "Time Point",
   - "Symbol shape" to "Infected With".

The samples still cluster by the infection type, and for the Dengue virus 2 infected cells, the samples also cluster by the time point post infection. This is consistent with a hypothesis that the expression profiles of Dengue virus 2 infected cells continued to differentiate as time progressed, whereas the expression profiles of mock infected cells were more constant over time (figure 17).

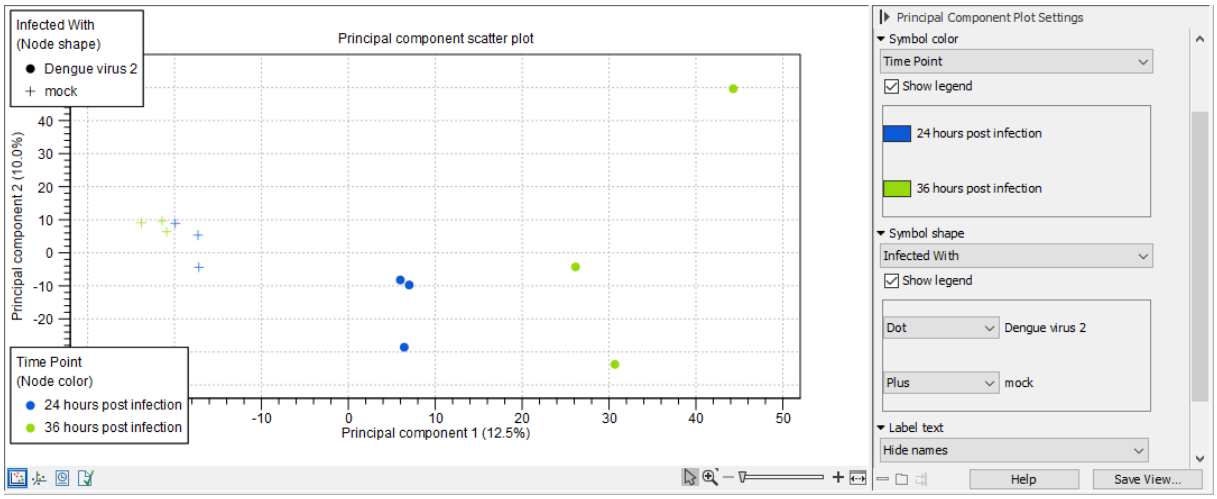

Figure 17: *The PCA plot when samples are colored by "Time Point" and shape is by "Infected With". Sample names have been hidden.*

To create a heat map,

1. Go to:

   Toolbox | () **RNA-Seq and Small RNA Analysis** | () **Create Heat Map for RNA-Seq**

2. Right-click on the "results" folder and choose **Add folder contents (recursively)** to add all 12 expression tracks.
3. Leave the default settings in the "Set clustering" wizard.
4. Click **Next**.
5. Leave the default settings in the "Set filtering" wizard.
6. Click **Next**.
7. Choose to "Open" the result.
8. Click **Finish**.
9. In the "Metadata" side panel section, set
    - "Metadata layer #1" to "Infected With",
    - "Metadata layer #2" to "Time Point".

The samples cluster just as for the PCA plot. Even though when inspecting the colors it looks as the mock infected cells cluster by the time point post infection, the hierarchical clustering shown as a tree above indicates that the two time points do not form distinct clusters.

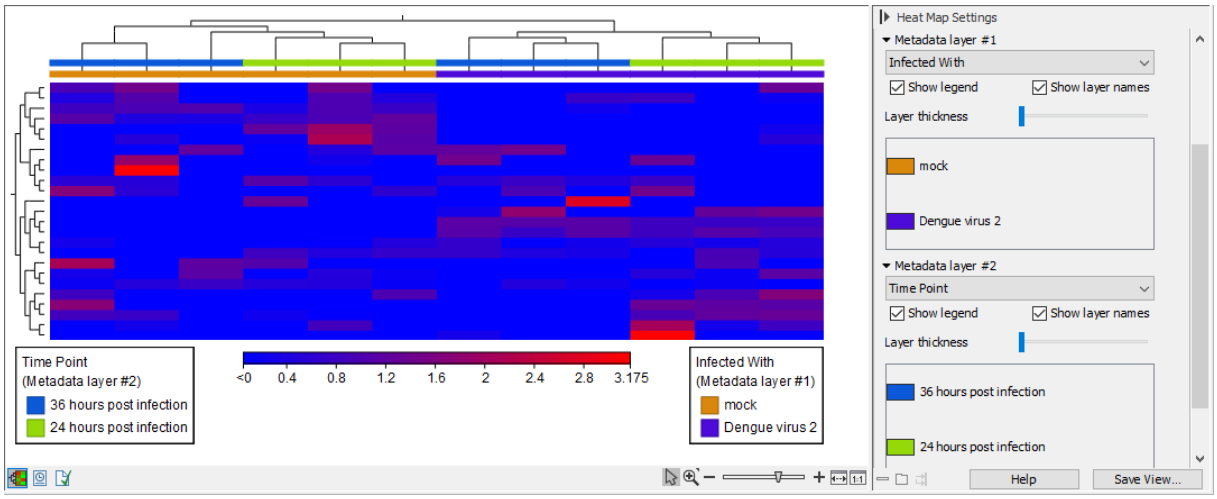

Figure 18: *The heat map when the "Infected With" and "Time Point" are added as metadata layers. The names of the samples and features are hidden to increase visibility.*

Automating and further customizing the analysis

We can further automate the analysis with just a few additions to the workflow used earlier. We provide the "RNA-Seq and IPA advanced analysis workflow" as an illustration of how this can be done. It is based on the "RNA-Seq and IPA analysis workflow", with a key addition being a new layer of **Iterate** and **Collect and Distribute** elements.

Open the "RNA-Seq and IPA advanced analysis workflow" by double clicking on its name in the Navigation Area.

The names of the control flow elements, as shown in figure 19, give an indication of their role in the workflow:

- Names ending in "over Accession" refer to running analysis steps on, and collecting results for *each sample*, where we are defining samples based on their accession.
- Names ending in "over Time Points" refer to running analysis steps on, and collecting results for *each group of samples*, where here each group represents a particular time point post infection.

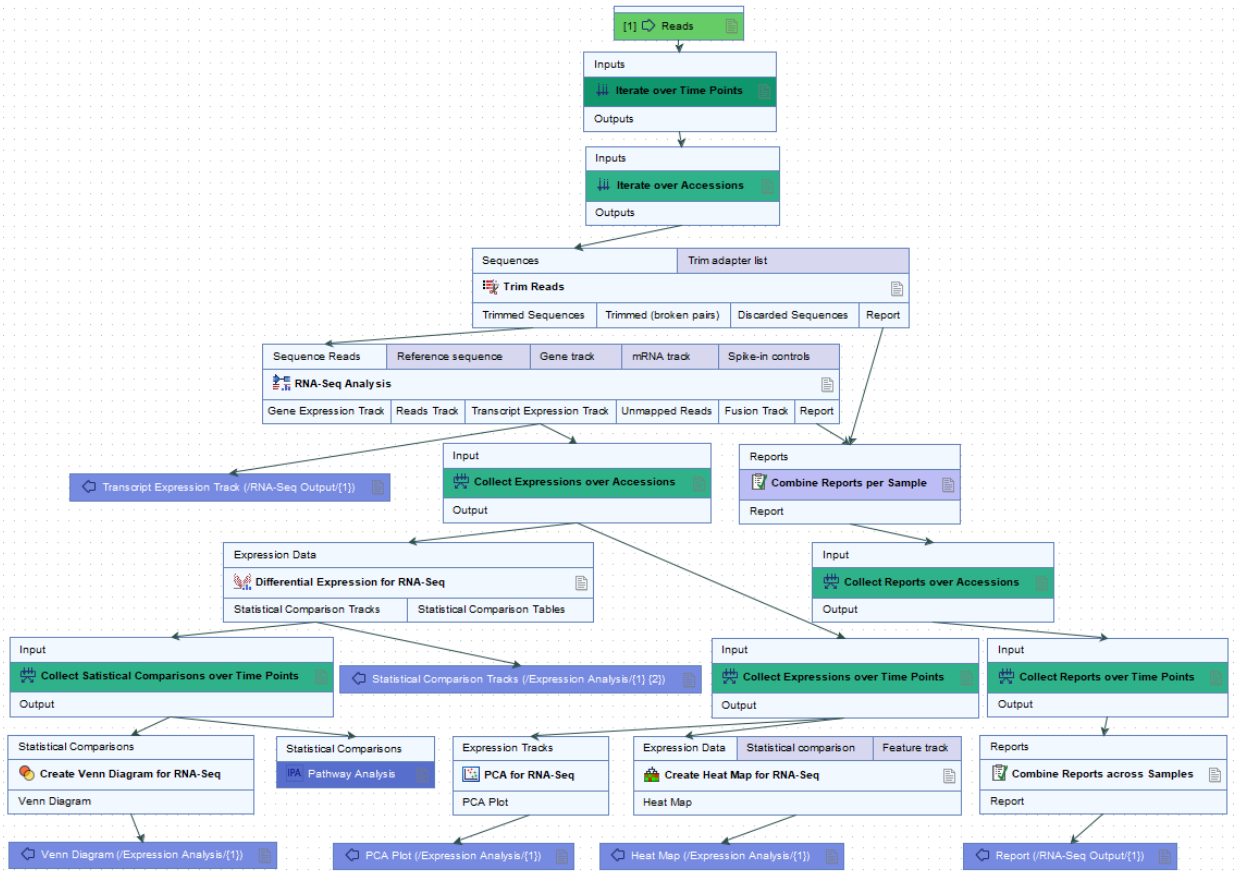

Figure 19: *"RNA-Seq and IPA advanced" workflow.*

The new **Iterate** and **Collect and Distribute** elements are the ones with "over Time Points" in their names.

A key benefit with this workflow design is that we can make use of results from multiple groups of samples in downstream steps of the same workflow. For example, here:

- We create a Venn diagram using the results from both the time points analyzed.
- We make a single submission to IPA with the results for both time points.
- We create a single PCA plot and a single heat map for the full data set. Previously we had to create separate PCA plots and heat maps for each time point.

Launching the "RNA-Seq and IPA advanced analysis workflow" is very much like launching the earlier workflow, except that:

- You should not check the "Batch" box.

  Checking the "Batch" box was needed to run the "RNA-Seq and IPA analysis workflow" once per time point, but in the "RNA-Seq and IPA advanced analysis workflow", the additional layer of **Iterate** and **Collect and Distribute** elements causes the appropriate sections of the workflow to be run once per time point.
- The iteration over time points needs to be configured in the workflow wizard and involves referring to the metadata table.

Feel free to launch the "RNA-Seq and IPA advanced analysis workflow" if you wish to.

Analyzing the full data set

If you wish to analyze the full data set, you can download the data from SRA directly in *CLC Genomics Workbench* (figure 20):

1. Go to:

   Download | () **Search for Reads in SRA**
2. Add PRJNA328591 in the search field.
3. Click () **Start search**.
4. Select all 12 experiments.
5. Click on () **Download Reads and Metadata**.

The downloaded metadata contains the "Run Accession", "Infected With" and "Time Point" needed to run the analysis.

Note that the tutorial data was down-sampled to approximately 3%. The full data is much larger and analyzing it will take considerably more time.

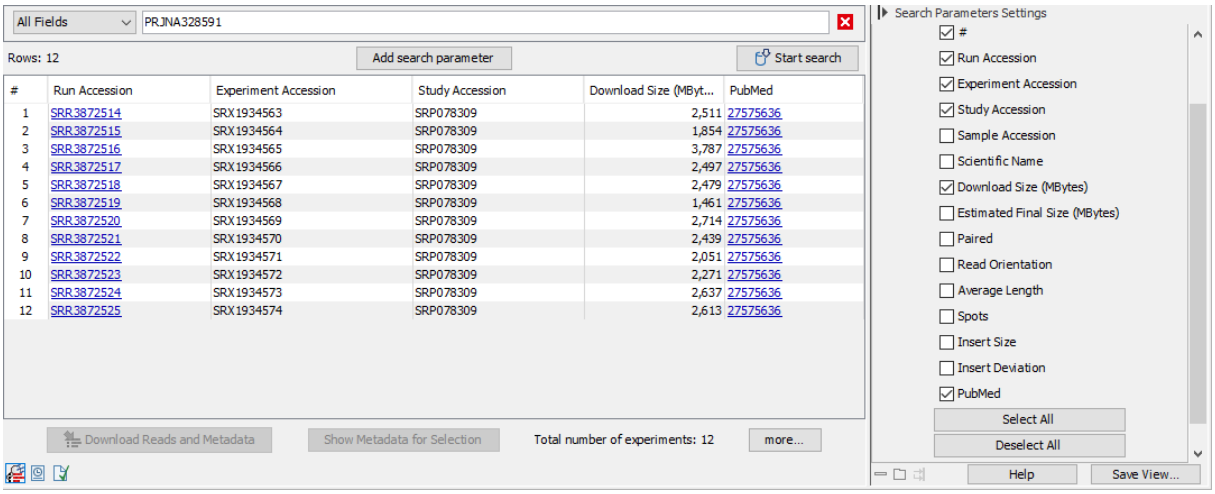

Figure 20: *The SRA results for project PRJNA328591.*