Tutorial

RNA-Seq and differential gene expression analysis

December 16, 2024

Sample to Insight

QIAGEN Aarhus A/S · Kalkværksvej 5, 11. · DK - 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

RNA-Seq and differential gene expression analysis

Introduction

The purpose of this tutorial is to illustrate how to harness the collaborative power of *CLC Genomics Workbench* and QIAGEN Ingenuity Pathway Analysis (IPA) for RNA-Seq data analysis and interpretation. We focus on the following:

- Import data.
- Analyze RNA-Seq data and identify differentially expressed genes using a template workflow.
- Interpret the results.
- Upload the results to IPA for biological interpretation.

Data used in this tutorial

This tutorial uses data from "The Dengue Virus NS5 Protein Intrudes in the Cellular Spliceosome and Modulates Splicing" by Maio et al 2016 (GSE84285, doi:10.1371/journal.ppat.1005841). The authors investigated transcriptional profiles of Dengue virus 2-infected cells.

The experimental setup included A549 human cells, either infected with **mock** or with **Dengue virus 2**, collected at two time points post infection (**24 hours** and **36 hours**). Three replicates were collected at each time point, resulting in a total of **12 samples**.

To complete the tutorial in a reasonable amount of time, only a subset of the reads mapping to chromosome 17 are used here. The data distributed for use with this tutorial contains:

- Fastq files containing reads from the 12 samples described above.
- An Excel file containing metadata for the samples.
- A reference sequence track for chromosome 17 of the human hg38 genome and corresponding gene, mRNA, and CDS tracks.
- A gene ontology table.
- A modified workflow for advanced RNA-Seq analysis including optional automatic upload to IPA.

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.

For uploading to IPA, you must have access to IPA and have the *Biomedical Genomics Analysis* plugin installed. You can request a free IPA trial by clicking on Request a trial. Installing plugins is described in the CLC Genomics Workbench manual.

General tips

- Throughout this tutorial, we provide links to relevant manual pages, which we recommend exploring for additional details.

- Tools and workflows can be found in the Toolbox, but it is often easier to launch them using Quick Launch (🚀), found in the top toolbar (shortcut Ctrl+Shift+T or ⌘ +Shift+T on Mac). Quick Launch displays the full Toolbox path, making it easy to identify the location of the tool or workflow if needed.

- The in-built manual can be accessed by clicking the **Help** button on wizards or by selecting the **Help** option under the **Help** menu.

- Within wizards, the **Reset** button can be used to change settings to their default values.

- Colors and gradients in plots can be changed by clicking on them in the Side Panel.

- Columns in tables can be hidden by unchecking 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, 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.

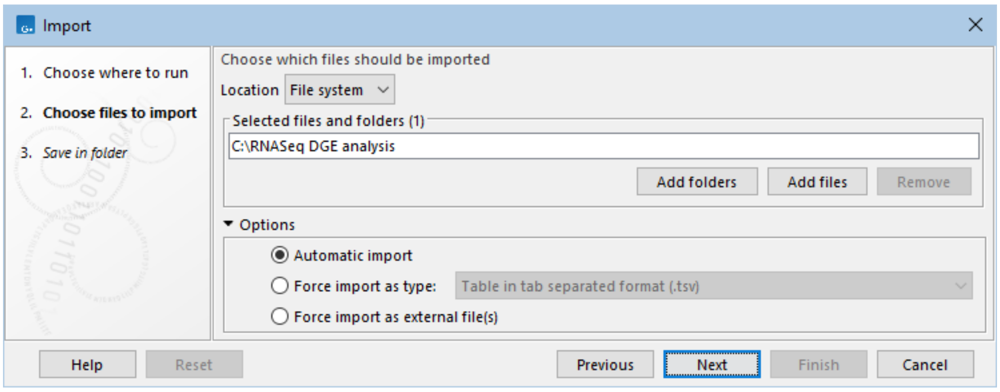
- Many data elements produced by *CLC Genomics Workbench* tools have multiple views, indicated as icons in the lower left corner of elements opened in the View Area. Clicking on one of the view icons while pressing the Ctrl (⌘ on Mac) key will open in split view such that both views are visible at the 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.

- Data can be imported prior to starting a workflow, or it can often be imported on the fly when the workflow is launched.

Import the data

We start by downloading and importing the tutorial data.

1. Download the tutorial data and unzip it.
2. Start the *CLC Genomics Workbench*.
3. Import the tutorial data:
   (a) Launch **Standard Import** (📥) using Quick Launch (🚀).
   (b) Locate the tutorial data using the **Add folders** button and select **Automatic import** (figure 1).


Import
1. Choose where to run
2. Choose files to import
3. Save in folder
Choose which files should be imported
Location: File system
Selected files and folders (1)
C:\RNASeq DGE analysis

Options
● Automatic import
○ Force import as type: Table in tab separated format (.tsv)
○ Force import as external file(s)

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

   (c) In the next step, select a suitable location in the Navigation Area to save the imported data and click on **Finish**.
   Since the tutorial data contains fastq files, a dialog will open (figure 2). Click on **Yes**. We will import these files separately in the next step. The imported data will be saved to a "RNASeq DGE analysis" folder in the selected location.

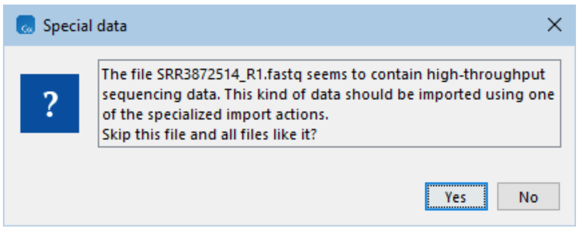
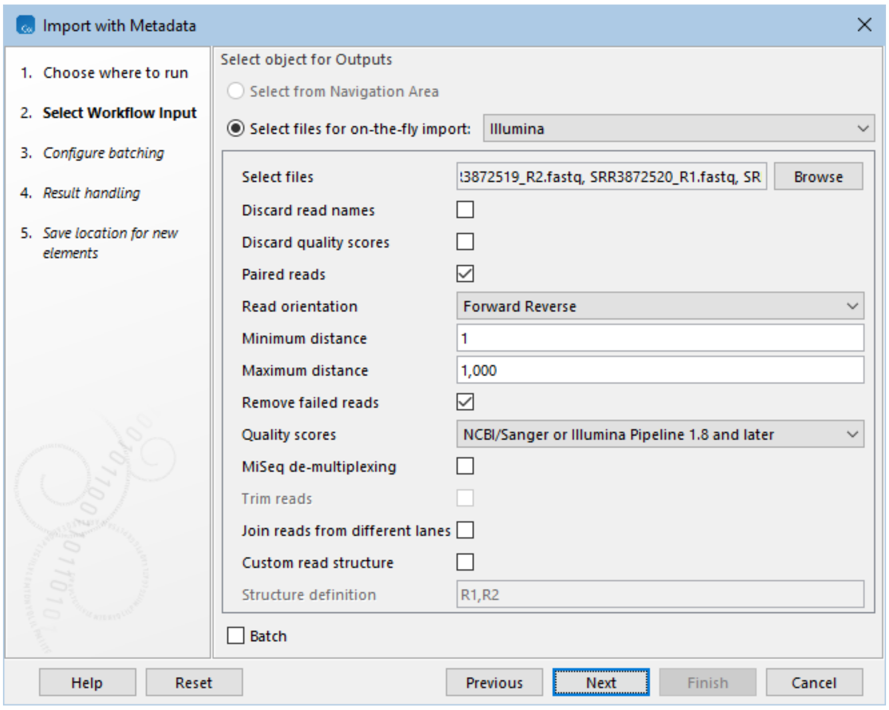

Special data
The file SRR3872514_R1.fastq seems to contain high-throughput sequencing data. This kind of data should be imported using one of the specialized import actions.
Skip this file and all files like it?

Figure 2: *A dialog opens when attempting to import fastq files with Standard Import.*

4. Import the tutorial fastq files:
   (a) Launch the Import with Metadata (📋) template workflow using Quick Launch (🚀).
   (b) In the first wizard step, "Select Workflow Input", the inputs must be specified (figure 3).
   Set "Select files for on-the-fly import" to **Illumina**.
   Click on the **Browse** button to the right of "Select files" and navigate to the tutorial data. Select the 24 fastq files located in the "Samples" subfolder.



Import with Metadata

1. Choose where to run
2. **Select Workflow Input**
3. *Configure batching*
4. *Result handling*
5. *Save location for new elements*

Select object for Outputs

○ Select from Navigation Area

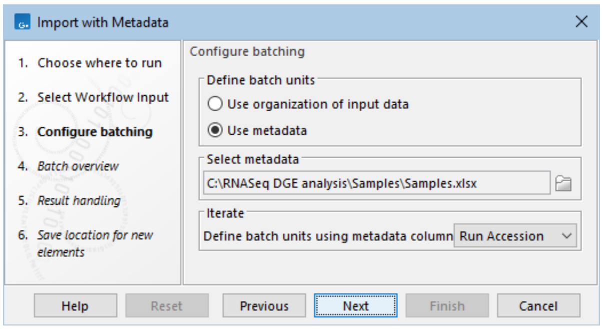
● Select files for on-the-fly import: Illumina

- Select files: :3872519_R2.fastq, SRR3872520_R1.fastq, SR [Browse]
- Discard read names: □
- Discard quality scores: □
- Paired reads: ☑
- Read orientation: Forward Reverse
- Minimum distance: 1
- Maximum distance: 1,000
- Remove failed reads: ☑
- Quality scores: NCBI/Sanger or Illumina Pipeline 1.8 and later
- MiSeq de-multiplexing: □
- Trim reads: □
- Join reads from different lanes: □
- Custom read structure: □
- Structure definition: R1,R2

□ Batch

Figure 3: *Import with Metadata configured to import the tutorial fastq files.*

(c) In the next step, "Configure batching", choose the **Use metadata** option (figure 4). Click on the (📁) button to the right of "Select metadata", navigate to the tutorial data, and select the "Samples.xlsx" file located in the "Samples" subfolder.
Under "Iterate", set "Define batch units using metadata column" to **Run Accession**.



Import with Metadata

1. Choose where to run
2. Select Workflow Input
3. **Configure batching**
4. *Batch overview*
5. *Result handling*
6. *Save location for new elements*

Configure batching

Define batch units
- ○ Use organization of input data
- ● Use metadata

Select metadata: C:\RNASeq DGE analysis\Samples\Samples.xlsx

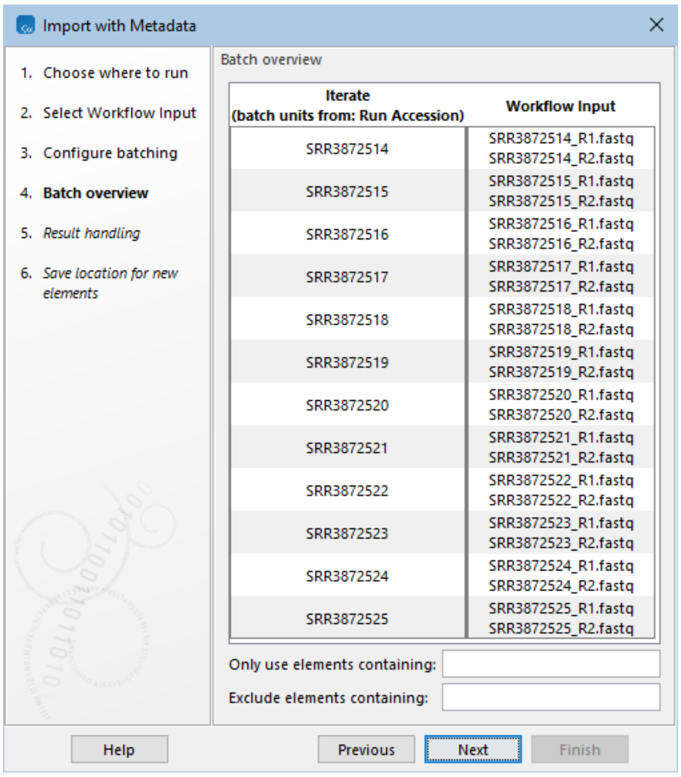
Iterate: Define batch units using metadata column: Run Accession

Figure 4: *Import with Metadata configured to use the "Samples.xlsx" file for batching.*

(d) The next step, "Batch overview", shows that 12 batch units are being imported, corresponding to the 12 samples (figure 5).
If the organization is not as expected, clicking on **Previous** returns to the "Configure batching" step where options can be adjusted.

(e) In the next step, "Result handling", choose **Save**.
Ensure that **Create workflow result metadata** is checked. This will create a `Workflow Result Metadata table` with information about the results.

(f) In the final step, choose to save the fastq files in the "Samples" subfolder of "RNASeq DGE analysis" in the Navigation Area.
Click on **Finish**.



Import with Metadata

1. Choose where to run
2. Select Workflow Input
3. Configure batching
4. **Batch overview**
5. *Result handling*
6. *Save location for new elements*

Batch overview

Iterate (batch units from: Run Accession)	Workflow Input
SRR3872514	SRR3872514_R1.fastq SRR3872514_R2.fastq
SRR3872515	SRR3872515_R1.fastq SRR3872515_R2.fastq
SRR3872516	SRR3872516_R1.fastq SRR3872516_R2.fastq
SRR3872517	SRR3872517_R1.fastq SRR3872517_R2.fastq
SRR3872518	SRR3872518_R1.fastq SRR3872518_R2.fastq
SRR3872519	SRR3872519_R1.fastq SRR3872519_R2.fastq
SRR3872520	SRR3872520_R1.fastq SRR3872520_R2.fastq
SRR3872521	SRR3872521_R1.fastq SRR3872521_R2.fastq
SRR3872522	SRR3872522_R1.fastq SRR3872522_R2.fastq
SRR3872523	SRR3872523_R1.fastq SRR3872523_R2.fastq
SRR3872524	SRR3872524_R1.fastq SRR3872524_R2.fastq
SRR3872525	SRR3872525_R1.fastq SRR3872525_R2.fastq

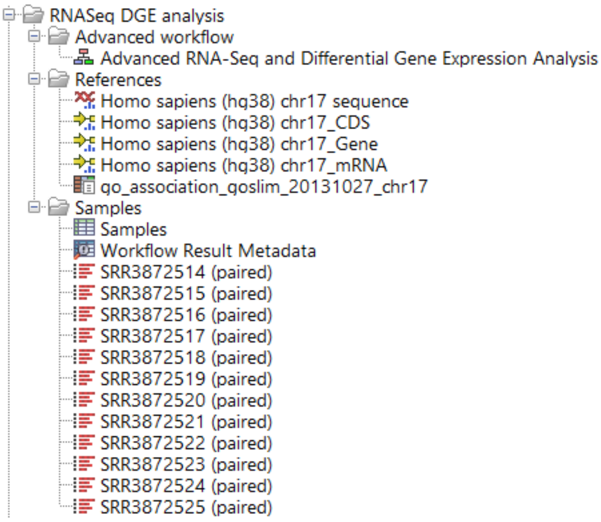
Only use elements containing:

Exclude elements containing:

Figure 5: *The Batch overview shows how the selected files are grouped in batch units.*

Once the import is completed, the folders and data elements are visible in the Navigation Area (figure 6).

The tools of *CLC Genomics Workbench* produce deterministic results, but changing the input 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.



RNASeq DGE analysis
- Advanced workflow
  - Advanced RNA-Seq and Differential Gene Expression Analysis
- References
  - Homo sapiens (hg38) chr17 sequence
  - Homo sapiens (hg38) chr17_CDS
  - Homo sapiens (hg38) chr17_Gene
  - Homo sapiens (hg38) chr17_mRNA
  - go_association_goslim_20131027_chr17
- Samples
  - Samples
  - Workflow Result Metadata
  - SRR3872514 (paired)
  - SRR3872515 (paired)
  - SRR3872516 (paired)
  - SRR3872517 (paired)
  - SRR3872518 (paired)
  - SRR3872519 (paired)
  - SRR3872520 (paired)
  - SRR3872521 (paired)
  - SRR3872522 (paired)
  - SRR3872523 (paired)
  - SRR3872524 (paired)
  - SRR3872525 (paired)

Figure 6: *The "RNASeq DGE analysis" folder in the Navigation Area after importing the tutorial data.*

Analyze the data

We will now use the RNA-Seq and Differential Gene Expression Analysis template workflow to analyze the tutorial data. If you run this workflow on your own data, please note that template workflows are provided as example workflows and may need to be customized to meet the specific requirements of your data. Specifically, for the RNA-Seq and Differential Gene Expression Analysis workflow, be aware that it does not contain trimming, and you should therefore ensure that your reads are trimmed beforehand, e.g., by using the Prepare Raw Data template workflow.

To see the content of the RNA-Seq and Differential Gene Expression Analysis workflow, locate the workflow in the Toolbox:

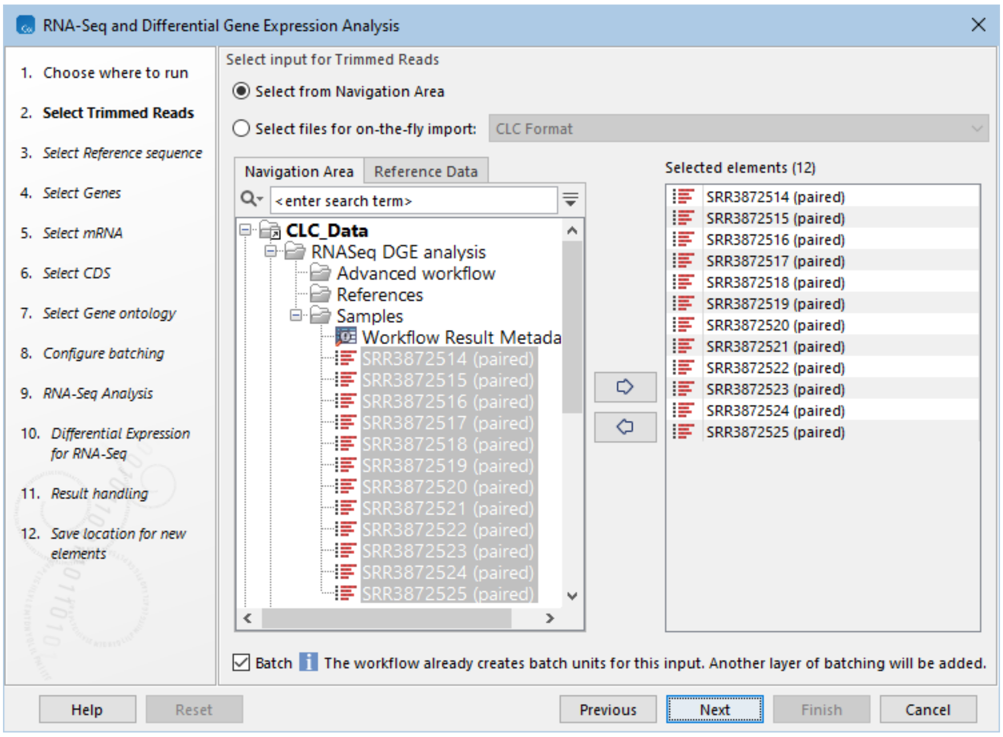
> **Workflows | Template Workflows | Basic Workflow Designs ( ) | RNA-Seq and Differential Gene Expression Analysis ( )**

right-click on its name and choose **Open Copy of Workflow**.

To run the workflow:

1. Launch the workflow using Quick Launch ( ).

2. In the first wizard step, "Select Trimmed Reads", specify the data to be analyzed by selecting the 12 samples in the "Samples" folder (figure 7). Make sure to check the **Batch** option below the data selection area.



RNA-Seq and Differential Gene Expression Analysis

1. Choose where to run
2. **Select Trimmed Reads**
3. Select Reference sequence
4. Select Genes
5. Select mRNA
6. Select CDS
7. Select Gene ontology
8. Configure batching
9. RNA-Seq Analysis
10. Differential Expression for RNA-Seq
11. Result handling
12. Save location for new elements

Select input for Trimmed Reads

● Select from Navigation Area

○ Select files for on-the-fly import: CLC Format

Selected elements (12)

SRR3872514 (paired)
SRR3872515 (paired)
SRR3872516 (paired)
SRR3872517 (paired)
SRR3872518 (paired)
SRR3872519 (paired)
SRR3872520 (paired)
SRR3872521 (paired)
SRR3872522 (paired)
SRR3872523 (paired)
SRR3872524 (paired)
SRR3872525 (paired)

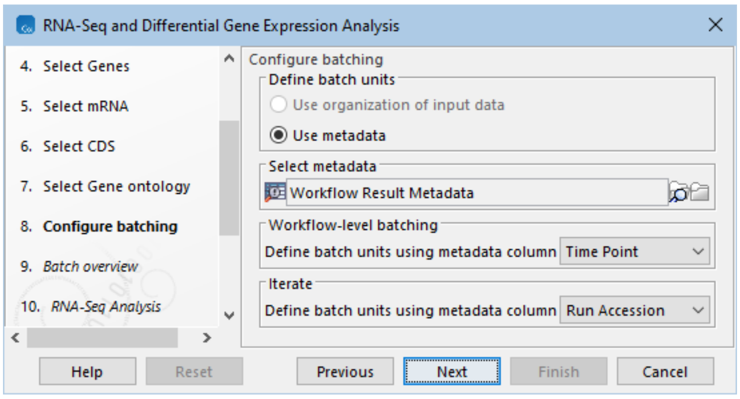
☑ Batch The workflow already creates batch units for this input. Another layer of batching will be added.

Figure 7: *All 12 samples are used as input. The batch option must be checked, leading to an info text appearing.*

An info text appears, indicating that the workflow already creates one layer of batch units and that checking the Batch option will create another layer of batching. This is precisely our intention, since we want to run the entire workflow using the 24 hour time point samples and then again using the 36 hour time point samples.

3. In the next five steps, select the **Reference sequence**, **Genes**, **mRNA**, **CDS**, and **Gene ontology** elements to use from the "References" folder in the Navigation Area.

4. In the next step, "Configure batching", define the batch units using the metadata table created previously in step 4e (figure 8).



Configure batching

Define batch units

○ Use organization of input data

● Use metadata

Select metadata: Workflow Result Metadata

Workflow-level batching: Define batch units using metadata column: Time Point

Iterate: Define batch units using metadata column: Run Accession

Figure 8: *The "Workflow Result Metadata" table is used to set up batching.*

Click on the (🔍) button at the right-hand side of "Select metadata" and choose the "Workflow Result Metadata" table from the "Samples" folder.

Under "Workflow-level batching", set "Define batch units using metadata column" to **Time Point**. This sets the workflow to run twice, once for each value in the Time Point metadata column ("24 hours post infection" and "36 hours post infection"), using input reads associated with metadata rows that have the corresponding value in this column.

Under "Iterate", set "Define batch units using metadata column" to **Run Accession**. This defines the iteration block, here corresponding to the sample, for each workflow run: input reads that are associated with metadata rows that have the same value in the Run Accession metadata column. Input reads forming an iteration block are analyzed together in the tools downstream of Iterate and upstream of Collect and Distribute.

5. The next step, "Batch overview", shows the organization of the input reads (figure 9).

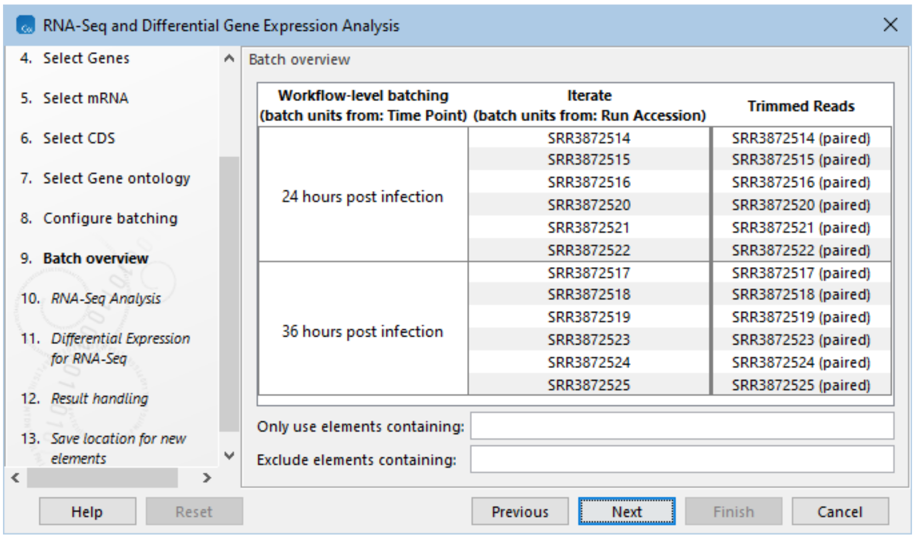
   If the organization is not as expected, clicking on **Previous** returns to the "Configure batching" step where options can be adjusted.

6. In the next step, "RNA-Seq Analysis", keep the default settings.

7. In the next step, "Differential Expression for RNA-Seq", set the following options (figure 10):

   - "Test differential expression due to" to **Infected With**.
   - "Comparisons" to **Against control group**.
   - "Control group" to **mock**.

   This specifies that we want to test differential expression between samples with different "Infected With" values, using samples labeled "mock" as the control group for comparison.



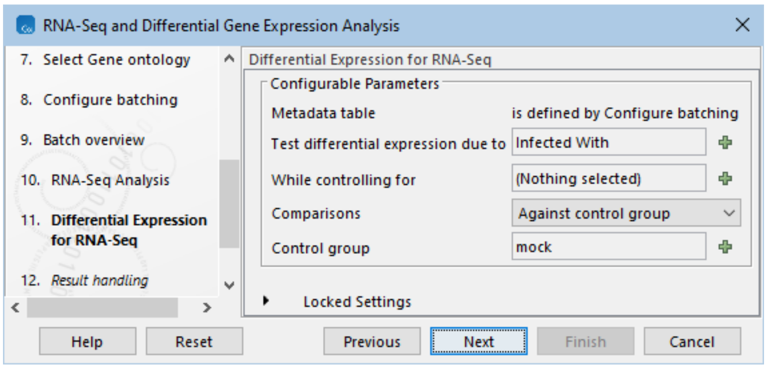
Batch overview

Workflow-level batching (batch units from: Time Point)	Iterate (batch units from: Run Accession)	Trimmed Reads
24 hours post infection	SRR3872514	SRR3872514 (paired)
	SRR3872515	SRR3872515 (paired)
	SRR3872516	SRR3872516 (paired)
	SRR3872520	SRR3872520 (paired)
	SRR3872521	SRR3872521 (paired)
	SRR3872522	SRR3872522 (paired)
36 hours post infection	SRR3872517	SRR3872517 (paired)
	SRR3872518	SRR3872518 (paired)
	SRR3872519	SRR3872519 (paired)
	SRR3872523	SRR3872523 (paired)
	SRR3872524	SRR3872524 (paired)
	SRR3872525	SRR3872525 (paired)

Only use elements containing:

Exclude elements containing:

Figure 9: *The Batch overview shows how the input reads are grouped in batch units.*



Differential Expression for RNA-Seq

Configurable Parameters

Metadata table: is defined by Configure batching

Test differential expression due to: Infected With

While controlling for: (Nothing selected)

Comparisons: Against control group

Control group: mock

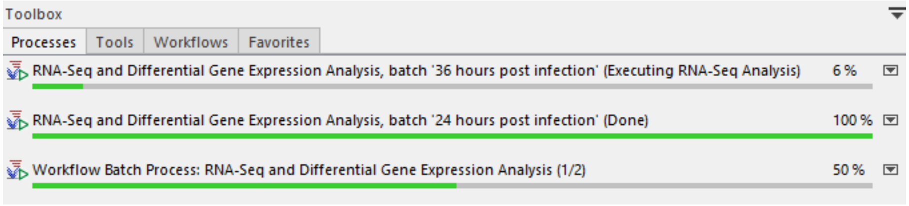
Figure 10: *Differential expression is tested due to "Infected With", using the "mock" group as control.*

8. In the next step, "Result handling", check the **Create subfolders per batch unit** and **Create workflow result metadata** options. This will create two subfolders for each of the time points, and a Workflow Result Metadata table.

9. In the last step, make a new subfolder in "RNASeq DGE analysis" called "Results" and choose to save the workflow results there.

   Click on **Finish**.

   The workflow will now execute. The progress can be monitored under the Processes tab in the Toolbox (figure 11). It will take some time for the workflow to run to completion.



RNA-Seq and Differential Gene Expression Analysis, batch '36 hours post infection' (Executing RNA-Seq Analysis) 6 %

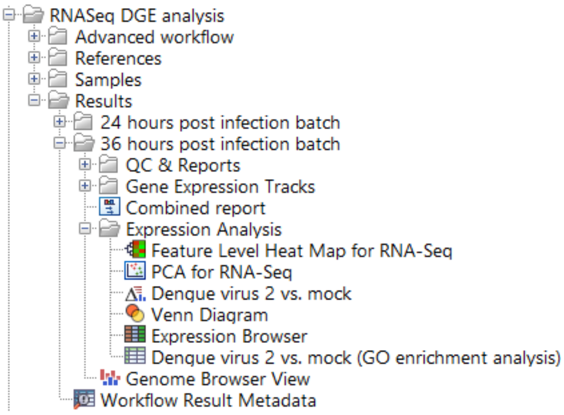
RNA-Seq and Differential Gene Expression Analysis, batch '24 hours post infection' (Done) 100 %

Workflow Batch Process: RNA-Seq and Differential Gene Expression Analysis (1/2) 50 %

Figure 11: *The "Processes" tab indicates how far the workflow execution has progressed.*

Interpret the results

Results from the workflow are placed in the "Results" folder (figure 12).



- RNASeq DGE analysis
  - Advanced workflow
  - References
  - Samples
  - Results
    - 24 hours post infection batch
    - 36 hours post infection batch
      - QC & Reports
      - Gene Expression Tracks
      - Combined report
      - Expression Analysis
        - Feature Level Heat Map for RNA-Seq
        - PCA for RNA-Seq
        - Dengue virus 2 vs. mock
        - Venn Diagram
        - Expression Browser
        - Dengue virus 2 vs. mock (GO enrichment analysis)
      - Genome Browser View
    - Workflow Result Metadata

Figure 12: *The "Results" folder in the Navigation Area.*

The results from each batch unit are saved in a separate subfolder, named after the batch units indicated in the "Batch overview" (figure 9). Within each of these folders are a `Combined report`, a `Genome Browser View`, and the following subfolders:

- **QC & Reports**: Contains a `QC Report` and an `RNA-Seq Report` for each of the samples.
- **Gene Expression Tracks**: Contains a `gene expression track` for each of the samples.
- **Expression Analysis**: Contains, among other elements, `differential expression analysis results`, a `PCA plot`, and a `feature level heat map`.

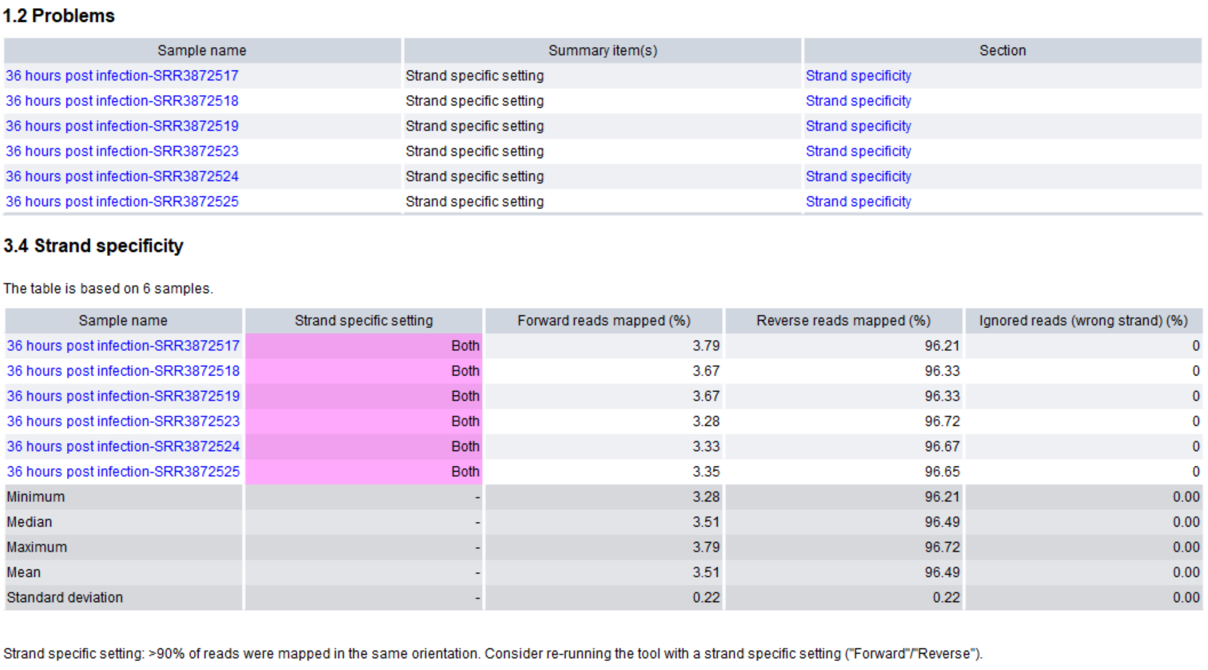
Additionally, a single `Workflow Result Metadata table` is generated. This can be particularly useful for finding all the output elements of a workflow when there are many batch runs involving many outputs.

Combined report

Open one of the **Combined reports** ().

The report summarizes information from the samples in that batch unit. It can be used to quickly review the results of `QC for Sequencing Reads` and `RNA-Seq Analysis`.

Samples identified as potentially problematic are listed in the Summary section and highlighted in the remaining sections in yellow or red. Here, the report shows that the "Strand specific setting" might have been set incorrectly (figure 13). This is indeed the case, since the workflow was run with "Strand specific setting" set to "Both" although the tutorial data was actually generated using a KAPA stranded RNA-Seq library kit. Ideally, the workflow should be re-run, setting the "Strand specific" option to **Reverse** in step 6. However, we will proceed with the current results, as the incorrect setting does not significantly impact the outcome here.



1.2 Problems

Sample name	Summary item(s)	Section
36 hours post infection-SRR3872517	Strand specific setting	Strand specificity
36 hours post infection-SRR3872518	Strand specific setting	Strand specificity
36 hours post infection-SRR3872519	Strand specific setting	Strand specificity
36 hours post infection-SRR3872523	Strand specific setting	Strand specificity
36 hours post infection-SRR3872524	Strand specific setting	Strand specificity
36 hours post infection-SRR3872525	Strand specific setting	Strand specificity

3.4 Strand specificity

The table is based on 6 samples.

Sample name	Strand specific setting	Forward reads mapped (%)	Reverse reads mapped (%)	Ignored reads (wrong strand) (%)
36 hours post infection-SRR3872517	Both	3.79	96.21	0
36 hours post infection-SRR3872518	Both	3.67	96.33	0
36 hours post infection-SRR3872519	Both	3.67	96.33	0
36 hours post infection-SRR3872523	Both	3.28	96.72	0
36 hours post infection-SRR3872524	Both	3.33	96.67	0
36 hours post infection-SRR3872525	Both	3.35	96.65	0
Minimum	-	3.28	96.21	0.00
Median	-	3.51	96.49	0.00
Maximum	-	3.79	96.72	0.00
Mean	-	3.51	96.49	0.00
Standard deviation	-	0.22	0.22	0.00

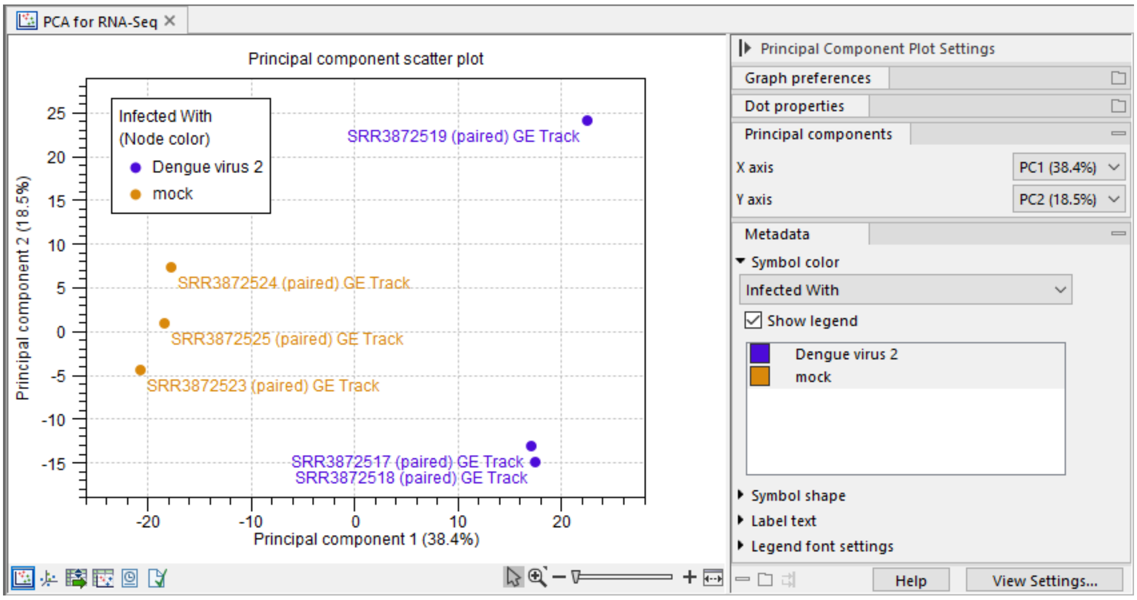
Strand specific setting: >90% of reads were mapped in the same orientation. Consider re-running the tool with a strand specific setting ("Forward"/"Reverse").

Figure 13: *The "Problems" section of a Combined report showing that the wrong "Strand specific setting" might have been used. Clicking on the blue "Strand specificity" text navigates to section 3.4 where the "Strand specific setting" can be inspected further.*

PCA plot and heat map

The PCA plots and heat maps provide a visual overview, helping to identify outliers among the samples and any interesting patterns in the data.

Open one of the **PCA for RNA-Seq** (📈) plots (figure 14), found in the "Expression Analysis" folders.


PCA for RNA-Seq
Principal component scatter plot
Infected With
(Node color)
● Dengue virus 2
● mock
SRR3872519 (paired) GE Track
SRR3872524 (paired) GE Track
SRR3872525 (paired) GE Track
SRR3872523 (paired) GE Track
SRR3872517 (paired) GE Track
SRR3872518 (paired) GE Track
Principal component 2 (18.5%)
25
20
15
10
5
0
-5
-10
-15
-20
-10
0
10
20
Principal component 1 (38.4%)
Principal Component Plot Settings
Graph preferences
Dot properties
Principal components
Metadata
Symbol shape
Label text
Legend font settings
Help
View Settings...

X axis: PC1 (38.4%)
Y axis: PC2 (18.5%)
Symbol color: Infected With
☑ Show legend
🟦 Dengue virus 2
🟧 mock

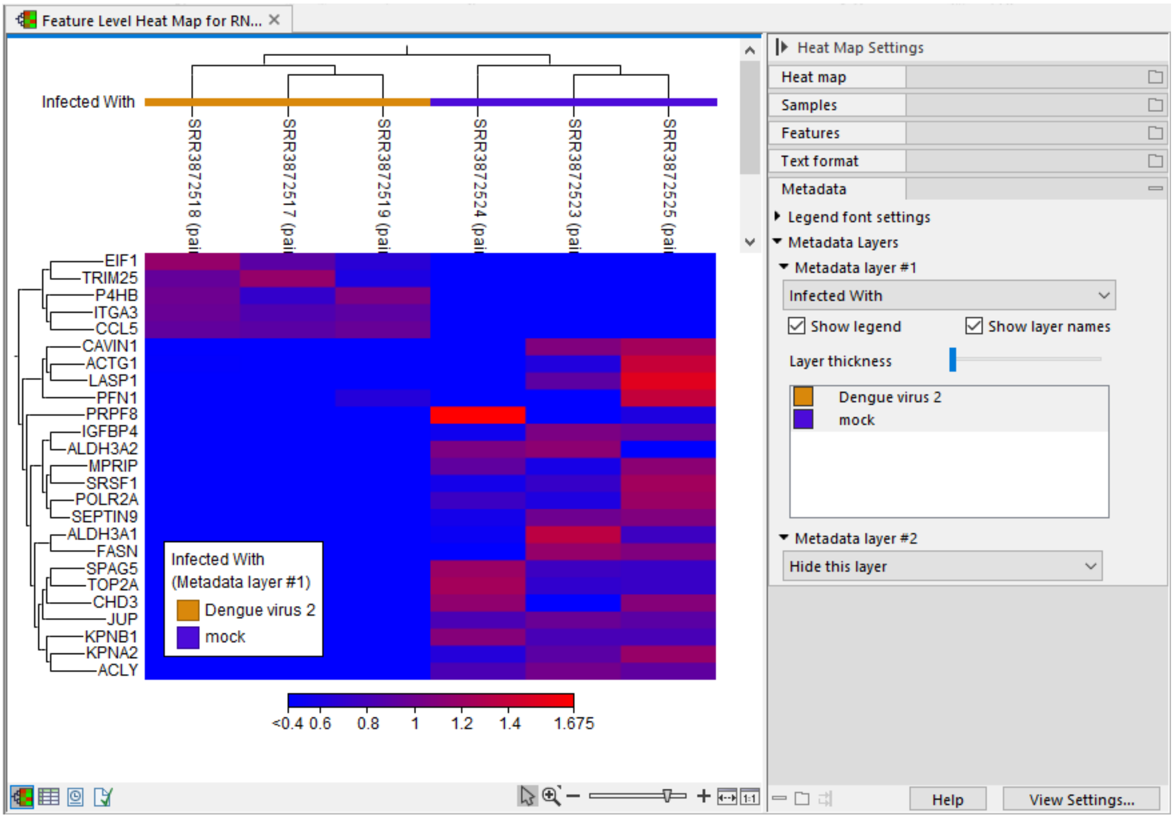
Figure 14: *PCA plot from the 36 hour post infection batch with samples colored by infection type.*

The plot is colored by "Infected With". The legend can be moved by clicking and dragging, and the view settings can be adjusted using the `Side Panel`. Note that the plot can also be visualized in `3D` by clicking on the ( ) icon.

We can see that the samples cluster by infection type (mock or Dengue virus 2), as expected. One of the "Dengue virus 2" samples (SRR3872519) is positioned farther from the other two "Dengue virus 2" samples, suggesting it may be an outlier. However, proper outlier detection requires more than three samples, and we will therefore proceed with our analysis without removing the sample.

Open one of the **Feature Level Heat Map for RNA-Seq** ( ) plots (figure 15), found in the "Expression Analysis" folders.

The vertical axis of the plot shows the 25 "most interesting" features (genes), i.e. those with the largest variance in expression value across the samples. The horizontal axis shows unsupervised clustering of the samples based on these genes. The legend can be moved by clicking and dragging, and the view settings can be adjusted using the `Side Panel`. To visualize that the samples cluster by infection type, add **Infected With** as a metadata layer in the "Metadata" palette.



Metadata layer #1: Infected With

☑ Show legend ☑ Show layer names

Metadata layer #2: Hide this layer

Figure 15: *Heat map from the 36 hour post infection batch with the "Infected With" metadata layer.*

Interestingly, the heat map shows that samples SRR3872519 and SRR3872517 cluster together, contradicting what we observed in the PCA plot. This may be due to the fact that the heat map clustering is based on only 25 genes, whereas the PCA plot is based on all genes.

Statistical comparison

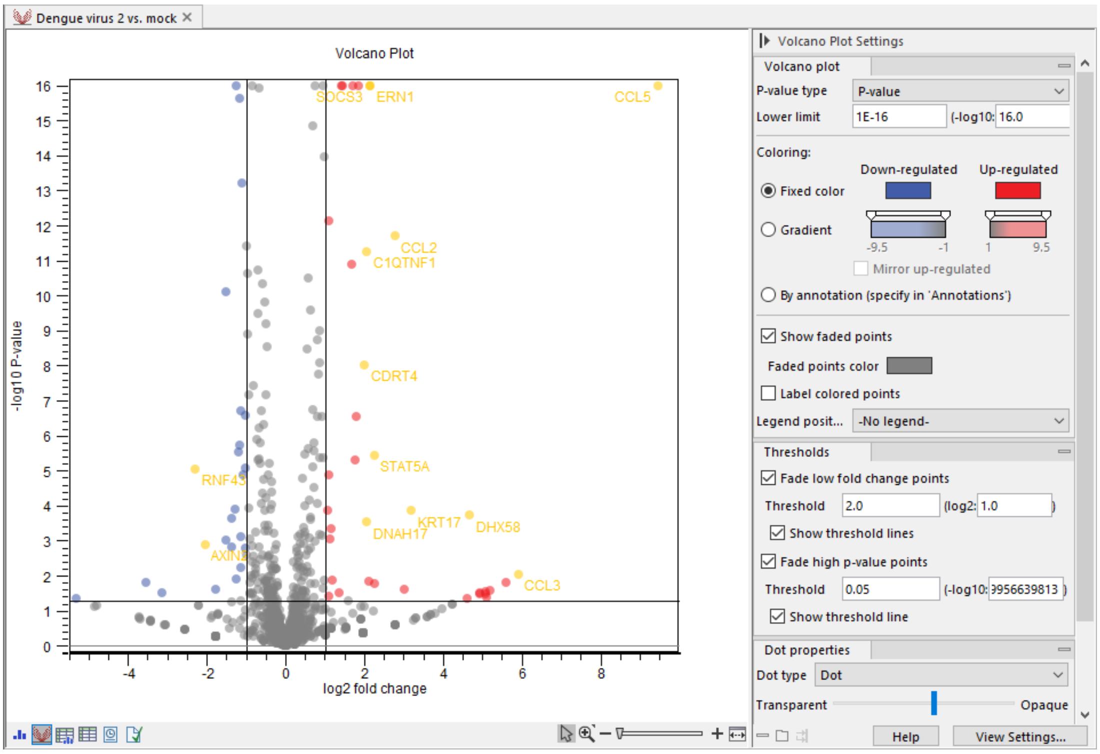
Results from the differential expression analysis can be used to identify genes with significant changes in expression, providing insights into biological differences between groups of samples.

Open one of the **Dengue virus 2 vs. mock** ( ) statistical comparison tracks, found in the "Expression Analysis" folders. The table view ( ) is shown by default.

We will first investigate the volcano plot. Click on the ( ) icon at the bottom of the view to open this plot. Improve the visibility by adjusting settings in the Side Panel (figure 16):

1. In the "Volcano plot" palette, set "P-value type" to **P-value**.
2. In the "Thresholds" palette, check the **Fade low fold change points** and **Fade high p-value points** options.
3. In the "Dot properties" palette, click-and-drag the bar for adjusting the dot transparency.

Gene names for points of interest can be viewed by hovering over them, which displays a tooltip with the name, or by clicking, which selects the point and shows the name. Alternatively, select multiple points by holding Ctrl (⌘) and clicking the points, or clicking and dragging to create a lasso selection.



Volcano Plot Settings

Volcano plot

P-value type: P-value

Lower limit: 1E-16 (-log10: 16.0

Coloring:

Down-regulated Up-regulated

● Fixed color

○ Gradient -9.5 -1 1 9.5

□ Mirror up-regulated

○ By annotation (specify in 'Annotations')

☑ Show faded points

Faded points color

□ Label colored points

Legend posit...: -No legend-

Thresholds

☑ Fade low fold change points

Threshold 2.0 (log2: 1.0)

☑ Show threshold lines

☑ Fade high p-value points

Threshold 0.05 (-log10: 956639813)

☑ Show threshold line

Dot properties

Dot type: Dot

Transparent Opaque

Figure 16: *Volcano plot from the 36 hour post infection batch. Gene names are displayed for points with absolute log2 fold change >2 and FDR p-value <0.01. Points with low fold changes or high p-values are faded.*

Comparison across time points

The results we have produced so far offer a rather limited view of the data, where the 24 and 36 hour time points cannot be compared. We will now create such a comparison.

Venn diagram across time points

First, we will create a Venn diagram of the statistical comparison tracks from the 24 and 36 hour time points, in order to compare differentially expressed genes between them.

1. Rename the two statistical comparison tracks to "Dengue virus 2 vs. mock (24 hours)" and "Dengue virus 2 vs. mock (36 hours)" by right-clicking on them in the Navigation Area and selecting **Rename** or by pressing F2. This is necessary because **Create Venn Diagram for RNA-Seq** does not accept two input elements with the same name.
2. Launch `Create Venn Diagram for RNA-Seq` () using Quick Launch ().
3. In the next step, "Select at least two statistical comparisons", select the two renamed statistical comparison tracks as input.
4. In the next step, "Result handling", select **Save**.
5. In the last step, choose to save the results in the "Results" folder and click on **Finish**.

The resulting Venn diagram shows how many genes are differentially expressed in the two statistical comparisons, and how many of these overlap. To further investigate the overlapping genes:

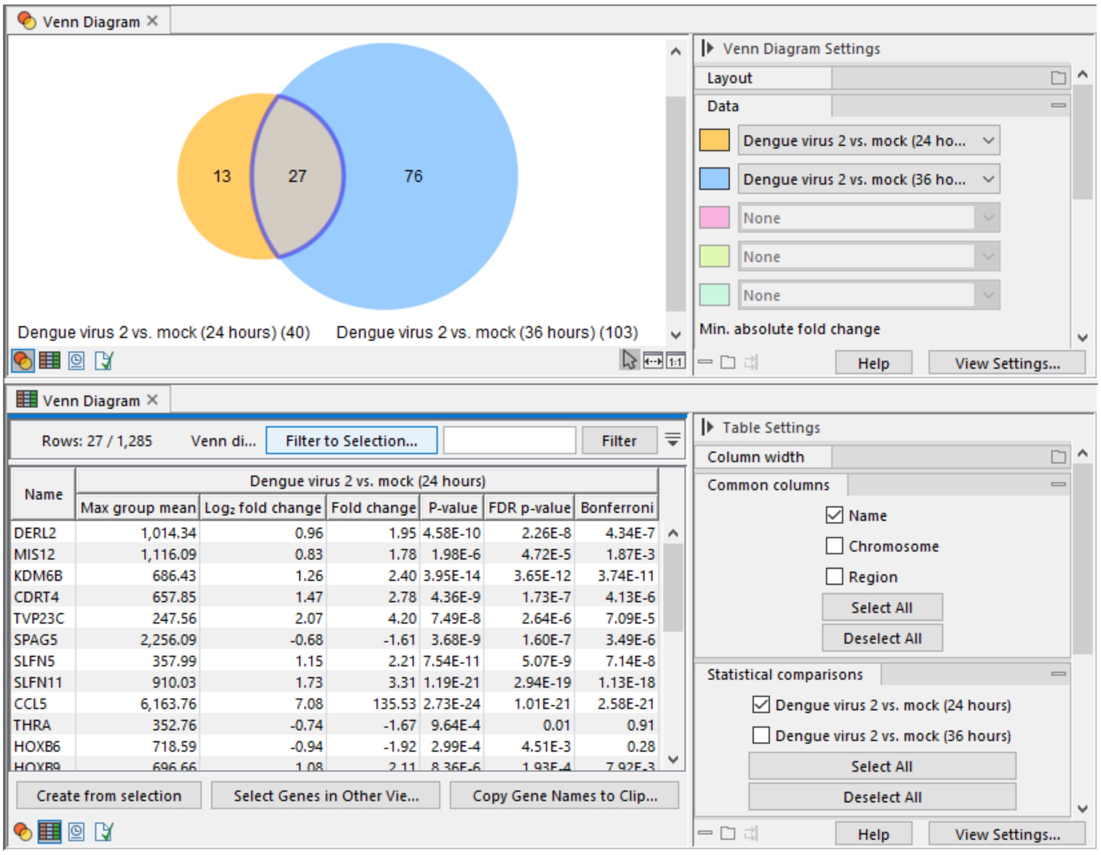
1. Select the intersection in the Venn diagram by clicking on it (figure 17).
2. Go to the table view () of the Venn diagram. The same genes will still be selected in the table. It may be hard to find them among all the genes, though.
3. To show only the selected genes, click on **Filter to Selection** at the top of the table and choose **Filter to selected rows** (figure 17).

PCA plot and heat map across time points

Next, we will create a PCA plot and a heat map based on all expression tracks from both the 24 and 36 hour time point to see how the samples cluster across time points.

To create the PCA plot across time points:

1. Launch `PCA for RNA-Seq` () using Quick Launch ().
2. In the next step, "Select at least 2 expression samples", right-click on the "Results" folder and choose **Add folder contents (recursively)** to add all 12 expression tracks.
3. In the next step, "Result handling", select **Save**.
4. In the last step, choose to save the results in the "Results" folder and click on **Finish**.



🟧 Dengue virus 2 vs. mock (24 ho...
🟦 Dengue virus 2 vs. mock (36 ho...
None
None
None

Rows: 27 / 1,285 Venn di...

Name	Dengue virus 2 vs. mock (24 hours)					
	Max group mean	$\log_2$ fold change	Fold change	P-value	FDR p-value	Bonferroni
DERL2	1,014.34	0.96	1.95	4.58E-10	2.26E-8	4.34E-7
MIS12	1,116.09	0.83	1.78	1.98E-6	4.72E-5	1.87E-3
KDM6B	686.43	1.26	2.40	3.95E-14	3.65E-12	3.74E-11
CDRT4	657.85	1.47	2.78	4.36E-9	1.73E-7	4.13E-6
TVP23C	247.56	2.07	4.20	7.49E-8	2.64E-6	7.09E-5
SPAG5	2,256.09	-0.68	-1.61	3.68E-9	1.60E-7	3.49E-6
SLFN5	357.99	1.15	2.21	7.54E-11	5.07E-9	7.14E-8
SLFN11	910.03	1.73	3.31	1.19E-21	2.94E-19	1.13E-18
CCL5	6,163.76	7.08	135.53	2.73E-24	1.01E-21	2.58E-21
THRA	352.76	-0.74	-1.67	9.64E-4	0.01	0.91
HOXB6	718.59	-0.94	-1.92	2.99E-4	4.51E-3	0.28
HOXB9	696.66	1.08	2.11	8.36E-6	1.93E-4	7.92E-3

☑ Name
☐ Chromosome
☐ Region

☑ Dengue virus 2 vs. mock (24 hours)
☐ Dengue virus 2 vs. mock (36 hours)

Figure 17: *Top: Venn diagram. Bottom: Table view of the same diagram, showing only the selected genes. Note that some columns have been unchecked in the Side Panel and are therefore not visible.*

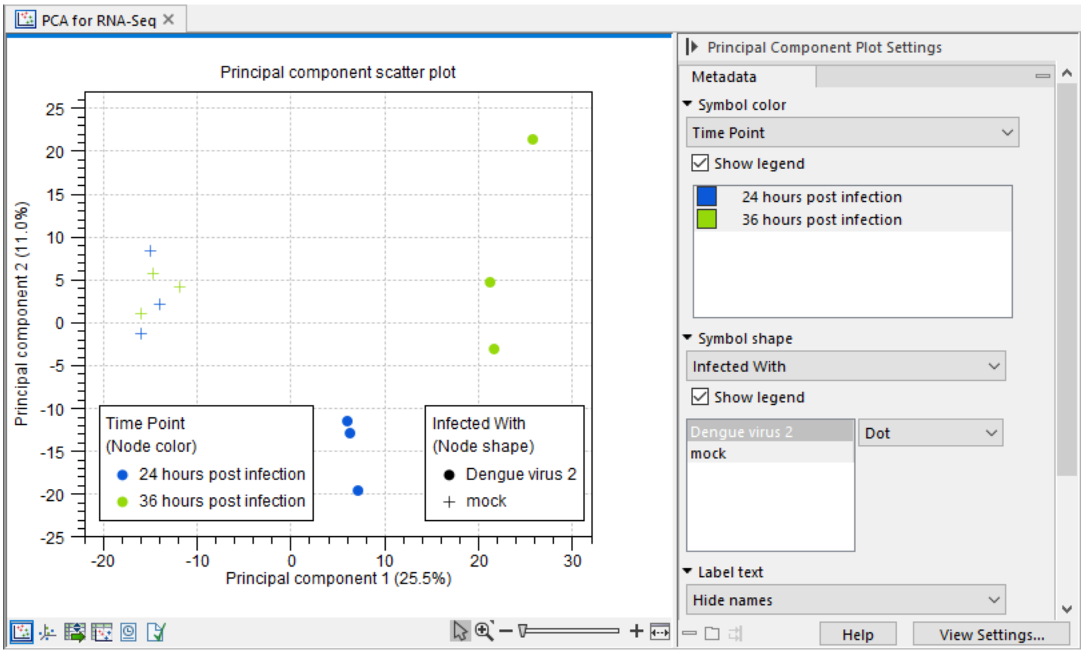
To visualize if the samples cluster across time points, configure the following options in the "Metadata" Side Panel palette of the resulting PCA plot:

1. Set "Symbol color" to **Time Point**.
2. Set "Symbol shape" to **Infected With**.
3. Above "Label text", click on "Dengue virus 2" to set the symbol shape for these samples, and choose **Dot**. Repeat for "mock" and set the symbol shape to **Plus**.
4. Under "Label text", select **Hide names**.

The samples cluster most strongly by infection type, which can be seen on the X-axis where the first principal component is displayed. Furthermore, the Dengue virus 2-infected cells cluster by time point, indicating 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 18).

To create the heat map across time points:

1. Launch `Create Feature Level Heat Map for RNA-Seq` ( ) using Quick Launch ( ).


PCA for RNA-Seq
Principal component scatter plot
Principal component 2 (11.0%)
Principal component 1 (25.5%)
25
20
15
10
5
0
-5
-10
-15
-20
-25
-20
-10
0
10
20
30
Time Point
(Node color)
● 24 hours post infection
● 36 hours post infection
Infected With
(Node shape)
● Dengue virus 2
+ mock
Principal Component Plot Settings
Metadata
Help
View Settings...

Symbol color

Time Point

☑ Show legend

🟦 24 hours post infection
🟩 36 hours post infection

Symbol shape

Infected With

☑ Show legend

Dengue virus 2 — Dot
mock

Label text

Hide names

Figure 18: *PCA plot showing the samples, with color determined by "Time Point" and shape determined by "Infected With". Label text has been hidden.*

2. In the next step, "Select at least 2 expression samples", right-click on the "Results" folder and choose **Add folder contents (recursively)** to add all 12 expression tracks.
3. In the next two steps, "Distances" and "Feature filters", keep the default settings.
4. In the next step, "Result handling", select **Save**.
5. In the last step, choose to save the results in the "Results" folder and click on **Finish**.

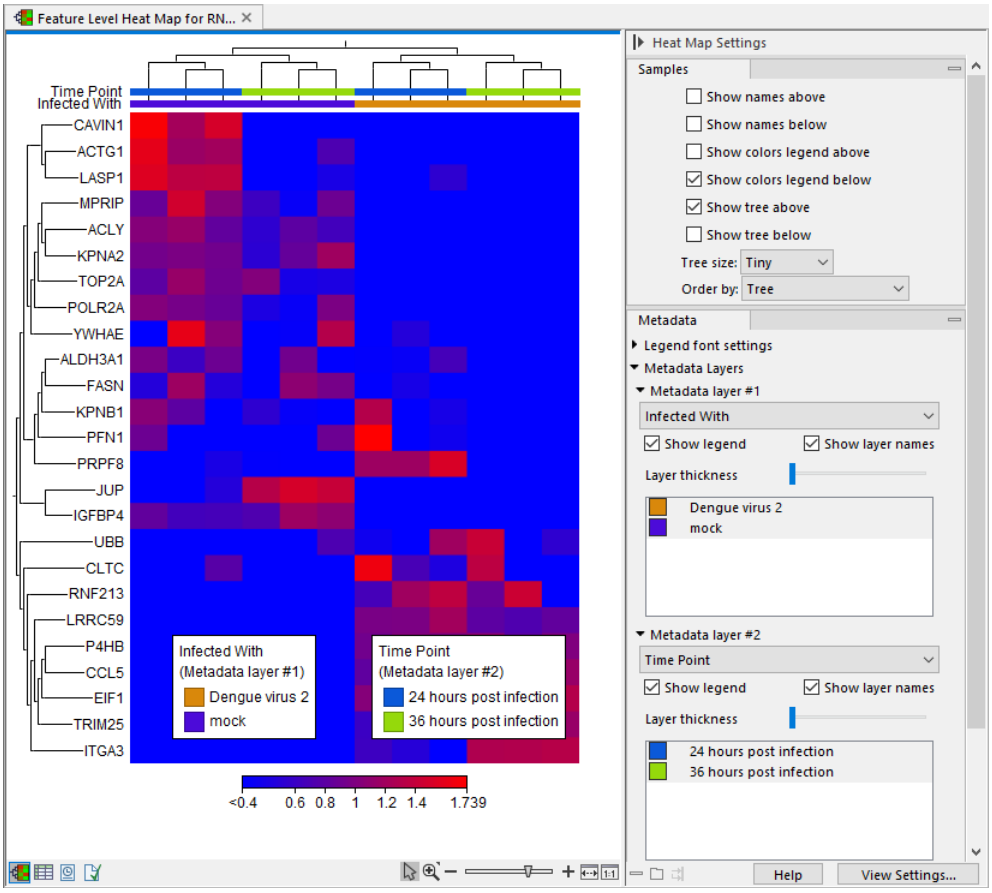
To visualize if the samples cluster across time points, configure the following options in the Side Panel of the resulting heat map:

1. In the "Samples" palette, uncheck **Show names above**.
2. In the "Metadata" palette, set "Metadata layer #1" to **Infected With** and "Metadata layer #2" to **Time Point**.

The samples cluster by infection type, similarly to the PCA plot. Here, both the mock-infected and Dengue virus 2-infected cells additionally cluster by time point, in contrast to what we observed in the PCA plot. Again this discrepancy may be explained by the fact that the heat map clustering is based on only 25 genes, whereas the PCA plot is based on all genes (figure 19).

Upload to IPA

For biological interpretation of the differentially expressed genes, statistical comparisons can be uploaded to `IPA` using the `Upload to IPA` ( ) tool. Recall from **Prerequisites** that this requires access to IPA and the *Biomedical Genomics Analysis* plugin to be installed.



Heat Map Settings

Samples

☐ Show names above

☐ Show names below

☐ Show colors legend above

☑ Show colors legend below

☑ Show tree above

☐ Show tree below

Tree size: Tiny

Order by: Tree

Metadata

Legend font settings

Metadata Layers

Metadata layer #1

Infected With

☑ Show legend ☑ Show layer names

Layer thickness

🟧 Dengue virus 2
🟪 mock

Metadata layer #2

Time Point

☑ Show legend ☑ Show layer names

Layer thickness

🟦 24 hours post infection
🟩 36 hours post infection

Figure 19: *Heat map with the sample names hidden and "Infected With" and "Time Point" metadata layers added.*

The statistical comparison tracks generated above contain all genes. Before uploading the results to IPA, we use the Filter on Custom Criteria ( ) tool to extract only the differentially expressed genes:

1. Launch **Filter on Custom Criteria** ( ) using Quick Launch ( ).

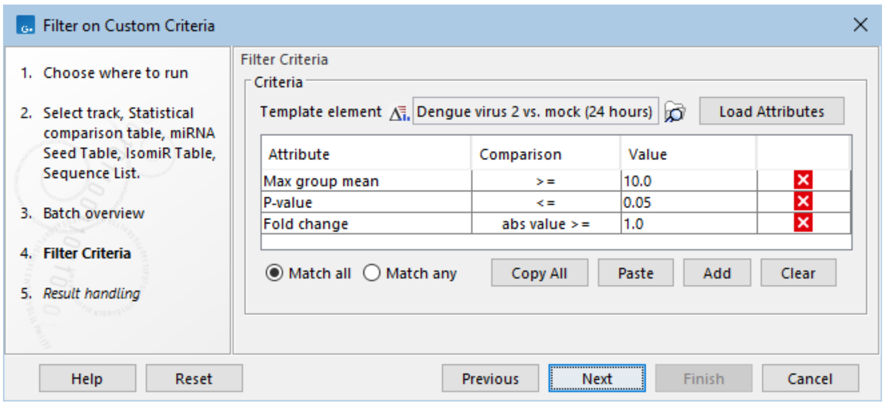
2. In the next step, "Select track, Statistical comparison table, miRNA Seed Table, IsomiR Table, Sequence List.", select the "Dengue virus 2 vs. mock (24 hours)" and "Dengue virus 2 vs. mock (36 hours)" statistical comparison tracks from the "Results" folder.

   Check the **Batch** option.

3. In the next step, "Batch overview", review the organization of the input data.

4. In the next step, "Filter Criteria", click on **Load Attributes**. Add the following criteria by clicking on **Add** and configuring the "Attribute", "Comparison", and "Value" of each criterion (figure 20):

   - Max group mean >= 10.0.
   - P-value <= 0.05.
   - Fold change abs value >= 1.0.



Filter on Custom Criteria

1. Choose where to run
2. Select track, Statistical comparison table, miRNA Seed Table, IsomiR Table, Sequence List.
3. Batch overview
4. **Filter Criteria**
5. *Result handling*

Filter Criteria

Criteria

Template element: Dengue virus 2 vs. mock (24 hours)

Attribute	Comparison	Value	
Max group mean	>=	10.0	☒
P-value	<=	0.05	☒
Fold change	abs value >=	1.0	☒

● Match all ○ Match any

Figure 20: *Filter on Custom Criteria configured to extract differentially expressed genes.*

5. In the next step, "Result handling", choose **Save in input folder**.

   Click on **Finish**.

When the tool has finished running, filtered versions of the statistical comparison tracks will be available in the "Results" folder.

To upload the results to IPA:

1. Launch **Upload to IPA** () using Quick Launch ().
2. In the next step, "Select statistical comparisons", select the "Dengue virus 2 vs. mock (24 hours, filtered)" and "Dengue virus 2 vs. mock (36 hours, filtered)" statistical comparison tracks from the "Results" folder.
3. In the next step, "Upload to IPA", click on "Log in" and log in to your IPA account.
4. In the next step, "Configuration", keep the default settings and click on **Finish**.

You will receive an email when the IPA analysis is complete.

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

Automation and further customization of the analysis

We can automate the analysis presented here by adding a few workflow elements to a `copy of the template workflow` used previously. See `the manual` for general information about editing workflows. The tutorial data contains such an edited workflow copy, "Advanced RNA-Seq and Differential Gene Expression Analysis". The key change is the addition of a new layer of Iterate and Collect and Distribute control flow elements, as well as the Upload to IPA element preceded by a `Fork` element.

Open the imported workflow by double clicking on its name in the "Advanced workflow" folder (figure 6).

The names of the new control flow elements give an indication of their role (figure 21):

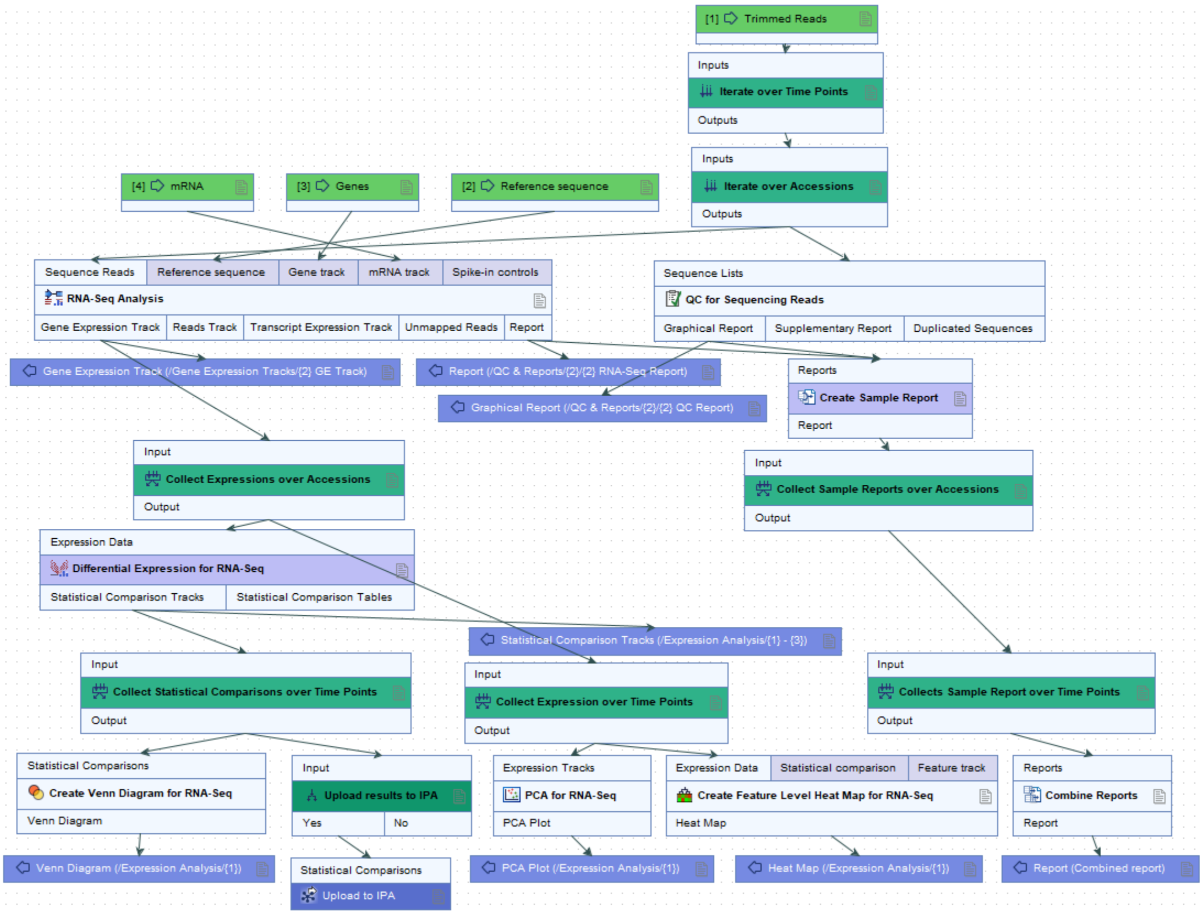


Figure 21: *Overview of the "Advanced RNA-Seq and Differential Gene Expression Analysis" workflow.*

- Names ending in "over Accessions" refer to running analysis steps on, and collecting results for, *each sample*, based on their "Run Accession" value.
- Names ending in "over Time Points" refer to running analysis steps on, and collecting results for, *each group of samples*, where each group represents a particular time point.

This workflow design allows joint analysis of results from multiple groups of samples in downstream steps of the same workflow. For example:

- We create a Venn diagram using the results from both analyzed 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.

Feel free to launch the "Advanced RNA-Seq and Differential Gene Expression Analysis" workflow if you wish to. Launching the workflow is very much like launching the "RNA-Seq and Differential Gene Expression Analysis" template workflow, as we did in step 1, except that:

- A new wizard step, "Specify workflow path" is present where you can choose whether to upload results to IPA. Only choose "Yes" if you have access to IPA.

- In the "Select Trimmed Reads" step, do *not* check the **Batch** option.

  Checking the Batch option was needed to run the "RNA-Seq and Differential Gene Expression Analysis" template once per time point, but in the "Advanced RNA-Seq and Differential Gene Expression Analysis" workflow, the additional layer of Iterate and Collect and Distribute elements leads to the appropriate sections of the workflow being run once per time point.

- In the "Configure batching" step:

  - Under "Iterate over Time Points" set "Define batch units using metadata column" to **Time Point**.
  - Under "Iterate over Accessions" set "Define batch units using metadata column" to **Run Accession**.

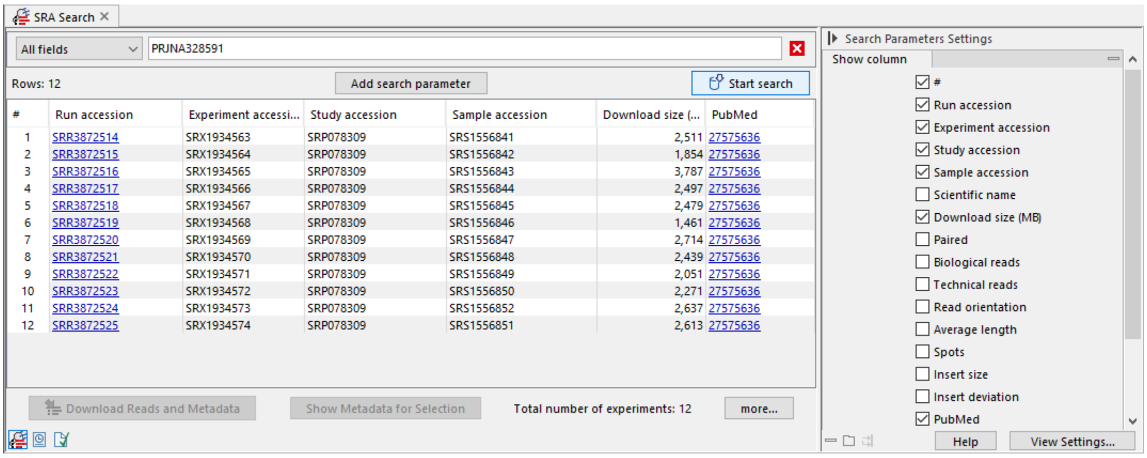
- If you selected "Yes" in the "Specify workflow path" step, a new "Upload to IPA" step is available. Click on "Log in" and log in to your IPA account.

The resulting Venn diagram, PCA plot, heat map, and statistical comparisons include all 12 samples: Dengue virus 2- and mock-infected samples, from 24 and 36 hour post infection time points.

If you chose to upload to IPA, the statistical comparisons are automatically uploaded as part of the workflow run. You will receive an email when the IPA analysis is complete.

Analyze the full data set

If you wish to analyze the full data set, the data can be downloaded from SRA directly in *CLC Genomics Workbench* (figure 22):



PRJNA328591

Rows: 12

#	Run accession	Experiment accessi...	Study accession	Sample accession	Download size (...	PubMed
1	SRR3872514	SRX1934563	SRP078309	SRS1556841	2,511	27575636
2	SRR3872515	SRX1934564	SRP078309	SRS1556842	1,854	27575636
3	SRR3872516	SRX1934565	SRP078309	SRS1556843	3,787	27575636
4	SRR3872517	SRX1934566	SRP078309	SRS1556844	2,497	27575636
5	SRR3872518	SRX1934567	SRP078309	SRS1556845	2,479	27575636
6	SRR3872519	SRX1934568	SRP078309	SRS1556846	1,461	27575636
7	SRR3872520	SRX1934569	SRP078309	SRS1556847	2,714	27575636
8	SRR3872521	SRX1934570	SRP078309	SRS1556848	2,439	27575636
9	SRR3872522	SRX1934571	SRP078309	SRS1556849	2,051	27575636
10	SRR3872523	SRX1934572	SRP078309	SRS1556850	2,271	27575636
11	SRR3872524	SRX1934573	SRP078309	SRS1556852	2,637	27575636
12	SRR3872525	SRX1934574	SRP078309	SRS1556851	2,613	27575636

Total number of experiments: 12

☑ #
☑ Run accession
☑ Experiment accession
☑ Study accession
☑ Sample accession
□ Scientific name
☑ Download size (MB)
□ Paired
□ Biological reads
□ Technical reads
□ Read orientation
□ Average length
□ Spots
□ Insert size
□ Insert deviation
☑ PubMed

Figure 22: *SRA search results for project PRJNA328591.*

1. Launch Search for Reads in SRA (🔍) using Quick Launch (🚀).
2. Enter "PRJNA328591" in the search field.
3. Click on **Start search** (🔍) or press Enter.
4. Select all 12 experiments (shortcut Ctrl+A or ⌘ +A on Mac).

5. Click on **Download Reads and Metadata** ( ) near the bottom, on the left-hand side. A wizard will open.

    (a) In the "Import Options" and "Edit Paired End Settings" steps, keep the default settings.

    (b) In the next step, "Result handling", select **Save**.

    (c) In the last step, choose to save the results in a suitable location in the Navigation Area.
    Click on **Finish**.

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

Note that the tutorial data was down-sampled to approximately 3%. The full data takes considerably more time to analyze.