

## Application Note

# RNA-seq analysis using long and short reads from pathogen-infected plant tissues

Dual RNA-seq analysis of plant–pathogen interactions using time-course experiments can uncover distinct transcriptional signatures during incompatible and compatible interactions and can help identify pathogen-effector candidates. Since short- and long-read technologies each have their inherent biases, a hybrid sequencing approach can be used to obtain a more complete picture. Instead of analyzing the plant RNA-seq data and the pathogen RNA-seq data separately, a combined analysis is more efficient and convenient and provides greater insight.

This application note describes the RNA-seq analysis of a plant–pathogen interaction using the expression analysis tools in QIAGEN® CLC Genomics Workbench Premium. This analysis can be used to find genes that are differentially expressed in a plant–microbe interaction. We cover the following steps for the efficient analysis of RNA-seq data:

1. Preparing a combined genomic reference for the simultaneous mapping of reads from both host and pathogen
2. Mapping the long and short reads to a genomic reference
3. Identifying differentially expressed genes using the statistical tools of the Workbench
4. Visualizing and extracting the expression data results using heat maps and Venn diagrams

5. Navigating the expression results in RNA-seq mappings along with the annotation information

The workflow covered in this application note is presented in Figure 1. The host and pathogen reference genomes are downloaded and combined into one synthetic reference using QIAGEN CLC Genomics Workbench tools. The short and long RNA-seq reads, along with the metadata information, are imported from the Sequence Read Archive (SRA) directly into the Workbench.

After running the RNA-seq analysis, the principal component analysis (PCA) tool is used to quality control the dataset and confirm that there are no outlying samples. The statistical tools are used to find differentially expressed genes, and the visualization tools (heat maps and Venn diagrams) are used to visualize and extract data. The tracks of differentially expressed genes are used to inspect the RNA-seq mappings.

## Data

The experimental data used in this analysis are from soybean leaves infected with Asian soybean rust, *Phakopsora pachyrhizi*, an obligate biotrophic fungal pathogen. The datasets were published by Elmore et al. 2020. A total of 32 RNA-seq samples were obtained from two different sequencing technologies, PacBio® and Illumina®. The leaf tissues were collected at four different

time points after infection: 3, 7, 10 and 14 days. The RNA-seq files are available at the SRA under project number [SRP221996](#).

The soybean reference genome is available at GenBank under the identifier *Glycine\_max\_v4.0*. The pathogen reference genome is available at the JGI MycoCosm database under the identifier *Phakopsora pachyrhizi* K8108 v2.0.

## Creating the synthetic reference genome

To analyze the RNA expression in both organisms simultaneously, we combined the host and pathogen genomes into one synthetic reference. This way, we did not need to run the RNA-seq expression analysis separately on two different references.

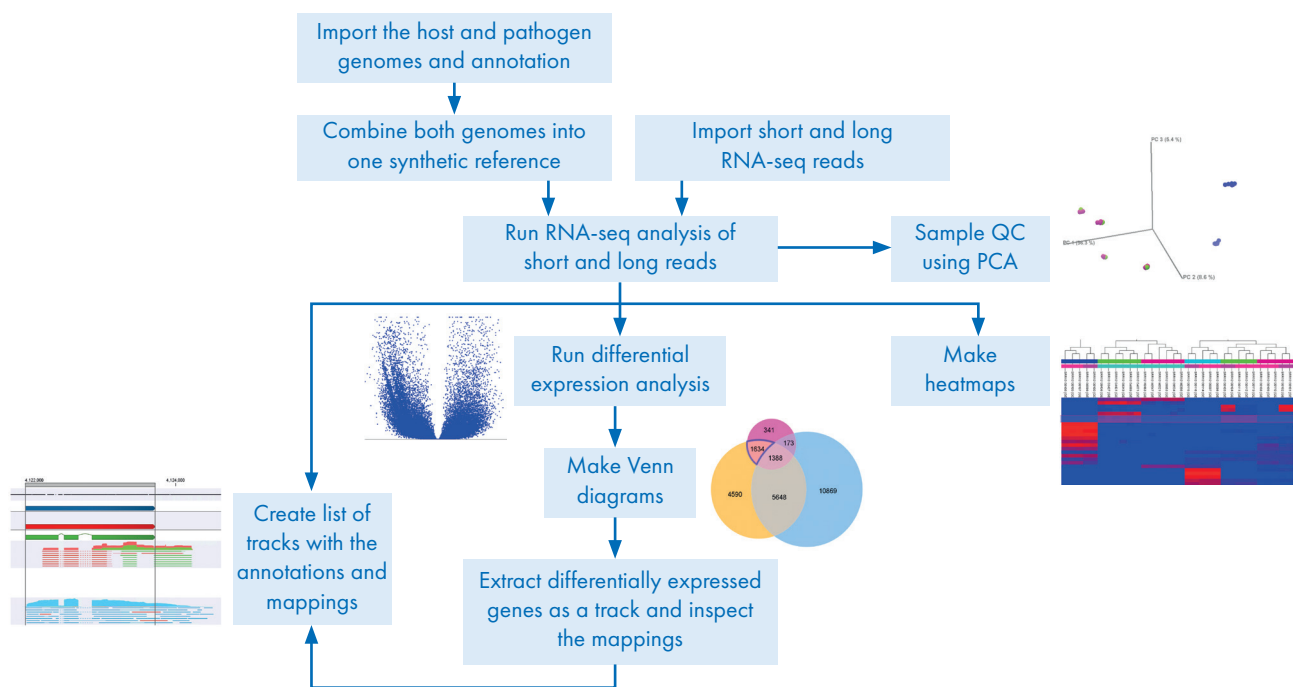


Figure 1. RNA-seq data analysis workflow.

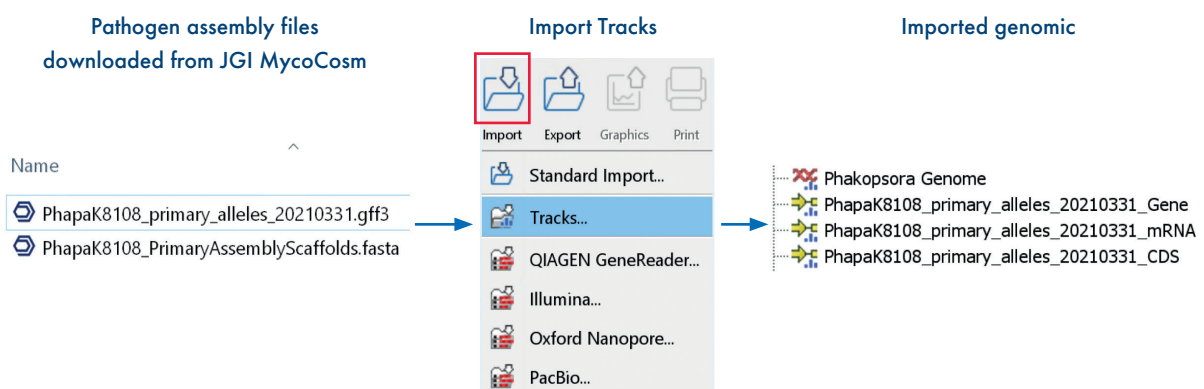


Figure 2. Importing the pathogen genome from JGI MycoCosm to QIAGEN CLC Genomics Workbench Premium.

The *Phakopsora* genomic reference and annotation data were downloaded from the JGI MycoCosm database and imported into the QIAGEN CLC Genomics Workbench Premium using the “Import Tracks” tool. We first imported the *Phakopsora* genome from the FASTA file, and then we imported annotation data from the GFF3 file. The data were subsequently available in the Workbench as four tracks: the Genome track, and three types of annotation tracks: Gene, mRNA and CDS (Figure 2).

The host genome, available at GenBank, was downloaded directly using the “Search for Sequences at NCBI...” action in the “Download” button’s menu. We selected all 20 chromosomes of soybean and imported them as a sequence list using “Download and Save”. This produced the “Glycine max Genome” sequence list file shown at the bottom-right of Figure 3.

When combining both genomes into one synthetic reference, the Workbench allowed us to concatenate sequence lists into one file. This function is available under *File* → *New* → *Sequence List...* Notably, the two genomic references we wished to import were in different formats. The soybean genome was a standalone file, whereas the pathogen genome consisted of multiple tracks (the last four shown at the bottom of Figure 3). The standalone file for soybean combined multiple layers of genomic information: the genomic sequence and all annotation types. In contrast, with multiple tracks, each track has just only one layer of information: a genome or some type of annotation. Standalone files are treated by QIAGEN CLC Genomics Workbench tools as a sequence list, whereas multiple tracks are not. To combine our two genomes, we first converted the multiple *Phakopsora* tracks into a standalone file. We used the “Convert from Tracks” tool in the “Track Conversion” folder to create this standalone file (Figure 4).

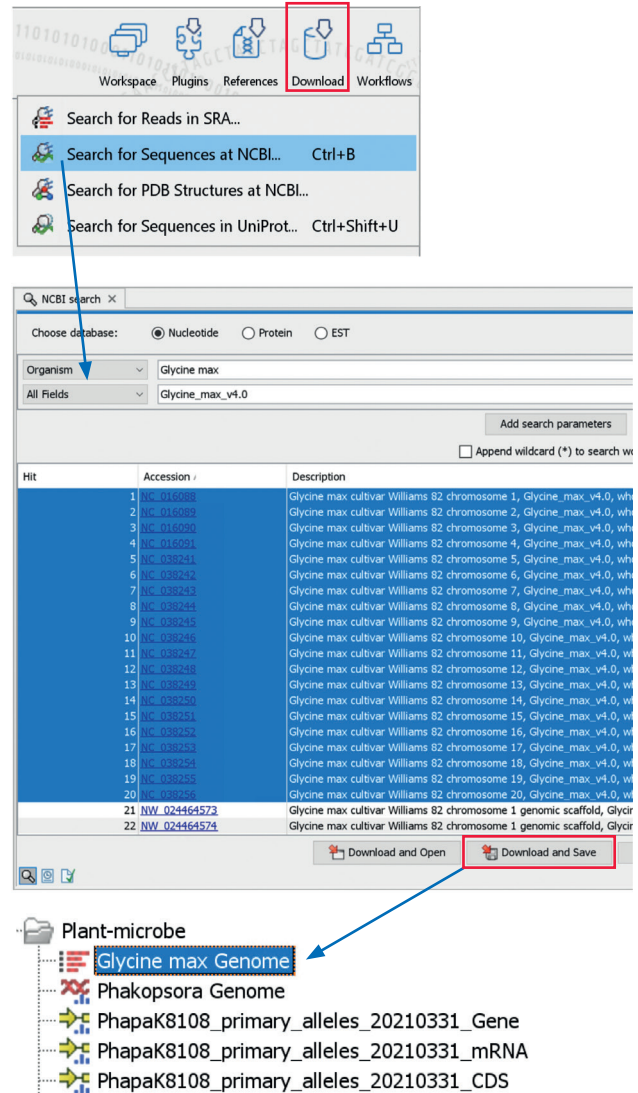


Figure 3. Downloading the host genome from GenBank.

When both the host and pathogen genomes were available as sequence lists (standalone files), they were combined into one synthetic genomic reference, as shown in Figure 5.

After generating the standalone file containing both genomes, the next consideration was that the RNA-seq analysis tool requires reference information in a form with multiple tracks. Thus, we converted the synthetic reference to multiple tracks using the “Convert to

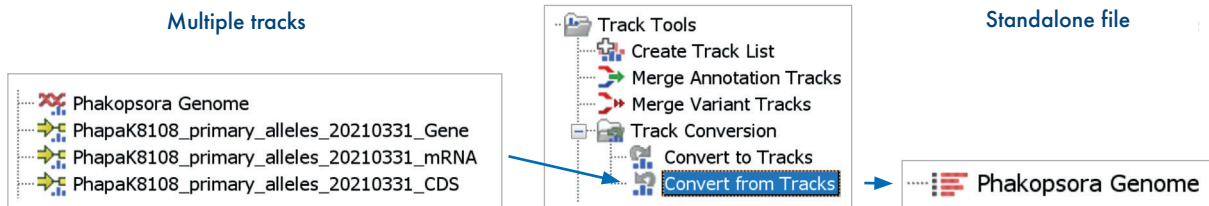


Figure 4. Converting genomic reference in multiple tracks to a standalone file (sequence list).

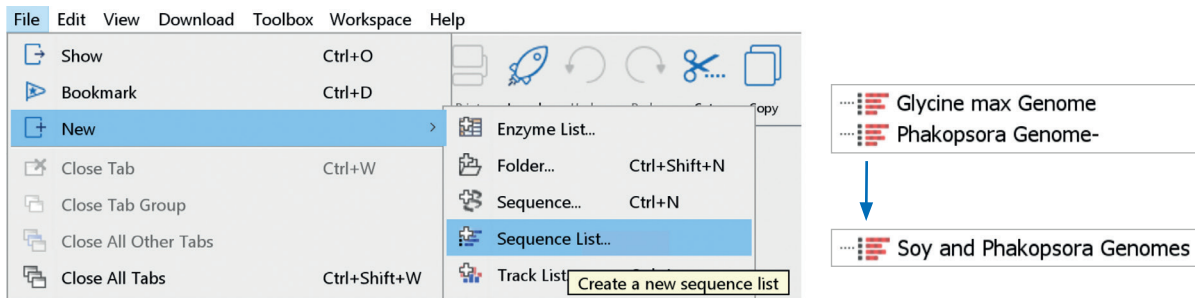


Figure 5. Combining reference files using New -> Sequence List...

Tracks” tool in the “Track Conversion” folder (Figure 6). The combined reference was then ready to be used for RNA-seq analysis.

### Importing RNA-seq reads and metadata

The RNA-seq reads were imported directly from GenBank using the “Search for Reads in SRA...” tool in the “Download” button’s menu. A search for the project identifier (SRP221996) returned all 32 RNA-seq data files. They were selected and then downloaded to the Workbench using “Download Reads and Metadata” (Figure 7).

The “Instrument” and “Development Stage” fields of the metadata table for the downloaded samples are shown Figure 8, and these values were used in the RNA-seq data comparison. As different instruments and technologies provide different sensitivity in the detection of RNA expression, we used data from both instruments to compare and identify genes that were differentially expressed.

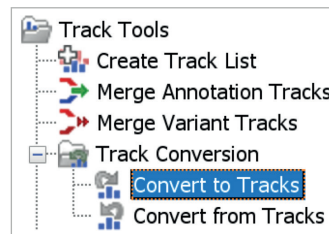


Figure 6. Converting the concatenated synthetic reference to multiple tracks.

### RNA-seq analysis of short and long reads

In this step, the RNA-seq reads were mapped and counted across genes and transcripts. To run an RNA-seq analysis on Illumina files, we used the “RNA-Seq Analysis” tool. For the PacBio files, we used the “RNA-Seq Analysis for Long Reads” tool (Figure 9). Both tools were run using the “Batch” option. This option creates a mapping track for each individual sample. In “Reference” settings, we selected the synthetic reference tracks (Figure 10). In the “Result handling” settings, we checked the “Create reads track” and “Create report” options.

For each of the 32 samples, the RNA-seq analysis (from either tool) produces four output files: counts per gene (GE), counts per transcript (TE), the RNA-seq report and an RNA-seq-reads mapping track (Figure 11). The counts-per-gene file shows the expression detected for each gene. The counts per transcript shows the expression detected for each transcript. The RNA-seq report contains summary mapping and expression statistics. The RNA-seq reads mapping track shows the location of each read maps in the reference.

### Sample quality control

Before conducting a differential expression analysis, we wanted to identify any outlying samples with unexpected expression behavior. The presence of such outliers would indicate experiment quality problems. To do this, we produce a principal component analysis (PCA) plot using all 32 gene expression files from the RNA-seq analysis output. PCA projects our high-dimensional dataset (one dimension per gene or transcript) onto just two or three dimensions. Any outlying samples identified in this analysis should not be included in differential expression analysis. We used the “PCA for RNA-seq” tool, which can be found in the “RNA-Seq and Small RNA Analysis” folder (Figure 12).

The results can be visualized as a two- or three-dimensional plot. The Illumina data contained four timepoints after infection (3, 7, 10 and 14 days) with five replicates for each timepoint. In the PacBio data, there were two timepoints (7 and 10 days after infection) with six replicates for each timepoint. The three-dimensional plot sometimes provides better spatial resolution. In our case, the two-dimensional plot, PacBio data from 7 and 10 days after infection were very near each other (red and green diamonds), whereas in the three-dimensional plot they are well-separated (two clusters

The screenshot shows the SRA Search interface. At the top, there's a search bar with 'SRP221996' entered. Below it, a table lists 32 rows of search results. The columns are: #, Run Accession, Experiment Accession, Study Accession, Sample Accession, Paired, and Average Length. A red box highlights the 'Download Reads and Metadata' button. Below the table, there's a menu with options: 'Search for Reads in SRA...', 'Search for Sequences at NCBI... Ctrl+B', 'Search for PDB Structures at NCBI...', and 'Search for Sequences in UniProt... Ctrl+Shift+U'.

Figure 7. Importing reads and metadata from the SRA.

The screenshot shows the SRA MetadataTable interface. It displays a table with 32 rows and 5 columns: Run Accession, Project Accession, Instrument, Development Stage, and Tissue. The data includes various run IDs, project IDs, instruments (PacBio RS II, Illumina HiSeq 3000, Illumina MiSeq), development stages (7 days after inoculation, 3 days after inoculation, 14 days after inoculation), and tissues (Infected leaf).

Run Accession	Project Accession	Instrument	Development Stage	Tissue
SRR10133427	SRP221996	PacBio RS II	7 days after inoculation	Infected leaf
SRR10134431	SRP221996	PacBio RS II	7 days after inoculation	Infected leaf
SRR10134669	SRP221996	PacBio RS II	7 days after inoculation	Infected leaf
SRR10134671	SRP221996	PacBio RS II	7 days after inoculation	Infected leaf
SRR10138343	SRP221996	PacBio RS II	7 days after inoculation	Infected leaf
SRR10139436	SRP221996	PacBio RS II	7 days after inoculation	Infected leaf
SRR10130103	SRP221996	Illumina HiSeq 3000	7 days after inoculation	Infected leaf
SRR10130104	SRP221996	Illumina HiSeq 3000	7 days after inoculation	Infected leaf
SRR10130111	SRP221996	Illumina HiSeq 3000	7 days after inoculation	Infected leaf
SRR10130112	SRP221996	Illumina HiSeq 3000	7 days after inoculation	Infected leaf
SRR10130113	SRP221996	Illumina HiSeq 3000	7 days after inoculation	Infected leaf
SRR10130097	SRP221996	Illumina MiSeq	3 days after inoculation	Infected leaf
SRR10130098	SRP221996	Illumina MiSeq	3 days after inoculation	Infected leaf
SRR10130114	SRP221996	Illumina MiSeq	3 days after inoculation	Infected leaf
SRR10130115	SRP221996	Illumina HiSeq 3000	3 days after inoculation	Infected leaf
SRR10130116	SRP221996	Illumina HiSeq 3000	3 days after inoculation	Infected leaf
SRR10130105	SRP221996	Illumina MiSeq	14 days after inoculation	Infected leaf
SRR10130106	SRP221996	Illumina MiSeq	14 days after inoculation	Infected leaf
SRR10130107	SRP221996	Illumina MiSeq	14 days after inoculation	Infected leaf
SRR10130099	SRP221996	Illumina HiSeq 3000	14 days after inoculation	Infected leaf
SRR10130100	SRP221996	Illumina HiSeq 3000	14 days after inoculation	Infected leaf
SRR10130101	SRP221996	Illumina HiSeq 3000	10 days after inoculation	Infected leaf
SRR10130102	SRP221996	Illumina HiSeq 3000	10 days after inoculation	Infected leaf
SRR10130108	SRP221996	Illumina MiSeq	10 days after inoculation	Infected leaf
SRR10130109	SRP221996	Illumina MiSeq	10 days after inoculation	Infected leaf
SRR10130110	SRP221996	Illumina MiSeq	10 days after inoculation	Infected leaf
SRR10139534	SRP221996	PacBio RS II	10 days after inoculation	Infected leaf
SRR10139853	SRP221996	PacBio RS II	10 days after inoculation	Infected leaf
SRR10140398	SRP221996	PacBio RS II	10 days after inoculation	Infected leaf
SRR10140483	SRP221996	PacBio RS II	10 days after inoculation	Infected leaf
SRR10140521	SRP221996	PacBio RS II	10 days after inoculation	Infected leaf
SRR10140547	SRP221996	PacBio RS II	10 days after inoculation	Infected leaf

Figure 8. The metadata table for the downloaded samples (20 Illumina, 12 PacBio).

For the 20 Illumina files

For the 12 PacBio files

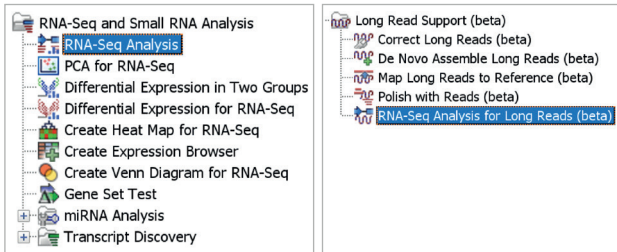


Figure 9. Tools for RNA-seq analysis of short and long reads.

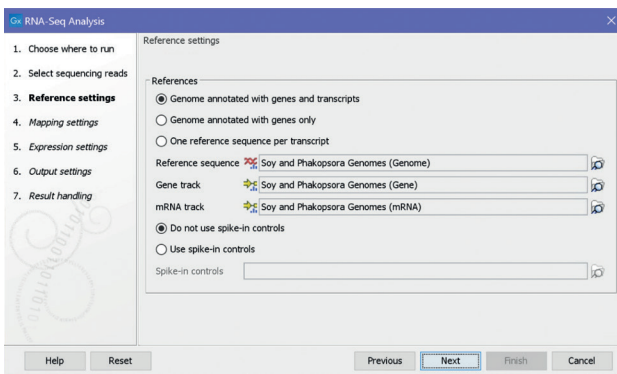


Figure 10. Selecting reference tracks for RNA-seq analysis.

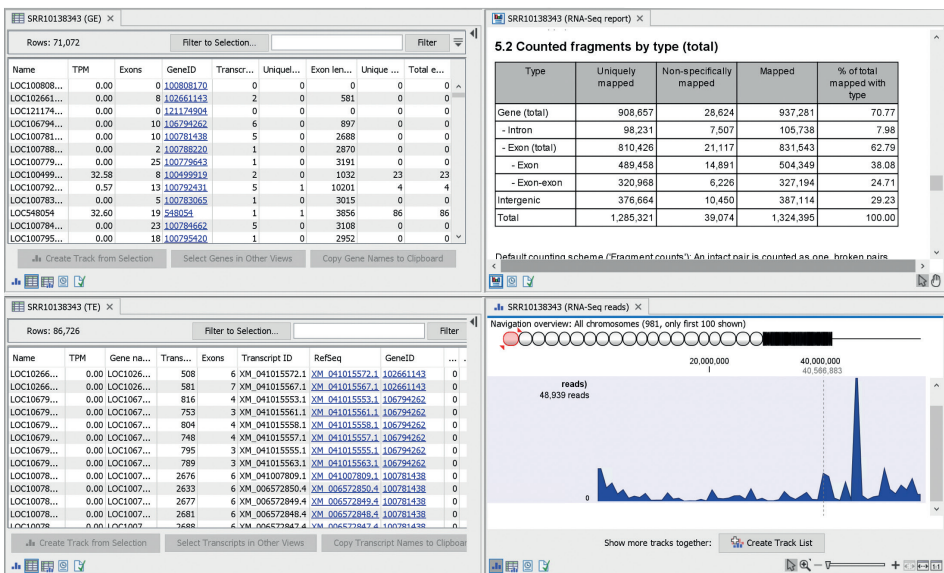
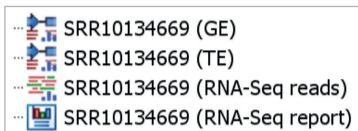


Figure 11. Output files from the RNA-seq analysis.

of cyan points). We also see that the samples are clustered by sequencing technology, not just by time after infection. However, inside each technology-timepoint pair, all samples are clustered together and we find no outliers (Figure 13).

Differential expression analysis

After quality controlling the samples, we proceeded to the differential expression analysis using the "Differential Expression for RNA-Seq" tool (Figure 12). For both technologies, we ran an independent analysis comparing 7 and 10 days after infection. The tool produces an output table, which can be visualized as a volcano plot to show the change in expression. The volcano plots look very different for the two technologies (Figure 14), indicating caution in drawing conclusions from just one technology. In this case, the depth of sequencing is significantly lower for the PacBio samples.

In a separate experiment, we looked for differentially expressed transcripts across all instruments and time-points. This analysis also used the "Differential Expression for RNA-Seq" tool. All 32 counts-per-transcript (TE) files from the RNA-seq analysis were tested for "development stage" while controlling for "instrument". In the output table, we selected just the pathogen transcripts by filtering for identifiers starting with "jgi". We set "fold change" to an absolute value of ">100" to return only transcripts

with over a 100-fold change in expression in either direction (up or down). This filtering returned 1373 such transcripts meeting these restrictions (Figure 15). Selected transcripts are highlighted in red in the volcano plot. We also created a differential expression track using the “Create Track from Selection” button.

## Heatmaps

Heatmaps can be used to visualize the RNA-seq counts files, clustering the data by sample and feature. We use the “Create Heat Map for RNA-Seq” tool (Figure 12). We chose the option for this tool that uses the statistical output from the “Differential Expression for RNA-Seq” tool to visualize the difference in expression across all samples.

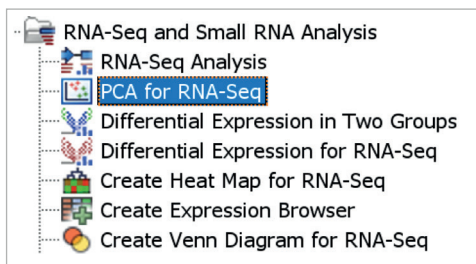


Figure 12. Principal component analysis tool.

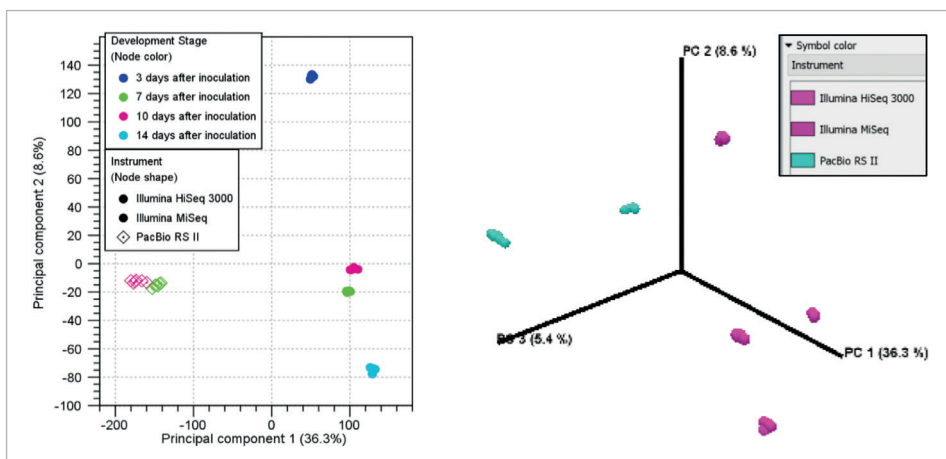


Figure 13. Principal component analysis output in two- (left) and three-dimensional (right) plots.

Heatmaps are interactive and allow us to visualize multiple layers of metadata. In Figure 16, the first layer is “development stage” and the second layer is “instrument”. There are four main branches for Illumina and two for PacBio. As expected, the most dramatic expression differences are found between 3 and 14 days after infection. Here, Illumina RNA-seq samples provide better resolution than PacBio.

## Venn diagrams

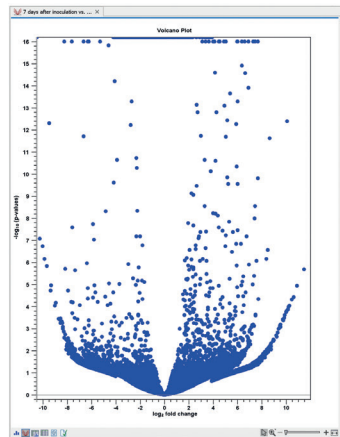
To find the differentially expressed genes that are detected in multiple experiments, we created a Venn diagram. The “Create Venn Diagram” tool computes the overlap of differentially expressed genes or transcripts in two or more statistical comparison tracks created by the “Differential Expression for RNA-Seq” tool. We created a Venn diagram with three different gene expression (GE) experiments:

- Differential expression in all 32 samples (13,260 genes)
- Differential expression in 7 versus 10 days in Illumina samples (3536 genes)
- Differential expression in 7 versus 10 days in PacBio samples (1883 genes)

We set “Minimum absolute fold change” to “3”, as shown in “Venn Diagram Settings”. The intersection of all three statistical comparisons is 803 genes (Figure 17). We selected four additional genes at the intersection of the two instrument-specific experi-

7 days after inoculation vs. 10 days after inoculation PacBio

Name	Chromosome	Max group ...	Log2 fold ...	Fold change	P-value	FDR p-value	Bonferroni	GeneID
LOC100305746	NC_038247	964.07	3.30	9.87	4.72E-14	2.35E-11	1.44E-9	100305746
LOC100790067	NC_038241	81.68	4.21	18.57	5.31E-14	2.99E-11	1.62E-9	100790067
LOC112998577	NC_038248	274.33	5.92	60.75	9.52E-14	4.56E-11	2.90E-9	112998577
SRC1	NC_038253	5,859.82	-2.27	-4.82	1.11E-13	5.24E-11	3.39E-9	547450
LOC100818915	NC_038247	264.52	3.82	14.11	1.63E-13	7.53E-11	4.97E-9	100818915
LOC100799873	NC_038254	126.27	5.18	36.21	3.12E-13	1.41E-10	9.50E-9	100799873
LOC10016231	NC_038254	64.93	7.69	206.52	3.48E-13	1.55E-10	1.06E-8	10016231
ERD15	NC_016089	782.60	-4.15	-17.78	5.55E-13	2.42E-10	1.69E-8	100796671
LOC100800844	NC_038249	58.90	6.02	64.82	6.54E-13	2.80E-10	1.99E-8	100800844
LCL3	NC_038255	58.53	5.24	37.67	6.64E-13	2.80E-10	2.02E-8	780540
LOC100807732	NC_038241	756.64	2.63	6.21	8.47E-13	3.51E-10	2.58E-8	100807732



7 days after inoculation vs. 10 days after inoculation Illumina

Name	Chromosome	Max group mean	Log2 fold ...	Fold change	P-value	FDR p-value	Bonferroni	GeneID
LOC100306510	NC_016088	623.35	3.42	10.72	0.00	0.00	0.00	100306510
ALDH11A1	NC_016089	184.67	3.19	9.11	0.00	0.00	0.00	100780041
LOC100800738	NC_016089	243.07	2.80	6.97	0.00	0.00	0.00	100800738
LOC100793702	NC_016089	296.72	4.89	29.62	0.00	0.00	0.00	100793702
CHLH	NC_016090	103.53	3.13	8.75	0.00	0.00	0.00	548033
MLOC100793702	NC_016090	118.97	5.92	60.56	0.00	0.00	0.00	728086
LOC100500448	NC_038241	220.80	3.90	14.96	0.00	0.00	0.00	100500448
LOC100788841	NC_038241	123.96	5.80	35.56	0.00	0.00	0.00	100788841
LOC100804547	NC_038241	613.75	2.97	7.84	0.00	0.00	0.00	100804547
LOC100782459	NC_038242	66.52	3.81	14.00	0.00	0.00	0.00	100782459
LCL4	NC_038243	74.43	3.79	13.81	0.00	0.00	0.00	100101859
COL21	NC_038243	150.47	3.59	12.02	0.00	0.00	0.00	100802113
LOC547630	NC_038245	156.72	3.63	12.36	0.00	0.00	0.00	547630

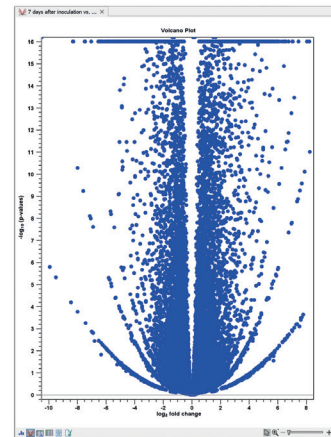


Figure 14. Tables and volcano plots from the differential expression analysis.

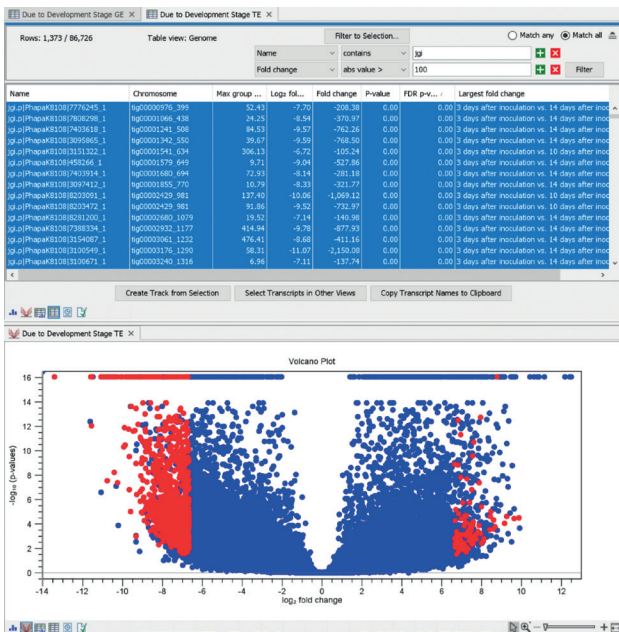


Figure 15. The table and volcano plot from the differential expression analysis in which the highly up- and down-regulated pathogen transcripts are indicated in red in the plot.

ments. Selecting genes in the Venn diagram also selects them in the table view.

### Inspection of RNA-seq mapping

Inspecting RNA-seq mappings allows us to identify problems in the reference annotation. This is important because analysis will be impaired if annotation is missing for a gene. To navigate RNA-seq mapping for just the genes selected in the Venn diagram, we had to first create a new track to contain genes with missing annotation. We selected the genes in the statistical analysis table and clicked "Create Track from Selection" (Figure 18).

To inspect the RNA-seq mappings for the differentially expressed genes, we created a list of tracks containing the reference genome, the RNA-seq mappings, the 807 differentially expressed genes, and the annotation tracks (Figure 19). This was accomplished using the



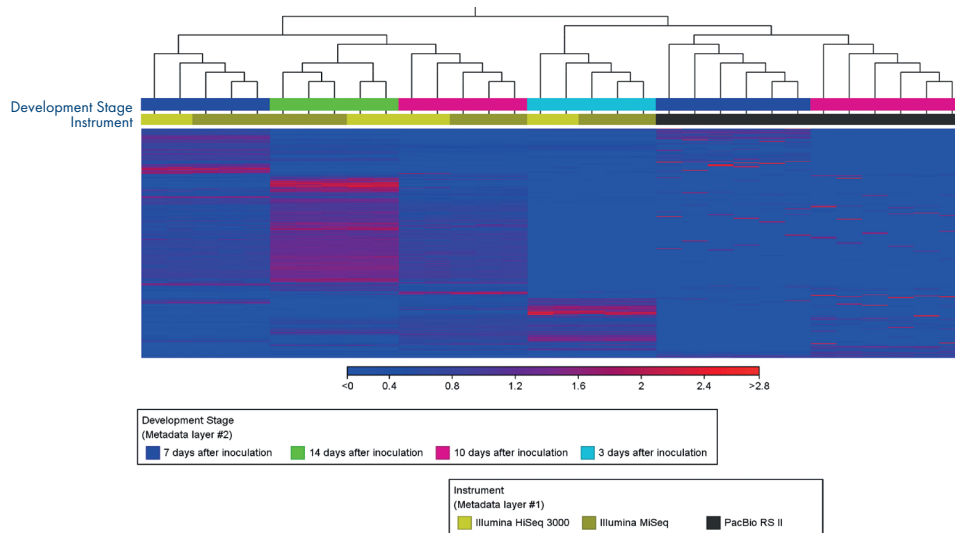


Figure 16. Heatmap of expression features across all 32 samples (one per column). Each row is an individual gene. Labels are available when interacting with the heatmap.

“Create Track List” tool under the “Track Tools” folder (Figure 4).

In Figure 19, the first track (a thin line) is the synthetic reference genome. The second track is gene annotation. The third track contains the 807 differentially expressed genes. The fourth track is the transcript annotation. The last four tracks are PacBio and Illumina mappings. Double-clicking the 807 genes track opens it in a table view. We filtered the table for “RCA11” to find one of the most dramatically down-regulated genes. The RCA11 gene is involved in plant photosynthesis, and it is dramatically downregulated between 7 and 10 days after infection when the host’s photosynthesis is shut down by the pathogen. The dramatic reduction is highlighted in red: these counts show the number of reads mapped to the selected gene in each mapping track (Figure 19).

The importance of inspecting read mappings becomes clear when we work with imperfectly annotated genomes (Figure 20). This figure shows that the read mappings extend beyond the annotated area of a soybean gene.

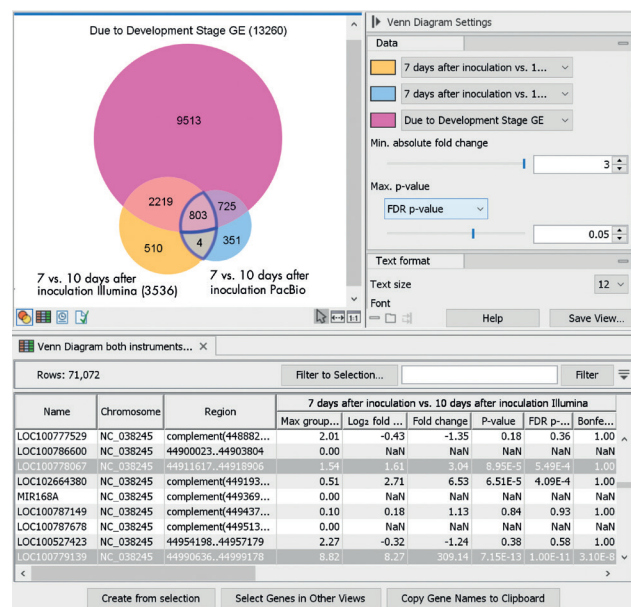


Figure 17. Venn diagram with 807 differentially expressed genes selected.

A few examples of truncated gene annotation were also found in the *Phakopsora* genome, shown in Figure 21. Annotation deficiencies become an even greater problem when an annotation is completely missing for a gene: RNA-seq reads mapped in an unannotated area are not counted in the expression analysis (Figure 22).

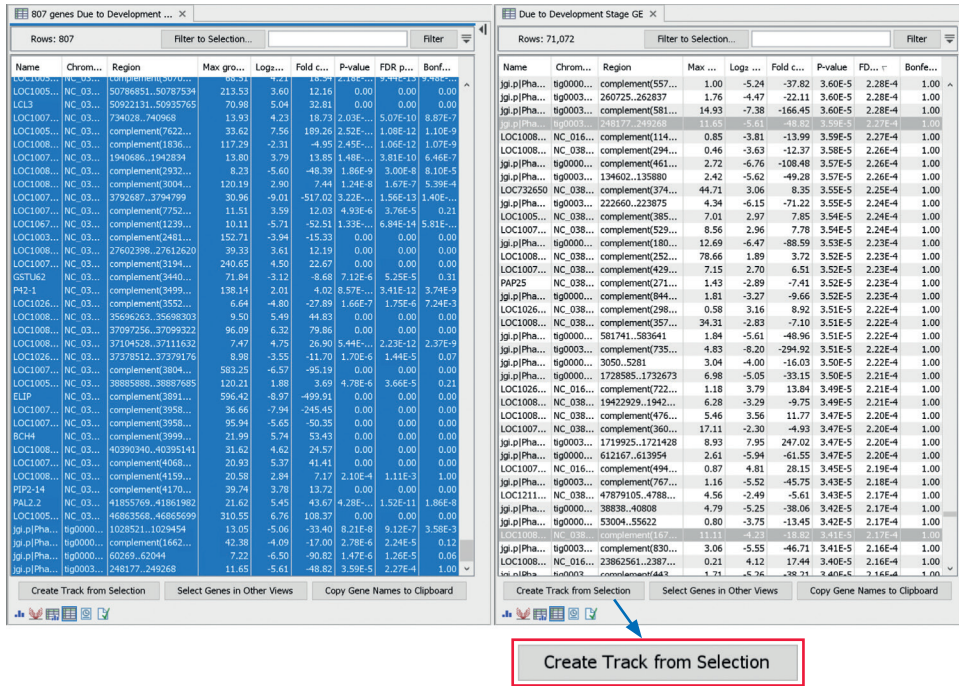


Figure 18. Creating a track with genes selected in the Venn diagram.

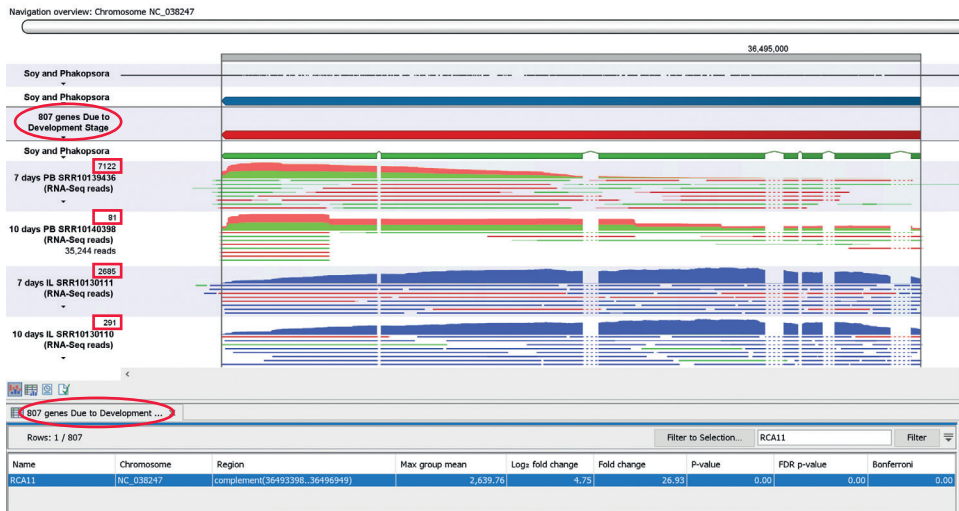


Figure 19. Selecting and navigating to particular genes in a list of tracks.

Problems of insufficient annotation can be addressed in the Workbench using the *ab initio* tools. This is done by creating new structural annotation tracks before running RNA-seq analysis. These tools are found in the “Transcript Discovery in the RNA-Seq and Small RNA Analysis” folder (Figure 9).

The “*ab initio* Transcript Discovery” tools are covered in detail in the Application Note “Improving structural annotation in complex genomes with QIAGEN CLC Genomics Workbench”.

## Summary

This application note covered the steps required for RNA-seq analysis of a host-pathogen system using short and long reads. The workflow described here started with the import of references from different sources, as well as 32 RNA-seq files from SRA. The analysis allowed the seamless identification of differentially expressed genes using various conditions and constraints. We used several options for data analysis and visualization, including graphical tools that allowed us to navigate and select data. Using the mapping browser, a number of annotation problems were identified in the reference data, and these were resolved using Workbench *ab initio* annotation tools.

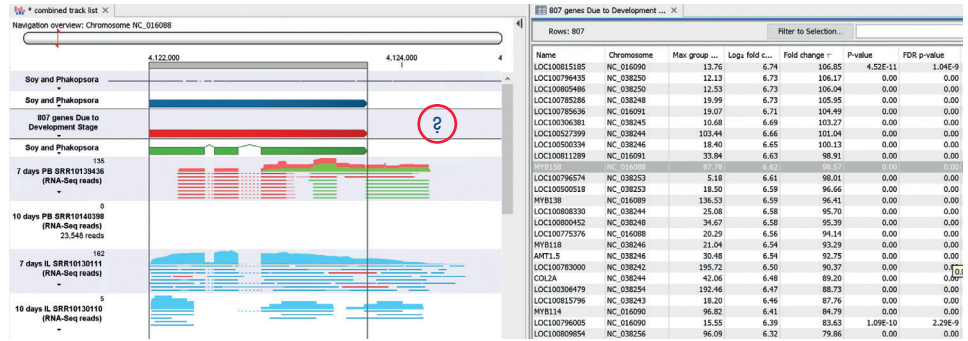


Figure 20. A soybean gene with incomplete annotation.

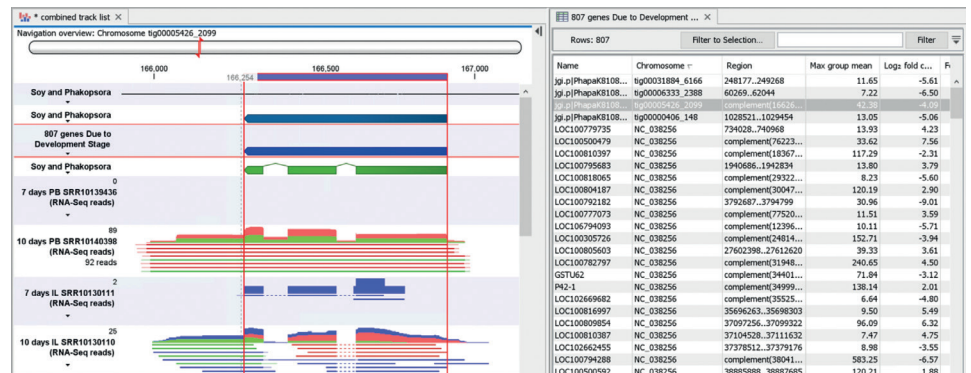


Figure 21. An example of a truncated annotation in a *Phakopsora* gene. The read mapping extends beyond the gene annotation on both ends.

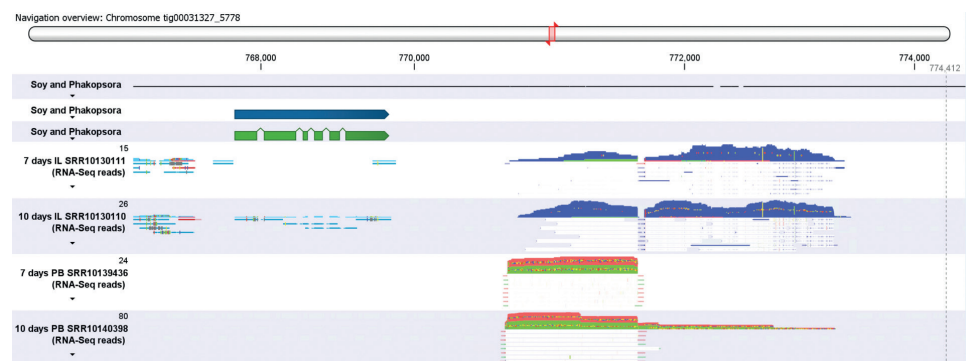



Figure 22. An example of an unannotated but expressed gene in *Phakopsora*.

## References

Elmore, M.G., et al. (2020) *De novo* transcriptome of *Phakopsora pachyrhizi* uncovers putative effector repertoire during infection. *Physio Molec Plant Path* 110. <https://doi.org/10.1016/j.pmpp.2020.101464>

 Learn more and request a free trial at [digitalinsights.qiagen.com/GXWBP](https://digitalinsights.qiagen.com/GXWBP).

These products are not intended for the diagnosis, prevention or treatment of a disease.

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN CLC Genomics product site. Further information can be requested from [ts-bioinformatics@qiagen.com](mailto:ts-bioinformatics@qiagen.com) or by contacting your local account manager.

Trademarks: QIAGEN®, Sample to Insight® (QIAGEN Group); Illumina® (Illumina, Inc.); PacBio® (Pacific Biosciences). Registered names, trademarks, etc. used in this document, even when not specifically marked as such, may still be protected by law.

01/2022 1126624 PROM-19784-001 © 2022 QIAGEN, all rights reserved.

Ordering [www.qiagen.com/bioinformatics](https://www.qiagen.com/bioinformatics) | Technical Support [digitalinsights.qiagen.com/support](https://digitalinsights.qiagen.com/support) | Website [digitalinsights.qiagen.com](https://digitalinsights.qiagen.com)