SEQUENCHER®

Tutorial for Windows and Macintosh

RNA-Seq

Gene Codes Corporation

TCA GENE
AGT CODES

Gene Codes Corporation
525 Avis Drive, Ann Arbor, MI 48108 USA
1.800.497.4939 (USA) +1.734.769.7249 (elsewhere)
+1.734.769.7074 (fax)
www.genecodes.com gcinfo@genecodes.com

RNA-Seq

About File Formats 3

Getting Started 3

Using the External Data Browser 3

Step 1 Re-aligning Your Reads using Cufflinks 4

Step 2 Merging the Cufflinks Output Files using Cuffmerge 6

Step 3 Quantifying Your Reads using Cuffquant 7

Step 4 Testing for Differential Expression with Cuffdiff 8

Step 5 Viewing Your Results 12

Conclusion 14

RNA-Seq Differential Expression using Cufflinks Suite

Next-Generation Sequencing requires new algorithms to process the large quantity of data produced. This ability has been harnessed to determine the levels of gene expression by directly sequencing RNA extracted under different experimental states such as time series or normal and disease-state tissues.

With **Sequencher**, you can choose to use either **GSNAP** or **BWA-MEM** to align your RNA-Seq sequences. You can then analyse the data for differential expression. Even if you have already aligned your data using another algorithm, you can still use **Sequencher** to analyse the results for differential expression.

Please see the 'Using DNA-Seq Tools with Sequencher' tutorial for detailed help in setting up your machine to use **GSNAP** and **BWA-MEM**, as well as the associated **Tablet** viewer.

Please note that the ability to perform RNA-Seq Differential Expression using the Cufflinks Suite in **Sequencher** is only supported on 64-bit operating systems.

About File Formats

In this tutorial, you will be provided with two BAM and two GTF files. If you want to use your own data, you will need to provide your own SAM or BAM files (referred to in the tutorial as SAM/BAM files) and a reference file in GTF format. If you have been working with a well-characterized genome, then you will probably be able to obtain such a file from the UCSC web site at http://genome.ucsc.edu/.

Getting Started

In this tutorial, you will use programs from the **Cufflink**s suite in **Sequencher** to analyze your aligned RNA-Seq reads. You do not need a special project unless you want to align your own reads from scratch. If you have not used **Sequencher** before for NGS data, please refer to the 'Next Gen Sequence Alignment' and 'Advanced Next Gen Sequence Alignment' tutorials. They will help you to align your FastQ reads to a reference genome which is required in order to generate a SAM file.

Using the External Data Browser

The **External Data Browser** is an important tool when managing any of your analyses using External Tools. You will find it especially useful for managing your RNA-Seq analyses. The browser allows you to add notes to any analysis you perform. These notes are attached to the Run folder containing the results of the analysis. This enables you to track your Runs, make notes about the data you used, add information about the parameters you used, or just add a title so you know what the experiment entailed. You will find these notes especially useful as you work through the RNA-Seq workflow. The **External Data Browser** will launch automatically anytime you perform a **DNA-Seq** or **RNA-Seq** assembly or alignment. When in **Sequencher**, do the following step to view your **DNA-Seq**, **RNA-Seq**, or **MUSCLE** assemblies and alignments:

- Go to the **Window** menu and choose **Open External Data Browser**.

As the various analyses are performed, you will see these appear in the **External Data Browser** dialog and you can follow the progress of the run by looking at the log file that is displayed in the bottom pane.

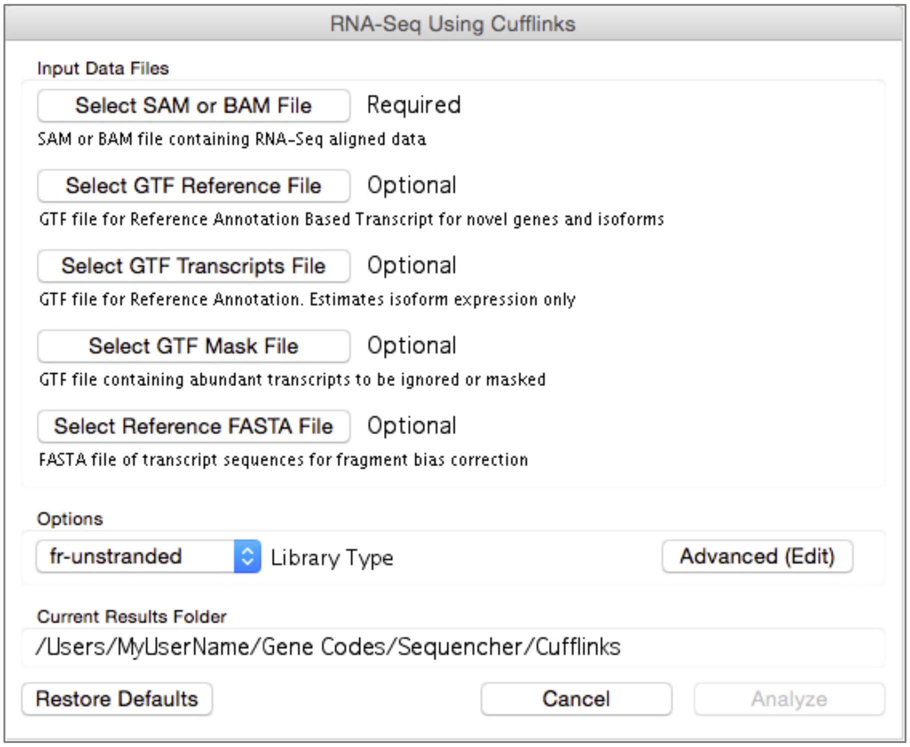
- To update the view in the **External Data Browser,** click on the **Refresh** button.

Step 1 Re-aligning Your Reads using Cufflinks

In Step 1, two SAM or BAM files are separately aligned to the GTF reference file. The output from this step is a transcripts.gtf file and will be used in Step 2 if you are performing Differential Expression analysis.

- Launch **Sequencher**.
- Go to the **Assemble** menu and select **RNA-Seq Using Cufflinks…**

The **RNA-Seq Using Cufflinks** dialog will open along with the **External Data Browser** if it is not already open.



RNA-Seq Using Cufflinks

Input Data Files

Select SAM or BAM File Required

SAM or BAM file containing RNA-Seq aligned data

Select GTF Reference File Optional

GTF file for Reference Annotation Based Transcript for novel genes and isoforms

Select GTF Transcripts File Optional

GTF file for Reference Annotation. Estimates isoform expression only

Select GTF Mask File Optional

GTF file containing abundant transcripts to be ignored or masked

Select Reference FASTA File Optional

FASTA file of transcript sequences for fragment bias correction

Options

fr-unstranded Library Type Advanced (Edit)

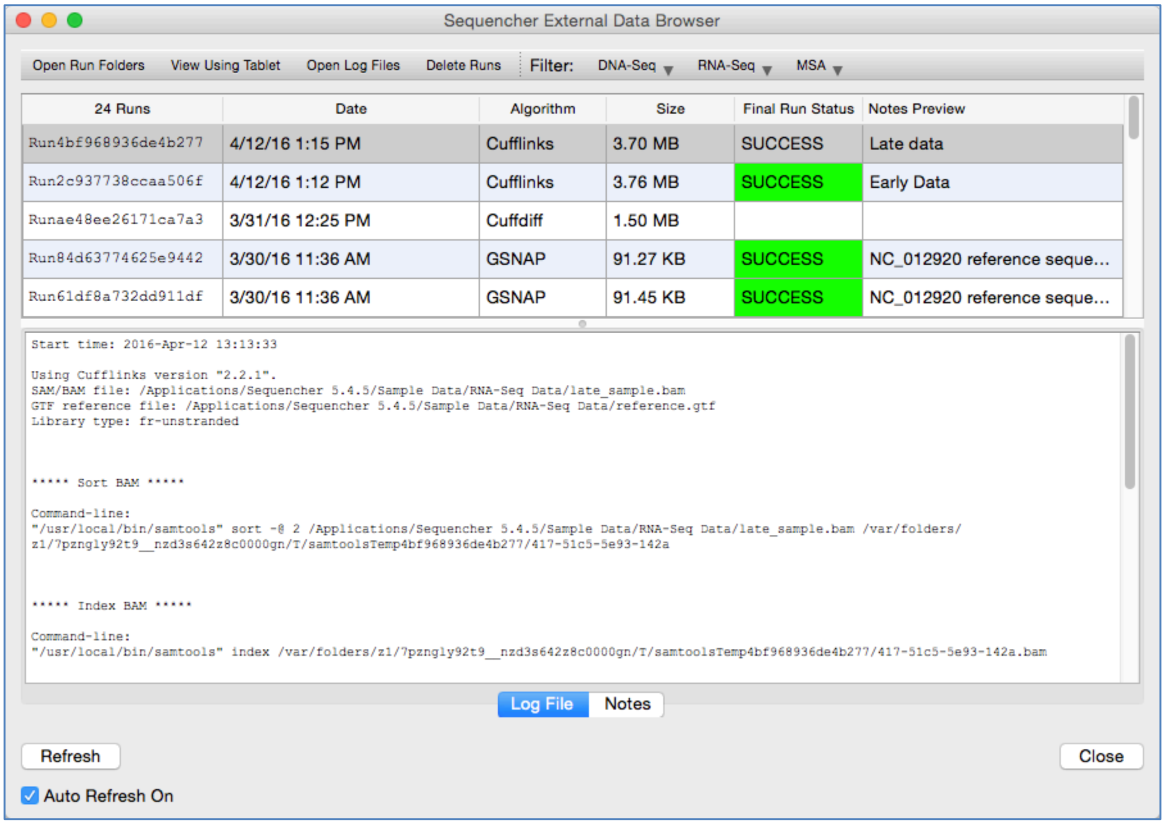
Current Results Folder

/Users/MyUserName/Gene Codes/Sequencher/Cufflinks

Restore Defaults Cancel Analyze

- Click on the **Select SAM or BAM File** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder.
- Navigate to the **RNA-Seq Data** folder.
- The folder contains two BAM files and two reference files. Double-click on the file called **early_sample.bam**.
- Click on the **Select GTF Reference File** button.

- **Sequencher** remembers the last place you navigated to. Double-click on the GTF reference file **reference.gtf**.
- Now click on the **Analyze** button.
- In the **External Data Browser**, follow the progress of the processing in the Log File tab. If the **Auto Refresh On** widget is checked and the currently running analysis Run is selected in the browser, refreshes will occur periodically for the current run. You can select **Refresh** at any time to get a progress update.
- Click on the **Notes** tab. Click in the Notes field and type "**Early data**" and then click on the **Save** button.
- When the analysis is completed, repeat the previous steps for the second SAM/BAM file called late_sample.bam. In the Notes field, you will need to add an annotation referring to "**Late data**" and save the note.



24 Runs	Date	Algorithm	Size	Final Run Status	Notes Preview
Run4bf968936de4b277	4/12/16 1:15 PM	Cufflinks	3.70 MB	SUCCESS	Late data
Run2c937738ccaa506f	4/12/16 1:12 PM	Cufflinks	3.76 MB	🟩 SUCCESS	Early Data
Runae48ee26171ca7a3	3/31/16 12:25 PM	Cuffdiff	1.50 MB		
Run84d63774625e9442	3/30/16 11:36 AM	GSNAP	91.27 KB	🟩 SUCCESS	NC_012920 reference seque...
Run61df8a732dd911df	3/30/16 11:36 AM	GSNAP	91.45 KB	🟩 SUCCESS	NC_012920 reference seque...

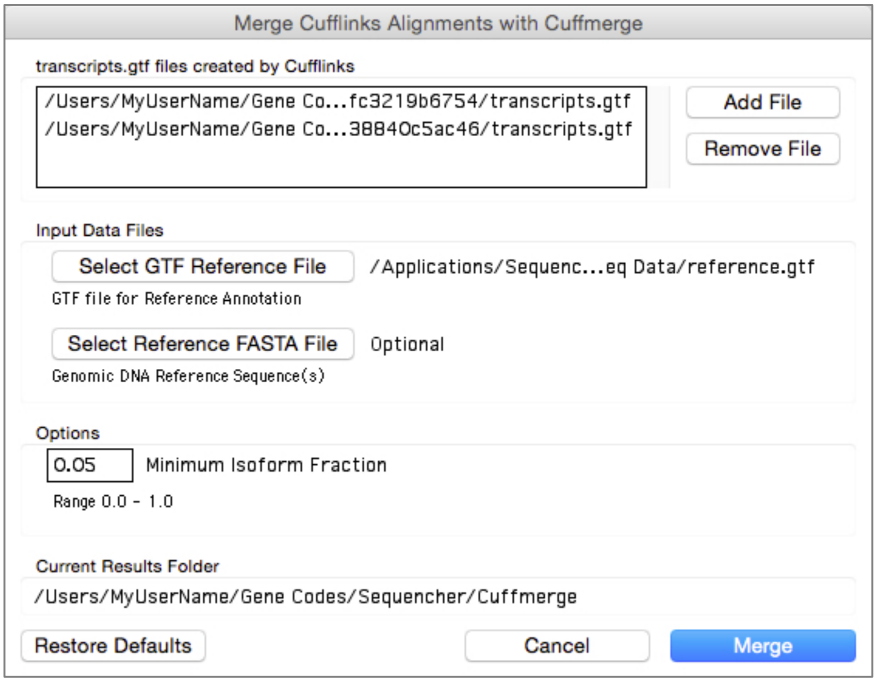
```
Start time: 2016-Apr-12 13:13:33

Using Cufflinks version "2.2.1".
SAM/BAM file: /Applications/Sequencher 5.4.5/Sample Data/RNA-Seq Data/late_sample.bam
GTF reference file: /Applications/Sequencher 5.4.5/Sample Data/RNA-Seq Data/reference.gtf
Library type: fr-unstranded



***** Sort BAM *****

Command-line:
"/usr/local/bin/samtools" sort -@ 2 /Applications/Sequencher 5.4.5/Sample Data/RNA-Seq Data/late_sample.bam /var/folders/
z1/7pzngly92t9__nzd3s642z8c0000gn/T/samtoolsTemp4bf968936de4b277/417-51c5-5e93-142a



***** Index BAM *****

Command-line:
"/usr/local/bin/samtools" index /var/folders/z1/7pzngly92t9__nzd3s642z8c0000gn/T/samtoolsTemp4bf968936de4b277/417-51c5-5e93-142a.bam
```

Step 2 Merging the Cufflinks Output Files using Cuffmerge

In Step 2, the output files from the **Cufflinks** runs (transcripts.gtf) are merged to create a 'consensus' file that provides the basis for calculating gene and transcript expression for each sample.

- Go to the **Assemble** menu and select **Merge Cufflinks Alignments with Cuffmerge…**

The **Merge Cufflinks Alignments with Cuffmerge** dialog will open along with the **External Data Browser** if it is not already open.



Merge Cufflinks Alignments with Cuffmerge

transcripts.gtf files created by Cufflinks

```
/Users/MyUserName/Gene Co...fc3219b6754/transcripts.gtf
/Users/MyUserName/Gene Co...38840c5ac46/transcripts.gtf
```

Input Data Files

Select GTF Reference File: /Applications/Sequenc...eq Data/reference.gtf

GTF file for Reference Annotation

Select Reference FASTA File: Optional

Genomic DNA Reference Sequence(s)

Options

0.05 Minimum Isoform Fraction

Range 0.0 - 1.0

Current Results Folder

/Users/MyUserName/Gene Codes/Sequencher/Cuffmerge

- Click on the **Add File** button.
- Navigate to the **Documents** folder, then to the **Gene Codes** folder, then to the **Sequencher** folder inside it. Within the **Sequencher** folder, you will see a **Cufflinks** folder. Navigate to it.

In the next step, use the information you added to the Notes in the **External Data Browser** to guide you to the correct Run folders.

- Navigate into the Run folder associated with the 'Early data' run.
- The folder contains a **transcripts.gtf** file. Select it and click on the **Open** button.
- **Sequencher** adds this file to the list.
- Repeat the steps above to add the **transcripts.gtf** file to the list for the 'Late data' run.
- **Sequencher** adds this file to the list also.
- Click on the **Select GTF Reference File** button.
- Navigate to the **RNA-Seq Data** folder you navigated to in Step 1 above, select the **reference.gtf** file, and then select the **Open** button.
- Now click on the **Merge** button.

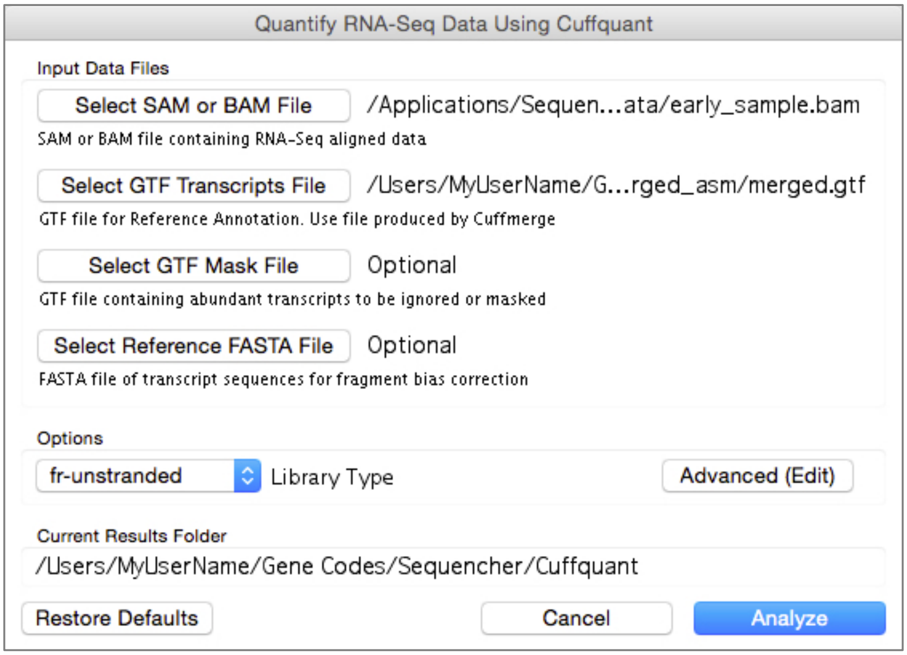
- In the **External Data Browser** dialog, follow the progress of the processing in the Log File tab using the **Refresh** button if **Auto Refresh On** is not checked.
- Click on the **Notes** tab for your **Cuffmerge** run. Type "**Merged Early and Late data**" in the notes field and click on the **Save** button.

Step 3 Quantifying Your Reads using Cuffquant

In Step 3, a SAM or BAM file is quantified against the GTF transcripts file. The output from this step is an abundances.cxb file and will be used in Step 4 if you are performing Differential Expression analysis. The reason for using **Cuffquant** is to help reduce the load on your computer by performing quantification as a separate step. You can skip this step and go directly to step 4 if you have a very powerful computer.

- Go to the **Assemble** menu and select **Quantify RNA-Seq Data Using Cuffquant**...

The **Quantify RNA-Seq Data Using Cuffquant** dialog will open along with the **External Data Browser** if it is not already open.



Quantify RNA-Seq Data Using Cuffquant

Input Data Files

Select SAM or BAM File: /Applications/Sequen...ata/early_sample.bam

SAM or BAM file containing RNA-Seq aligned data

Select GTF Transcripts File: /Users/MyUserName/G...rged_asm/merged.gtf

GTF file for Reference Annotation. Use file produced by Cuffmerge

Select GTF Mask File: Optional

GTF file containing abundant transcripts to be ignored or masked

Select Reference FASTA File: Optional

FASTA file of transcript sequences for fragment bias correction

Options

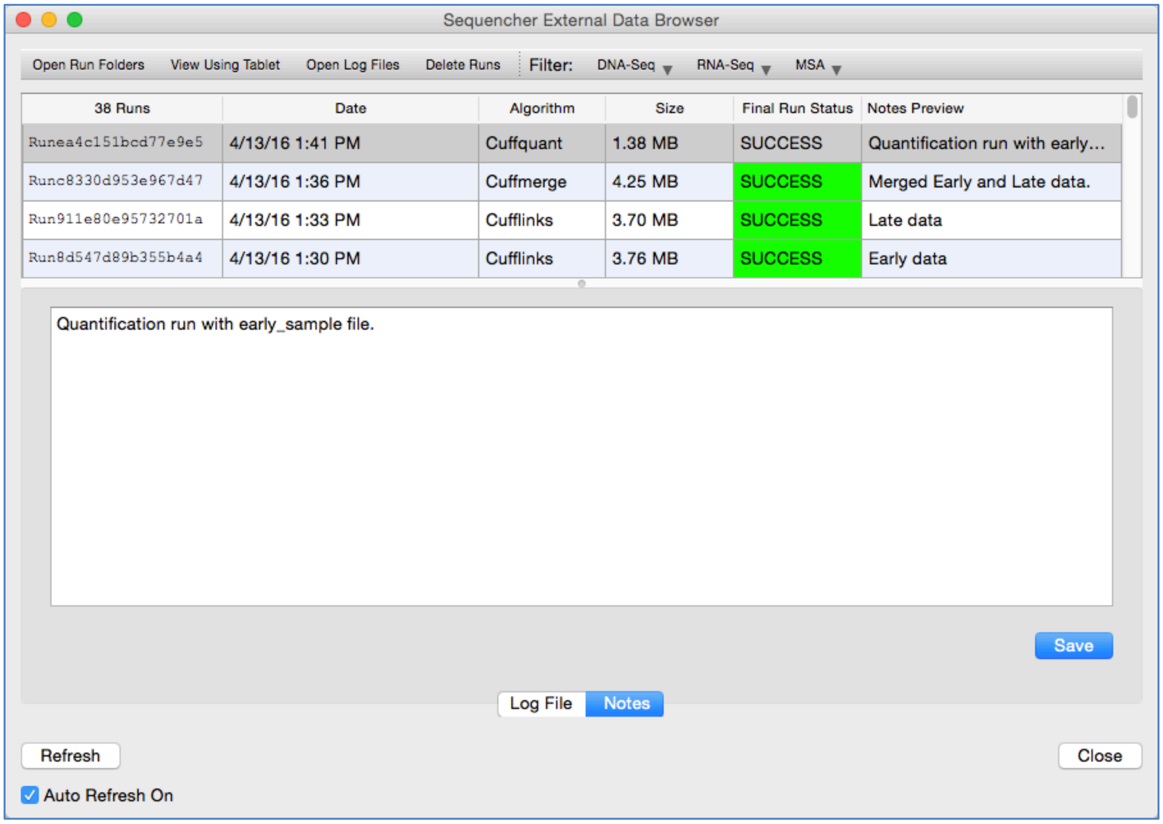
Library Type: fr-unstranded

Current Results Folder

/Users/MyUserName/Gene Codes/Sequencher/Cuffquant

- Click on the **Select SAM or BAM File** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder.
- Navigate to the **RNA-Seq Data** folder.
- The folder contains two BAM files and two reference files. Double-click on the file called **early_sample.bam**.
- Click on the **Select GTF Transcripts File** button.

- Navigate to the **Documents** folder, then to the **Gene Codes** folder, then to the **Sequencher** folder inside it. Within the **Sequencher** folder, you will see a **Cuffmerge** folder. Navigate to it.
- Navigate into the appropriate Run folder for the earlier **Cuffmerge** run and then into its **merged_asm** folder.
- Double-click on the **merged.gtf** file.
- Now click on the **Analyze** button.
- In the **External Data Browser** dialog, follow the progress of the processing in the Log File tab using the **Refresh** button if **Auto Refresh On** is not checked.
- Click on the **Notes** tab. Type "**Quantification run with early_sample file.**" in the notes field and click on the **Save** button.



38 Runs	Date	Algorithm	Size	Final Run Status	Notes Preview
Runea4c151bcd77e9e5	4/13/16 1:41 PM	Cuffquant	1.38 MB	SUCCESS	Quantification run with early…
Runc8330d953e967d47	4/13/16 1:36 PM	Cuffmerge	4.25 MB	SUCCESS	Merged Early and Late data.
Run911e80e95732701a	4/13/16 1:33 PM	Cufflinks	3.70 MB	SUCCESS	Late data
Run8d547d89b355b4a4	4/13/16 1:30 PM	Cufflinks	3.76 MB	SUCCESS	Early data

Quantification run with early_sample file.

☑ Auto Refresh On

- You will now need to repeat the previous steps for the second SAM/BAM file called late_sample.bam. In the Notes field, you will need to add an annotation referring to "**Quantification run with late_sample file.**" and **Save** that note.

Step 4 Testing for Differential Expression with Cuffdiff

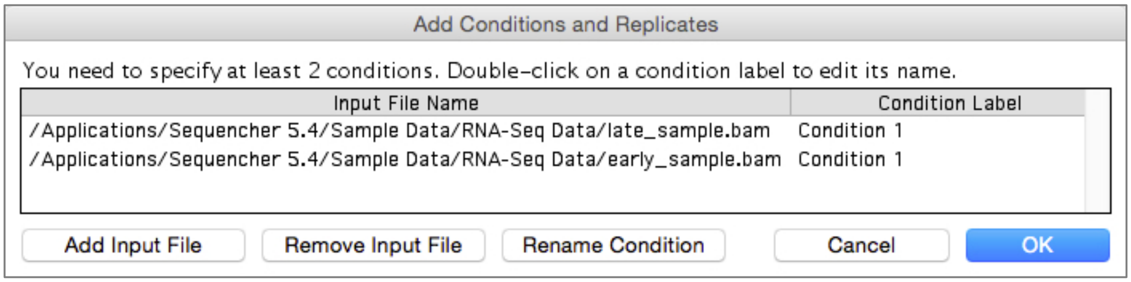
The final step in this analysis is testing for Differential Expression. In this step, you will use the original SAM/BAM files you used in Step 1. You will also use the merged.gtf file that was created in Step 2. It may be the case that certain files contain no results; this is not a failure of the program. It may indicate that no differential expression was detected. If you are following this tutorial with your own data, note that this can also happen if the GTF file you are using lacks the *tss_id* and *p_id* tags. Since the GTF file is a text file, this is easily checked. Note that all isoforms of a gene must have the *p_id* tag, otherwise Differential Expression will not be performed.

- Go to the **Assemble** menu and select **RNA-Seq Differential Expression Using Cuffdiff…**

The **DIFFERENTIAL EXPRESSION USING CUFFDIFF** dialog will open along with the **EXTERNAL DATA BROWSER** if it is not already open.

If you skipped Step 3, Quantifying Your Reads Using Cuffquant, continue with these steps:

- Click on the **Add/Remove Input Files** button.
- In the **Add Conditions and Replicates** dialog, click on the **Add Input File** button.
- Navigate to the **Sample Data** folder and then to the **RNA-Seq Data** folder inside it.
- The folder contains two BAM files and reference files. Double-click on the file **early_sample.bam**.
- Click on the **Add Input File** button again.
- Navigate again to the **Sample Data** folder and then to the **RNA-Seq Data** folder inside it.
- Double-click on the file **late_sample.bam**. The names of both files should now be listed in the **Add Conditions and Replicates** dialog.



You need to specify at least 2 conditions. Double-click on a condition label to edit its name.

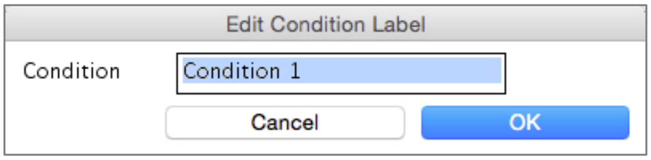
Input File Name	Condition Label
/Applications/Sequencher 5.4/Sample Data/RNA-Seq Data/late_sample.bam	Condition 1
/Applications/Sequencher 5.4/Sample Data/RNA-Seq Data/early_sample.bam	Condition 1

If you performed Step 3, Quantifying Your Reads Using Cuffquant, continue with these steps:

- Click on the **Add/Remove Input Files** button.
- In the **Add Conditions and Replicates** dialog, click on the **Add Input File** button.
- Navigate to the **Run** folder you annotated "**Quantification run with early_sample file.**"
- Change the file filter from **SAM BAM Files** to **CXB Files** and double-click on the file called **abundances.cxb.**
- Click on the **Add Input File** button again.
- Navigate to the **Run** folder you annotated "**Quantification run with late_sample file.**"
- Double-click on the file called **abundances.cxb** file.

Whether you used Step 3 or not, continue by renaming your conditions.

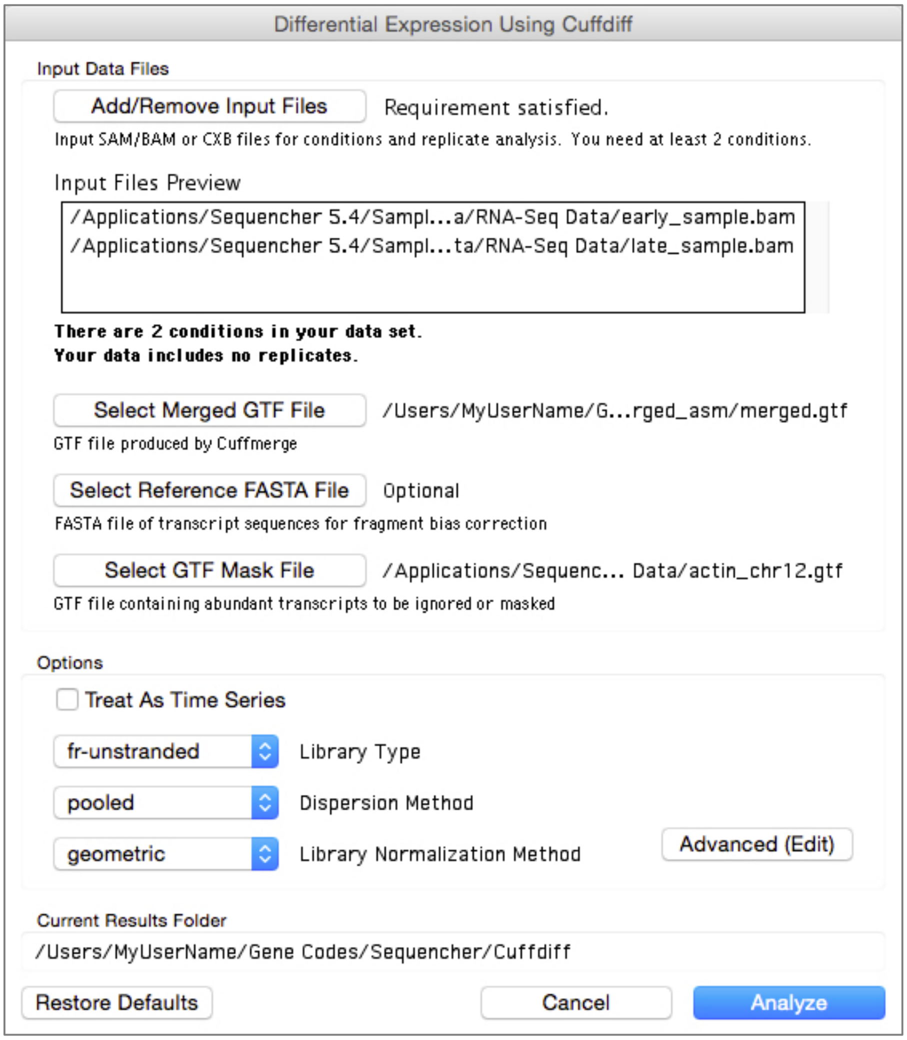
- You must have at least 2 uniquely named conditions in order to perform a differential expression. Double-click on the first condition for the first sample in the list to rename its label.



- Replace the name **Condition 1** with the name **Early** and click on the **OK** button.
- Double-click on the other condition in the list and change the name from **Condition 1** to **Late**.
- Click on the **OK** button.
- Back in the **Add Conditions and Replicates** dialog, click on the **OK** button.

In the next step, use the information you added to the Notes in the **External Data Browser** to guide you to the correct **Cuffmerge** Run folder.

- Click on the **Select Merged GTF File** button.
- Navigate to the **Documents** folder, then to the **Gene Codes** folder, then to the **Sequencher** folder inside it. Within the **Sequencher** folder, you will see a **Cuffmerge** folder. Navigate to it.
- Navigate into the appropriate Run folder for the earlier **Cuffmerge** run and then into its **merged_asm** folder.
- Double-click on the **merged.gtf** file.
- Click on the **Select GTF Mask File** button.
- Navigate to the **Sample Data** folder and open the **RNA-Seq Data** folder.
- Double-click on the **actin_chr12.gtf** file.



Differential Expression Using Cuffdiff

Input Data Files

Requirement satisfied.

Input SAM/BAM or CXB files for conditions and replicate analysis. You need at least 2 conditions.

Input Files Preview

```
/Applications/Sequencher 5.4/Sampl...a/RNA-Seq Data/early_sample.bam
/Applications/Sequencher 5.4/Sampl...ta/RNA-Seq Data/late_sample.bam
```

There are 2 conditions in your data set.
Your data includes no replicates.

Select Merged GTF File: /Users/MyUserName/G...rged_asm/merged.gtf
GTF file produced by Cuffmerge

Select Reference FASTA File: Optional
FASTA file of transcript sequences for fragment bias correction

Select GTF Mask File: /Applications/Sequenc... Data/actin_chr12.gtf
GTF file containing abundant transcripts to be ignored or masked

Options

☐ Treat As Time Series

fr-unstranded — Library Type

pooled — Dispersion Method

geometric — Library Normalization Method

Current Results Folder

/Users/MyUserName/Gene Codes/Sequencher/Cuffdiff

Note that if you ran Step 3, the **Input Files Preview** window in the **Differential Expression Using Cuffdiff** dialog will reflect the .cxb files you selected instead of the .bam files shown in the above image.

- Click on the **Analyze** button.
- In the **External Data Browser**, follow the progress of the processing in the Log File tab. If the **Auto Refresh On** widget is checked and the currently running analysis Run is selected in the browser, refreshes will occur periodically for the current run. You can select **Refresh** at any time to get a progress update.
- Click on the **Notes** tab for your **Cuffdiff** run. Type "**Differential expression run with actin mask file. Labeled Early and Late**." in the notes field and click on the **Save** button.

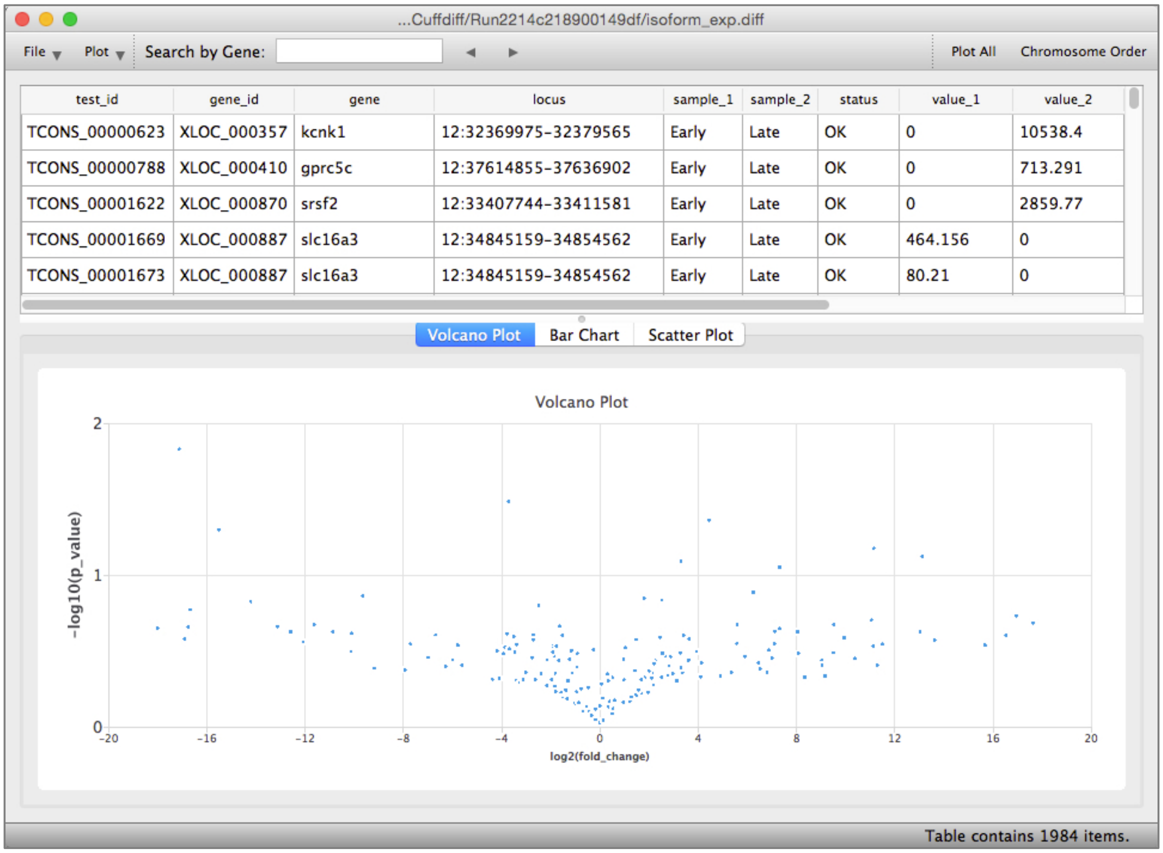
Step 5 Viewing Your Results

The final step in the analysis is to view the results. You can view most of the results files as tables. There are two file types you can view as plots or charts.

The most useful type of plot for obtaining an overall view of your results is the Volcano Plot; it plots two values in the form of a scatter plot. The plot takes the p-value for each result in the diff file and converts it to its –log10 value. It then takes the corresponding log2(fold_change) in expression directly from the table and places the result on the plot. The plot shows a spray of dots emanating from a central point. The most statistically significant changes appear higher on the plot while those with the greatest magnitude change in expression appear to the extreme left or right of the plot.

- Go to the **View** menu and select **Display RNA-Seq Data & Plots...**
- Browse to your **External Data Home** folder.
- Open the **Gene Codes > Sequencher** folder and open the most recent run folder in the **Cuffdiff** folder. The file picker automatically filters out any data that cannot be viewed as a table, plot, or chart.
- Choose the **isoform_exp.diff** file and select the **Open** button.

The .diff file is displayed in a window containing two panes. The top pane contains your data in tabular form and the lower pane contains the plots and charts. It opens in the **Volcano Plot** view by default.

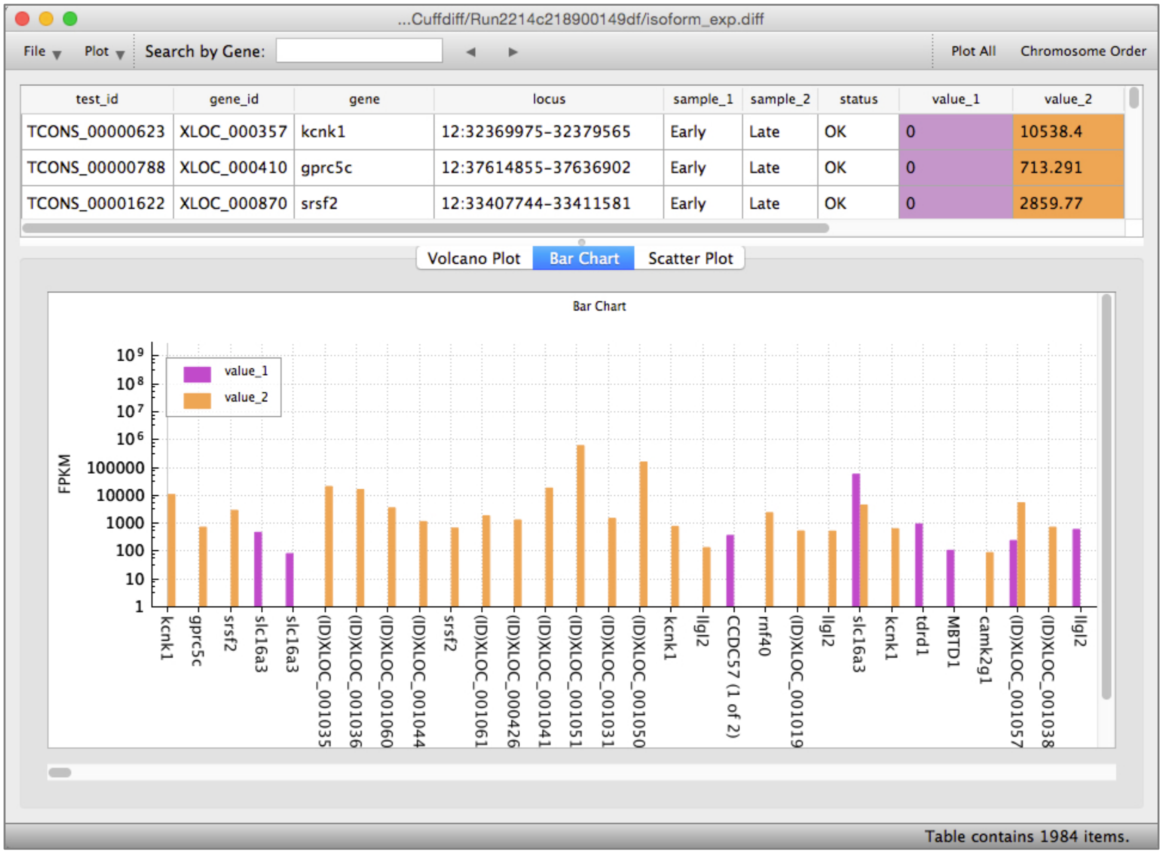


test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2
TCONS_00000623	XLOC_000357	kcnk1	12:32369975-32379565	Early	Late	OK	0	10538.4
TCONS_00000788	XLOC_000410	gprc5c	12:37614855-37636902	Early	Late	OK	0	713.291
TCONS_00001622	XLOC_000870	srsf2	12:33407744-33411581	Early	Late	OK	0	2859.77
TCONS_00001669	XLOC_000887	slc16a3	12:34845159-34854562	Early	Late	OK	464.156	0
TCONS_00001673	XLOC_000887	slc16a3	12:34845159-34854562	Early	Late	OK	80.21	0

Table contains 1984 items.

Another chart that is useful is the bar chart. In this chart, the values from the two different samples are plotted side-by-side for each gene. The difference in expression is immediately visible as you scroll across the chart from left to right and see locations where one bar is elevated relative to the other or may be completely missing.

- To view a bar chart, click on the **Bar Chart** tab above the plot area.
- The chart is created and displayed and you can view the data by scrolling left and right.
- Look closely at any position where there is either only one bar or one bar is significantly higher than its partner from the other sample file.
- Where the gene has a name, you can obtain more information from the table by locating its row.

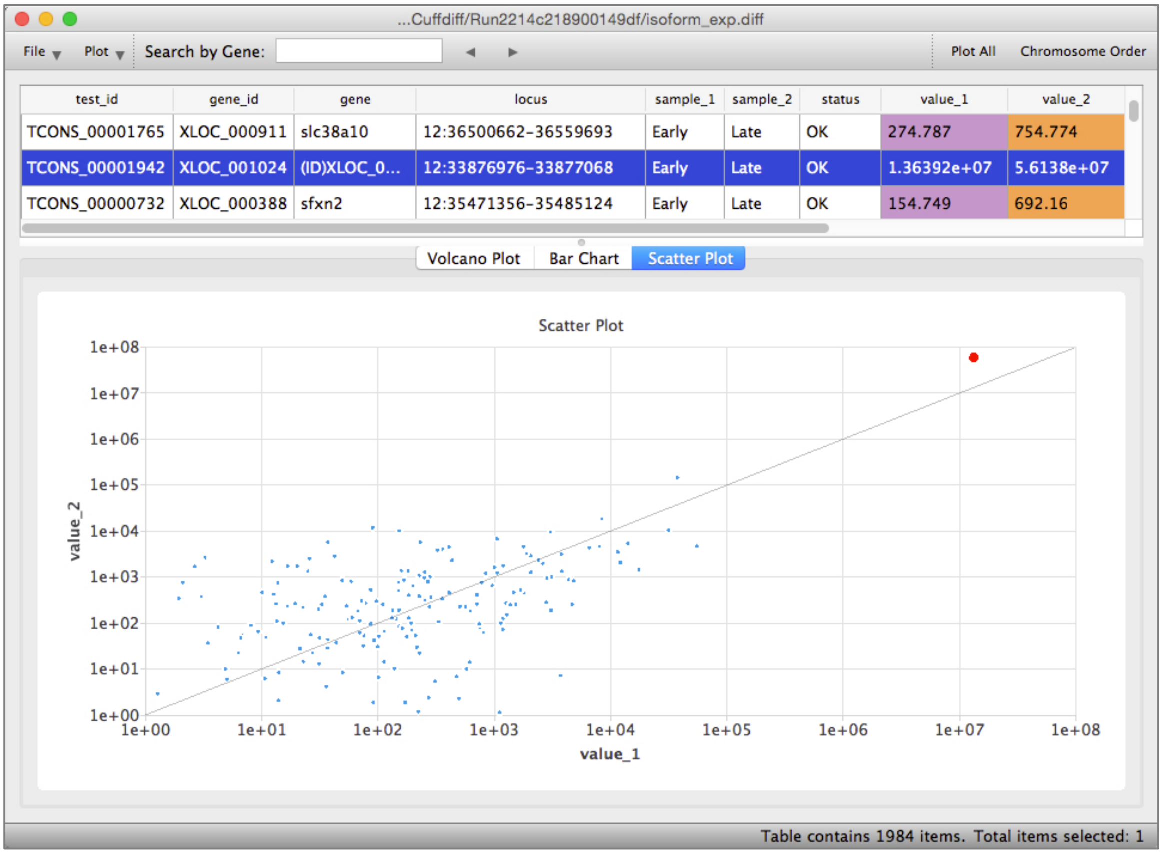


test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2
TCONS_00000623	XLOC_000357	kcnk1	12:32369975-32379565	Early	Late	OK	0	10538.4
TCONS_00000788	XLOC_000410	gprc5c	12:37614855-37636902	Early	Late	OK	0	713.291
TCONS_00001622	XLOC_000870	srsf2	12:33407744-33411581	Early	Late	OK	0	2859.77

Table contains 1984 items.

The final plot type is the Scatter Plot. In this plot, the FPKM values from each sample are plotted on a Log10 scale. This will allow you to see whether there are significant differences in expression based on the FPKM metric.

To view a scatter plot, click on the **Scatter Plot** tab.



...Cuffdiff/Run2214c218900149df/isoform_exp.diff

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2
TCONS_00001765	XLOC_000911	slc38a10	12:36500662-36559693	Early	Late	OK	274.787	754.774
TCONS_00001942	XLOC_001024	(ID)XLOC_0...	12:33876976-33877068	Early	Late	OK	1.36392e+07	5.6138e+07
TCONS_00000732	XLOC_000388	sfxn2	12:35471356-35485124	Early	Late	OK	154.749	692.16

Table contains 1984 items. Total items selected: 1

Conclusion

In this tutorial, you have worked with **Sequencher's RNA-Seq** tools. You have learned how to use the **External Data Browser** to annotate your analytical steps. You have also learned how to add the different file types to each step of the **RNA-Seq** workflow. You have learned how to use **Cuffquant** to break up the workflow and reduce your computational load. Finally, you have learned how to visualize the results files as tables and as plots.