

Expression Analysis using RNA-Seq

June 27, 2019

Sample to Insight

_

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com · ts-bioinformatics@qiagen.com



Expression Analysis using RNA-Seq

This tutorial will show how the RNA-Seq Analysis tools facilitate the expression analysis of RNA-Seq data.

Prerequisites For this tutorial, you must be working with the CLC Genomics Workbench 10.0 or higher.

Download and import data

This tutorial will use RNA-Seq data for male and female *Drosophila melanogaster* from 3 different strains, collected at different time points during the year, and sequenced using 2 different RNA preparation protocols. Our goal is to find genes that are differentially expressed between the different strains, while controlling for the effects of the other factors.

To allow the completion of the tutorial in a reasonable amount of time, we have reduced the samples to only 10 individuals out of the 726 present in the original dataset [Lin et al., 2016], and have reduced each sample to 250,000 reads mapping to chromosome 2R.

The data set includes:

- The reads from 10 Drosophila samples.
- An Excel spreadsheet that contains the metadata associated with each individual.
 - SRR_ID is an SRA identifier unique for each individual.
 - DGRP_Number describes the strain of the fly.
 - Sex stated as M for males and F for females.
 - Environment stated as 2 and 3 for different calendar times for collecting the flies.
 - RNA_Prep_Method using QIAGEN RNeasy kit in all cases but following either the centrifuge or the vacuum based protocol.
 - Lane of the sequencer on which the sample was loaded.
- A workflow to rapidly and efficiently process statistical analyses on the RNA-Seq results.
- A GO annotation file for Drosophila.

Go through the following steps to download and import the data into the Workbench.

- 1. Download the sample data from our website: http://resources.qiagenbioinformatics. com/testdata/RNA_Seq_Droso2.zip.
- 2. Unzip the zip file on your local computer.
- 3. Start the workbench.
- 4. First, import the workflow and the gene association file we will need later in this tutorial: **File** | **Import** () | **Standard Import** () and choose the files called "RNA-Seq statistics workflow" and "gene_association.fb". Save the imported data in a new folder you can create in your Navigation Area (called for example RNA-Seq Drosophila).



5. Next, import the reads via the toolbar: File | Import () | Illumina (). Choose the 10 *.fastq files, and leave the option "Paired reads" unchecked. You can however "Discard read names", "Discard quality scores" and "Remove failed reads" (figure 1). Click Next.

Gx Illumina High-Throughpu	ut Sequencing Import
Choose where to run Import files and options	Select files of types Ilumina files (.txt/.fasta/.fg/.qpeq) Look in: Image: Constraint of the selection of the select
Out of the	General potions Paired reads Discard read names Discard read names Discard quality scores Minimum distance 200 Maximum distance 550
Carlo I Promission	Illumina options Illumina options CEI/Sanger or Illumina Pipeline 1.8 and later MISeq de multiplexing Trim reads
? 🥱	Previous Next Finish Cancel

Figure 1: In the Import wizard, select all 10 reads files.

- 6. Now import the metadata table via the toolbar: File | Import ((4) | Import Metadata (1).
- 7. A wizard opens. In the first dialog, select the spreadsheet saved on your local computer that contains the sample information "DrosophilaMetadata.xls". The contents of the Excel spreadsheet populates the table situated at the bottom of the dialog (figure 2). Click **Next**.

. Select spreadsheet	Selection of metad Spreadsheet wit C:\Users\		eq\DrosophilaN	1etadata.xlsx			6
	Metadata previe	w					
	SRR_ID	DGRP_Number	Sex	Environment	RNA_Prep_Me	Lane	
	SRR 1543488	DGRP-93	F	2	Centrifuge	4	
	SRR 1543627	DGRP-93	м	2	Vacuum	3	_
	SRR 1543821	DGRP-93	M	3	Centrifuge	4	
	SRR 1543532	DGRP-320	M	2	Centrifuge	5	
and and and	SRR 1543624	DGRP-320	M	2	Vacuum	4	=
Marian Carling	SRR 1543751	DGRP-320	м	3	Vacuum	7	-
10 J	SRR 1543802	DGRP-320	м	3	Centrifuge	3	
1.0	SRR 1543519	DGRP-370	м	2	Centrifuge	5	
	SRR 1543733	DGRP-370	м	3	Vacuum	6	
C Manaaralut	SRR 1543767	DGRP-370	F	3	Vacuum	2	*
Post 1175							
2 6				Previous	Next		ncel

Figure 2: In the Import Metadata wizard, select the spreadsheet with the sample information.

- Click on the Navigation button next to "Location of data", and find the imported reads in your Navigation Area. Right-click the name of the folder and select "Add folder contents" (figure 3). Click OK.
- 9. In the wizard, now you can see that the data association is successful by the fact that the metadata has filled each row of the table located at the bottom of the dialog. Click **Next**.
- 10. Select the "RNA-Seq Drosophila" folder to save the resulting metadata table.
- 11. Now open the metadata table you just saved (called "Samples" as can be seen in the upper part of figure 4). Select all rows in the table, and click **Find Associated Data**. This will show that all rows are associated correctly with the corresponding reads.



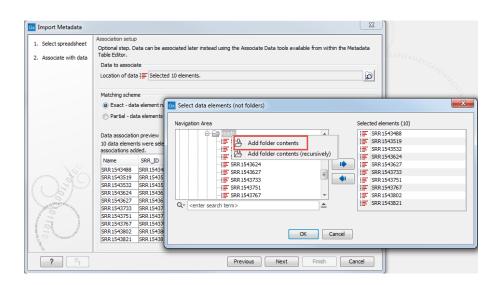


Figure 3: In the Import Metadata wizard, select the reads.

Rows: 10	Metadata						Filter
R_ID	DGRP_Num	ber Se	x	Environment	RNA_Prep_Method	Lane	
R1543488	DGRP-93				Centrifuge	4	
Rows: 10	DGRP-370	F anage Data		3	Vacuum	6 2 Addition	
21543767	DGRP-370	F anage Data		3	Vacuum	2	
1543767 Set Up Table. Rows: 10	DGRP-370	F anage Data Elements	Find Assoc	3 iated Data	Vacuum	2 Addition Filt	
Set Up Table. Rows: 10 SRR_ID	DGRP-370	F anage Data Elements Type	D Find Assoc	3 iated Data	Vacuum	2 Addition	
 Ret Up Table. Rows: 10 SRR_ID SRR1543488 	DGRP-370	Elements	Name SRR 1543488	3 iated Data Path CLC_Data /I CLC_Data /I	Vacuum	2 Addition Filt	
 Rows: 10 SRR_ID SRR1543488 SRR1543627 	DGRP-370 Metadata Role Sample Sample	Elements	Name SRR 1543488 SRR 1543627	ated Data Path C.C. Data //	Vacum Associate Data Associate Data RNA Seq Drosophila / reads RNA Seq Drosophila / reads RNA Seq Drosophila / reads	2 P Addition Filt	
✗1543767 ✗ Set Up Table. Rows: 10 SRR_ID SRR 1543488 SRR 1543627 SRR 1543627 SRR 1543521 SRR 1543521 SRR 1543524	DGRP-370 Metadata Role Sample Sample Sample Sample	Elements	Name SRR 1543488 SRR 1543627 SRR 1543821 SRR 1543627 SRR 1543624	ated Data Path CLC_Data // CLC_Data // CLC_Data // CLC_Data // CLC_Data // CLC_Data //	Vacum Associate Data RNA Seq Drosophila / reads	2 P Addition Filt	
 Set Up Table. Rows: 10 SRR_ID SRR1543488 SRR1543821 SRR1543821 SRR1543522 SRR1543524 SRR1543524 SRR1543524 SRR1543524 	DGRP-370 Metadata Metadata Role Sample Sample Sample Sample Sample	Elements	Name SR 1543488 SR 1543627 SR 1543627 SR 1543521 SR 1543521 SR 1543524 SR 1543524	Path CLC_Data // CLC_Data // CLC_Data // CLC_Data // CLC_Data // CLC_Data //	Vacum Associate Data Associate Data RNA Seq Drosophila / reads	2 Addition Filt	
✗1543767 ✗ Set Up Table. Rows: 10 SRR 1543488 SRR 1543488 SRR 1543821 SRR 1543521 SRR 1543523 SRR 1543525 SRR 1543524 SRR 1543525	DGRP-370 Metadata Metadata Role Sample Sample Sample Sample Sample	Elements	Name SRR 1543488 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543627 SR 1543624 SR 1543624 SR 1543626 SR 1543627	ated Data Path CLC_Data //	Vacuum Associate Data RNA Seq Drosophila / reads	2 Addition Filt	
✗1543767 ✗ Set Up Table. Rows: 10 SRR_ID SRR 1543488 SRR 1543821 SRR 1543821 SRR 1543821 SRR 1543824 SRR 1543751	DGRP-370 Metadata Role Sample Sample Sample Sample Sample Sample Sample Sample	Elements	Name SR 1543488 SR 1543488 SR 1543627 SR 1543627 SR 1543624 SR 1543523 SR 1543523 SR 1543624 SR 1543624	3 Path CLC_Data //	Vacuum Associate Data RNA Seq Drosophila / reads	2 P Addition Filt	
✗1543767 ✗ Set Up Table. Rows: 10 SRR 1543488 SRR 1543488 SRR 1543821 SRR 1543521 SRR 1543523 SRR 1543525 SRR 1543524 SRR 1543525	DGRP-370 Metadata Metadata Role Sample Sample Sample Sample Sample	Elements	Name SRR 1543488 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543627 SR 1543624 SR 1543624 SR 1543626 SR 1543627	ated Data Path C.C. Data // C.C	Vacuum Associate Data RNA Seq Drosophila / reads	2 P Addition Filt	

Figure 4: In the Metadata table, check that associations between reads and the metadata have happened successfully.

Data management configuration

In order to go through this tutorial, you need the reference sequence for *Drosophila melanogaster*. In order to access this data, you can download the relevant data directly from the workbench.

- 1. Open the **Reference Data Manager** (1) in the top right corner of the Workbench, and choose the tab **Download Genomes** (2).
- In the wizard, open the section Animals (others) (3), and select Drosophila melanogaster (4) from the drop down menu (figure 5). Check "Download genome sequence" (5) and add



a check to the row "Genome Annotations" (6). Click on **Download** (6). You can check the Download process at the bottom of the wizard.

					Workspace Plugins (References Downloa	Hand Workflows
Gx Manage Reference Data					1	×
Download Genomes Public Repositories	Custom Sets Reference Data Sets	Imported Data Imported Reference Data			Manage Reference Data: Free space in CLC_References locatio Free space in temporary folder locatio	
Animal (mammals) Animal (others) Anolis carolinensis Select to download latest data for Anolis carolinen Ceenorhabditis elegans Select to download latest data for Caenorhabditis		Drosophila melanogas Download genome Select how to get reference © Download genome sequ © Use existing genome set	sequence 4			
Danio rerio Select to download latest data for Danio rerio	3				Download size for selection: 44	
		Name	5 Version	Provider	Size	E
Drosophila melanogaster Select to download latest data for Drosophila mela	nogaster	Sequence	BDGP6	Ensembl	39.8 MB	
Gallus gallus Select to download latest data for Gallus gallus		Genome Annotations	BDGP6.94	Ensembl	5.0 MB 79.8 MB	
Taeniopygia guttata Select to download latest data for Taeniopygia gut	ttata	Previous downloads	In	nporting data	6 Downloa	ad
▶ Other		Date	Size	Versions	Folder	te 🗸
Help						Close

Figure 5: Download reference data for Drosophila melanogaster.

3. Click Download .

Once the data and references have been downloaded and imported, you should see the folder and files in the Navigation Area as shown in figure 6.

🕀 🔓 CLC_Data
B RNA Seg Drosophila
E-Carreads
SRR1543488
SRR 1543519
•••
SRR 1543532
SRR 1543767
SRR 1543802
SRR 1543821
- Joi Samples
😥 🚰 references
⊞… 📆 Recycle bin (0)
CLC_References
Genomes
🖶 🏝 Drosophila_melanogaster-2018-10-18-11-06
Trosophila melanogaster sequence
Drosophila melanogaster _three_prime_UTR
Drosophila melanogaster _tRNA
Drosophila melanogaster _ncRNA_gene
Drosophila melanogaster _rRNA
the state in the state of the s

Figure 6: The data set has been imported.

The reference tracks are now saved in the CLC_References | Genomes | Drosophila_melanogaster folder accessible from the Navigation Area.



RNA-Seq quantification step

To map the reads to the reference, we will use the RNA-Seq tool. Go to:

```
Toolbox | RNA-Seq Analysis (🚘) | RNA-Seq Analysis (🛬)
```

- 1. Depending on your setup you may be asked where to run the tool. Choose the appropriate option for you and click **Next**.
- 2. In the next window, **check the "Batch" box** and select the 10 reads files (figure 7).

Gx RNA-Seq Analysis		×
1. Choose where to run	Select sequencing reads	
1. Choose where to run	Navigation Area	Selected elements (10)
2. Select sequencing reads	E-G CLC_Data	SRR 1543488
	RNA Seq Drosophila	E SRR 1543519
	🖨 🗁 reads	E SRR 1543532
	SRR 1543488	E SRR 1543624
		E SRR 1543627
		E SRR 1543733
		E SRR 1543751
		SRR1543767
	SRR 1543733	
10	SRR 1543751	↓ SRR 1543821
C) FOR C	SRR 1543767	
	SRR 1543821	
Manara Maria	English CLC_References	
11400 Harris		
74		
O Distance	Q- <enter search="" term=""></enter>	≜
TO STATE ATTEND		
	V Batch	
?		Previous Next Finish Cancel

Figure 7: Do not forget to check the "Batch" mode option to make sure that the 10 samples are processed independently of each other, and not as the reads of a single individual.

- 3. In the next wizard window review the batch units and click **Next**. Failing to see this window means you have not checked the "Batch" option during the previous step.
- 4. Set up the wizard step so that it looks like in figure 8. Keep "Genome annotated with genes and transcripts" checked. The Reference sequence, Gene track and mRNA track were previously downloaded using the Download manager, and can now be found in your "references folder".

Click Next.

- 5. Use default parameters for the "Mapping options" dialog, except for the Strand specific option that should be set to **Reverse** (figure 9). Click **Next**.
- 6. In the "Expression level options" dialog, leave all parameters as default as well. You can use the **Reset** button if you are not sure whether you have previously changed parameters for this tool. Click **Next**.
- 7. Save the results in a specified location (a new RNA-Seq results folder for example) and create subfolders per batch unit.

After starting the tool, you can use the progress bar to see how it is progressing (figure 10). The mapping for these 10 samples takes around 20 minutes on a standard laptop computer.



Gx RNA-Seq Analysis		×
1. Choose where to run	Select reference References	
2. Select sequencing reads	Genome annotated with genes and transcripts	
3. Batch overview	Genome annotated with genes only	
4. Select reference	One reference sequence per transcript	
	Reference sequence 🎇 Drosophila melanogaster sequence	Ø
	Gene track Drosophila melanogaster_Gene	ø
Cr. M	mRNA track	Ø
South States of the states of	Spike-in controls (a) Do not use spike-in controls	
1 feel	Use spike-in controls	
TON YOL MANNING WITH WITH MANNING WITH WITH MANNING WITH WITH MANNING WITH WITH WITH WITH WITH WITH WITH WITH	Spike-in controls	\$
?	Previous Next Finish Can	cel

Figure 8: Genome annotated with genes and transcripts is checked.

Gx RNA-Seq Analysis			X
1. Choose where to run	Mapping options Read alignment		
2. Select sequencing reads	Mismatch cost	2	
3. Batch overview	Insertion cost	3	
3. Batch overview	Deletion cost	3	
4. Select reference	Length fraction	0.8	
5. Mapping options	Similarity fraction	0.8	
6	Global alignment		
OP.	✓ Color space alignment		
	Color error cost 3		
Marine Marine	✓ Auto-detect paired distances		
1 1 guarn	Strand specific	Reverse -	
1010	Maximum number of hits for a rea	d 10	
ALL AND RIGHTS			
<u> </u>			
?		Previous Next Finish C	Cancel

Figure 9: The RNA-Seq was sequenced using a reverse protocol, so set the strand specificity to "Reverse" for the mapping.

When the RNA-Seq Analysis tool has completed, you can click on the Refresh button of the Metadata Elements table (bottom part) of the "Samples" file. You can see that after analysis, we now have many more files associated to the initial samples (figure 11).

PCA plot

We now use the PCA plot to check for outlying samples of poor quality that may need excluding from further analysis. We will also use the plot to see which factors listed in the metadata may confound the detection of genes that are differentially expressed due to fly strain.

- 1. Click the "Role" column title of the Metadata Elements table to sort the elements by title and select the 10 rows whose role is "Gene expression" (figure 11). Alternatively, you can enter "Gene expression" in the Filter box and select all remaining rows.
- 2. Go to Toolbox | RNA-Seq Analysis (🚔) | PCA for RNA-Seq ([]).



and the state of t		
eq Analysis, batch 'SRR 1543488' (M	. 21 %	▼
rocess: RNA-Seq Analysis (0/10)	0 %	▼
ad Genome (Done)	100 %	▼
Toolbox Favorites		
	rocess: RNA-Seq Analysis (0/10) ad Genome (Done)	rocess: RNA-Seq Analysis (0/10) 0 % ad Genome (Done) 100 %

Figure 10: The process called "Batch Process" indicates how many batches have been completed, while the one called "RNA-Seq Analysis" shows the analysis progress of a particular batch unit.

Rows: 10	Metadata				Filter
RR_ID	DGRP_Number	Sex	Environment	RNA_Prep_Method	Lane
RR 1543488	DGRP-93	F	2	Centrifuge	4
RR 1543767					
🧷 Set Up Table	e 🧷 Manage Dat	ta 😡 Find	Associated Data	Associate Data	Additional Filterin
		ta 💭 🔊 Find	Associated Data	Associate Data	Ø Additional Filterin Filter
Set Up Table Rows: 50 SRR_ID	Manage Dai	ta Digi Find	Associated Data	Associate Data	Ø Additional Filterin Ø Filter
Rows: 50	Metadata Elements	Туре		Path	
Rows: 50	Metadata Elements	Туре	Name	Path	Filter =
Rows: 50 SRR_ID SRR1543488	Metadata Elements Role / Gene express to	Туре	Name SRR 1543488 (GE)	Path CLC_Dat CLC_Dat	Filter ₹
Rows: 50 SRR_ID SRR 1543488 SRR 1543627	Metadata Elements Role / Gene express to Gene expression 2	Type	Name SRR1543488 (GE) SRR1543627 (GE)	Path CLC_Dat CLC_Dat	Filter =
Rows: 50 SRR_ID SRR 1543488 SRR 1543627 SRR 1543821	Metadata Elements Role / Gene expression Gene expression Gene expression		Name SRR1543488 (GE) SRR1543627 (GE) SRR1543821 (GE)	Path CLC_Dat CLC_Dat CLC_Dat	Filter = ia / RNA Seq Drosophik a / RNA Seq Drosophik a / RNA Seq Drosophik a / RNA Seq Drosophik
Rows: 50 SRR_ID SRR1543488 SRR1543627 SRR1543821 SRR1543532	Metadata Elements Role , Gene expression Gene expression Gene expression		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543532 (GE)	Path CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat	Filter Filter Ta / RNA Seq Drosophik ta / RNA Seq Drosophik ta / RNA Seq Drosophik ta / RNA Seq Drosophik
Rows: 50 SRR_ID SRR 1543488 SRR 1543627 SRR 1543821 SRR 1543522 SRR 1543624	Metadata Elements Role / Gene expression Gene expression Gene expression Gene expression Gene expression		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543532 (GE) SRR 1543522 (GE)	Path CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat	Filter Fil
Rows: 50 SRR_ID SRR.1543488 SRR.1543827 SRR.1543821 SRR.1543822 SRR.1543524 SRR.1543525 SRR.1543525 SRR.1543519	Metadata Elements Role , Gene expression 2 Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543532 (GE) SRR 1543571 (GE)	Path CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat	Filter Fil
Rows: 50 SRR_ID SRR 1543488 SRR 1543627 SRR 1543627 SRR 1543521 SRR 1543521 SRR 1543521 SRR 1543627 SRR 1543802	Metadata Elements Role / Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543521 (GE) SRR 1543551 (GE) SRR 1543751 (GE) SRR 1543802 (GE)	Path CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat	Filter Filter
Rows: 50 SRR_ID SRR.1543488 SRR.1543827 SRR.1543821 SRR.1543822 SRR.1543524 SRR.1543525 SRR.1543525 SRR.1543519	Metadata Elements Role , Gene expression 2 Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression		Name SRR1543488 (GE) SRR1543627 (GE) SRR1543521 (GE) SRR1543522 (GE) SRR1543524 (GE) SRR1543502 (GE) SRR1543302 (GE)	Path CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat	Filter Image: Comparison of the compar
Rows: 50 SRR_ID SRR 1543488 SRR 1543627 SRR 1543624 SRR 1543624 SRR 1543624 SRR 1543751 SRR 1543751 SRR 1543753	Metadata Elements Role 7 Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543522 (GE) SRR 1543522 (GE) SRR 1543522 (GE) SRR 1543512 (GE) SRR 1543519 (GE) SRR 1543519 (GE)	Path CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat	Filter Fil
Rows: 50 SRR_ID SRR 1543488 SRR 1543627 SRR 1543627 SRR 1543624 SRR 1543524 SRR 1543624 SRR 1543624 SRR 1543751 SRR 1543733 SRR 1543767	Metadata Elements Role / Gene expression Gene expression		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543532 (GE) SRR 1543571 (GE) SRR 1543624 (GE) SRR 1543701 (GE) SRR 1543733 (GE) SRR 1543767 (GE)	Path CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat CLC_Dat	Filter Image: Comparison of the compar
Rows: 50 SRR_ID SRR 1543488 SRR 1543627 SRR 1543624 SRR 1543624 SRR 1543624 SRR 1543624 SRR 1543629 SRR 154367 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543621	Metadata Elements Role , Gene expression 2 Gene expression Gene expres		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543522 (GE) SRR 1543522 (GE) SRR 1543522 (GE) SRR 1543502 (GE) SRR 1543519 (GE) SRR 1543519 (GE) SRR 1543519 (GE) SRR 154362 RNA-56 SRR 1543627 RNA-56 SRR 1543621 RNA-56	Path CLC_Dat CLC_DAT CLC_DA	Filter Fil
Rows: 50 SRR_ID SRR 1543488 SRR 1543627 SRR 1543622 SRR 1543624 SRR 1543522 SRR 1543525 SRR 1543753 SRR 1543763 SRR 1543763 SRR 1543488 SRR 1543488	Metadata Elements Role , Gene expression Gene		Name SRR 1543488 (GE) SRR 1543627 (GE) SRR 1543522 (GE) SRR 1543522 (GE) SRR 1543522 (GE) SRR 1543512 (GE) SRR 1543702 (GE) SRR 1543703 (GE) SRR 1543703 (GE) SRR 1543488 RNA-Se SRR 1543627 RNA-Se	Path CLC_Dat CLC_DAT CLC_DA	Filter Fil

Figure 11: Finding all results having "Gene Expression" as Role using the metadata table.

- 3. In the first wizard window, all Gene Expression files (having the (GE) suffix) should be pre-selected.
- 4. Choose to Save the PCA plot.

Open the PCA plot. The samples are plotted according to their DGRP_Number. Choose in turn each of the different values from the side panel option: **Metadata | Symbol color**. It becomes evident that the principal variation between the samples (PC1) is due to their sex, while the second greatest source of variation (PC2) emphasizes the importance of using the same RNA-Seq preparation protocol throughout an experiment (figure 12). The variation due to the RNA-Seq

preparation protocol is an example of a batch effect. Batch effects can arise for many reasons, e.g. between different lanes of a sequencer, or samples prepared on different days. It is good practice to log as much metadata as possible about an experiment, so that all potential batch effects can be detected and corrected for in the statistics.

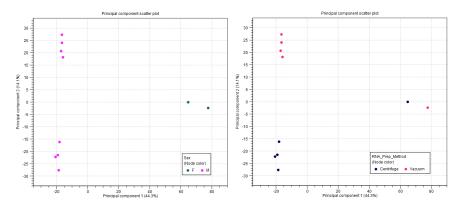


Figure 12: Sex of the flies, and RNA preparation protocols are the 2 main factors explaining the differences between individuals.

You can also investigate more remote components of the PCA by choosing to display the plot with another principal component as axis using the drop down menu in the side panel. The differences in the fly strains appear as PC3 and PC4. You can also see the PCA in 3D using the icon in the lower left corner (figure 13).

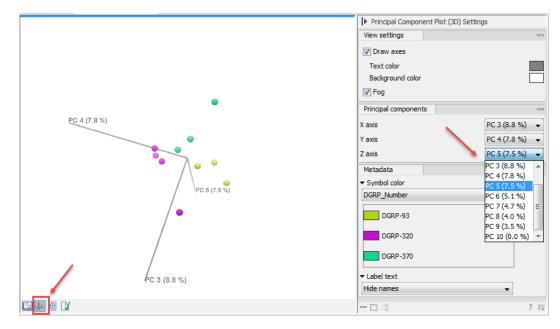


Figure 13: Choose another Principal Component as axis x or y (and z) for your PCA plot, and view the PCA plot in 3D.

Going back to the initial representation of the PCA (figure 12), it is evident that fly sex is so dominant that we would benefit from removing this large confounding effect for further analyses. In general, it is often a matter of judgement whether to keep samples and try to correct for the factor in the statistics, or to remove samples so that no correction is necessary. In this particular case, we can proceed by removing the female individuals without sacrificing much data. We will



thus perform the rest of the analysis on the 8 male samples. As for the other large confounding effect, the RNA preparation protocol, we will show how to correct for it in the next section.

Using a workflow to perform additional statistical analyses

To pursue our investigation of the data, we will use the workflow we imported earlier. Open the workflow by double clicking on its name in the Navigation Area (figure 14).

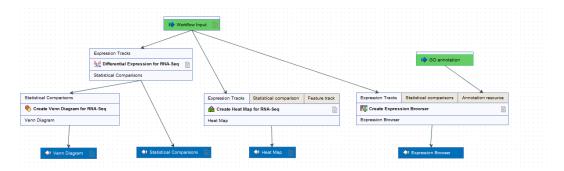


Figure 14: View of the RNA-Seq statistics workflow.

We can run the workflow from the Navigation Area, but in order to be able to use the functionalities of the metadata table to help select the relevant input files, it is recommended to install the workflow in your workbench.

- 1. In the lower right corner of the View Area, click on the button "Installation".
- 2. A pop up window opens containing the workflow information. Click on **Next**.
- 3. In the following window, choose to "Install the workflow on your local computer" before clicking **Finish**.

The workflow is now part of your toolbox in the "Workflow" folder. Before running it, we will remove the two female individuals from the dataset. This can be done easily from the metadata table "Samples".

- 1. In the upper table of the "Samples" view, click on the arrow next to "Filter" to open up the advanced filter box.
- 2. Select **Sex** in the first field, **=** in the second, and type **M** in the last.
- 3. Click on the button "Filter".
- 4. Select the 8 remaining rows and click on the button "Find Associated Data".
- 5. Now, only the Metadata Elements associated with male individuals remain in the lower table as well.
- 6. Select all elements whose role is "Gene expression", and right click on the selection.
- 7. In the menu that just opened in your View Area, go to **Toolbox | Workflows | RNA-Seq** statistics workflow (figure 15)



		Ľ	Sex 🗸 =	▼ M	🚹 🔀 🛛 Filter
R_ID	DGRP_Number	Sex	Environment	RNA_Prep_Method	Lane
R1543627	DGRP-93			Vacuum	
R1543733	DGRP-370	м		Vacuum	
	🖉 Set Up Table	Manage Data	Find Associated Data	Associate Data	Additional Filtering
	·				-
		T			
SRR_ID	Role /	Type	Name	Path	
SRR_ID SRR 1543627	Role z	i ype	Name SRR 1543627 (GE)	CLC Data / RNA Seg Drosophila /	references / RNA-Seg results / \$
-					
SRR 1543627	Gene expression Gene expression Gene expression	File +	SRR 1543627 (GE)	CLC_Data / RNA Seq Drosophila /	
SRR 1543627 SRR 1543821	Gene expression Gene expression	2 €	SRR 1543627 (GE) SRR 1543821 (GE)	CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila /	references / RNA-Seq results / { references / RNA-Seq results / {
SRR 1543627 SRR 1543821 SRR 1543532	Gene expression Gene expression Gene expression	File +	SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543532 (GE)	CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila /	references / RNA-Seq results / references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543532 SRR 1543624	Gene expression Gene expression Gene expression Gene expression 3	File → Edit → View →	SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543532 (GE) SRR 1543624 (GE) SRR 1543751 (GE)	CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila /	references / RNA-Seq results / references / RNA-Seq results / references / RNA-Seq results / references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543532 SRR 1543624 SRR 1543751 SRR 1543802 SRR 1543519	Gene expression Gene expression Gene expression Gene expression Gene expression	File Edit View Toolbox ↓	SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543532 (GE) SRR 154352 (GE) SRR 1543624 (GE) SRR 1543751 (GE) Ingenuity Pathway Analysis	CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543532 SRR 1543624 SRR 1543751 SRR 1543802 SRR 1543519 SRR 1543733	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression	File File File File File File File File	SRR 1543627 (GE) SRR 1543821 (GE) SRR 154352 (GE) SRR 154362 (GE) SRR 154362 (GE) SRR 154362 (GE) SRR 1543751 (GE) Ingenuity Pathway Analysis Classical Sequence Analysis	 (C.CData / RNA Seq Drosophila / 	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543532 SRR 1543524 SRR 1543524 SRR 1543519 SRR 1543519 SRR 1543513 SRR 1543527	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report	File Edit View Toolbox Show	SRR 1543627 (GE) SRR 1543621 (GE) SRR 1543522 (GE) SRR 1543624 (GE) SRR 1543624 (GE) A Ingenuity Pathway Analysis Classical Sequence Analysis Molecular Biology Tools	CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543532 SRR 1543624 SRR 1543624 SRR 1543802 SRR 1543802 SRR 1543519 SRR 1543733 SRR 1543627 SRR 1543821	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report	File File Edit View Toolbox Show	SRR 1543627 (GE) SRR 1543621 (GE) SRR 1543522 (GE) SRR 1543624 (GE) SRR 1543624 (GE) A Ingenuity Pathway Analysis Classical Sequence Analysis Molecular Biology Tools	QLC_Data RNA Seq Drosophila /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543621 SRR 1543624 SRR 1543624 SRR 1543751 SRR 1543751 SRR 1543753 SRR 1543733 SRR 1543627 SRR 1543821 SRR 1543822	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report Report	File File Edit View Toolbox Show	SRR 1543627 (GE) SRR 1543621 (GE) SRR 1543622 (GE) SRR 1543624 (GE) SRR 1543624 (GE) SRR 1543624 (GE) Classical Sequence Analysis Molecular Biology Tools BLAST	CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543822 SRR 1543824 SRR 1543824 SRR 1543827 SRR 1543827 SRR 1543827 SRR 1543821 SRR 1543822 SRR 1543822	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report Report Report	File File Edit View Show	SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543822 (GE) SRR 154352 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 154352 (GE) SRR 154352 (GE) SRR 154352 (GE) SRR 154362 (GE) SRR 15	CLC_Data / RNA Seq Drosophila / CLC_Data / RNA Seq Drosophila /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543822 SRR 1543822 SRR 1543824 SRR 1543624 SRR 1543629 SRR 1543829 SRR 1543821 SRR 1543821 SRR 1543824 SRR 1543524	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report Report Report Report Report	File Edit View Toolbox Show	SRR 1543627 (GE) SRR 1543821 (GE) SRR 1543822 (GE) SRR 154352 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 1543524 (GE) SRR 154352 (GE) SRR 154352 (GE) SRR 154352 (GE) SRR 154362 (GE) SRR 15	CLC Data / RNA Seq Drosophia / CLC Data /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543822 SRR 1543824 SRR 1543824 SRR 1543827 SRR 1543827 SRR 1543827 SRR 1543821 SRR 1543822 SRR 1543822	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report Report Report	File File Edit View Show	SRR: 1543627 (GE) SRR: 1543821 (GE) SRR: 1543821 (GE) SRR: 1543524 (GE) SRR: 1543524 (GE) SRR: 15435751 (GE) Classical Sequence Analysis Classical Sequence Analysis Molecular Biology Tools BLAST NGS Core Tools Track Tools	CLC Data / RNA Seq Drosophia / CLC Data /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543522 SRR 1543522 SRR 1543524 SRR 1543624 SRR 1543625 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543624 SRR 1543524 SRR 1543524 SRR 1543624	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report Report Report Report Benort	File Edit View Show Show	SRR: 1543627 (GE) SRR: 1543821 (GE) SRR: 1543821 (GE) SRR: 1543524 (GE) SRR: 1543524 (GE) SRR: 1543571 (GE) Classical Sequence Analysis Molecular Biology Tools BLAST NGS Core Tools Track Tools Resequencing Analysis	C.C. Data / RNA Seq Drosophia / C.C. Data / RNA Seq Drosophia /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543522 SRR 1543522 SRR 1543524 SRR 1543624 SRR 1543625 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543624 SRR 1543524 SRR 1543524 SRR 1543624	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report Report Report Report Benort	File File Edit View View Show Groups and Gro	SRR.1543627 (GE) SRR.1543621 (GE) SRR.1543521 (GE) SRR.1543521 (GE) SRR.1543524 (GE) SRR.1543751 (GE) Classical Sequence Analysis Classical Sequence Analysis BLAST NGS Core Tools Track Tools Resequencing Analysis Transcriptomics Analysis	CLC Data / RNA Seq Drosophia / CLC Data /	references / RNA-Seq results / references / RNA-Seq results /
SRR 1543627 SRR 1543821 SRR 1543522 SRR 1543522 SRR 1543524 SRR 1543624 SRR 1543625 SRR 1543627 SRR 1543627 SRR 1543627 SRR 1543624 SRR 1543524 SRR 1543524 SRR 1543624	Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Gene expression Report Report Report Report Report Benort	File Edit View Show Show	SRR.1543627 (GE) SRR.1543621 (GE) SRR.1543521 (GE) SRR.1543521 (GE) SRR.1543524 (GE) SRR.1543751 (GE) Classical Sequence Analysis Classical Sequence Analysis BLAST NGS Core Tools Track Tools Resequencing Analysis Transcriptomics Analysis	C.C. Data / RNA Seq Drosophia / C.C. Data / RNA Seq Drosophia /	references / RNA-Seq results / references / RNA-Seq results /

Figure 15: Start the workflow from the metadata table with only a subset of your previous results.

- 8. A wizard window opens for the workflow. If you are asked, choose where to execute the workflow and click **Next**.
- 9. Click **Next** again to accept the selected tracks. Note that if you start the workflow from the Toolbox, the tracks selected in the metadata table will also be automatically used as input, but not if you start the workflow from the Navigation Area.
- 10. In the "Select GO annotation" dialog, select the file "gene_association.fb" you imported earlier. Click **Next**.
- 11. In the "Differential Expression for RNA-Seq" window, set the parameters as shown in figure 16. We want to find genes that are differentially expressed between the different strains of flies (DGRP_Number) while controlling for the RNA preparation protocol (RNA_Prep_Method), that we now know introduces a bias in the data. We want to produce a comparison between each pair of fly strains, so we select the 'All group pairs' comparison. When finished, click **Next**.
- 12. In the "Result handling" window, choose a location to save the results (for example a new subfolder you can call "statistical analyses"), and click **Finish**.

The workflow will now execute.

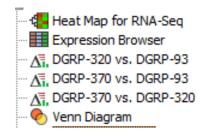
Results visualization and interpretation

When the workflow is completed, several new objects are created (figure 17)



Change where he was	Differential Expression for RNA-Seq		
. Choose where to run	Configurable Parameters		
Select input for Workflow Input	Metadata table	5 Samples	j,
51	Test differential expression due to	DGRP_Number	4
Differential Expression	While controlling for	RNA_Prep_Method	4
11 /	Comparisons	All group pairs	•
	 Locked Settings 		





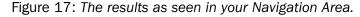




Figure 18: An example of Venn Diagram. Note that the values you obtain in this diagram could differ slightly from the ones seen in this figure.

Venn diagram Open the Venn diagram (figure 18).

The Venn diagram shows how many transcripts were detected to be differentially expressed in each comparison.

Select one of the diagram's intersections, where genes are differentially expressed in two comparisons. With the selection maintained, go to the table view of the Venn diagram. The same genes will still be selected in the table. It can be hard to find them, however you can create a



new selection only containing the selected genes by clicking the "Create from selection" button. This will open a new view containing only the genes of the section that was selected in the Venn diagram.

Statistical comparison Now we will show you how to track the genes of interest throughout your statistical results.

1. Make a new Track List containing all 3 statistics results, as well as the sequence, gene and mRNA reference tracks: go to **File | New | Track list** and select the statistical comparisons and the reference as indicated in figure 19. Click **Finish**.

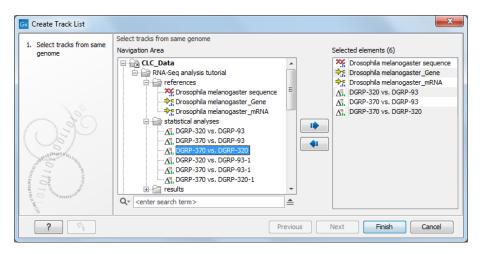


Figure 19: Creating a list of tracks with some references and the statistical comparisons.

- 2. Open a table view of one of the statistical comparison by double-clicking on the name of the track.
- 3. Now re-open the Venn diagram, and select one of the "shared by two" sections.
- 4. Go to the table view of the Venn diagram and click "Select Genes in Other views". Because diagram and table are linked, the genes present in the central section of the Venn diagram are automatically selected in the table view, and by clicking on "Select genes in other view", they are now also selected in the table view of the statistical comparison track.
- 5. Going back to the statistical comparison we opened from the track list, switch to the volcano plot view (figure 20).

We have demonstrated that a selection you make in the table will also be selected in the volcano plot, but note that it also works the other way round, so a gene selected in the volcano plot will also be selected in the table. The fact that table and plot are linked allows us to highlight the genes in the plots by filtering them in the table. Also, if you select only one gene in the volcano plot, the track list should jump to the gene (use the **Zoom to Selection** () button for optimal rendering of the track list).



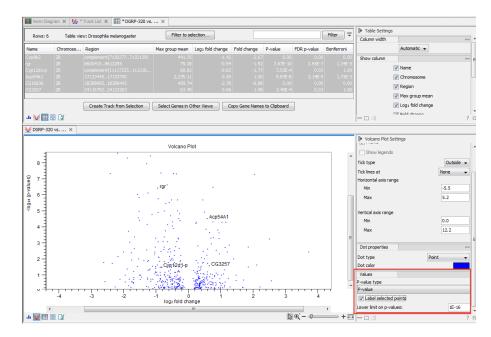


Figure 20: The genes selected in the Venn diagram are now highlighted in red in the volcano plot. To enhance visibility, check the "Label selected points" in the side panel settings to the right of the volcano plot.

Expression Browser The Expression Browser makes it possible to inspect gene and transcript expression level counts and statistics for many samples at the same time. In addition, the GO biological process column will list for each gene or transcript one or several biological processes.

Again, it is possible to highlight in the Expression Browser genes that were selected as being differentially expressed between the different strains in the Venn diagram.

- 1. Open the Expression Browser (III) from the Navigation Area.
- 2. Now re-open the Venn diagram where one of the "shared by two" sections should still be selected.
- 3. In the table view of the Venn diagram, click **Select Genes in Other Views**.
- 4. The selected genes are now also selected in the Expression Browser table, but they can be hard to find. Click on the "Filter to selection" button to only see the selected rows.

The Expression Browser now only displays genes with the names of the previously selected genes (figure 21). This means that if two genes have the same name, both will be in that table, even if only one was in your original selection. Check that the number of rows displayed in the Expression Browser is identical to the sum of the genes selected in the Venn diagram (in the table below we selected all 3 sections of genes differentially expressed between strains).

You can then click on the process name to open the corresponding page on the Consortium for Gene Ontology webpage. It is also possible to access additional online information by clicking on the PMID, RefSeq, HGNC or UniProt accession number when available.



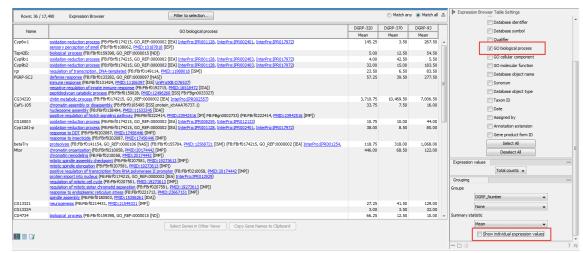


Figure 21: Expression browser table displaying the genes found to be differentially expressed between the different strains. The table is set up to only show the name of the gene, the GO biological processes associated to this particular gene, and the mean of total counts for each strain.

Heat map Finally, open the Heat Map for RNA-Seq from the Navigation Area. Note that the HeatMap was obtained without controlling for the preparation method, as can be seen in the structure of the workflow. The hierarchical clustering algorithm is pre-configured in the workflow to select the 25 "most interesting" genes in the heat map based on the coefficient of variation (relative standard deviation). The samples are also clustered on the horizontal axis, in unsupervised clustering.

In the right hand side panel, choose each metadata element in turn (see figure 22). You will notice that the 25 most interesting genes are differentially expressed due to the RNA Preparation method. This is what we should expect, because we know from the PCA plot that variation due to RNA preparation is greater than between the different strains.

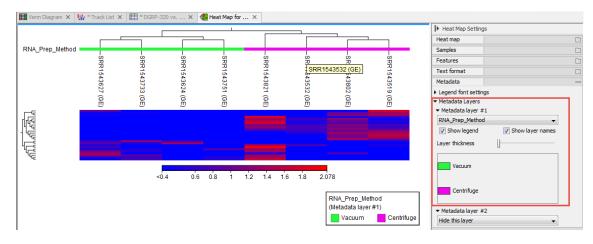


Figure 22: Choose among the different metadata layers.

One of the options of the Heat Map for RNA-Seq tool is to "Filter by statistics": using any of our statistical results with this option would lead to a heat map where samples clustered by fly strain.

- 1. Go back to the "samples" metadata table, and make sure that the "Gene expression" files for the 8 male individuals are selected.
- 2. From the toolbox, or directly from a right click on the selected elements in the metadata table, choose to run the "Create Heat Map for RNA-Seq" tool (figure 23). This opens the tool's wizard.

R_ID	DGRP_Number	Sex /	Environment	RNA_Prep_Method	Lane
1543488	DGRP-93	F	2	Centrifuge	4
1543767	DGRP-370	F	3	Vacuum	2
1543821					
1543532 1543624					
1543751					
1543802					
1543519					
1	Set Up Table 🧪 Mana	Genome Finish	ons	Shift+T	Additional Filtering
Rows: 40	Metadata Elements	🛅 🛛 GeneMark Gen	e Finding	•	Filter 🗮
SRR ID	Role / T	🚘 Microbial Gene	omics Module	•	
SRR 1543627	Gene expression	🗟 Classical Seque	ence Analysis	•	
SRR 1543627 SRR 1543821	Gene expression	Molecular Biol	oav Tools	•	Â
SRR 1543532		BLAST	- , ,		
SRR 1543624		_			E
SRR 1543751		NGS Core Too	ls	•	
SRR 1543802		🚡 Track Tools		•	
SRR 1543519 SRR 1543733		🙀 Resequencing	Analysis	•	
SRR 1543627	Table filters	RNA-Seq Anal	ysis	RNA-Seq Anal	ysis
SRR 1543821	Remove Association(s)	Microarray and	d Small RNA Analysis	Create Combin	ned RNA-Seg Report
SRR 1543532	File	Epigenomics A		PCA for RNA-	
SRR 1543624			· · · · · · · · · · · · · · · · · · ·		
SRR 1543751 SRR 1543802	Edit	De Novo Sequ	encing		oression for RNA-Seq
ISRR 1343002	View	🖬 Workflows			ap for RNA-Seq
	Toolbox	Legacy Tools		Crosto Heat	Map for RNA-Seq

Figure 23: Run the "Create Heat Map for RNA-Seq" tool on the 8 samples we chose to keep in the statistical analyses.

- 3. In the first window of the wizard, the elements are automatically pre-selected. Click Next.
- 4. In the next window, leave the parameters as they are selected by default (see figure 24) and click **Next**.
- In the "Set filtering" window, choose the option "Filter by statistics". Then click on the Navigation icon next to the "Statistical comparison" field and select one of the comparison generated by the workflow (figure 25). Click OK. Leave all other values as default and click Next.
- 6. Choose to **Open** the result and click on **Finish** to start the tool.

The newest heat map opens as in figure 26, and it is now apparent that the samples are clustered by strains. There are also more differentially expressed genes included in the heat map, and these are different from the ones output by the tool when we did not correct for statistics.



Gx Create Heat Map for R	NA-Seq
 Choose where to run Select at least two expression tracks Set parameters 	Set parameters Distance Euclidean distance Manhattan distance 1 - Pearson correlation
A Construction of the second s	Clusters Single linkage Average linkage Complete linkage
?	Previous Next Finish Cancel

Figure 24: Choose "Eucledian distances" and "Complete linkage" to run the "Create Heat Map for RNA-Seq" tool.

A. Set htterng Freed number of features Z5 Minimum counts in at least one sample 10 Fiter by statistics Statistical comparison AL_DCRP-370 vs. DGRP-320 Minimum absolute fold change 1.5	Create Heat Map for I Choose where to run Select at least two expression tracks Set parameters	Set filtering Filter settings Filter settings Filter by statistic	s ,	
P-value Produce Produce	4. Set filtering	Minimum counts in at least one s Filter by statistics Statistical comparison	Sample 10	6
		P-value Prop -value Bonferroni Threshold Specify features Feature track	0.05 0.05	AT, DGRP-370 vs. DGRP-320

Figure 25: Add a statistical comparison to filter by statistics.

Using the RNA-Seq analysis tools, we have been able not only to to detect batch effects, but also to control for these during subsequent analyses, allowing us to find strain-specific differences that were much smaller than differences due to gender or sample preparation.



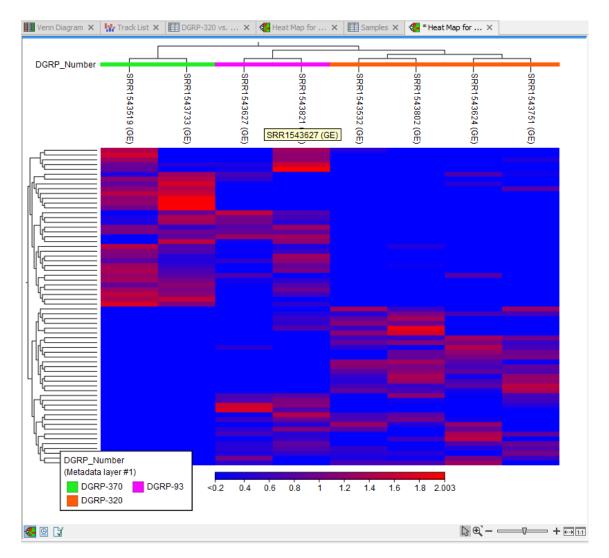


Figure 26: The heat map after correcting for the sample preparation method bias.



Bibliography

[Lin et al., 2016] Lin, Y., Golovnina, K., Chen, Z.-X., Lee, H. N., Negron, Y. L. S., Sultana, H., Oliver, B., and Harbison, S. T. (2016). Comparison of normalization and differential expression analyses using rna-seq data from 726 individual drosophila melanogaster. *BMC genomics*, 17(1):1.