

Análise de Expressão Diferencial

Cálculo da expressão diferencial e resultados

Dr. Pablo Rodrigo Sanches

Departamento de Genética – FMRP/USP

psanches@usp.br

Roteiro de análise

1. DESeq2 (normalização e cálculo da expressão diferencial)
 - a) tratamento vs. controle
2. Obter os resultados:
 - a) PCA
 - b) Distância entre amostras
 - c) Histograma de p valor
 - d) Genes significativamente DE
 - e) Tabelas de genes significativamente DE

Renomear arquivos Gene counts no Galaxy (saída do software StringTie)

The screenshot shows the Galaxy web interface at `usegalaxy.org/datasets/edit`. The main area is titled "Edit dataset attributes" and contains several sections:

- Name:** A text input field containing "0h II StringTie on data 106 and data 109: Gene counts". A black text box above it displays the target name: "3h II RNA STAR...; 3h III RNA STAR...; 12h II RNA STAR...; ...".
- Info:** A large empty text area for additional information.
- Annotation:** A large empty text area for dataset annotations.
- Database/Build:** A dropdown menu currently set to "Additional Species Are Below".
- Number of comment lines:** A text input field.

On the right side, the "History" panel shows a list of datasets. The dataset "126: StringTie on data 106 and data 109: Gene counts" is highlighted in yellow, matching the dataset being edited. A yellow arrow points to the "Save" button in the top right of the edit area.

DESeq2 – Via Galaxy

The screenshot displays the Galaxy web interface for the DESeq2 tool. The main configuration area is titled "DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.11.40.7+galaxy2)".

Factor level 1:

- Experiment name: **Nome do experimento** (input field: effects_drug_srt)
- Treatment name: **Nome do tratamento** (input field: 3h-SRT)
- Counts file(s): A list of files for treatment replicates, including "209: StringTie on data 3 and data 164: Gene counts", "208: StringTie on data 3 and data 164: Assembled transcripts", "207: StringTie on data 3 and data 160: Transcript counts", "206: StringTie on data 3 and data 160: Gene counts", "205: StringTie on data 3 and data 160: Assembled transcripts", "204: StringTie on data 3 and data 156: Transcript counts", and "203: StringTie on data 3 and data 156: Gene counts". A yellow callout points to this list: "Selecione os arquivos de contagem das réplicas do tratamento".

Factor level 2:

- Treatment name: **Nome do tratamento** (input field: 3h-SAB)
- Counts file(s): A list of files for treatment replicates, including "227: StringTie on data 3 and data 168: Gene counts", "226: StringTie on data 3 and data 168: Assembled transcripts", "225: StringTie on data 3 and data 184: Transcript counts", "224: StringTie on data 3 and data 184: Gene counts", "223: StringTie on data 3 and data 184: Assembled transcripts", "222: StringTie on data 3 and data 180: Transcript counts", and "221: StringTie on data 3 and data 180: Gene counts". A yellow callout points to this list: "Selecione os arquivos de contagem das réplicas do tratamento".

The right sidebar shows a "History" panel with a search bar and a list of previous jobs, including "242: DESeq2 plots on data 218, data 215, and others" and "241: DESeq2 result file on data 218, data 215, and others".

DESeq2 – Via Galaxy (cont.)

Galaxy

usegalaxy.org/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fbioc%2Fdeseq2%2Fdeseq2%2F2.11.40.7%2Bgalaxy2&version=2.11.40.7%20galaxy2

Galaxy

Workflow Visualize Shared Data Help User

Using 90%

! The US Galaxy Team is exploring the possibility of hosting user-focused US-centric Galaxy meetings for the presentation of new research use Galaxy, building relationships, and facilitating communication between the user community and Galaxy developers. Please help us out by completing a **very short survey** so that we can gauge interest.

Tools

deseq

Upload Data

Show Sections

DESeq2 Determines differentially expressed features from count tables

Annotate DESeq2/DEXSeq output tables Append annotation from GTF to differential expression tool outputs

WORKFLOWS

All workflows

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.11.40.7+galaxy2)

Files have header?

Yes

If this option is set to Yes, the first row of the input file is assumed to be a header row.

Choice of Input data

Count data (e.g. from HTSeq-count, featureCounts or StringTie)

Advanced options

Output options

Job Resource Parameters

Use default job resource parameters

Additional Options

Email notification

No

Send an email notification when the job completes.

Run Tool

Help

What it does

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution

Inputs

Count Files

DESeq2 takes count tables generated from **featureCounts**, **HTSeq-count** or **StringTie** as input. Count tables must be generated for each sample individually. One header row is assumed, but files with no header (e.g. from HTSeq) can be input with the **Files have header?** option set to No. DESeq2 is capable of handling multiple factors that affect your experiment. The first factor you input is considered as the primary factor that affects gene expressions. Optionally, you can input one or more secondary factors that might influence your experiment. But the final output will be changes in genes due to primary factor in presence of secondary

History

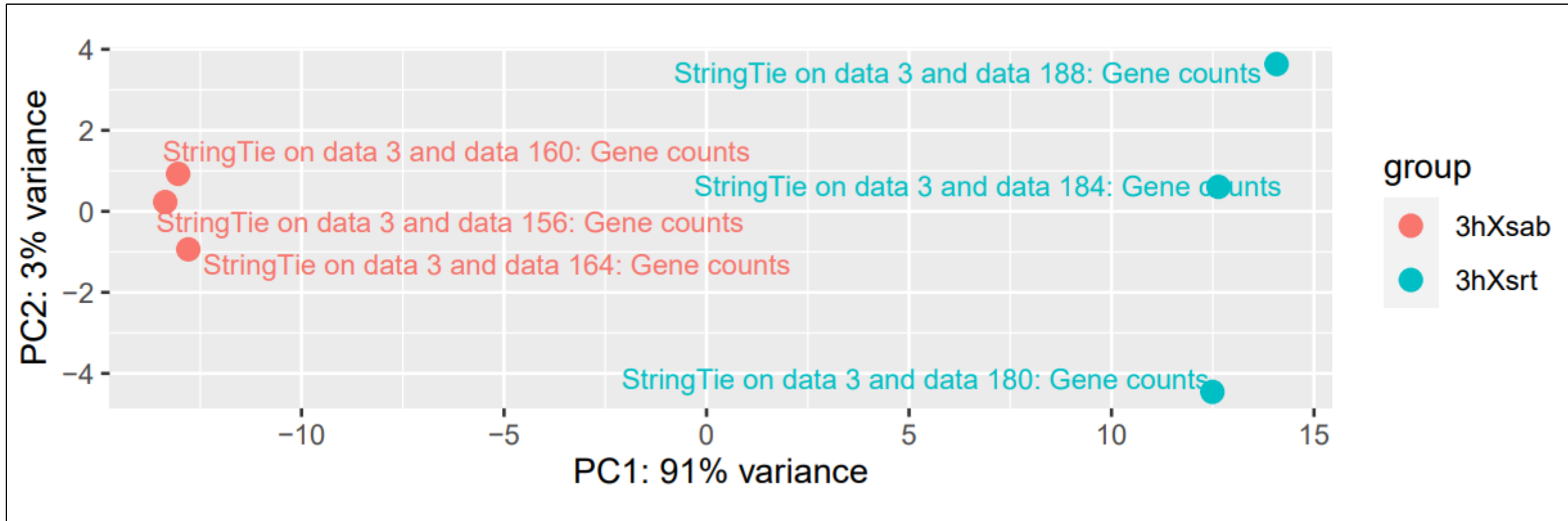
search datasets

SRT-Aluno

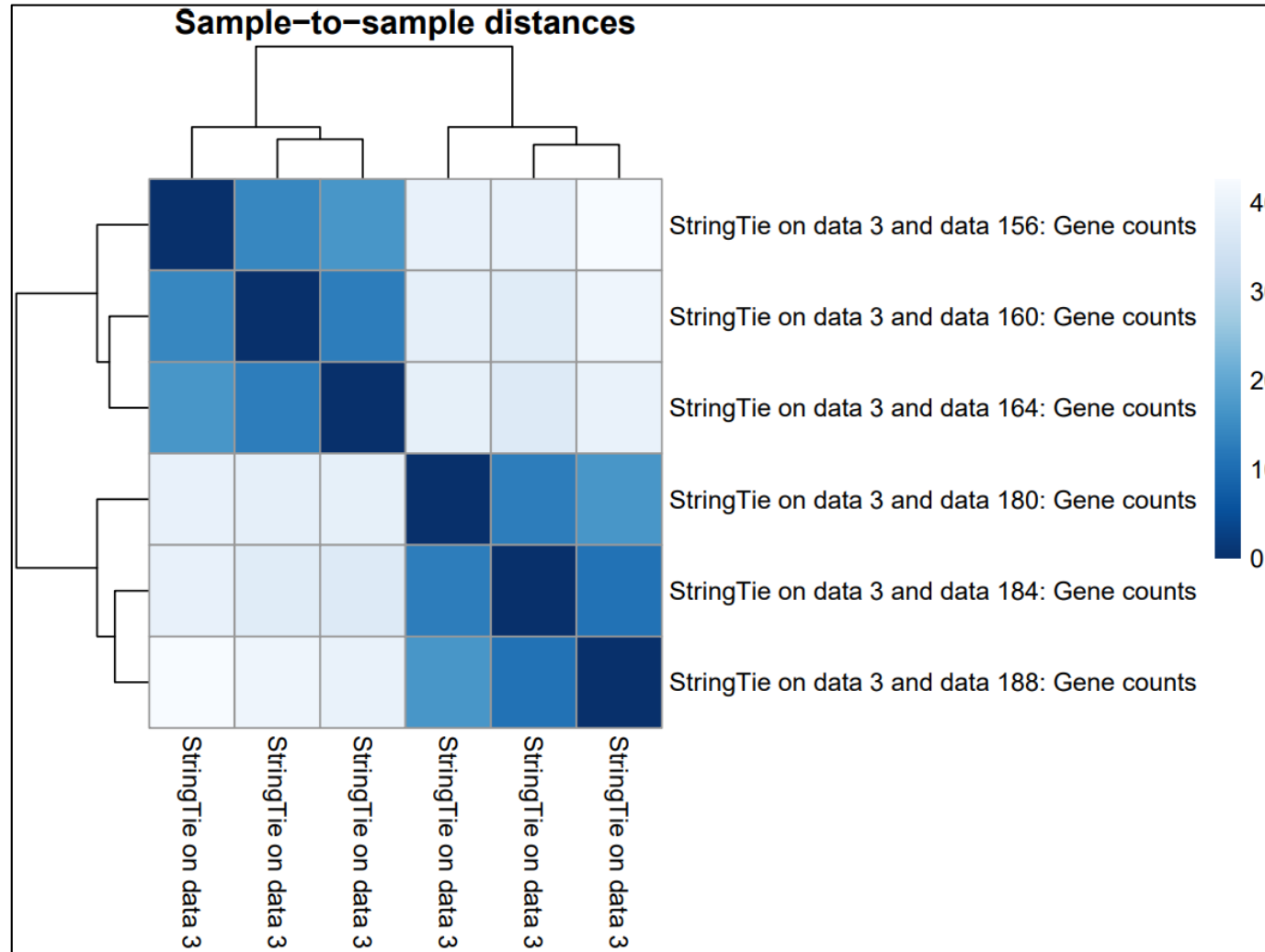
242 GB 186 55

- 242 : DESeq2 plots on data 218, data 215, and others
- 241 : DESeq2 result file on data 218, data 215, and others
- 240 : DESeq2 plots on data 209, data 206, and others
- 239 : DESeq2 result file on data 209, data 206, and others
- 237 : StringTie on data 3 and data 206: Transcript counts
- 236 : StringTie on data 3 and data 206: Gene counts
- 235 : StringTie on data 3 and data 206: Assembled transcripts
- 234 : StringTie on data 3 and data 196: Transcript counts
- 233 : StringTie on data 3 and data 196: Gene counts
- 232 : StringTie on data 3 and data 196: Assembled transcripts

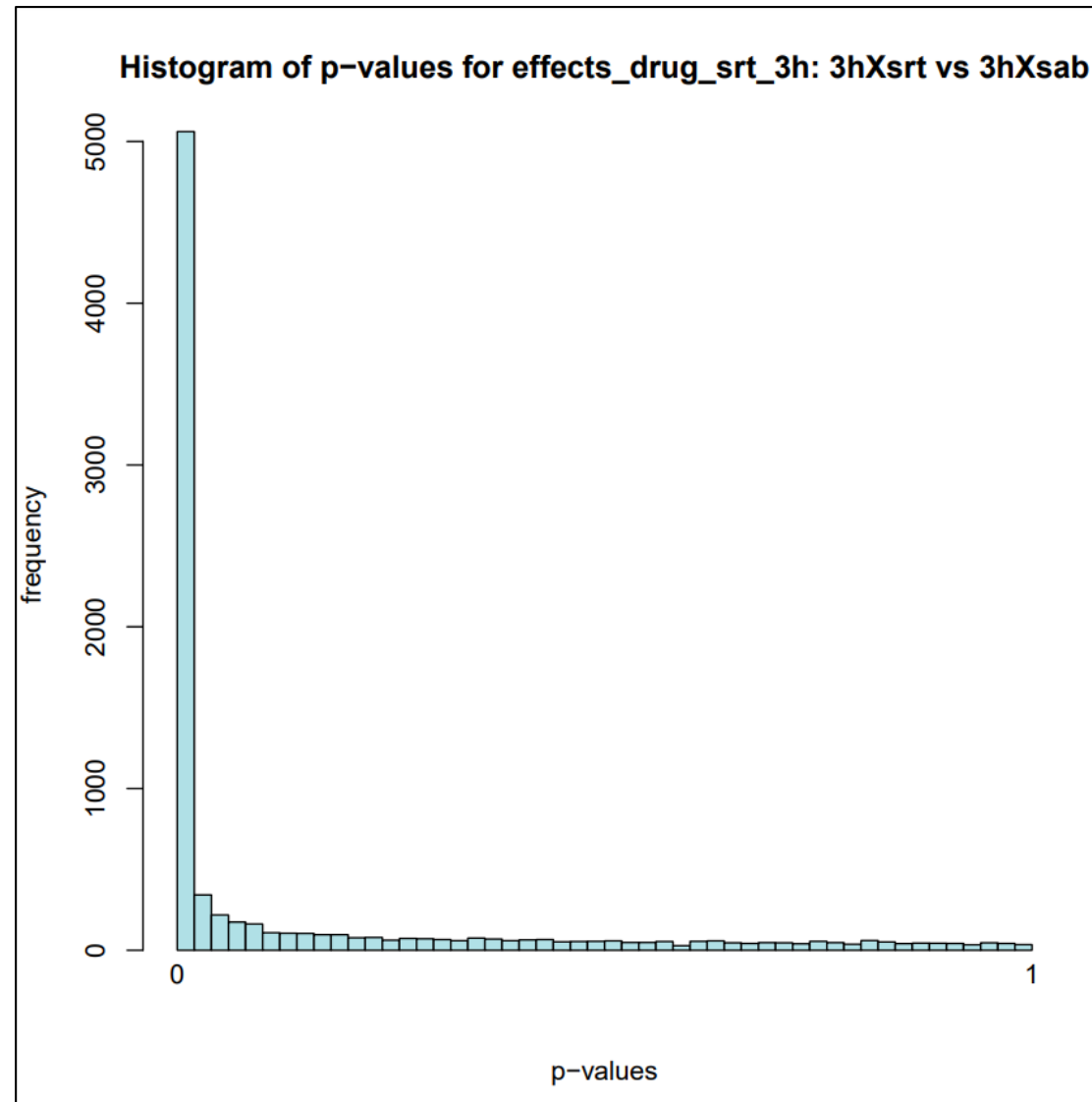
Exemplo de Resultado DESeq2 (PCA)



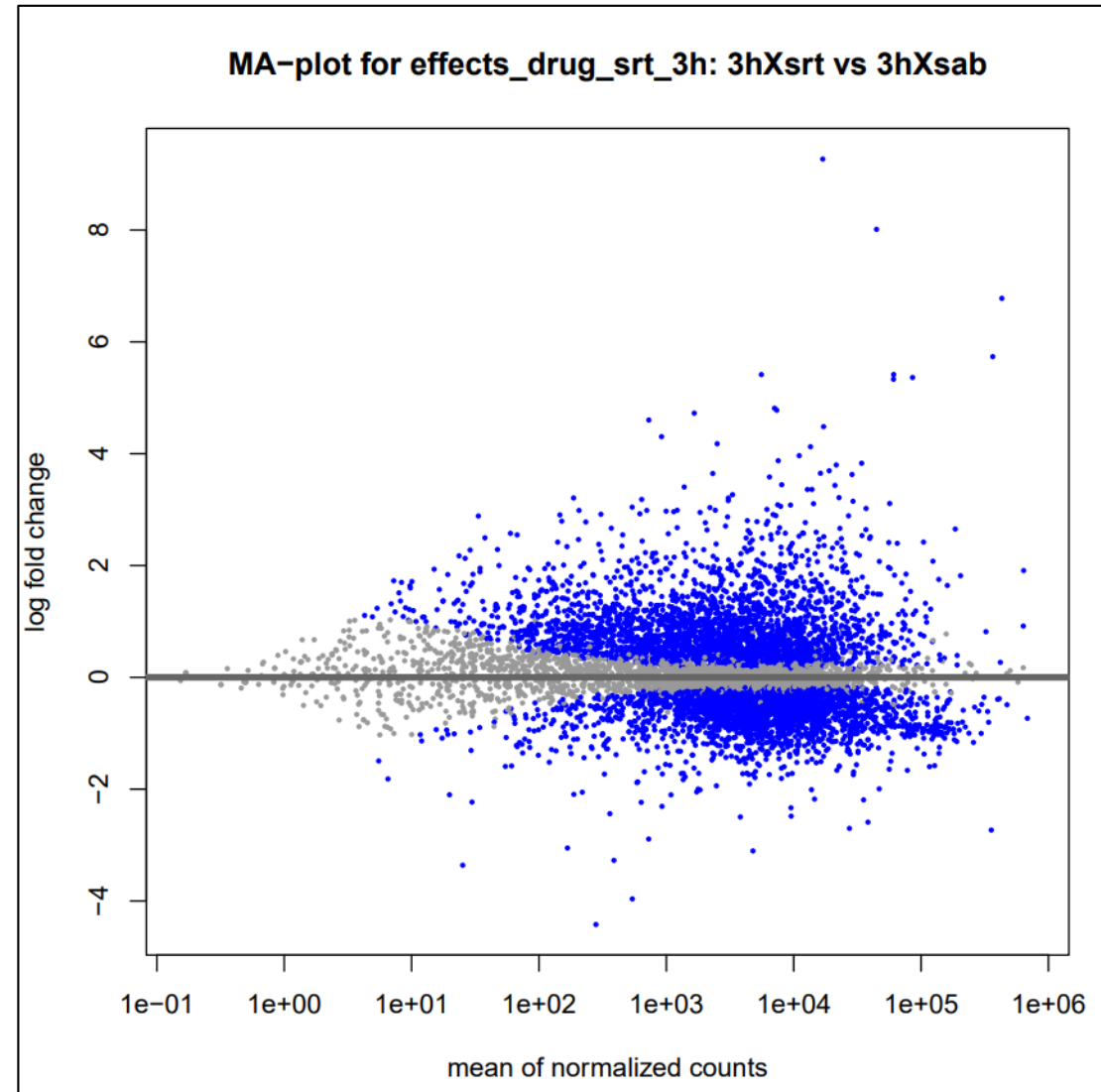
Exemplo de Resultado DESeq2 (Distância entre amostras)



Exemplo de Resultado DESeq2 (Histograma de p valor)



Exemplo de Resultado DESeq2 (DEG)



Exemplo de Resultado DESeq2 (Tabelas DEG)

The screenshot shows the Galaxy web interface. The main panel displays a table of differentially expressed genes (DEG) results from a DESeq2 analysis. The table has the following columns: GeneID, Base mean, log2(FC), StdErr, Wald-Stats, P-value, and P-adj. The table lists 30 genes, with the first few rows showing GeneID, Base mean, log2(FC), StdErr, Wald-Stats, P-value, and P-adj.

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
TERG_04937	60894.141693394	5.41644203570673	0.110759302035499	48.9028184194476	0	0
TERG_04952	430607.920123861	6.77874544684291	0.14063961661639	48.1994022021029	0	0
TERG_07830	5571.12234052747	5.41472389206682	0.13191940262864	41.0456974802223	0	0
TERG_08041	85730.1696860086	5.36264676418039	0.104038869705649	51.5446465282888	0	0
TERG_08751	365274.12353951	5.73514717500478	0.132353851585295	43.3319250351305	0	0
TERG_01782	44707.7653042112	8.0121275174711	0.223592974143738	35.8335390553035	3.31911302080352e-261	4.75407621679757e-278
TERG_01543	60686.2721653453	5.3320312578694	0.150476948503839	35.4342064408181	5.07914873177471e-275	6.23574345726741e-272
TERG_04003	6890.85630730244	2.91161051618903	0.0848161525935421	34.3284908258232	2.9496868784941e-258	3.16870112922229e-255
TERG_01113	26952.2952467852	2.88914318805371	0.0864529654205205	33.4186707650856	7.34316843034225e-245	7.01190994337348e-242
TERG_00010	1655.6099033574	4.72543076219962	0.149916619097384	31.5203930734993	4.56551098817562e-218	3.92360014323813e-215
TERG_04950	6450.53043574702	3.58323535061473	0.117273236368155	30.5545874027549	4.91437169416522e-205	3.83946457633236e-202
TERG_08195	19002.0868009922	3.69518805983612	0.122213620976518	30.2354846400151	8.09656188130387e-201	5.79848773399379e-198
TERG_06347	17160.6355489803	4.48379855177771	0.152912977632212	29.3225507815445	5.35099487377477e-189	3.53741922655541e-186
TERG_07465	7556.96227185147	3.87391351808623	0.136118647111639	28.4598297168574	3.6828134937616e-178	2.26072136895623e-175
TERG_06779	34133.8522493733	3.8291794740426	0.135431893876213	28.27388363437401	7.25023367096762e-176	4.15390054455305e-173
TERG_03829	7069.58107920694	4.81372839755162	0.17065975091292	28.2220905721388	3.13316967442757e-175	1.6829037615461e-172
TERG_01460	21084.3356667712	3.43362326481343	0.126808029197375	27.0773332457445	1.82096121969719e-161	9.20549454239861e-159
TERG_01572	7243.0420514299	2.8916571323994	0.107088226299974	27.0025681842869	1.37875437280344e-160	6.58278615548487e-158
TERG_07556	3066.9462190158	3.20722828371639	0.121092416030572	26.4857898524932	1.41305224598829e-154	6.39145842211756e-152
TERG_11731	56706.247779122	3.10988394777939	0.118994821083347	26.1346159393033	1.47429213021608e-150	6.33503328354711e-148
TERG_04953	15011.3948196924	2.59515000638616	0.100808712322913	25.7433107376008	3.83025068842721e-146	1.56748449601636e-143
TERG_08283	916.681787157463	4.30428816473097	0.169980076370613	25.3223098649879	1.81444383845478e-141	7.08787743076382e-139
TERG_06701	12816.911199228	3.3600427191289	0.133597144934813	25.1505578264296	1.39347733861163e-139	5.20675836870798e-137
TERG_12230	2199.67961004117	3.03622773959201	0.121071214467827	25.0780315778434	8.63761627006312e-139	3.09298642603843e-136
TERG_07094	6248.88829132656	2.79152631652153	0.113025766190292	24.6981410576917	1.11981623005762e-134	3.84948027244608e-132
TERG_04007	12330.1820847007	2.76703048182618	0.113018350056601	24.4016631000403	1.8303478511351e-132	6.87075142840583e-130

The right sidebar shows a history of datasets. The entry "239 : DESeq2 result file on data 209, data 206, and others" is highlighted in red and pointed to by an arrow.

Sobre as colunas da tabela de resultados DESeq2

Coluna	Descrição
1	GeneID = Identificador do gene
2	Base mean = média das contagens normalizadas tomadas em todas as amostras
3	$\log_2(\text{FC})$ = mudança \log_2 vezes entre os grupos. Por exemplo, valor 2 significa que a expressão aumentou 4 vezes
4	stdErr = erro padrão da estimativa $\log_2\text{FoldChange}$
5	Wald-Stats = estatística de Wald
6	P-value = valor p do teste de Wald
7	P-adj = valor p ajustado de Benjamini-Hochberg (Taxa de Falsas Descobertas - FDR)

Download dos resultados

The screenshot shows the Galaxy web interface. The main content is a table of DESeq2 results. The table has the following columns: GeneID, Base mean, log2(FC), StdErr, Wald-Stats, P-value, and P-adj. The table contains 30 rows of data. On the right side, there is a 'History' panel showing a list of datasets. The dataset '239 : DESeq2 result file on data 209, data 206, and others' is highlighted with a red box, and a black arrow points to the download icon (a square with a downward arrow) in the bottom right corner of its entry. The download icon is also visible in the bottom right corner of the table area.

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
TERG_04937	60894.141693394	5.41644203570673	0.110759302035499	48.9028184194476	0	0
TERG_04952	430607.920123861	6.77874544684291	0.14063961661639	48.19940220201209	0	0
TERG_07830	5571.12234052747	5.41472389206682	0.13191940262864	41.0456974002223	0	0
TERG_08041	85730.1696860086	5.36264676418039	0.104038869705649	51.5446465282888	0	0
TERG_08751	365274.12353951	5.73514717500478	0.132353851585295	43.3319250351305	0	0
TERG_01782	44707.7653042112	8.012127514711	0.223592974143738	35.8335390553035	3.31911302080352e-281	4.75407621679757e-278
TERG_01543	60686.2721653453	5.3320312578694	0.150476948503839	35.4342064406181	5.07914873177471e-275	6.23574345726741e-272
TERG_04003	6890.85630730244	2.91161051618903	0.0848161525935421	34.3284908258232	2.9496868784941e-258	3.16870112922229e-255
TERG_01113	26952.2952467852	2.88914318805371	0.0864529654205205	33.4186707650856	7.34316843034225e-245	7.01190994337348e-242
TERG_00010	1655.6099033574	4.72543076219962	0.149916619097384	31.5203930734993	4.56551098617562e-218	3.92360014323813e-215
TERG_04950	6450.53043574702	3.58323535061473	0.117273236368155	30.5545874027549	4.91437169416522e-205	3.83946457633236e-202
TERG_08195	19002.0868009922	3.69518805983612	0.122213620976518	30.2354846400151	8.09656188130387e-201	5.79848773399379e-198
TERG_06347	17160.6355489803	4.48379855177771	0.152912977632212	29.3225507815445	5.35099487377477e-189	3.53741922655541e-186
TERG_07465	7556.96227185147	3.87391351808623	0.136118647111639	28.4598297168574	3.6828134937616e-178	2.26072136895623e-175
TERG_06779	34133.8522493733	3.8291794740426	0.135431893876213	28.2738383437401	7.25023367096762e-176	4.15390054455305e-173
TERG_03829	7069.58107920694	4.81372839755162	0.170565975091292	28.2220905721388	3.13316967474257e-175	1.6829037615461e-172
TERG_01460	21084.3356667712	3.43362326481343	0.126808029197375	27.0773332457445	1.82096121969719e-161	9.20549454239861e-159
TERG_01572	7243.0420514299	2.8916571323994	0.107088226299974	27.0025681842869	1.37875437280344e-160	6.58278615548487e-158
TERG_07556	3066.9462190158	3.20722828371639	0.121092416030572	26.4857898524932	1.41305224598829e-154	6.39145842211756e-152
TERG_11731	56706.2477779122	3.10988394777939	0.118994821083347	26.1346159393033	1.47429213021808e-150	6.33503328354711e-148
TERG_04953	15011.3948196924	2.59515000638816	0.100808712322913	25.7433107376008	3.83025068842721e-146	1.56748449601636e-143
TERG_08283	916.681787157463	4.30428816473097	0.169980076370613	25.3223098649879	1.81444383845478e-141	7.08787743076382e-139
TERG_06701	12816.911199228	3.3600427191289	0.133597144934813	25.1505578264296	1.39347733861163e-139	5.20675836870798e-137
TERG_12230	2199.67961004117	3.03622773959201	0.121071214467827	25.0780315778434	8.63761627006312e-139	3.09298642603843e-136
TERG_07094	6248.88829132656	2.791526316252153	0.113025766190292	24.6981410576917	1.11981623005762e-134	3.84948027244608e-132
TERG_04007	12320.1820847007	2.78703045182618	0.113018350056601	24.4010671000403	1.83034765113511e-132	6.07075142840582e-130