

EBSeq: An R package for differential expression analysis using RNA-seq data

Galaxy Manual

Jeea Choi and Ning Leng

Table of Contents

1. Galaxy.....	1
2. Preparation for the analysis.....	2
3. Gene level DE test across two conditions.....	2
4. Get Normalized Expressions	6
5. Get All Possible Patterns in a Multiple Condition Design.....	6
6. Choose Patterns of interest in a Multiple Condition Design.....	7
7. Gene level DE test across multiple Condition Design	8
8. Get Ig vector from gene-isoform mapping for isoform level DE analysis	10
9. Isoform level DE test across two conditions.....	11
10. Isoform level DE test across multiple conditions	13
11. Problem shooting	14
Reference:	14

1. Galaxy

The empirical Bayes model in Leng et al., 2013 is implemented in Galaxy (<http://galaxy.morgridge.net/>). This manual is a guideline for using the galaxy EBSeq interface, which will allow a user to run EBSeq without directly using R. Files can be uploaded in tab delimited format. Guideline for two condition gene DE analysis can be found in section 3, guidelines for multiple condition gene DE analysis can be found in section 5-7, and guidelines for two/multiple condition isoform DE analysis can be found in section 9/10.

2. Preparation for the analysis

As shown below, upload the file

Galaxy / MIR Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools

search tools

- NGS: RNA Analysis
- Fluidigm analysis package: ANOVA and PCA
- Reference database tools
- RNA-Seq File Manipulation
- RNA DE of Genes
- EBSeq
- Additional Downstream Analysis
- Get Data**
 - Upload File from your computer**
 - UCSC Main table browser
 - UCSC Test table browser
 - UCSC Archaea table browser
 - BX table browser
 - Get Microbial Data
 - BioMart Central server
 - BioMart Test server
 - CBI Rice Mart rice mart
 - GrameneMart Central server
 - modENCODE fly server
 - Flymine server
 - Flymine test server
 - modENCODE modMine server

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
Choose File No file chosen
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<i>Please create or log in to a Galaxy account to view files uploaded via FTP.</i>		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at galaxydev.morgridge.net using your Galaxy credentials (email address and password).

Convert spaces to tabs:
☐ Yes
Use this option if you are entering intervals by hand.

Genome:
Click to Search or Select

Execute

3. Gene level DE test across two conditions

Input requirement:

The input file formats supported by EBSeq are .csv, .tab, or .txt (tab delimited). In your input file, the rows should be the genes either with or without column names. In other words, the first column shows your gene names. Note: This example does not use isoform level expression data. An example of isoform expression analysis is shown in Section 9.

Example data set in .csv format:

GeneMat.csv (without column names): we will use this to show following screenshot

Galaxy / MIR Galaxy		Analyze DataWorkflowShared DataVisualizationHelpUser										
Tools												
search tools												
NGS: RNA Analysis		"Gene_1"	1879	2734	2369	2636	2188	9743	9932	10099	9829	9831
Fluidigm analysis package: ANOVA and PCA		"Gene_2"	24	40	22	27	31	118	108	144	117	113
Reference database tools		"Gene_3"	3291	3259	3214	3407	3298	1058	960	679	605	662
RNA-Seq File Manipulation		"Gene_4"	97	124	146	114	126	33	19	31	22	36
RNA DE of Genes		"Gene_5"	485	485	469	428	475	128	135	103	118	110
EBSeq		"Gene_6"	113	92	64	96	137	39	16	23	30	16
Additional Downstream Analysis		"Gene_7"	886	687	771	786	768	3002	2768	2861	2979	3104
Get Data		"Gene_8"	84	25	67	62	61	277	246	297	241	212
Upload File from your computer		"Gene_9"	68	63	94	70	64	255	260	233	293	299
UCSC Main table browser		"Gene_10"	802	874	863	853	937	212	201	236	232	176
UCSC Test table browser		"Gene_11"	3713	3620	3805	3682	3629	917	902	855	982	935
UCSC Archaea table browser		"Gene_12"	144	172	109	98	146	25	33	24	23	15
BX table browser		"Gene_13"	19	16	15	25	30	3	6	12	5	6
Get Microbial Data		"Gene_14"	12488	13374	13208	13298	13286	3413	2949	3408	3414	3384
		"Gene_15"	928	1396	1192	830	962	4535	4490	4612	4581	4473
		"Gene_16"	3445	3424	3567	3256	3299	711	795	723	830	902
		"Gene_17"	32	23	25	24	31	96	106	110	133	78
		"Gene_18"	2465	2574	2269	2382	2286	555	599	586	556	505
		"Gene_19"	575	497	459	706	713	2036	2007	2120	2246	2093
		"Gene_20"	2391	2547	2639	2677	2551	524	749	598	520	504

GeneMat.csv (with column names): It fails if the first cell from the first row is empty

Galaxy / MIR Galaxy		Analyze Data	Workflow	Shared Data	Visualization	Help	User				
Tools	gene	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
NGS: RNA Analysis	Gene_1	1879	2734	2369	2636	2188	9743	9932	10099	9829	9831
Fluidigm analysis package: ANOVA and PCA	Gene_2	24	40	22	27	31	118	108	144	117	113
Reference database tools	Gene_3	3291	3259	3214	3407	3298	1058	960	679	605	662
RNA-Seq File Manipulation	Gene_4	97	124	146	114	126	33	19	31	22	36
RNA DE of Genes	Gene_5	485	485	469	428	475	128	135	103	118	110
EBSeq	Gene_6	113	92	64	96	137	39	16	23	30	16
▪ Gene level DE test across two conditions Runs EBSeq to find DE genes across two conditions	Gene_7	886	687	771	786	768	3002	2768	2861	2979	3104
	Gene_8	84	25	67	62	61	277	246	297	241	212
	Gene_9	68	63	94	70	64	255	260	233	293	299
	Gene_10	802	874	863	853	937	212	201	236	232	176
▪ Get Normalized Expressions Calculate normalization factors and get the normalized expression matrix	Gene_11	3713	3620	3805	3682	3629	917	902	855	982	935
	Gene_12	144	172	109	98	146	25	33	24	23	15
	Gene_13	19	16	15	25	30	3	6	12	5	6
▪ Get All Possible Patterns in a Multiple Condition Design Get all possible patterns in a	Gene_14	12488	13374	13208	13298	13286	3413	2949	3408	3414	3384
	Gene_15	928	1396	1192	830	962	4535	4490	4612	4581	4473
	Gene_16	3445	3424	3567	3256	3299	711	795	723	830	902

EBSeq in Galaxy:

Go to EBSeq and click Gene level DE test across two conditions

Galaxy / MIR Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools

RNA-Seq File Manipulation

RNA DE of Genes

EBSeq

- Gene level DE test across two conditions Runs EBSeq to find DE genes across two conditions
- Get Normalized Expressions Calculate normalization factors and get the normalized expression matrix
- Get All Possible Patterns in a Multiple Condition Design Get all possible patterns in a multiple condition design
- Choose Patterns of Interest in a Multiple Condition Design Choose patterns of interest in a multiple condition design
- Gene level DE test across multiple conditions Runs EBSeq to find DE genes across multiple (more than two) conditions
- Get lq vector from gene-isoform mapping for isoform level DE analysis Get lq vector from gene-isoform mapping for isoform level DE analysis

Gene level DE test across two conditions (version 1.0.0)

Gene Expression (tab delimited, please use the unnormalized values, e.g. expected counts form RSEM):

88: GeneMat.csv

If header starts with #, first run Workflow Remove '#' from beginning of first line!

The First Row is Sample Names?:

No

Enter which condition each sample belongs to (separated by comma, no space please):

C1,C1,C1,C1,C1,C2,C2,C2,C2,C2

Sample with condition that comes first alphabetically will be in numerator (C1 relative to C2).

Target FDR:

0.05

Execute

Link to view or download documentation written by Ning:

http://www.morgridge.net/upload/files/cshafer/Docs/EBSeq_v1_6_0.pdf

http://www.morgridge.net/upload/files/cshafer/Docs/EBSeq_Vignette_postv1_5_2.pdf

The input Conditions should have exactly two levels. The length of the Condition vector should be exactly the same as the number of columns in the data file (except the gene names column).

Four output files will be generated. Each of the first 3 files contains Posterior probability of being DE (PPDE), Fold Change (RealFC), Posterior Fold Change (PostFC) and normalized gene expressions. The Four files are: Genes with the same order as in input file; Genes sorted by PPDE; DE Genes under target FDR (PPDE>=TargetFDR) and sorted by PPDE; Library size factor for each sample.

Next, a user can customize:

1. The input file. All uploaded files will be shown to choose input file.
2. Select whether the first row is sample names. "No" in this example.
3. Enter which condition each sample belongs to: The number of typed condition separated by comma (,) needs to be matched the number of columns in GeneMat.csv. Here we call the first condition as C1 and the second condition as C2. In this example, the first 5 samples are from condition 1 and the other 5 are from condition 2.
4. Target false discovery rate (FDR); the default is 0.05
5. Press "Execute" button

Galaxy / MIR Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools

RNA-Seq File Manipulation

RNA DE of Genes

EBSeq

- Gene level DE test across two conditions Runs EBSeq to find DE genes across two conditions
- Get Normalized Expressions Calculate normalization factors and get the normalized expression matrix
- Get All Possible Patterns in a Multiple Condition Design Get all possible patterns in a multiple condition design
- Choose Patterns of Interest in a Multiple Condition Design Choose patterns of interest in a multiple condition design
- Gene level DE test across multiple conditions Runs EBSeq to find DE genes across multiple (more than two) conditions
- Get lq vector from gene-isoform mapping for isoform level DE analysis Get lq vector from gene-isoform mapping for isoform level DE analysis

Gene level DE test across two conditions (version 1.0.0)

Gene Expression (tab delimited, please use the unnormalized values, e.g. expected counts form RSEM):

88: GeneMat.csv

If header starts with #, first run Workflow Remove '#' from beginning of first line!

The First Row is Sample Names?:

No

Enter which condition each sample belongs to (separated by comma, no space please):

C1,C1,C1,C1,C1,C2,C2,C2,C2,C2

Sample with condition that comes first alphabetically will be in numerator (C1 relative to C2).

Target FDR:

0.05

Execute

Link to view or download documentation written by Ning:

http://www.morgridge.net/upload/files/cshafer/Docs/EBSeq_v1_6_0.pdf

http://www.morgridge.net/upload/files/cshafer/Docs/EBSeq_Vignette_postv1_5_2.pdf

The input Conditions should have exactly two levels. The length of the Condition vector should be exactly the same as the number of columns in the data file (except the gene names column).

Four output files will be generated. Each of the first 3 files contains Posterior probability of being DE (PPDE), Fold Change (RealFC), Posterior Fold Change (PostFC) and normalized gene expressions. The Four files are: Genes with the same order as in input file; Genes sorted by PPDE; DE Genes under target FDR (PPDE>=TargetFDR) and sorted by PPDE; Library size factor for each sample.

Explaining the Outputs

Five files will be generated for GeneMat.csv example (screenshot below):

- (1) Normalization factors: estimated library size for each sample that is used for median-by-ratio normalization.
- (2) Sorted output with FDR cutoff: Columns are posterior probability of being DE (PPDE), Fold Change (RealFC), posterior Fold Change (PosteriorFC), and median-by-ratio normalized gene expressions (cell orders are the same with input cell order). Only genes with $PPDE \geq 1 - \text{Target_FDR}$ are listed. The RealFC calculates $(\text{mean_condition1} + 0.01) / (\text{mean_condition2} + 0.01)$, in which the within condition mean is calculated using normalized data. The posterior fold changes are estimated from the empirical bayes model. The posterior FC estimations will give less extreme values for low expressers. e.g. if gene1 has $\text{mean1} = 5000$ and $\text{mean2} = 1000$, its FC and PostFC will both be 5. If gene2 has $\text{mean1} = 5$ and $\text{mean2} = 1$, its FC will be 5 but its PostFC will be < 5 and closer to 1. Therefore, when we sort the PostFC, gene2 will be less significant than gene1.
- (3) Output sorted by PPDE: Columns are the same as in (1). Genes are sorted decreasingly by PPDE.
- (4) Output: Columns are the same as in (1) and (2). Rows are the genes in the same order as the input file.
- (5) Version Info: Shows EBSeq version info. Input parameters (e.g. FDR chosen, list of conditions) are shown.

The screenshot shows the Galaxy / MIR Galaxy interface. A green notification box in the center states: "The following job has been successfully added to the queue: 98: Version Info, 99: Output, 100: Output sorted by PPDE, 101: Sorted output with FDR cutoff, 102: Normalization factors". Below this, a message says: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."

The History pane on the right shows a list of jobs. A red circle highlights the following jobs:

- 102: Normalization factors
- 101: Sorted output with FDR cutoff
- 100: Output sorted by PPDE
- 99: Output
- 98: Version Info

Below these jobs, the "GeneMat.csv" file is shown with a table of gene expression data. The table has 6 columns: 1. Chrom, 2. Start, 3. End, 4, 5, 6. The rows represent different genes.

	1. Chrom	2. Start	3. End	4	5	6
"Gene_1"	1879	2734	2369	2636	2188	
"Gene_2"	24	48	22	27	31	
"Gene_3"	3291	3259	3214	3487	3298	
"Gene_4"	97	124	146	114	126	
"Gene_5"	485	485	469	428	475	
"Gene_6"	113	92	64	96	137	

4. Get Normalized Expressions

Normalized expressions can be obtained using the 'Get Normalized expressions' module without performing DE analysis (Section 3).

Input requirement:

The input file formats are similar to that of Section 3.

Explaining the Outputs

Three files will be generated for GeneMat.csv example:

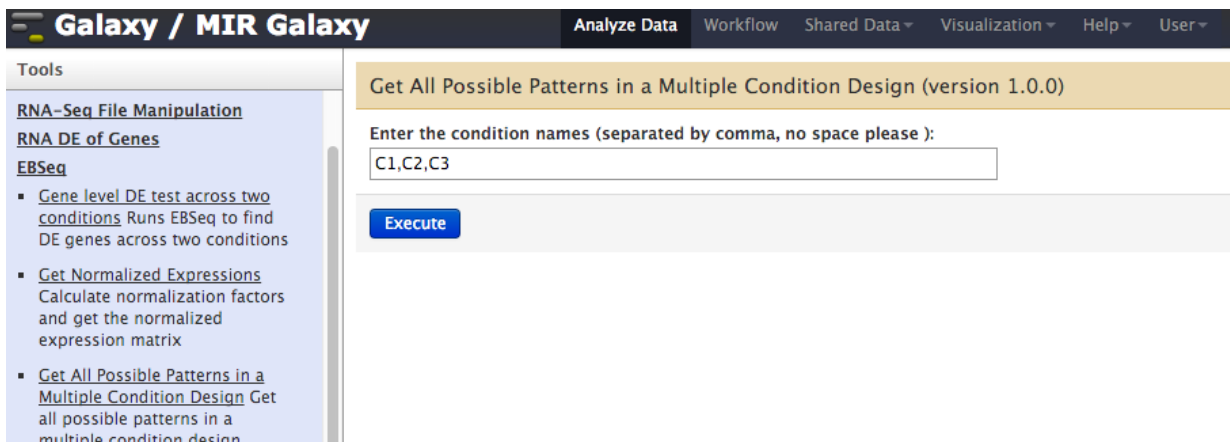
- (1) Normalization factors: Library size for each sample that is used for median-by-ratio normalization
- (2) Boxplots: Boxplot of normalized expression for each sample. Y axis shows normalized expression values.
- (3) Normalized expression: median-by-ratio normalized (library size adjusted) gene expressions

5. Get All Possible Patterns in a Multiple Condition Design

In a two condition DE analysis, a gene can only be either DE or EE. In a case with more than 2 conditions, more than multiple patterns are possible. For example, in a case with 3 conditions, there are total 5 possible patterns: $C1 = C2 = C3$, $C1 = C2 \neq C3$, $C1 \neq C2 = C3$, $C2 \neq C1 = C3$, $C1 \neq C2 \neq C3$. Before performing DE test across multiple conditions, we need to construct the possible patterns, first. This can be done by the "Get All Possible Patterns" module.

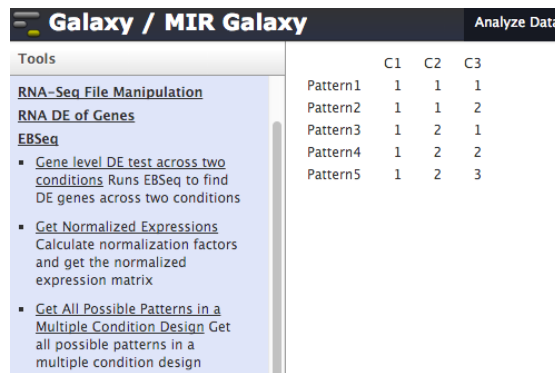
Input:

Enter the condition names: Comma separated multiple conditions need to be typed



The screenshot shows the Galaxy / MIR Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. On the left, a 'Tools' sidebar lists categories like 'RNA-Seq File Manipulation', 'RNA DE of Genes', and 'EBSeq'. The main panel displays the 'Get All Possible Patterns in a Multiple Condition Design (version 1.0.0)' module. It features a text input field with the placeholder 'Enter the condition names (separated by comma, no space please):' and the example text 'C1,C2,C3'. Below the input field is a blue 'Execute' button.

Explaining the Outputs



	C1	C2	C3
Pattern1	1	1	1
Pattern2	1	1	2
Pattern3	1	2	1
Pattern4	1	2	2
Pattern5	1	2	3

Output shows the possible patterns in a multiple condition design

For example, the first pattern is $C1 = C2 = C3$

the second pattern is $C1 = C2 \neq C3$

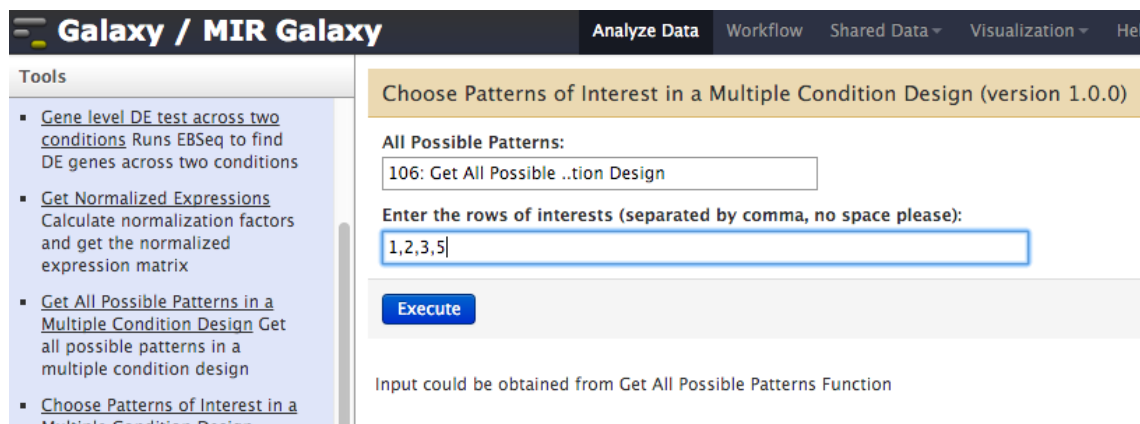
For data sets with more than 3 conditions, the number of possible patterns increase exponentially. To reduce the runtime and to make the output more easily interpreted, we suggest the user to choose a small set of patterns that related to the experimental hypothesis (Note we suggest the user to always include the all EE pattern and all DE pattern, usually the first pattern and the last pattern).

A user can choose patterns that is of interest directly from this output before continuing on to further analysis (Section 6)

6. Choose Patterns of interest in a Multiple Condition Design

Input:

Use the output from Section 5 and enter the rows of interests. If we are interested in identifying genes that follows pattern 2, 3, and 5 but not 4, we can type "1,2,3,5". Note it is always suggested to include the all EE pattern (the first one here) and the all DE pattern (the last one) as background patterns. (screenshot below)



Galaxy / MIR Galaxy Analyze Data Workflow Shared Data Visualization Help

Tools

- Gene level DE test across two conditions Runs EBSeq to find DE genes across two conditions
- Get Normalized Expressions Calculate normalization factors and get the normalized expression matrix
- Get All Possible Patterns in a Multiple Condition Design Get all possible patterns in a multiple condition design
- Choose Patterns of Interest in a Multiple Condition Design

Choose Patterns of Interest in a Multiple Condition Design (version 1.0.0)

All Possible Patterns:
106: Get All Possible ..tion Design

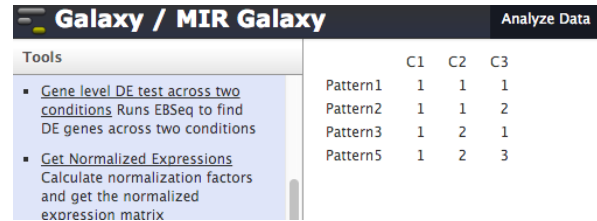
Enter the rows of interests (separated by comma, no space please):
1,2,3,5

Execute

Input could be obtained from Get All Possible Patterns Function

Output:

Output shows the chosen patterns.



The screenshot shows the Galaxy / MIR Galaxy interface. On the left, under the 'Tools' section, the 'Gene level DE test across two conditions' tool is highlighted. On the right, the 'Analyze Data' section displays a table with the following data:

	C1	C2	C3
Pattern1	1	1	1
Pattern2	1	1	2
Pattern3	1	2	1
Pattern5	1	2	3

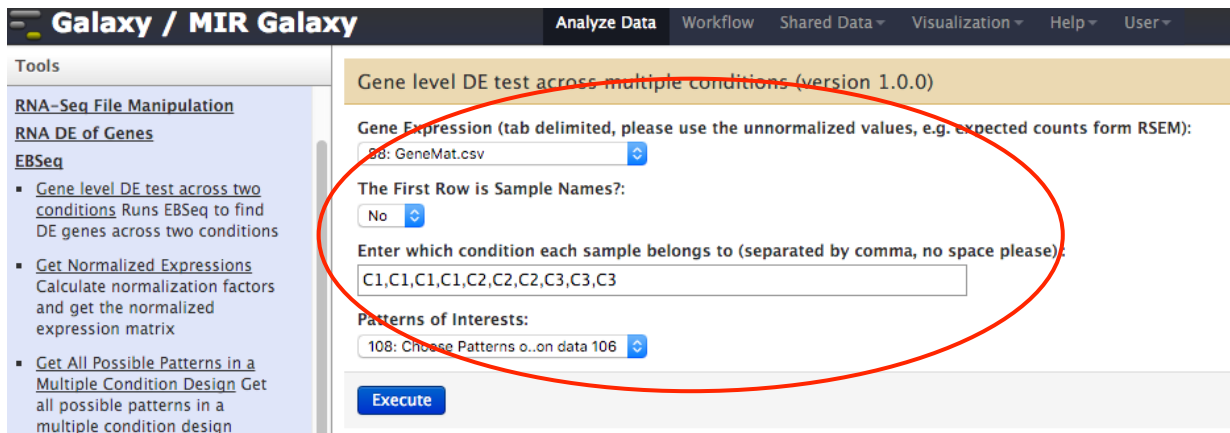
7. Gene level DE test across multiple Condition Design

Input:

The input file formats are similar to that of Section 3.

Next, a user can customize:

1. The input file. All uploaded files will be shown to choose input file.
2. Select whether the first row is sample names. "No" in this example.
3. Enter which condition each sample belongs to: The number of typed condition separated by comma (,) needs to be matched by the number of columns in GeneMat.csv
4. Patterns of interest. Constructed patterns from Section 5 or Section 6 can be chosen.
5. Press "Execute" button.



The screenshot shows the Galaxy / MIR Galaxy interface with the 'Gene level DE test across multiple conditions (version 1.0.0)' tool selected. The configuration fields are as follows:

- Gene Expression (tab delimited, please use the unnormalized values, e.g. expected counts form RSEM):** 88: GeneMat.csv
- The First Row is Sample Names?:** No
- Enter which condition each sample belongs to (separated by comma, no space please):** C1,C1,C1,C1,C2,C2,C2,C3,C3,C3
- Patterns of Interests:** 108: Choose Patterns o...on data 106
- Execute** button

Explaining the Outputs

Four files will be generated for GeneMat.csv example (screenshot below):

- (1) Normalization factors: Library size for each sample that is used for median-by-ratio normalization
- (2) Pattern with highest PP: Column 1 shows the pattern with the highest posterior probability for each gene (MAP). The other columns are the median-by ratio normalized gene expressions. Rows are the genes with the same order as input.

- (3) PP of each pattern: Columns are posterior probability of being each pattern. Rows are the genes with the same order as input. The higher the PP is, the more likely that this gene is following this specific pattern.
- (4) Version Info: Shows EBSeq version info. Input parameters (e.g. FDR chosen, list of conditions) are shown.

The screenshot displays the Galaxy / MIR Galaxy web interface. The top navigation bar includes 'Galaxy / MIR Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A status bar on the right indicates 'Using 2.2 MB'.

On the left, the 'Tools' panel lists several RNA-Seq analysis tools, including 'RNA-Seq File Manipulation', 'RNA DE of Genes', and 'EBSeq'.

The central area shows a green notification box with a checkmark icon, stating: 'The following job has been successfully added to the queue:'. Below this, the job details are listed:

- 109: Version Info
- 110: PP of each pattern
- 111: Pattern with highest PP
- 112: Normalization factors

A note below the job details reads: 'You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.'

On the right, the 'History' panel is visible, showing a list of jobs. The job '112: Normalization factors' is highlighted with a red circle. The history list includes:

- 112: Normalization factors (2.2 MB)
- 111: Pattern with highest PP
- 110: PP of each pattern
- 109: Version Info

8. Get Ig vector from gene-isoform mapping for isoform level DE analysis

For isoform level analysis, an *Ig* vector is required (see Leng *et al.*, 2013, or the EBSeq vignette for details on *Ig*). If you have the *Ig* vector file generated from RSEM, please ignore this subsection.

Input:

Again, csv, xls, or xlsx files are accepted. The first column specifies the isoform names and the second column specifies the corresponding gene names.

Example data set in .csv format:
igvec.csv

Galaxy / MIR Galaxy		Analyze Data
Tools		
Calculate normalization factors and get the normalized expression matrix	"Iso_1_1"	"Gene_1"
	"Iso_1_2"	"Gene_2"
	"Iso_1_3"	"Gene_3"
	"Iso_1_4"	"Gene_4"
	"Iso_1_5"	"Gene_5"
	"Iso_1_6"	"Gene_6"
	"Iso_1_7"	"Gene_7"
	"Iso_1_8"	"Gene_8"
	"Iso_1_9"	"Gene_9"
	"Iso_1_10"	"Gene_10"
	"Iso_1_11"	"Gene_11"
	"Iso_1_12"	"Gene_12"
	"Iso_1_13"	"Gene_13"
	"Iso_1_14"	"Gene_14"
	"Iso_1_15"	"Gene_15"
	"Iso_1_16"	"Gene_16"
	"Iso_1_17"	"Gene_17"

Galaxy / MIR Galaxy		Analyze Data	Workflow	Shared Data	Visualization	Help	User
Tools							
Calculate normalization factors and get the normalized expression matrix							
Get All Possible Patterns in a Multiple Condition Design							
Choose Patterns of Interest in a Multiple Condition Design							
Gene level DE test across multiple conditions							
Get Ig vector from gene-isoform mapping for isoform level DE analysis							

Get Ig vector from gene-isoform mapping for isoform level DE analysis (version 1.0.0)

input:

121: igvec.csv

Input should be no-header, first column is isoform names, second column is gene names. See below for more information.

Execute

⚠ If you are using the 'RSEM to EBSeq' tool for extracting the correct RSEM output columns to be input directly to EBSeq, your isoforms here must be in alphabetical (A-Z) order, or the Ig vector will not correspond to the correct isoform! If you did NOT use 'RSEM to EBSeq' but did the cuts and joins on RSEM output manually, your genes (rather than isoforms) should be in alphabetical (A-Z) order.

TIP: To go directly from RSEM output to the input for this tool (does the cuts and sorting in one step), use the Ready_for_Ig_vector Workflow ONLY IF YOU MAKE THE EBSEQ INPUT FILE USING 'RSEM TO EBSE' TOOL. The first column should be isoform names, and the second column should be gene names, such as what one gets from using the 'Text Manipulation: Cut' tool for c1,c2 on RSEM output file isoforms.results. Please make sure there is no header on the input. Header can easily be removed with the 'Text Manipulation: Remove beginning' tool to remove the first line. The aforementioned Workflow does all of the above in one step.

Explaining the Outputs

"Ig vector" file will be generated

9. Isoform level DE test across two conditions

Input:

The *Ig* vector file from Section 4.1 or RSEM `rsem-generate-ngvector` function (<http://deweylab.biostat.wisc.edu/rsem/rsem-generate-ngvector.html>).

The data input could be .csv, .txt, or .tab files (tab delimited).

Rows are isoforms and columns are samples.

- The first column shows the isoform names

Example data set

IsoMat.csv

Galaxy / MIR Galaxy		<div>Analyze DataWorkflowShared DataVisualizationHelpUser</div>										
Tools		"Iso_1_1"	176	212	164	142	180	687	681	737	446	527
RNA-Seq File Manipulation		"Iso_1_2"	789	915	919	942	892	3334	3211	2641	3371	3382
RNA DE of Genes		"Iso_1_3"	1300	1377	1408	1376	1395	383	440	367	378	369
EBSeq		"Iso_1_4"	474	487	483	473	499	1587	1671	1437	1598	1668
▪ Gene level DE test across two conditions	Runs EBSeq to find DE genes across two conditions	"Iso_1_5"	1061	949	816	1040	897	211	266	289	231	275
		"Iso_1_6"	346	348	426	392	488	1452	1751	1487	1310	1370
▪ Get Normalized Expressions	Calculate normalization factors and get the normalized expression matrix	"Iso_1_7"	2604	3284	2643	2705	2651	794	823	827	789	808
		"Iso_1_8"	859	981	894	793	913	235	223	244	312	263
▪ Get All Possible Patterns in a Multiple Condition Design	Get all possible patterns in a multiple condition design	"Iso_1_9"	2598	1990	2720	2700	2354	10108	11481	5625	8481	7759
		"Iso_1_10"	322	448	451	328	314	683	794	1429	1302	1137
▪ Choose Patterns of Interest in a Multiple Condition Design	Choose patterns of interest in a multiple condition design	"Iso_1_11"	514	668	654	423	611	155	107	143	247	141
		"Iso_1_12"	18	20	17	21	20	52	80	47	52	48
		"Iso_1_13"	12119	12260	11659	14918	12126	2785	3393	4700	3876	3461
		"Iso_1_14"	577	615	726	623	637	1921	2102	2348	2411	2265
		"Iso_1_15"	2802	3111	3574	3438	3635	860	789	1030	981	1042
		"Iso_1_16"	22	25	18	15	21	49	83	66	94	76
		"Iso_1_17"	187	182	221	198	244	662	778	872	648	913
		"Iso_1_18"	355	324	324	350	316	1324	1234	1033	1350	990

A user can customize:

1. The input file. All uploaded files will be shown to choose input file.
2. Select whether the first row is sample names. "No" in this example.
3. Enter which condition each sample belongs to: The number of typed condition separated by comma (,) needs to be matched the number of columns in IsoMat.csv
4. The name for the *Ig* vector file.
5. Target false discovery rate (FDR); the default is 0.05

Galaxy / MIR Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools

RNA-Seq File Manipulation

RNA DE of Genes

EBSeq

- Gene level DE test across two conditions Runs EBSeq to find DE genes across two conditions
- Get Normalized Expressions Calculate normalization factors and get the normalized expression matrix
- Get All Possible Patterns in a Multiple Condition Design Get all possible patterns in a multiple condition design
- Choose Patterns of Interest in a Multiple Condition Design Choose patterns of interest in a multiple condition design
- Gene level DE test across multiple conditions Runs EBSeq to find DE genes across multiple (more than two) conditions
- Get Ig vector from gene-isoform mapping for isoform level DE analysis Get Ig vector from gene-isoform mapping for isoform level DE analysis

Isoform level DE test across two conditions (version 1.0.0)

Isoform Expression (tab delimited, please use the unnormalized values, e.g. expected counts from RSEM):

123: IsoMat.csv

The First Row is Sample Names?:

No

Enter which condition each sample belongs to (separated by comma, no space please):

C1,C1,C1,C1,C1,C2,C2,C2,C2

Sample with condition that comes first alphabetically will be in numerator (C1 relative to C2).

Ig Vector to define the uncertainty groups:

122: Ig vector

Target FDR:

0.05

Execute

The input Conditions should have exactly two levels. The length of the Condition vector should be exactly the same as the number of columns in the data file (except the isoform names column).

The Ig Vector could be generated by the GetIg function or obtained from RSEM output.

Four output files will be generated. Each of the first 3 files contains Posterior probability of being DE (PPDE), Fold Change (RealFC), Posterior Fold Change (PostFC) and normalized isoform expressions. The four files are:

Isoforms with the same order as in input file; Isoforms sorted by PPDE; DE Isoforms under target FDR (PPDE>=TargetFDR) and sorted by PPDE; Library size factor for each sample.

Explaining the Outputs

Five files will be generated:

- (1) Normalization factors: Library size for each sample that is used for median-by-ratio normalization.
- (2) Sorted output with target FDR: Columns are posterior probability of being DE (PPDE), Fold Change (RealFC), posterior Fold Change (PosteriorFC), and median-by ratio normalized gene expressions. Only genes with $PPDE \geq 1 - Target_FDR$ are listed.
- (3) Output sorted by PP: Columns are the same as in (1). Isoforms are sorted decreasingly by PPDE
- (4) Output: Columns are the same as in (1) and (2). Rows are the isoforms in the same order as the input file
- (5) Version Info: Shows EBSeq version info. Input parameters (e.g. FDR chosen, list of conditions) are shown.

10. Isoform level DE test across multiple conditions

Input:

The input file formats are similar to that of Section 9.

Next, a user can customize:

1. The input file. All uploaded files will be shown to choose input file.
2. Select whether the first row is sample names. “No” in this example.
3. Enter which condition each sample belongs to: The number of typed condition separated by comma (,) needs to be matched the number of columns in IsoMat.csv
4. The name for the *Ig* vector file.
5. Pattern of interests. Constructed patterns from Section 5 or Section 6 can be chosen.
6. Press “Execute” button.

Galaxy / MIR Galaxy Analyze Data Workflow Shared Data Visualization Help User

Tools

RNA-Seq File Manipulation

RNA DE of Genes

EBSeq

- [Gene level DE test across two conditions](#) Runs EBSeq to find DE genes across two conditions
- [Get Normalized Expressions](#) Calculate normalization factors and get the normalized expression matrix
- [Get All Possible Patterns in a Multiple Condition Design](#) Get all possible patterns in a multiple condition design
- [Choose Patterns of Interest in a Multiple Condition Design](#) Choose patterns of interest in a multiple condition design
- [Gene level DE test across multiple conditions](#) Runs EBSeq to find DE genes across multiple (more than two) conditions
- [Get Ig vector from gene-isoform mapping for isoform level DE analysis](#) Get Ig vector from gene-isoform mapping for isoform level DE analysis

Isoform level DE test across multiple conditions (version 1.0.0)

Isoform Expression (tab delimited, please use the unnormalized values, e.g. expected counts from RSEM):

123: IsoMat.csv

The First Row is Sample Names?:

No

Enter which condition each sample belongs to (separated by comma, no space please):

C1,C1,C1,C1,C2,C2,C2,C3,C3,C3

Ig Vector to define the uncertainty groups:

122: Ig vector

Patterns of Interests:

108: Choose Patterns o..on data 106

Execute

The input file and conditions should have more than two levels (for exactly two levels, please use the 'isoform level DE test across two conditions' tool). Format of the Isoform Expression file should be transcript_id, EC1, EC2 ... Please do not include the gene name in this file. The length of the Condition vector should be exactly the same as the number of columns in the data file (except the isoform names column).

The patterns of interests could be obtained by function Get All Possible Patterns (and optionally, if there are too many patterns generated, the function Choose Patterns could be used to choose only subset of the patterns.)

The Ig Vector could be generated by the GetIg function or obtained from RSEM output.

Three output files will be generated. The first file contains the Posterior probability of being each pattern. The second file contains the pattern with the highest PP for each isoform and the normalized expressions. Isoforms are with the same order as in input file. The last file provides the library size factor for each sample.

Explaining the Outputs

Four files will be generated for IsoMat.csv example (screenshot below):

- (1) Normalization factors: Library size for each sample that is used for median-by-ratio normalization
- (2) Pattern with highest PP: Column 1 shows the pattern with the highest posterior probability for each isoform (MAP). The other columns are median-by-ratio normalized isoform expressions. Rows are the isoforms with the same order as input.
- (3) PP of each pattern: Columns are posterior probability of being each pattern. Rows are the isoforms with the same order as input.
- (4) Version Info: Shows EBSeq version info. Input parameters (e.g. FDR chosen, list of conditions) are shown.

11. Trouble shooting

More details of the EBSeq implementation can be found at http://www.biostat.wisc.edu/~kendzior/EBSEQ/EBSeq_Vignette.pdf.

If you have additional questions not addressed in this manual regarding the EBSeq interface, please see the Q&A section on the EBSeq website biostat.wisc.edu/~kendzior/EBSEQ, or contact us at sswanson@morgridge.org.

Format problem: It fails if the first cell from the first row is empty.

Reference:

Leng, N., J.A. Dawson, J.A. Thomson, V. Ruotti, A.I. Rissman, B.M.G. Smits, J.D. Haag, M.N. Gould, R.M. Stewart, and C. Kendzior. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments, *Bioinformatics*, [e-pub ahead of print 21 February 2013] [[Download](#)].

Li, B., V. Ruotti, R.M. Stewart, J.A. Thomson, and C. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26(4): 493-500, 2010.