

MeV Quickstart Guide For RNA-Seq Data

Getting started with RNASeq Data

Preface: This guide is an introduction to using the new RNASeq functions in MeV. The guide contains a brief tour of the new RNASeq file loader and a demonstration of a few of the new functions we have added specifically to support RNASeq data. The guide will first walk you through loading the data using the new RNA-Seq file loader. Then it will describe using an RNA-Seq-optimized module, EdgeR, to find differentially expressed genes between two groups of samples. Finally, it will demonstrate how to examine these differentially expressed genes for functional themes using the new module GOSeq.

These new options were added in MeV v4.7. If you already have MeV v4.7 installed, you can skip the **Setup** step and go directly to **Loading a Data Set**.

Setup

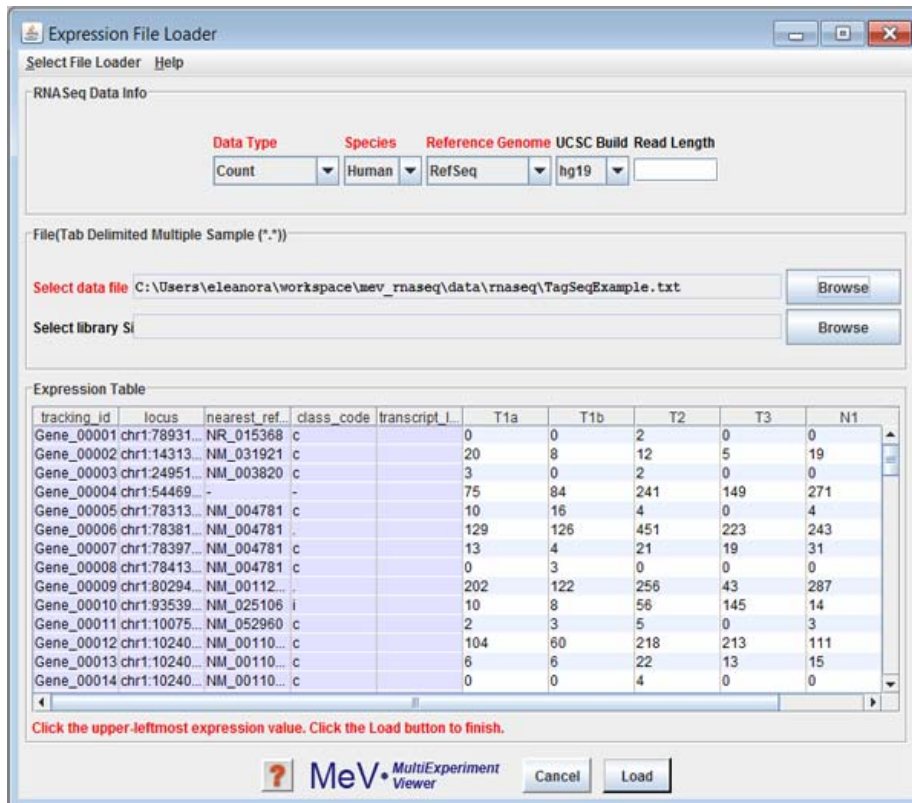
I. Installing MeV

1. First make sure that Java is properly installed on your computer. Java v1.6 or higher for a Windows PC/Linux and v1.5 or higher for Mac OSX needs to be installed in order for MeV to work. Go to <http://java.com/> to get the latest version. Certain MeV modules also require Java 3D, which can be found here: <http://java.sun.com/products/java-media/3D/download.html>
2. Download the [RNASeq Pilot project](#), if you have not done so already.

3. A screen should pop up that asks what you want to do with the files. Whether using a PC, Mac, or Linux, open the files and download them.
4. Once downloaded, open the folder and unzip the file.
5. The unzipped folder can be copied to any convenient location on the hard drive.
6. Open the MeV_4_7_0 folder. Double click the file called tmev.bat to run the program.

Loading a Data Set

1. In the Multiple Array Viewer, go to *File* → *Load Data*.
2. When the window titled Expression File Loader appears, click *Select File loader* → *RNASeq DGE Files*. The RNASeq file loader screen will appear.
3. Click the *Browse* button at the upper right side of the screen. In the file browser that appears, navigate to the MeV folder, then open the *data/rnaseq* folder. Choose the file *TagSeqExample.txt*. This file contains raw count data¹.



The new RNASeq data loader accepts raw count data, RPKM or FPKM, mapped to either ENSEMBL IDs or RefSeq IDs.

4. Choose the appropriate parameters for each of the drop-down menus at the top of the file loader screen. For the data file we have selected, choose the *Data Type*

¹ MeV can also load RPKM data, or combined RPKM/count data. The data file *isoforms.fpkm_cnt_Ref.txt* is an example of this file format.

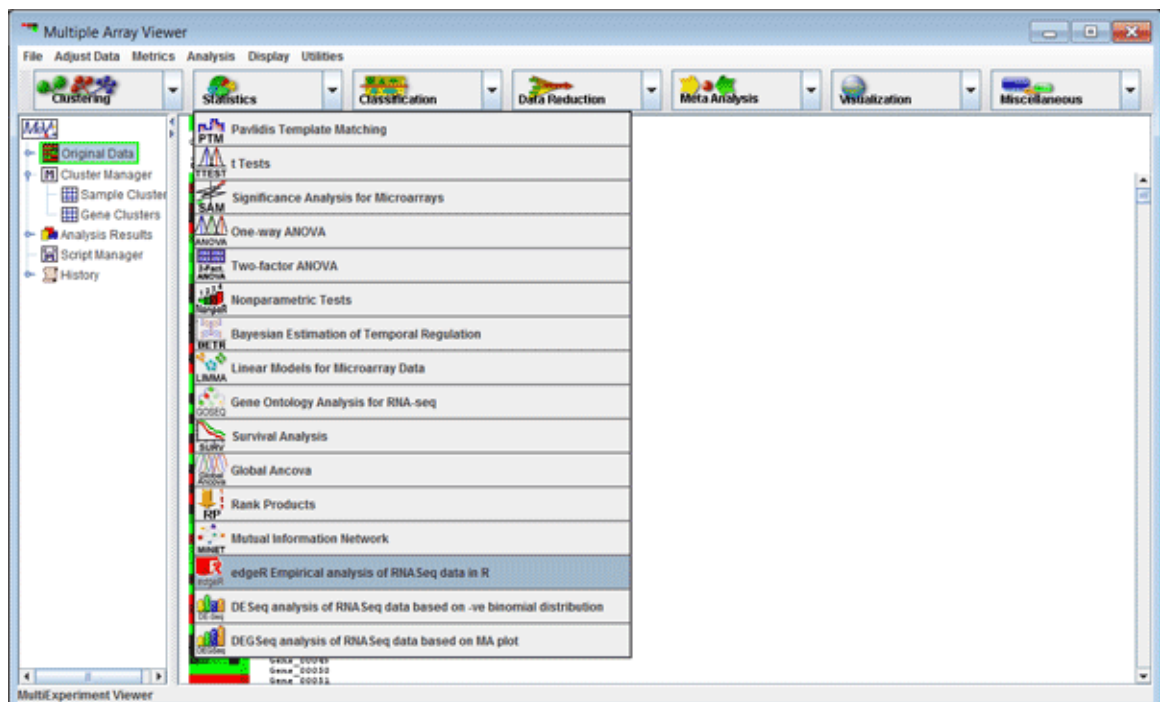
- Count, the *Species* Human, the *Reference Genome* RefSeq, and the *UCSC build* hg19. Leave *Read Length* blank.
5. Click the *Load* button.

RNA-Seq Analysis

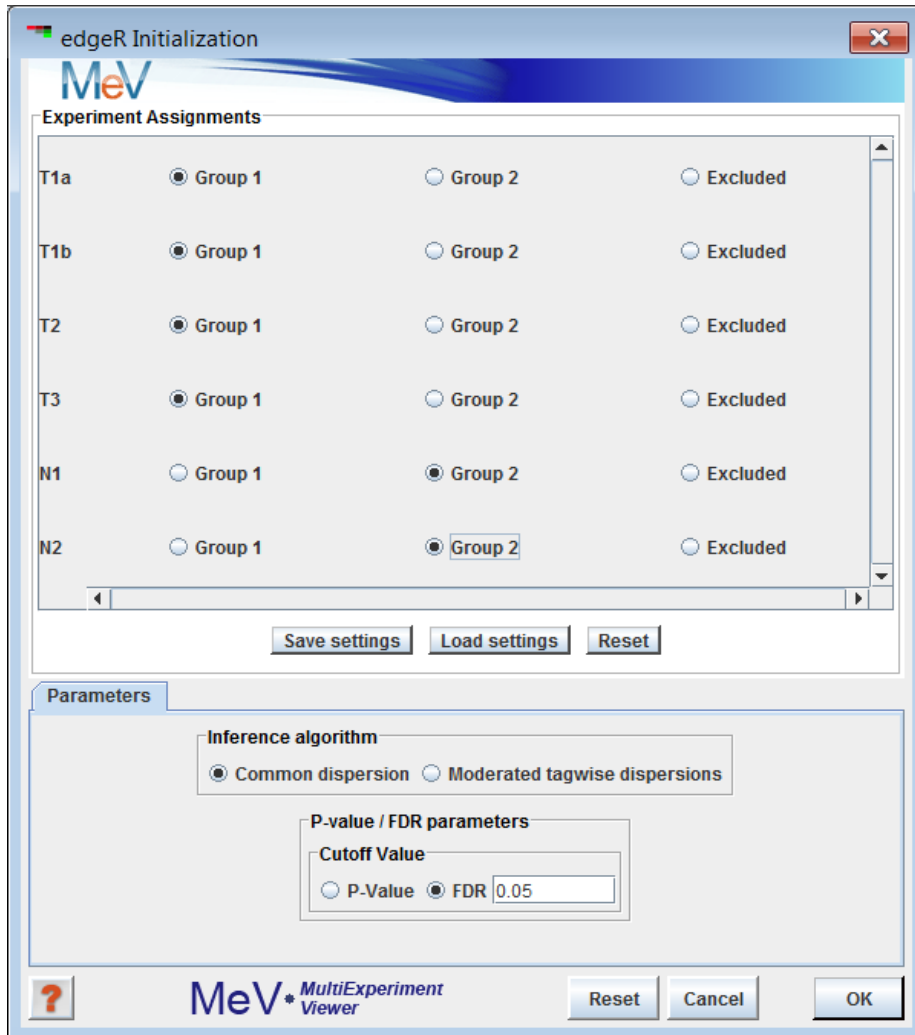
Differential Expression Detection

Begin your RNASeq analysis by testing for differential expression of all of the unique reads. To do this, we will use a module called edgeR, based on the [Empirical Analysis of Digital Gene Expression data in R](#) package written by Mark Robinson.

1. In the row of colorful buttons across the top of the MultiExperiment Viewer window, click the one labeled Statistics. Choose *Empirical Analysis of Digital Gene Expression data in R* (edgeR). An initialization dialog will appear.
2. Select the group membership for each of the six samples. Click "Group 1" for the first four samples, and "Group 2" for the remaining two samples.
3. Leave the default values for the Inference Algorithm and p-value/FDR parameters.
4. Click Ok. The analysis will run and display the results in the result tree, on the left of the Multiple Array Viewer window.



The edgeR module can be found in the Statistics drop-down menu.



The edgeR initialization dialog.

Differential Expression Results

1. Open up the result node labeled edgeR, and expand the nodes to find one labeled *Significant Gene List*. Click on this node to select it and display the list of genes found to be differentially expressed between the two sample groups you selected in the previous section. You can click on the links to launch a web browser displaying more information about individual genes.
2. Right-click on the window in a cell with no links (the *Stored Color* column is a good bet). Choose *Store entire cluster* and click Ok to label each of the genes in this window with a color. This color label will be visible anywhere a gene display is shown in MeV - even in the results of other modules.

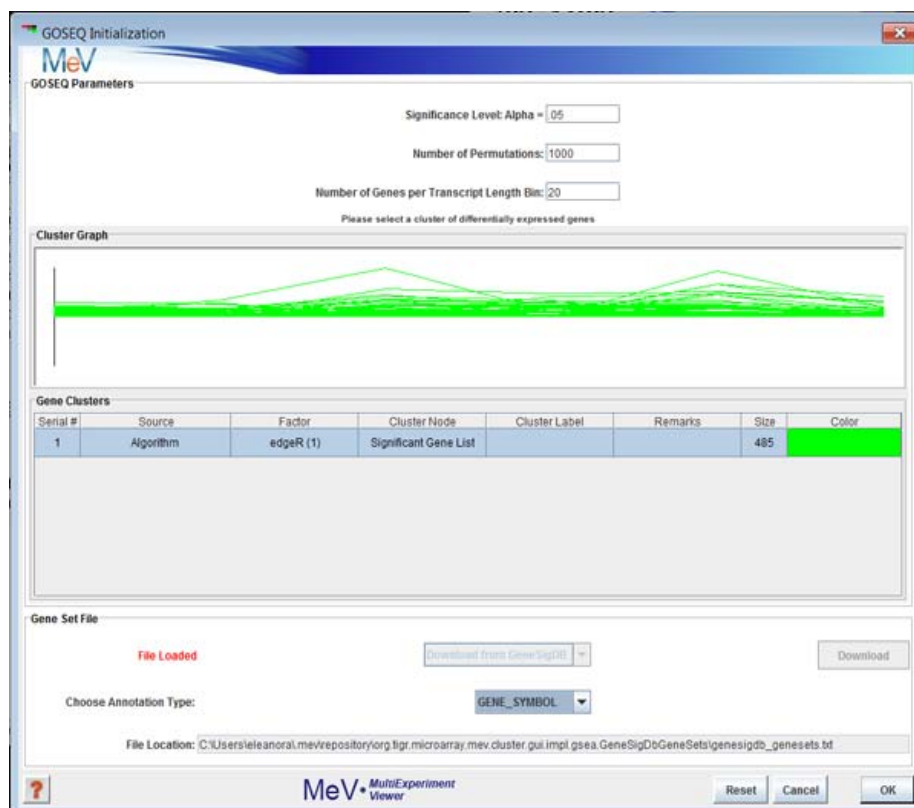
tracking_id	locus	nearest_ref	class_code	transcript	GENE_TITLE	TX_END	CHR	LOCA	TX_START	GC
Gene_01354	chr1:18664	NM_000963	NA	prastatinalar	186643876	chr1:1866401	186643510			
Gene_10880	chr2:22936	NM_001113	c	act-GAP_dom	228384737	chr2:2283335	228384697			
Gene_12457	chr2:18399	NM_138295	c	ncleosome	183993176	chr2:1832892	183993090			
Gene_01353	chr1:18633	NM_003292	c	ncleosome	186332124	chr1:1862802	186331985			
Gene_14803	chr3:96037	-	NA	NA	96047148	chr3:950375	96037545	NA		
Gene_00855	chr1:31215	NM_006762	c	lgsasomat	31215386	chr1:312053	31215346			
Gene_12473	chr2:45004	NM_001003	c	NA	450040712	chr2:4500016	450040652			
Gene_12473	chr2:45004	NM_001003	c	NA	450041395	chr2:4500016	450041271			
Gene_12473	chr2:45004	NM_001003	c	NA	450041815	chr1:4714626	47181862			
Gene_12473	chr2:45004	NM_001003	c	NA	450042581	chr2:239715	23972433			
Gene_12473	chr2:45004	NM_001003	c	NA	450042826	chr2:232387	232388784			
Gene_12473	chr2:45004	NM_001003	c	NA	450042694	chr1:2114332	211462567			
Gene_12473	chr2:45004	NM_001003	c	NA	45004494	chr1:784476	7904362			
Gene_12473	chr2:45004	NM_001003	c	NA	450045437	chr3:331383	331181410			
Gene_12473	chr2:45004	NM_001003	c	NA	450045840	chr1:1586301	158645635			
Gene_12473	chr2:45004	NM_001003	c	NA	450048052	chr2:112812	112874759			
Gene_12473	chr2:45004	NM_001003	c	NA	450045265	chr1:151512	151555143			
Gene_12473	chr2:45004	NM_001003	c	NA	450044598	chr2:540912	54164554			
Gene_12473	chr2:45004	NM_001003	c	NA	450045209	chr3:160118	160152284			
Gene_12473	chr2:45004	NM_001003	c	NA	450043637	chr3:33015	3318349			
Gene_12473	chr2:45004	NM_001003	c	NA	4500499510	chr1:24266	24296469			
Gene_12473	chr2:45004	NM_001003	c	NA	4500470680	chr1:118148	118164936			
Gene_12473	chr2:45004	NM_001003	c	NA	4500479342	chr3:179322	179342062			
Gene_12473	chr2:45004	NM_001003	c	NA	4500400725	chr2:55399	55400532			
Gene_12473	chr2:45004	NM_001003	c	NA	4500478480998	chr2:178479	178480932			
Gene_12473	chr2:45004	NM_001003	c	NA	4500459537891	chr2:159213	159537790			
Gene_12473	chr2:45004	NM_001003	c	NA	4500418852	chr1:150335	150418710			
Gene_12473	chr2:45004	NM_001003	c	NA	4500411532827	chr1:115312	115322787			
Gene_12473	chr2:45004	NM_001003	c	NA	4500449062408	chr3:490617	49062352			
Gene_12473	chr2:45004	NM_001003	c	NA	45004102508814	chr2:102314	102508683			
Gene_12473	chr2:45004	NM_001003	c	NA	4500439945517	chr1:392471	39945476			
Gene_12473	chr2:45004	NM_001003	c	NA	4500413326	chr1:24702	247012963			
Gene_12473	chr2:45004	NM_001003	c	NA	4500479125121	chr1:781154	78124810			
Gene_12473	chr2:45004	NM_001003	c	NA	45004380066	chr3:46340	46340663			

Results of the edgeR module, showing significantly differentially expressed genes/transcripts. Right-click to reveal a context menu with many powerful options.

Examining the differential expression list for signature themes

Now that we have a list of differentially expressed genes, we can examine it for themes. To do this, we will use the GOSeq module. This module is based on the R package [GOSeq](#), by Matthew Young. It is designed to find enriched gene groups in length-biased data, such as RNASeq data. Compare it to tools like EASE for microarray data.

1. From the *Statistics* drop-down menu, choose the item *Gene Ontology analysis for RNA-seq*.
2. Leave the GOSeq parameters *Significance Level: Alpha*, *Number of Permutations* and *Number of Genes per Transcript Length Bin* set at their default values.
3. You should have a cluster pre-selected in the cluster selector dialog. If you have more than one cluster available in this dialog, choose the one you want to examine for geneset enrichment.
4. Choose *Download from GeneSigDb* from the drop-down menu. Click the *Download* button.
5. Check that the *Choose Annotation Type* drop-down menu is set at *GENE_SYMBOL*.
6. Leave the *File Location* field blank.
7. Click *Ok*. GOSeq will run.

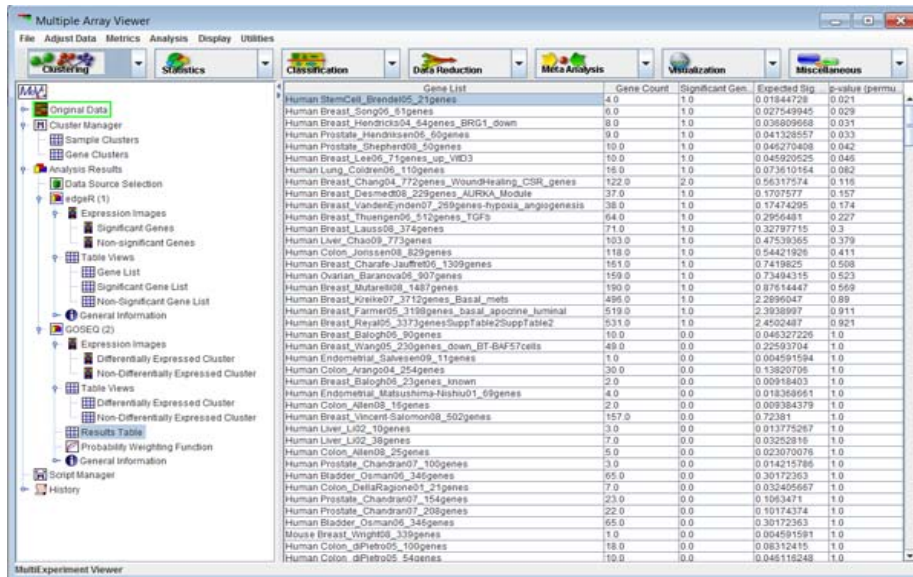


The GOSeq initialization dialog.

Signature theme results

In the Result Tree, you will see a new result node named *GOSEQ*.

1. Open this node and select the node labeled *Results Table*. This table contains the complete list of genelists downloaded from the GeneSigDb database, as well as a rating for each list as to whether the contents of that list is enriched in the selected group of differentiated genes used to run GOSeq.
2. Double-click on the header labeled *p-value* to sort the list. Those gene lists with low p-values, like Human StemCell_Brendel05_21genes, listed here, are enriched in the set of differentially expressed genes we found in our previous edgeR analysis. You can explore this gene list by going to the GeneSigDb website.



The screenshot shows the Multiple Array Viewer (MeV) software interface. The main window displays a table of gene lists and their enrichment statistics. The table has the following columns: Gene List, Gene Count, Significant Genes, Expected Sig., and p-value (permuted). The table is sorted by p-value in ascending order. The first few rows are:

Gene List	Gene Count	Significant Genes	Expected Sig.	p-value (permuted)
Human StemCell_Brendel05_21genes	4.0	1.0	0.01844729	0.021
Human Breast_Sony05_51genes	8.0	1.0	0.027549945	0.029
Human Breast_Hendricks04_54genes_BRG1_down	8.0	1.0	0.036809958	0.031
Human Prostate_Hendriksen05_60genes	9.0	1.0	0.041328557	0.033
Human Prostate_Shepherd03_50genes	10.0	1.0	0.046270408	0.042
Human Breast_Lee05_71genes_up_V03	10.0	1.0	0.045629555	0.045
Human Lung_Coldren06_110genes	18.0	1.0	0.073610154	0.082
Human Breast_Chang04_772genes_WoundHealing_CSIR_genes	122.0	2.0	0.5637574	0.119
Human Breast_Desmaisons05_229genes_ALPK4_Module	37.0	1.0	0.1707577	0.157
Human Breast_VandenEynden07_250genes-hypoxia_angiogenesis	38.0	1.0	0.17474205	0.174
Human Breast_Thuenen05_512genes_TGFs	64.0	1.0	0.2956481	0.227
Human Breast_Laus05_374genes	71.0	1.0	0.32797715	0.3
Human Liver_Cha09_773genes	183.0	1.0	0.47539365	0.379
Human Colon_Jonsson08_829genes	118.0	1.0	0.54421926	0.411
Human Breast_Charafe-Jauffret06_1309genes	161.0	1.0	0.7419825	0.508
Human Ovarian_Baranova05_307genes	159.0	1.0	0.73494315	0.523
Human Breast_Butcher09_1487genes	190.0	1.0	0.87614447	0.569
Human Breast_Kreike07_3712genes_Basal_mets	495.0	1.0	2.2896047	0.89
Human Breast_Farmer05_3199genes_basal_apocrine_luminal	519.0	1.0	2.3938997	0.911
Human Breast_Ray05_3373genesSuppTable2SuppTable2	531.0	1.0	2.4502487	0.921
Human Breast_Baloz05_80genes	10.0	0.0	0.045327226	1.0
Human Breast_Wang05_230genes_down_BT-BAF57cells	49.0	0.0	0.22593704	1.0
Human Endometrial_Sabesan09_11genes	1.0	0.0	0.004591594	1.0
Human Colon_Aran04_254genes	36.0	0.0	0.13620709	1.0
Human Breast_Baloz05_21genes_inown	2.0	0.0	0.00915403	1.0
Human Endometrial_Matsushima-Hisui01_69genes	4.0	0.0	0.018369551	1.0
Human Colon_Allen09_15genes	2.0	0.0	0.009384379	1.0
Human Breast_Vincard-Carmon08_502genes	157.0	0.0	0.72381	1.0
Human Liver_Lu02_10genes	3.0	0.0	0.013775267	1.0
Human Liver_Lu02_38genes	7.0	0.0	0.03252816	1.0
Human Liver_Lu02_25genes	5.0	0.0	0.023070076	1.0
Human Prostate_Chandran07_102genes	3.0	0.0	0.014215786	1.0
Human Bladder_Osman06_345genes	65.0	0.0	0.30172363	1.0
Human Colon_DellaRagione01_21genes	7.0	0.0	0.032405667	1.0
Human Prostate_Chandran07_154genes	23.0	0.0	0.1063471	1.0
Human Prostate_Chandran07_209genes	22.0	0.0	0.10114374	1.0
Human Bladder_Osman06_345genes	65.0	0.0	0.30172363	1.0
Mouse Breast_Wright08_339genes	1.0	0.0	0.004591591	1.0
Human Colon_dPietro05_100genes	18.0	0.0	0.08312415	1.0
Human Colon_dPietro05_54genes	10.0	0.0	0.045116248	1.0

Gene signatures, published in GeneSigDb, with enrichment in the list of selected genes. Future plans include adding links from this display directly to the gene signature web page, where the list of genes in the signature and the source publication can be found.

From here, you can continue examining gene signatures of interest by searching the GeneSigDb website, or continue on with another analysis by simply selecting it from one of the drop-down menus. For this pilot, most of the standard MeV modules are available to use. A few of them, like the EASE and GSEA modules, require specific annotation files that are currently only available for DNA microarray data. Part of the full RNASeq implementation project will be to adapt MEV to fully support RNASeq analysis in all modules. However, that support is not yet available.

Additional Resources

For more information on the different modules, consult the MeV Manual, found in the “documentation” folder, or go to <http://mev.tm4.org>.