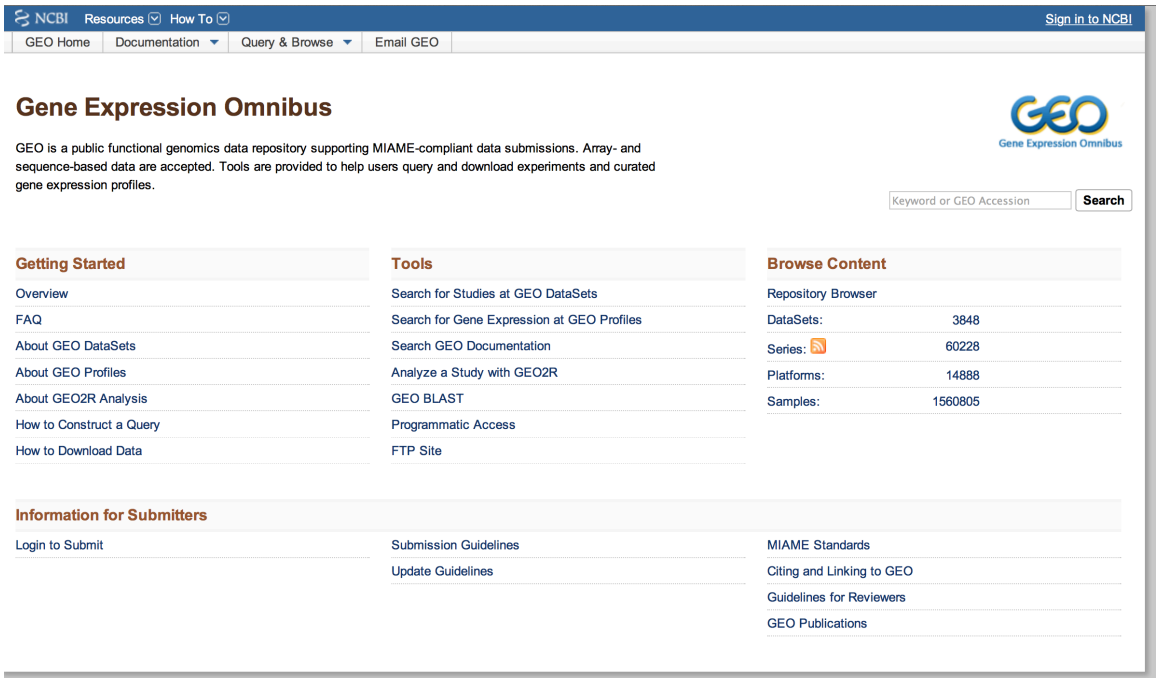
Secondary Analysis of GEO expression data
Short Course on Genetics of Addiction
August 26th, 2015

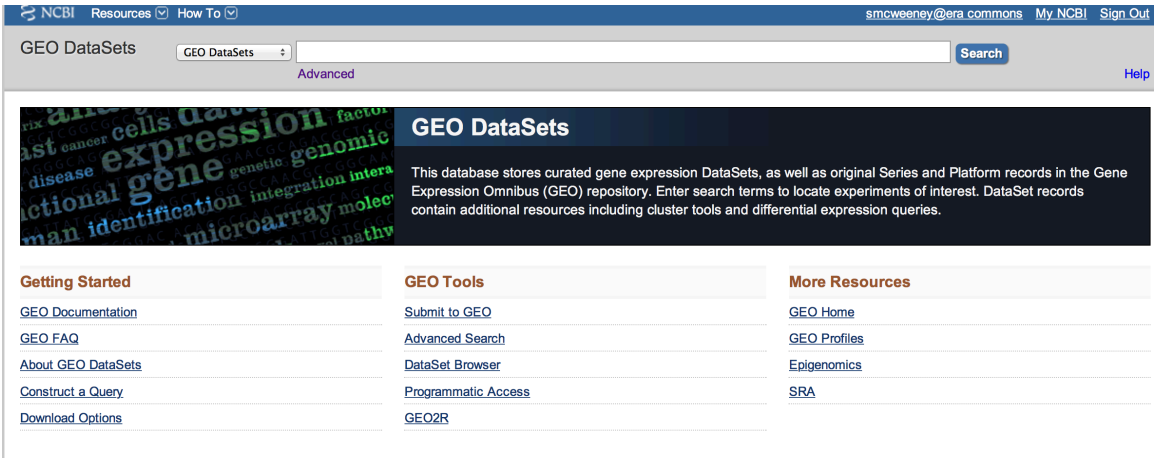
The Gene Expression Omnibus (GEO) is a public repository for functional genomics data (both array and massively parallel sequencing). When you publish functional genomics data, you should submit your data to GEO for others to access. Most journals now require this. As of this workshop, GEO hosts data on over 1.5 million samples. This makes GEO a powerful resource for secondary data analysis of public data to answer new questions or to augment your own primary data. We will highlight how to use GEO and the associated tools provided with it to re-analyze or query public data sets.

Please go to http://www.ncbi.nlm.nih.gov/geo/



Finding a Gene Expression Study

Under tools, Click on “Search for Studies at GEO DataSets”. GEO hosts both curated data sets as well as the associated series and platform records. By searching the data sets, you can use key words or search terms to find data sets of interest.



For this example, we will use the following search terms:

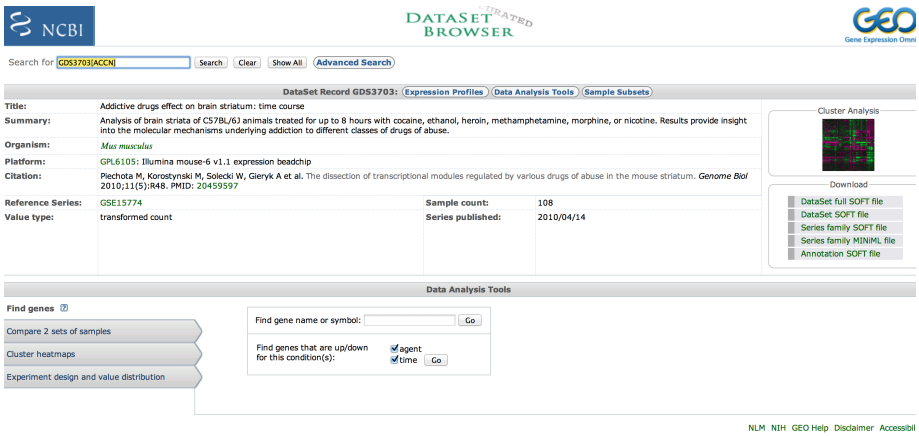
mouse[organism] AND (cocaine)

After we press the Search button, this simple search that we have done is automatically translated by GEO to:

("Mus"[Organism] OR "Mus musculus"[Organism]) AND ("cocaine"[MeSH Terms] OR cocaine[All Fields])

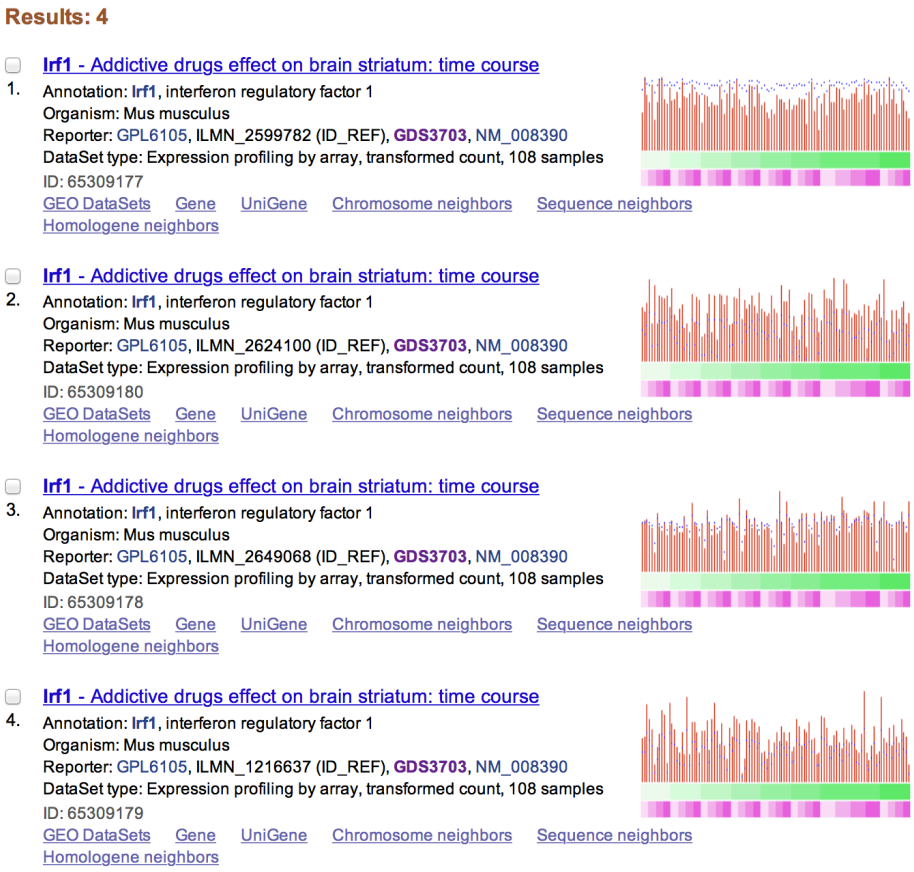
It will return all GEO data sets where organism noted as Mouse or Mus Musculus and cocaine appears either the MeSH Term or any field are returned.

In the list of returned results, find a study named "**Addictive drugs effect on brain striatum: time course**". This is a time series that examines gene expression changes after treatment with different drugs (cocaine, ethanol, heroin, methamphetamine, morphine, or nicotine). Click on the hyperlink for the name of the study. This will take you to the Curated Dataset Browser.



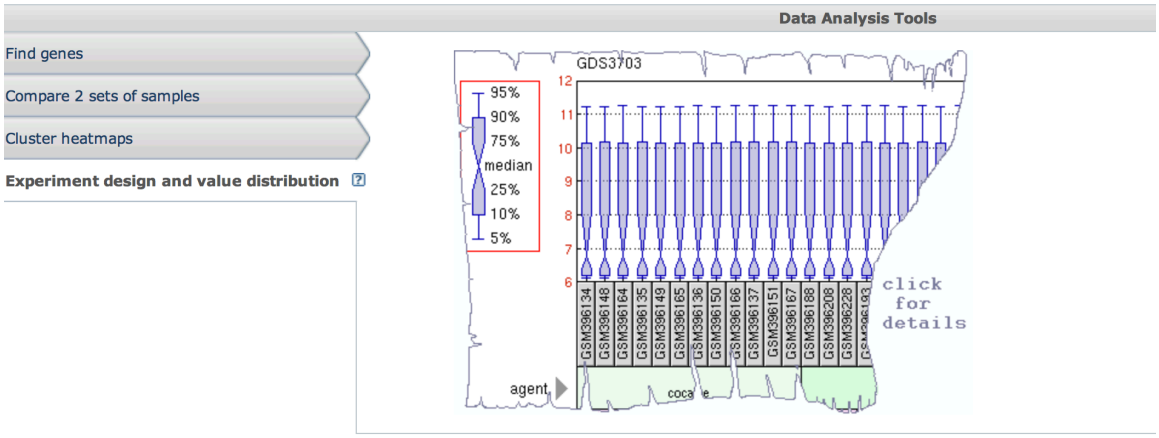
Querying a Study

At the bottom of the page, under Data Analysis Tools, you will see that you can enter any gene of interest to see the expression of that gene in this study. If you enter "IRF1", you will see that you get back several entries. Why is this the case?

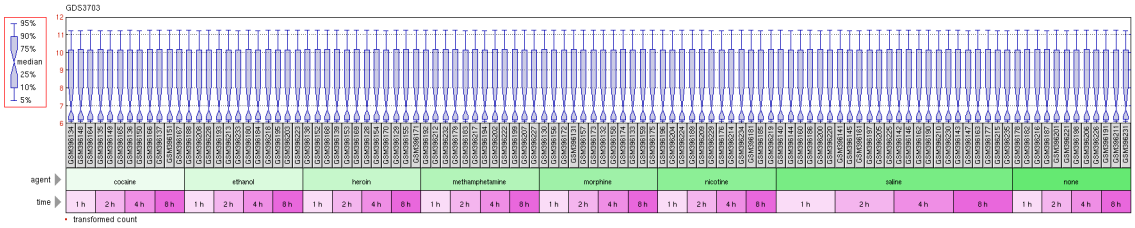


Overview of the Experiment

Back on the Curated data Browser, click on "Experiment design and value distribution". You will see a snapshot of the distribution plot. Click on the image to see the full details.

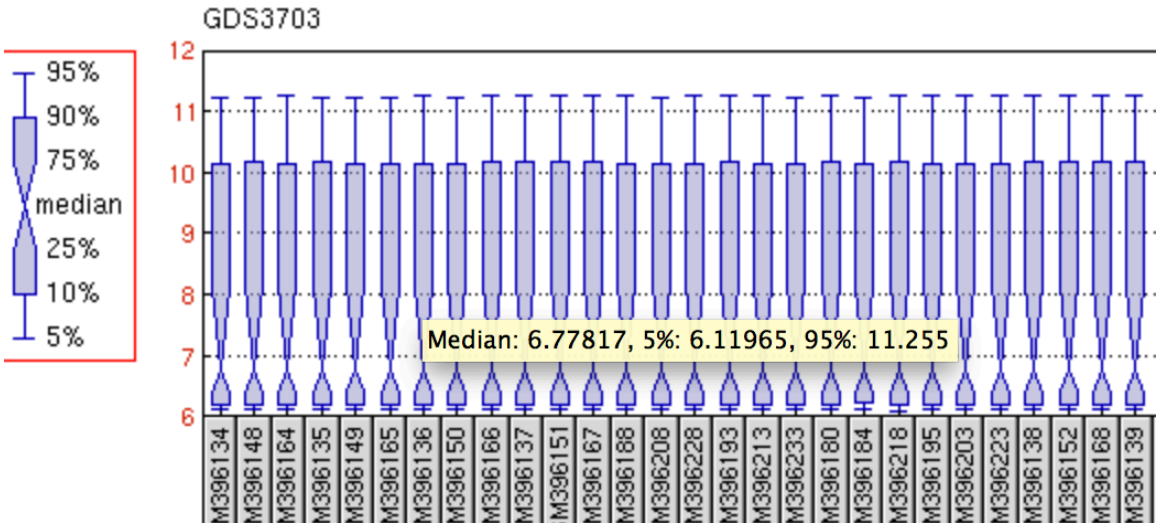


This will allow you to see the expression distribution for every sample in the study, as well as the annotation for treatment and time point. This study has 4 time points (1, 2, 4 and 8 hours) and 8 treatments (6 drugs, and 2 controls (saline, none)).

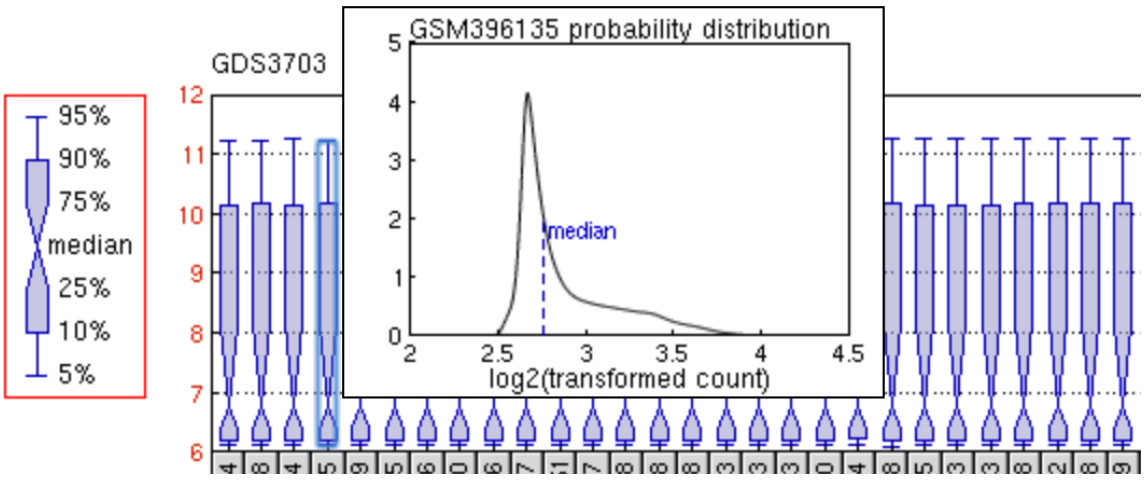


You should note the values of the expression data and distributions across the samples. This data is log2 normalized expression using quantile normalization. How would raw expression differ from this?

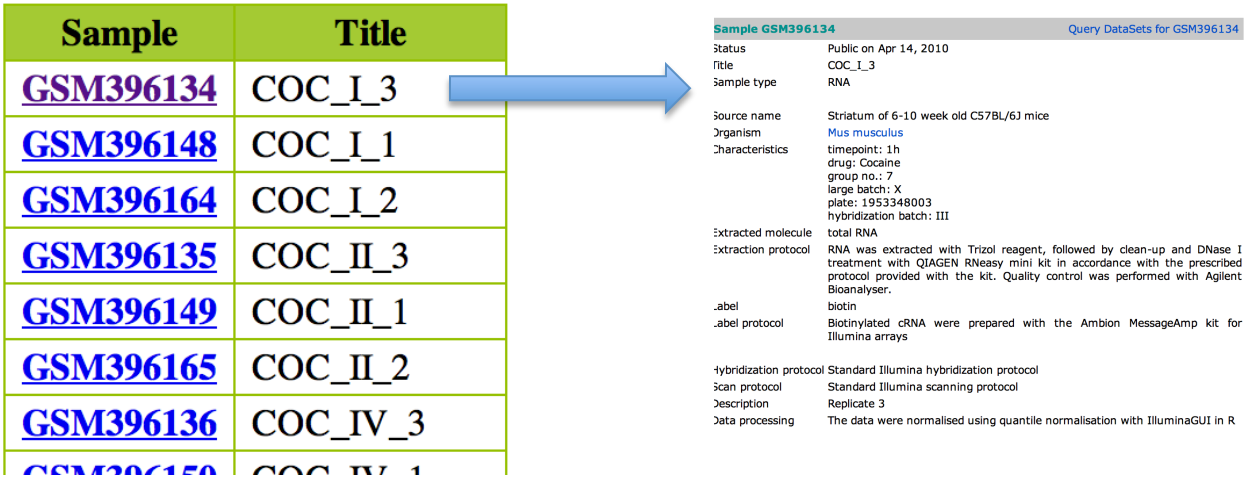
If you mouse over a sample, it will give you the summary statistics (median and tails) for that sample.



If you click on the sample, it will compute the histogram.

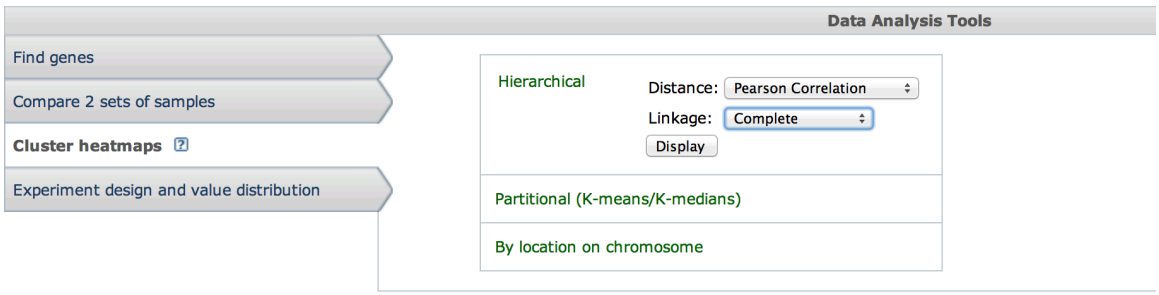


The sample table, allows you to click on the hyperlink for each sample to view the associated details.

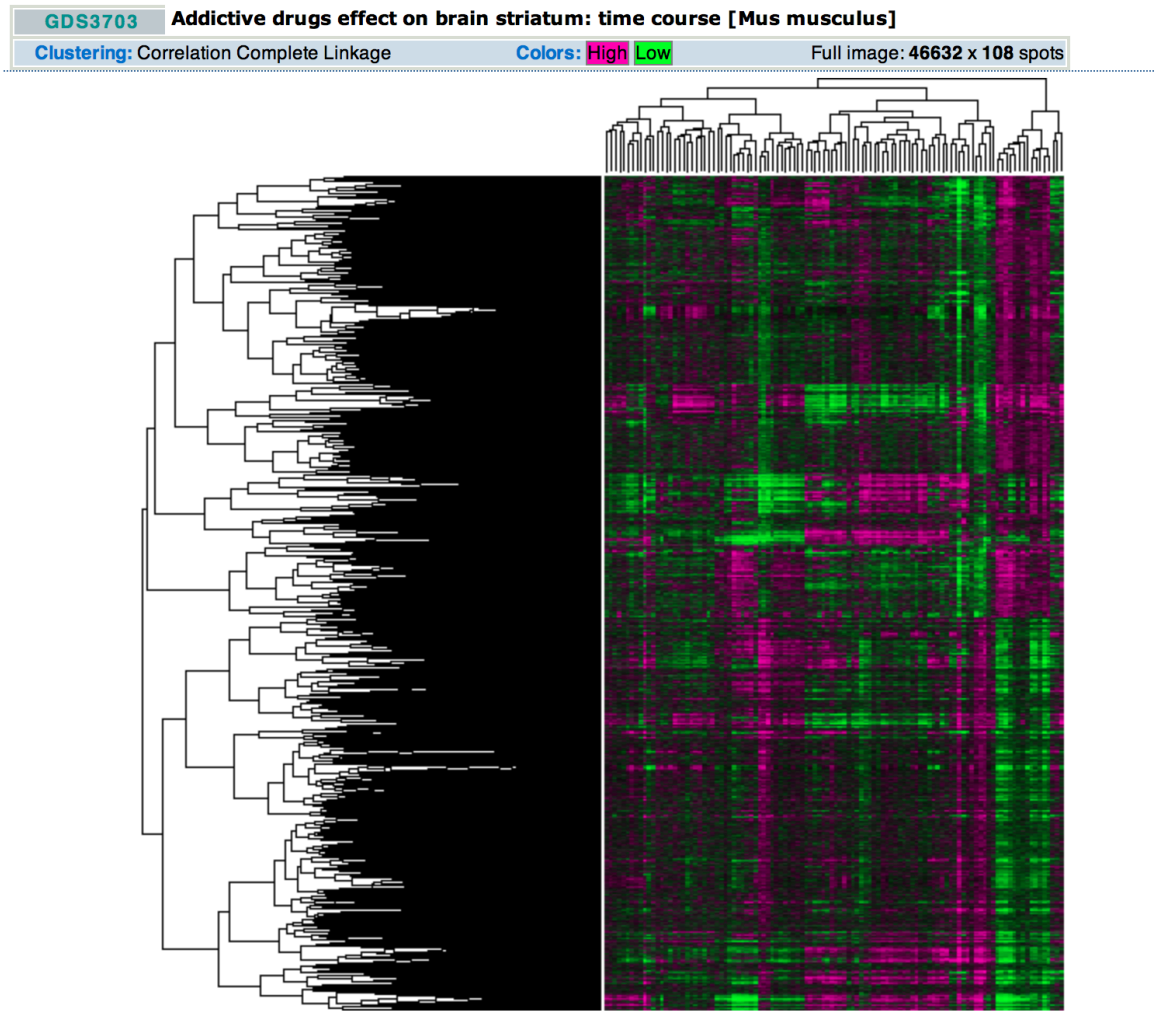


Visualizing Gene Expression: Heatmaps

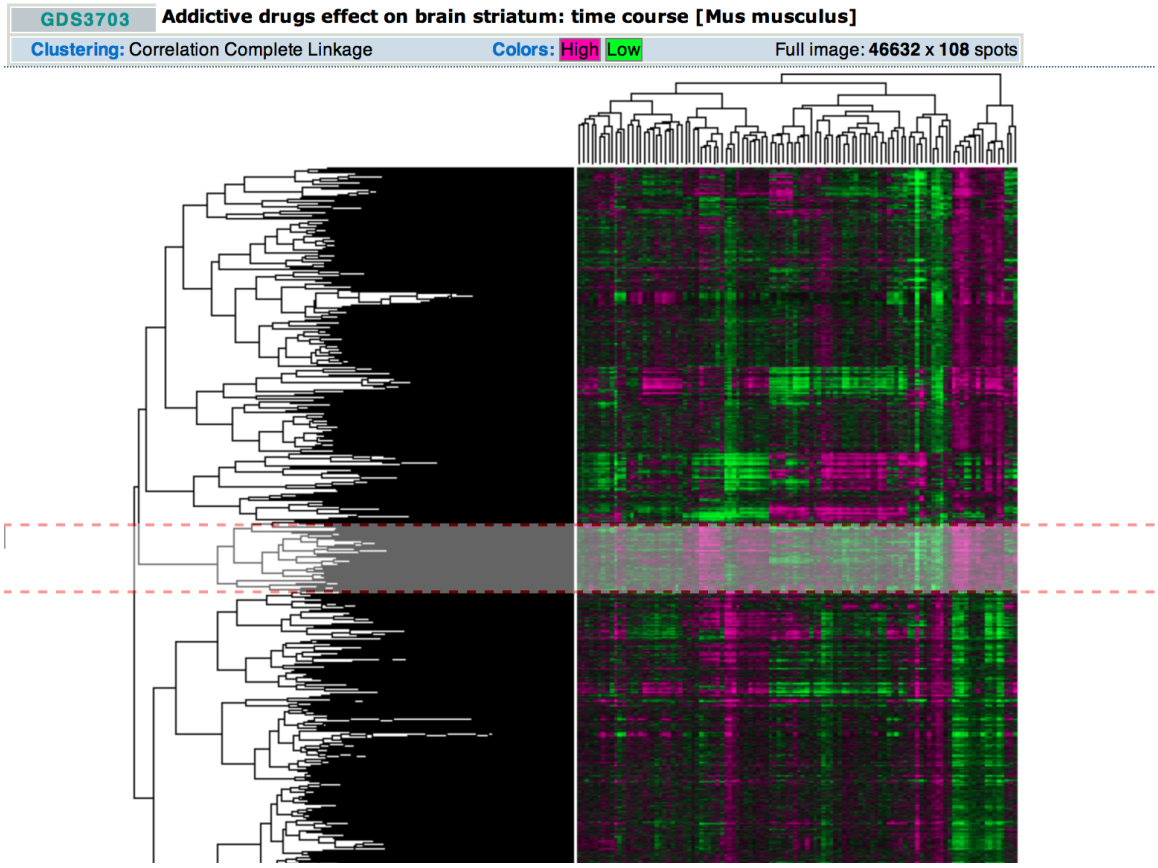
If you return to the Curated DataSet Browser page, click on Cluster heatmaps button under Data Analysis Tools. Clustering can be an effective way to visualize the data.



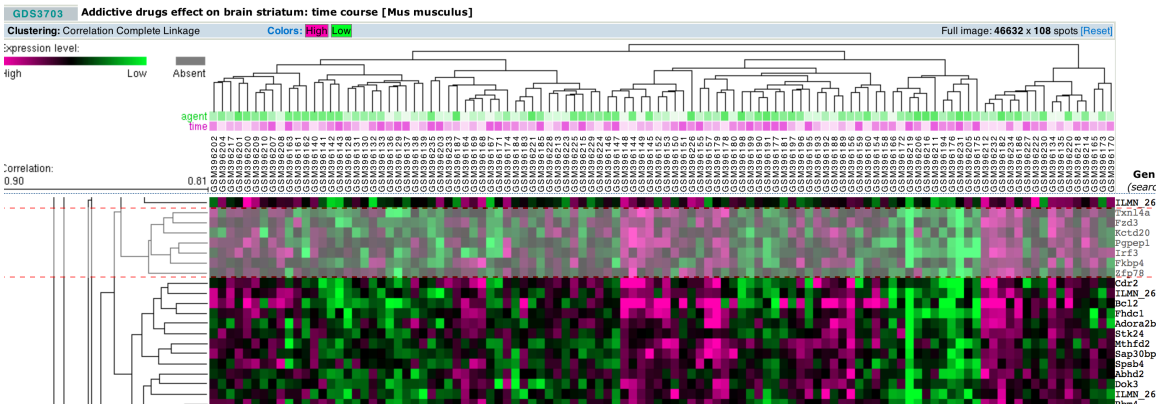
Your choice of distance measure and algorithm are key. For this example, we select Pearson Correlation and Complete Linkage. You should examine other parameters to see how it impacts the clusters (for example Euclidean versus Correlation). After you click the "Display" button, the heatmap is generated.

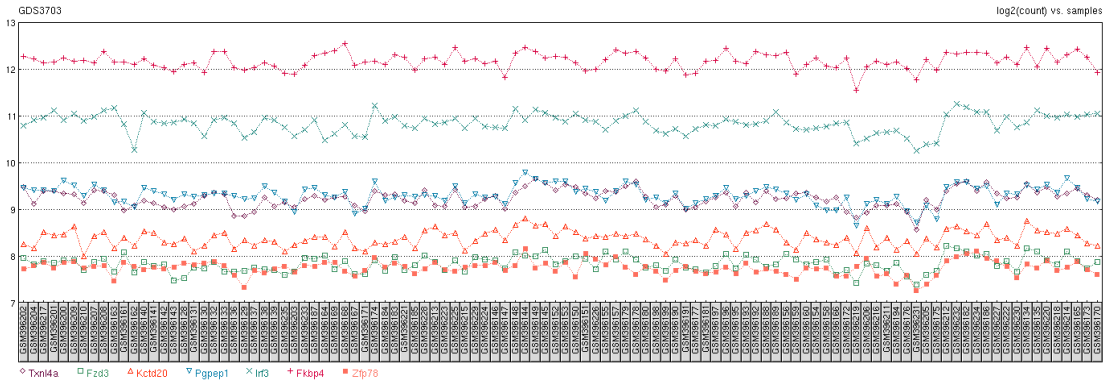


You can select a region of interest and then click “Display values” to see the expression for that subset of genes.



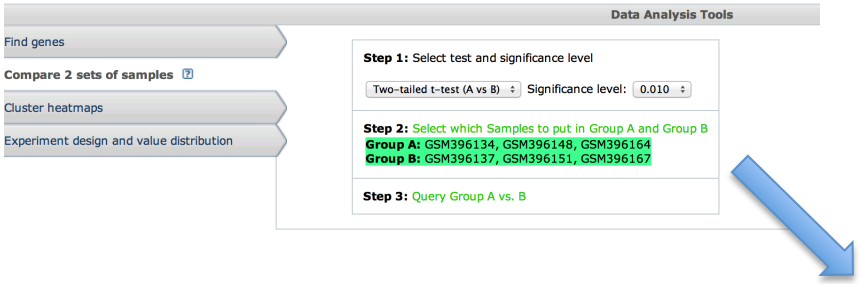
After you click plot values, you will see that you have too many genes (most likely). If you then click “Show heatmap region”, you will get a blow-up of the heatmap. Select a smaller subset of genes and select “Plot values”. This will allow you to view the expression of these genes across the samples from the cluster you identified. Note you can download this data or view the genes in Entrez for more information.





Conducting Differential Expression (Approach 1)

Finally, we can examine differential expression under the Compare 2 sets of Samples tab under Data Analysis Tools on the Curated Dataset Browser main page.
You will need to select which groups you want to compare. For this example, we examine 1 hour vs. 8 hour gene expression in the cocaine treated animals.



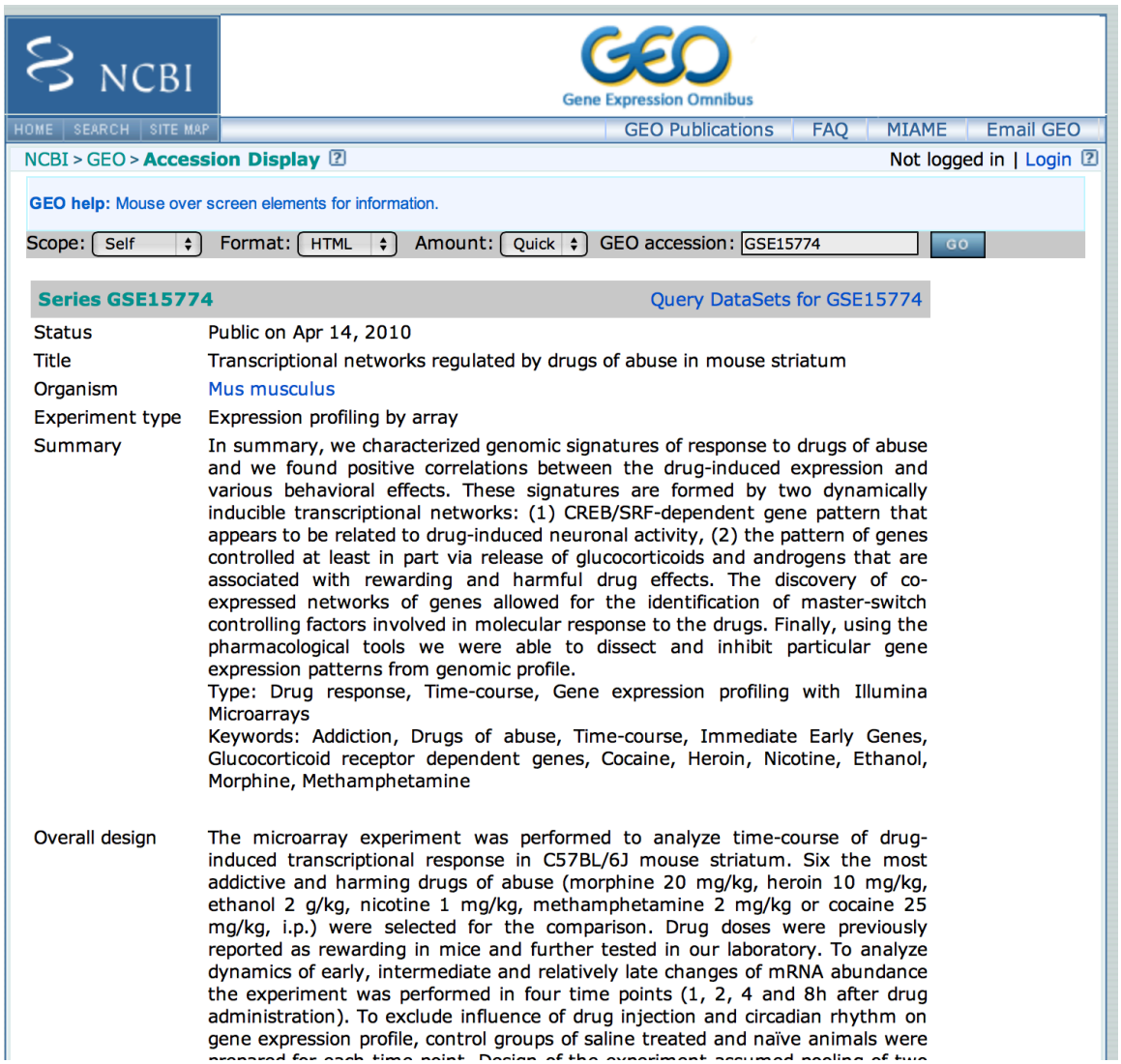
Click on accessions to select samples individually, click on colored blocks and then on blinking arrows to select groups of samples.

Samples, Group A	Factors		Samples, Group B
	agent	time	
GSM396134	cocaine	1 h	GSM396134
GSM396148			GSM396148
GSM396164			GSM396164
GSM396135		2 h	GSM396135
GSM396149			GSM396149
GSM396165			GSM396165
GSM396136		4 h	GSM396136
GSM396150			GSM396150
GSM396166			GSM396166
GSM396137		8 h	GSM396137
GSM396151			GSM396151
GSM396167			GSM396167
GSM396188	ethanol	1 h	GSM396188
GSM396208			GSM396208
GSM396228			GSM396228
GSM396193		2 h	GSM396193
GSM396213			GSM396213
GSM396233			GSM396233
GSM396180		4 h	GSM396180
GSM396184			GSM396184
GSM396218			GSM396218
GSM396195		8 h	GSM396195
GSM396203			GSM396203
GSM396223			GSM396223
GSM396138		1 h	GSM396138
GSM396152			GSM396152

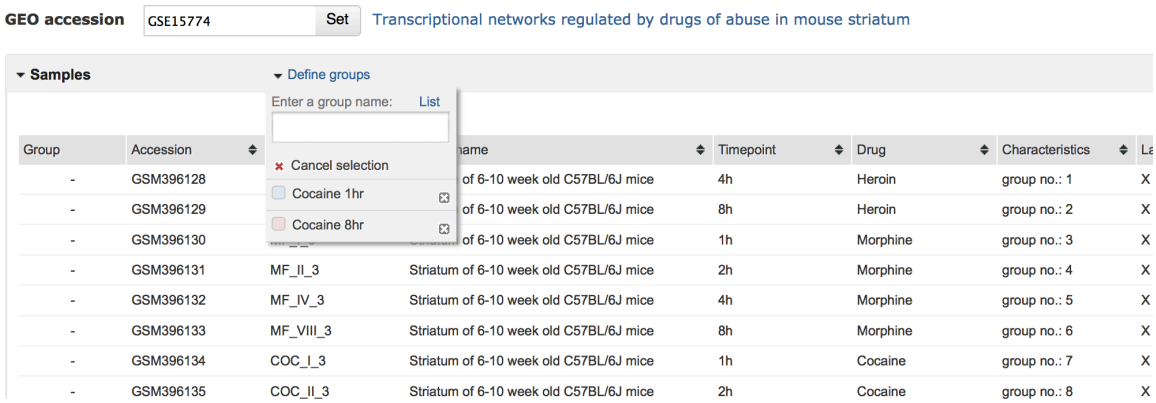
When you select Query Group A vs. B, it returns 652 genes that are differentially expressed. Note: the genes are returned as profiles. You can download this data or find pathways for them. Note the profile pathways functionality can be very slow depending upon the number of genes. This is not the most robust way to perform differential expression analysis so we will utilize the Geo2R package instead.

Conducting Differential Expression (Approach 2)

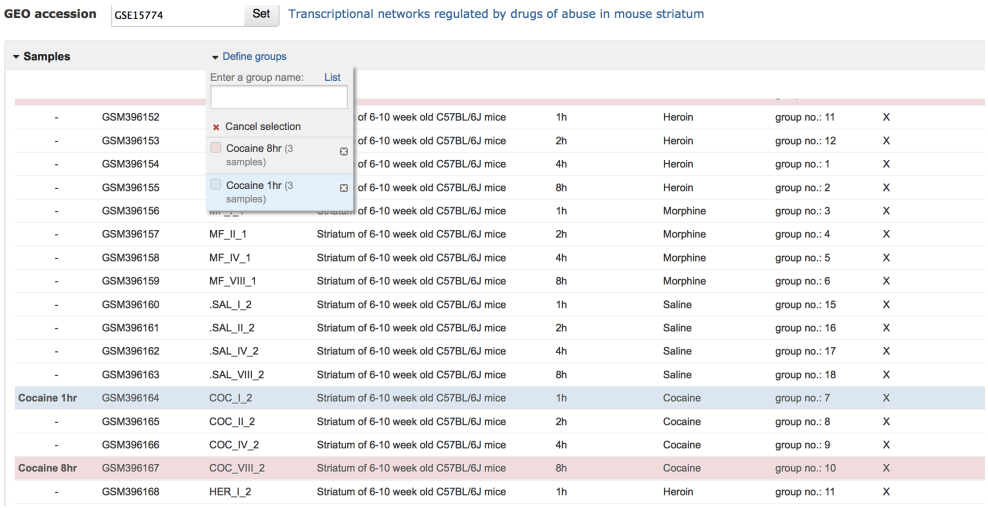
On the main page for the Curated Dataset Browser, you will see a reference series for this dataset: GSE15774. Click on that hyperlink. This will take you to the GEO entry for this series.



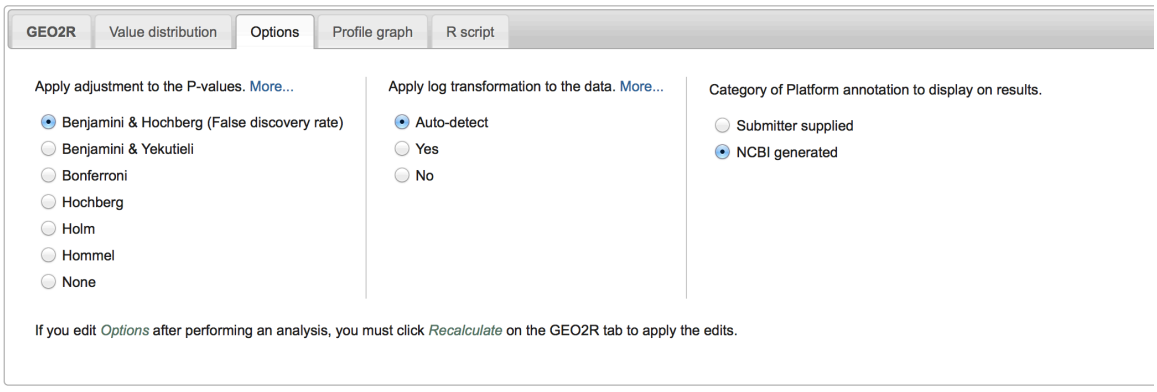
If you scroll to the bottom of the page, you will see a link for **Analyze with Geo2R**. Click on that and you will see that this series has been populated in the Search menu. Click on the Set button and all of the samples will be displayed. Click on Define groups and give labels for the groups you want to define.



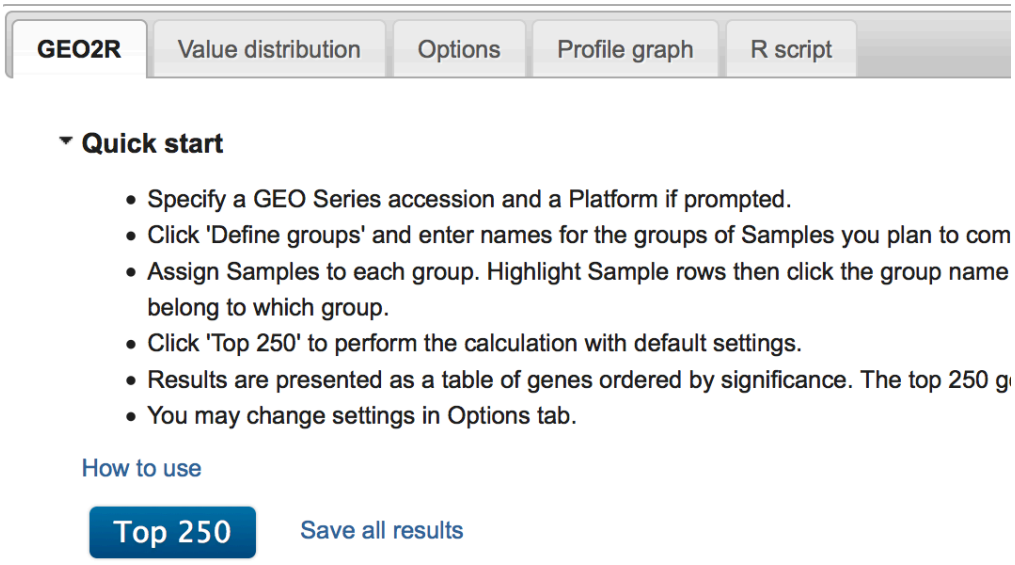
Then select the samples you want to add to each group. Click on a sample or samples, then select the group to which it should be assigned.



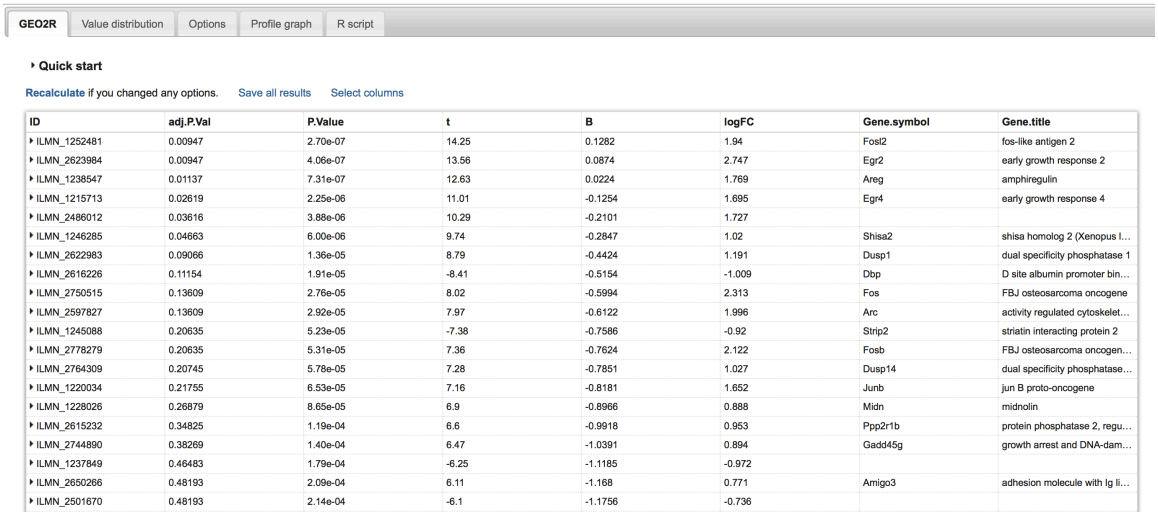
At the bottom, on the option tab, you can select the method used for adjusting the p-values for multiple testing. We will use Benjamini and Hochberg for this example. You should examine how other types of adjustments or no adjustment changes the results.



On the Geo2R tab, click on Top 250.

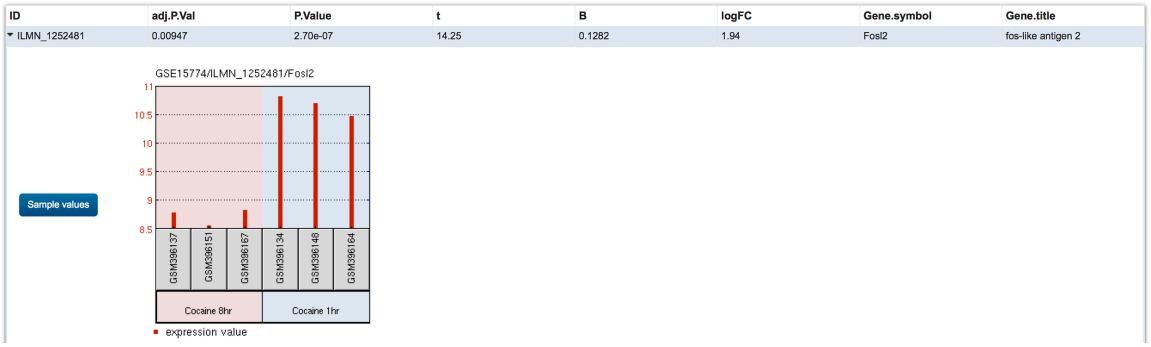


You will be presented back with the table of results. This includes the following statistics: ID (this case probe-set ID), adjusted P-value, Raw P-value, the Moderated t-statistic, the B-statistic (a log-odds that the gene is differentially expressed), the log2 Fold Change, Gene Symbol and Gene name. There is often confusion about the B-statistic. If B =1.5, the odds of differential expression is exp(1.5)=4.48, i.e., about four and a half to one. The probability that the gene is differentially expressed is 4.48/(1+4.48)=0.82, i.e., the probability is about 82% that this gene is differentially expressed. A B-statistic of zero corresponds to a 50-50 chance that the gene is differentially expressed.

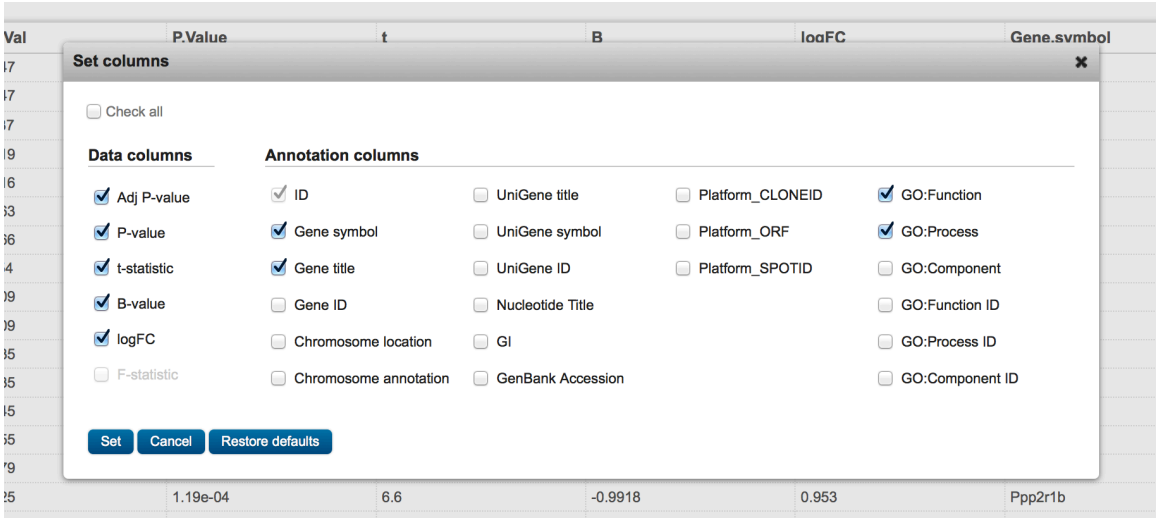


ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
ILMN_1252481	0.00947	2.70e-07	14.25	0.1282	1.94	Fosl2	fos-like antigen 2
ILMN_2623984	0.00947	4.06e-07	13.56	0.0874	2.747	Egr2	early growth response 2
ILMN_1238547	0.01137	7.31e-07	12.63	0.0224	1.769	Areg	amphiregulin
ILMN_1215713	0.02619	2.25e-06	11.01	-0.1254	1.695	Egr4	early growth response 4
ILMN_2486012	0.03616	3.88e-06	10.29	-0.2101	1.727		
ILMN_1246285	0.04663	6.00e-06	9.74	-0.2847	1.02	Shisa2	shisa homolog 2 (Xenopus l…
ILMN_2622983	0.09066	1.36e-05	8.79	-0.4424	1.191	Dusp1	dual specificity phosphatase 1
ILMN_2616226	0.11154	1.91e-05	-8.41	-0.5154	-1.009	Dbp	D site albumin promoter bin…
ILMN_2750515	0.13609	2.76e-05	8.02	-0.5994	2.313	Fos	FBJ osteosarcoma oncogene
ILMN_2597827	0.13609	2.92e-05	7.97	-0.6122	1.996	Arc	activity regulated cytoskelet…
ILMN_1245088	0.20635	5.23e-05	-7.38	-0.7586	-0.92	Strip2	striatin interacting protein 2
ILMN_2778279	0.20635	5.31e-05	7.36	-0.7624	2.122	Fosb	FBJ osteosarcoma oncogen…
ILMN_2764309	0.20745	5.78e-05	7.28	-0.7851	1.027	Dusp14	dual specificity phosphatase…
ILMN_1220034	0.21755	6.53e-05	7.16	-0.8181	1.652	Junb	jun B proto-oncogene
ILMN_1228026	0.26879	8.65e-05	6.9	-0.8966	0.888	Midn	midnolin
ILMN_2615232	0.34825	1.19e-04	6.6	-0.9918	0.953	Ppp2r1b	protein phosphatase 2, regu…
ILMN_2744890	0.38269	1.40e-04	6.47	-1.0391	0.894	Gadd45g	growth arrest and DNA-dam…
ILMN_1237849	0.46483	1.79e-04	-6.25	-1.1185	-0.972		
ILMN_2650266	0.48193	2.09e-04	6.11	-1.168	0.771	Amigo3	adhesion molecule with Ig li…
ILMN_2501670	0.48193	2.14e-04	-6.1	-1.1756	-0.736		

If you click on an individual gene, it will show you the expression distribution across the samples in the 2 groups.



If you click on select columns, you can add or remove columns from the table.



ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title	GO.Function	GO.Process
ILMN_1252481	0.00947	2.70e-07	14.25	0.1282	1.94	Fosl2	fos-like antigen 2	DNA binding///RNA p…	positive regulation of f…
ILMN_2623984	0.00947	4.06e-07	13.56	0.0874	2.747	Egr2	early growth response 2	DNA binding///HMG b…	Schwann cell differen…
ILMN_1238547	0.01137	7.31e-07	12.63	0.0224	1.769	Areg	amphiregulin	cytokine activity///epid…	G-protein coupled rec…
ILMN_1215713	0.02619	2.25e-06	11.01	-0.1254	1.695	Egr4	early growth response 4	DNA binding///RNA p…	cellular response to c…
ILMN_2486012	0.03616	3.88e-06	10.29	-0.2101	1.727				
ILMN_1246285	0.04663	6.00e-06	9.74	-0.2847	1.02	Shisa2	shisa homolog 2 (Xen…		multicellular organism…
ILMN_2622983	0.09066	1.36e-05	8.79	-0.4424	1.191	Dusp1	dual specificity phosp…	MAP kinase tyrosine/…	cell cycle///dephosph…
ILMN_2616226	0.11154	1.91e-05	-8.41	-0.5154	-1.009	Dbp	D site albumin promot…	DNA binding///RNA p…	circadian rhythm///po…
ILMN_2750515	0.13609	2.76e-05	8.02	-0.5994	2.313	Fos	FBJ osteosarcoma on…	DNA binding///DNA bi…	SMAD protein signal t…
ILMN_2597827	0.13609	2.92e-05	7.97	-0.6122	1.996	Arc	activity regulated cyto…	actin binding///protein…	anterior/posterior patt…
ILMN_1245088	0.20635	5.23e-05	-7.38	-0.7586	-0.92	Strip2	striatin interacting pro…	molecular_function	cell migration///cytosk…
ILMN_2778279	0.20635	5.31e-05	7.36	-0.7624	2.122	Fosb	FBJ osteosarcoma on…	DNA binding///double…	cellular response to c…
ILMN_2764309	0.20745	5.78e-05	7.28	-0.7851	1.027	Dusp14	dual specificity phosp…	MAP kinase tyrosine/…	dephosphorylation///p…
ILMN_1220034	0.21755	6.53e-05	7.16	-0.8181	1.652	Junb	jun B proto-oncogene	DNA binding///DNA bi…	cellular process///cell…

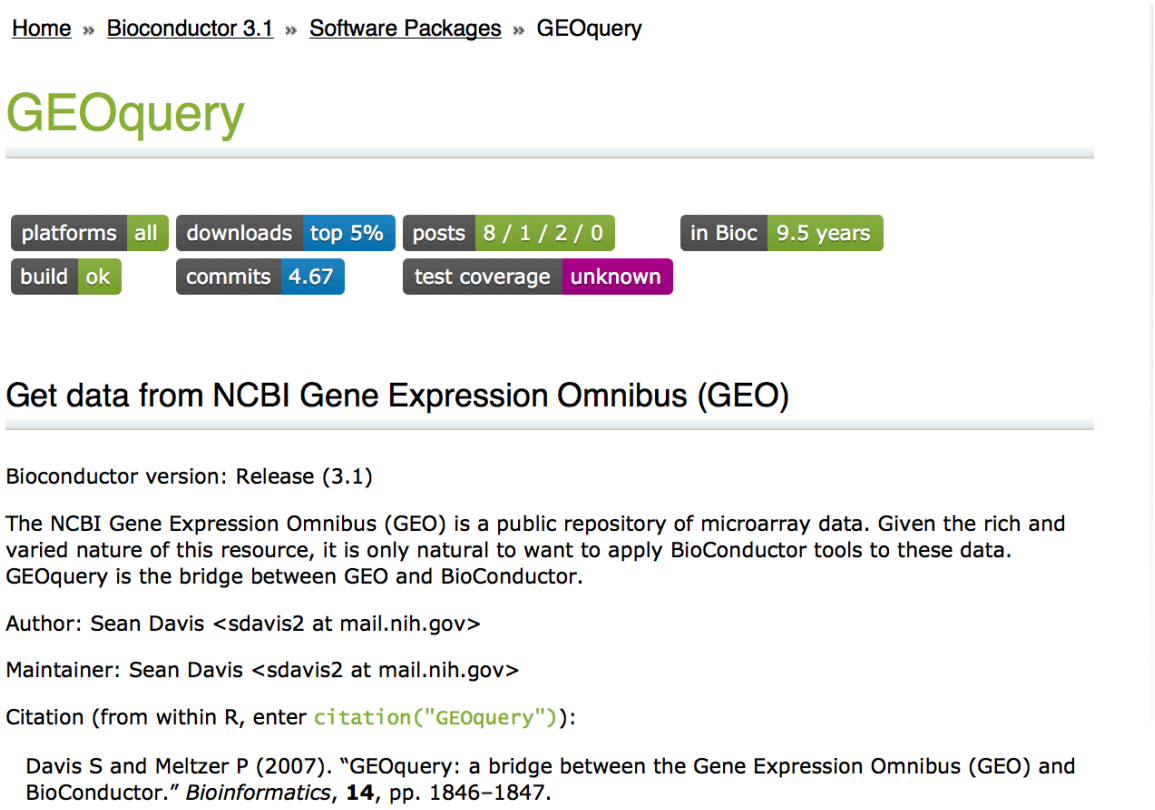
You can save this table by clicking on “Save All Results”.

Integrating GEO into your analysis workflow

If you have experience with the statistical programming language R, you can do all of this from the R environment.

The R script to retrieve the data and perform DE is included (ExampleGeo2Rlab.R).

You will need to install the R package, GEOQuery from Bioconductor (bioconductor.org) as well as other packages for analysis (such as Biobase, Limma etc).



Home » Bioconductor 3.1 » Software Packages » GEOquery

GEOquery

platforms all | downloads top 5% | posts 8 / 1 / 2 / 0 | in Bioc 9.5 years

build ok | commits 4.67 | test coverage unknown

Get data from NCBI Gene Expression Omnibus (GEO)

Bioconductor version: Release (3.1)

The NCBI Gene Expression Omnibus (GEO) is a public repository of microarray data. Given the rich and varied nature of this resource, it is only natural to want to apply BioConductor tools to these data. GEOquery is the bridge between GEO and BioConductor.

Author: Sean Davis <sdavis2 at mail.nih.gov>

Maintainer: Sean Davis <sdavis2 at mail.nih.gov>

Citation (from within R, enter `citation("GEOquery")`):

Davis S and Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics*, **14**, pp. 1846–1847.

Caveats about Secondary Analysis of Public Data (Major Points - not exhaustive)

1. It is critical that you review all of the provided metadata and experimental details. This should be done before you develop your analysis plan.
2. If you plan to combine your data with this data or perform meta-analysis across studies, you must examine if the platforms, treatments, tissues etc are comparable. For NGS experiments, you must examine sample and library preparation in particular.
3. You should carefully QA/QC the data and develop an appropriate normalization strategy. Confounders or batch effects can impact the analysis

and interpretation. Normalization across raw data from different experiments does not guarantee that batch effects are eliminated, particularly if not all of the metadata was provided.

4. A common issue is that key metadata may not have been provided which can impact your planned analyses. You can contact the authors to attempt to obtain these details if they are not provided. Your timeline for analysis should include this possibility.
5. Annotation is dynamic. You will see differences between the annotation reported and the current annotation. Track versions and be consistent in the annotation used for all samples.
6. If you plan to use the processed data provided by the authors, it is critical that you have read the methods and understand how the data has been analyzed to that point. For most secondary analysis plans, it is best to re-analyze the data from the raw data so that you are aware of any issues that may impact your question.
7. You must still consider power and sample size for secondary analysis. Spend your time upfront assessing feasibility, finding data sets and determining power before conduct the analysis.
8. Remember that the data were generated to ask a different question and there may be issues of imbalance, bias etc when you attempt to use this data to answer a different question. Replication and validation are key.