

Secondary Analysis of Transcriptomics Data
Short Course on Genetics of Addiction
September 13th 2016

What is Secondary Analysis?

Secondary analysis is when we re-use data that was already generated for another study and for a different purpose. It is important to assess the data to determine if it is appropriate for your question of interest prior to conducting your own analysis.

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.

What is GEO?

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 1.9 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/>

The screenshot shows the homepage of the Gene Expression Omnibus (GEO) at <http://www.ncbi.nlm.nih.gov/geo/>. The top navigation bar includes links for NCBI Resources, How To, GEO Home, Documentation, Query & Browse, and Email GEO. On the right, there are user account links for smcweeney@era-commons, My NCBI, and Sign Out. The main content area features the GEO logo and a brief description of the repository's purpose: "GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles." Below this, there are three main sections: "Getting Started" (with links to Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, and How to Download Data), "Tools" (with links to Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, GEO BLAST, Programmatic Access, and FTP Site), and "Browse Content" (listing Repository Browser, DataSets: 4348, Series: 73099, Platforms: 16343, and Samples: 1923835). At the bottom, there is a section for "Information for Submitters" with links to Login to Submit, Submission Guidelines, Update Guidelines, MIAME Standards, Citing and Linking to GEO, Guidelines for Reviewers, and GEO Publications.

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.

NCBI

Dataset Browser

Search for **GDS3703[ACCN]**

DataSet Record GDS3703: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Addictive drugs effect on brain striatum: time course		
Summary:	Analysis of brain striata of C57BL/6J animals treated for up to 8 hours with cocaine, ethanol, heroin, methamphetamine, morphine, or nicotine. Results provide insight into the molecular mechanisms underlying addiction to different classes of drugs of abuse.		
Organism:	<i>Mus musculus</i>		
Platform:	GPL6105: Illumina mouse-6 v1.1 expression beadchip		
Citations:	Piechota M, Korostynski M, Solecki W, Gieryk A et al. The dissection of transcriptional modules regulated by various drugs of abuse in the mouse striatum. <i>Genome Biol</i> 2010;11(5):R48. PMID: 20459597 Korostynski M, Piechota M, Dzbejk J, Miernaski W et al. Novel drug-regulated transcriptional networks in brain reveal pharmacological properties of psychotropic drugs. <i>BMC Genomics</i> 2013 Sep 8;14:606. PMID: 24010892		
Reference Series:	GSE15774	Sample count:	108
Value type:	transformed count	Series published:	2010/04/14

Cluster Analysis


Data Analysis Tools

Find genes

Compare 2 sets of samples

Cluster heatmaps

Experiment design and value distribution

Find gene name or symbol:

Find genes that are up/down for this condition(s): agent time

NLM NIH GEO Help Disclaimer Accessibility

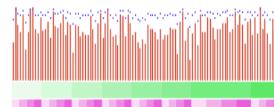
Querying a Study

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?

Results: 4

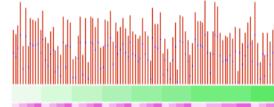
[Irif1 - Addictive drugs effect on brain striatum: time course](#)

1. Annotation: *Irif1*, interferon regulatory factor 1
Organism: *Mus musculus*
Reporter: *GPL6105*, *ILMN_2599782* (ID_REF), [GDS3703](#), *NM_008390*
DataSet type: Expression profiling by array, transformed count, 108 samples
ID: 65309177
- [GEO DataSets](#) [Gene](#) [UniGene](#) [Chromosome neighbors](#) [Sequence neighbors](#)
[Homologene neighbors](#)



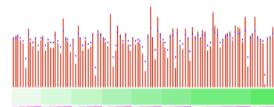
[Irif1 - Addictive drugs effect on brain striatum: time course](#)

2. Annotation: *Irif1*, interferon regulatory factor 1
Organism: *Mus musculus*
Reporter: *GPL6105*, *ILMN_2624100* (ID_REF), [GDS3703](#), *NM_008390*
DataSet type: Expression profiling by array, transformed count, 108 samples
ID: 65309180
- [GEO DataSets](#) [Gene](#) [UniGene](#) [Chromosome neighbors](#) [Sequence neighbors](#)
[Homologene neighbors](#)



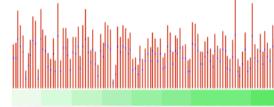
[Irif1 - Addictive drugs effect on brain striatum: time course](#)

3. Annotation: *Irif1*, interferon regulatory factor 1
Organism: *Mus musculus*
Reporter: *GPL6105*, *ILMN_2649068* (ID_REF), [GDS3703](#), *NM_008390*
DataSet type: Expression profiling by array, transformed count, 108 samples
ID: 65309178
- [GEO DataSets](#) [Gene](#) [UniGene](#) [Chromosome neighbors](#) [Sequence neighbors](#)
[Homologene neighbors](#)



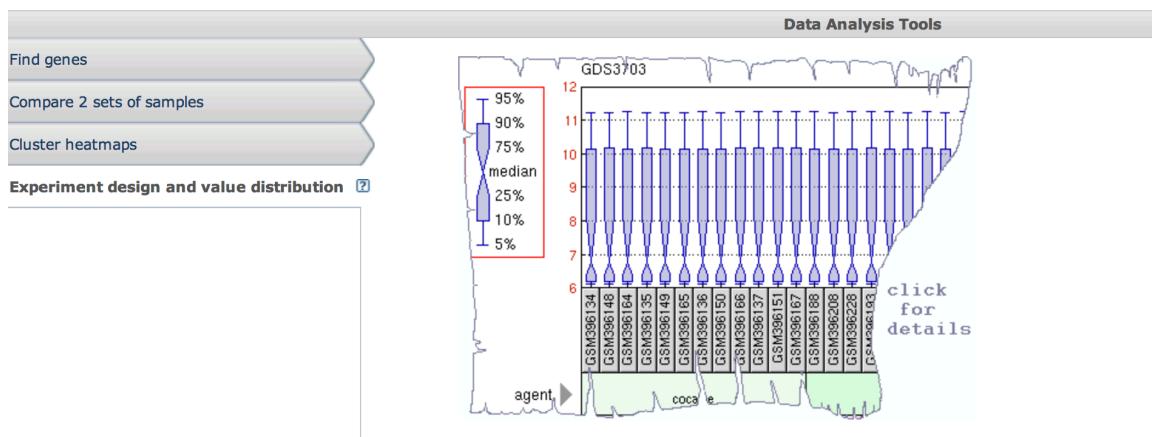
[Irif1 - Addictive drugs effect on brain striatum: time course](#)

4. Annotation: *Irif1*, interferon regulatory factor 1
Organism: *Mus musculus*
Reporter: *GPL6105*, *ILMN_1216637* (ID_REF), [GDS3703](#), *NM_008390*
DataSet type: Expression profiling by array, transformed count, 108 samples
ID: 65309179
- [GEO DataSets](#) [Gene](#) [UniGene](#) [Chromosome neighbors](#) [Sequence neighbors](#)
[Homologene neighbors](#)

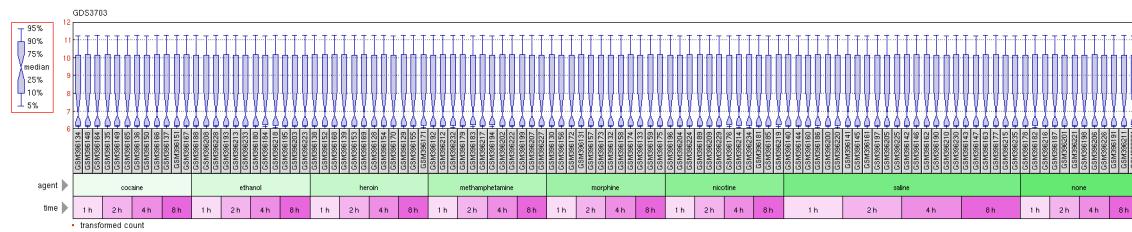


Overview of the Experiment

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.

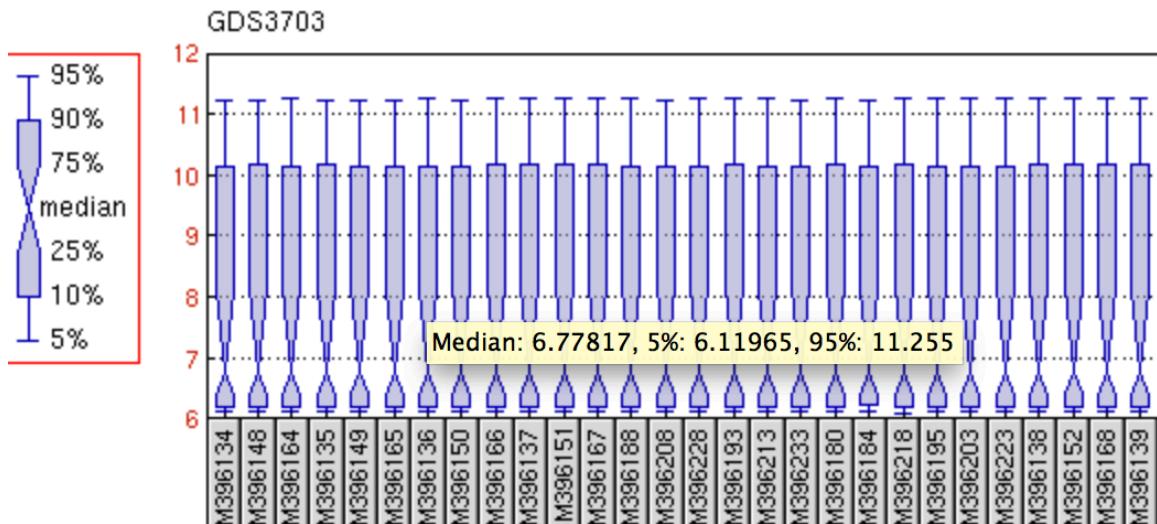


This will allow you to see the expression distribution for every sample in the study, as well as the annotation for treatment and time point. This study has 4 time points (1, 2, 4 and 8 hours) and 8 treatments (6 drugs, and 2 controls (saline, none)).

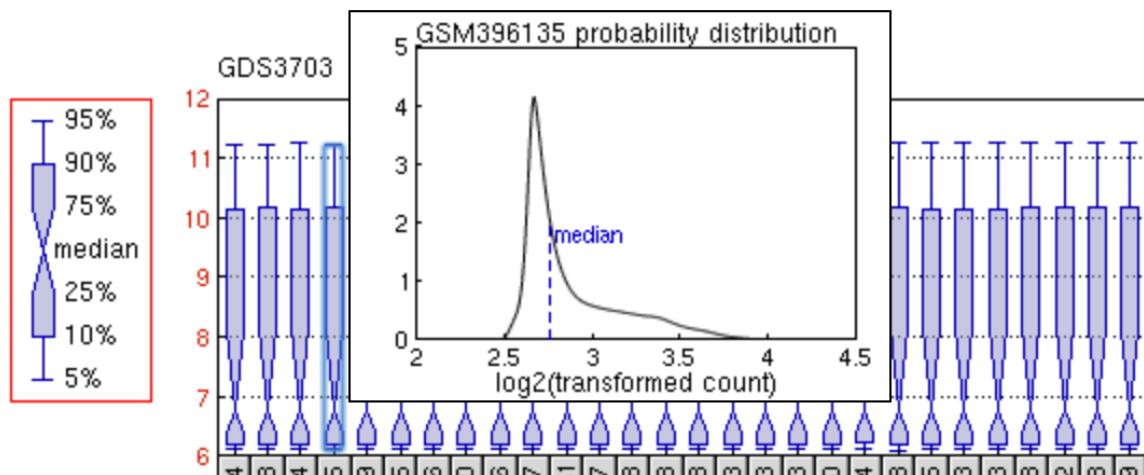


You should note the values of the expression data and distributions across the samples. This data is log2 normalized expression using quantile normalization. How would raw expression differ from this?

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.



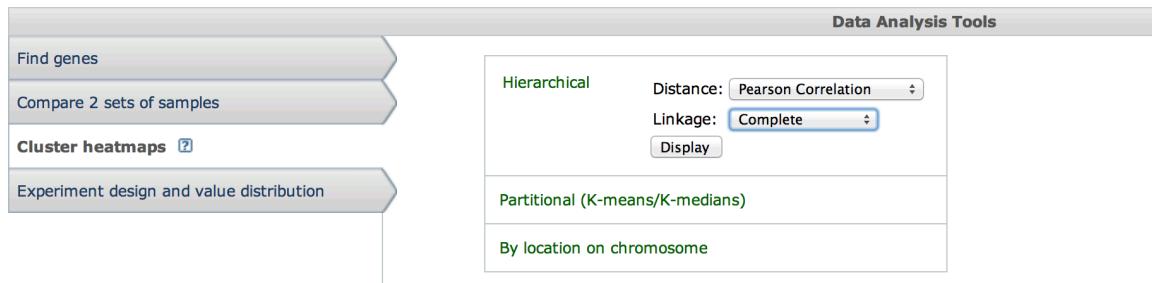
Sample	Title	
GSM396134	COC_I_3	
GSM396148	COC_I_1	
GSM396164	COC_I_2	
GSM396135	COC_II_3	
GSM396149	COC_II_1	
GSM396165	COC_II_2	
GSM396136	COC_IV_3	
GSM396150	COC_IV_1	

Sample GSM396134 [Query DataSets for GSM396134](#)

Status	Public on Apr 14, 2010
Title	COC_I_3
Sample type	RNA
Source name	Striatum of 6-10 week old C57BL/6J mice
Organism	<i>Mus musculus</i>
Characteristics	timepoint: 1h drug: Cocaine group no.: 7 large batch: X plate: 1953348003 hybridization batch: III
Extracted molecule	total RNA
Extraction protocol	RNA was extracted with Trizol reagent, followed by clean-up and DNase I treatment with QIAGEN RNeasy mini kit in accordance with the prescribed protocol provided with the kit. Quality control was performed with Agilent Bioanalyzer.
.ab1	biotin
.ab1 protocol	Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays
Hybridization protocol	Standard Illumina hybridization protocol
Scan protocol	Standard Illumina scanning protocol
Description	Replicate 3
Data processing	The data were normalised using quantile normalisation with IlluminaGUI in R

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.



The screenshot shows the 'Data Analysis Tools' section of the Curated DataSet Browser. On the left, there are four buttons: 'Find genes', 'Compare 2 sets of samples', 'Cluster heatmaps' (which is highlighted with a blue background), and 'Experiment design and value distribution'. On the right, there are two main sections: 'Hierarchical' and 'Partitional (K-means/K-medians)'. Under 'Hierarchical', there are dropdown menus for 'Distance' (set to 'Pearson Correlation') and 'Linkage' (set to 'Complete'), and a 'Display' button. Below these are two more sections: 'Partitional (K-means/K-medians)' and 'By location on chromosome'.

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.

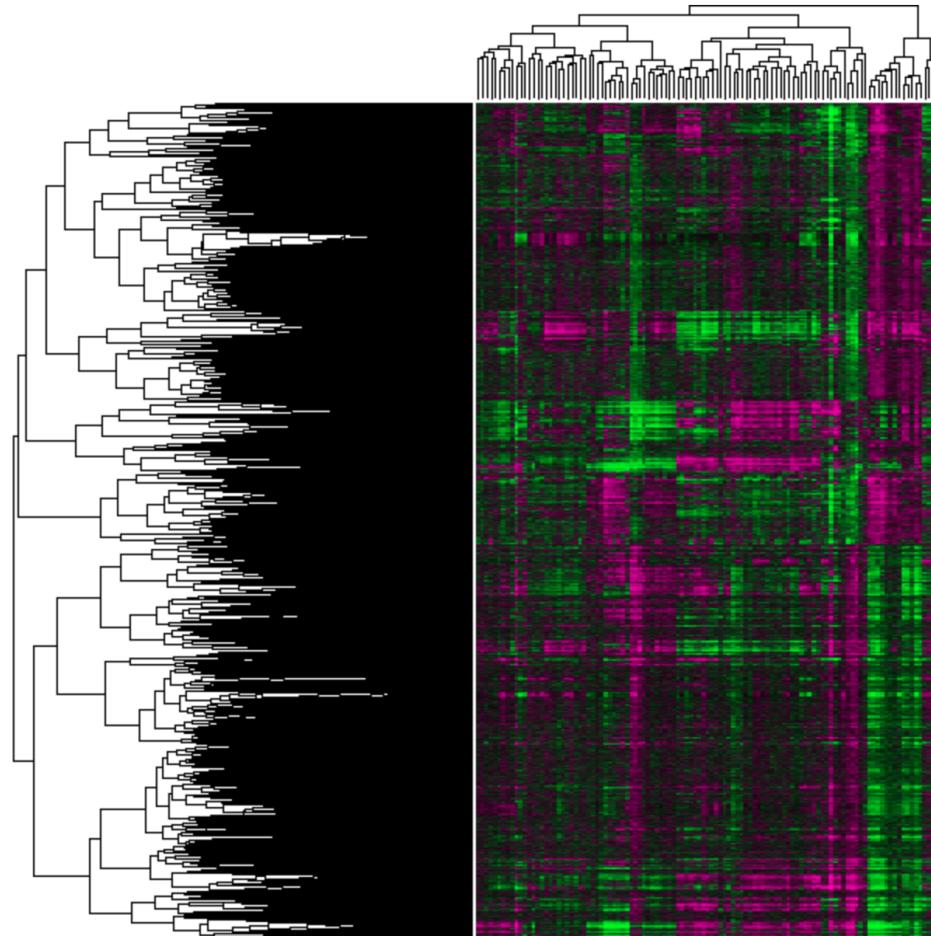
GDS3703

Addictive drugs effect on brain striatum: time course [Mus musculus]

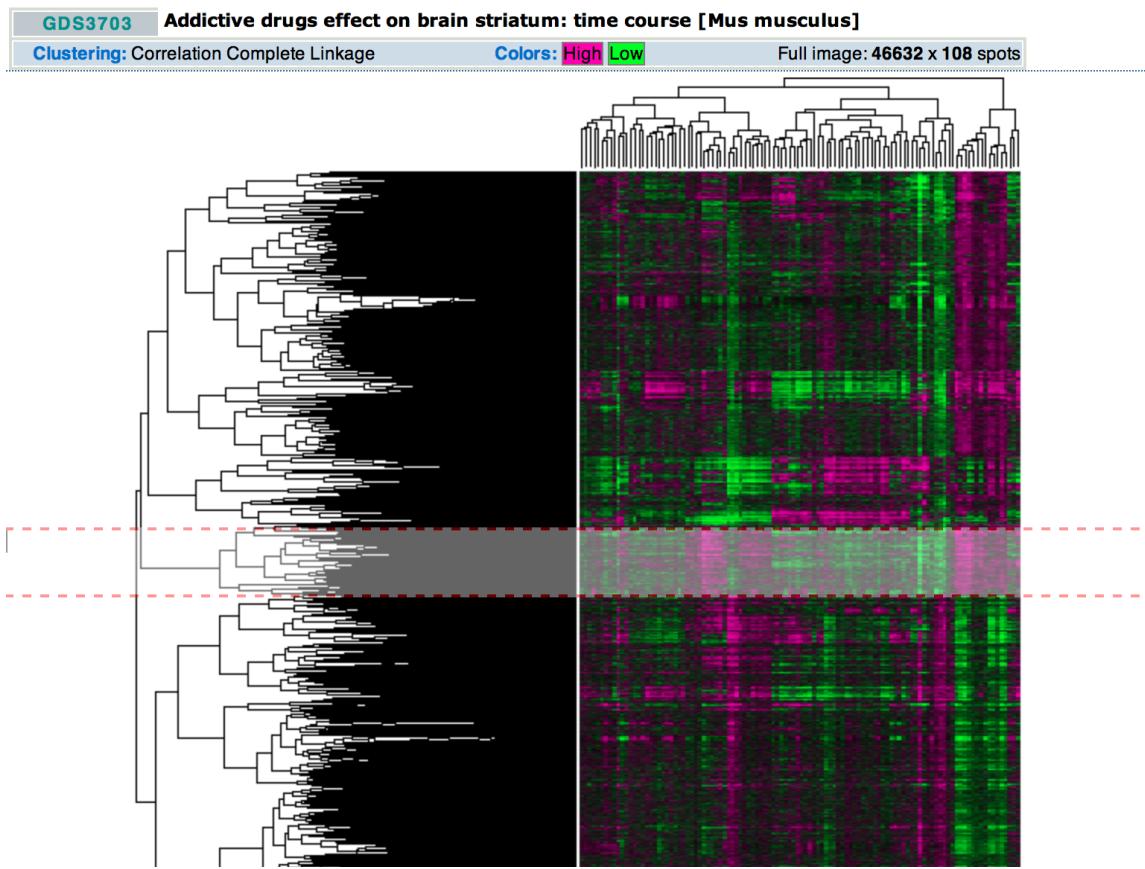
Clustering: Correlation Complete Linkage

Colors: High Low

Full image: 46632 x 108 spots

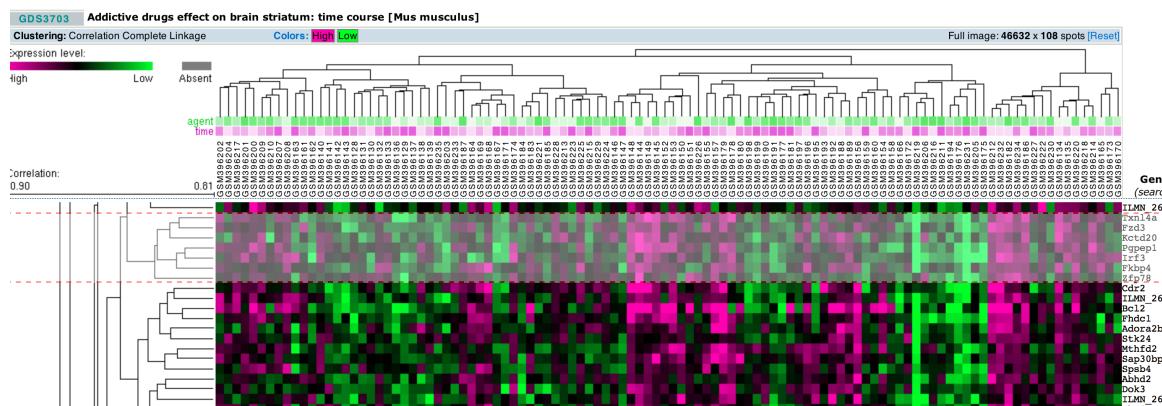


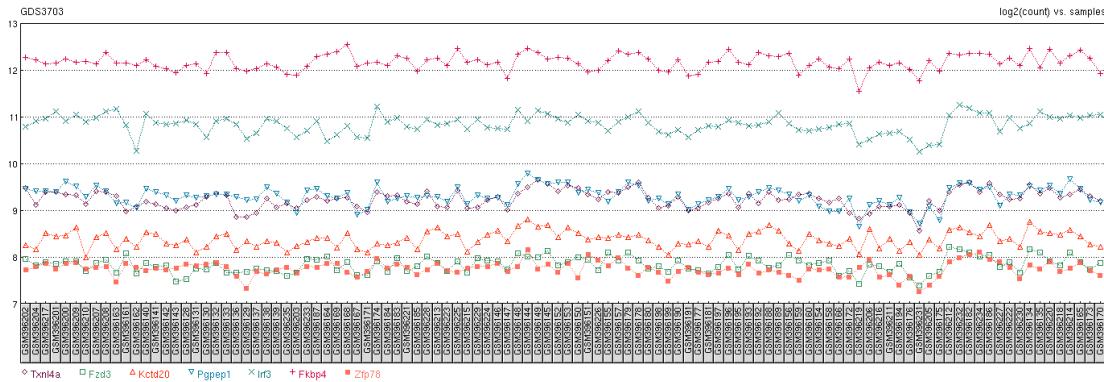
You can select a region of interest and then click “Display values” to see the expression for that subset of genes.



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. **Also you must select your level of significance!**

[Find genes](#)

[Compare 2 sets of samples](#)

[Cluster heatmaps](#)

[Experiment design and value distribution](#)

Step 1: Select test and significance level
 Two-tailed t-test (A vs B) Significance level: 0.010

Step 2: Select which Samples to put in Group A and Group B
 Group A: GSM396134, GSM396148, GSM396164
 Group B: GSM396137, GSM396151, GSM396167

Step 3: Query Group A vs. B

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
GSM396148		GSM396148
GSM396164		GSM396164
GSM396135		GSM396135
GSM396149		GSM396149
GSM396165		GSM396165
GSM396136		GSM396136
GSM396150		GSM396150
GSM396166		GSM396166
GSM396137		GSM396137
GSM396151	GSM396151	
GSM396167	GSM396167	
GSM396188	ethanol	1 h
GSM396208		GSM396208
GSM396228		GSM396228
GSM396193		GSM396193
GSM396213		GSM396213
GSM396233		GSM396233
GSM396180		GSM396180
GSM396184		GSM396184
GSM396218		GSM396218
GSM396195		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.

The screenshot shows the NCBI GEO Accession Display page for dataset GSE15774. The top navigation bar includes links for HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. A user is logged in as "Not logged in | Login". The main content area displays the following details:

Series GSE15774		Query DataSets for GSE15774
Status	Public on Apr 14, 2010	
Title	Transcriptional networks regulated by drugs of abuse in mouse striatum	
Organism	Mus musculus	
Experiment type	Expression profiling by array	
Summary	In summary, we characterized genomic signatures of response to drugs of abuse and we found positive correlations between the drug-induced expression and various behavioral effects. These signatures are formed by two dynamically inducible transcriptional networks: (1) CREB/SRF-dependent gene pattern that appears to be related to drug-induced neuronal activity, (2) the pattern of genes controlled at least in part via release of glucocorticoids and androgens that are associated with rewarding and harmful drug effects. The discovery of co-expressed networks of genes allowed for the identification of master-switch controlling factors involved in molecular response to the drugs. Finally, using the pharmacological tools we were able to dissect and inhibit particular gene expression patterns from genomic profile. Type: Drug response, Time-course, Gene expression profiling with Illumina Microarrays Keywords: Addiction, Drugs of abuse, Time-course, Immediate Early Genes, Glucocorticoid receptor dependent genes, Cocaine, Heroin, Nicotine, Ethanol, Morphine, Methamphetamine	
Overall design	The microarray experiment was performed to analyze time-course of drug-induced transcriptional response in C57BL/6J mouse striatum. Six the most addictive and harming drugs of abuse (morphine 20 mg/kg, heroin 10 mg/kg, ethanol 2 g/kg, nicotine 1 mg/kg, methamphetamine 2 mg/kg or cocaine 25 mg/kg, i.p.) were selected for the comparison. Drug doses were previously reported as rewarding in mice and further tested in our laboratory. To analyze dynamics of early, intermediate and relatively late changes of mRNA abundance the experiment was performed in four time points (1, 2, 4 and 8h after drug administration). To exclude influence of drug injection and circadian rhythm on gene expression profile, control groups of saline treated and naïve animals were prepared for each time point. Design of the experiment assumed pooling of two	

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.

Samples		Define groups					
Group	Accession						
-	GSM396128	Enter a group name: List	of 6-10 week old C57BL/6J mice	4h	Heroin	group no.: 1	X
-	GSM396129	<input type="checkbox"/> Cancel selection	of 6-10 week old C57BL/6J mice	8h	Heroin	group no.: 2	X
-	GSM396130	<input checked="" type="checkbox"/> Cocaine 1hr	of 6-10 week old C57BL/6J mice	1h	Morphine	group no.: 3	X
-	GSM396131	<input checked="" type="checkbox"/> Cocaine 8hr	of 6-10 week old C57BL/6J mice	2h	Morphine	group no.: 4	X
-	GSM396132		Striatum of 6-10 week old C57BL/6J mice	4h	Morphine	group no.: 5	X
-	GSM396133		Striatum of 6-10 week old C57BL/6J mice	8h	Morphine	group no.: 6	X
-	GSM396134		Striatum of 6-10 week old C57BL/6J mice	1h	Cocaine	group no.: 7	X
-	GSM396135		Striatum of 6-10 week old C57BL/6J mice	2h	Cocaine	group no.: 8	X

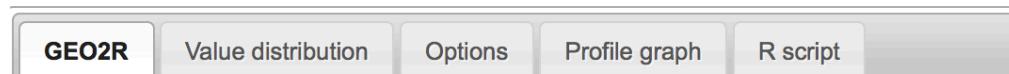
Then select the samples you want to add to each group. Click on a sample or samples, then select the group to which it should be assigned.

Samples		Define groups					
Group	Accession						
-	GSM396152	Enter a group name: List	of 6-10 week old C57BL/6J mice	1h	Heroin	group no.: 11	X
-	GSM396153	<input type="checkbox"/> Cancel selection	of 6-10 week old C57BL/6J mice	2h	Heroin	group no.: 12	X
-	GSM396154	<input checked="" type="checkbox"/> Cocaine 8hr (3 samples)	of 6-10 week old C57BL/6J mice	4h	Heroin	group no.: 1	X
-	GSM396155	<input checked="" type="checkbox"/> Cocaine 1hr (3 samples)	of 6-10 week old C57BL/6J mice	8h	Heroin	group no.: 2	X
-	GSM396156		of 6-10 week old C57BL/6J mice	1h	Morphine	group no.: 3	X
-	GSM396157		Striatum of 6-10 week old C57BL/6J mice	2h	Morphine	group no.: 4	X
-	GSM396158		Striatum of 6-10 week old C57BL/6J mice	4h	Morphine	group no.: 5	X
-	GSM396159		Striatum of 6-10 week old C57BL/6J mice	8h	Morphine	group no.: 6	X
-	GSM396160		Striatum of 6-10 week old C57BL/6J mice	1h	Saline	group no.: 15	X
-	GSM396161		Striatum of 6-10 week old C57BL/6J mice	2h	Saline	group no.: 16	X
-	GSM396162		Striatum of 6-10 week old C57BL/6J mice	4h	Saline	group no.: 17	X
-	GSM396163		Striatum of 6-10 week old C57BL/6J mice	8h	Saline	group no.: 18	X
Cocaine 1hr	GSM396164	<input checked="" type="checkbox"/> COC_I_2	Striatum of 6-10 week old C57BL/6J mice	1h	Cocaine	group no.: 7	X
-	GSM396165	<input checked="" type="checkbox"/> COC_II_2	Striatum of 6-10 week old C57BL/6J mice	2h	Cocaine	group no.: 8	X
-	GSM396166	<input checked="" type="checkbox"/> COC_IV_2	Striatum of 6-10 week old C57BL/6J mice	4h	Cocaine	group no.: 9	X
Cocaine 8hr	GSM396167	<input checked="" type="checkbox"/> COC_VIII_2	Striatum of 6-10 week old C57BL/6J mice	8h	Cocaine	group no.: 10	X
-	GSM396168	<input checked="" type="checkbox"/> HER_I_2	Striatum of 6-10 week old C57BL/6J mice	1h	Heroin	group no.: 11	X

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.

GEO2R	Value distribution	Options	Profile graph	R script
Apply adjustment to the P-values. More...		Apply log transformation to the data. More...		
<input checked="" type="radio"/> Benjamini & Hochberg (False discovery rate) <input type="radio"/> Benjamini & Yekutieli <input type="radio"/> Bonferroni <input type="radio"/> Hochberg <input type="radio"/> Holm <input type="radio"/> Hormmel <input type="radio"/> None		<input checked="" type="radio"/> Auto-detect <input type="radio"/> Yes <input type="radio"/> No		
Category of Platform annotation to display on results.				
<input type="radio"/> Submitter supplied <input checked="" type="radio"/> NCBI generated				
If you edit Options after performing an analysis, you must click Recalculate on the GEO2R tab to apply the edits.				

On the Geo2R tab, click on Top 250.



▼ Quick start

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare.
- Assign Samples to each group. Highlight Sample rows then click the group name to which group.
- Click 'Top 250' to perform the calculation with default settings.
- Results are presented as a table of genes ordered by significance. The top 250 genes.
- You may change settings in Options tab.

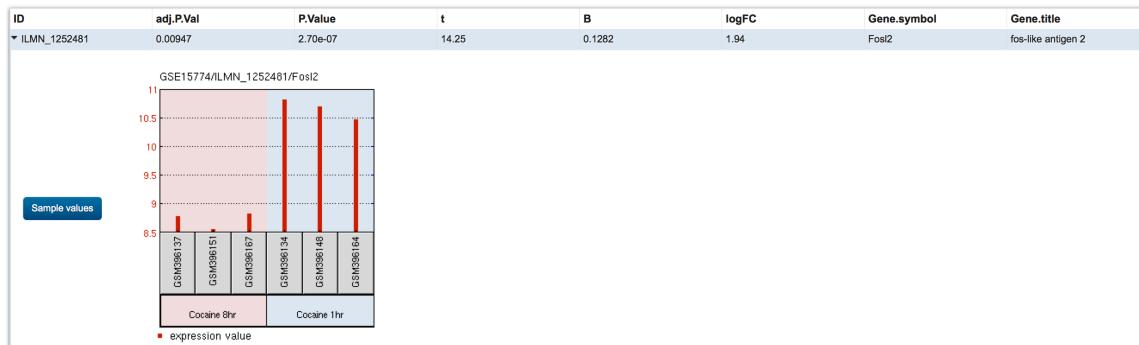
How to use

Top 250 Save all results

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.

Geo2R							
Value distribution Options Profile graph R script							
Quick start Recalculate if you changed any options. Save all results Select columns							
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	4.08e-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 laevis)
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 cytoskeletal...
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	Mdn	midnolin
ILMN_2615232	0.34825	1.19e-04	6.6	-0.9918	0.953	Ppp2rlb	protein phosphatase 2, regul...
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.

Check all

Data columns	Annotation columns
<input checked="" type="checkbox"/> Adj P-value	<input checked="" type="checkbox"/> ID
<input checked="" type="checkbox"/> P-value	<input checked="" type="checkbox"/> Gene symbol
<input checked="" type="checkbox"/> t-statistic	<input checked="" type="checkbox"/> Gene title
<input checked="" type="checkbox"/> B-value	<input type="checkbox"/> Gene ID
<input checked="" type="checkbox"/> logFC	<input type="checkbox"/> Chromosome location
<input type="checkbox"/> F-statistic	<input type="checkbox"/> Chromosome annotation

Val	P.Value	t	B	logFC	Gene.symbol
17					
17					
19					
16	1.19e-04	6.6	-0.9918	0.953	Ppp2r1b

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//lepid...	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				
ILMN_2622983	0.09066	1.36e-05	8.79	-0.4424	1.191	Dusp1	shisa homolog 2 (Xen...	dual specificity phosph...	MAP kinase tyrosine/...
ILMN_2616226	0.11154	1.91e-05	-8.41	-0.5154	-1.009	Dbp	D site albumin promot...	DNA binding//RNA p...	multicellular organism...
ILMN_2750515	0.13609	2.76e-05	8.02	-0.5994	2.313	Fos	FBX osteosarcoma on...	DNA binding//DNA bi...	circadian rhythm//po...
ILMN_2597827	0.13609	2.92e-05	7.97	-0.6122	1.996	Arc	activity regulated cyto...	actin binding//protein...	cell cycle//dephosph...
ILMN_1245088	0.20635	5.23e-05	-7.38	-0.7586	-0.92	Strip2	stratin interacting pro...	molecular_function	cell migration//cytosk...
ILMN_2778279	0.20635	5.31e-05	7.36	-0.7624	2.122	Fosb	FBX 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 phosph...	MAP kinase tyrosine/...	dephosphorylation//p...
ILMN_1299074	0.21755	6.53e-05	7.16	-0.8181	1.699	Irun	iun R nnnnnnnnnnnn	DNA binding//DNA N	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 (Geo2R_AddictionLab.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).

The screenshot shows the Bioconductor website with the following details:

- Header:** Bioconductor logo, "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS", and navigation menu with links to Home, Install, Help, Developers, and About. A search bar is also present.
- Breadcrumbs:** Home > Bioconductor 3.3 > Software Packages > GEOquery
- Section:** GEOquery (highlighted in green)
- Statistics:** platforms: all, downloads: top 5%, posts: 7 / 0.6 / 0.3 / 0, in Bioc: 10.5 years, build: ok, commits: 1.33, test coverage: 75%
- Social:** Facebook and Twitter icons
- Section:** Get data from NCBI Gene Expression Omnibus (GEO)
- Text:** Bioconductor version: Release (3.3)
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 <sda...>
Maintainer: Sean Davis <sda...>
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.
- Section:** Installation
To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("GEOquery")
```
- Documentation:** Documentation, Bioconductor, Package vignettes and manuals, Workflows for learning and use, Course and conference material, Videos, Community resources and tutorials, R / CRAN packages and documentation
- Support:** Support site - for questions about Bioconductor packages, Bioc-devel mailing list - for package developers

WGCNA

WGCNA and methods like it require R programming experience, which was not a pre-requisite for this course. If you are interested in learning more about these methods, please see the WGCNA page maintained by the Horvath Lab <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/> which includes:

Step by Step Tutorials:

<https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/index.html>

Installation Instructions:

<https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/InstallationInstructions.html>

Recommended Introductory Papers to WGCNA:

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008, 9:559 ([link to paper](#))

Langfelder P, Luo R, Oldham MC, Horvath S (2011) Is my network module preserved and reproducible? PLoS Comp Biol. 7(1): e1001057