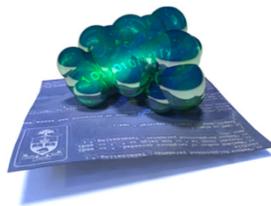


A
BIOINFORMATICS
COURSE

DIFFERENTIAL EXPRESSION

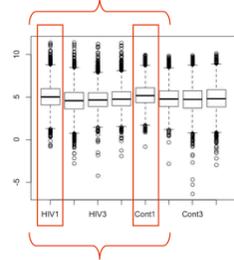


BORIS STEIPE

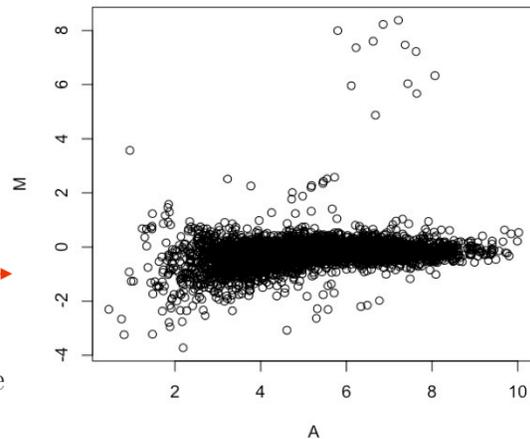
*DEPARTMENT OF BIOCHEMISTRY – DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO*

DIFFERENTIAL EXPRESSION ANALYSIS: THE QUESTION

$$(\text{Hiv1} + \text{Cont1}) / 2 = A$$



$$\text{Hiv1} - \text{Cont1} = M$$



Commonly we create M/A plots to evaluate Differential Expression:

M (minus) is the log ratio.

A (average) is overall intensity.

In this example of differential expression measurements data was analyzed from 7680 genes in CD4-T-cell lines at time $t = 24$ hours after infection with HIV type 1 virus. 4 replicates were performed for infected cells, and non-infected controls. (Data courtesy of Sohrab Shah).

The raw data was log-transformed. The y-axis of this “M/A plot” shows “minus” calculations i.e. $\log(\text{sample}) - \log(\text{control})$, which corresponds to a log-ratio measurement of the untransformed data. Genes with high increases of expression are on top, genes that are repressed are on the bottom. Most values show no differential expression and cluster around zero, because the values for sample and control are approximately the same, thus their ratio is around one.

The x-axis shows “average” calculations, i.e. the mean of sample and control. Highly expressed genes are on the right, genes with low expression levels are on the left. Note that the noise is larger for genes that have dim spots on the microarray slide.

The question now is: which of these genes would we consider **Differentially Expressed**, so we can follow up on the hypothesis that they contribute to disease mechanisms?

NCBI: GEO AND GEO2R

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. [Full instructions](#) [YouTube](#)

GEO accession

GEO2R

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, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, source and characteristics) columns to help determine which Samples belong 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 are presented and may be viewed as profile graphs. Alternatively, the complete results table may be saved.
- You may change settings in Options tab.

[How to use](#)

[Save all results](#)

The GEO database provides tools to support differential expression analysis in data sets.

NCBI: GEO

Browser: "yeast cell cycle" - GEO DataS...
 URL: https://www.ncbi.nlm.nih.gov/gds/?term="yeast+cell+cycle"

NCBI Resources How To steipe My NCBI Sign Out

GEO DataSets "yeast cell cycle" Search

Summary 20 per page Sort by Default order Send to: Filters: Manage Filters

Entry type DataSets (2)
 Series (6)
 Samples (63)
 Platforms (0)

Organism Customize ...

Study type Expression profiling by array
 Methylation profiling by array
 Customize ...

Author Customize ...

Attribute name tissue (0)
 strain (3)
 Customize ...

Publication dates 30 days
 1 year
 Custom range...

Clear all Show additional filters

Search results
 Items: 1 to 20 of 71

1. [Wild type strain across two cell cycles \(II\)](#)
 Analysis of wild type W303a cells across two cell cycles, a length of 2 hours after synchronization with alpha factor. Gene expression examined at 5-minute intervals.
 Organism: Saccharomyces cerevisiae
 Type: Expression profiling by array, log ratio, 25 time sets
 Platform: GPL1914 Series: GSE4987 50 Samples
 Download data: GEO (GPR)
 DataSet Accession: GDS2350 ID: 2350
[PubMed](#) [Full text in PMC](#) [Similar studies](#) [GEO Profiles](#) [Analyze DataSet](#)

2. [Wild type strain across two cell cycles \(I\)](#)
 Analysis of wild type W303 cells across two cell cycles, a length of 2 hours after synchronization with alpha factor. Results compared to those from an experiment using a yox1 yhp1 double mutant strain (GDS2318).
 Organism: Saccharomyces cerevisiae
 Type: Expression profiling by array, log ratio, 13 time sets
 Platform: GPL1914 Series: GSE3635 13 Samples
 Download data: GEO (GPR)
 DataSet Accession: GDS2347 ID: 2347
[PubMed](#) [Full text in PMC](#) [Similar studies](#) [GEO Profiles](#) [Analyze DataSet](#)

3. [Checkpoints Couple Transcription Network Oscillator Dynamics to Cell-Cycle Progression](#)
 (Submitter supplied) **Yeast cell cycle** transcript dynamics in three S. cerevisiae strains grown at 30 °C in Glycerol (GDS2347, GDS2348) (submitted with GPR entries GPR1914, GPR1915, GPR1916)

Top Organisms [Tree]
 Saccharomyces cerevisiae (71)
 Schizosaccharomyces pombe (3)
 Saccharomyces paradoxus (1)

Find related data
 Database: Select
 Find items

Search details
 "yeast cell cycle"[All Fields]
 Search See more...

Recent activity
 Turn Off Clear
 "yeast cell cycle" (71) GEO DataSets

A keyword search yields candidate data sets.

NCBI: GEO BROWSER

GEO DataSet Browser

https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2347

NCBI DATASET BROWSER GEO Gene Expression Omnibus

Search for GDS2347[ACCN] Search Clear Show All Advanced Search

DataSet Record GDS2347: Expression Profiles Data Analysis Tools Sample Subsets

Title: Wild type strain across two cell cycles (1)

Summary: Analysis of wild type W303 cells across two cell cycles, a length of 2 hours after synchronization with alpha factor. Results compared to those from an experiment using a yox1 yhp1 double mutant strain (GDS2318).

Organism: *Saccharomyces cerevisiae*

Platform: GPL1914: FHCRC Yeast Amplicon v1.1

Citation: Pramila T, Miles S, GuhaThakurta D, Jemiolo D et al. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. *Genes Dev* 2002 Dec 1;16(23):3034-45. PMID: 12464633

Reference Series: GSE3635 Sample count: 13

Value type: log ratio Series published: 2006/06/01

Cluster Analysis

Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINIML file
- Annotation SOFT file

Data Analysis Tools

Find genes ?

Compare 2 sets of samples

Cluster heatmaps

Experiment design and value distribution

Find gene name or symbol: Go

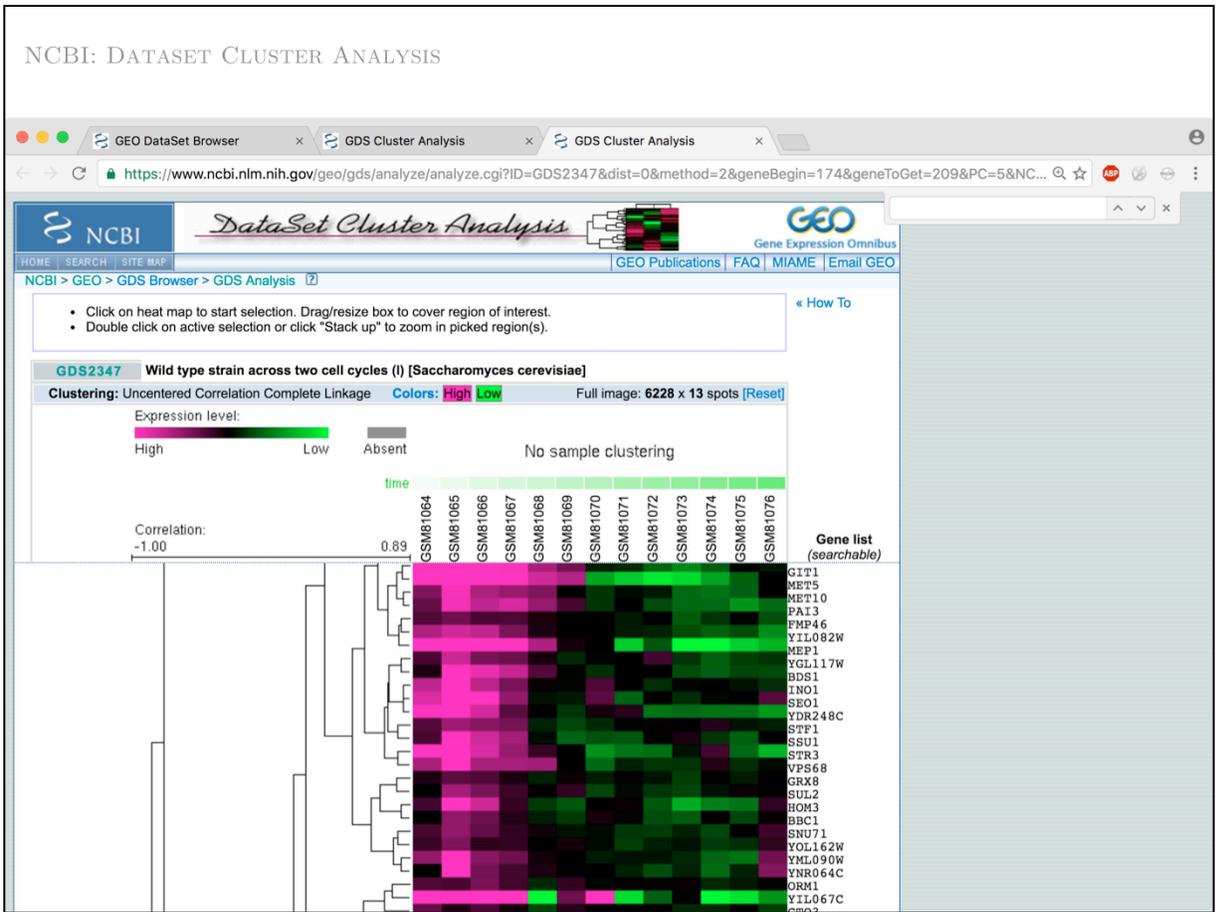
Find genes that are up/down for this condition(s): time Go

NLM NIH GEO Help Disclaimer Accessibility

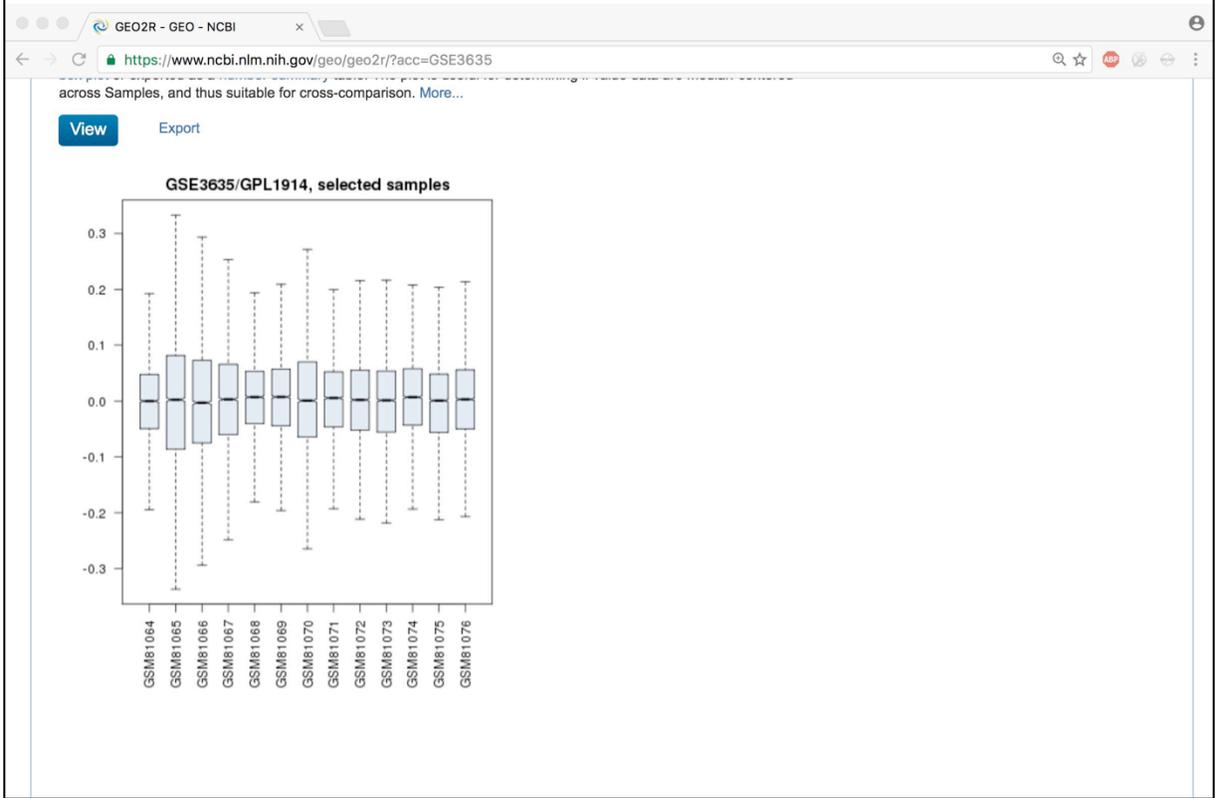
The dataset browser displays information about the experiment.



Hierarchical clustering shows genes with similar expression profiles. Regions of the dense gene tree can be expanded (i.e. the region between the dotted red lines).



The expanded region of the expression “heatmap” identifies the individual genes and shows how they are up- and down regulated in the individual experiments of the data set. Since this is a cell cycle study, the columns correspond to time points. We see genes that are highly expressed at the beginning of the cycle and poorly expressed towards its end.



The GEO2R tool supports detailed analysis and gene discovery: first, the quality of the samples is determined, to allow identifying problematic experiments that do not follow the general distribution of the others ...

GEO accession: [Saccharomyces cerevisiae alpha factor cell cycle](#)

▼ Samples Selected 0 out of 13 samples

Define groups

Enter a group name:

Group	Accession	Title	Source name 1	Source name 2	Yeast cell cycle-time point 0 min (Ch1)	Yeast asynchronous culture (Ch2)	Yeast cell cycle-time point 1 (Ch1)
-	GSM81064	Yeast cell cycl	000.rfm	Yeast cell cycle-time point 0 min	Yeast asynchronous culture		
-	GSM81065	Yeast cell cycl	0010.rfm	Yeast cell cycle-time point 10 min	Yeast asynchronous culture		
-	GSM81066	Yeast cell cycl	0020.rfm	Yeast cell cycle-time point 20 min	Yeast asynchronous culture		
-	GSM81067	Yeast cell cycl	0030.rfm	Yeast cell cycle-time point 30 min	Yeast asynchronous culture		
-	GSM81068	Yeast cell cycl	0040.rfm	Yeast cell cycle-time point 40 min	Yeast asynchronous culture		
-	GSM81069	Yeast cell cycl	0050.rfm	Yeast cell cycle-time point 50 min	Yeast asynchronous culture		
-	GSM81070	Yeast cell cycl	0060.rfm	Yeast cell cycle-time point 60 min	Yeast asynchronous culture		
-	GSM81071	Yeast cell cycl	0070.rfm	Yeast cell cycle-time point 70 min	Yeast asynchronous culture		
-	GSM81072	Yeast cell cycl	0080.rfm	Yeast cell cycle-time point 80 min	Yeast asynchronous culture		
-	GSM81073	Yeast cell cycle-time point 90 min 2001-04-11_0090.rfm	0090.rfm	Yeast cell cycle-time point 90 min	Yeast asynchronous culture		
-	GSM81074	Yeast cell cycle-time point 100 min 2001-04-11_0100.rfm	0100.rfm	Yeast cell cycle-time point 100 min	Yeast asynchronous culture		
-	GSM81075	Yeast cell cycle-time point 110 min 2001-04-11_0110.rfm	0110.rfm	Yeast cell cycle-time point 110 min	Yeast asynchronous culture		
-	GSM81076	Yeast cell cycle-time point 120 min 2001-04-11_0120.rfm	0120.rfm	Yeast cell cycle-time point 120 min	Yeast asynchronous culture		

Calculate the distribution of value data for the Samples you have selected. Distributions may be viewed graphically as a

... then groups are defined and differential expression values are computed between groups, to identify the most significant differentially expressed genes.

<http://steipe.biochemistry.utoronto.ca/abc>

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA