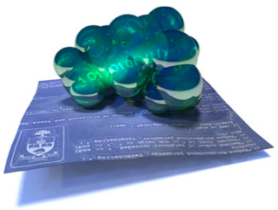
A
Bioinformatics
Course

Differential Expression

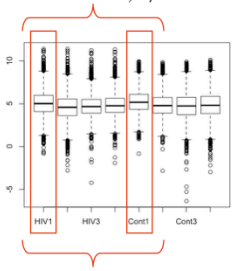


Boris Steipe

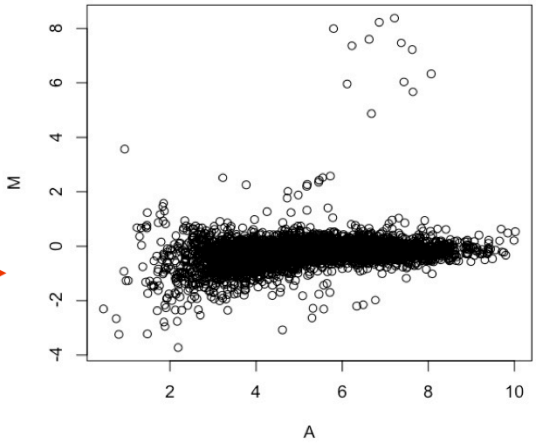
department of biochemistry – department of molecular genetics
University of Toronto

Differential Expression Analysis: the Question

(Hiv1 + Cont1) / 2 = A



Hiv1 – Cont1 = M



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

M (minus) is the log ratio.

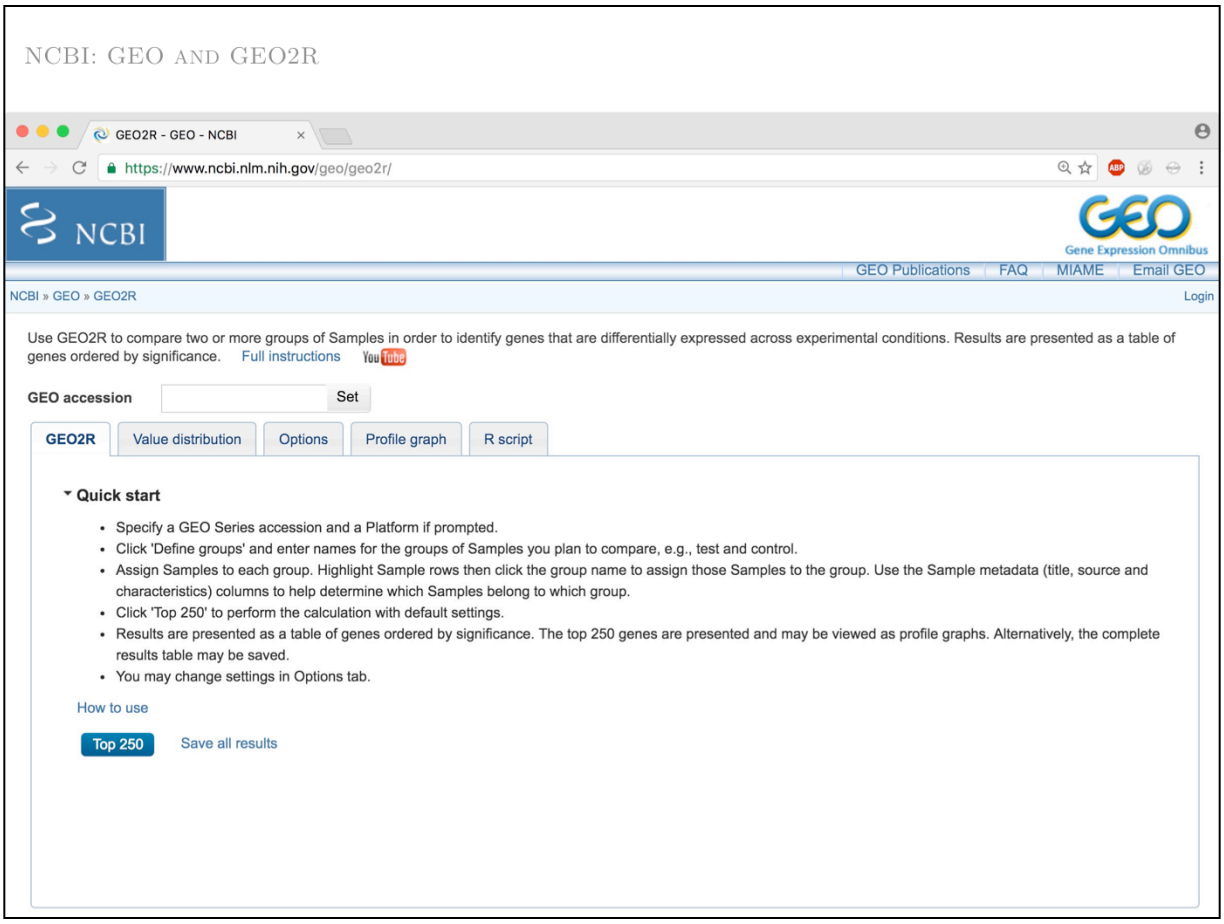
A (average) is overall intensity.

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).

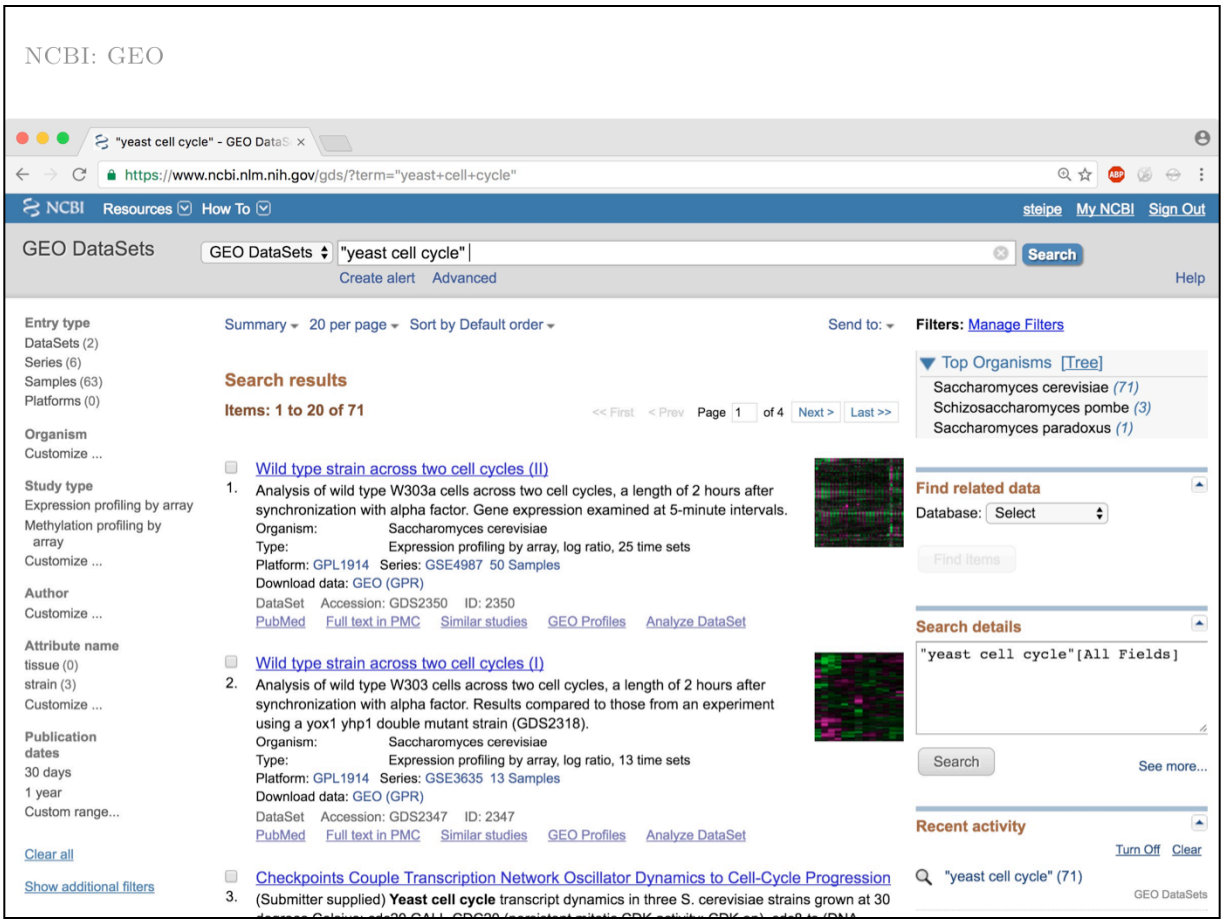
The raw data was log-transformed. The y-axis of this "M/A plot" shows "minus" calculations i.e. log(sample) – log(control), which corresponds to a log-ratio measurement of the untransformed data. Genes with high increases of expression are on top, genes that are repressed are on the bottom. Most values 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.

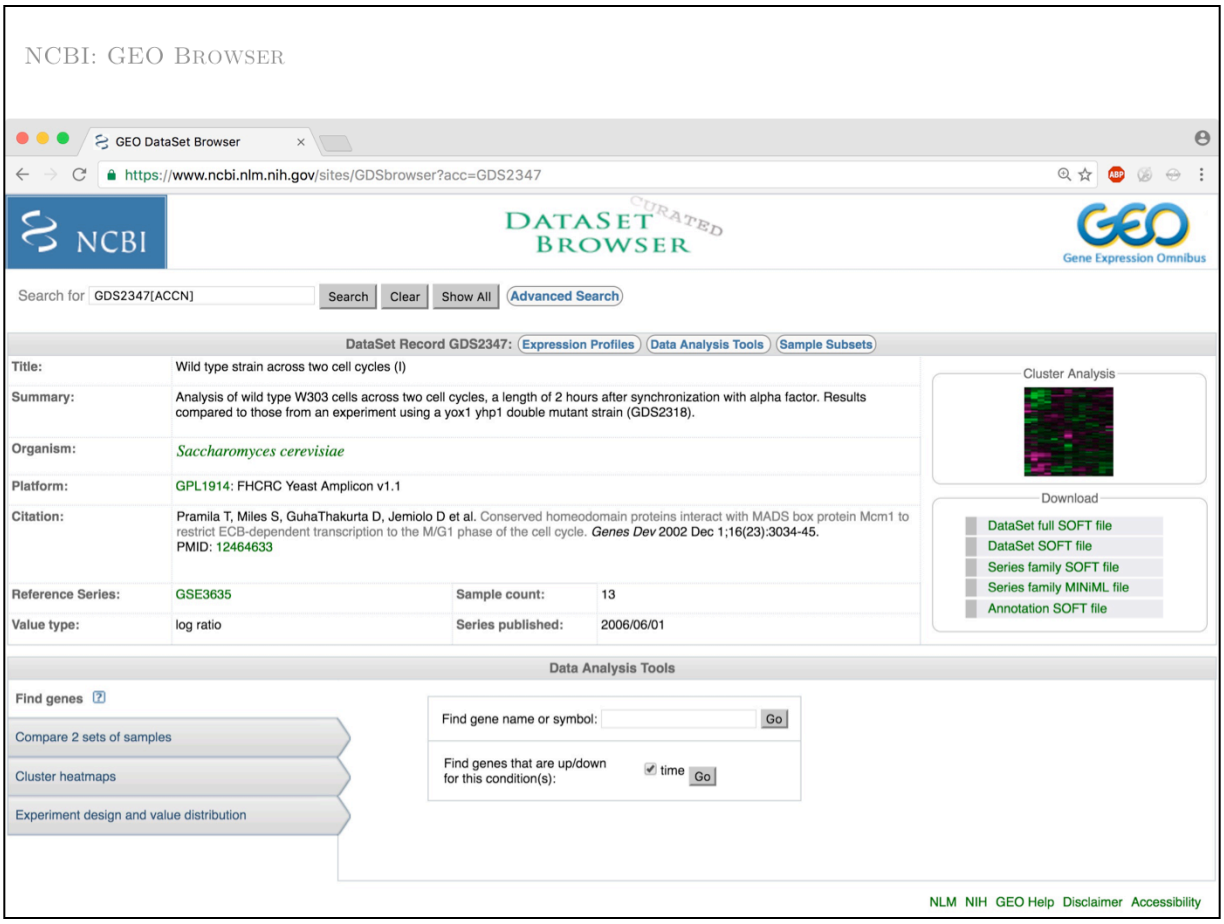
The question now is: which of these genes would we consider **Differentially Expressed**, so we can follow up on the hypothesis that they contribute to disease mechanisms?



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

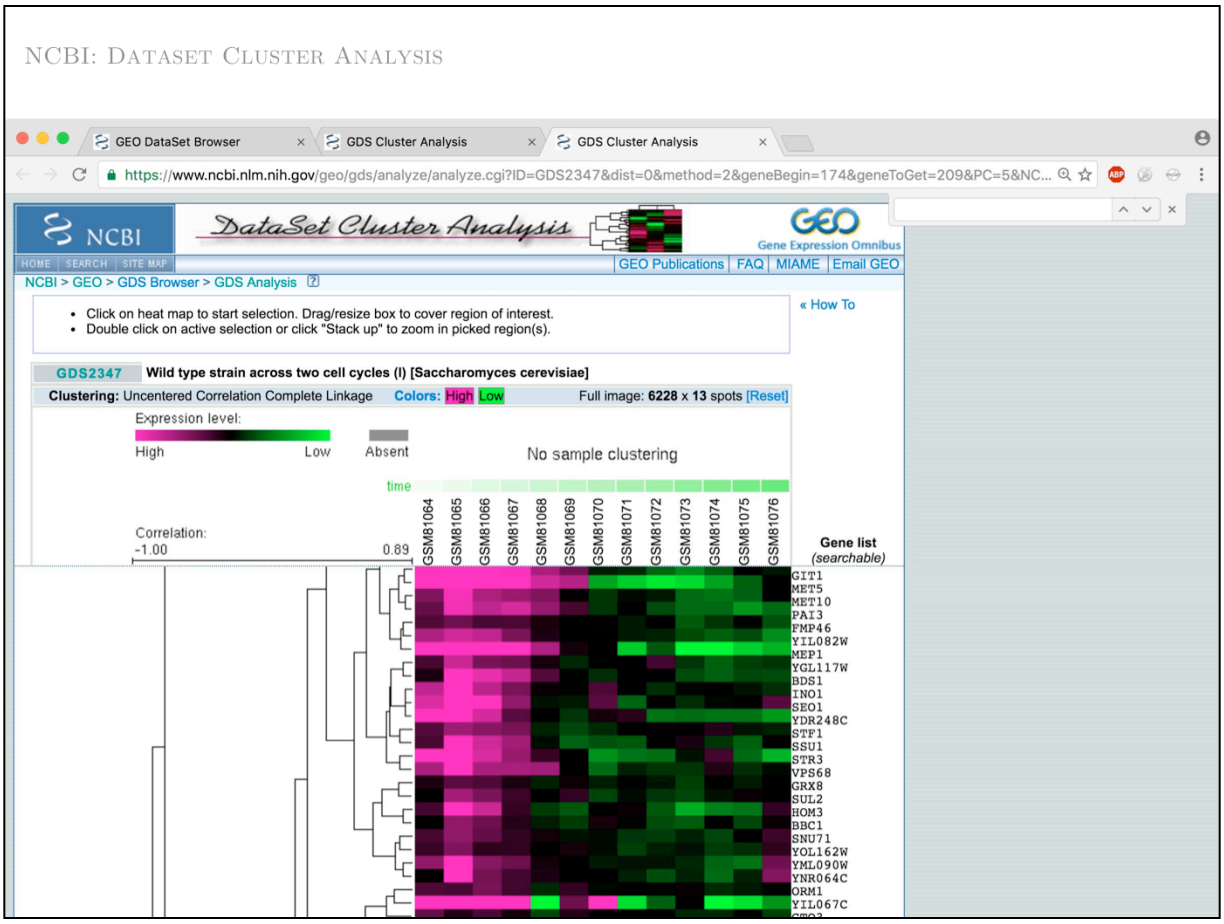


A keyword search yields candidate data sets.



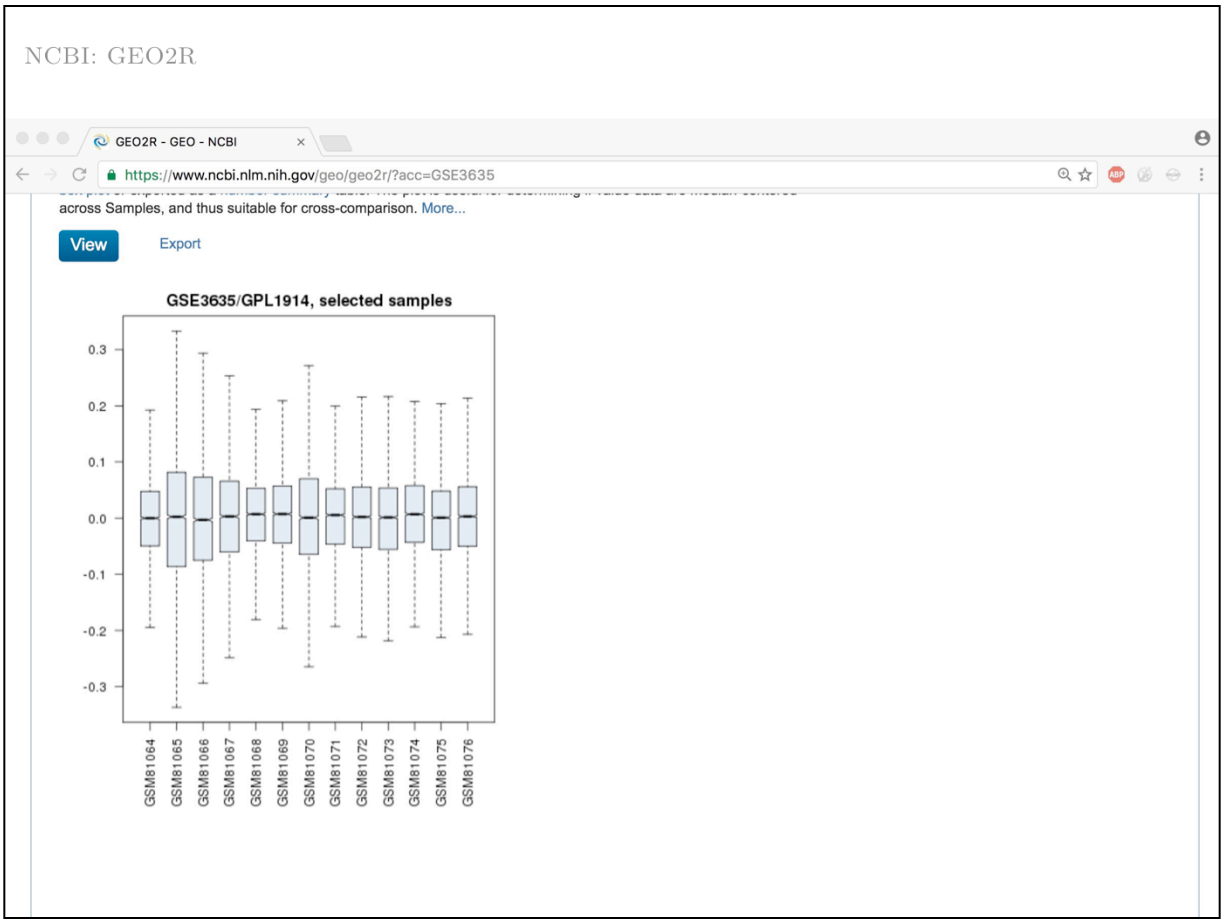
The dataset browser displays information about the experiment.

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).



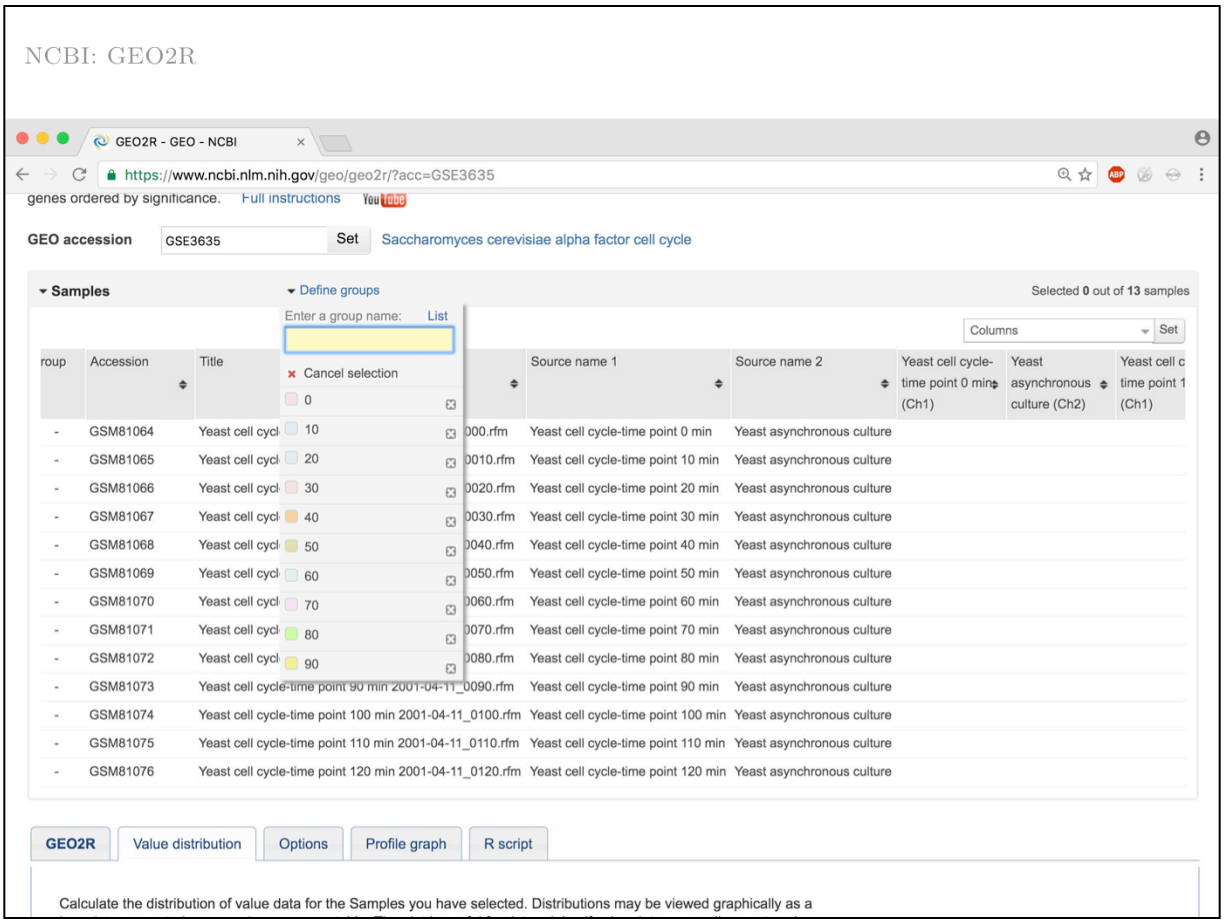
The expanded region of the expression "heatmap" identifies the individual genes and shows how they are up- and down regulated in the individual experiments of the data set. Since this is a cell cycle study, the columns correspond to time points. We see genes that are highly expressed at the beginning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

NCBI: GEO2R



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 ...

NCBI: GEO2R



... 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