# Gene Expression Analysis – Processing, Querying and Visualizing Gene Expression Data CANB 7640

Aik Choon Tan, Ph.D.
Associate Professor of Bioinformatics
Division of Medical Oncology
Department of Medicine
aikchoon.tan@ucdenver.edu

10/16/2018

http://tanlab.ucdenver.edu/labHomePage/teaching/CANB7640/

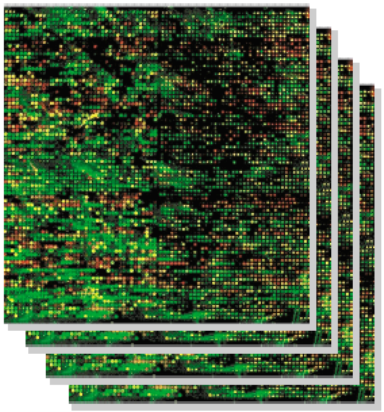
# Outline

- Processing Raw Data
  - Normalization Concepts
  - Affymetrix Power Tools
- Querying Public Gene Expression Database
  - NCBI GEO
  - Rules for Reusing Public Gene Expression Data
- Visualizing matrix as heat map
  - Heat map concept
  - matrix2png

# Analytical Process in Microarray Experiments

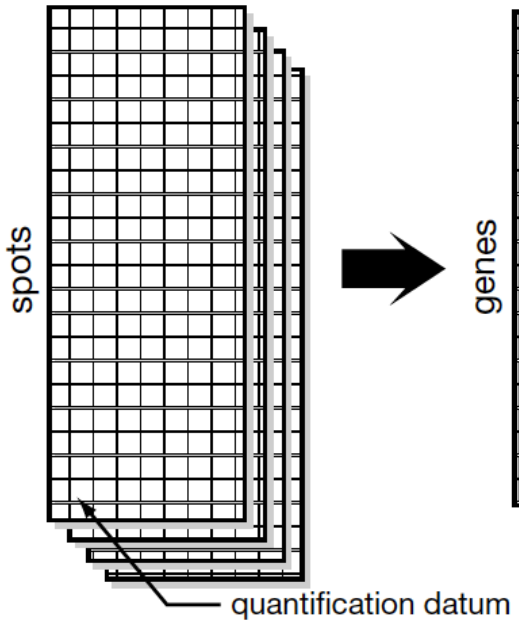
**raw data**

array scans



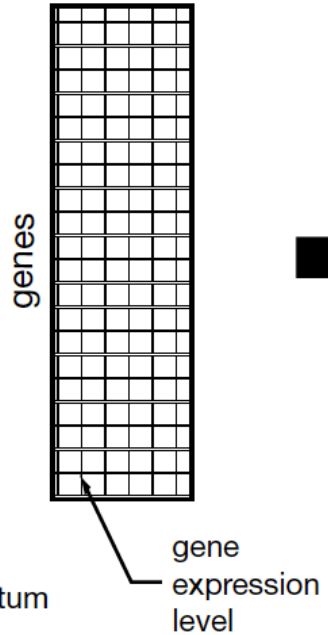
**quantification matrices**

quantifications




**gene expression data matrix**

samples




**gene expression data analysis**

- Cluster analysis
- Candidate gene analysis
- Gene Set analysis

**gene expression secondary analysis**

- Clusters
- Gene Lists
- Gene Sets

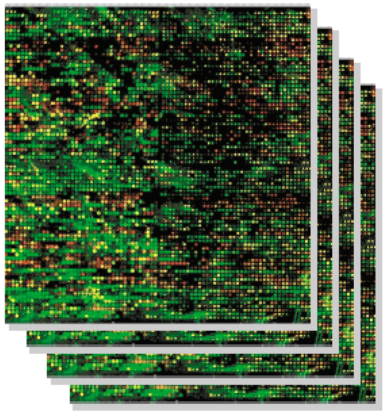
**gene expression visualization**

- Clusters
- Gene Lists
- Gene Sets

# So far, Tools Covered in the Analytical Process in Microarray Experiments

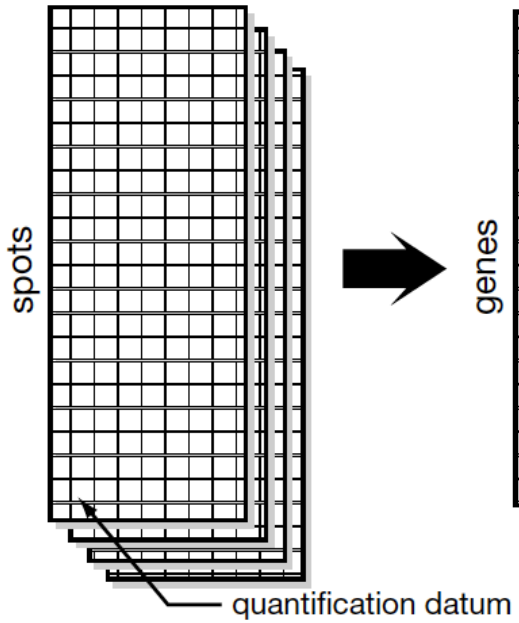
**raw data**

array scans



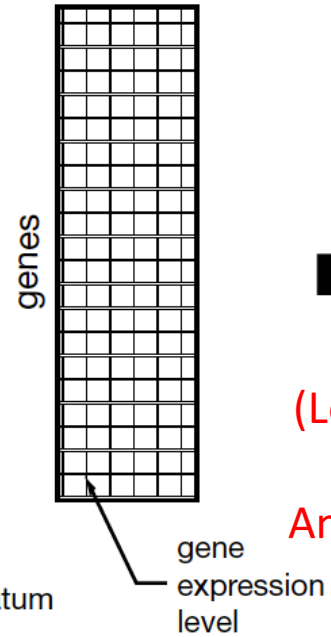
**quantification matrices**

quantifications




**gene expression data matrix**

samples




**gene expression data analysis**

- Cluster analysis (CLUSTER 3.0)
- Candidate gene analysis (SAM)
- Gene Set analysis (GSEA)

**gene expression secondary analysis**

(Leading Edge Analysis)

- Clusters
- Gene Lists
- Gene Sets

**gene expression visualization**

(Java TreeView)

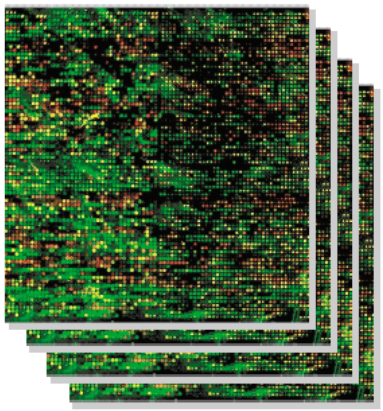
(NIH DAVID, KEGG)

- Clusters
- Gene Lists
- Gene Sets

# Complete Analytical Process in Microarray Experiments

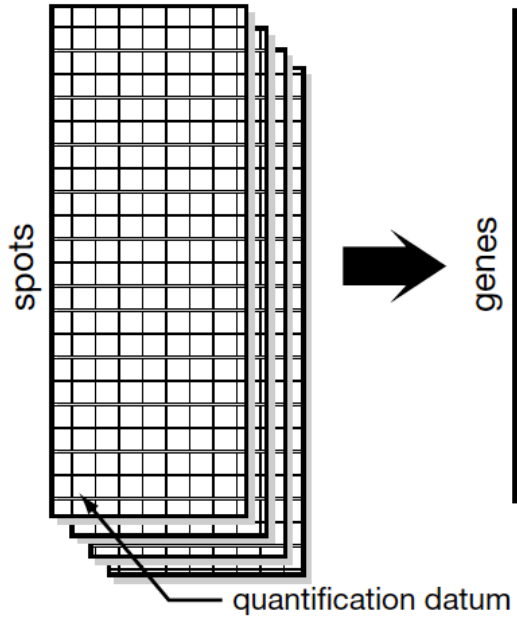
**raw data**

array scans



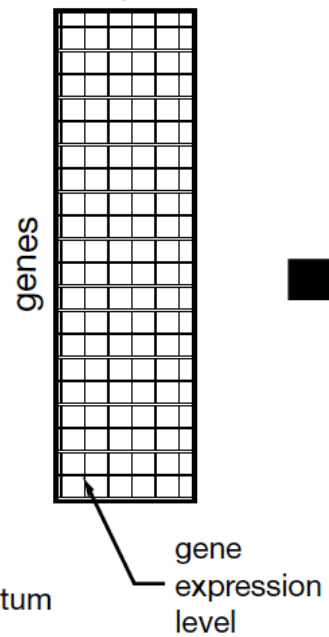
**quantification matrices**

quantifications




**gene expression data matrix**

samples




**gene expression data analysis**

- Cluster analysis
- Candidate gene analysis
- Gene Set analysis

**gene expression secondary analysis**

- Clusters
- Gene Lists
- Gene Sets

**gene expression visualization**

- Clusters
- Gene Lists
- Gene Sets

**publication and data sharing**

- Publication
- Deposit into Public Databases

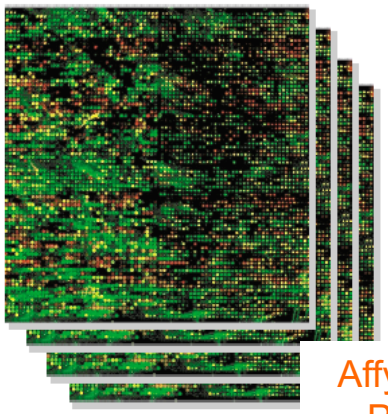
**reuse and re-discovery**

- Published Papers
- Public Databases

# Tools to Learn Today in the Complete Analytical Process in Microarray Experiments

**raw data**

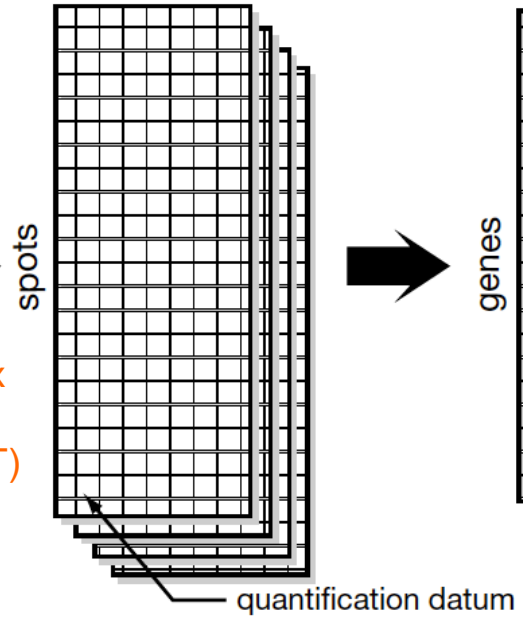
array scans



Affymetrix Power Tools (APT)

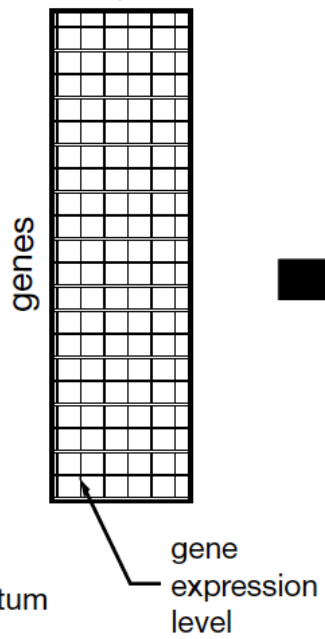
**quantification matrices**

quantifications




**gene expression data matrix**

samples




**gene expression data analysis**

- Cluster analysis
- Candidate gene analysis
- Gene Set analysis

**gene expression secondary analysis**

- Clusters
- Gene Lists
- Gene Sets

**gene expression visualization**

- Clusters
- Gene Lists
- Gene Sets

Heat map tool -matrix2png

**publication and data sharing**

- Publication
- Deposit into Public Databases

NCBI GEO

**reuse and re-discovery**

- Published Papers
- Public Databases

NCBI GEO

# Extracting Signals from Microarray

- Image file to pixel
  - Potential biases in scanner, distribution of the fluorescence dyes, over/under exposure of a particular corner in the array (Batch)
  - Potential biases in sample preparation steps (Technical)
- Normalization is a way to account for these biases and hope to generate a comparable measurements across samples

# Normalization Methods in Microarray

- Image file to pixel
  - Potential biases in scanner, distribution of the fluorescence dyes, over/under exposure of a particular corner in the array (Batch)
  - Potential biases in sample preparation steps (Technical)
- Normalization is a way to account for these biases and hope to generate a comparable measurements across samples for downstream analysis

# Robust Multi-array Average (RMA)

*Biostatistics* (2003), **4**, 2, *pp*. 249–264
*Printed in Great Britain*

## Exploration, normalization, and summaries of high density oligonucleotide array probe level data

RAFAEL A. IRIZARRY*
*Department of Biostatistics, Johns Hopkins University, Baltimore MD 21205, USA*
rafa@jhu.edu

BRIDGET HOBBS
*Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia*

FRANCOIS COLLIN
*Gene Logic Inc., Berkeley, CA, USA*

YASMIN D. BEAZER-BARCLAY, KRISTEN J. ANTONELLIS, UWE SCHERF
*Gene Logic Inc., Gaithersburg, MD, USA*

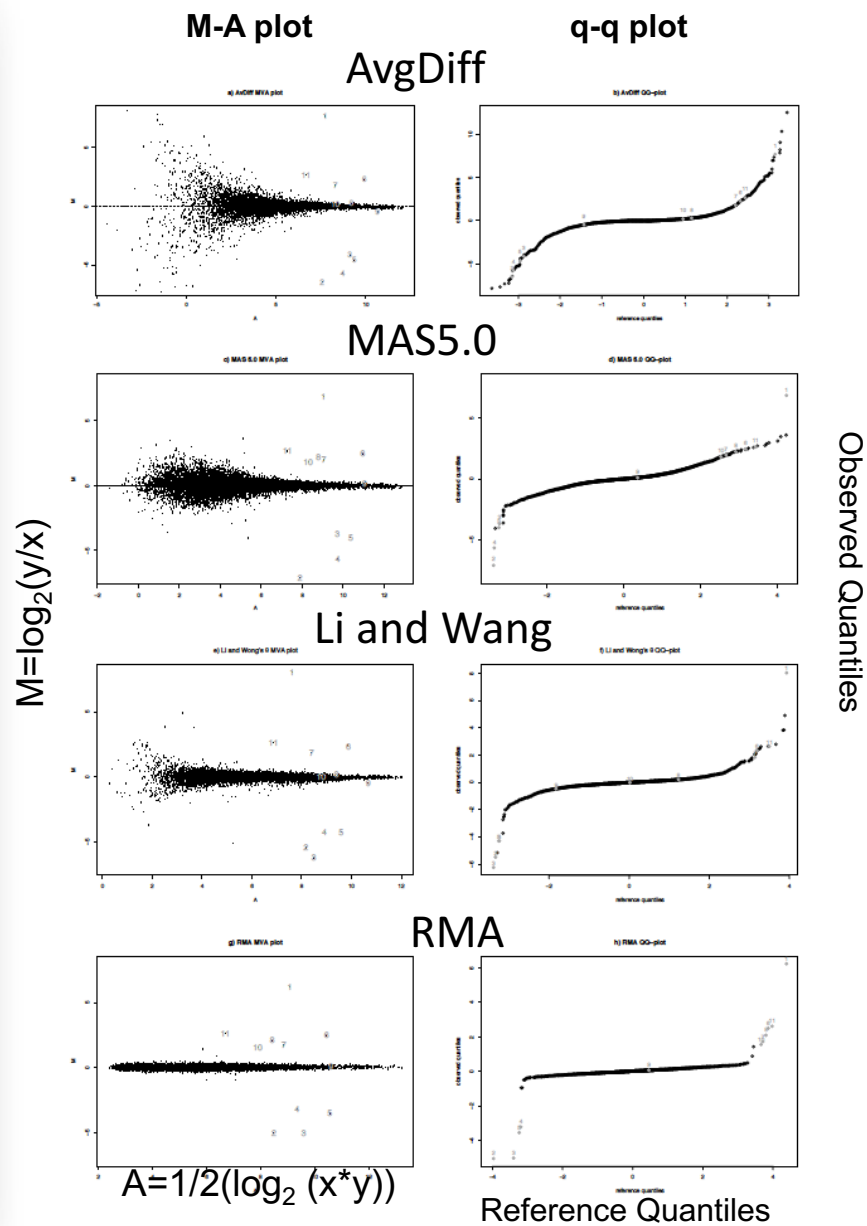
TERENCE P. SPEED
*Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia. Department of Statistics, University of California at Berkeley*

SUMMARY

In this paper we report exploratory analyses of high-density oligonucleotide array data from the Affymetrix GeneChip® system with the objective of improving upon currently used measures of gene expression. Our analyses make use of three data sets: a small experimental study consisting of five MGU74A mouse GeneChip® arrays, part of the data from an extensive spike-in study conducted by Gene Logic and Wyeth's Genetics Institute involving 95 HG-U95A human GeneChip® arrays; and part of a dilution study conducted by Gene Logic involving 75 HG-U95A GeneChip® arrays. We display some familiar features of the perfect match and mismatch probe (*PM* and *MM*) values of these data, and examine the variance–mean relationship with probe-level data from probes believed to be defective, and so delivering noise only. We explain why we need to normalize the arrays to one another using probe level intensities. We then examine the behavior of the *PM* and *MM* using spike-in data and assess three commonly used summary measures: Affymetrix's (i) average difference (AvDiff) and (ii) MAS 5.0 signal, and (iii) the Li and Wong multiplicative model-based expression index (MBEI). The exploratory data analyses of the probe level data motivate a new summary measure that is a robust multi-array average (RMA) of background-adjusted, normalized, and log-transformed *PM* values. We evaluate the four expression summary measures using the dilution study data, assessing their behavior in terms of bias, variance and (for MBEI and RMA) model fit. Finally, we evaluate the algorithms in terms of their ability to detect known levels of differential expression using the spike-in data. We conclude that there is no obvious downside to using RMA and attaching a standard error (SE) to this quantity using a linear model which removes probe-specific affinities.

(Cited by 6094)

*To whom correspondence should be addressed



# Affymetrix Power Tools

- Affymetrix Power Tools (APT) are a set of ***cross-platform command line programs*** that implement ***algorithms for analyzing and working*** with ***Affymetrix GeneChip® arrays***.
- APT is an ***open-source*** project licensed under the GNU General Public License (GPL). (Developers who need a non-GPL license may purchase a commercial license from Affymetrix.)
- APT programs are intended for "***power users***" who prefer programs that can be utilized in scripting environments and are sophisticated enough to handle the complexity of extra features and functionality.
- The vision is that APT provides a platform for developing and deploying new algorithms without waiting for the GUI implementations.

# How to get Affymetrix Power Tools?

Login | Affymetrix

https://www.affymetrix.com/estore/user/login.jsp?toURL=/estore/

Fellowships ...rd Medicine | Concur -- Login | TP53 | Rank-Based ...search Grid | PAR-11-208:...atics (R01) | Science/AAA... Resources

affymetrix
Biology for a better world

Welcome, **Guest** | Login | Register

Quick Order | 0 Items

Search our site by entering a Product name, Part number, CAS #, etc.

Home | Products | Brands | Community | Support | Partners & Programs | About Affymetrix | Careers | NetAffx | JP CN

Customer Login
Contact Us

See the whole picture.
Get 100% PCR recovery with USB® ExoSAP-IT® single-tube solution.
More >

Home > Customer Login

## Customer Login

### Returning Customer

Already registered? Please login.

**Email ID:**
**Password:**

Need login assistance?

Submit

### New Customer

Are you a new customer? Register Now to start enjoying all the benefits of the Affymetrix website.

New Customer

### Why Register?

- View pricing for your organization
- Purchase products online

e information
e, sample data, protocols and more...

Webpage: [www.affymetrix.com](http://www.affymetrix.com)
Register for Free
Login

# User Manual (Help Page)

http://media.affymetrix.com/support/developer/powertools/changelog/apt-probeset-summarize.html

## MANUAL: apt-probeset-summarize (1.14.3)

### Contents



### Introduction

`apt-probeset-summarize` is a program for doing background subtraction, normalization and summarizing probe sets from Affymetrix expression microarrays. It implements analysis algorithms such as RMA, Plier, and DABG (detected above background).

The main features of apt-probeset-summarize not common in other implementations are:

- Quantile normalization using a subset (sketch) of the data which results in much smaller memory usage.
- Analyze microarray in chunks of probesets, which when combined with the sketch quantile normalization above allows the analysis thousands of chips on a single computer or in parallel across a cluster of computers.
- Save target normalizations and probe (feature) effects for use later.
- Ability to group probesets into larger "meta" probesets as specified by the user via a text file. This is particularly for the exon level microarrays where gene or transcript level estimates are desired.

# Usage

```
hslib@HSL-TL-04 /cygdrive/c/Program Files/Affymetrix Power T
ools/APT-14.2/Bin
$ ./apt-probeset-summarize.exe -a ANALYSIS_TYPE -d CDF -o OUTPUT *.CEL
```

Call the program

Type of analysis

Chip annotations:
-d chip.CDF
or
-p chip.pgf
-c chip.clf

Output Folder

Input CEL files

```
usage:
   apt-probeset-summarize -a rma-sketch -a plier-mm-sketch \
        -p chip.pgf -c chip.clf -o output-dir *.cel
```

# NCBI GEO

NCBI Resources How To Sign in to NCBI

GEO Home | Documentation | Query & Browse | Email GEO

## Gene Expression Omnibus

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.

Gene Expression Omnibus

Keyword or GEO Accession Search

### Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

### Tools

- 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
- FTP Site

### Browse Content

- Repository Browser

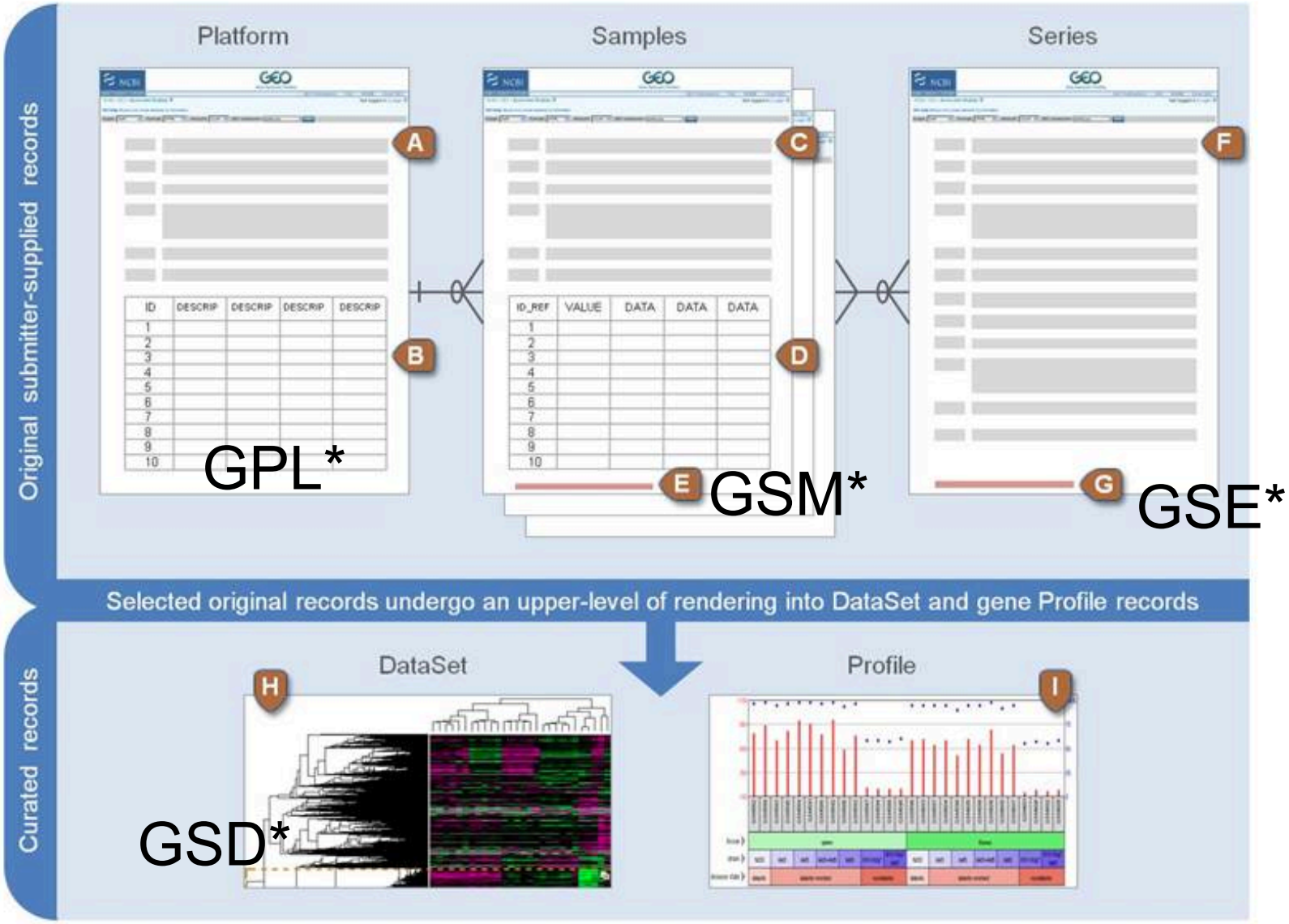
| | |
|---|---|
| DataSets: | 3847 |
| Series: | 51159 |
| Platforms: | 13429 |
| Samples: | 1245651 |

### Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

http://www.ncbi.nlm.nih.gov/geo/

# NCBI GEO Data Formats



# NCBI GEO Data Formats

| | | | |
|---|---|---|---|
| **Platform** | Platform records are supplied by submitters<br>A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template.Each Platform record is assigned a unique and stable GEO accession number (GPLxxx). A Platform may reference many Samples that have been submitted by multiple submitters.<br>Example Platform record » | **A** | **Text description of the array or sequencer** |
| | | **B** | **Text tab-delimited table of the array template** |
| **Sample** | Sample records are supplied by submitters<br>A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx). A Sample entity must reference only one Platform and may be included in multiple Series.<br>Example Sample record » | **C** | **Text description of the biological sample and protocols to which it was subjected** |
| | | **D** | **Text tab-delimited table of processed hybridization result** (may optionally include raw data columns) |
| | | **E** | **Original raw data file, or processed sequence data file** |
| **Series** | Series records are supplied by submitters<br>A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx).<br>Example Series record » | **F** | **Text description of the overall experiment** |
| | | **G** | **Tar archive of original raw data files, or processed sequence data files** |

(e.g HG-U133 Plus 2.0 microarray)

(e.g Sample1)

(e.g Cancer_vs_Normal)

# NCBI GEO Data Formats

**DataSet**

DataSet records are assembled by GEO curators

As explained above, A GEO Series record is an original submitter-supplied record that summarizes an experiment. These data are reassembled by GEO staff into GEO Dataset records (GDSxxx).

A DataSet represents a curated collection of biologically and statistically comparable GEO Samples and forms the basis of GEO's suite of data display and analysis tools.

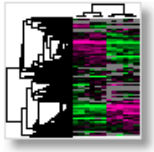
Samples within a DataSet refer to the same Platform, that is, they share a common set of array elements. Value measurements for each Sample within a DataSet are assumed to be calculated in an equivalent manner, that is, considerations such as background processing and normalization are consistent across the DataSet. Information reflecting experimental factors is provided through DataSet subsets.

Both Series and DataSets are searchable using the GEO DataSets interface, but only DataSets form the basis of GEO's advanced data display and analysis tools including gene expression profile charts and DataSet clusters. Not all submitted data are suitable for DataSet assembly and we are experiencing a backlog in DataSet creation, so not all Series have corresponding DataSet record(s).

For more information, see About GEO DataSets page.

Example DataSet record »

H



**Profile**

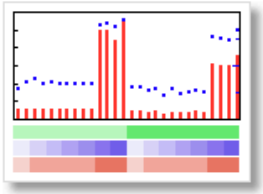
Profiles are derived from DataSets

A Profile consists of the expression measurements for an individual gene across all Samples in a DataSet. Profiles can be searched using the GEO Profiles interface.

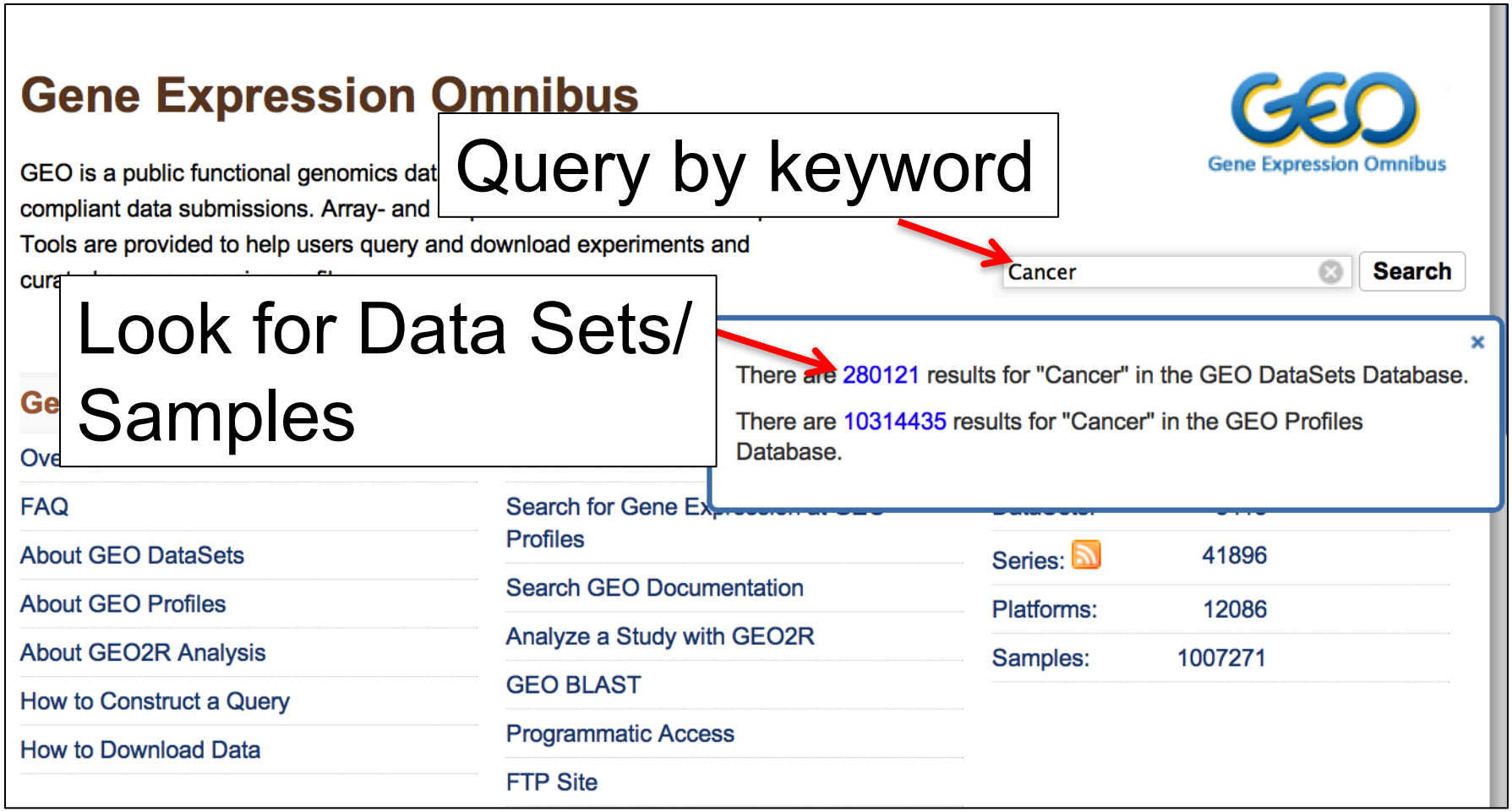
For more information, see About GEO Profiles page.

Example Profile records »

I



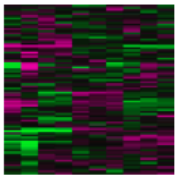
# Querying NCBI GEO



**Results: 1 to 20 of 280121**



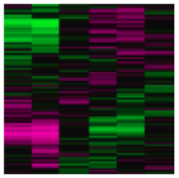
1. Tuberous sclerosis complex 1 deficient naïve CD4 and CD8 T cells



Analysis of naïve CD4 and CD8 T cells deficient for Tuberous sclerosis complex 1 (Tsc1). Tsc1 tumor suppressor is an mTOR signaling modulator involved in naive T cell quiescence and immune homeostasis regulation. Results identify a Tsc1-dependent gene signature in naive T cells.

Organism: Mus musculus
Type: Expression profiling by array, transformed count, 2 cell type, 2 genotype/variation sets
Platform: GPL11180 Series: GSE29797 10 Samples
Download data: GEO (CEL)
DataSet Accession: GDS4572 ID: 4572
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

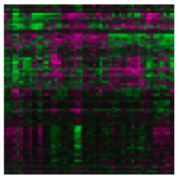
2. EphB2-selected populations of intestinal crypt cells



Analysis of intestinal epithelial cells FACS-sorted according to high, medium or low EphB2 receptor levels. Intestinal stem cells (ISC) highly express EphB2 receptor which becomes gradually silenced as cells differentiate. Results provide insight into a molecular program specific for normal ISCs.

Organism: Mus musculus
Type: Expression profiling by array, transformed count, 3 cell type sets
Platform: GPL1261 Series: GSE27605 6 Samples
Download data: GEO (CEL)
DataSet Accession: GDS4514 ID: 4514
PubMed Similar studies GEO Profiles Analyze DataSet

3. Clinical outcome of stage UICC II colon **cancer** patients



Analysis of tumor cells from sporadic stage UICC II colon **cancer** patients who were treated by elective standard oncological resection but developed relapse during follow-up. Results provide insight into the challenges of constructing molecular signatures predictive for patient outcome.

Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 2 disease state, 2 other sets
Platform: GPL570 Series: GSE18088 53 Samples
Download data: GEO (CEL)
DataSet Accession: GDS4513 ID: 4513
PubMed Similar studies GEO Profiles Analyze DataSet

# Read and Locate Data from NCBI GEO

Int J Colorectal Dis (2011) 26:847–858
DOI 10.1007/s00384-011-1176-x

ORIGINAL ARTICLE

## Molecular profiles and clinical outcome of stage UICC II colon cancer patients

Jörn Gröne · Dido Lenze · Vindi Jurinovic · Manuela Hummel · Henrik Seidel · Gabriele Leder · Georg Beckmann · Anette Sommer · Robert Grützmann · Christian Pilarsky · Ulrich Mansmann · Heinz-Johannes Buhr · Harald Stein · Michael Hummel

Accepted: 3 March 2011 / Published online: 5 April 2011


**Abstract**
*Purpose* Published multigene classifiers suggesting outcome prediction for patients with stage UICC II colon cancer have not been translated into a clinical application so far. Therefore, we aimed at validating own and published gene expression signatures employing methods which enable their reconstruction in routine diagnostic specimens.
*Methods* Immunohistochemistry was applied to 68 stage UICC II colon cancers to determine the protein expression of previously published prognostic classifier genes (CDH17, LAT, CA2, EMR3, and TNFRSF11A). RNA from macrodissected tumor samples from 53 of these 68 patients was profiled on Affymetrix GeneChips (HG-U133 Plus 2.0). Prognostic signatures were generated by "nearest shrunken centroids" with cross-validation. Previously published gene signatures were applied to our data set using "global tests" and leave-one-out cross-validation
*Results* Correlation of protein expression with clinical outcome failed to separate patients with disease-free follow-up (group DF) and relapse (group R). Although gene expression profiling allowed the identification of differentially expressed genes ("DF" vs. "R"), a stable classification/prognosis signature was not discernable. Furthermore, the application of previously published gene signatures to our data was unable to predict clinical outcome (prediction rate 75.5% and 64.2%; n.s.). T-stage was the only independent prognostic factor for relapse with established clinical and pathological parameters including microsatellite status (multivariate analysis).
*Conclusions* Our protein and gene expression analyses do not support application of molecular classifiers for prediction of clinical outcome in current routine diagnostic as a basis for patient-orientated therapy in stage UICC II colon cancer. Further studies are needed to develop prognosis signatures applicable in patient care.

**Keywords** Colon cancer · Immunohistochemistry · Gene expression signature · Prognosis

J. Gröne (✉) · H.-J. Buhr
Department of General, Vascular and Thoracic Surgery, Charité University Medicine Berlin,
Campus Benjamin Franklin, Hindenburgdamm 30,
12200 Berlin, Germany
e-mail: joern.groene@charite.de

D. Lenze · H. Stein · M. Hummel
Institute of Pathology, Charité University Medicine Berlin,
Campus Benjamin Franklin, Hindenburgdamm 30,
12200 Berlin, Germany

V. Jurinovic · U. Mansmann
Institut für Medizinische Informatik Biometrie
Epidemiologie (IBE),
Munich, Germany

M. Hummel
Core Facilities-Microarray Unit, Centre for Genomic Regulation,
C/Dr. Aiguader 88,
08003 Barcelona, Spain

H. Seidel · G. Leder · G. Beckmann · A. Sommer
Target Discovery, Bayer Schering Pharma AG,
Müllerstr. 178,
13353 Berlin, Germany

R. Grützmann · C. Pilarsky
Department of Visceral, Thoracic and Vascular Surgery,
University Hospital Dresden,
Fetscherstr. 74,
01307 Dresden, Germany

parameters (age, gender, tumor localization, grading, T-stage, microsatellite status), available scores were then tested in multivariate Cox regression analysis. Correlation of expression of selected proteins (CDH17 and EMR3, one probe set each; TNFRSF11A and LAT, two probe sets each) and corresponding RNA expression data was demonstrated by scatter plots.

### Microarray analyses

*Tumor sample preparation and array hybridization* For microarray analyses, snap frozen tissue specimens were cut into 7-µm-thick sections that were stained with H&E. Stained sections were reviewed by a pathologist to identify areas of vital tumor cells and to ensure a tumor content of 80–90%. Corresponding tumor areas were macrodissected by vertical 3-mm incision into the frozen tissue with a sterile blade. Incision was followed by a series of ten 20-µm frozen sections. Separated tumor areas were harvested by sterile micropipette tip and collected in buffer (RLT buffer, RNeasy Mini Kit; Qiagen, Hilden, Germany). Each series of ten sections was followed by a 7-µm H&E-stained section to control tissue composition. The number of tissue sections used to extract RNA was dependent on the expanse of the area of individual tumor tissue.

Total RNA was isolated using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions and quantified using the Nanodrop ND-1000 UV–vis spectrophotometer (Nanodrop Technologies, USA). The quality of the RNA was controlled using the BioAnalyzer (Agilent Technologies, USA), and exclusively high quality RNA (RIN≥7.6) was used for further analysis. For Affymetrix GeneChip analysis, 3 µg total RNA of each sample was converted to biotin-labeled cRNA and hybridized on HG-U133 Plus 2.0 arrays (Affymetrix, USA), following the manufacturer's recommendations.

*Microarray data analysis* The quality of all microarrays was reviewed by inspection of scatter plots (MvA plots) [25]. Variation of non-biological origin between the arrays were reduced by normalization (variance stabilization) using the *vsn* package in R (language and environment for statistical computing and graphics). "vsn" is a robust method for normalization of large-scale gene expression data. When running experiments that involve multiple high-density oligonucleotide arrays, it is important to remove sources of variation between arrays of non-biological origin. Normalization is a process for reducing this variation that works also on values that are negative after background subtraction [10]. For construction of a classifier for relapse (yes/no), the method of "nearest shrunken centroids" was applied [26] based on all stage UICC II patients and on the subgroup of microsatellite stable (MSS) patients. To avoid overfitting, a repeated double cross-validation procedure was used [27]. The data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE18088 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=hncvtygaygqmghg&acc=GSE18088).

Data of previously published prognostic gene expression signatures involving patients with stage UICC II colon cancer were analyzed by testing their power to separate between patients with relapse or disease-free patients in our data set using "global test." This test can determine whether the global expression pattern of a group of genes is significantly related to clinical variable [28] (Table 2). The two data sets of Lin et al. [22] were validated as published by the authors (New Zealand data: support vector machine; German data set: three nearest neighbor classificator, leave-one-out cross-validation, permutation approach).

### Results

Our study comprised paraffin-embedded and formalin-fixed tissues from 68 patients all of which have been employed for immunohistochemistry (IHC) detection of protein expression ("protein collection"). In addition, frozen tissue specimens were available for 53 of these 68 patients (78%).

# Querying NCBI GEO by GSE ID

## Gene Expression Omnibus

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.

Query by GSE ID

GSE18088

Search

### Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

### Tools

- 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
- FTP Site

### Browse Content

- Repository Browser

| | |
|---|---|
| DataSets: | 4348 |
| Series: | 74244 |
| Platforms: | 16460 |
| Samples: | 1960540 |

### Information for Submitters

- My GEO Submissions
- My GEO Profile
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

NCBI > GEO > **Accession Display** Not logged in | Login

**GEO help:** Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE18088 GO

## Series GSE18088

Query DataSets for GSE18088

| | |
|---|---|
| Status | Public on Apr 10, 2011 |
| Title | Correlation of molecular profiles and clinical outcome of stage UICC II colon cancer patients |
| Organism | Homo sapiens |
| Experiment type | Expression profiling by array |
| Summary | Background Published multi-gene classifiers suggested outcome prediction for patients with stage UICC II colon cancer based on different gene expression signatures. However, there is currently no translation of these classifiers for application in routine diagnostic. Therefore, we aimed at validating own and published gene expression signatures employing methods which enable RNA and protein detection in routine diagnostic specimens. Results Immunohistochemistry was applied to 68 stage UICC II colon cancers to determine the protein expression of five selected previously published classifier genes (CDH17, LAT, CA2, EMR3, and TNFRSF11A). Correlation of protein expression data with clinical outcome within a 5-year post-surgery course failed to separate patients with a disease-free follow-up [Group DF] and relapse [Group R]). In addition, RNA from macrodissected tumor samples from 53 of these 68 patients was profiled on Affymetrix GeneChips (HG-U133 Plus 2.0). Prognostic signatures were generated by Nearest Shrunken Centroids with cross-validation. Although gene expression profiling allowed the identification of differentially expressed genes between the groups DF and R, a stable classification and prognosis signature was not discernable in our data. Furthermore, the application of previously published gene signatures consisting of 22 and 19 genes, respectively, to our gene expression data set using 'global tests' and leave-one-out cross-validation was unable to predict clinical outcome (prediction rate 75.5% and 64.2%; n.s.). T-stage was the only independent prognostic factor for relapse in multivariate analysis with established clinical and pathological parameters including microsatellite status. Conclusions Our protein and gene expression analyses currently do not support application of molecular classifiers for prediction of clinical outcome in routine diagnostic as a basis for patient-orientated therapy in stage UICC II colon cancer. Further studies are needed to develop prognosis signatures applicable in patient care. |
| Overall design | 53 patients with primary stage UICC II colon cancer treated by elective standard oncological resection were selected. None of the patients received adjuvant chemotherapy. Patients with susceptibility for hereditary colorectal cancer or inflammatory bowel disease were excluded from this study. Routine histopathologic staging of resected specimen was performed by experienced pathologists. |
| Contributor(s) | Gröne J, Lenze D, Jurinovic V, Hummel M, Seidel H, Leder G, Beckmann G, Sommer A, Grützmann R, Pilarsky C, Mansmann U, Buhr H, Stein H, Hummel M |
| Citation(s) | Gröne J, Lenze D, Jurinovic V, Hummel M et al. Molecular profiles and clinical outcome of stage UICC II colon cancer patients. *Int J Colorectal Dis* 2011 Jul;26(7):847-58. PMID: 21465190 |

| | |
|---|---|
| Submission date | Sep 11, 2009 |
| Last update date | Sep 12, 2013 |
| Contact name | Dido Lenze |
| E-mail | dido.lenze@charite.de |
| Organization name | Charité-Universitätsmedizin Berlin |
| Department | Pathologie, Campus Benjamin Franklin |
| Street address | Hindenburgdamm 30 |
| City | Berlin |
| ZIP/Postal code | 12200 |
| Country | Germany |

Platforms (1) GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (53) GSM452148 C1
+ More... GSM452149 C2
GSM452150 C3

**Relations**

BioProject PRJNA119367

**Analyze with GEO2R**

| Download family | Format |
|---|---|
| SOFT formatted family file(s) | SOFT |
| MINiML formatted family file(s) | MINiML |
| Series Matrix File(s) | TXT |

| Supplementary file | Size | Download | File type/resource |
|---|---|---|---|
| GSE18088_RAW.tar | 243.6 Mb | (http)(custom) | TAR (of CEL) |

*Raw data provided as supplementary file*

*Processed data included within Sample table*

**Sample GSM452148** Query DataSets for GSM452148

| | |
|---|---|
| Status | Public on Apr 10, 2011 |
| Title | C1 |
| Sample type | RNA |
| Source name | colon |
| Organism | Homo sapiens |
| Characteristics | localization: proximal<br>gender: male<br>relapse: no<br>microsatellite status: MSI-high<br>age at diagnosis, years: 61<br>grading: G2<br>pt: 3 |
| Extracted molecule | total RNA |
| Extraction protocol | For microarray analyses, snap frozen tissue specimens were cut into 7 µm-thick sections that were stained with haematoxylin & eosin (H&E). Stained sections were reviewed by a pathologist to identify areas of vital tumor cells and to ensure a tumor content of 80-90%. Corresponding tumor areas were macrodissected by vertical 3 mm incision into the frozen tissue specimens with a sterile blade. Incision was followed by a series of ten 20 µm frozen sections. Separated tumor areas were harvested by sterile micropipette tip and collected in a buffer (RLT buffer, RNeasy Mini Kit; Qiagen, Hilden, Germany). Each series of ten sections was followed by a 7 µm H&E stained section to control tissue composition. The number of tissue sections used to extract RNA was dependent on the expanse of the area of individual tumor tissue. Total RNA was isolated using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions and quantified using the Nanodrop ND-1000 UV-vis spectrophotometer (Nanodrop Technologies, USA). The quality of the RNA was controlled using the BioAnalyzer (Agilent Technologies, USA) and exclusively high quality RNA was used for further analysis. |
| Label | biotin |
| Label protocol | Biotinylated cRNA were prepared according to the standard Affymetrix protocol from 3ug total RNA |
| Hybridization protocol | Following fragmentation, 10 ug of cRNA were hybridized for 16 hr at 45C on HG-U133 Plus 2.0 arrays (Affymetrix). GeneChips were washed and stained with standard protocols in the Affymetrix Fluidics Station 450. |
| Scan protocol | GeneChips were scanned using the Scanner G3000 |
| Description | none |
| Data processing | Data were normalized with VSN (Bioconductor package vsn 3.2.1) and summarized with the median polish method. Data analysis and statistical computing were performed with R (2.6.0). |

| | | |
|---|---|---|
| Platform ID | GPL570 | |
| Series (1) | GSE18088 | Correlation of molecular profiles and clinical outcome of stage UICC II colon cancer patients |

**Data table header descriptions**

| | |
|---|---|
| **ID_REF** | |
| **VALUE** | normalized log2 signal intensity |

**Data table**

| ID_REF | VALUE |
|---|---|
| 1007_s_at | 7.558720057 |
| 1053_at | 6.750650907 |
| 117_at | 5.71742226 |
| 121_at | 6.437156704 |
| 1255_g_at | 4.07892685 |
| 1294_at | 6.3153491 |
| 1316_at | 4.739850728 |
| 1320_at | 4.81556091 |
| 1405_i_at | 7.221498026 |
| 1431_at | 3.814676333 |
| 1438_at | 6.472361804 |
| 1487_at | 6.680293336 |
| 1494_f_at | 4.565978445 |
| 1552256_a_at | 6.826440737 |
| 1552257_a_at | 6.726811022 |
| 1552258_at | 4.075673415 |
| 1552261_at | 4.763925276 |
| 1552263_at | 5.262063729 |
| 1552264_a_at | 6.742858328 |
| 1552266_at | 4.125112975 |

Total number of rows: **54675**

Table truncated, full table size **1211 Kbytes**.

View full table...

| Supplementary file | Size | Download | File type/resource |
|---|---|---|---|
| GSM452148.CEL.gz | 4.1 Mb | (ftp)(http) | CEL |

*Raw data provided as supplementary file*

*Processed data included within Sample table*

Platform ID GPL570

Series (1) GSE28146 Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease

**Relations**

Affiliated with GSM21215

**Data table header descriptions**

**ID_REF** Probe Set

**VALUE** MAS5-calculated Signal intensity

**PROBABILITY** the probability that a probe set is absent, ranging from 0 to 1 (in this study, probe set with $p < 0.05$ were considered present)

**Data table**

| ID_REF | VALUE | PROBABILITY |
|---|---|---|
| 215306_at | 123.3 | 0.366211 |
| 214927_at | 143.1 | 0.303711 |
| 211606_at | 85.8 | 0.432373 |
| 212678_at | 363.5 | 0.023926 |
| 213929_at | 26.8 | 0.466064 |
| 213567_at | 648.6 | 0.010742 |
| 217011_at | 13.2 | 0.904785 |
| 217483_at | 180.8 | 0.056152 |
| 215769_at | 14.8 | 0.870361 |
| 222271_at | 236.9 | 0.149658 |
| 217536_x_at | 1839.9 | 0.000244 |
| 216539_at | 12.5 | 0.828613 |
| 216158_at | 306.8 | 0.018555 |
| 216394_x_at | 111.7 | 0.870361 |
| 217136_at | 26.8 | 0.432373 |
| 214722_at | 1136.6 | 0.129639 |
| 216965_x_at | 344.2 | 0.095215 |
| 212017_at | 1056 | 0.111572 |
| 213392_at | 3198.6 | 0.001221 |

Total number of rows: **54675**

Table truncated, full table size **1344 Kbytes**.

# Different Normalization Method (MAS5.0)

| | | |
|---|---|---|
| Platform ID | GPL550 | |
| Series (2) | GSE763 | Cell-type specific responses to chemotherapeutics in breast cancer |
| | GSE1647 | Prediction of toxicant-specific gene expression signatures following chemotherapeutic treatment of breast cell lines |

**Data table header descriptions**

| | |
|---|---|
| **ID_REF** | |
| **VALUE** | same as UNF_VALUE but with flagged values removed |
| **SPOT** | spot number on array |
| **CH1_MEAN** | channel 1 mean intensity |
| **CH1_SD** | standard deviation of channel 1 intensity |
| **CH1_BKD_MEDIAN** | channel 1 background median intensity |
| **CH1_BKD_SD** | standard deviation of channel 1 background median intensity |
| **CH2_MEAN** | channel 2 mean intensity |
| **CH2_SD** | standard deviation of channel 2 intensity |
| **CH2_BKD_MEDIAN** | channel 2 background median intensity |
| **CH2_BKD_SD** | standard deviation of channel 2 background median intensity |
| **TOT_BPIX** | number of background pixels |
| **TOT_SPIX** | number of spot pixels |
| **CH2BN_MEDIAN** | channel 2 normalized background median intensity |
| **CH2IN_MEAN** | channel 2 normalized mean intensity |
| **CH1DL_MEAN** | channel 1 Lowess_normalized mean intensity |
| **CH2DL_MEAN** | channel 2 Lowess_normalized mean intensity |
| **LOG_RAT2N_MEAN** | log2_ratio of (CH2IN_MEAN - CH2BN_MEDIAN) over (CH1_MEAN - CH1_BKD_MEDIAN), CH2IN_MEAN and CH2BN_MEDIAN are global-normalized intensities |
| **CORR** | correlation coefficient among pixels |
| **FLAG** | Spot flag. 0:not flagged; negative:flagged as bad spots; positive:flagged as good spots |
| **CONTROL** | Y: control gene; N: not control |
| **UNF_VALUE** | LOG_RAT2L_MEAN; log2_ratio of CH2DL_MEAN over CH1DL_MEAN |

**Data table**

| ID_REF | VALUE | SPOT | CH1_MEAN | CH1_SD | CH1_BKD_MEDIAN | CH1_BKD_SD | CH2_MEAN | CH2_SD | CH2_BKD_MEDIAN | CH2_BKD_SD | TOT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .027 | 3480 | 103 | 27 | 94 | 30 | 80 | 43 | 59 | 26 | 540 |
| 2 | -.389 | 2598 | 153 | 38 | 139 | 31 | 88 | 45 | 66 | 23 | 579 |
| 3 | -1.117 | 3482 | 97 | 29 | 90 | 30 | 64 | 24 | 55 | 18 | 527 |
| 4 | 1.132 | 2600 | 142 | 34 | 138 | 31 | 87 | 36 | 64 | 21 | 592 |
| 5 | -1.168 | 3484 | 108 | 30 | 91 | 24 | 73 | 33 | 57 | 19 | 567 |
| 6 | -.254 | 2602 | 179 | 57 | 136 | 30 | 115 | 63 | 64 | 21 | 543 |
| 7 | .485 | 3486 | 464 | 376 | 89 | 23 | 709 | 677 | 56 | 19 | 641 |
| 8 | .05 | 2604 | 240 | 115 | 135 | 31 | 196 | 150 | 62 | 21 | 728 |
| 9 | | 7008 | 103 | 28 | 109 | 27 | 73 | 29 | 62 | 20 | 548 |
| 10 | -1.402 | 6126 | 178 | 62 | 146 | 31 | 88 | 41 | 67 | 21 | 543 |
| 11 | .575 | 7010 | 165 | 67 | 109 | 27 | 165 | 118 | 58 | 22 | 556 |
| 12 | -.08 | 6128 | 195 | 73 | 145 | 32 | 130 | 96 | 66 | 22 | 512 |
| 13 | .524 | 7012 | 112 | 27 | 108 | 26 | 76 | 32 | 60 | 20 | 569 |
| 14 | -1.166 | 6130 | 185 | 55 | 145 | 30 | 94 | 42 | 66 | 23 | 564 |
| 15 | .376 | 7014 | 247 | 154 | 109 | 29 | 274 | 229 | 57 | 20 | 712 |
| 16 | -.896 | 6132 | 162 | 45 | 140 | 138 | 89 | 43 | 67 | 27 | 713 |
| 17 | .65 | 10536 | 3277 | 1870 | 118 | 31 | 7460 | 4435 | 60 | 44 | 565 |
| 18 | -.429 | 9654 | 177 | 58 | 140 | 34 | 105 | 62 | 64 | 21 | 515 |
| 19 | .448 | 10538 | 4261 | 2807 | 117 | 49 | 8607 | 6005 | 61 | 99 | 542 |

Total number of rows: **20163**

# Other Platform (different array)

# Other formats (RNAseq)

**Series GSE48213** Query DataSets for GSE48213

| | |
|---|---|
| Status | Public on Aug 20, 2013 |
| Title | Transcriptional profiling of a breast cancer cell line panel using RNAseq technology |
| Organism | Homo sapiens |
| Experiment type | Expression profiling by high throughput sequencing |
| Summary | 56 breast cancer cell lines were profiled to identify patterns of gene expression associated with subtype and response to therapeutic compounds. |
| Overall design | Cell lines were profiled in their baseline, unperturbed state. |
| Contributor(s) | Gray JW, Heiser LM |
| Citation missing | *Has this study been published? Please login to update or notify GEO.* |
| Submission date | Jun 21, 2013 |
| Last update date | Aug 21, 2013 |
| Contact name | Laura M. Heiser |
| E-mail | heiserl@ohsu.edu |
| Organization name | OHSU |
| Street address | 3181 SW Sam Jackson Park Rd. |
| City | Portland |
| State/province | OR |
| ZIP/Postal code | 97239 |
| Country | USA |

Platforms (1) GPL10999 Illumina Genome Analyzer IIx (Homo sapiens)

Samples (56) GSM1172844 184A1, RNA-Seq
More... GSM1172845 184B5, RNA-Seq
GSM1172846 21MT1, RNA-Seq

This SubSeries is part of SuperSeries:
GSE48216 Modeling precision treatment of breast cancer

**Relations**

| | |
|---|---|
| BioProject | PRJNA210428 |
| SRA | SRP026537 |

| Download family | Format |
|---|---|
| SOFT formatted family file(s) | SOFT |
| MINiML formatted family file(s) | MINiML |
| Series Matrix File(s) | TXT |

| Supplementary file | Size | Download | File type/resource |
|---|---|---|---|
| GSE48213_RAW.tar | 13.5 Mb | (http)(custom) | TAR (of TXT) |
| SRP/SRP026/SRP026537 | | (ftp) | SRA Study |

*Raw data provided as supplementary file*

*Processed data provided as supplementary file*

**Series GSE4335** Query DataSets for GSE4335

| | |
|---|---|
| Status | Public on Mar 07, 2006 |
| Title | Norway/Stanford Breast Tumors |
| Organism | Homo sapiens |
| Experiment type | Expression profiling by array |
| Summary | Characterization of patterns of gene expression measured by cDNA microarrays to subclassify tumors into clinically relevant subgroups. In this study, we have refined the previously defined subtypes of breast tumors that could be distinguished by their distinct patterns of gene expression. A total of 115 malignant breast tumors and 7 benign tissues were analyzed by hierarchical clustering based on patterns of expression of 534 "intrinsic" genes and shown to subdivide into a basal epithelial-like, an ERBB2-overexpressing, two luminal epithelial-like and a normal breast tissue-like subgroup. The genes used for classification were selected based on their similar expression levels between pairs of consecutive samples taken from the same tumor separated by 15 weeks of neoadjuvant treatment.<br>A disease state experiment design type is where the state of some disease such as infection, pathology, syndrome, etc is studied.<br>Keywords: disease_state_design |
| Overall design | Computed |
| Web link | http://genome-www.stanford.edu/breast_cancer/robustness/ |
| Contributor(s) | Perou C, Sorlie T |
| Citation(s) | Sorlie T, Tibshirani R, Parker J, Hastie T et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003 Jul 8;100(14):8418-23. PMID: 12829800 |
| Submission date | Mar 03, 2006 |
| Last update date | Mar 17, 2012 |
| Organization | Stanford Microarray Database (SMD) |
| E-mail | array@genome.stanford.edu |
| Phone | 650-498-6012 |
| URL | http://genome-www5.stanford.edu/ |
| Department | Stanford University, School of Medicine |
| Street address | 300 Pasteur Drive |
| City | Stanford |
| State/province | CA |
| ZIP/Postal code | 94305 |
| Country | USA |
| Platforms (7) ⊟ Less... | GPL180 SVC<br>GPL2776 SVL_SVM_SVN_SVO<br>GPL2777 SVJ<br>GPL2778 SHAC<br>GPL3045 SHBG<br>GPL3047 SHBY<br>GPL3147 SHAZ |
| Samples (122) ⊞ More... | GSM1844 BC402B-BE<br>GSM1845 BC709B-BE<br>GSM1846 BC107B-BE |

# Different Array layouts within a Data Set

# Types of Data Available for Download

Platforms (1) GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (53) GSM452148 C1
More... GSM452149 C2
GSM452150 C3

**Relations**
BioProject PRJNA119367

**Analyze with GEO2R**

| Download family | Format |
|---|---|
| SOFT formatted family file(s) | SOFT |
| MINiML formatted family file(s) | MINiML |
| Series Matrix File(s) | TXT |

| Supplementary file | Size | Download | File type/resource |
|---|---|---|---|
| GSE18088_RAW.tar | 243.6 Mb | (http)(custom) | TAR (of CEL) |

*Raw data provided as supplementary file*
*Processed data included within Sample table*

**SOFT data**

**MINiML data**

**SERIES data**

**RAW data**

# SOFT Format

SOFT/: This directory contains files in “Simple Omnibus Format in Text” (SOFT). SOFT files are generated for DataSet entries, as well as for Series and Platform entries (subdirectories are included for each entry type). The Series and Platform files are actually “family files” that include the metadata and complete data tables of all related entries in the family. In contrast, the DataSet SOFT files include the metadata of the DataSet entry only, plus a matrix table containing the extracted gene annotations and Sample values used in GEO Profiles.

# MINiML format

MINiML/: This directory includes files in MINiML (MIAME Notation in Markup Language) format. MINiML is essentially an XML rendering of SOFT format, and the files provided here are the XML-equivalents of the Series and Platform family files provided in the SOFT/ directory.

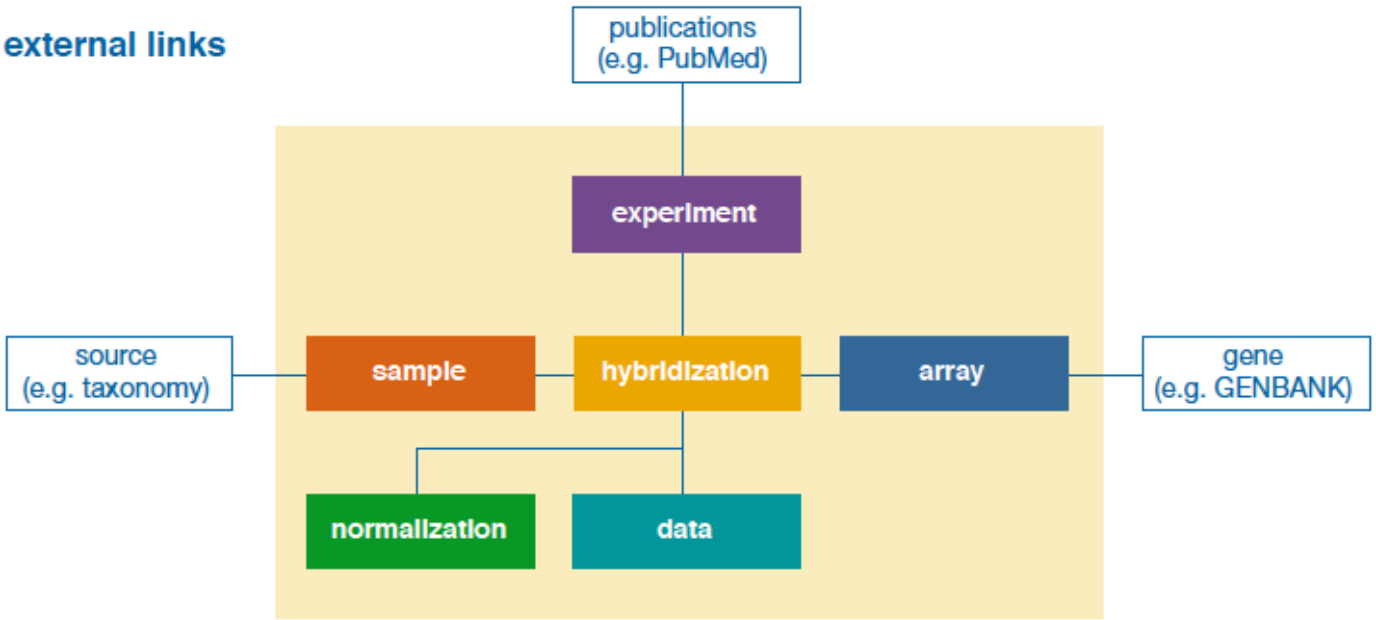



## Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma[1], Pascal Hingamp[2], John Quackenbush[3], Gavin Sherlock[4], Paul Spellman[5], Chris Stoeckert[6], John Aach[7], Wilhelm Ansorge[8], Catherine A. Ball[4], Helen C. Causton[9], Terry Gaasterland[10], Patrick Glenisson[11], Frank C.P. Holstege[12], Irene F. Kim[4], Victor Markowitz[13], John C. Matese[4], Helen Parkinson[1], Alan Robinson[1], Ugis Sarkans[1], Steffen Schulze-Kremer[14], Jason Stewart[15], Ronald Taylor[16], Jaak Vilo[1] & Martin Vingron[17]

Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

# MIAME

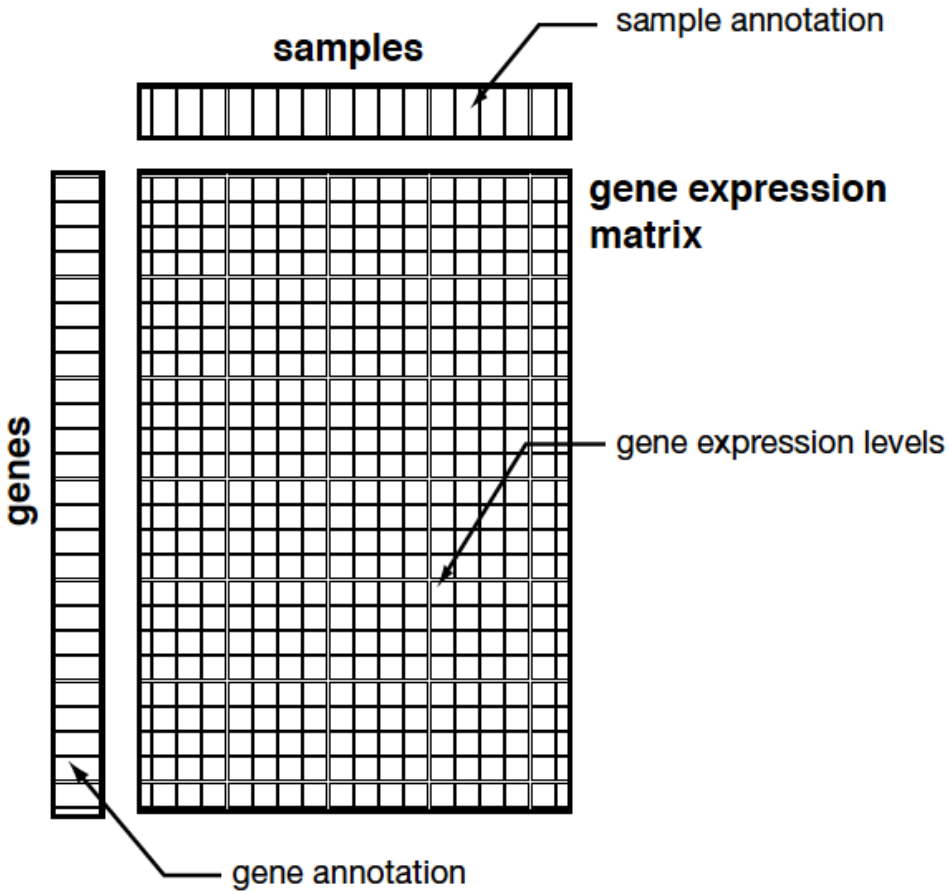



**Six Parts of MIAME**

1. Experimental design: the set of hybridization experiments as a whole
2. Array design: each array used and each element (spot, feature) on the array
3. Samples: samples used, extract preparation and labeling
4. Hybridizations: procedures and parameters
5. Measurements: images, quantification and specifications
6. Normalization controls: types, values and specifications

# Series Matrix Format

SeriesMatrix/: This directory contains tab-delimited value-matrices generated from the VALUE column of the Sample tables of each Series entry. Files also include Series and Sample metadata and are ideal for opening in spreadsheet applications such as MicrosoftExcel. Most users find SeriesMatrix files the most convenient format for handling data that have not been assembled into a DataSet



# Series Information

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | !Series_title | Correlation of molecular profiles and clinical outcome of stage UICC II colon cancer patients | | | | | | | |
| 2 | !Series_geo_accession | GSE18088 | | | | | | | |
| 3 | !Series_status | Public on Apr 10 2011 | | | | | | | |
| 4 | !Series_submission_date | Sep 11 2009 | | | | | | | |
| 5 | !Series_last_update_date | Sep 12 2013 | | | | | | | |
| 6 | !Series_pubmed_id | 21465190 | | | | | | | |
| 7 | !Series_summary | Background  Published multi-gene classifiers suggested outcome prediction for patients with stage UICC II colon cancer based on different gene ex | | | | | | | |
| 8 | !Series_overall_design | 53 patients with primary stage UICC II colon cancer treated by elective standard oncological resection were selected. None of the patients receivec | | | | | | | |
| 9 | !Series_type | Expression profiling by array | | | | | | | |
| 10 | !Series_contributor | JÃ¶rn,,GrÃ¶ne | | | | | | | |
| 11 | !Series_contributor | Dido,,Lenze | | | | | | | |
| 12 | !Series_contributor | Vindi,,Jurinovic | | | | | | | |
| 13 | !Series_contributor | Manuela,,Hummel | | | | | | | |
| 14 | !Series_contributor | Henrik,,Seidel | | | | | | | |
| 15 | !Series_contributor | Gabriele,,Leder | | | | | | | |
| 16 | !Series_contributor | Georg,,Beckmann | | | | | | | |
| 17 | !Series_contributor | Anette,,Sommer | | | | | | | |
| 18 | !Series_contributor | Robert,,GrÃ¼tzmann | | | | | | | |
| 19 | !Series_contributor | Christian,,Pilarsky | | | | | | | |
| 20 | !Series_contributor | Ulrich,,Mansmann | | | | | | | |
| 21 | !Series_contributor | Heinz-Johannes,,Buhr | | | | | | | |
| 22 | !Series_contributor | Harald,,Stein | | | | | | | |
| 23 | !Series_contributor | Michael,,Hummel | | | | | | | |
| 24 | !Series_sample_id | GSM452148 GSM452149 GSM452150 GSM452151 GSM452152 GSM452153 GSM452154 GSM452155 GSM452156 GSM452157 GSM452158 GSM4 | | | | | | | |
| 25 | !Series_contact_name | Dido,,Lenze | | | | | | | |
| 26 | !Series_contact_email | dido.lenze@charite.de | | | | | | | |
| 27 | !Series_contact_department | Pathologie, Campus Benjamin Franklin | | | | | | | |
| 28 | !Series_contact_institute | CharitÃ©-UniversitÃ¤tsmedizin Berlin | | | | | | | |
| 29 | !Series_contact_address | Hindenburgdamm 30 | | | | | | | |
| 30 | !Series_contact_city | Berlin | | | | | | | |
| 31 | !Series_contact_zip/postal_code | 12200 | | | | | | | |
| 32 | !Series_contact_country | Germany | | | | | | | |
| 33 | !Series_supplementary_file | ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/series/GSE18088/GSE18088_RAW.tar | | | | | | | |
| 34 | !Series_platform_id | GPL570 | | | | | | | |
| 35 | !Series_platform_taxid | 9606 | | | | | | | |
| 36 | !Series_sample_taxid | 9606 | | | | | | | |
| 37 | !Series_relation | BioProject: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA119367 | | | | | | | |

# Sample Descriptions

| | | | | |
|---|---|---|---|---|
| 39 | !Sample_title | C1 | C2 | C3 |
| 40 | !Sample_geo_accession | GSM452148 | GSM452149 | GSM452150 |
| 41 | !Sample_status | Public on Apr 10 2011 | Public on Apr 10 2011 | Public on Apr 10 2011 |
| 42 | !Sample_submission_date | Sep 11 2009 | Sep 11 2009 | Sep 11 2009 |
| 43 | !Sample_last_update_date | Apr 10 2011 | Apr 10 2011 | Apr 10 2011 |
| 44 | !Sample_type | RNA | RNA | RNA |
| 45 | !Sample_channel_count | 1 | 1 | 1 |
| 46 | !Sample_source_name_ch1 | colon | colon | colon |
| 47 | !Sample_organism_ch1 | Homo sapiens | Homo sapiens | Homo sapiens |
| 48 | !Sample_characteristics_ch1 | localization: proximal | localization: distal | localization: distal |
| 49 | !Sample_characteristics_ch1 | gender: male | gender: female | gender: female |
| 50 | !Sample_characteristics_ch1 | relapse: no | relapse: no | relapse: no |
| 51 | !Sample_characteristics_ch1 | microsatellite status: MS | microsatellite status: MSS | microsatellite status: M |
| 52 | !Sample_characteristics_ch1 | age at diagnosis, years: 6 | age at diagnosis, years: 65 | age at diagnosis, years |
| 53 | !Sample_characteristics_ch1 | grading: G2 | grading: G3 | grading: G3 |
| 54 | !Sample_characteristics_ch1 | pt: 3 | pt: 3 | pt: 3 |
| 55 | !Sample_molecule_ch1 | total RNA | total RNA | total RNA |
| 56 | !Sample_extract_protocol_ch1 | For microarray analyses, | For microarray analyses, snap fr | For microarray analyse |
| 57 | !Sample_label_ch1 | biotin | biotin | biotin |
| 58 | !Sample_label_protocol_ch1 | Biotinylated cRNA were | Biotinylated cRNA were prepare | Biotinylated cRNA wer |
| 59 | !Sample_taxid_ch1 | 9606 | 9606 | 9606 |
| 60 | !Sample_hyb_protocol | Following fragmentatior | Following fragmentation, 10 ug | Following fragmentatic |
| 61 | !Sample_scan_protocol | GeneChips were scanne | GeneChips were scanned using t | GeneChips were scann |
| 62 | !Sample_description | none | none | none |
| 63 | !Sample_data_processing | Data were normalized w | Data were normalized with VSN | Data were normalized |
| 64 | !Sample_platform_id | GPL570 | GPL570 | GPL570 |
| 65 | !Sample_contact_name | Dido,,Lenze | Dido,,Lenze | Dido,,Lenze |
| 66 | !Sample_contact_email | dido.lenze@charite.de | dido.lenze@charite.de | dido.lenze@charite.de |
| 67 | !Sample_contact_department | Pathologie, Campus Ben | Pathologie, Campus Benjamin Fr | Pathologie, Campus Be |
| 68 | !Sample_contact_institute | CharitÃ©-UniversitÃ¤tsr | CharitÃ©-UniversitÃ¤tsmedizin | CharitÃ©-UniversitÃ¤t |
| 69 | !Sample_contact_address | Hindenburgdamm 30 | Hindenburgdamm 30 | Hindenburgdamm 30 |
| 70 | !Sample_contact_city | Berlin | Berlin | Berlin |
| 71 | !Sample_contact_zip/postal_code | 12200 | 12200 | 12200 |
| 72 | !Sample_contact_country | Germany | Germany | Germany |
| 73 | !Sample_supplementary_file | ftp://ftp.ncbi.nlm.nih.go | ftp://ftp.ncbi.nlm.nih.gov/pub/g | ftp://ftp.ncbi.nlm.nih.g |
| 74 | !Sample_data_row_count | 54675 | 54675 | 54675 |

# Start of the table (!series_matrix_table_begin)

| | | | | | |
|---|---|---|---|---|---|
| 75 | !series_matrix_table_begin | | | | |
| 76 | ID_REF | GSM452148 | GSM452149 | GSM452150 | GSM452151 |
| 77 | 1007_s_at | 7.558720057 | 9.126071305 | 8.734507244 | 8.25817423 |
| 78 | 1053_at | 6.750650907 | 6.601878355 | 6.923114063 | 6.93849575 |
| 79 | 117_at | 5.71742226 | 5.605266492 | 5.015487439 | 5.19466256 |
| 80 | 121_at | 6.437156704 | 6.434629124 | 6.392032573 | 6.62520124 |
| 81 | 1255_g_at | 4.07892685 | 3.713068215 | 3.794982838 | 3.6812607 |
| 82 | 1294_at | 6.3153491 | 6.306886327 | 6.049185465 | 6.76594234 |
| 83 | 1316_at | 4.739850728 | 4.970590666 | 4.956611321 | 4.75673162 |
| 84 | 1320_at | 4.81556091 | 4.712215864 | 4.801955899 | 4.92741649 |
| 85 | 1405_i_at | 7.221498026 | 5.352830756 | 5.790997115 | 5.50934298 |
| 86 | 1431_at | 3.814676333 | 3.84964452 | 3.925627406 | 4.08005801 |
| 87 | 1438_at | 6.472361804 | 6.164358643 | 5.835383402 | 8.01234779 |
| 88 | 1487_at | 6.680293336 | 7.382490981 | 7.5320706 | 7.3796595 |
| 89 | 1494_f_at | 4.565978445 | 4.625585548 | 4.540350169 | 4.61935429 |
| 90 | 1552256_a_at | 6.826440737 | 7.688037915 | 7.309412689 | 6.93808976 |
| 91 | 1552257_a_at | 6.726811022 | 7.425983787 | 7.664644639 | 7.65690367 |
| 92 | 1552258_at | 4.075673415 | 4.054877581 | 4.195442427 | 4.19607764 |
| 93 | 1552261_at | 4.763925276 | 4.466654127 | 4.453573101 | 4.62730665 |
| 94 | 1552263_at | 5.262063729 | 5.645246606 | 5.258001921 | 5.15663966 |
| 95 | 1552264_a_at | 6.742858328 | 6.970926255 | 6.461754975 | 6.77675133 |
| 96 | 1552266_at | 4.125112975 | 3.98170158 | 4.08473216 | 4.3781455 |
| 97 | 1552269_at | 3.789323493 | 4.135743538 | 4.426834426 | 3.86267082 |
| 98 | 1552271_at | 5.264483914 | 5.225509609 | 5.375904363 | 5.22877832 |
| 99 | 1552272_a_at | 5.284224098 | 5.168305352 | 5.153969844 | 5.10496151 |

# End of the table (!series_matrix_table_end)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 54741 | AFFX-r2-Ec-bioC-5_at | 7.602566865 | 7.976068619 | 8.443252771 | 7.86176609 | 7.51330307 |
| 54742 | AFFX-r2-Ec-bioD-3_at | 9.972820987 | 10.14050028 | 10.58824722 | 10.2102976 | 9.81051338 |
| 54743 | AFFX-r2-Ec-bioD-5_at | 9.740375139 | 9.830162642 | 10.34988347 | 10.0079244 | 9.48508604 |
| 54744 | AFFX-r2-P1-cre-3_at | 11.69035502 | 11.88191979 | 12.26538416 | 11.9396896 | 11.5212214 |
| 54745 | AFFX-r2-P1-cre-5_at | 11.56705477 | 11.73901537 | 12.19634907 | 11.9348203 | 11.3533448 |
| 54746 | AFFX-ThrX-3_at | 4.427476199 | 4.431568982 | 4.315637461 | 4.31604016 | 4.27621105 |
| 54747 | AFFX-ThrX-5_at | 3.845228081 | 3.74339138 | 3.92022782 | 3.92688029 | 3.8072115 |
| 54748 | AFFX-ThrX-M_at | 3.547658219 | 3.565593578 | 3.583823938 | 3.56543057 | 3.50535699 |
| 54749 | AFFX-TrpnX-3_at | 3.371747344 | 3.416933903 | 3.370492506 | 3.23021951 | 3.31562569 |
| 54750 | AFFX-TrpnX-5_at | 4.167570905 | 4.079075467 | 4.162768097 | 4.07082138 | 3.95383253 |
| 54751 | AFFX-TrpnX-M_at | 3.880334531 | 3.759008208 | 3.716777878 | 3.69124187 | 3.6254122 |
| 54752 | !series_matrix_table_end | | | | | |

# GEO2R Analysis Pipeline

Platforms (1) GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (53) GSM452148 C1
+ More... GSM452149 C2
GSM452150 C3

**Relations**
BioProject PRJNA119367

**GEO2R Analysis Pipeline**

**Analyze with GEO2R**

| Download family | Format |
|---|---|
| SOFT formatted family file(s) | SOFT |
| MINiML formatted family file(s) | MINiML |
| Series Matrix File(s) | TXT |

| Supplementary file | Size | Download | File type/resource |
|---|---|---|---|
| GSE18088_RAW.tar | 243.6 Mb | (http)(custom) | TAR (of CEL) |

*Raw data provided as supplementary file*
*Processed data included within Sample table*

# GEO2R - Analysis

NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » GEO2R » GSE18088

User: aikchoontan | My submissions | Logout

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** GSE18088 Set Correlation of molecular profiles and clinical outcome of stage UICC II colon cancer patients

**Samples** Define groups Selected **0** out of **53** samples

**GEO2R** | Value distribution | Options | Profile graph | R script

**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

**Top 250** Save all results

| NLM | NIH | Email GEO | Disclaimer | Accessibility |

# GEO2R – Samples in the Series

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

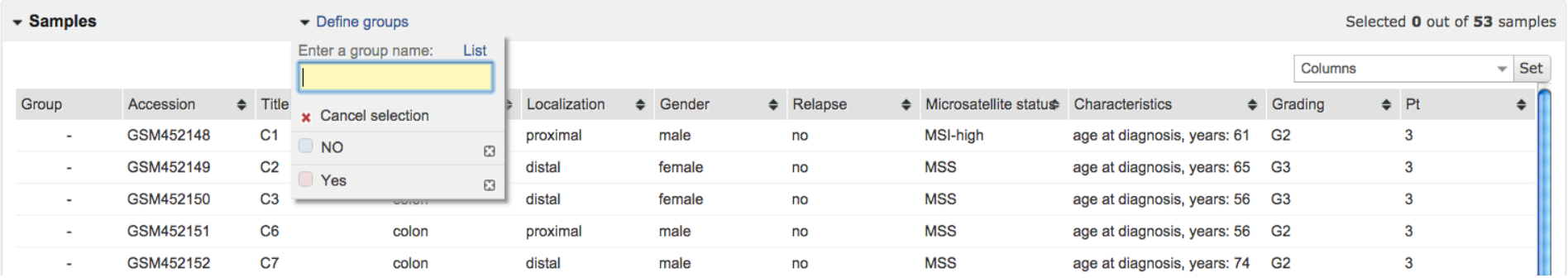
**GEO accession** GSE18088 Set Correlation of molecular profiles and clinical outcome of stage UICC II colon cancer patients

**Samples** Define groups Selected **0** out of **53** samples

Columns Set

| Group | Accession | Title | Source name | Localization | Gender | Relapse | Microsatellite status | Characteristics | Grading | Pt |
|---|---|---|---|---|---|---|---|---|---|---|
| - | GSM452148 | C1 | colon | proximal | male | no | MSI-high | age at diagnosis, years: 61 | G2 | 3 |
| - | GSM452149 | C2 | colon | distal | female | no | MSS | age at diagnosis, years: 65 | G3 | 3 |
| - | GSM452150 | C3 | colon | distal | female | no | MSS | age at diagnosis, years: 56 | G3 | 3 |
| - | GSM452151 | C6 | colon | proximal | male | no | MSS | age at diagnosis, years: 56 | G2 | 3 |
| - | GSM452152 | C7 | colon | distal | male | no | MSS | age at diagnosis, years: 74 | G2 | 3 |
| - | GSM452153 | C8 | colon | proximal | male | no | MSI-high | age at diagnosis, years: 56 | G2 | 4 |
| - | GSM452154 | C10 | colon | distal | female | no | MSS | age at diagnosis, years: 75 | G2 | 3 |
| - | GSM452155 | C11 | colon | proximal | female | no | MSS | age at diagnosis, years: 75 | G3 | 3 |
| - | GSM452156 | C13 | colon | proximal | male | no | MSS | age at diagnosis, years: 58 | G2 | 3 |
| - | GSM452157 | C20 | colon | proximal | female | no | MSI-high | age at diagnosis, years: 75 | G3 | 3 |
| - | GSM452158 | C22 | colon | proximal | male | no | MSS | age at diagnosis, years: 68 | G3 | 3 |
| - | GSM452159 | C23 | colon | proximal | female | yes | MSI-high | age at diagnosis, years: 72 | G2 | 4 |
| - | GSM452160 | C24 | colon | distal | female | no | MSS | age at diagnosis, years: 59 | G2 | 3 |
| - | GSM452161 | C25 | colon | proximal | male | yes | MSS | age at diagnosis, years: 66 | G2 | 4 |
| - | GSM452162 | C26 | colon | proximal | female | no | MSS | age at diagnosis, years: 70 | G2 | 3 |
| - | GSM452163 | C27 | colon | proximal | female | no | MSI-high | age at diagnosis, years: 74 | G3 | 4 |
| - | GSM452164 | C28 | colon | proximal | female | yes | MSS | age at diagnosis, years: 60 | G2 | 3 |
| - | GSM452165 | C29 | colon | distal | male | yes | MSS | age at diagnosis, years: 74 | G2 | 4 |
| - | GSM452166 | C30_2 | colon | proximal | female | no | MSS | age at diagnosis, years: 71 | G3 | 3 |

# GEO2R – Define Groups



# GEO2R – Define Samples

Samples

Define groups

Selected **53** out of **53** samples

Enter a group name: List

Cancel selection

NO (40 samples)

Yes (13 samples)

Columns Set

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| NO | GSM452189 | C68 | colon | distal | male | no | MSI-high | age at diagnosis, years: 85 | G2 | 4 |
| NO | GSM452193 | C74 | colon | proximal | male | no | MSS | age at diagnosis, years: 71 | G2 | 3 |
| NO | GSM452195 | C81 | colon | distal | female | no | MSS | age at diagnosis, years: 73 | G3 | 3 |
| NO | GSM452196 | C82 | colon | distal | male | no | MSS | age at diagnosis, years: 81 | G2 | 3 |
| NO | GSM452197 | C83 | colon | distal | male | no | MSI-high | age at diagnosis, years: 40 | G3 | 3 |
| NO | GSM452198 | C94 | colon | distal | male | no | MSS | age at diagnosis, years: 62 | G1 | 3 |
| NO | GSM452199 | C102 | colon | distal | female | no | MSS | age at diagnosis, years: 70 | G3 | 3 |
| Yes | GSM452159 | C23 | colon | proximal | female | yes | MSI-high | age at diagnosis, years: 72 | G2 | 4 |
| Yes | GSM452161 | C25 | colon | proximal | male | yes | MSS | age at diagnosis, years: 66 | G2 | 4 |
| Yes | GSM452164 | C28 | colon | proximal | female | yes | MSS | age at diagnosis, years: 60 | G2 | 3 |
| Yes | GSM452165 | C29 | colon | distal | male | yes | MSS | age at diagnosis, years: 74 | G2 | 4 |
| Yes | GSM452171 | C38 | colon | distal | male | yes | MSS | age at diagnosis, years: 65 | G2 | 3 |
| Yes | GSM452175 | C43 | colon | distal | male | yes | MSI-high | age at diagnosis, years: 34 | G3 | 4 |
| Yes | GSM452177 | C46 | colon | distal | male | yes | MSS | age at diagnosis, years: 81 | G3 | 3 |
| Yes | GSM452178 | C48 | colon | proximal | female | yes | MSS | age at diagnosis, years: 55 | G2 | 4 |
| Yes | GSM452190 | C69 | colon | distal | female | yes | MSS | age at diagnosis, years: 82 | G2 | 3 |
| Yes | GSM452191 | C70 | colon | distal | male | yes | MSS | age at diagnosis, years: 62 | G2 | 3 |
| Yes | GSM452192 | C71 | colon | distal | female | yes | MSS | age at diagnosis, years: 70 | G2 | 3 |
| Yes | GSM452194 | C80 | colon | distal | female | yes | MSS | age at diagnosis, years: 40 | G2 | 3 |
| Yes | GSM452200 | C113 | colon | distal | female | yes | MSI-high | age at diagnosis, years: 61 | G2 | 3 |

# GEO2R – Run Analysis

GEO2R | Value distribution | Options | Profile graph | R script

**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

Top 250 Save all results

# GEO2R – Top250 DEG

GEO2R | Value distribution | Options | Profile graph | R script

Quick start

**Recalculate** if you changed any options. Save all results Select columns

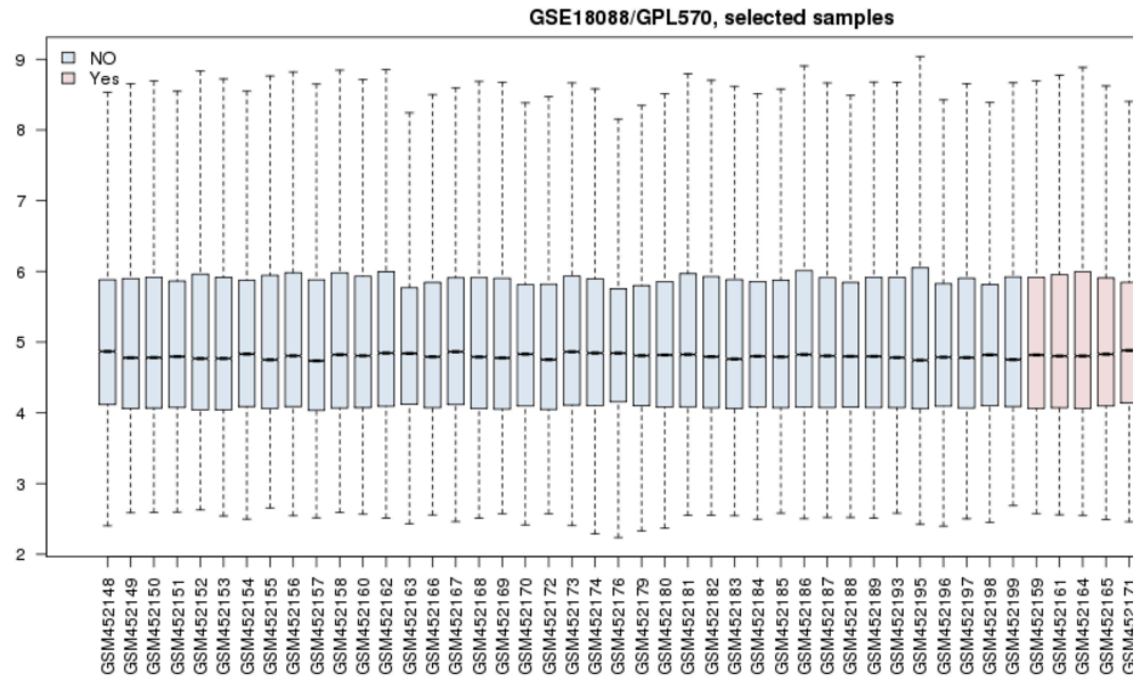
| ID | adj.P.Val | P.Value | t | B | logFC | Gene.symbol | Gene.title |
|---|---|---|---|---|---|---|---|
| 1557706_at | 0.153 | 0.00000289 | 5.23 | 3.49218 | 0.18 | ZHX2 | zinc fingers and homeob... |
| 218714_at | 0.153 | 0.00000559 | 5.04 | 2.99582 | 0.341 | PRR14 | proline rich 14 |
| 202400_s_at | 0.206 | 0.00002048 | 4.67 | 2.01298 | 0.197 | SRF | serum response factor |
| 231917_at | 0.206 | 0.00002244 | -4.65 | 1.94369 | -0.47 | GFM2 | G elongation factor mitoc... |
| 237582_at | 0.206 | 0.00002244 | -4.65 | 1.9434 | -0.139 | | |
| 1556090_at | 0.206 | 0.00002928 | 4.57 | 1.74157 | 0.206 | | |
| 213597_s_at | 0.206 | 0.00002993 | -4.56 | 1.7247 | -0.206 | CTDSPL | CTD small phosphatase ... |
| 224992_s_at | 0.206 | 0.0000301 | 4.56 | 1.72055 | 0.377 | CMIP | c-Maf inducing protein |
| 37226_at | 0.206 | 0.00003399 | -4.53 | 1.62825 | -0.212 | BNIP1 | BCL2 interacting protein 1 |
| 223354_x_at | 0.269 | 0.00004928 | -4.42 | 1.34562 | -0.423 | MFF | mitochondrial fission factor |
| 202675_at | 0.292 | 0.00006107 | -4.35 | 1.18246 | -0.494 | SDHB | succinate dehydrogenas... |
| 236404_at | 0.292 | 0.00006412 | 4.34 | 1.14534 | 0.2 | | |
| 213076_at | 0.301 | 0.0000716 | 4.3 | 1.06139 | 0.249 | ITPKC | inositol-trisphosphate 3-k... |
| 231918_s_at | 0.331 | 0.00008481 | -4.25 | 0.93241 | -0.563 | GFM2 | G elongation factor mitoc... |
| 236227_at | 0.35 | 0.00010092 | -4.2 | 0.79991 | -0.32 | TMEM161B | transmembrane protein ... |
| 35776_at | 0.35 | 0.00011038 | 4.17 | 0.73167 | 0.239 | ITSN1 | intersectin 1 |
| 203297_s_at | 0.35 | 0.00011293 | 4.17 | 0.71425 | 0.286 | JARID2 | jumonji and AT-rich inter... |
| 226087_at | 0.35 | 0.00011527 | -4.16 | 0.6986 | -0.416 | LZIC | leucine zipper and CTNN... |
| 200978_at | 0.354 | 0.00012303 | -4.14 | 0.64898 | -0.442 | MDH1 | malate dehydrogenase 1 |
| 217932_at | 0.358 | 0.00014766 | -4.09 | 0.50993 | -0.422 | MRPS7 | mitochondrial ribosomal ... |

# GEO2R – Box Plot Distributions

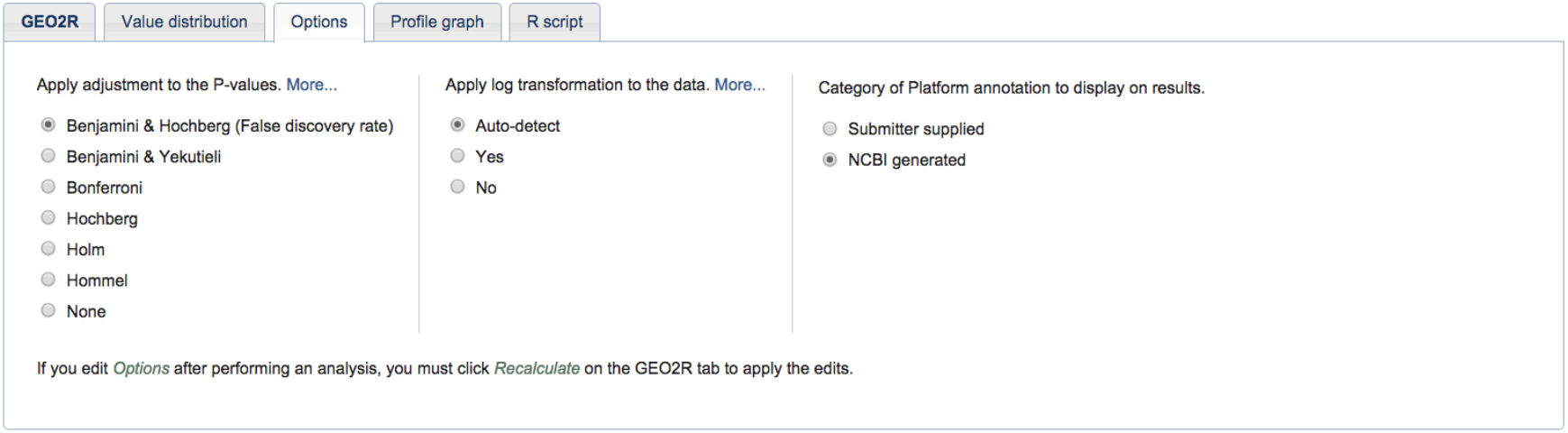
GEO2R | Value distribution | Options | Profile graph | R script

Calculate the distribution of value data for the Samples you have selected. Distributions may be viewed graphically as a box plot or exported as a number summary table. The plot is useful for determining if value data are median-centered across Samples, and thus suitable for cross-comparison. More...

View  Export



# GEO2R – Options for Analysis



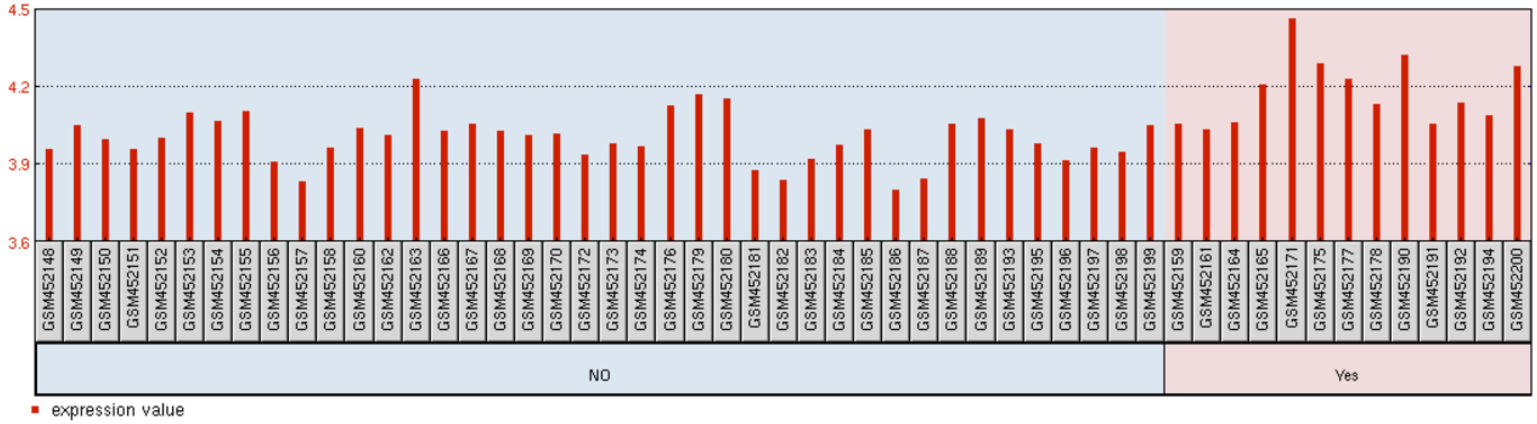
# GEO2R – Profile Graph

GEO2R | Value distribution | Options | Profile graph | R script

Enter ID: 1557706_at Set View data for GPL570

Sample values

GSE18088 / GPL570 / 1557706_at




This tab allows you to view a specific gene expression profile graph by entering the corresponding identifier from the ID column of the Platform record. This feature does not perform any calculations; it merely displays the expression values of the gene across Samples. Sample groups may or may not be defined for this feature to work.

# GEO2R – R Script

GEO2R | Value distribution | Options | Profile graph | R script

```
# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.36.0, limma 3.26.8
# R scripts generated  Tue Oct 18 11:47:26 EDT 2016

################################################################
#   Differential expression analysis with limma
library(Biobase)
library(GEOquery)
library(limma)

# load series and platform data from GEO

gset <- getGEO("GSE18088", GSEMatrix =TRUE, AnnotGPL=TRUE)
if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group names for all samples
gsms <- "0000000000010100110000010001011000000000011101000001"
sml <- c()
for (i in 1:nchar(gsms)) { sml[i] <- substr(gsms,i,i) }

# log2 transform
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
          (qx[6]-qx[1] > 50 && qx[2] > 0) ||
          (qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)
if (LogC) { ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex) }

# set up the data and proceed with analysis
sml <- paste("G", sml, sep="")    # set group names
fl <- as.factor(sml)
gset$description <- fl
design <- model.matrix(~ description + 0, gset)
colnames(design) <- levels(fl)
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(G1-G0, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)

tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title"))
write.table(tT, file=stdout(), row.names=F, sep="\t")
```

# Reuse of public genome-wide gene expression data

*Nature Reviews Genetics* | AOP, published online 27 December 2012; doi:10.1038/nrg3394

REVIEWS

## Reuse of public genome-wide gene expression data

*Johan Rung and Alvis Brazma*

Abstract | Our understanding of gene expression has changed dramatically over the past decade, largely catalysed by technological developments. High-throughput experiments — microarrays and next-generation sequencing — have generated large amounts of genome-wide gene expression data that are collected in public archives. Added-value databases process, analyse and annotate these data further to make them accessible to every biologist. In this Review, we discuss the utility of the gene expression data that are in the public domain and how researchers are making use of these data. Reuse of public data can be very powerful, but there are many obstacles in data preparation and analysis and in the interpretation of the results. We will discuss these challenges and provide recommendations that we believe can improve the utility of such data.

# Major Microarray Data Repositories

| Database | Description | URL | Refs |
|---|---|---|---|
| *Public repositories* | | | |
| ArrayExpress (from EBI) | Any functional genomic data | http://www.ebi.ac.uk/arrayexpress | 8 |
| Gene Expression Omnibus (GEO; from NCBI) | Any functional genomic data | http://www.ncbi.nlm.nih.gov/geo | 9 |
| DDBJ Omics Archive | Any functional genomic data | http://trace.ddbj.nig.ac.jp/dor | 10 |
| Stanford Microarray Database | Any functional genomic data | http://smd.stanford.edu | 104 |

*(Rung & Brazma, Nat Rev Genetics 2013)*

# Disease Specific Microarray Data Databases

| *Topical databases* | | | |
|---|---|---|---|
| Oncomine | Cancer | http://www.oncomine.org | 34 |
| Pancreatic Expression DB | Pancreatic expression | http://www.pancreasexpression.org | 31 |
| ParkDB | Parkinson's disease | http://www2.cancer.ucl.ac.uk/Parkinson_Db2 | 32 |
| ProfileChaser | Expression similarity | http://profilechaser.stanford.edu | 26 |
| PlexDB | Plants | http://www.plexdb.org | 37 |
| GXD | Mice | http://www.informatics.jax.org/expression.shtml | 41 |
| TFGD | Tomatoes | http://ted.bti.cornell.edu | 38 |
| miRGator | microRNA | http://mirgator.kobic.re.kr | 28 |
| COXPRESdb | Multi-species comparisons | http://coxpresdb.jp | 25 |
| OryzaExpress | Rice; co-expression | http://bioinf.mind.meiji.ac.jp/OryzaExpress | 21 |
| GDP | Glaucoma | http://glaucomadb.jax.org/glaucoma | 33 |
| aGEM | Anatomical | http://agem.cnb.csic.es | 44 |
| Atted-II | Plants; co-expression | http://atted.jp | 22 |
| ArraySearch | *Arabidopsis thaliana* | http://arraysearch.org | 24 |
| GUDMAP | Genitourinary system | http://www.gudmap.org | 36 |
| EMAGE | Mouse *in situ* expression | http://www.emouseatlas.org/emage | 42 |
| 4DXpress | Multi-species anatomical | http://4dx.embl.de/4DXpress | 43 |
| GCOD | Cancer | http://compbio.dfci.harvard.edu/tgi/cgi-bin/tucan/tucan.pl | 35 |

*(Rung & Brazma, Nat Rev Genetics 2013)*

# Other Microarray Data Repositories

| *Added-value databases* | | | |
|---|---|---|---|
| Gene Expression Atlas | Gene expression in different cell types, organism parts, developmental stages, disease states, sample treatments and other biological or experimental conditions | http://www.ebi.ac.uk/gxa | 16 |
| GeneChaser | Differential expression | http://genechaser.stanford.edu | 17 |
| BioGPS | Tissue expression | http://biogps.org | 40 |
| Genevestigator | Commercial; wide range of data and analysis types | https://www.genevestigator.com/gv | 105 |
| Gene Expression Barcode | Tissue expression | http://barcode.luhs.org | 18 |
| Nextbio | Commercial; wide range of data and analysis types | http://www.nextbio.com | |

| *Integrative databases* | | | |
|---|---|---|---|
| Wormbase | *Caenorhabditis elegans* — genes, genomes, phenotypes, genetic variation, proteins, antibodies and developmental stages | http://www.wormbase.org | 49 |
| IntOGen | Cancer — gene expression, copy number alteration and mutations | http://www.intogen.org | 45 |
| canSAR | Cancer — gene expression, proteins, structures, interactions and compounds | http://cansar.icr.ac.uk | 47 |
| CMAP | Drug response, gene expression and diseases | http://www.broadinstitute.org/cmap | 46 |
| Cistrome | Gene expression regulation by DNA-binding proteins | http://cistrome.org | 27 |

*(Rung & Brazma, Nat Rev Genetics 2013)*

# Things to Consider when Reusing of Raw Data from Public Databases

1. *Quality control.*
   - Public archives store data as they have been received from the submitter.
   - Include only arrays that pass quality-control criteria in further analysis.
   - Be aware that some studies, often from the same laboratory, contain identical raw data files, such as when a set of control samples has been used independently in two different studies.
2. *Revise annotation.*
   - Annotation of public data can be incomplete, not-up-to-date or conflicting (different terms to annotate samples and experimental factors).
3. *Array selection.*
   - Experiments that reuse data need experimental design (like new experiment).
   - Include only arrays in the study that address the intended question.
   - By excluding non-informative arrays, decrease the data heterogeneity and improve the conditions for accurate statistical tests concerning the goal of the study.

*(Rung & Brazma, Nat Rev Genetics 2013)*

# Things to Consider when Reusing of Raw Data from Public Databases

4. *Define and annotate probe sets.*
   - Different platforms, and sometimes even different versions of arrays, may define probe sets or the individual probe sequences differently.
   - Original manufacturer annotation may be outdated.
   - This may have a serious impact on the data analysis
5. *Normalize and analyze across all arrays and experimental conditions as if it were a single data set.*
   - Cross-platform normalization needs special attention and should be dealt with carefully, or the biases introduced may outweigh the benefit of combining many samples.
   - In downstream analysis, adjusting for study effect and other biases may be necessary.

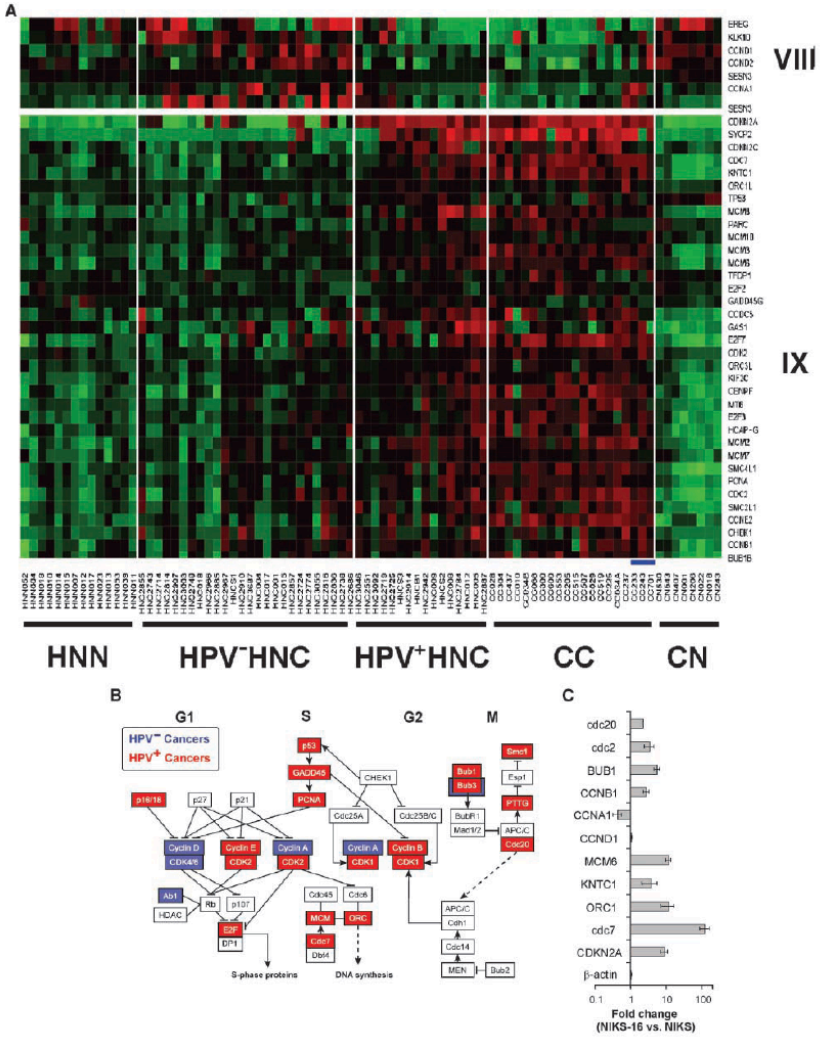
*(Rung & Brazma, Nat Rev Genetics 2013)*

# Presenting the Microarray Results: Gene List vs Heat Map

| Probe set ID* | Gene title | Gene symbol | *t* statistic | Overlaps† |
|---|---|---|---|---|
| 207039_at | *Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)* | *CDKN2A* | 6.73 | T/N, CC/HNC |
| 228286_at | *Hypothetical protein FLJ40869* | *FLJ40869* | 5.45 | CC/HNC |
| 218397_at | *Fanconi anemia, complementation group L* | *FANCL* | 5.63 | CC/HNC |
| 203358_s_at | *Enhancer of zeste homologue 2 (Drosophila)* | *EZH2* | 6.41 | CC/HNC |
| 218783_at | *DKFZP434B168 protein* | *DKFZP434B168* | 6.00 | CC/HNC |
| 206316_s_at | *Kinetochore associated 1* | *KNTC1* | 6.26 | T/N, CC/HNC |
| 201555_at | *MCM3 minichromosome maintenance deficient 3 (S. cerevisiae)* | *MCM3* | 5.88 | T/N, CC/HNC |
| 221677_s_at | *Downstream neighbor of SON* | *DONSON* | 6.08 | T/N, CC/HNC |
| 204510_at | *CDC7 cell division cycle 7 (S. cerevisiae)* | *CDC7* | 6.42 | T/N, CC/HNC |
| 227255_at | *Casein kinase* | *LOC149420* | 5.59 | CC/HNC |
| 222201_s_at | *CASP8 associated protein 2* | *CASP8AP2* | 5.09 | T/N, CC/HNC |
| 224428_s_at | *Cell division cycle associated 7* | *CDCA7* | 4.36 | CC/HNC |
| 219306_at | *Kinesin-like 7* | *KNSL7* | 5.45 | CC/HNC |
| 212621_at | *KIAA0286 protein* | *KIAA0286* | 4.60 | T/N |
| 229551_x_at | *Zinc finger protein 367* | *ZNF367* | 6.29 | T/N, CC/HNC |
| 222848_at | *Leucine zipper protein FKSG14* | *FKSG14* | 4.37 | T/N, CC/HNC |
| 228401_at | — | — | 4.49 | T/N, CC/HNC |
| 225655_at | *Ubiquitin-like, containing PHD and RING finger domains, 1* | *UHRF1* | 4.69 | T/N, CC/HNC |
| 227350_at | *Helicase, lymphoid-specific* | *HELLS* | 5.13 | T/N, CC/HNC |
| 228033_at | *E2F transcription factor 7* | *E2F7* | 4.36 | T/N, CC/HNC |
| 218585_s_at | *RA-regulated nuclear matrix-associated protein* | *RAMP* | 4.99 | T/N, CC/HNC |
| 209172_s_at | *Centromere protein F, 350/400ka (mitosin)* | *CENPF* | 4.51 | T/N, CC/HNC |
| 226456_at | *Hypothetical protein MGC24665* | *MGC24665* | 6.23 | T/N |
| 202589_at | *Thymidylate synthetase* | *TYMS* | 5.51 | T/N |
| 239680_at | — | — | 5.19 | CC/HNC |
| 236513_at | — | — | 4.85 | CC/HNC |
| 224320_s_at | *MCM8 minichromosome maintenance deficient 8* | *MCM8* | 5.73 | T/N |
| 202532_s_at | *Dihydrofolate reductase* | *DHFR* | 5.24 | None |
| 210371_s_at | *Retinoblastoma binding protein 4* | *RBBP4* | 4.73 | T/N, CC/HNC |
| 201970_s_at | *Nuclear autoantigenic sperm protein (histone-binding)* | *NASP* | 6.42 | T/N, CC/HNC |
| 223542_at | *Ankyrin repeat domain 32* | *ANKRD32* | 4.40 | T/N, CC/HNC |
| 209337_at | *PC4 and SFRS1 interacting protein 1* | *PSIP1* | 6.01 | CC/HNC |
| 205961_s_at | *PC4 and SFRS1 interacting protein 1* | *PSIP1* | 5.59 | CC/HNC |
| 206542_s_at | *SWI/SNF related, matrix associated, actin-dep chromatin regulator* | *SMARCA2* | 4.88 | None |
| 242471_at | — | — | 4.97 | None |
| 229442_at | *Hypothetical protein MGC33382* | *MGC33382* | 4.45 | T/N, CC/HNC |
| 203482_at | *Chromosome 10 open reading frame 6* | *C10orf6* | 6.24 | CC/HNC |
| 201448_at | *TIA1 cytotoxic granule-associated RNA binding protein* | *TIA1* | 5.60 | None |
| 221264_s_at | *TAR DNA binding protein* | *TARDBP* | 5.57 | None |
| 214093_s_at | *Far upstream element (FUSE) binding protein 1* | *FUBP1* | 4.78 | None |
| 209285_s_at | *Retinoblastoma-associated protein 140* | *RAP140* | 5.56 | None |
| 230120_s_at | *Plasminogen-like* | *PLGL* | 5.39 | None |
| 217122_s_at | *Solute carrier family 35, member E2* | *SLC35E2* | 7.47 | None |
| 228466_at | *Clone IMAGE:111714 mRNA sequence* | — | 5.59 | None |
| 212179_at | *Chromosome 6 open reading frame 111* | *C6orf111* | 5.31 | None |
| 235919_at | — | — | 5.10 | None |
| 215731_s_at | *M-phase phosphoprotein 9* | *MPHOSPH9* | 4.64 | None |
| 229886_at | *FLJ32363 protein* | *FLJ32363* | 5.87 | None |
| 228174_at | — | — | 6.44 | None |
| 212774_at | *Zinc finger protein 238* | *ZNF238* | 4.65 | None |
| 226478_at | *Transmembrane 7 superfamily member 3* | *TM7SF3* | 4.64 | None |
| 42361_g_at | *Chromosome 6 open reading frame 18* | *C6orf18* | 5.76 | CC/HNC |
| 202726_at | *Ligase I, DNA ATP-dependent* | *LIG1* | 6.26 | None |
| 231931_at | *PR domain containing 15* | *PRDM15* | 7.15 | CC/HNC |
| 230777_s_at | *PR domain containing 15* | *PRDM15* | 6.54 | CC/HNC |
| 229468_at | *Cyclin-dependent kinase 3* | *CDK3* | 5.45 | None |
| 230653_at | — | — | 5.15 | None |

(Continued on the following page)

vs



# Heat map

- Heat maps represent ***two-dimensional*** tables of numbers as ***shades of colors***.
- This is a ***popular plotting technique*** in biology, used to ***depict gene expression*** and other ***multivariate data***.
- The ***dense*** and ***intuitive display*** makes heat maps well-suited for presentation of ***high-throughput data***.
- Hundreds of rows and columns can be displayed on a screen.
- Heat maps rely fundamentally on ***color encoding*** and on ***meaningful reordering*** of the ***rows and columns***.
- When either of these components is compromised, the utility of the visualization suffers.

*(Gehlenborg & Wong, Nat Methods 2012)*

# Examples of Heat maps

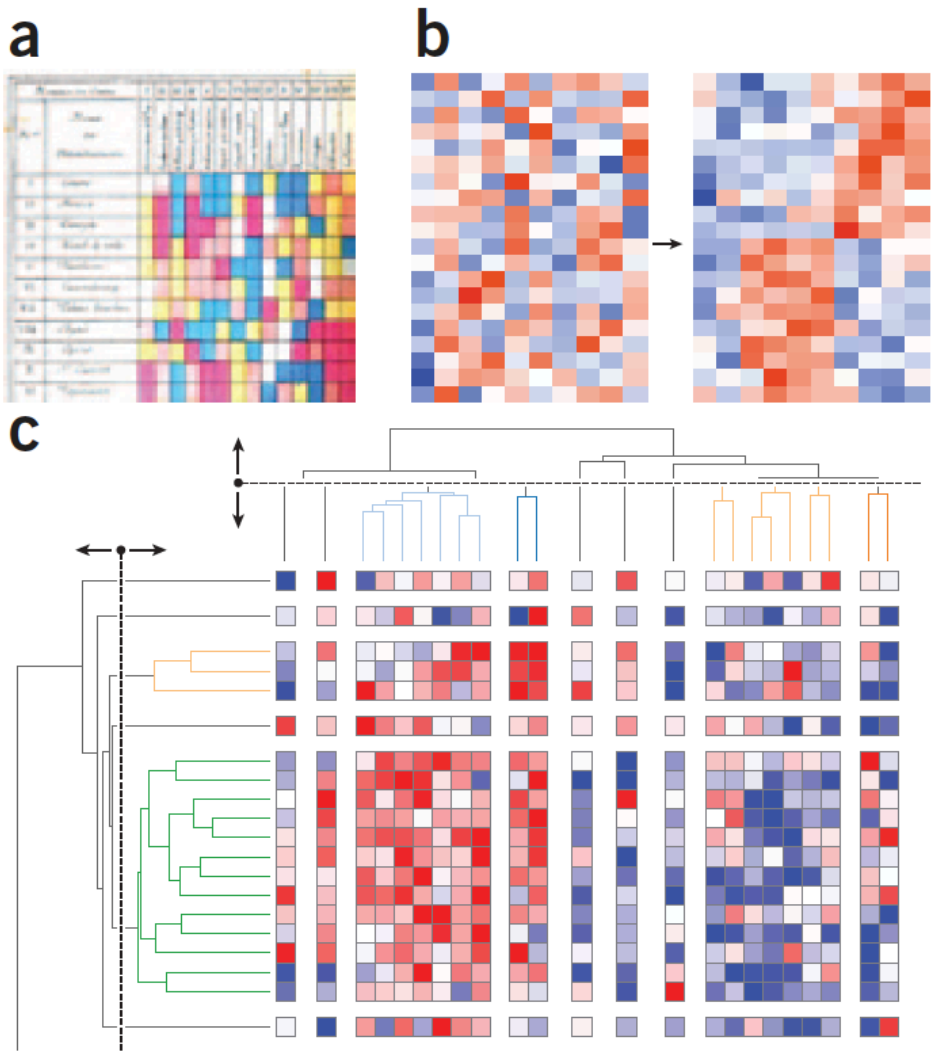



**Figure 1** | Heat maps. (**a**) An example of a colored table from ref. 1. (**b**) Clustering brings like next to like items to reveal patterns in the data. (**c**) Adding gaps according to the hierarchical cluster tree helps emphasize relationships in the matrix.

*(Gehlenborg & Wong, Nat Methods 2012)*

# matrix2png - Heat map Generation Tool

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 19 no. 2 2003*
*Pages 295–296*

## *Matrix2png: a utility for visualizing matrix data*

*Paul Pavlidis[1,*] and William Stafford Noble[2]*

*[1]Columbia Genome Center, Columbia University, New York, USA and [2]Department of Genome Sciences, University of Washington, Seattle WA, USA*

revised on August 27, 2002; accepted on September 2, 2002

http://www.chibi.ubc.ca/matrix2png/

**ABSTRACT**
We describe a simple software tool, 'matrix2png', for creating color images of matrix data. Originally designed with the display of microarray data sets in mind, it is a general tool that can be used to make simple visualizations of matrices for use in figures, web pages, slide presentations and the like. It can also be used to generate images 'on the fly' in web applications. Both continuous-valued and discrete-valued (categorical) data sets can be displayed. Many options are available to the user, including the colors used, the display of row and column labels, and scale bars. In this note we describe some of matrix2png's features and describe some places it has been useful in the authors' work.
**Availability:** A simple web interface is available, and Unix binaries are available from http://microarray.cpmc.columbia.edu/matrix2png. Source code is available on request.
**Contact:** pp175@columbia.edu

Figure 1A

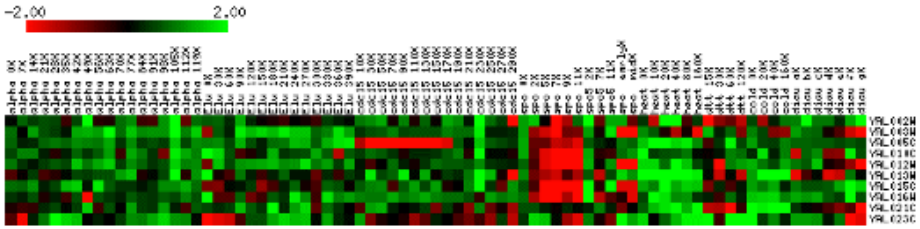
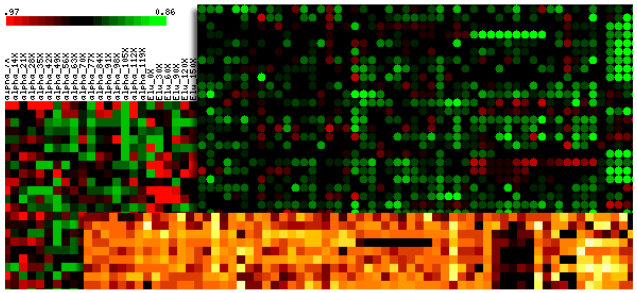



Figure 1B

# matrix2png

## Matrix2png

http://www.chibi.ubc.ca/matrix2png/




Matrix2png is a simple but powerful program for making visualizations of microarray data and many other data types. It generates PNG formatted images from text files of data. It is fast, easy to use, and reasonably flexible. It can be used to generate publication-quality images, or to act as a image generator for web applications. Our group has found it useful for imaging all kinds of matrix-based data, not just microarray data.

### Jump to the web interface.

The current version of matrix2png is 1.2.1 (February 2011). See the version history here

### Features

- PNG output only.
- Simple UNIX command line interface. Has been tested under Linux and Solaris.
- Uses a simple text format to import data.
- Set range, contrast, outlier trimming.
- Generates scale bars, row labels, and column labels.
- Use rectangles or ellipses.
- Convenient color selection from a preset palette as well as popular color maps.
- Can handle missing values and data in the form of discrete (categorical) values

### Web interface to matrix2png

Follow this link to use a simple web interface to matrix2png with your own data files.

### Download

Please visit the download page for source code.

### Documentation

**NOTE!** Matrix2png generates PNG (portable network graphics), not gifs or jpegs. The PNG format is supported by the major web browsers as well as image processing software such as Adobe Photoshop, Macromedia Fireworks, etc. Read about the PNG format here. For many users, the web interface to matrix2png will suffice. If you want to install and use matrix2png on your system, see this page.

Gallery of examples.

Detailed documentation.

### How to cite matrix2png

If you use images created with matrix2png for publication or presentation, please cite:

**Pavlidis, P. and Noble W.S. (2003) Matrix2png: A Utility for Visualizing Matrix Data. Bioinformatics 19: 295-296 (abstract).**

**Readers of the Bioinformatics application note:** Here is the color version of the figure from the paper (pdf format).

# Examples of matrix2png

http://www.chibi.ubc.ca/matrix2png/

The first examples use this data file, which is part of the "Eisen" data set from Stanford.

## Example 1

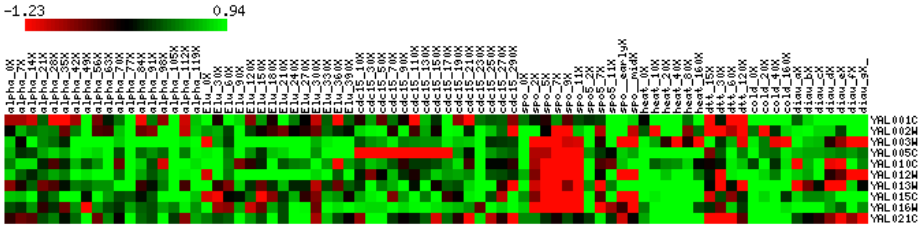
```
matrix2png -data testdata.rdb -size 8:8 -map 1  -range -2:2 -numr 10 >! example1.png
```

About the options: -size determines the size of each feature, 8x8 pixels in this case. -map 1 selects the color scheme from a presets. -range -2:2 specifies that the ends of the color range correspond to values of -2 and 2. This means that values less than -2 and greater than 2 will be 'clipped' and displayed as the 'mincolor' and 'maxcolor' respectively. -numr 10 determines that only 10 rows of data will be shown.

## Example 2

Use the -r, -c, and -s options to get row labels, column labels, and a scalebar. The -trim 5 limits the color range to the middle 90% of the data values - the highest and lowest 5% are 'clipped'.

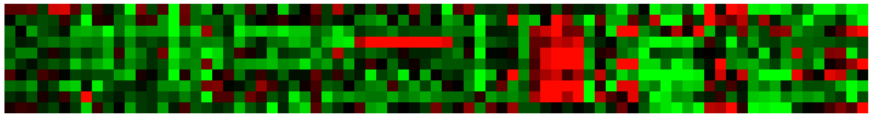
```
matrix2png -data testdata.rdb -size 8:8 -r -c -s -map 3  -trim 5 -numr 10 >! example2.png
```




## Example 3

This is example 2 without trimming.

```
matrix2png -data testdata.rdb -size 8:8 -map 3 -range -2:2 -numr 10 >! example3.png
```



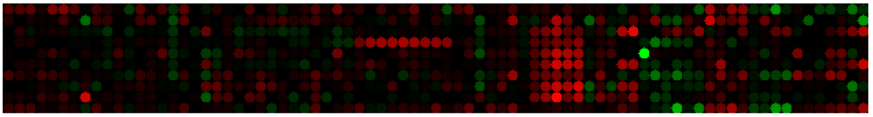
# Examples of matrix2png

## Example 4

You can (partly) select the colors you want to use instead of a map, using -mincolor and -maxcolor. The use of -b makes the color range go through black in the middle. Use -e to get ellipses. The low contrast in this image comes from the failure to use -range or -trim or -con.

```
matrix2png -data testdata.rdb -size 8:8 -mincolor red -maxcolor green -e -b -bkgcolor black >! example4.png
```

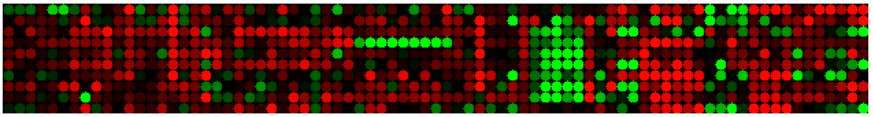
Note that similar results can be obtained with -map 3 instead of specifying -mincolor, -maxcolor, and -b.



## Example 5

The -bkgcolor option sets the background color. This is preset map number 4.

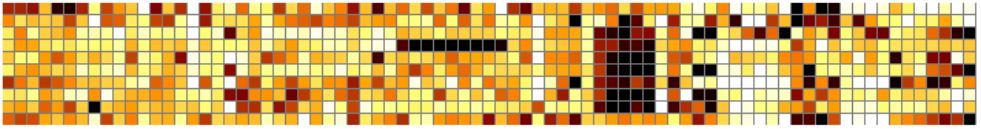
```
matrix2png -data testdata.rdb -size 8:8 -map 4 -e -range -2:2 -bkgcolor black -numr 10 >! example5.png
```



## Example 6

Add dividers between each block with -d.

```
matrix2png -data testdata.rdb -size 8:8 -map 1 -d -range -2:2 -numr 10 >! example6.png
```



## Example 7

This one is of data that is categorical, not continuous, from a data file representing a gene structure prediction compared to a reference structure. It uses a map file, NM_000041.map.

```
matrix2png -size 1:16 -s -r -dmap NM_000041.map -data NM_000041.ds.mtx >! NM_000041.10.png
```


http://www.chibi.ubc.ca/matrix2png/

matrix2png home

# Examples of Heat maps

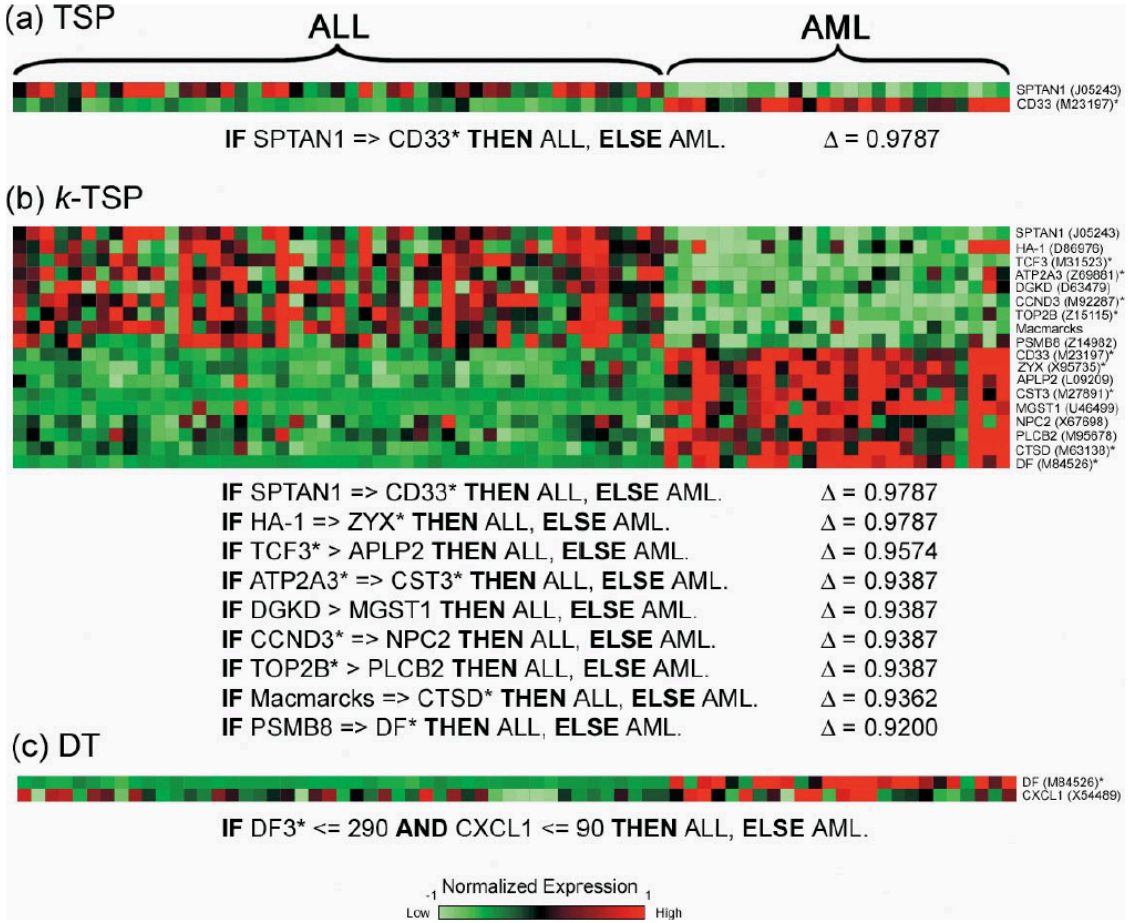
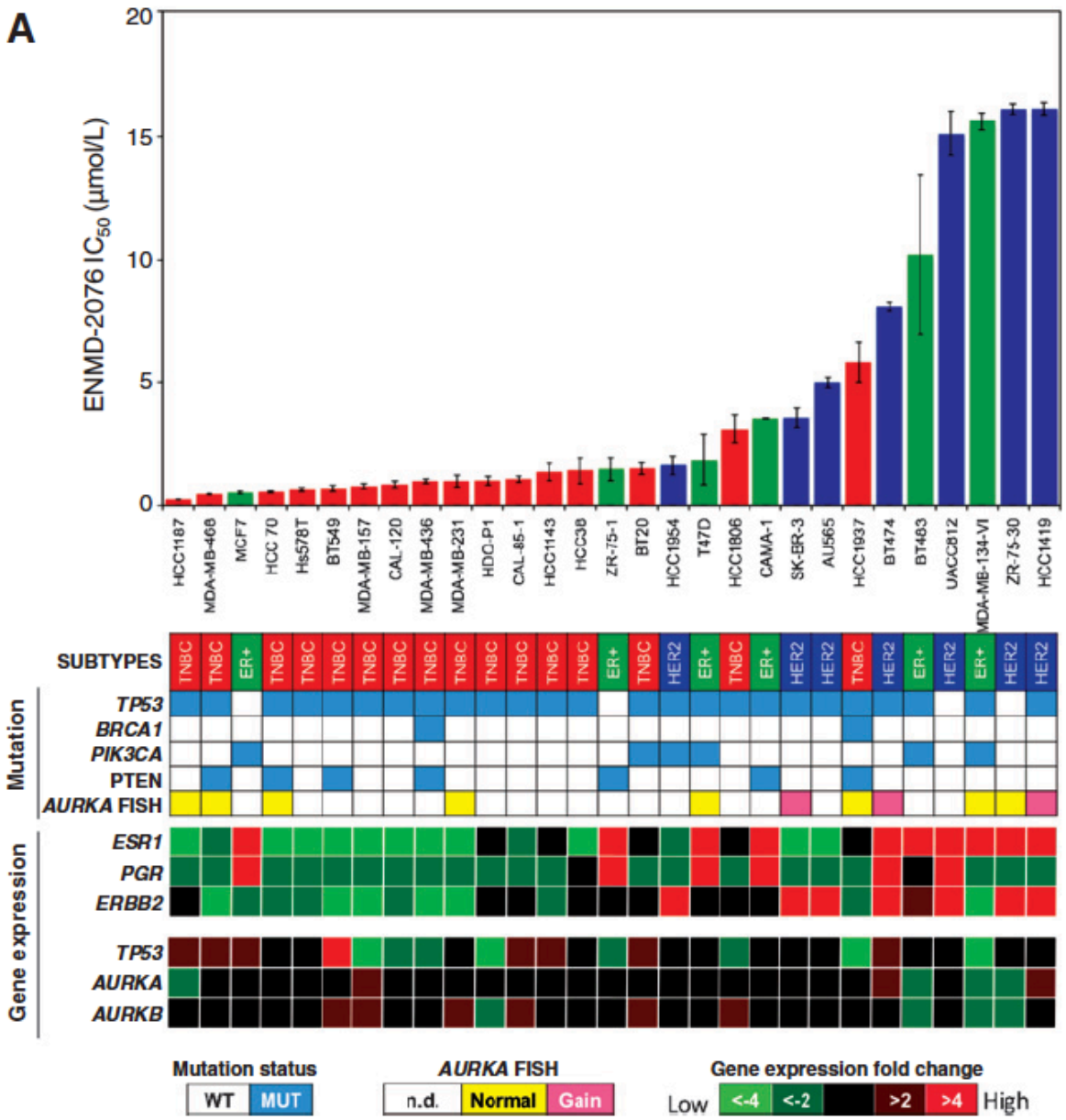



**Fig. 2.** Genes that distinguish ALL from AML. Each row corresponds to a gene and each column corresponds to a sample array. Genes labeled with an asterisk (*) were identified in Golub *et al.* (1999). This heat map is generated by using the matrix2png software (Pavlidis and Noble, 2003). The expression level for each gene is normalized across the samples such that the mean is 0 and the standard deviation (SD) is 1. Genes with expression levels greater than the mean are colored in red and those below the mean are colored in green. The scale indicates the number of SDs above or below the mean. In (a–c), the discriminative genes and decision rules in three cases are shown: (**a**) TSP Classifier, (**b**) *k*-TSP Classifier and (**c**) Decision tree (DT) classifier.

*(Tan et al, Bioinformatics 2005)*

# Examples of Heat maps




*(Diamond et al, Clinical Cancer Research 2013)*

# Take Home Message

*Know your data*

- Know how to extract and normalize your microarray gene expression data.
- If data set was downloaded from public databases, know how the data were processed. Best to download raw data and do your own normalization.

*Know where to find data*

- Know how to locate public data from published papers.
- Know how to download data sets for correlative analysis / meta-analysis.
- Know how to integrate data for hypothesis generation / discovery.

*Know how to present your data analysis*

- Heat map is one of the methods to visualize gene expression and other high-throughput data.
- Need to know how to add value to heat map.