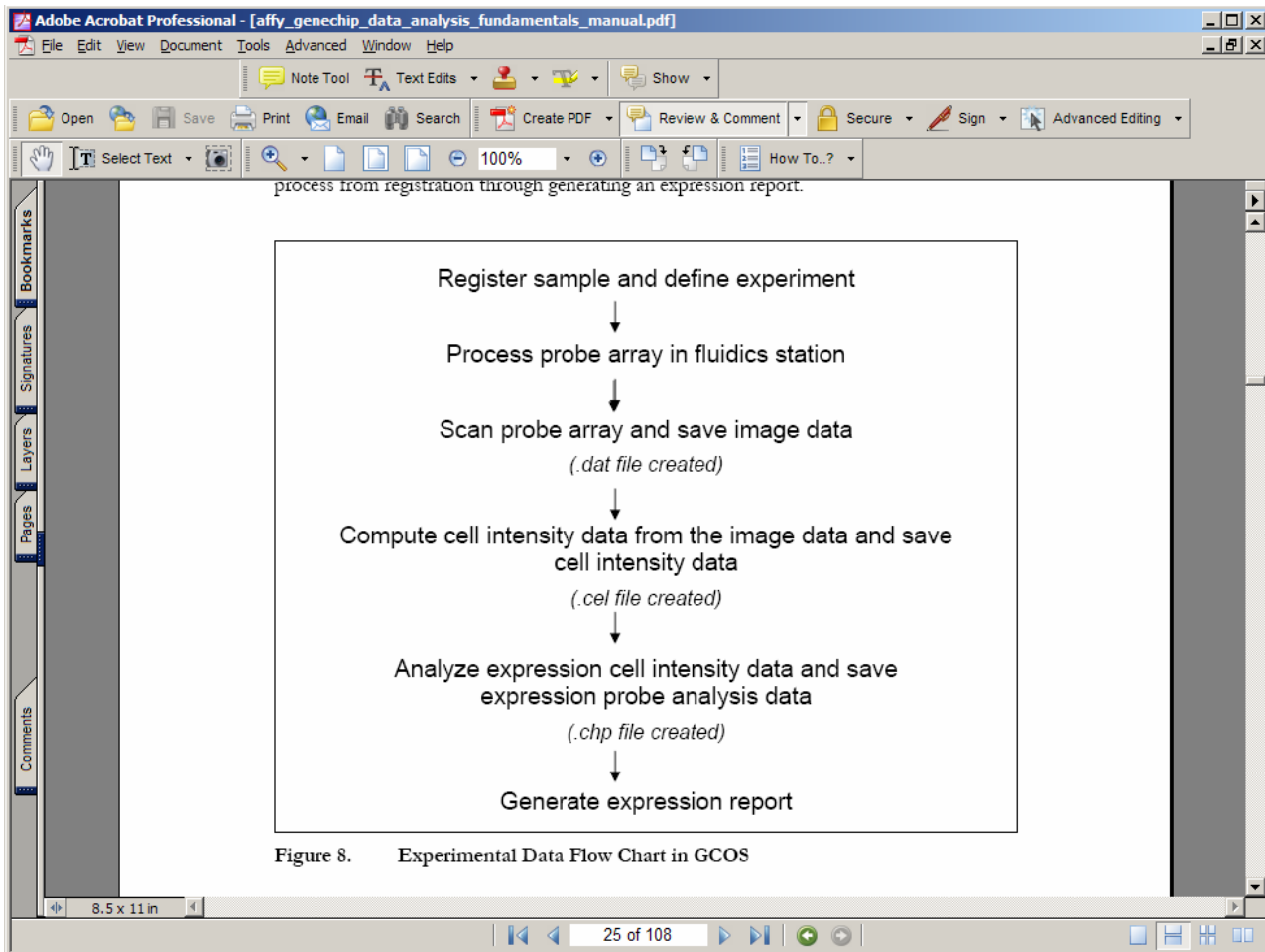


# Notes written during work on my Math 501 project and paper on cancer accuracy using PCA

By Nasser Abbasi

This is my scratch notes files where I kept notes during work on the math 501 project (Spring 2007, CSUF)



Geneship, it looks like I need a .chp file

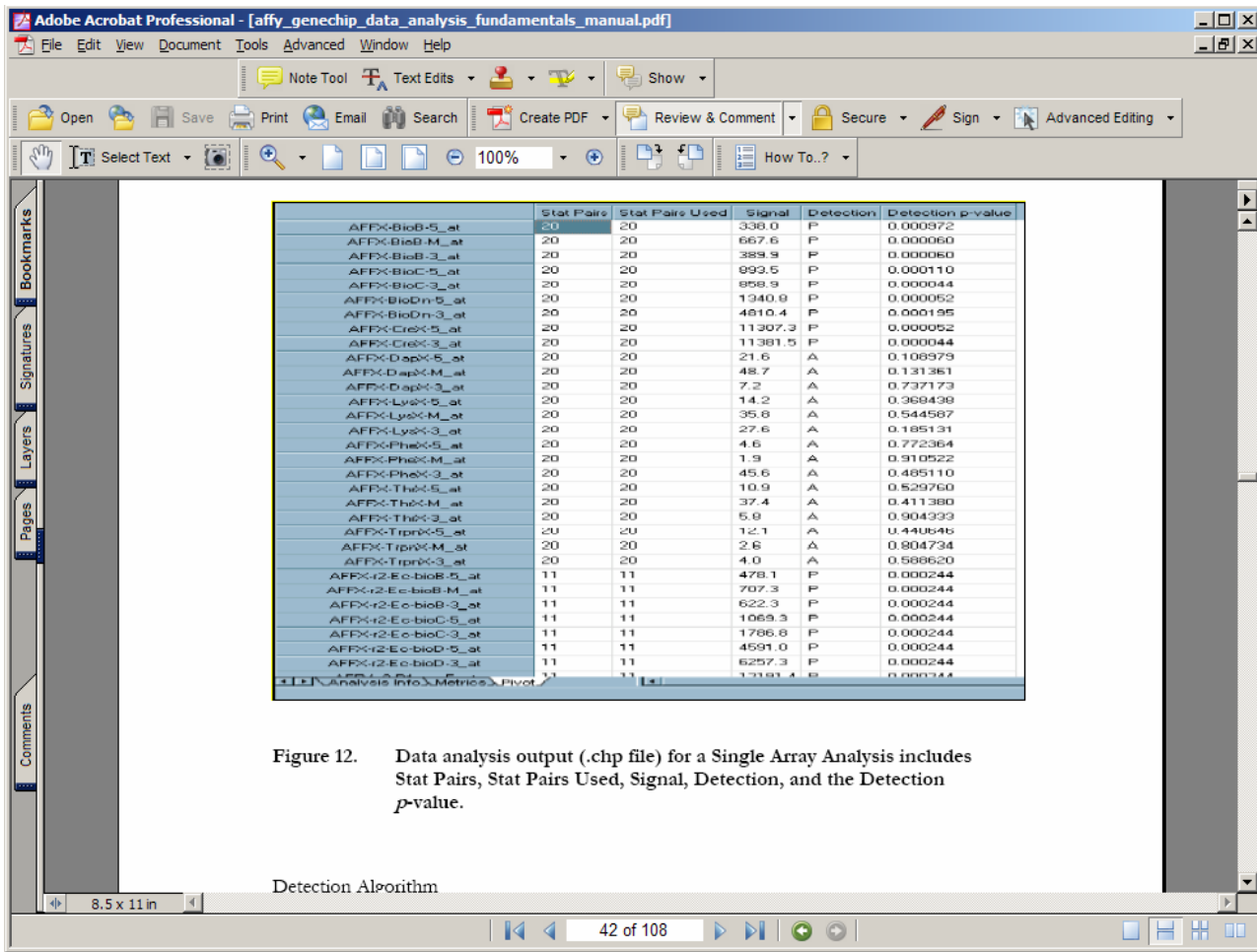


Figure 12. Data analysis output (.chp file) for a Single Array Analysis includes Stat Pairs, Stat Pairs Used, Signal, Detection, and the Detection p-value.

Detection Algorithm

<http://www.affymetrix.com/products/fos/cancer.affx> see this for links to paper dealing with using affy geneships for cancer studies

idea: use ICA for : "identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction)."

-----  
Check this sometimes:

PCA disjoint models for multiclass cancer analysis using gene expression data

[Bicciato S,](#)

[Luchini A,](#)

[Di Bello C.](#)

Department of Chemical Process Engineering, University of Padova, via Marzolo, 9, 35131, Padova, Italy.  
silvio.bicciato@unipd.it

-----  
[http://www.affymetrix.com/support/technical/manual/expression\\_manual.affx](http://www.affymetrix.com/support/technical/manual/expression_manual.affx)

affy technical manuals.

-----  
**Standard Data Formats**

Axon Instruments' GAL and GPR file formats are standards in the microarray industry.

## GenePix Array List (GAL) Files

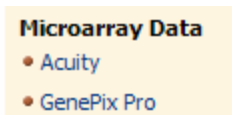
Easily construct lists describing the position and content of each spot on an array from plain text files. Substance name and ID lists can be pasted directly into the array settings file to create a GenePix Array List (GAL) file. GAL files can also be created from a collection of microwell plate files using the GenePix Array List Generator. Add as many columns as you like to the GAL file to track any sample information along with GenePix results. See the [GenePix File Formats](#) page for instructions.

## File Formats

- Full support for 16-bit grayscale TIFF images.
- Results (GPR) file containing substance names, feature locations and all extracted data saved in ASCII file format for easy import into advanced analysis packages.
- Export Results in MAGE-ML format.
- Array List (GAL) files allow user-defined columns.

[http://www.moleculardevices.com/pages/software/gn\\_genepix\\_pro.html](http://www.moleculardevices.com/pages/software/gn_genepix_pro.html)

contains microarray data



[http://www.moleculardevices.com/pages/software/gn\\_acuity.html](http://www.moleculardevices.com/pages/software/gn_acuity.html)

to download microarray data files

[http://cebs.niehs.nih.gov/cebs-browser/help/microarray/help-datafile\\_download.html](http://cebs.niehs.nih.gov/cebs-browser/help/microarray/help-datafile_download.html)

nih microarray

<http://dir.niehs.nih.gov/microarray/>

it looks like data is agilent data or affymetrix data.

This site talks about microarray data download

<http://www.arabidopsis.org/help/tutorials/micro7.jsp>

TAIR includes data using both cDNA arrays and Affymetrix GeneChips technology.

How to download and view MA data

<http://www.arabidopsis.org/help/tutorials/micro7.jsp>

DATA

<ftp://ftp.arabidopsis.org/home/tair/Microarrays/>

microarray analysis

<http://www.statsci.org/micrarra/>

databases for microarray


- [Y. F. Leung's comparison of database systems](#)
- [BASE](#). BioArray Software Environment.
- [BASE Plug-in for limma package](#) created by Oja Spjuth, Uppsala University.
- [GeneX](#). Open source database repository of gene expression data.
- [Iobion Informatics](#). Produce GeneTraffic server based system.

<http://chip.dfc.harvard.edu/stats/data.php#download> describes the files used by genechip afymax

## Chip File (.CHP)

The .CHP file contains Signal values and Presence Calls for each probe set on the microarray and can be viewed using MAS 5.0. The Statistical Algorithm is used to calculate the Signal values and Presence Calls from the probe-level fluorescence intensities contained in the .CEL file.

## Cell Intensity File (.CEL)

The .CEL file contains fluorescence intensities for each probe on the microarray. When the .CEL file is opened in either MAS 5.0 or dChip, these probe-specific intensity values are used to reconstruct the scanned image of the hybridized array. It is recommended that the investigator view the .CEL images for each sample to make sure there are no obvious chip defects. For information on how the values in the .CEL file are calculated from the original scan, see the Affymetrix  MAS 5.0 [Probe-level Analysis](#) section of this website.

The probe-specific intensities in the .CEL file are also used in the Statistical Algorithm to calculate the probe-set-level Signals and Presence Calls recorded in the .CHP file. For information on how the intensity values in the .CEL file are used to calculate the .CHP file information, see the [Presence Calls](#) and [Expression Estimates](#) sections of this website.

<http://www.affymetrix.com/products/software/specific/mas.affx> affymetric website

<http://www.gene-chips.com/GeneChips.html#Datamining> places to find data

<http://nciarray.nci.nih.gov/> center of cancer research

<http://nciarray.nci.nih.gov/cgi-bin/gipo> to get data from above

K Kudoh, M Ramanna, R Ravatn, AG Elkahloun, ML Bittner, PS Meltzer, JM Trent, WS Dalton, KV Chin, Monitoring the expression profiles of doxorubicin-induced and doxorubicin-resistant cancer cells by cDNA microarray, *Cancer Research* 60: 15 (AUG 1 2000) Pages 4161-4166

<http://discover.nci.nih.gov/> cancer and microarray

<http://discover.nci.nih.gov/datasets.jsp> has cancer data?

<http://discover.nci.nih.gov/cellminer/loadDownload.do> contains download of .cel files (cancer data)

<http://www.molbiolcell.org/cgi/content/full/13/6/1929> link to chen paper, where earlier paper by Dr Lee references and used data from.

<http://www.ncbi.nlm.nih.gov/projects/geo/> gene expression database query

<http://smd-www.stanford.edu/> standford microarray DB

<http://genome-www.stanford.edu/hcc/Figures/ArrayInformation.htm> array for chen paper

<http://genome-www.stanford.edu/hcc/> chen site liver cancer gene expression

	No. of Cases	Category	Subcategory
Adenoma	3	Adenoma	Liver
FNH	4	FNH	Liver
HCC	102	Primary tumors	Liver
Non-tumor liver	74	non-tumor tissues	Liver
Liver cancer cell lines	10	Cell-line	Liver

<http://smd.stanford.edu/cgi-bin/data/viewDetails.pl?fullID=10029GENEPIX0> chen liver data here.

<http://www.bio.davidson.edu/projects/GCAT/SMDdirections.html> directions on finding GCAT Data on the Stanford Microarray Database

microarray databases

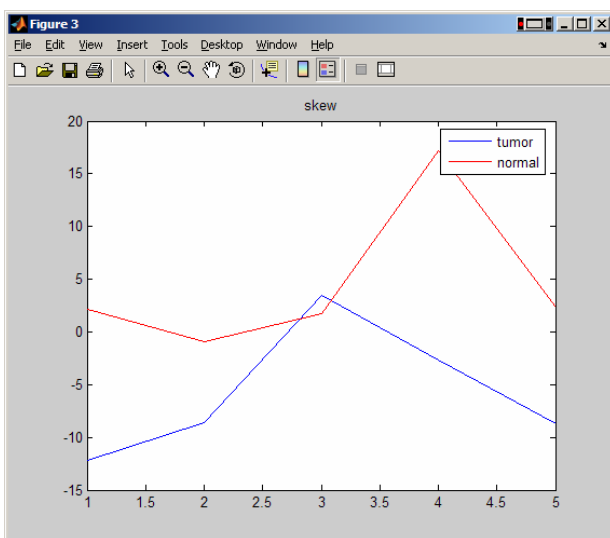
<http://smd-www.stanford.edu/resources/databases.shtml>

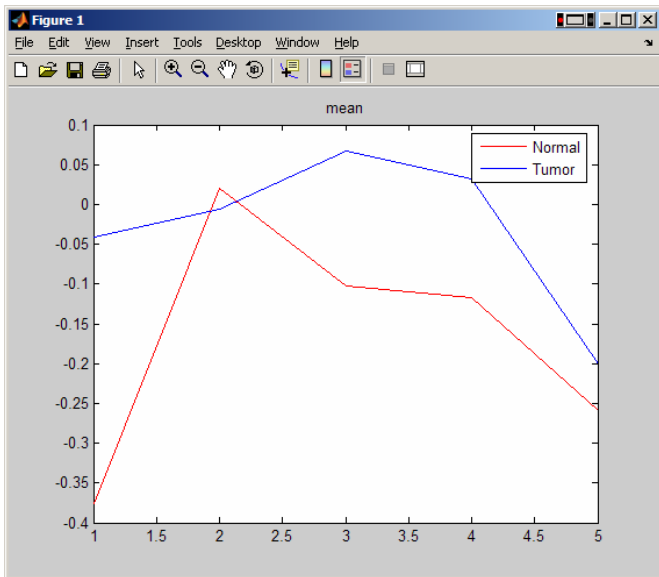
how to download chen data

go to <http://smd.stanford.edu/cgi-bin/search/QuerySetup.pl> and select experimenter as XINCHEN

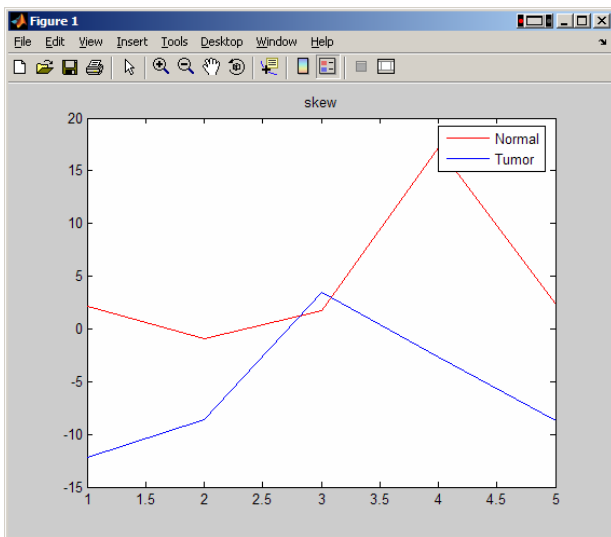
K>> size(BladderData)

6688 126





<http://www.molbiolcell.org/cgi/content/full/14/8/3208> chen second paper



Links from Dr Lee:

<http://jnci.oxfordjournals.org/cgi/content/full/94/17/1320#SEC1>

<http://www.pnas.org/cgi/content/abstract/101/25/9309?view=abstract>

We collected and analyzed 40 published cancer microarray data sets, comprising 38 million gene expression measurements from >3,700 cancer samples.

40 data sets were publicly available and compiled; in total, 37,901,459 gene measurements from 3,762 microarray experiments. Most data sets were of two general formats, either single-channel intensity data, usually corresponding to Affymetrix microarrays, or dual-channel ratio data, usually corresponding to spotted cDNA microarrays, and in the majority of cases, a single composite data file was provided by the study authors and incorporated into our database.

What does this mean?

**Fig. 3.** Meta-signature of undifferentiated cancer. Sixty-nine genes that are overexpressed in undifferentiated cancer relative to well differentiated cancer ( $Q < 0.10$ ) in at least four of seven signatures representing six types of cancer. See [Fig. 2](#) legend for description.

<http://www.pnas.org/content/vol101/issue25/images/large/zpq0240451350003.jpeg>

cancer profiling database:

<http://www.oncomine.org/main/index.jsp> =====> this one

I logged into the above, here is some data I downloaded

### ***Bladder cancer***

**Title:** [Gene expression in the urinary bladder: a common carcinoma in situ gene expressionsignature exists disregarding histopathological classification.](#) **Organization:** Department of Clinical Biochemistry, Aarhus University Hospital, Skejby, AarhusN, Denmark.

**Reference:** Cancer Res 2004/06/02 **Tissue:** Bladder **Array Type:** Affymetrix, Human Genome U133A Array **Study Description:** **Sample Description:** Normal Bladder - Biopsy (9), Normal Bladder Mucosa - Cystectomy (5), [Carcinoma In Situ](#) (4), [NA](#) (2), Superficial Transitional Cell Carcinoma (27), Invasive Transitional Cell Carcinoma (13)

**Data Link:** <http://www.mdl.dk/Files/supplementary%20information%20-%20CIS.pdf>  
[http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds\\_browse.cgi?gds=1479](http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=1479)  
<http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE3167>

Dr Chen Xin send me this:

Hi Nasser:

You should be able to find the arrays at:

[http://smd.stanford.edu/cgi-bin/publication/viewPublication.pl?pub\\_no=107](http://smd.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=107)

Also at GEO:

<http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE3500>

Good luck with your research!

Best,  
Xin

Here it is in GEO also:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds&term=GSE3500\[Accession\]&cmd=search](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds&term=GSE3500[Accession]&cmd=search)  
here it below in GEO (above), I think I only need the liver one

1: GSE3500 record: Gene expression patterns in human liver cancers [Homo sapiens]

Summary: (Submitter supplied) All the array experiments published in "Gene expression patterns in human liver cancers" by Chen X, e Hepatocellular carcinoma (HCC) is a leading cause of death worldwide. Using cDNA microarrays to characterize patterns of gene expression in HCC, we found consistent differences between the expression patterns in HCC compared with those in nontumor liver tissues. The expression patterns in HCC were also readily distinguished from those associated with tumors metastatic to liver. The global gene expression patterns intrinsic to each tumor were sufficiently distinctive that multiple modules from the same patient could usually be recognized and distinguished from all the others in the large sample set on basis of their gene expression patterns alone. [more...](#)

[13 related Platforms](#)

Samples: 207 (listing 18)

<a href="#">GSM79784</a> : Liver (SF8)	<a href="#">GSM79785</a> : HCC (HCV+, SF30)	<a href="#">GSM79786</a> : Met. Colon Cancer (SF37)
<a href="#">GSM79787</a> : Met. Colon Cancer (SF38)	<a href="#">GSM79788</a> : Met. Ovarian Granulosa Cell Tumor..	<a href="#">GSM79789</a> : Met. Granulosa Cell Tumor (SF39, ..
<a href="#">GSM79790</a> : HCC (HBV+, SF18)	<a href="#">GSM79791</a> : Liver (HBV+, HK10, S2)	<a href="#">GSM79792</a> : Liver (HBV+, HK24)
<a href="#">GSM79793</a> : HCC (HBV+, SF1, 1)	<a href="#">GSM79794</a> : Liver (HBV+, SF1)	<a href="#">GSM79795</a> : Met. Colon Cancer (SF5)
<a href="#">GSM79796</a> : HCC (HBV+, HK8)	<a href="#">GSM79797</a> : HCC (HBV+, SF7)	<a href="#">GSM79798</a> : Liver (HBV+, HK13)
<a href="#">GSM79799</a> : HCC (HBV+, SF35, 3)	<a href="#">GSM79800</a> : Liver (HBV+, HK47)	<a href="#">GSM79801</a> : Hep3B cells

## [Sample GSM79784](#) Simple annotation: Non-tumor tissue, Liver

This is meta data description of data published in Chen paper Liver cancer 2002.

```
<publication>
!Citation=Chen et al.  MBC in Press, published April 3, 2002 as 10.1091/mbc.02-02-0023
!Title=Gene expression patterns in human liver cancers
!PubMedID=
  <experiment_set>
    !Name=HCCpaper_All_arrays
    !ExptSetNo=1162
    !Description=All the array experiments published in "Gene expression
patterns in human liver cancers" by Chen X, et al.
  </experiment_set>
  <experiment_set>
    !Name=HCCpaper_Figure1_arrays
    !ExptSetNo=1164
    !Description=Experiments used in Figure one of the HCC paper by Chen X, et
al.
  </experiment_set>
  <experiment_set>
    !Name=HCCpaper_Figure3_arrays
    !ExptSetNo=1165
    !Description=Experiments used in Figure 3 of the HCC paper by Chen X, et
al.
  </experiment_set>
  <experiment_set>
    !Name=HCCpaper_Figure6_arrays
    !ExptSetNo=1166
    !Description=Experiments used in the Figure 6 of HCC paper by Chen X, et
al
  </experiment_set>
</publication>
```

To read SOFT files

[http://www2.warwick.ac.uk/fac/sci/moac/currentstudents/peter\\_cock/r/geo/](http://www2.warwick.ac.uk/fac/sci/moac/currentstudents/peter_cock/r/geo/)

soft file format description

<http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html#SOFTformat>

## Sample data table headers and content

The first row in the file must be a header line that identifies the content of each column. The two required columns are listed below. In addition to the required columns, submitters are encouraged to supply any number of auxiliary non-standard columns describing, for example, supporting measurements and calculations, quality



evaluations or flags. Columns may appear in any order after the ID\_REF column. In this way, GEO is a flexible and open system, allowing you to provide all information necessary to thoroughly describe your hybridization results.

- **ID\_REF:** (Required) Identifier reference - these should match the unique identifiers given in the identifier (ID) column of the corresponding Platform data table.
- **VALUE:** (Required) These values should be the final, normalized quantification measurements that are comparable across rows and Samples, and preferably processed as described in any accompanying manuscript. Values that should be discarded (e.g., background higher than count, or otherwise flagged as 'bad') should either be left blank or labeled as "null".
  - For single channel data, this column should contain normalized (scaled) signal count data.
  - For dual channel data, this column should contain normalized log ratio data (preferably test/reference).


Matlab bioinformatics toolbox

[http://www.mathworks.com/access/helpdesk\\_r13/help/toolbox/bioinfo/a1052335616.html](http://www.mathworks.com/access/helpdesk_r13/help/toolbox/bioinfo/a1052335616.html)

<http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html>

Free Hotmail RealPlayer TOSHIBA Access Windows Marketplace Windows Media

## The Colors of a Microarray



Reproduced with permission from the Office of Science Education, the National Institutes of Health.

In this schematic:

**GREEN** represents **Control DNA**, where either DNA or cDNA derived from normal tissue is hybridized to the target DNA.

**RED** represents **Sample DNA**, where either DNA or cDNA is derived from diseased tissue hybridized to the target DNA.

**YELLOW** represents **a combination of Control and Sample DNA**, where both hybridized equally to the target DNA.

**BLACK** represents areas where **neither the Control nor Sample DNA** hybridized to the target DNA.

Each spot on an array is associated with a particular gene. Each color in an array represents either healthy (control) or diseased (sample) tissue. Depending on the type of array used, the location and intensity of a color will tell us whether the gene, or mutation, is present in either the control and/or sample DNA. It will also provide an estimate of the expression level of the gene(s) in the sample and control DNA.

Some learning :

<http://learn.genetics.utah.edu/units/basics/tour/>

soft file:

```
!Sample_series_id = GSE3500
!Sample_data_row_count = 24192
#ID_REF = ID_REF
```

```
#CH1I_MEAN = Mean feature pixel intensity at wavelength 532 nm.; Type: integer; Scale: linear_scale
```

```
#CH2I_MEAN = Mean feature pixel intensity at wavelength 635 nm.; Type: integer; Scale: linear_scale
```

```
#CH1B_MEDIAN = The median feature background intensity at wavelength 532 nm.; Type: integer; Scale: linear_scale; Channel: Cy3 Channel; Background
```

```
#CH2B_MEDIAN = The median feature background intensity at wavelength 635 nm.; Type: integer; Scale: linear_scale; Channel: Cy5 channel; Background
```

#CH1D\_MEAN = The mean feature pixel intensity at wavelength 532 nm with the median background subtracted.; Type: integer; Scale: linear\_scale; Channel: Cy3 Channel

#CH2D\_MEAN = .The mean feature pixel intensity at wavelength 635 nm with the median background subtracted.; Type: integer; Scale: linear\_scale; Channel: Cy5 channel

#CH1I\_MEDIAN = Median feature pixel intensity at wavelength 532 nm.; Type: integer; Scale: linear\_scale

#CH2I\_MEDIAN = Median feature pixel intensity at wavelength 635 nm.; Type: integer; Scale: linear\_scale

#CH1B\_MEAN = The mean feature background intensity at wavelength 532 nm.; Type: integer; Scale: linear\_scale; Background

#CH2B\_MEAN = The mean feature background intensity at wavelength 635 nm.; Type: integer; Scale: linear\_scale; Background

#CH1D\_MEDIAN = The median feature pixel intensity at wavelength 532 nm with the median background subtracted.; Type: integer; Scale: linear\_scale

#CH2D\_MEDIAN = The median feature pixel intensity at wavelength 635 nm with the median background subtracted.; Type: integer; Scale: linear\_scale

#CH1\_PER\_SAT = The percentage of feature pixels at wavelength 532 nm that are saturated.; Type: integer; Scale: linear\_scale

#CH2\_PER\_SAT = The percentage of feature pixels at wavelength 635 nm that are saturated.; Type: integer; Scale: linear\_scale

#CH1I\_SD = The standard deviation of the feature intensity at wavelength 532 nm.; Type: integer; Scale: linear\_scale; Channel: Cy3 Channel

#CH2I\_SD = The standard deviation of the feature pixel intensity at wavelength 635 nm.; Type: integer; Scale: linear\_scale; Channel: Cy5 channel

#CH1B\_SD = The standard deviation of the feature background intensity at wavelength 532 nm.; Type: float; Scale: linear\_scale; Channel: Cy3 Channel; Background

#CH2B\_SD = The standard deviation of the feature background intensity at wavelength 635 nm.; Type: integer; Scale: linear\_scale; Channel: Cy5 channel; Background

#PERGTBCH1I\_1SD = The percentage of feature pixels with intensities more than one standard deviation above the background pixel intensity, at wavelength 532 nm.; Type: integer; Scale: linear\_scale

#PERGTBCH2I\_1SD = The percentage of feature pixels with intensities more than one standard deviation above the background pixel intensity, at wavelength 635 nm.; Type: integer; Scale: linear\_scale

#PERGTBCH1I\_2SD = The percentage of feature pixels with intensities more than two standard deviations above the background pixel intensity, at wavelength 532 nm.; Type: integer; Scale: linear\_scale

#PERGTBCH2I\_2SD = The percentage of feature pixels with intensities more than two standard deviations above the background pixel intensity, at wavelength 635 nm.; Type: integer; Scale: linear\_scale

#SUM\_MEAN = The sum of the arithmetic mean intensities for each wavelength, with the median background subtracted.; Type: integer; Scale: linear\_scale

#SUM\_MEDIAN = The sum of the median intensities for each wavelength, with the median background subtracted.; Type: integer; Scale: linear\_scale

#RAT1\_MEAN = Ratio of the arithmetic mean intensities of each spot for each wavelength, with the median background subtracted. Channel 1/Channel 2 ratio, (CH1I\_MEAN - CH1B\_MEDIAN)/(CH2I\_MEAN - CH2B\_MEDIAN) or Green/Red ratio.; Type: float; Scale: linear\_scale

#RAT2\_MEAN = The ratio of the arithmetic mean intensities of each feature for each wavelength, with the median background subtracted.; Type: float; Scale: linear\_scale

#RAT2\_MEDIAN = The ratio of the median intensities of each feature for each wavelength, with the median background subtracted.; Type: float; Scale: linear\_scale

#PIX\_RAT2\_MEAN = The geometric mean of the pixel-by-pixel ratios of pixel intensities, with the median background subtracted.; Type: float; Scale: linear\_scale

#PIX\_RAT2\_MEDIAN = The median of pixel-by-pixel ratios of pixel intensities, with the median background subtracted.; Type: float; Scale: linear\_scale

#RAT2\_SD = The geometric standard deviation of the pixel intensity ratios.; Type: float; Scale: linear\_scale

#TOT\_SPIX = The total number of feature pixels.; Type: integer; Scale: linear\_scale

#TOT\_BPIX = The total number of background pixels.; Type: integer; Scale: linear\_scale

#REGR = The regression ratio of every pixel in a 2-feature-diameter circle around the center of the feature.; Type: float; Scale: linear\_scale

#CORR = The correlation between channel1 (Cy3) & Channel 2 (Cy5) pixels within the spot, and is a useful quality control parameter. Generally, high values imply better fit & good spot quality.; Type: float; Scale: linear\_scale

#DIAMETER = The diameter in um of the feature-indicator.; Type: integer; Scale: linear\_scale

#X\_COORD = X-coordinate of the center of the spot-indicator associated with the spot, where (0,0) is the top left of the image.; Type: integer; Scale: linear\_scale

#Y\_COORD = Y-coordinate of the center of the spot-indicator associated with the spot, where (0,0) is the top left of the image.; Type: integer; Scale: linear\_scale

#TOP = Box top:  $\text{int}(((\text{centerX} - \text{radius}) - \text{Xoffset}) / \text{pixelSize})$ .; Type: integer; Scale: linear\_scale

#BOT = Box bottom:  $\text{int}(((\text{centerX} + \text{radius}) - \text{Xoffset}) / \text{pixelSize})$ .; Type: integer; Scale: linear\_scale

#LEFT = Box left:  $\text{int}(((\text{centerY} - \text{radius}) - \text{yoffset}) / \text{pixelSize})$ .; Type: integer; Scale: linear\_scale

#RIGHT = Box right:  $\text{int}(((\text{centerY} + \text{radius}) - \text{yoffset}) / \text{pixelSize})$ .; Type: integer; Scale: linear\_scale

#FLAG = The type of flag associated with a feature: -100 = user-flagged null spot; -50 = software-flagged null spot; 0 = spot valid.; Type: integer; Scale: linear\_scale

#CH2IN\_MEAN = Normalized value of mean Channel 2 (usually 635 nm) intensity (CH2I\_MEAN/Normalization factor).; Type: integer; Scale: linear\_scale; Channel: Cy5 channel

#CH2BN\_MEDIAN = Normalized value of median Channel 2 (usually 635 nm) background (CH2B\_MEDIAN/Normalization factor).; Type: integer; Scale: linear\_scale; Channel: Cy5 channel; Background

#CH2DN\_MEAN = Normalized value of mean Channel 2 (usually 635 nm) intensity with normalized background subtracted (CH2IN\_MEAN - CH2BN\_MEDIAN).; Type: integer; Scale: linear\_scale; Channel: Cy5 channel

#RAT2N\_MEAN = Type: float; Scale: linear\_scale

#CH2IN\_MEDIAN = Normalized value of median Channel 2 (usually 635 nm) intensity (CH2I\_MEDIAN/Normalization factor).; Type: integer; Scale: linear\_scale

#CH2DN\_MEDIAN = Normalized value of median Channel 2 (usually 635 nm) intensity with normalized background subtracted (CH2IN\_MEDIAN - CH2BN\_MEDIAN).; Type: integer; Scale: linear\_scale

#RAT1N\_MEAN = Ratio of the means of Channel 1 (usually 532 nm) intensity to normalized Channel 2 (usually 635 nm) intensity with median background subtracted (CH1D\_MEAN/CH2DN\_MEAN). Channel 1/Channel 2 ratio normalized or Green/Red ratio normalized.; Type: float; Scale: linear\_scale

#RAT2N\_MEDIAN = Channel 2/Channel 1 ratio normalized, RAT2\_MEDIAN/Normalization factor or Red/Green median ratio normalized.; Type: float; Scale: linear\_scale

#VALUE = Log (base 2) of the ratio of the mean of Channel 2 (usually 635 nm) to Channel 1 (usually 532 nm) [log (base 2) (RAT2N\_MEAN)].; Type: float; Scale: log\_base\_2

#LOG\_RAT2N\_MEDIAN = Log (base 2) of the ratio of the median of Channel 2 (usually 635 nm) to Channel 1 (usually 532 nm) [log (base 2) (RAT2N\_MEDIAN)].; Type: float; Scale: log\_base\_2  
!sample\_table\_begin

The Gene Expression Omnibus (**GEO**) repository at the National Center for Biotechnology Information (NCBI) archives and freely disseminates microarray and other forms of high-throughput data generated by the scientific community.

<http://nar.oxfordjournals.org/cgi/content/full/gkl887?ijkey=ysG9Li2nfUYJvdZ&keytype=ref>

chen data for liver, 2002 paper:

<http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSM79784>

control is on channel 1, tumor on channel 2.

Platform ID [GPL3009](#) Series (1)

[GSE3500](#) Gene expression patterns in human liver cancers

Software format descriptions. Also shows relationship between platform/samples/series

<http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html#SOFTformat>

The distinctive gene expression patterns are characteristic of the tumors and not the patient

~1640 genes that are differentially expression in HCC and non-tumor liver

Top 600 genes that are differentially expressed in HCC and Non-tumor Liver

Microarray procedure

23075 cDNA clones, representing about 17,400 genes, were mechanically printed onto treated glass microscope slides

```
EDU>> getgeodata('GSM79795','ToFile','GSM79795.txt');
```

```
EDU>> GEOSOFTData = geosoftread('GSM79795.txt');
```

### **GEO Platform (GPL)**

These files describe a particular type of microarray. They are annotation files.

### **GEO Sample (GSM)**

Files that contain all the data from the use of a single chip. For each gene there will be multiple scores including the main one, held in the VALUE column.

### **GEO Series (GSE)**

Lists of GSM files that together form a single experiment.

### **GEO Dataset (GDS)**

These are curated files that hold a summarised combination of a GSE file and its GSM files. They contain the expression level for each gene from each sample (i.e. just the VALUE field from the GSM file).

Format for the PLATFORM file. It looks I might need access to this also to know which gene goes with which spot

#ID =

#METACOLUMN =

#METAROW =  
#COLUMN =  
#ROW =  
#SPOT\_ID =  
#CONTROL =  
#SEQUENCE DESCRIPTION =  
#POLYMER =  
#TYPE =  
#GenBank =

Describes formats

<http://www.ncbi.nlm.nih.gov/projects/geo/info/overview.html>

platform record

### Data table header descriptions

<b>ID</b>	Affymetrix Probe Set ID
<b>Species Scientific Name</b>	The genus and species of the organism represented by the probe set.
<b>Annotation Date</b>	The date that the annotations for this probe array were last updated. It will generally be earlier than the date when the annotations were posted on the Affymetrix web site.
<b>GB_LIST</b>	GenBank Accession Number
<b>SPOT_ID</b>	Sequence Type: Indicates whether the sequence is an Exemplar, Consensus or Control sequence. An Exemplar is a single nucleotide sequence taken directly from a public database. This sequence could be an mRNA or EST. A Consensus sequence, is a nucleotide sequence assembled by Affymetrix, based on one or more sequence taken from a public database.
<b>Sequence Source</b>	The database from which the sequence used to design this probe set was taken.
<b>Representative Public ID</b>	The accession number of a representative sequence. Note that for consensus-based probe sets, the representative sequence is only one of several sequences (sequence sub-clusters) used to build the consensus sequence and it is not directly used to derive the probe sequences. The representative sequence is chosen during array design as a sequence that is best associated with the transcribed region being interrogated by the probe set. Refer to the "Sequence Source" field to determine the database used.
<b>Gene Title</b>	Title of Gene represented by the probe set.
<b>Gene Symbol</b>	A gene symbol, when one is available (from UniGene).
<b>Entrez Gene</b>	Entrez Gene database UID
<b>RefSeq Transcript ID</b>	References to multiple sequences in RefSeq. The field contains the ID and Description for each entry, and there can be multiple entries per ProbeSet.
<b>RGD Name</b>	Rat Genome Database
<b>Gene Ontology Biological Process</b>	Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence". The description corresponds

directly to the GO ID. The evidence can be "direct", or "extended".

**Gene Ontology Cellular Component** Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence". The description corresponds directly to the GO ID. The evidence can be "direct", or "extended".

**Gene Ontology Molecular Function** Gene Ontology Consortium Molecular Function derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence". The description corresponds directly to the GO ID. The evidence can be "direct", or "extended".

Fields in PLATFORM record

**ID MetaRow** Number of metarow in the metagrid where the spot is located  
**MetaColumn** Number of metacolumn in the metagrid where the spot is located  
**Row** Number of row in the subgrid where the spot is located  
**Column** Number of column in the subgrid where the spot is located  
**Array ID** Array identifier  
**Accession** Accession of TIGR Gene Indices contig or GenBank number  
**GB\_ACC** GenBank accession  
**SPOT\_ID SPOT ID** Spot identifier  
**Description** Gene description based on current top blast hit  
**SEQUENCE** Sequence

Look for: AA225741,AA259201,AI732153,AI820965

```
Platform GPL2831
9 1 1 9 1 IMAGE:1008379 DNA cDNA_clone

Platform GPL2938 (44,500 rows)
33301 4 9 21 26 IMAGE:1008379 DNA cDNA_clone

PLATFORM = GPL3007
705 1 1 5 26 IMAGE:1008379 DNA cDNA_clone

PLATFORM = GPL3009
585 1 1 25 21 IMAGE:1008379 DNA cDNA_clone
```

The following Platforms that has AA225741,AA259201,AI732153,AI820965 on location

```
729 1 1 1 27
GPL2648,GPL2649, GPL2868, GPL2906, GPL2935, GPL2948 , GPL3008, GPL3010,GPL3011
```

Look at spot ID:

```
1491 2 1 7 27 IMAGE:1286706 DNA cDNA_clone AA740767
735 1 1 7 27 IMAGE:1286706 DNA cDNA_clone AA740767
711 1 1 11 26 IMAGE:1286706 DNA cDNA_clone AA740767
```

<http://smd.stanford.edu/cgi-bin/data/viewData.pl?fullID=12416GENEPIX0>  
 use SMD, not GEO.  
 SMD allows better control on what fields to download

To see effective of filtering on data, go to <http://smd.stanford.edu/cgi-bin/data/grids.pl?fullID=12406GENEPIX0> and select different filters.

24,192 is the total number of spots on the microarray. However, not all spots can be loaded with genes, some can be empty. Hence the number of genes does not necessarily the same as number of spots.

SMD scheme !

<https://genome.unc.edu/cgi-bin/SMD/tableSpecifications?table=RESULT>

**SECTOR** spot sector grid coordinate

**SECTORROW** spot row grid coordinate

**SECTORCOL** spot column grid coordinate

From <http://www.microarray.org/sfgf/common/misc/faq.jsp>

- **How are the features(spots) arranged on the SFGF microarray?**

**The current MM array edition contains 48 sectors of 30 rows and 30 columns of spots in each sector, printed at 146 micron spacing. The recent SH arrays have been printed with 30 columns per sector and with 29 or 30 rows in each of the 48 sectors. The spot diameter is 80 - 90 microns. For details about your particular array batch, please refer to the QC notes page of the website(you must have an SFGF account and be logged in to access the QC notes page).**

**The total possible number of spots that we can fit into this configuration is 43,200, and there is one array per slide.**

**If one orients the array with the barcode at the bottom, the sectors are numbered with #1 in the upper left-hand corner, 2-4 proceeding to the right, 5-8 in the second row of sectors, and so on.**

**Sector numbering for my project**

**1 2 3 4  
5 6 7 8  
9 10 11 12  
13 14 15 16  
17 18 19 20  
21 22 23 24  
25 26 27 28  
29 30 31 32**