# GS01 0163 Analysis of Microarray Data

Keith Baggerly and Kevin Coombes Section of Bioinformatics Department of Biostatistics and Applied Mathematics UT M. D. Anderson Cancer Center kabagg@mdanderson.org kcoombes@mdanderson.org

August 31, 2006

# Lecture 2: The Basics of dChip

- So, why are we here?
- Getting the stuff required
- Using dChip
  - Loading Data
  - Looking at Data
  - Normalizing Data
  - Model Fitting
  - Exporting Results
- The Real World...

# So, why are we here?

We want to learn about dChip.

The freeware package dChip has become quite widely used for the analysis of Affymetrix gene chip data. We're going to look at using it now.

The main web page for dChip is

http://biosun1.harvard.edu/complab/dchip/

where you can download the software, get links to some publicly available data, and browse through the online manual.

Much of this lecture will follow the manual, and the associated "Short tutorial" and "Lab" with my editorial comments.

# Step 1: Get dChip

This step is fairly trival; simply download the latest version (dchip2006.exe as of August 30, 2006) and put the application somewhere (eg, D:Program Files/dChip2006/). We keep this application on a shared drive at

/data/bioinfo/affymetrix/00 Affymetrix Info/DChip Files

The entire application is about 1.7M in size. At present, dChip only runs on Windows platforms. Some success has been reported using windows emulators on the Mac, but there is a performance hit.

# A Biological Example

There is a genetic translocation that occurs in ALL, associated with a mixed-lineage leukemia gene (MLL). Patients with this translocation have noticably worse outcomes. It is thought that this translocation may make the disease qualitatively different, and somewhat closer to AML. If the disease is different, we may want to adjust the therapy as well.

Using Affymetrix gene chips, can we identify differences between ALL, MLL, and AML?

# **Step 2: Get CEL Files**

The Lab page supplies a link to the example data CEL files on leukemia (ALL,MLL,AML) from Dana Farber.

http://www.broad.mit.edu/cgi-bin/cancer/
publications/pub\_paper.cgi?mode=view&paper\_id=63

The CEL files are available as gzipped tar files, which WinZip should be able to uncompress. There are 6 CEL file collections at this site, each about 35-41M in size, or 100-127M in size when uncompressed. These files contain about 10-12 CEL files each.

The suffixes on these files should be .tar.gz, but for some reason they are tar.tar. This latter suffix needs to be changed so that the file type will be recognized.

# Step 2: Get CEL Files (cont)

If you are working with CEL files stored in more than one location, it is often useful to assemble a "data list file" specifying the locations of the files. This file should be a text file (and end in .txt). Every row should contain either a specific file name or a directory. An example from the manual:

E:\Affy data\dan\CA-H.cel E:\Affy data\dan\CA-HR.cel E:\Affy data\dan\zugen E:\Affy data\dan\PC-C.cel

Here, the AML samples were run later, so we put them in a different directory.

# **Step 2A: Digression**

The Dana Farber web site also supplies the quantifications that they used in their analyses, as

expression\_data.txt

or

expression\_data\_plus\_APcalls.txt

These data were initially quantified using MAS4.0 (AvDiff). We prefer to work with the CEL files as raw data and to construct our own quantifications.

# **Step 3: Get Explanatory Files**

Also at the above site, there are files describing the sample-to-chip mapping in more detail:

scaling\_factors\_and\_fig\_key.txt

and a link to the paper that appeared in Nature Genetics describing the biological context of the problem.

# **Step 4: Find the CDF file**

This requires that we know what type of Affy chip was used. In this case (according to the paper), the chips were U95A.

For this example, a compressed version of the CDF file can be downloaded from the dChip site; more generally, we have a collection of CDF files for the chip types we use in

data/bioinfo/affymetrix/00 Affymetrix Info/CDF Files

A warning – the cdf extension is also used for "channel files" by Microsoft, so don't worry if you see a weird icon.

### **Step 4A: Digression**

Actually, the CDF file for these chips is a bit tricky.

There is a set of U95 chips, U95A,U95B,...,U95E that contain probes for all genes in the genome. The probes were assembled using the 95th build of the Unigene database to define what a "gene" was. However, while these chips surveyed the genome, most of the probes corresponding to "interesting" genes were put on the A chip, so most people just bought those as opposed to the set.

Soon after the U95A release, some mistakes were noted in the probe design, and Affy released the U95Av2, which is the type we have encountered more frequently here at MDA.

Can you tell them apart?

# Step 5: Get the Gene Info file(s)

Every chip type has a fixed set of probesets printed on it, but the probeset identifiers are typically not enough to suggest anything (1389\_at?). We need more context – is there a common name for the associated gene? Which chromosome is it on, and where? Is the gene known or thought to be part of a functional family (eg, cytoskeleton)? Are there IDs that can let us look up more information in national databases?

The above information for each chip type has been collected and assembled into GeneInfo files available at the dChip website. *These files are tab-delimited text files, but they've had an xls extension placed on them so that Excel is the default program for opening them.* 

These info files can change over time!

# Step 5: Get the Gene Info file(s) (cont)

Actually, when we download the zip file from the dChip web site, we get 3 files:

HG-U95Av2 gene info2.xls HG-U95Av2 gene info2 Gene Ontology.xls HG-U95Av2 gene info2 Protein Domain.xls

We're going to look at each of these in turn, but I want to quickly note that these files are for the U95Av2 chip, as opposed to the U95A chip. In terms of the probesets that were used, the overlap is so large (12600 of 12625) that working with these should be fine.

These files are on our system in

/data/bioinfo/affymetrix/00 Affymetrix Info/DChip Files

## HG-U95Av2 gene info2.xls

The first few entries:

```
Probe Set Name : Identifier : LocusLink :
 Name : Gene Ontology.xls : Protein Domain.xls :
 Pathway : Chromosome : Description
1000_at : X60188 : 5595 :
 mitogen-activated protein kinase 3 : 7165
 7154 | 6935 | 42330 | 9605 | 6928 | 8151 | 4707 | 4702 | 4674 |
 4672 | 16301 | 3824 | 16773 | 16772 | 16740 | 5057 | 4871 | :
 2290 719 3527 : : 16 16p 16p 12 :
 X60188 /FEATURE=mRNA /DEFINITION=HSERK1 Human ERK1
 mRNA for protein serine/threonine kinase
1001_at : X60957 : 7075 :
 tyrosine kinase with immunoglobulin and epidermal
 growth factor homology domains : |7498|9888|
```

# HG-U95Av2 gene info2 Gene Ontology.xls

Term ID	Term Name Frequency	
3	reproduction 101	
18	regulation of DNA recombination 9	
41	transition metal transport 16	
67	DNA replication and chromosome cycle 10	3
70	mitotic chromosome segregation 7	
72	M-phase specific microtubule process 8	
74	regulation of cell cycle 330	
75	cell cycle checkpoint 35	
76	DNA replication checkpoint 8	

### HG-U95Av2 gene info2 Protein Domain.xls

Frequency Term ID Term Name Kringle 16 1 Cdc20/Fizzy 2 4 Retinoid X receptor 3 15 4 Saposin type B 5 Helix-turn-helix, AraC type 5 11 Vertebrate metallothionein, family 1 6 6 7 Tubby - 7 C2 domain 8 84 Cysteine proteases inhibitor 10 18

"what ghastly names they all have..." E. J. (Ernest John) Moncrieff

Most of the files that we have worked with so far have described properties associated with a given chip type, not with the samples we have used. We can also supply and use sample-specific information in a tab-delimited text file. The first few entries here:

scan name	sample_	name	type
CL2001011101AA	ALL_1	A	
CL2001011104AA	ALL_2	A	
CL2001011105AA	ALL_3	A	
CL2001011108AA	ALL_4	A	
CL2001011109AA	ALL_5	A	
CL2001011111AA	ALL_6	A	
CL2001011112AA	ALL_7	A	
CL2001011116AA	ALL_8	A	
CL2001011113AA	ALL_9	A	

# Step 6: Get the Sample Info file (cont)

The header row and the first two columns are required, but any columns beyond that are at our discretion. By default, column values are treated as factors, but adding the string "(numeric)" to a column name will override this.

What else could we have included?

- Presence/absence of other translocations
- train/test status
- specimen type (diagnostic, relapse)
- run date...

# **Step 7: write a README file**

Strictly speaking, this is not mentioned in the Lab or Tutorial, but I'll put it here, right before actually running the program.

What is the biological question you are seeking to address? What contrasts of data samples will allow you to address this?

Sending a brief description of this type off to the investigator before running the analysis can save some time...

### Step 8: run dChip

Nice, friendly, unexciting...

🗈 dChip	
<u>Analysis View Data Image</u>	<u>C</u> lustering Ch <u>r</u> omosome <u>P</u> athway <u>T</u> ools <u>H</u> elp
Analysis	Welcome to dChip 2006 (DNA-Chip Analyzer), Build date: Aug 29 2006         Select 'Help/Website' for manual and updates.         This version has format 4 for CDF.BIN and DCP files. If you use it to analyze dChip data in previous format, CDF or CEL
	files will be re-extracted 08/30/06, 12:46

Now, we need to tell it where to find the data for analysis. Go to Analysis/Open Group.

# Finding files, part 1

Working on a group of arrays	×
Data files Other information	
Group name arrayCourseLec2  Delete Data source Data directory or Data file list .txt D:\dChipExample06\InfoFiles\data_file_list.txt File type: DAT CEL DCP(dChip) Suffix of TXT call file: .txt	
<ul> <li>Read unnormalized probe data</li> <li>Read in expression values from TXT file</li> <li>Ignore existing DCP file</li> <li>Perform 'Analysis/Normalize &amp; MBEI' afterwards</li> <li>Help</li> </ul>	
OK Cancel Apply	,

#### assign a group name, locate data files

# Finding files, part 2

Working on a group	p of arrays		
Data files Other information			
CDF file (Chip desc	ription file)		
Select:	D:\dChipExample06\CDFFile\HG_U95A.CDF	Help	
Ignore existing .	cdf.bin file		
Subarray CDF:	None	Help	
Probe sequence	None		
Probeset mask file:	None	Help	
Array type:	Expression 💌		
- Information files			
Gene or SNP:	Gene or SNP: D:\dChipExample06\InfoFiles\HG-U95Av2 gene		
(Do not specify genome information file			
Sample:	D:\dChipExample06\InfoFiles\sample_info.xls_}	<u>Help</u>	
(Probe set mask file, gene/SNP and sample information file are optional)			
	OK Cancel	Apply	

locate CDF, gene info, and sample info files. Under Options, we can set the working directory where results should be stored.

### **Reading files**

arrayCourseLec2	
<u>A</u> nalysis <u>V</u> iew <u>D</u> ata <u>I</u> mage	<u>Clustering</u> Chromosome Pathway Tools Help
Analysis View Data Image Analysis CEL Image Plots	Gene 12600: A Found D:\dChipExample\CELFiles\CL2001011153AA.CEL Reading CEL file Line 400000 Cannot find Affymetrix analysis call file 'D:\dChipExample\CELFiles\CL2001011153AA.bt' No matching TXT file; calculating Presence calls Gene 12600: A Found D:\dChipExample\CELFiles\CL2001011154AA.CEL Reading CEL file Line 400000 Cannot find Affymetrix analysis call file 'D:\dChipExample\CELFiles\CL2001011154AA.bt' No matching TXT file; calculating Presence calls Gene 12600: A Cancot find Affymetrix analysis call file 'D:\dChipExample\CELFiles\CL2001011154AA.bt' No matching TXT file; calculating Presence calls Gene 12600: P Calculating background Reading sample information file 'D:\dChipExampleO6\InfoFiles\sample_ info xis' Read 1 sample information columns (besides the first two columns) for 42 arrays Writing array summary file D:\dChipExampleO6\UntpHies\HG-U95Av2 gene info2 Xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info2 Protein Domain xis' Read in the annotation index file 'D:\dChipExampleO6\InfoFiles\HG-U95Av2 gene info
	Finished in 00 hours 02 minutes 17 seconds)
<	
Click an icon in this window to activ	ivate the corresponding menu

### What has been wrought?

For each CEL file, a binary "dcp" file has been produced:

CL2001011101AA.CEL	10,393 KB
CL2001011101AA.dcp	1,764 KB
CL2001011102AA.CEL	10,324 KB
CL2001011102AA.dcp	1,764 KB

 $(2*640^2)*2 = 1638400$ 

Keep the means as 16-bit integers, and allocate space for 2 CEL equivalents in each dcp file – 1 for the raw data, and 1 for the processed data.

This saves space, and uses an intelligent data structure.

### What has been wrought?

A binary version of the CDF file has been produced for quicker processing.

HG\_U95A.CDF 29,814 KB HG\_U95A.CDF.bin 7,092 KB

## What has been wrought?

3 interim files have been produced:

dChip.ini arrayCourseLec2.ini arrayCourseLec2 arrays.xls

The first two are configuration files, and are stored with the exe file. The last summarizes some aspects of the files examined, and is stored in the working directory.

# The dChip.ini file

#### dChip.ini

- CDF\_FILE=
- $READ_DAT=0$
- READ\_CEL=1
- $READ_DCP=0$
- DATA\_PATH=D:\Program Files\dChip2006 WORKING\_DIR=D:\Program Files\dChip2006 GOSURFER\_DIR=D:\Program Files\dChip2006 USE\_UNNORM=0 MAS5\_SIGNAL=0

# The arrayCourseLec2.ini file

```
arrayCourseLec2.ini
```

CDF\_FILE=D:\dChipExample06\CDFFile\HG\_U95A.CDF

READ\_DAT=0

READ\_CEL=1

READ\_DCP=0

```
DATA_PATH=D:\dChipExample06\InfoFiles\data_file_list
WORKING_DIR=D:\dChipExample06\Output
GOSURFER_DIR=D:\Program Files\dChip2006
USE_UNNORM=0
MAS5 SIGNAL=0
```

# The arrayCourseLec2 arrays.xls file

arrayCourseLec2 arrays.xls

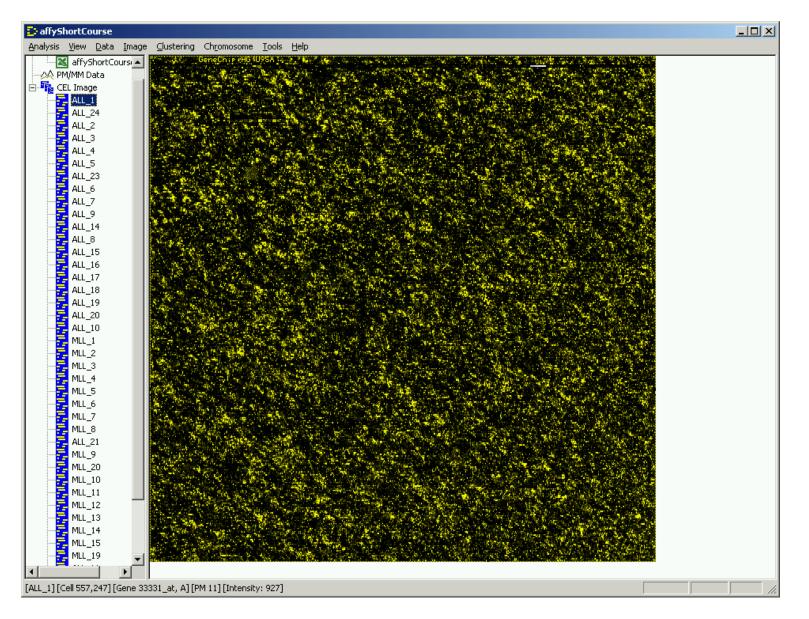
- Number : Array : File Name : Median Intensity
   (unnormalized) : P call %
- 1 : ALL\_1 : D:\dChipExample06\CELFiles\ CL2001011101AA.CEL : 1519 : 48.2
- 2 : ALL\_24 : D:\dChipExample06\CELFiles\ CL2001011102AA.CEL : 1202 : 38.3
- 3 : ALL\_2 : D:\dChipExample06\CELFiles\ CL2001011104AA.CEL : 1795 : 49.5
- 4 : ALL\_3 : D:\dChipExample06\CELFiles\ CL2001011105AA.CEL : 1106 : 36.9

# Look at the Chips

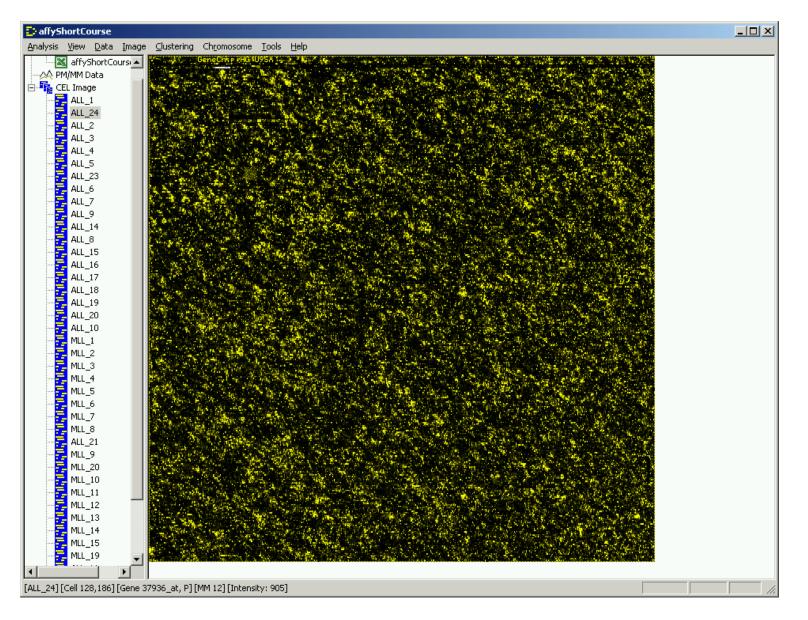
The "Short Tutorial" next suggests going to "View/CEL Image" to look at the data. Unfortunately, this is for an earlier version of dChip, as this pulldown option no longer exists.

So, we click on the "CEL Image" icon at the left of the display and cycle through. If you click on one of the file names, the up and down arrows will let you cycle through them, or Page Up/Page Down also works. The display range covers from the 1st percentile (black) to the 95th (bright yellow).

### Look at the First Chip: ALL\_1



### Look at the Second Chip: ALL\_24



# Zoom In

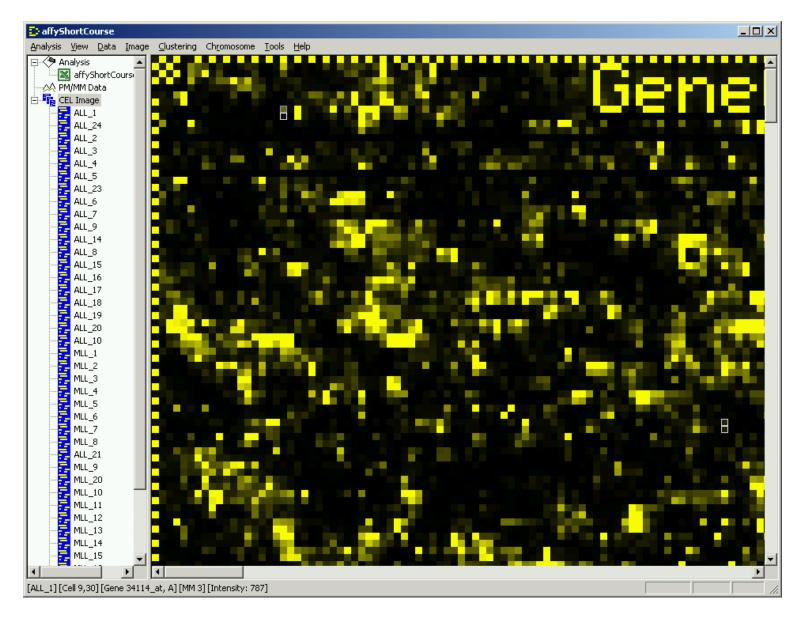
If you click on a part of the image, you select the corresponding probe set. The arrow keys will let you zoom in on the image to look at that spots more closely.

Down arrow: zoom in lots Up arrow: zoom out lots Right arrow: zoom in a little Left arrow: zoom out a little Scrollbars move about

Page Up and Page Down cycle you through the set of chips.

32

### Zoom In: ALL\_1



### Normalize the data

go to Analysis/Normalize & Model

dChip will pick one array in the set to normalize all of the others to; by default it will choose the array with the median overall feature intensity.

(This can make a difference. Trying it with at least two different chips is recommended.)

For each chip, dChip then calculates an "invariant set" of features whose ranks do not change a great deal, and uses those to define a normalization curve.

Functionally, this often works like quantile normalization to the target chip.

# Choosing from the menu...

Normalization and model-based expression/signal	X	
Baseline array [median probe intensity (brigtness)]	1	
ALL_21 [1465]		
(The default baseline array has median brightness. Click CEL image to ensure the baseline is free of image contaminaion and gradient)		
The first array 'ALL_1' has been normalized using baseline array 'ALL_21'		
✓ Perform <u>n</u> ormalization		
☐ ⊻iew normalization plot		
Compute model-based expression/signal values <u>Help</u>		
If 'Normalized' or 'Modelled' shows at lower-right corner, you don't need to perform that step, unless dataset or options have changed.		
Options OK Cancel		

#### Is it normalized?

arrayCourseLec2		
<u>A</u> nalysis <u>V</u> iew <u>D</u> ata Image	<u>C</u> lustering Ch <u>r</u> omosome <u>P</u> athway <u>T</u> ools <u>H</u> elp	
E P Analysis	Searching Invariant-set: 11771	~
arrayCourseLec	Median probe intensity: 1663 -> 1474	
PM/MM Data	MLL_14	
	Accessing 'D:\dChipExample06\CELFiles\CL2001011144AA.dcp' (file format 4)	
ALL_1	Searching Invariant-set: 10003	
ALL_24	Median probe intensity: 1267 -> 1503	
ALL_2	MLL_15	
ALL_3	Accessing 'D:\dChipExample06\CELFiles\CL2001011146AA.dcp' (file format 4)	
ALL_4	Searching Invariant-set: 12445	
ALL_5	Median probe intensity: 1852 -> 1452	
	MLL_19 Accessing 'D:\dChipExample06\CELFiles\CL2001011149AA.dcp' (file format 4)	
ALL_6	Searching Invariant-set: 11989	
ALL_9	Median probe intensity: 1684 -> 1459	
ALL_14	ALL 11	
ALL_14	ALL_11 Accessing 'D:\dChipExample06\CELFiles\CL2001011150AA.dcp' (file format 4)	
ALL_8	Searching Invariant-set: 12065	
ALL_16	Median probe intensity: 1898 -> 1476	
ALL_10	ALL 22	
	Accessing 'D:\dChipExample06\CELFiles\CL2001011151AA.dcp' (file format 4)	
ALL_10	Searching Invariant-set: 11426	
ALL_20	Median probe intensity: 1592 -> 1446	
	MLL_18	
MLL_1	Accessing 'D:\dChipExample06\CELFiles\CL2001011152AA.dcp' (file format 4)	
MLL_2	Searching Invariant-set: 11979	
MLL_3	Median probe intensity: 1777 -> 1489	
	ALL 12	
MLL_5	Accessing 'D:\dChipExample06\CELFiles\CL2001011153AA.dcp' (file format 4)	
MLL_6	Searching Invariant-set: 12725	
MLL_7	Median probe intensity: 1071 -> 1486	
	ALL 13	
ALL_21	Accessing 'D:\dChipExample06\CELFiles\CL2001011154AA.dcp' (file format 4)	
📕 📃 міц_9	Searching Invariant-set: 13396	
MLL_20	Median probe intensity: 1235 -> 1481	
MLL_10	······································	
	Calculating background	
MLL_7 MLL_8 ALL_21 MLL_9 MLL_20 MLL_20 MLL_10 MLL_11 MLL_12		
MLL_13	Finished in 00 hours 01 minutes 16 seconds}	-
MLL_14 🤍	· · · · · · · · · · · · · · · · · · ·	=
	J	Normalized
Analysis outputs		

## Fit the Model 1

go to Analysis/Normalize & Model

Normalization and model-based expression/signal
Baseline array [median probe intensity (brigtness)]
ALL_21 [1465]
(The default baseline array has median brightness. Click CEL image to ensure the baseline is free of image contaminaion and gradient)
The first array 'ALL_1' has been normalized using baseline array 'ALL_21'
Perform <u>n</u> ormalization <u>Help</u>
☐ ⊻iew normalization plot
Compute model-based expression/signal values <u>Help</u>
If 'Normalized' or 'Modelled' shows at lower-right corner, you don't need to perform that step, unless dataset or options have changed.
Options OK Cancel

## Fit the Model 2

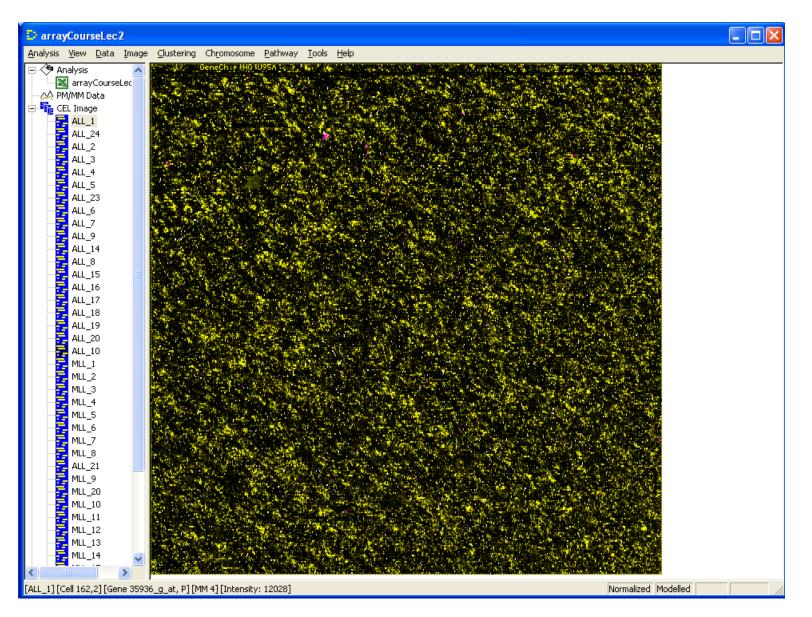
Choose "Options" and select the PM-only model

Options 🛛
Clustering Analysis Model Chromosome
Model-based expression/signal value
Model method: Model-based expression
Background subtraction: 5th percentile of region (PM-only)
Check single, array and probe outliers
Do not call all replicate arrays as array outlier
Exclude 0 5' probes (For degraded or two-round amplified
Compute signals separately for A and B allele for SNP arra
Probe sensitivity index (PSI) file
Usage: Do not use 🗨 Help
File: D:\Program Files\dChip2006\arrayCourseLec2.psi
Normalization
Use selected probes: Invariant set
Probe set file: None
Smoothing method: Running median
Reset Default         Print Settings         OK         Cancel         Apply

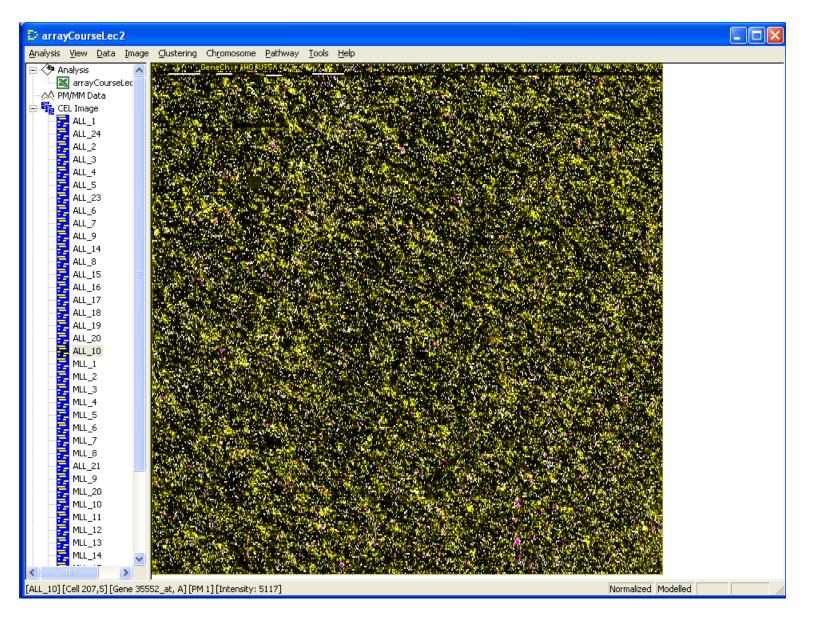
#### Fit the Model 3

arrayCourseLec 2		
<u>A</u> nalysis <u>V</u> iew <u>D</u> ata Image <u>C</u> lusterin	ng Ch <u>r</u> omosome <u>P</u> athway <u>T</u> ools <u>H</u> elp	
Analysis Analysis Aracia Ac	ag diploted by the set of the	
Click an icon in this window to activate the o	corresponding menu Normalized Mode	elled

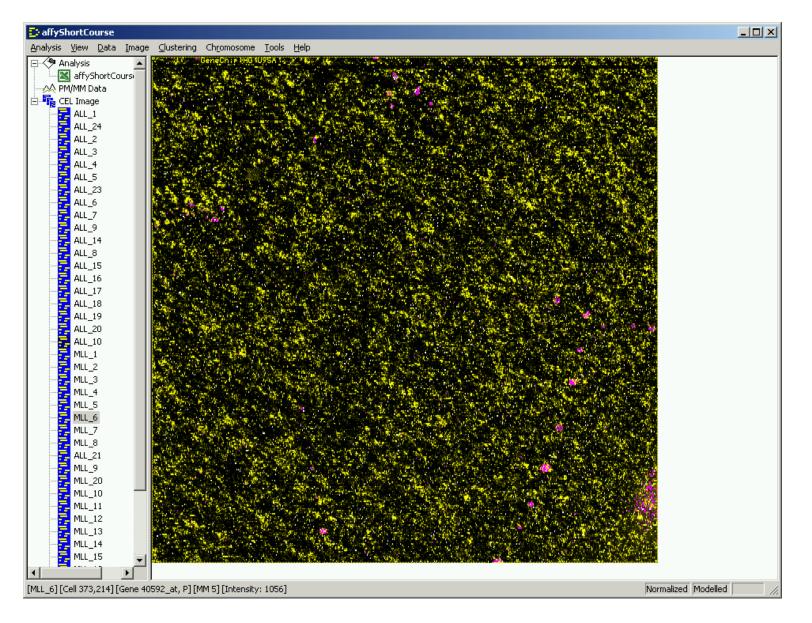
#### Look at the Chips, with Cues



#### Look at the Chips, with Cues



#### Look at the Chips, with Cues



# **Residual Checking is Useful**

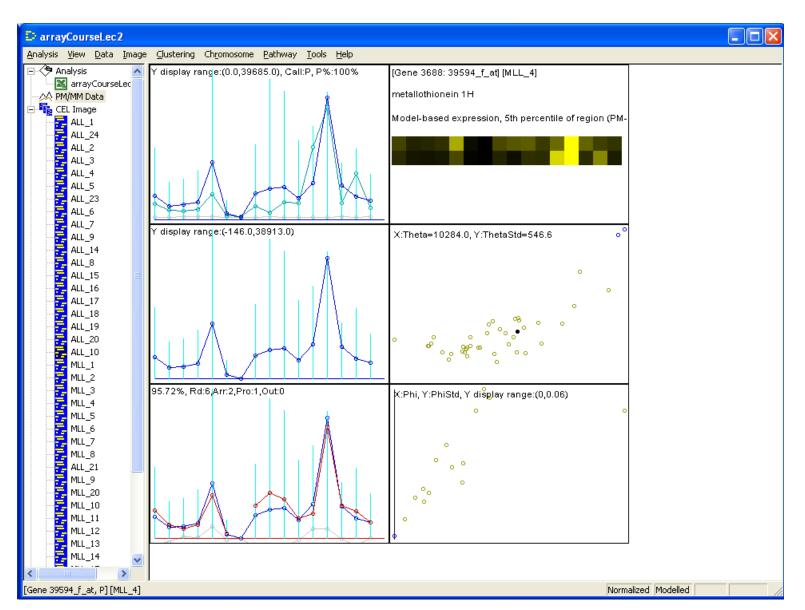
Hitting the "o" key toggles the display of outliers, which can let us look at the values underneath to see if we can spot what the model is picking up.

The file

arrayCourseLec2.xls

has been updated in the model-fitting process to record the number of "array outliers" (high standard errors, in white) and "single outliers" (discounted measurements, in purple). Model fitting is performed in a robust fashion.

So, what does a probeset look like?



#### Look at a Probeset

#### Look at a Probeset

Various panels show

- The PM/MM values for this probeset in this array
- A heatmap view of the same thing
- The target PM-MM values or PM-BG values in this array
- The MBEI values, plotted against their standard errors
- The target values, fitted values, and residuals
- The probe sensitivity indices, plotted against their standard errors

outliers are indicated with colored dots.

#### Look at a Probeset

Cycling through the different chips can be accomplished using the Page Up or Page Down keys. The arrow keys zoom in and out as before, but this feature is less useful here.

Holding down the Page Down key produces an animation effect, which can also be achieved using Data/Animate.

The samples are sorted in order of increasing MBEI values, so cycling through produces a differential effect.

For the sample in question, there were 2 array outliers, 1 probe outlier, and 0 single outliers. The model explained 95.72% of the variation, and iterative fitting took 6 rounds.

So, which probesets are "interesting"?

## **Find Interesting Genes**

Go to Analysis/Compare Samples

Choose the groups using "Select by Category"; this exploits the information that we supplied in the Sample Info file.

One group is "Baseline", the other "Experiment"

Filter using the lower bound of fold change

Filter on absolute differences

## **Find Interesting Genes: Panel 1**

Compare Samples	×
Compare samples Combine compar	isons
Baseline (B)	Experiment (E)
MLL_12 [33] MLL_13 [34] MLL_14 [35] MLL_15 [36] MLL_19 [37] ALL_11 [38] ALL_22 [39] MLL_18 [40] ALL_12 [41]	MLL_10 [31] MLL_11 [32] MLL_12 [33] MLL_13 [34] MLL_14 [35] MLL_15 [36] MLL_19 [37] ALL_11 [38] ALL_22 [39]
Select by category	Select by category
Comparison criteria (1)	✓ or B / E > 1.2
	dence bound of fold change ▼ or B · E > 100
For logged data, use (2)	
(3) 🔲 (P value for testing E	= B) <= 0.05
(4)	% and P call of E >= 20 %
(5) 🔲 (P value for paired t-t	iest) <= 0.05
Help	
	OK Cancel Apply

# **Find Interesting Genes**

Look at "Combine Comparisons"

See where the comparison results will be sent

Estimate FDR using permutations

## **Find Interesting Genes: Panel 2**

Compare Samples 🔀							
Compare samples Combine comparisons							
Combine ty	pe						
C And	C And not	Insert c	omparis	on D	elete e	entry	
💿 Or	🔿 Or not	Insert p	arenthe	sis			
Combine	Baseline	Experiment	E/	or B/E>	U	E-B>	or B
	1,2,3,4,	20,21,22,23	1	1.200	L	10	100.
,							
<							>
Compare on gene list: using all genes							
Oper		):\dChipExample	:06\0u	tout\arravC	oursel	ec2 co	-
Open     D:\dChipExample06\Output\arrayCourseLec2 co       Ouput all genes     Image: Ouput expression values							
✓ Permute samples to assess False Discovery Rate (FDR)     50 times     Help     Options							
		01	<	Canc	el	A	pply

## **Find Interesting Genes – Voila!**

arrayCourseLec 2	
<u>A</u> nalysis <u>V</u> iew <u>D</u> ata <u>I</u> mage	<u>C</u> lustering Chromosome Pathway Tools Help
Analysis vew bata intege	2234242 (326, 513, 10324, 149, 37, 31, 227, 35, 6], genes obtained: 4 Permutation 36: [26, 6, 16, 8, 2, 22, 6, 42, 34, 37, 41, 9, 14, 24, 40, 29, 15, 27, 17, 4, 21, 10, 39, 19 vs. 35, 25, 20, 1, 3, 13, 36, 11, 32, 28, 30, 18, 33, 13, 12, 23, 38, 7], genes obtained: 6 Permutation 36: [28, 65, 14, 1, 38, 13, 36, 25, 52, 21, 16, 17, 19, 15, 27, 31, 3, 30, 40, 11, 42, 12, 35 vs. 24, 4, 33, 10, 14, 39, 29, 37, 20, 22, 82, 23, 21, 86, 72, 82, 29, 34, 83, 22, 25, 31, 23, 30, 8, 9, 36 vs. 26, 37, 41, 35, 15, 24, 33, 20, 144, 11, 32, 10, 12, 40, 16, 6, 39], genes obtained: 10 Permutation 36: [11, 72, 31, 63, 93, 72, 93, 442, 7, 15, 38, 18, 41, 53, 13, 40, 21, 22, 12, 19, 27, 26, 33 vs. 83, 13, 62, 49, 5, 2, 28, 64, 14, 11, 32, 10, 25, 20, 30, 3], genes obtained: 13 Permutation 39: [11, 73, 36, 63, 24, 10, 9], genes obtained: 13 Permutation 39: [11, 73, 36, 63, 74, 11, 29, 73, 33, 28, 9, 74, 38, 11, 44, 35, 11, vs. 64, 23, 91, 63, 33, 33, 26, 23, 41, 14, 24, 25, 56, 40, 21, 15, 14, 11, 33, 51, 1 vs. 64, 23, 39, 16, 33, 33, 36, 42, 41, 144, 21, 71, 19, 39, 41, 42, 26, 58, 40, 21, 15, 14, 11, 13, 35, 11 vs. 64, 23, 39, 16, 36, 37, 31, 29, 27, 33, 28, 9, 74, 142, 158, 24, 11, 38, 4, 28, 7, 32 vs. 35, 36, 16, 23, 24, 31, 39, 26, 34, 17, 29, 73, 32, 89, 74, 142, 158, 24, 11, 38, 4, 28, 7, 32 vs. 35, 36, 16, 23, 24, 31, 39, 26, 34, 17, 29, 73, 32, 89, 71, 17, 23, 18, 35, 59, 26, 742, 38 vs. 16, 43, 21, 11, 33, 63, 39, 14, 25, 40, 16, 55, 22, 37, 24, 30, 127, 17, 15, 29, 34, 24, 71, 28, 36, 20, 24, 24, 34, 0, 27, 7 vs. 22, 15, 19, 11, 18, 39, 30, 9, 35, 62, 53, 15, 41, 16, 29, 10], genes obtained: 18 Permutation 44; [33, 24, 39, 10, 28, 11, 17, 15, 29, 34, 24, 31, 12, 28, 72, 27, 20, 25, 13, 36, 14, 24, 25, 108, 31, 48, 29, 12, 17, 27, 30, 25, 13, 36, 14, 24, 21, 30, 21, 18, 34, 80, 29, 12, 17, 27, 30, 25, 13, 36, 14, 34, 41, 98, 29, 12, 17, 28, 36, 20, 24, 24, 34, 04, 27, 7 vs. 22, 15, 19, 1, 11, 8, 39, 30, 9, 35, 62, 53, 15, 41, 14, 34, 28, 12, 14, 16, 14, 19, 27, 30, 25, 13, 36, 22, vs. 18, 20, 46, 9, 35,
Analysis outputs	Normalized Modelled

## **Find Interesting Genes**

Results are exported to

arrayCourseLec2 compare result.xls

[COMPARE\_CRITERIA\_V2] \$NUM\_OPTION\_LINE=5 \$ARRAY\_LIST\_FILE= \$COMPARE\_ON\_GENE\_LIST= \$COMPARE\_ON\_USE\_LIST=1 \$AVERAGE\_USING\_STANDARD\_ERROR=Yes \$OMIT\_AFFY\_CONTROL\_GENE=Yes \$NUM\_CRITERION=1

#### More compare result.xls (1)

\$Parenthesis : Combine : Baseline : Experiment :

E/B> : or B/E> : Use : E-B> : or B-E> P value <= : P call % of B >= :

and P call % of E >= : % Pair P value <=

No : and : 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,18,19,28,38,39,41,42 :

20,21,22,23,24,25,26,27,29,30,31,32,33,34,35, 36,37,40 :

1.200 : 1.200 : Lower Bound : 100.000 : 100.000 NA : NA : NA : NA

## More compare result.xls (2)

[COMPARE\_RESULT]
probe set : gene : Accession : LocusLink
 Description : ALL\_1 ALL\_24 ALL\_2 ... :
 baseline mean :
 MLL\_1 MLL\_2 MLL\_3 MLL\_4 ... :
 experiment mean :
 fold change : lower bound of FC : upper bound
 of FC : difference of means : filtered

#### More compare result.xls (3)

31407\_at : protease, serine, 7 (enterokinase) :

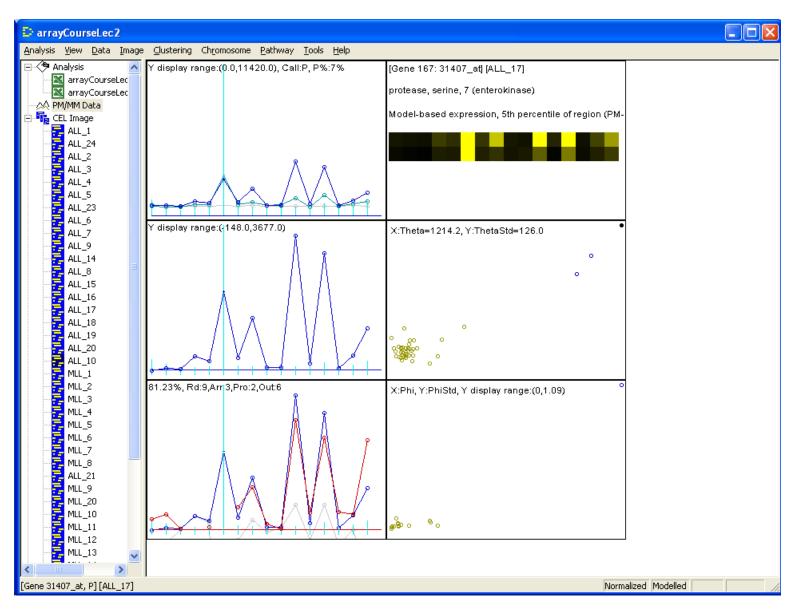
- : U09860 : 5651 : Cluster Incl. U09860:Human enterokinase mRNA, complete cds /cds=(40,3099) /gb=U09860 /gi= 746412 /ug=Hs.158333 /len=3696 : 988.74 158.31 296.43 76.82 427.5 ... : 256.93 : 100.29 64.72 157.82 111.28 110.88 ... : 128.5 :
- -2.15 : -1.28 : -3.09 : -148.05 : \*

#### **Find This Gene**

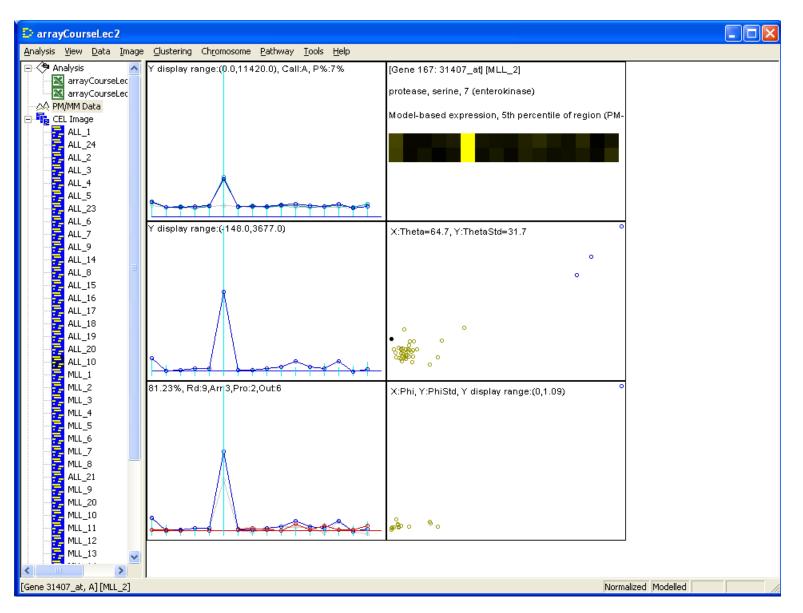
in Probeset View, use View/Find Gene

Find probe set or gene	· ×
Change one of the fo	llowing to search:
Probe set number:	3688
Probe set name:	31407_at
Gene keywords:	metallothionein 1H
	e.g. "receptor * kinase" matches with "receptor both "receptor [1-9]" and "receptor ?" match
0	K Cancel

# Find This Gene: ALL\_17, High End



## Find This Gene: MLL\_17, Low End



## **Other Exports: Expression Results**

Tools/Export Expression Value...

Export expression values	X
Gene list file all genes Use the probe set name in the Arrays to be exported	e 2nd column
MLL_9 MLL_20 MLL_10 MLL_11 MLL_12 MLL_13 MLL_14 MLL_15 MLL_19 ALL_11 ALL_22 MLL_18 ALL_12 ALL_12 ALL_12 Select by category	D:\dChipExample06\Output\arrayCo urseLec2 expression.xls Has both signal and call Has standard error Format: Tab delimited text file Gene names in the last column Include header information Append to this file
Help Options	OK Cancel

## **Export all Expression Results (2)**

#### produces arrayCourseLec2 expression.xls

probe set gene Accession LocusLink Description ALL\_1 ALL\_24 ALL\_2 ALL\_3 ALL\_4 ALL\_5 ALL\_23 ALL\_6 ALL\_7 ... AFFX-MurIL2\_at M16762 Mouse interleukin 2 (IL-2) gen M16762 M16762 Mouse interleukin 2 (IL-2) gen 1324.22 1766.49 1562.23 1739.9 1486.82

1624.63 1759.31 1763.18 1558.21 1555.06

• • •

AFFX-MurIL10\_at interleukin 10 M37897 16153 M37897 Mouse interleukin 10 mRNA, complete cds 917.868 1360.26 1067.69 1380.64 1037.5 1074.34 1294.49 1109.37 1181.09 1090.53 1121.5

## **Other Exports: Probe Results**

Tools/Export Probe Set...

produces

arrayCourseLec2 31407\_at probe data.xls

Probeset	Probe	Array	PM	MM	Bkgrd
Theta	Theta_Std	Phi	I	PhiStd	
31407_at	0	0	985	805	842
988.743	85.1642	0.221	123 (	).1212	87
31407_at	0	1	976	786	812
158.308	29.8064	0.2211	123 (	).1212	87

## **Other Exports: PSIs**

Keep the PSIs? Analysis/Model-based Expression, Options, Usage: Write

Options	X
Clustering Analysis Model Chromosome	
Model-based expression index         Method       PM-only model         Image: Check single outlier       Image: Check array outlier         Image: Check array outlier       Image: Check array outlier         Image: Treat image spikes as single outlier       Image: Do not call all replicate arrays as array outlier         Image: Exclude       0       5' probes	
Probe sensitivity index (PSI) file Usage: Write Help File: D:\Program Files\dChip2004\affyShortC	
SNP array	
Reset Default OK Cancel Apply	

# So, Did We Find What They Did?

#### Well...

It turns out that half of the chips used were U95A, and the rest (including all of the AML samples) were U95Av2. By default, dChip does not combine results from different chip types. However, since the difference is not large (25 probesets out of 12625), we can mask the ones that don't overlap and get it to fit anyway.

## **Combine the Chip Types**

Wo	orking on a group	of arrays	×	
D	ata files Other inform	nation		
	- CDF file (Chip descr	iption file)		
	Select:	D:\dChipExample06\CDFFile\HG_U95Av2.CDF	Help	
	I Ignore existing .c	df.bin file		
	Subarray CDF:	None	Help	
	Probe sequence	None		
	Probeset mask file:	D:\dChipExample06\InfoFiles\hg_u95av2 probe	Help	
	Array type:	Expression		
	- Information files			
	Gene or SNP:	D:\dChipExample06\InfoFiles\HG-U95Av2 gene	Help	
	(Do not specify geno	ome information file		
	Sample:	D:\dChipExample06\InfoFiles\sample_info.xls	<u>Help</u>	
(Probe set mask file, gene/SNP and sample information file are optional)				
		OK Cancel	Apply	

# the mask file is from the dChip web site, and we use the U95Av2 CDF file.

# **Do We Find What They Did Now?**

#### Well...

It turns out that the paper reported gene names and gene symbols, but did not specify the Affymetrix probe ids. Unfortunately, some of the annotation has changed over time.

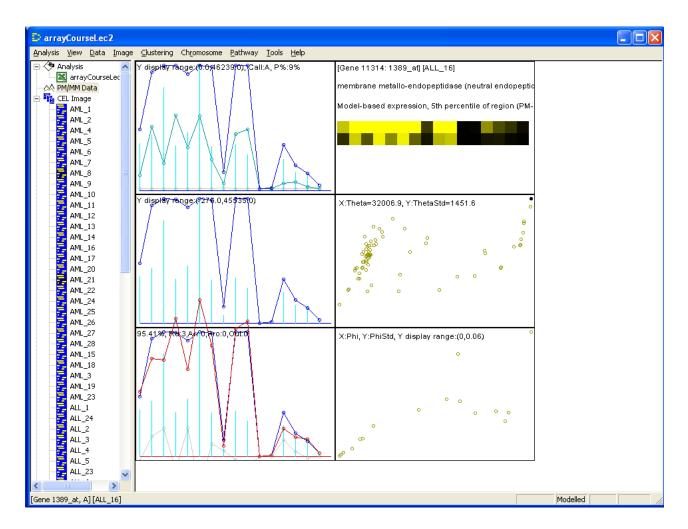
If we look for

J03779 (gene accession number), aka CD10 (gene symbol)

in the expression tables supplied with the paper, it's not there. But if we look in the gene info files supplied with dChip, it *is* there (it's 1389\_at).

#### And?

#### FC: -3.91, CI: (-3.36,-4.5), Diff: -16956.1. Different!



## Summary

We know what files to track down

We know how to load them in for processing

We know how to normalize and fit models

We know how to export results

We've seen how finicky indexing can be.

And we struck biology!

Thus endeth the lesson...