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

8 November 2005

Lecture 19: A Two-Color Case Study

- Case Study Biology
- Getting Data
- Inferences from GPR Files
- Quality Checks
- Further Analysis

The Biology

Working with a case study. this follows Chapter 4 of Gentleman et al (2005), “Preprocessing Two-Color Spotted Arrays”, by Y.H. Yang and A.C. Paquet.

The dataset used here is a subset of a larger dataset described in Rodriguez et al (2004), “Differential gene expression by integrin β7+ and β7- memory T helper cells”, BMC Immunology, 5:13.

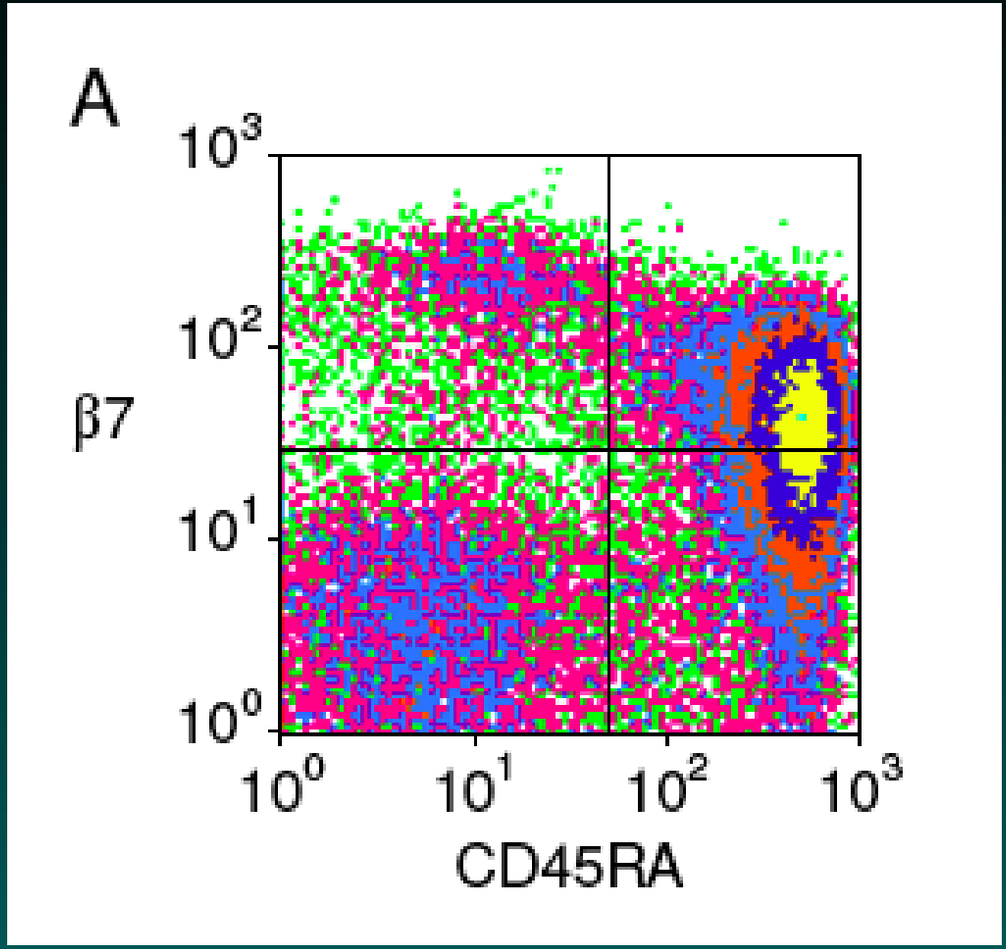
In that paper, they asked whether different types of helper cells were associated with the adhesion or migration of T cells.

How do we Get Cells?

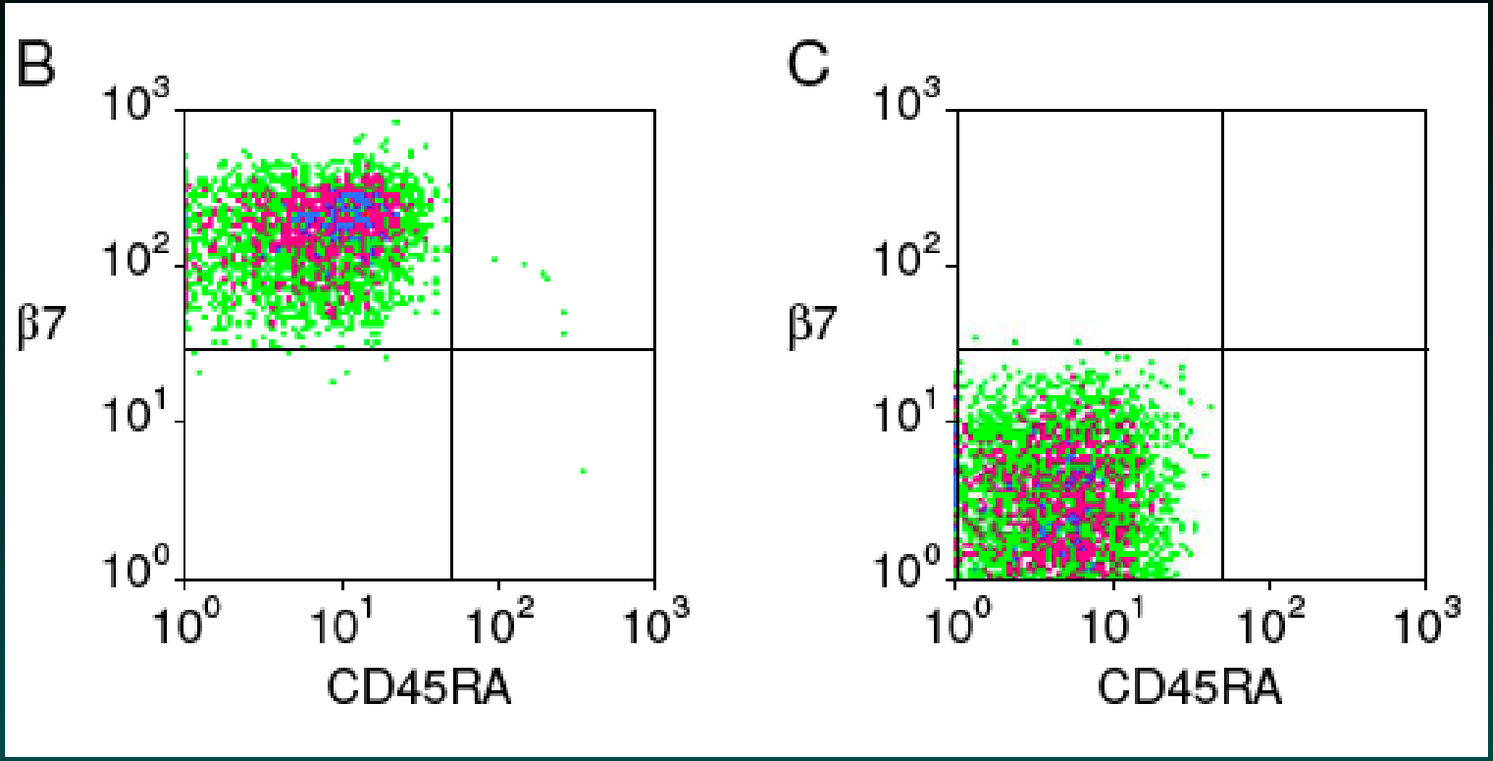
Extract CD4+ T cells, and derive enriched subpopulations that are $\beta 7$+ and $\beta 7$-. Cell subpopulations were obtained using flow cytometry.

Initially, cells are sorted by their levels of $\beta 7$ and CD45RA. High levels of CD45RA are not as interesting here, as their adhesion targets are already known. We want to focus on $\beta 7$ and see if we see separations there.

Cells Before Filtering



Cells After Filtering



After purification, the distributions are separated into our target groups.

Samples are Paired!

Extraction was done with samples from 9 individuals, so there is a natural data pairing.

Given the pairing, individual arrays were used to contrast the two by hybridizing β7+ in one channel and β7- in the other.

In all, 27 arrays were run, including at least 2 for each patient in a dye-swap arrangement.

The actual data is available from the Gene Expression Omnibus (GEO) maintained by the NCBI, with accession number GSE1039 (We'll come back to this).

Stuff Inferrable from GEO

sample, channel 1 (635nm), channel 2 (532nm), Patient ID, Gender (or is ch1 Cy3 and ch2 Cy5?)

GSM16665 - + 001 F GPL976 Hs_004_187_2
GSM16675 + - 001 F GPL976 Hs_004_186_2
GSM16679 - + 006 F GPL976 Hs_004_235
GSM16680 - + 009 F GPL976 Hs_004_189_1
GSM16681 + - 009 F GPL976 Hs_004_188
GSM16685 - + 001 F GPL978 6Hs.094
GSM16686 - + 001 F GPL978 6Hs.195.1 **
GSM16687 + - 003 F GPL978 6Hs.168 **

and so on. The ones with asterisks are contained in the subset we will look at today.

More on Methods

No data from patients 2 and 5.

The arrays used 70-mer oligos from Operon; there were 23184 spots on the arrays. Two different chip platforms were used when the experiment was run; these are available from GEO as

GPL976 UCSF 4Hs Human v.2 Oligo Array
GPL978 UCSF 6Hs Human v.2 Oligo Array

The RNA was subjected to 2 rounds of amplification using kits from Ambion.

All of the arrays were quantified using Axon's GenePix software, so we have gpr quantification files. The TIFF files are also available for download.

More on Methods, and our Subset

What other information would we like to have?
Run date? (scan date is available; this should be close)
Date of blood draw? (this is given in the TargetBeta7.txt file)
Gene information? (some of this is here)
Patient age? (this was there)

The data used here involves a subset of 6 arrays from this experiment. All 6 were of a single platform type, and had a common layout format.

Why were these 6 chosen?

Getting the Data

Let's get the 6 gpr files, and some TargetInfo and SpotInfo files

http://www.bioconductor.org/workshops/2005/BioC2005/labs/lab01/Data/integrinbeta7.zip

This zip file includes 6 gpr files, and a text file, TargetBeta7.txt, that contains sample information (eg, phenoData information). Eg:

```
FileNames         Subject ID #    Cy3      Cy5
6Hs.195.1.gpr     001             b7 -     b7 +
Hyb buffer        Hyb Temp (deg C)Hyb Time (h)
Ambion Hyb Slide        55        40
Date of Blood Draw      Amplification
2002.10.11              R2 aRNA
```

Using R

The first step is simply to load a whole bunch of packages:

```
> library("marray");
> library("mclust");
> library("convert");
> library("arrayQuality");
> library("colorspace");
> library("grid");
> library("hexbin");
```

Getting the Sample Info

```
> TargetInfo <- read.marrayInfo("TargetBeta7.txt")
> TargetInfo
An object of class "marrayInfo"
@maLabels
[1] "6Hs.195.1.gpr" "6Hs.168.gpr"   "6Hs.166.gpr"   "
[5] "6Hs.194.gpr"   "6Hs.243.1.gpr"

@maInfo
      FileNames Subject ID #  Cy3  Cy5        Hyb buff
1 6Hs.195.1.gpr            1 b7 - b7 + Ambion Hyb Sli
2   6Hs.168.gpr            3 b7 + b7 - Ambion Hyb Sli
  Hyb Time (h) Date of Blood Draw Amplification  Slid
1           40          2002.10.11       R2 aRNA  Amino
2           40          2003.01.16       R2 aRNA  Amino
```

Getting the Numerical Info

Grab the data from the gpr files:

```
mraw <- read.GenePix(targets = TargetInfo);

# Note: this works on my PC. On my Mac laptop,
# I get the following error messages:

> mraw <- read.GenePix(targets = TargetInfo)
Error in if (skip > 0) readLines(file, skip) :
missing value where TRUE/FALSE needed
In addition: Warning messages:
1: input string 32 is invalid in this locale in:
  grep(pattern, x, ignore.case, extended, value, fixe
2: input string 32 is invalid in this locale in:
  grep(pattern, x, ignore.case, extended, value, fixe
```

What Can be Inferred?

So, what does our marrayRaw object contain at this point?

Let's take a look at the individual slots here.

```
> slotNames(mraw)
[1] "maRf"       "maGf"       "maRb"       "maGb"
[6] "maLayout"   "maGnames"   "maTargets" "maNotes"
```

Of these, the first 5 are the basic quantification information, extracted from the gpr files. All of them are 23184 by 6 in size. The others are the associated layout and annotation files. Let's extract these and find out a bit more about them.

Summary, Part 1 – Layout

```
> summary(mraw)
Pre-normalization intensity data:
  Object of class marrayRaw.

Number of arrays:  6 arrays.

A) Layout of spots on the array:
Array layout:  Object of class marrayLayout.

Total number of spots: 23184
Dimensions of grid matrix: 12 rows by 4 cols
Dimensions of spot matrices: 23 rows by 21 cols

Currently working with a subset of 23184spots.
```

More Layout

```
Control spots:
There are   5 types of controls :

   Buffer    Empty Negative Positive   probes
        3     1328      225      204    21424
```

Notes on layout:

The layout can be inferred from the gpr files! This is not too suprising, as every row of a gpr file contains entries for grid row, grid col, spot row, and spot col. As a side note, what is the precise order?

Layout Ordering

```
> zedL <- mraw@maLayout
> zedLSC <- maSpotCol(zedL); zedLSR <- maSpotRow(zedL
> zedLGR <- maGridRow(zedL); zedLGC <- maGridCol(zedL
> zedLcoords <- cbind(zedLGR,zedLGC,zedLSR,zedLSC)

> zedLcoords[1:25,]
      zedLGR zedLGC zedLSR zedLSC
 [1,]      1      1      1      1
 [2,]      1      1      1      2
 [3,]      1      1      1      3
...
[20,]      1      1      1     20
[21,]      1      1      1     21
[22,]      1      1      2      1
```

Summary Part 2 – Sample Info

```
B) Samples hybridized to the array:
Object of class marrayInfo.

        maLabels     FileNames SubjectID  Cy3  Cy5 Daf
1 6Hs.195.1.gpr 6Hs.195.1.gpr         1 b7 - b7 +
2   6Hs.168.gpr   6Hs.168.gpr         3 b7 + b7 -
..
  Date of Scan
1   2003.07.25
2   2003.08.07
..
```

Since we supplied the marrayInfo file in the call to read.GenePix, this is imported from there.

Summary Part 3 – Array Summaries

```
C) Summary statistics for log-ratio distribution:
                  Min. 1st Qu. Median  Mean 3rd Qu. Max
6Hs.195.1.gpr -6.13   -1.00  -0.52 -0.50   -0.08 5.95
6Hs.168.gpr   -7.08   -0.80  -0.21 -0.23    0.34 5.19
6Hs.166.gpr   -7.07   -1.25  -0.64 -0.62   -0.02 6.15
6Hs.187.1.gpr -9.81   -0.92  -0.60 -0.55   -0.25 5.00
6Hs.194.gpr   -5.93    0.00   0.44  0.53    0.90 7.74
6Hs.243.1.gpr -6.38   -1.13  -0.69 -0.64   -0.21 7.05
```

Log ratios – what direction is the default? Cy3/Cy5? Cy5/Cy3?
(the latter, according to documentation)

Summary Part 4 – Notes

D) Notes on intensity data:
GenePix Data

Ok, that dealt with most of the microarray structure itself.

What happens if we ask about the gene names? This is what we really want, so that we can understand the biology.

Annotation

```
> mraw@maGnames[1:2,]
An object of class "marrayInfo"
@maLabels
[1] "H200000297" "H200000303"
@maInfo
                      ID
H200000297 H200000297
                    Name
H200000297 OVGP1 - Oviductal glycoprotein 1, 120kD (m
@maNotes
[1] ""
```

again, these are read in from the gpr files. The first column here, the maLabels, is the Operon-supplied identifier for that specific oligo, and as such it should be unique.

Getting the Data: TMTOWTDI

Assembling an marrayRaw object need not be hard.

So, what if you're working with a Mac?

This marrayRaw object and a few other things are available as a package from BioConductor called "beta7". I had to run a search at the top level of BioConductor to find this; it is part of the "Data" page associated with the monograph. I downloaded the gzipped tar (.tar.gz) file and did an install from local source.

http://www.bioconductor.org/docs/mogr/data

library("beta7"); data(beta7);

loads an marrayRaw object (called beta7) with info on the 6 selected arrays.

How was Data Reported?

Table 1: Gene transcripts with higher expression in β7⁺ versus β7⁻ CD4⁺ CD45RA⁻ T helper cells*

Symbol	Name	Accession	Fold Difference	P value
CCR9	chemokine (C-C motif) receptor 9	NM_031200	+3.0	< 0.01
CCL5	chemokine (C-C motif) ligand 5	NM_002985	+2.4	< 0.01
RAM2	transcription factor RAM2	NM_018719	+2.2	< 0.01
LRRN3	leucine rich repeat neuronal 3	AL442092	+2.1	< 0.01
GFI1	growth factor independent 1	NM_005263	+1.8	< 0.01
ITGA4	integrin, alpha 4 (CD49D)	NM_000885	+1.7	< 0.01
CD1C	CD1C antigen, c polypeptide	NM_001765	+1.7	< 0.01
KLRB1	killer cell lectin-like receptor subfamily B, member 1	NM_002258	+1.7	< 0.01
LAIR1	leukocyte-associated Ig-like receptor 1	NM_002287	+1.7	< 0.01
RRM2	ribonucleotide reductase M2 polypeptide	NM_001034	+1.6	< 0.01
--	Homo sapiens cDNA FLJ32290 fis, clone PROST2000463	AK056852	+1.6	< 0.01
HHL	expressed in hematopoietic cells, heart, liver	NM_014857	+1.6	0.02
IL18RAP	interleukin 18 receptor accessory protein	NM_003853	+1.6	< 0.01
SREBF1	sterol regulatory element binding transcription factor 1	NM_004176	+1.6	< 0.01
KLRG1	killer cell lectin-like receptor subfamily G, member 1	NM_005810	+1.5	< 0.01
LGALS2	lectin, galactoside-binding, soluble, 2 (galectin 2)	NM_006498	+1.5	0.01

* Includes all transcripts with fold difference ≥+1.5 and adjusted P < 0.05. Positive fold difference values indicate higher expression on β7⁺ cells.

There are some unique identifiers here!

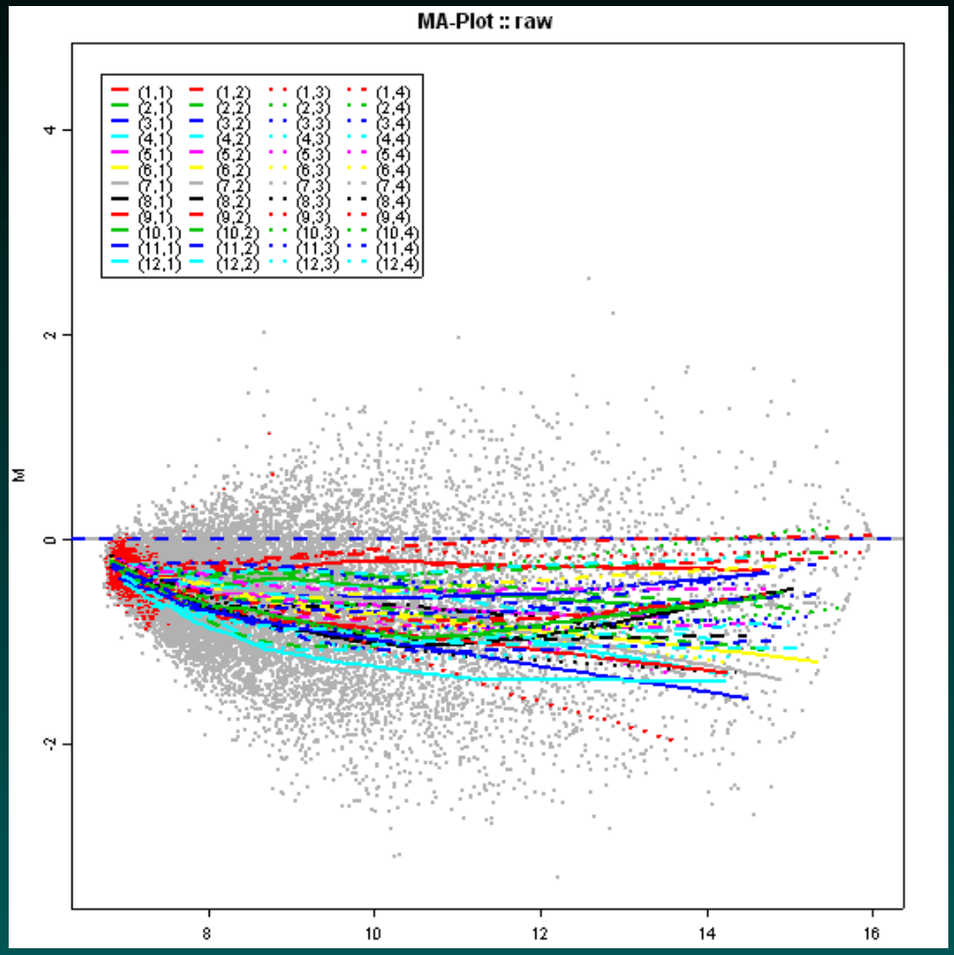
Checking the Data

Ok, now we have the raw data. What do we want to try next? Well, checking array quality would be nice.

```
> maQualityPlots(mraw); # again, works on PC only
save as diagPlot..6Hs.195.1.png
save as diagPlot..6Hs.168.png
save as diagPlot..6Hs.166.png
save as diagPlot..6Hs.187.1.png
save as diagPlot..6Hs.194.png
save as diagPlot..6Hs.243.1.png
```

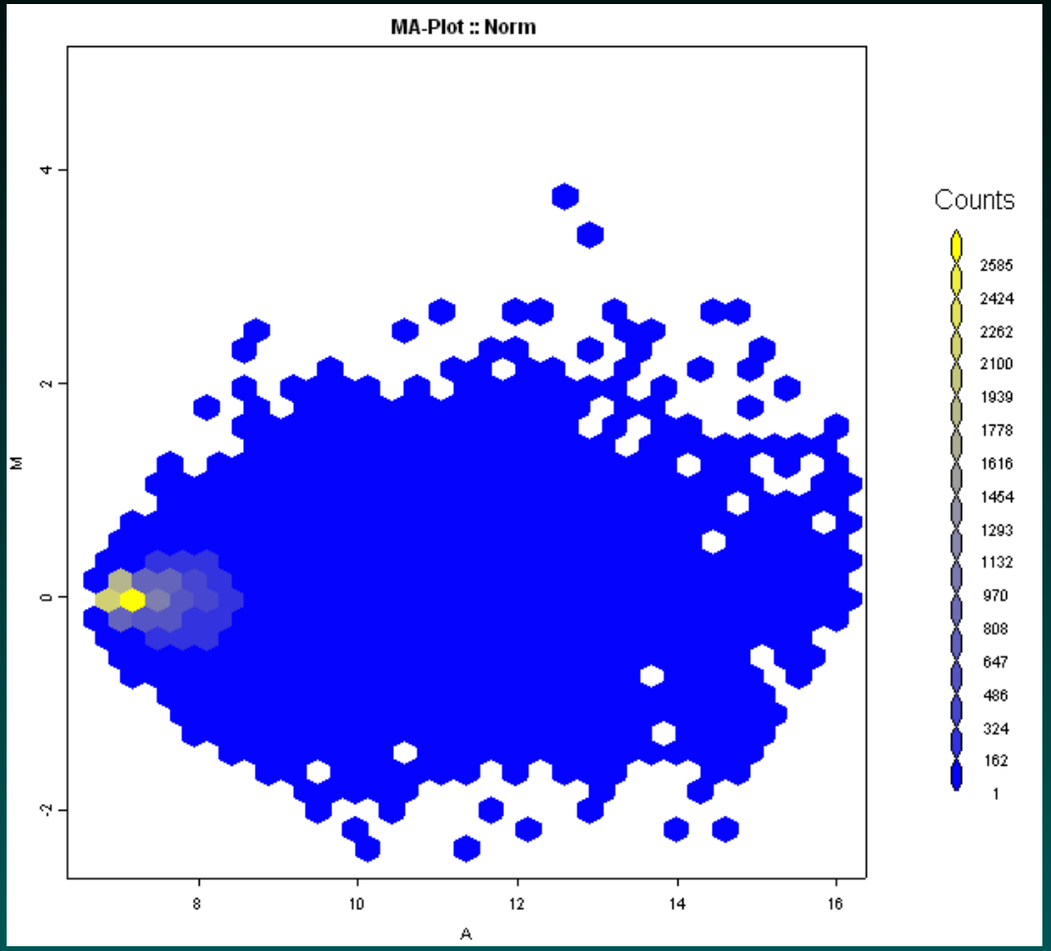
what does this produce? One large png file for each array. This plot has 8 panels...

Panel (a)



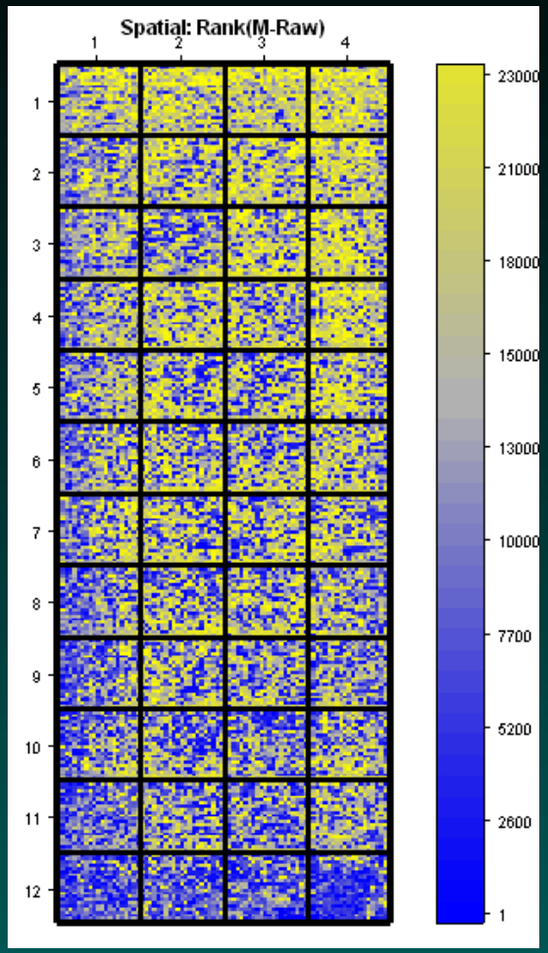
(a) an MA-plot for the raw data, with loess traces for each pin

Panel (b)



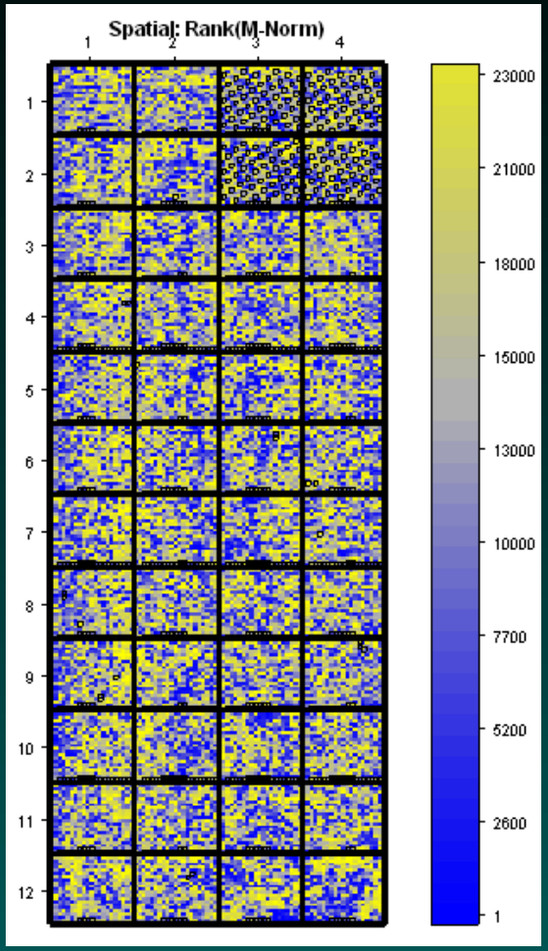
(b) an MA-plot for the data after print-tip loess normalization, displayed using hexbin.

Panel (c)



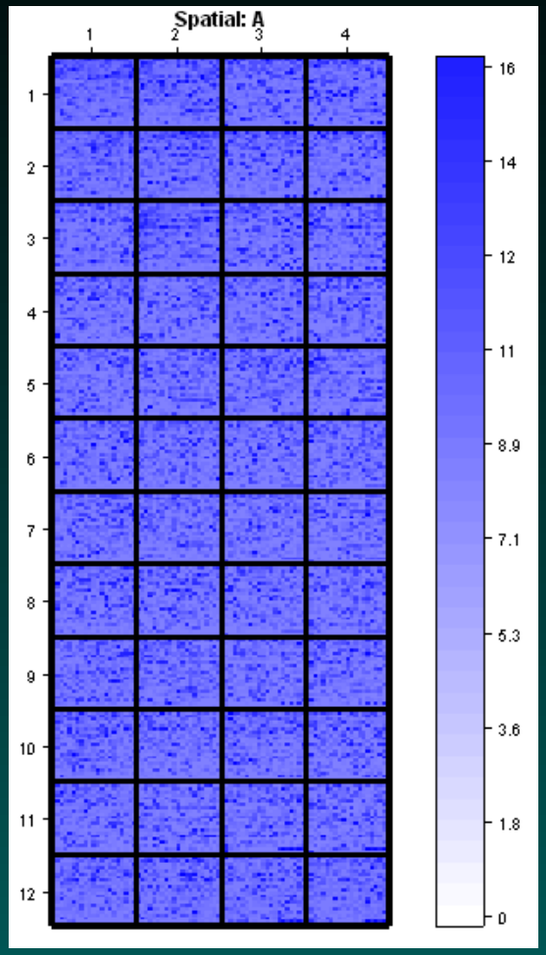
(c) a spatial plot of ranks of the M-Raw differences

Panel (d)



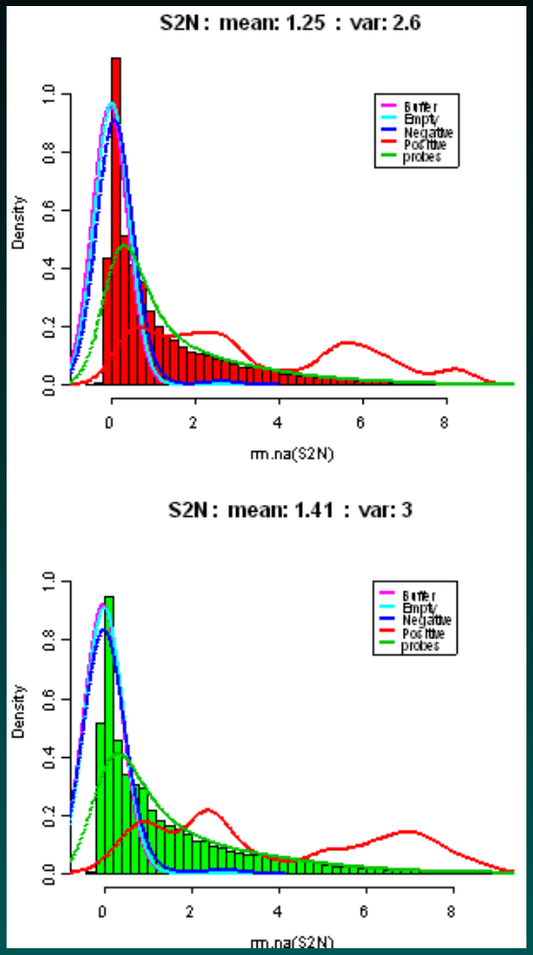
(d) a spatial plot of ranks of the M-Norm differences, with outliers flagged

Panel (e)



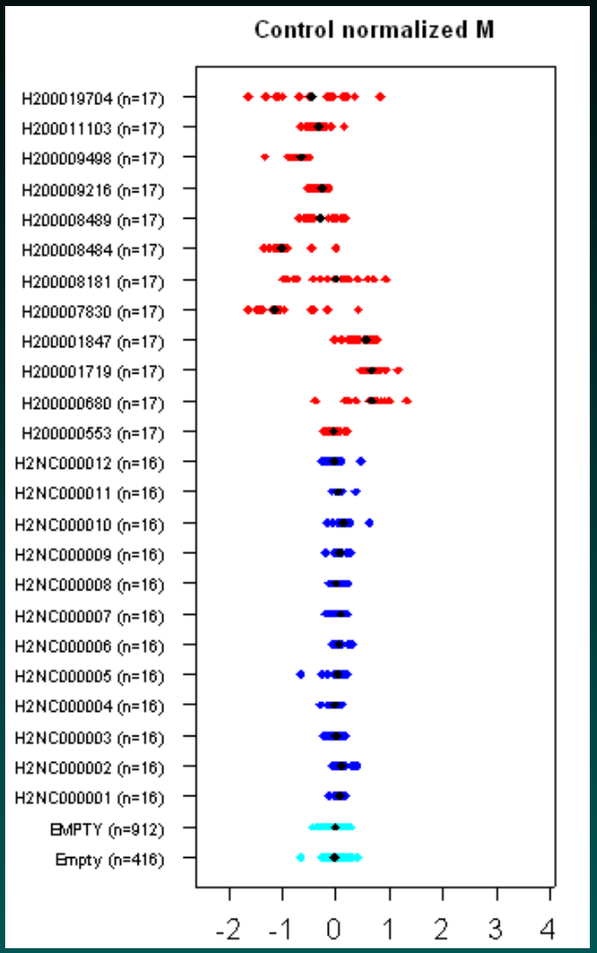
(e) a spatial plot of the A values

Panel (f)



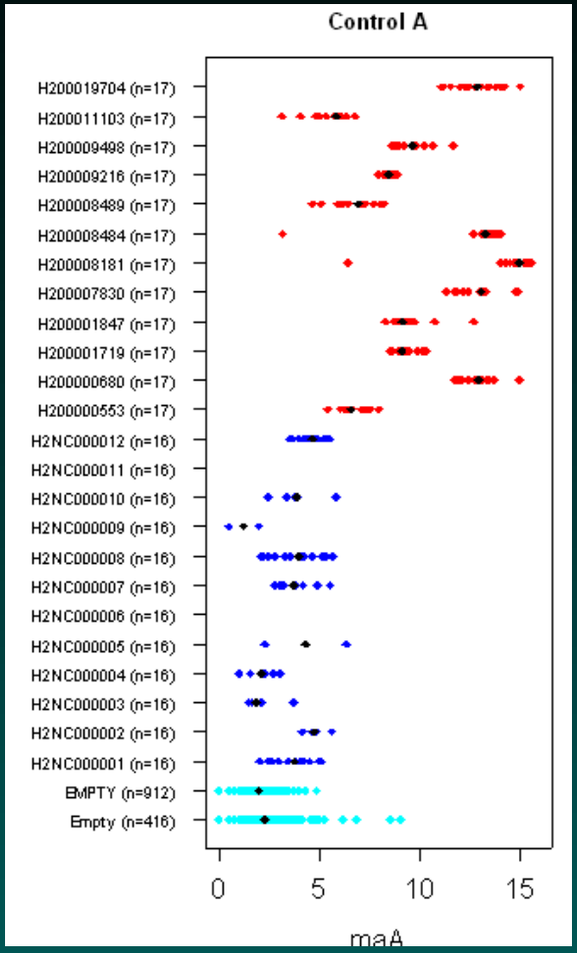
(f) signal to noise distribution plots for each channel (presumably assessed on the raw data)

Panel (g)



(g) M distributions for replicated controls using the normalized values

Panel (h)



(h) A distributions for replicated controls using the normalized values

What next?

Ok, given that the arrays look ok, we'd like to do some numerical contrasts. What needs to be done before we do this?

What next?

Ok, given that the arrays look ok, we'd like to do some numerical contrasts. What needs to be done before we do this?

Go from an marrayRaw object to an marrayNorm object.

```
> normdata <- maNorm(mraw);
```

by default, this will invoke print-tip loess as the processing method.

Exporting the Data

```
write.marray(normdata);
```

This will create a file “maRawResults.xls”, even though the normalized data was used. This will give grid R,C, spot R,C, the spot ID, the gene name, and the associated log ratio values. It presumes that we know which direction the ratios are taken in (it’s Cy5/Cy3).

Using the Data Further

```
library("convert");
mdata <- as(normdata,"exprSet");
```

This would seem to coerce our marrayNorm object into an exprSet, which we can then act upon to get more information. This is partially correct.

The gene names are not retained or passed, so keeping track of the annotation must be done by index value or attached separately.

How was the Data Analyzed?

According to the methods, they worked just with the foreground measurements; no background was subtracted.

Print-tip loess was used to normalize the array data, and log ratios were computed.

Differentially expressed genes were estimated using a linear model (and the limma package). The model:

$$Y_{ij} = \mu + A_i + \epsilon_{ij}$$

The individual (b7+/b7-) log ratio values for each array are expressed in terms of an overall level, a patient effect, and a chip effect. The patient effect lets them deal with replicates intelligently.

More Analysis

For each gene, a “moderated t-test” was performed using an empirical Bayes approach, pooling information about the variance to make the results more stable.

The genes had to be significant at a 0.01 level after a Bonferroni correction, and the mean fold change had to be more than 1.5.

What Other Info was Provided?

Together with the paper, and the data posted to GEO (the layouts of the arrays used, the gpr files, and more information about what the genes are), there was also a supplementary information file giving a MIAME-compliant list of information.

This list was important, as it specified which samples were labeled with Cy5, and which with Cy3. What is recorded in GEO is simply “Channel 1” and “Channel 2”.