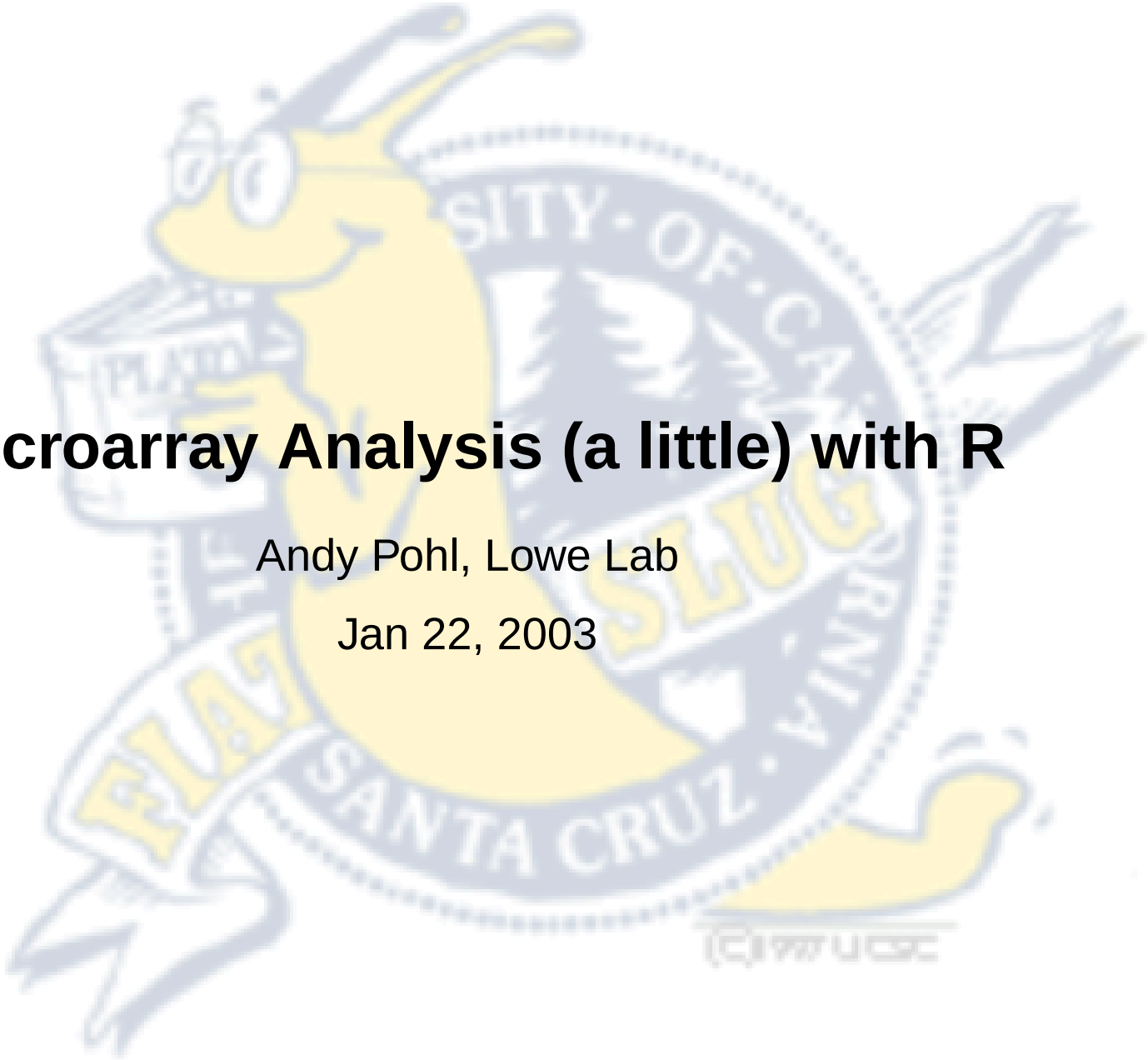


Microarray Analysis (a little) with R

Andy Pohl, Lowe Lab

Jan 22, 2003

Overview

1. Basic Analysis Strategy
2. R
3. An R package for microarrays: Bioconductor and using the marray packages.
4. Bioconductor stuff I don't have time to cover, but I'll briefly introduce.
5. A short demonstration of R and Bioconductor on some Lowe Lab data.

These slides are available from:

http://lowelab.ucsc.edu/andy/biocond-intro.pdf

Analysis

- The main goal of analysis is to look for genes expressed similarly across a range of conditions in a *guilt by association* way.
- The big problems include the obligatory signal to noise issue of microarrays, and the problem of choosing a method of clustering, etc.
- The basic steps of analysis:
  1. Experiment (Raw data)
  2. Image Analysis (GenePix)
  3. Pre-normalization analysis
     - Background analysis and see how different types of spots e.g. controls compared to the rest.
     - Spatial intensity analysis (by sector).
     - Analyze pin and plate variations.

Analysis

4. Normalization and filtering
   - Scale normalization
   - Intensity dependent location normalization (e.g. using controls).
5. Post-normalization stuff:
   - Clustering
   - Hypotheses testing
   - Classification

- Because there is so much data, it's necessary to visualize the analysis with a bunch of plots.
- R is a good tool for making plots.

What is R?

- R was created in 1996 as an open-source alternative to the S-PLUS statistics program (commercial).
- Since then, many people around the world have contributed add-on packages for R to accomplish various statistical or visual tasks. Usually these contributors are professors or grad students promoting something they publish.
- R has a command-based interface (like a UNIX terminal prompt), but different packages sometimes include GUI widgets, usually to assist input.

Advantages of R

- It will definately be popular for many many more years.
- It's what the world's top biostatisticians are using, so all the obvious advantages are there: reproducibility of analyses, common means of communicating methods, etc.
- Highly satisfying output.
- Many add-on packages like Bioconductor.
- No shortage of manuals, tutorials, etc.
- Available for Windows and Linux. It can also run on Apples with MacOS X (although not natively).

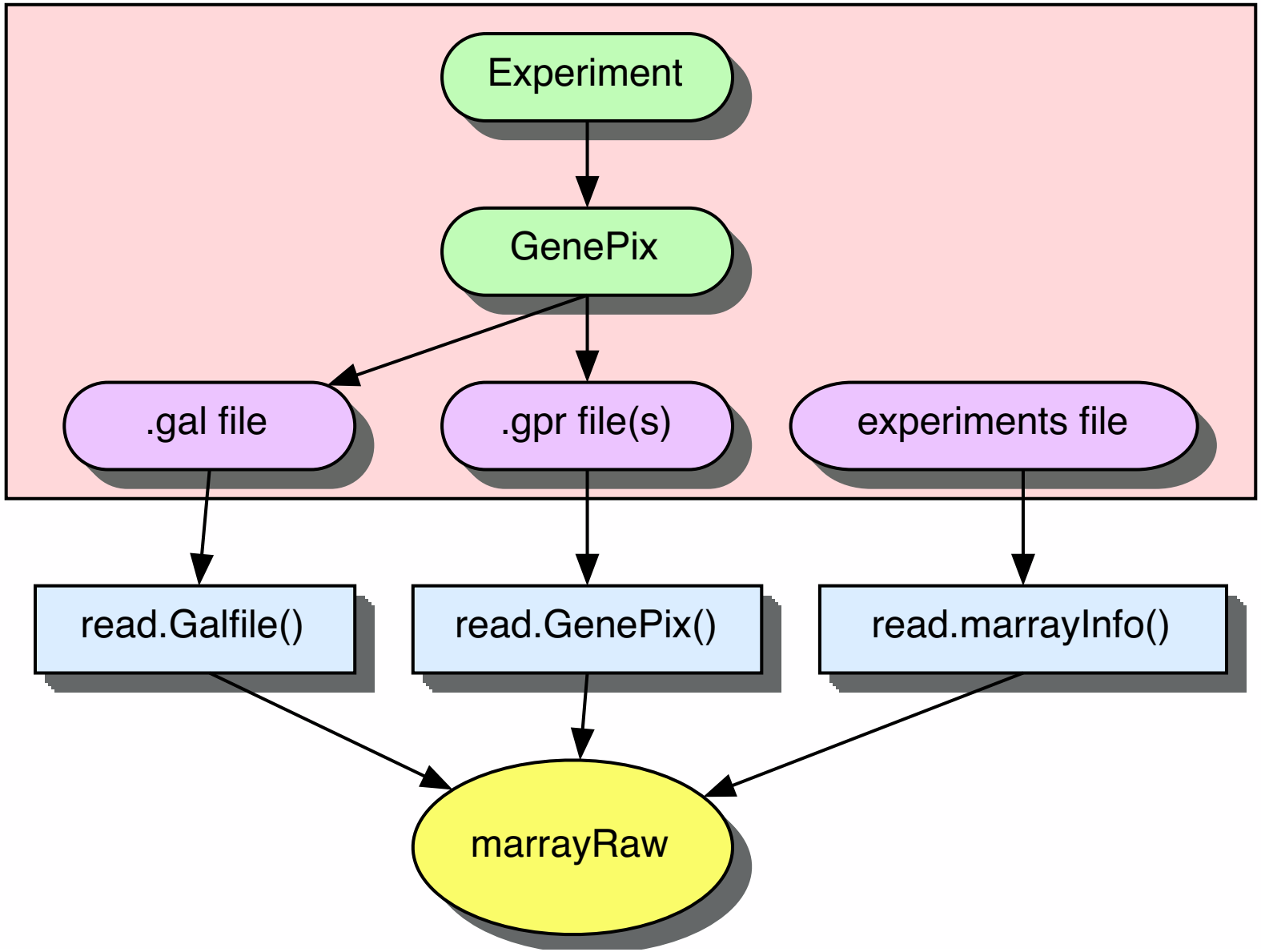
Disadvantages of R

- **NOT** easy to learn. For people who have some background with data structures and programming it's not so bad, but there is still a learning curve. I plan to deal with this problem by creating a web page detailing using Bioconductor from Jeff's point of view.
- Some of the documentation can be confusing, seem incomplete, or expect a high level of expertise with statistics.

Bioconductor

- Bioconductor is a fairly new add-on package for R. It's first release was in March, 2002.
- Headed by Robert Gentleman (Harvard), who is one of the two original creators of R, and still heads the development of that.
- It has routines for both Affymetrix and spotted arrays.
- It tries to handle all the steps of analysis after GenePix, but right now the step that has the most potential for expanding is the post-normalization stuff.

Bioconductor: Inputting data



Bioconductor: After data is inputted

- Now that there is an **marrayRaw** object in R's memory, there are two directions we can go:
  1. Make plots.
  2. Normalize.
- I'll start with making a few plots. There are three basic types of plots included in the **maPlots** portion of Bioconductor:
  1. Image plot: shows things like background intensity across all sectors.
  2. Scatter plot: spot intensity in Cy3 vs. Cy5 foreground, M vs. A, whatever.
  3. Boxplot: mainly for showing pin-intensity variation.

Bioconductor: A few notes

$$
\begin{aligned}
M &= \log_2\left(\frac{R}{G}\right) \\
A &= \log_2\left(\sqrt{RG}\right) = \frac{\log_2 R + \log_2 G}{2}
\end{aligned}
$$

- These equations come up over and over. M is just the familiar log-ratio.
- The “experiments” file just contains information about the labels used in experiments (.gpr files).

Side example of M vs. A

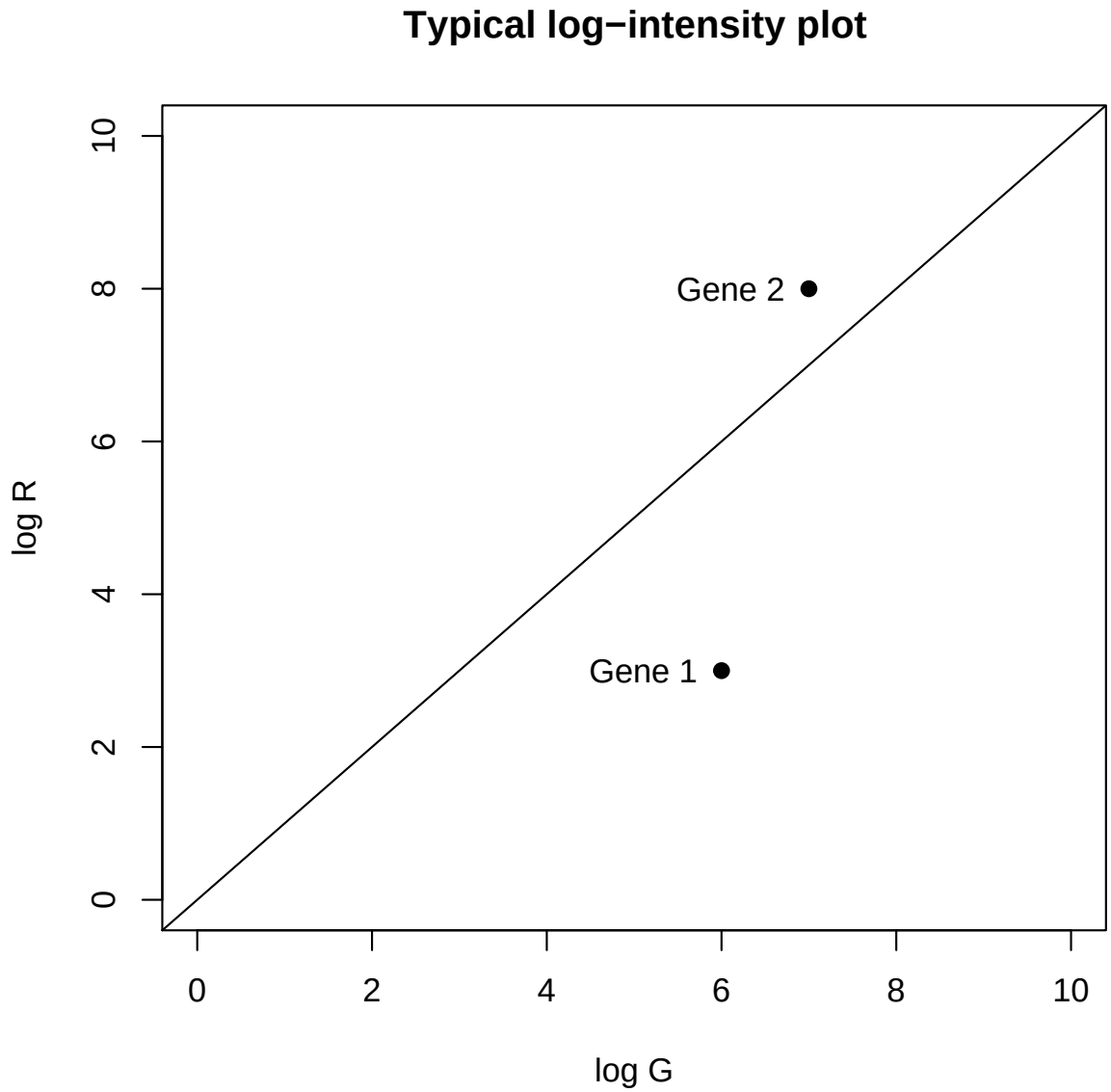
We have two genes. The foreground intensity information follows:

1. Gene 1: Cy5 = 8, Cy3 = 64.
2. Gene 2: Cy5 = 256, Cy3 = 128.

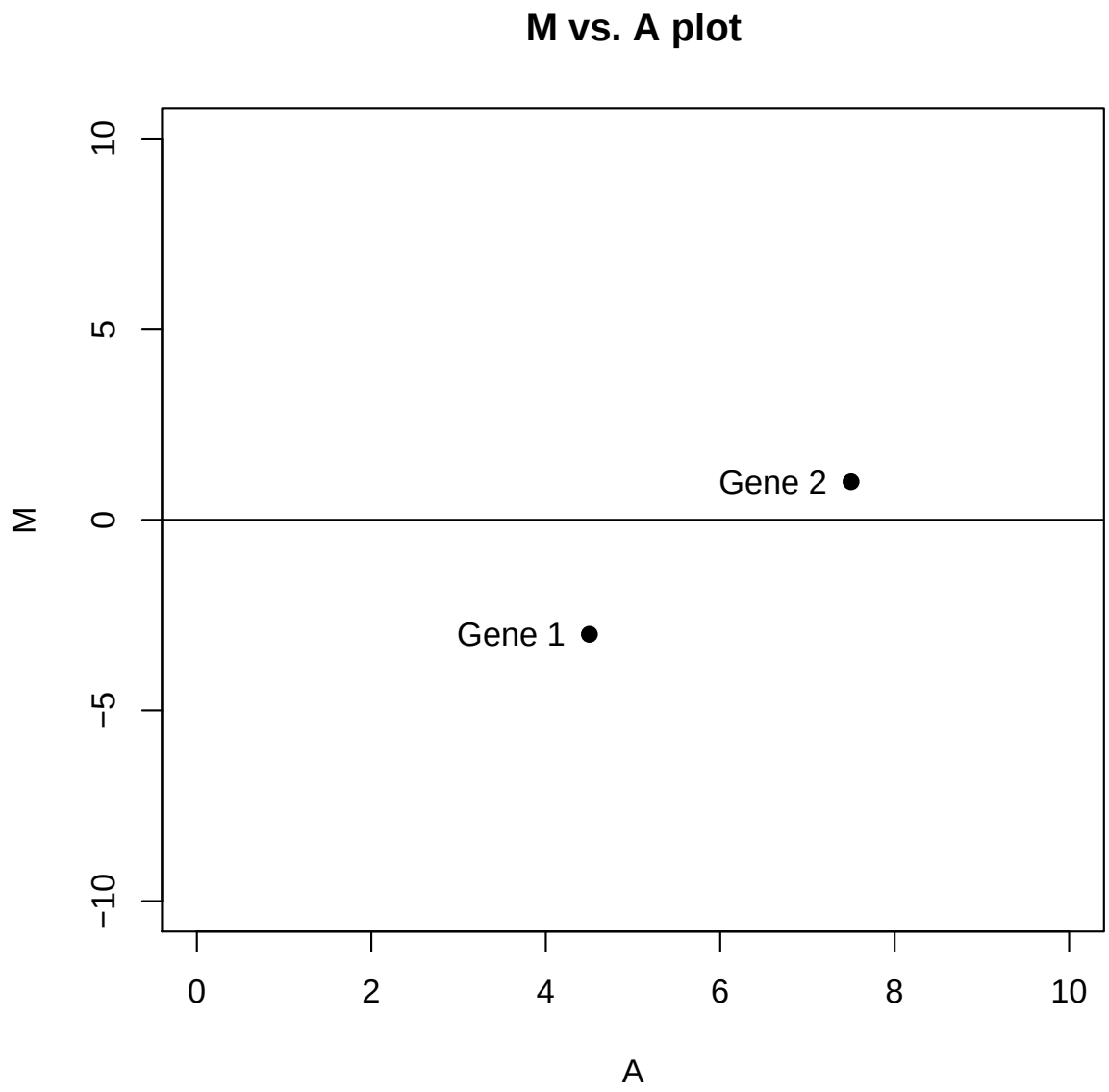
Now the calculations go as follows:

$$
\begin{aligned}
M_1 &= \log_2\left(\frac{8}{64}\right) = 3 - 6 = -3 \\
A_1 &= \log_2\sqrt{8 \times 64} = \frac{\log_2 8 + \log_2 64}{2} = \frac{3+6}{2} = 4.5 \\
M_2 &= \log_2\left(\frac{256}{128}\right) = 8 - 7 = 1 \\
A_2 &= \log_2\sqrt{256 \times 128} = \frac{\log_2 256 + \log_2 128}{2} = \frac{8+7}{2} = 7.5
\end{aligned}
$$

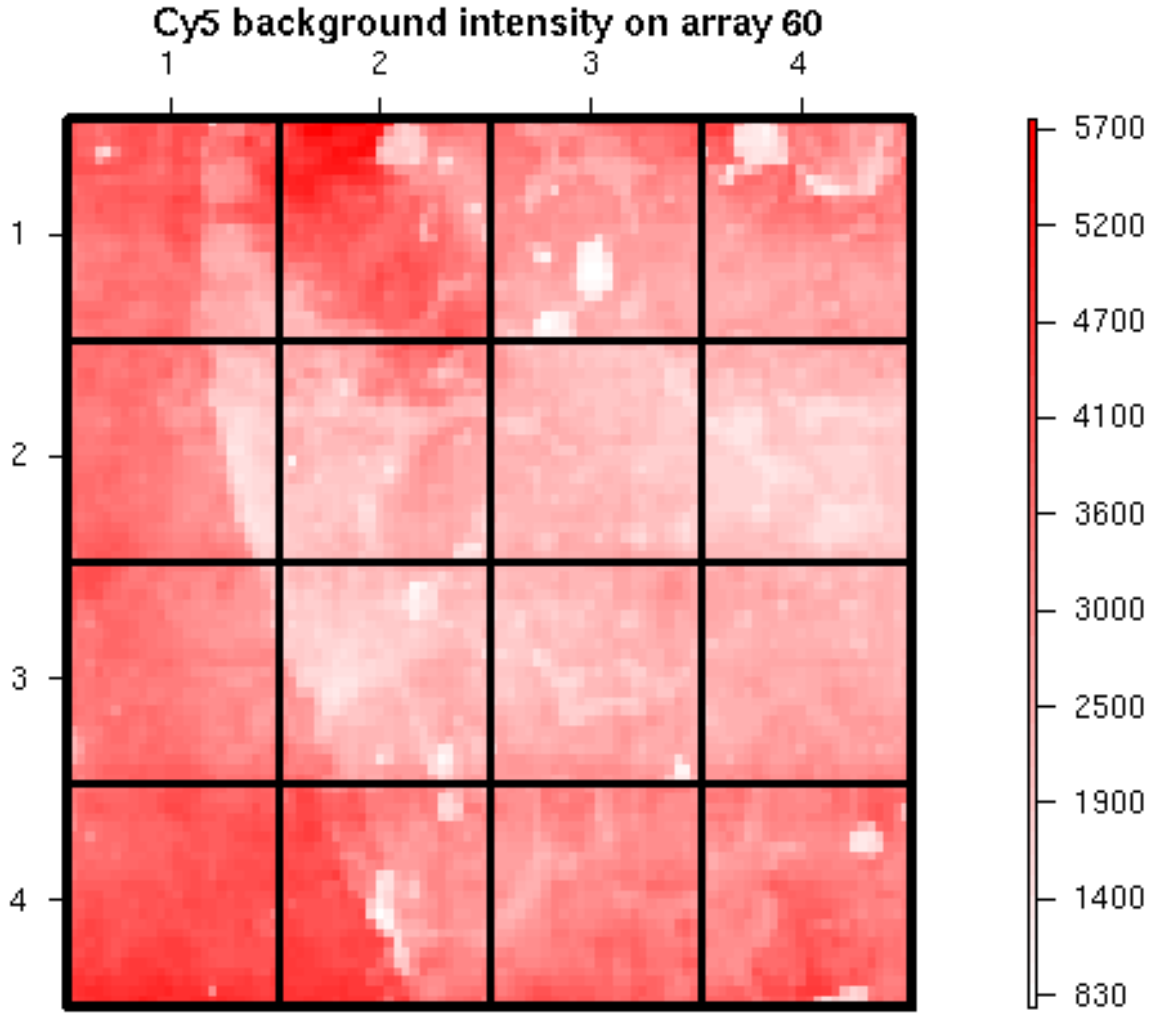
Side example of M vs. A



Side example of M vs. A



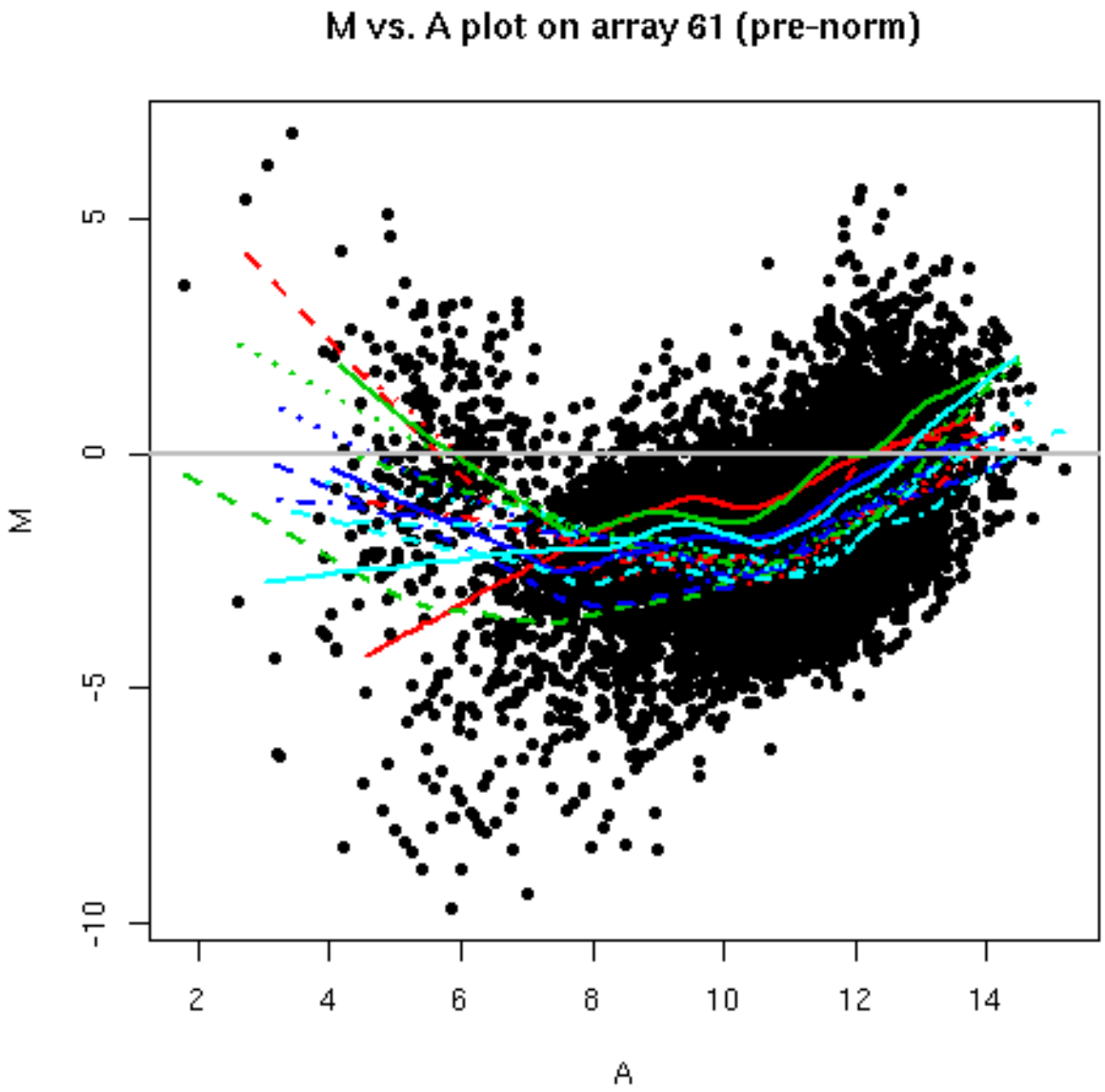
Bioconductor: Example of maImage plot



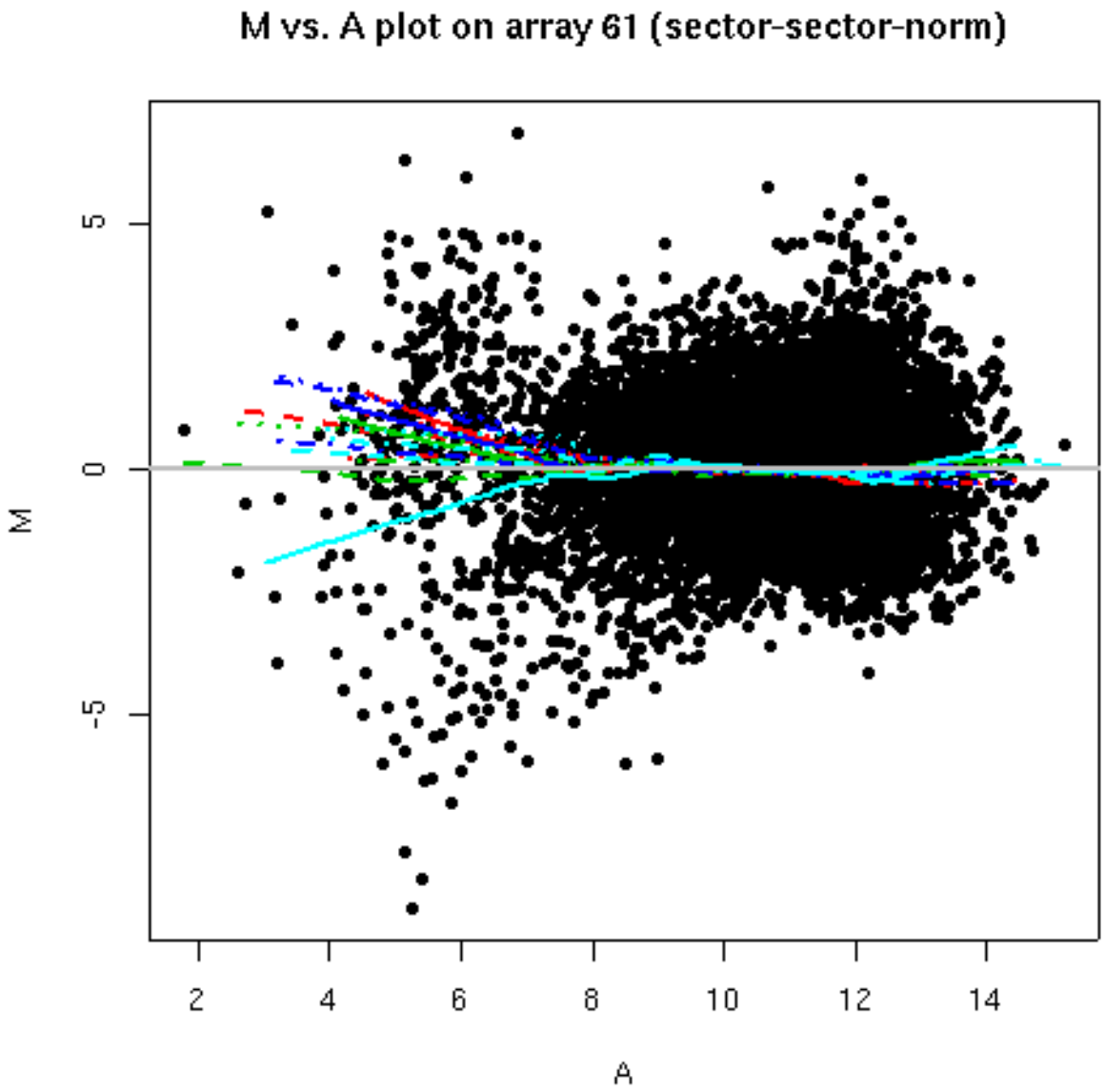
Bioconductor: Normalization

- The types of normalization built-in fall into two different categories:
  1. Location normalization using the local regression function *loess* or using the median:
     - by sector
     - by plate
     - by row-column coordinate
     - by positive and negative control spots
  2. Scale normalization: MAD scaling of log ratios.

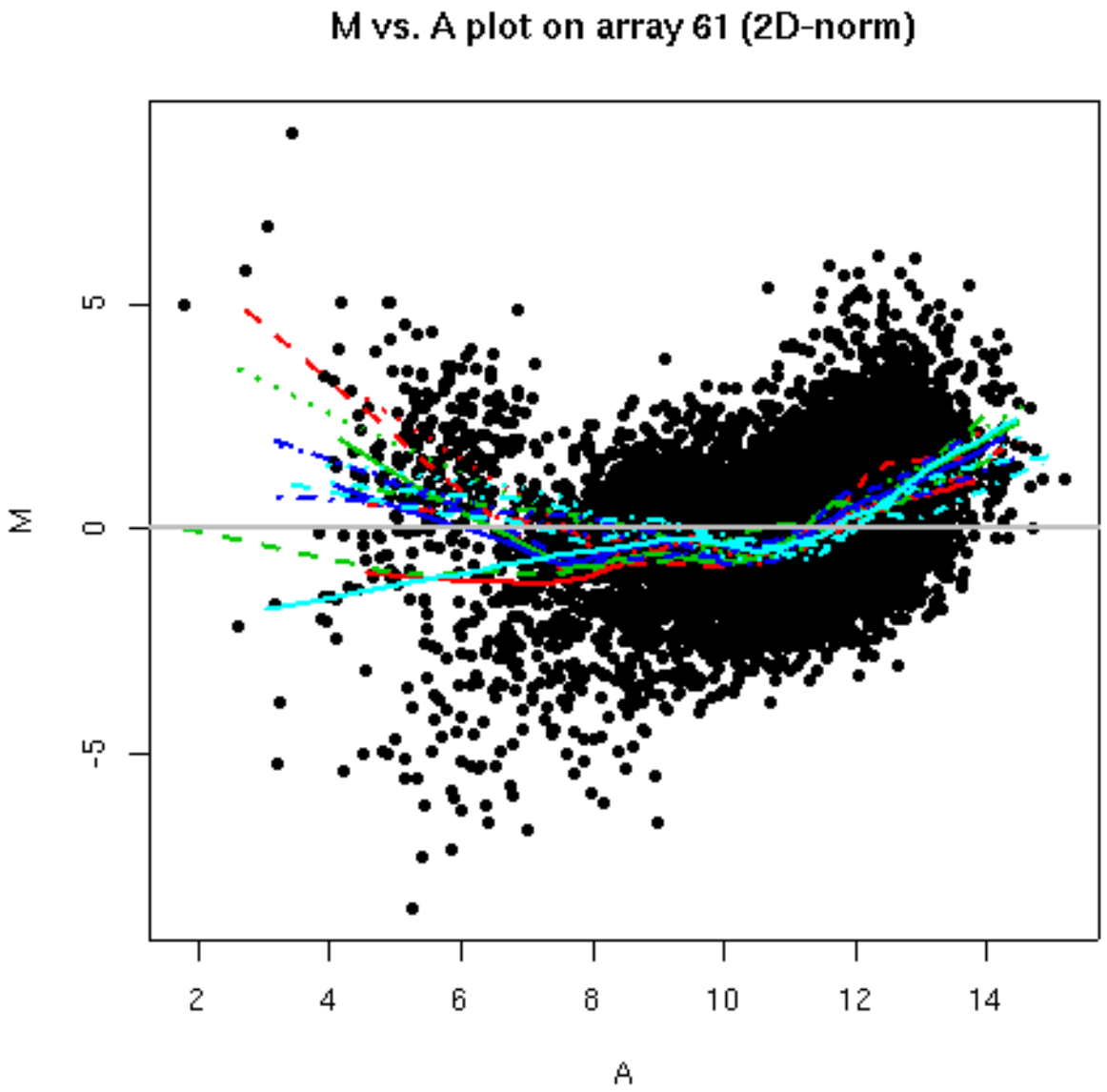
Bioconductor: Scatter plot



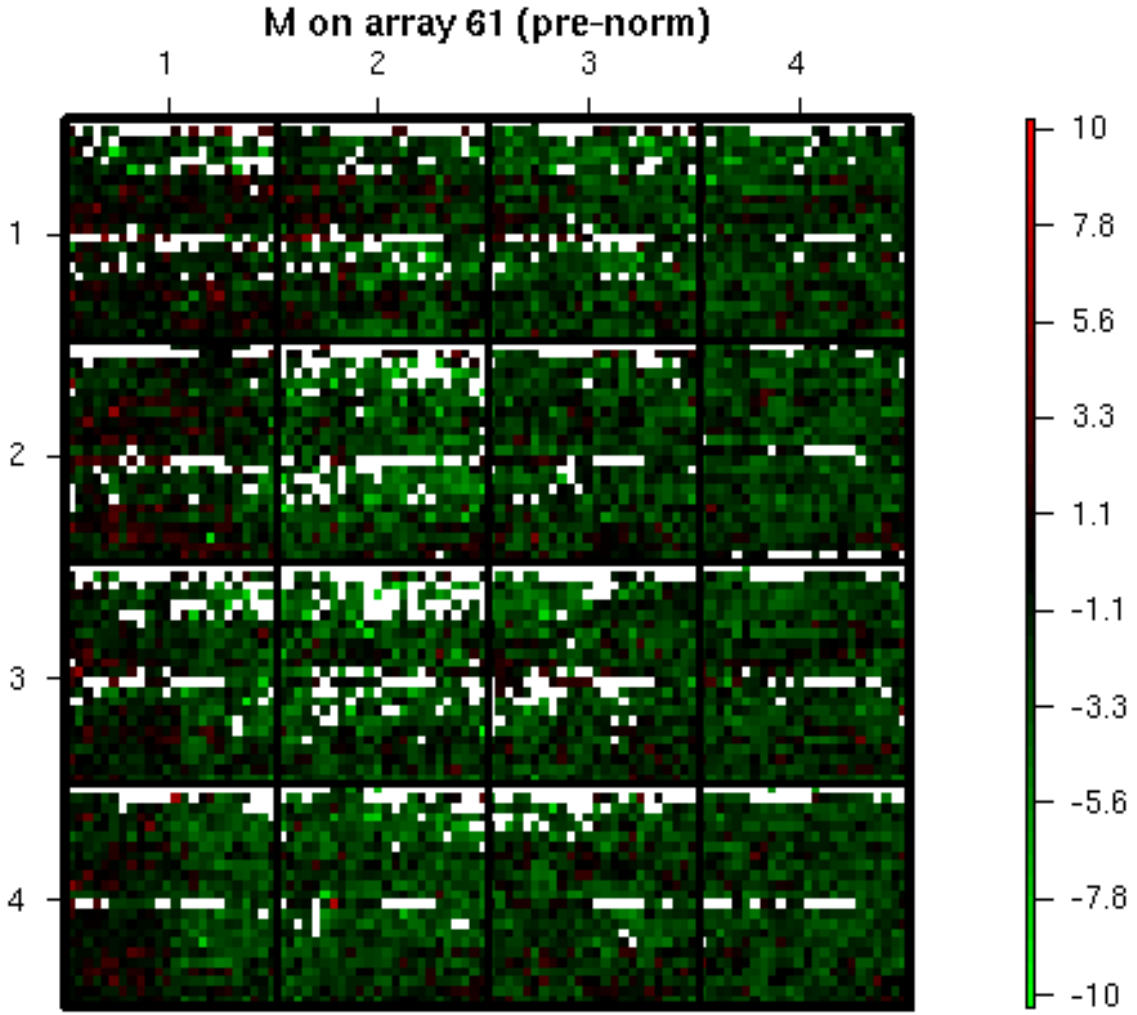
Bioconductor: Scatter plot



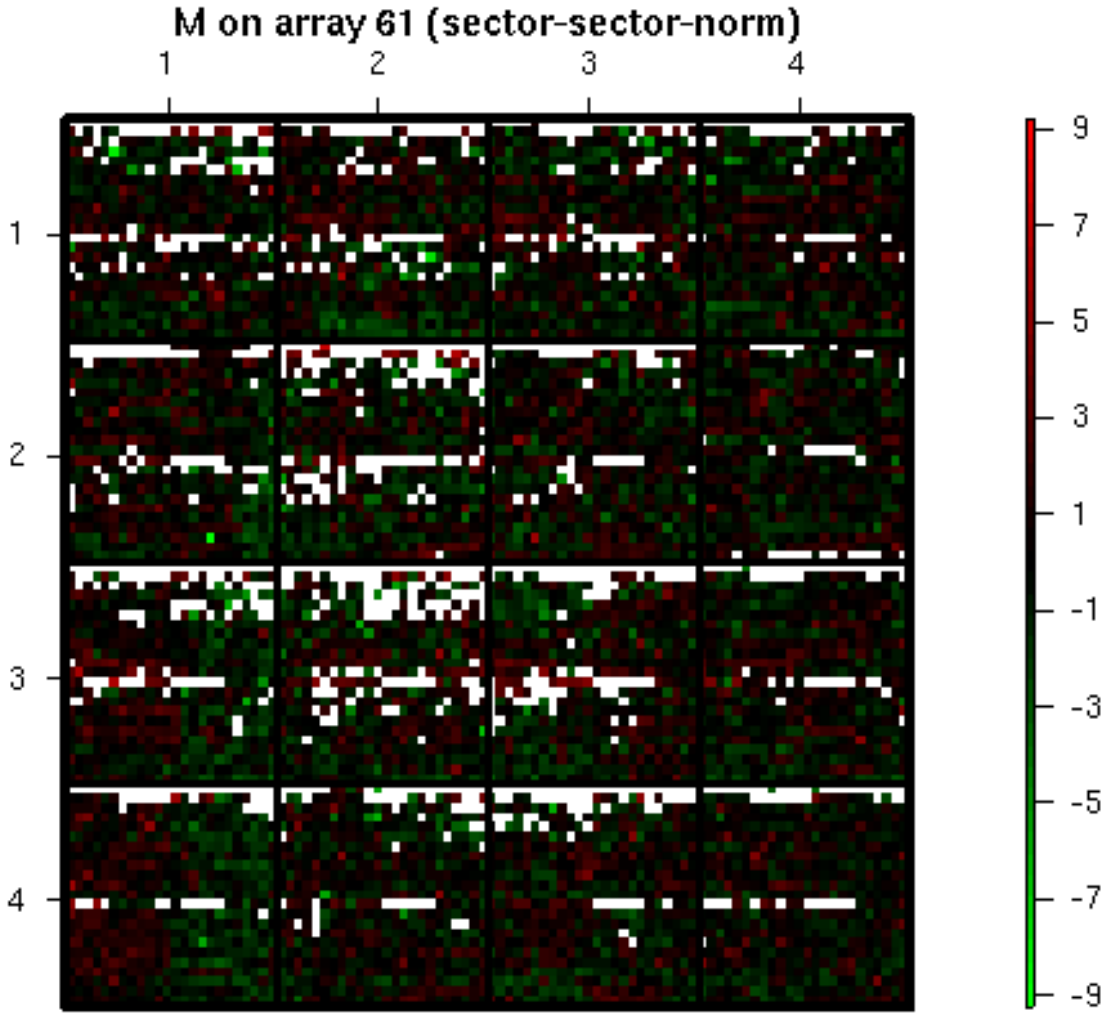
Bioconductor: Scatter plot



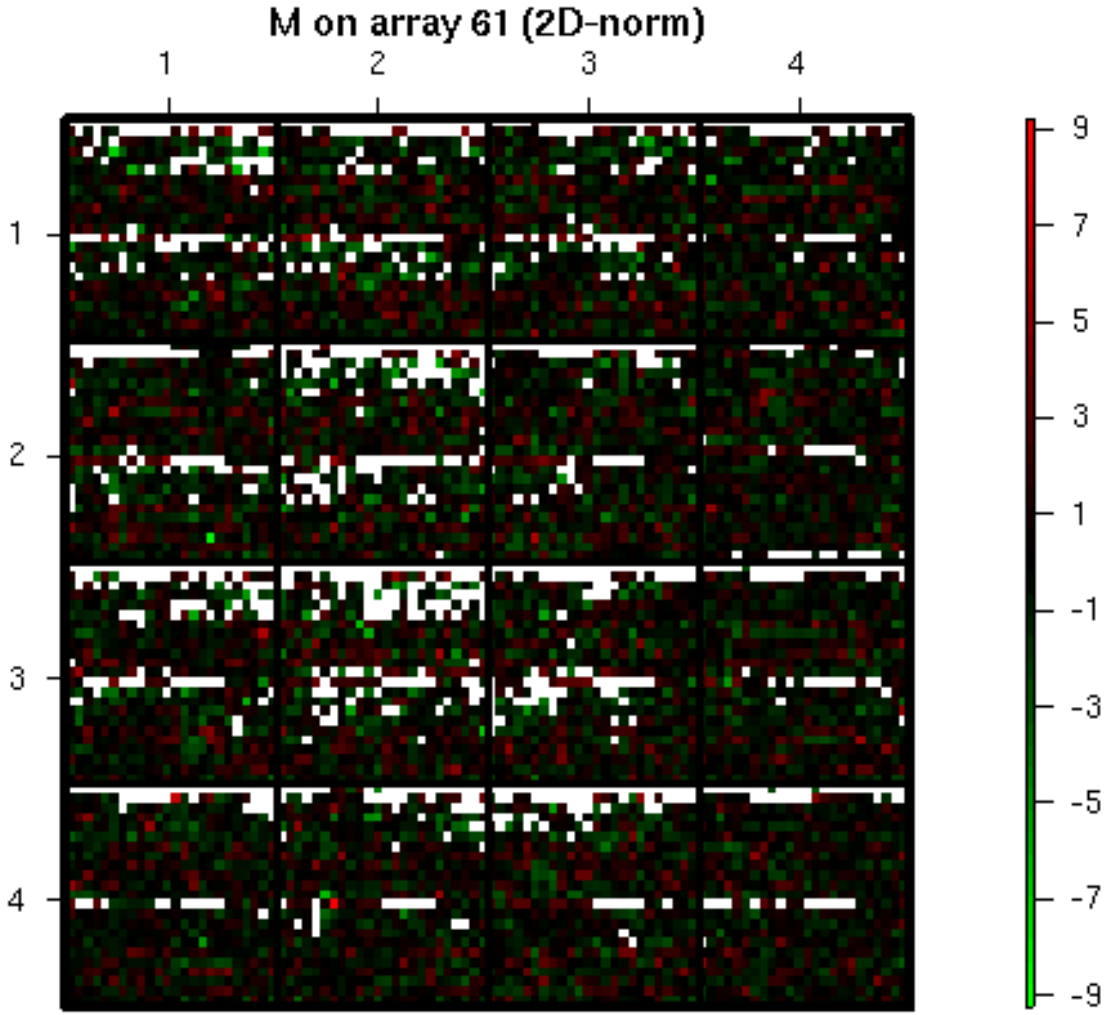
Bioconductor: Image plot of M



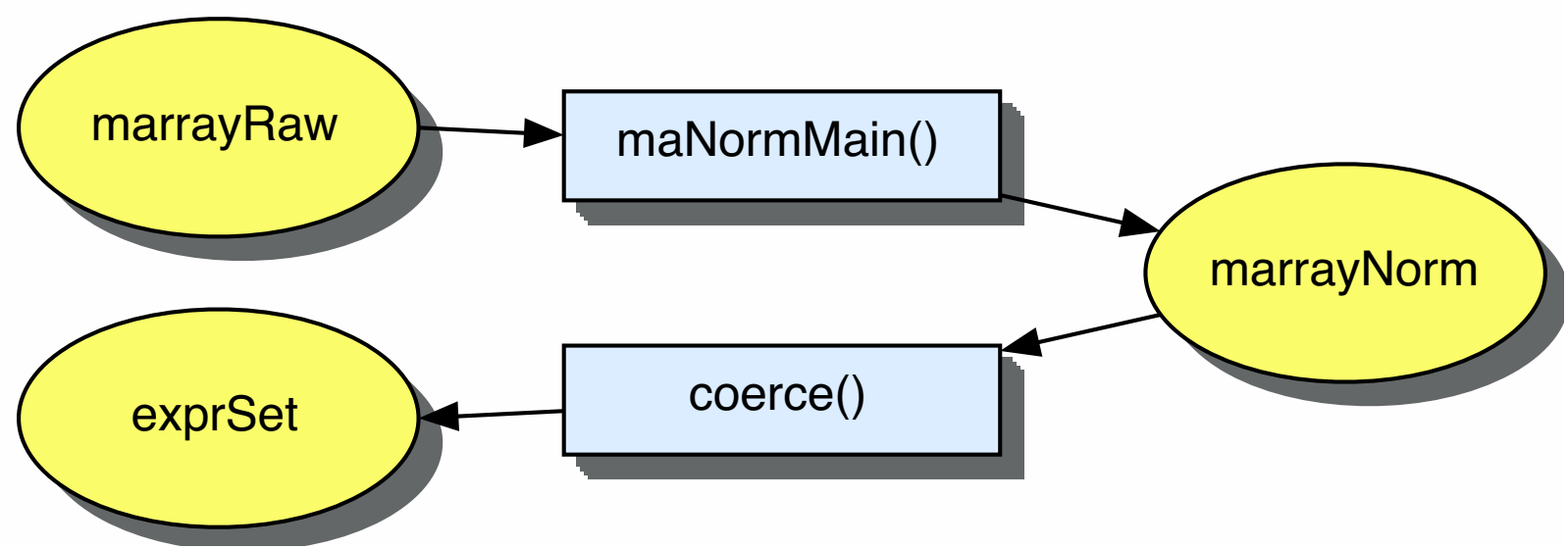
Bioconductor: Image plot of M



Bioconductor: Image plot of M



Bioconductor: After normalization

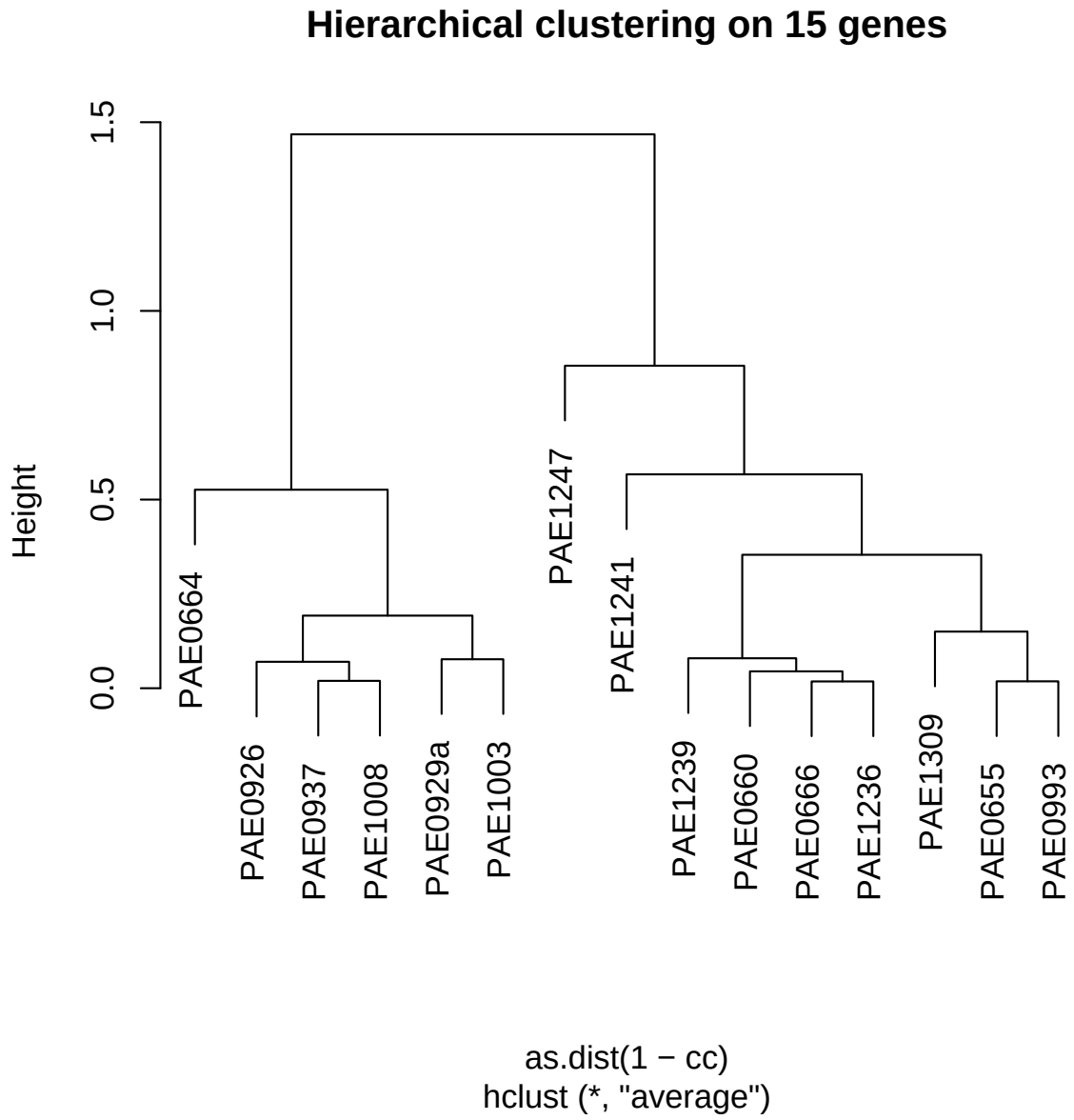


- An **exprSet** data structure is necessary for further analysis, so some manipulation needs to be done.

Bioconductor: After normalization

- Get a **exprSet** object by coercing the **marrayNorm** object.
- Filter the resulting set of genes into something suitable for analysis, using:
  - thresholds
  - filter out genes with NA values (any or all)
  - ANOVA
  - Cox regression, etc.
- Do analysis i.e. cluster, etc. (not covered today).

Bioconductor: Hierarchical clustering



Thanks

- Lily Shiue
- Prof. Todd Lowe
- Prof. Sandrine Dudoit
- Jeff Savas