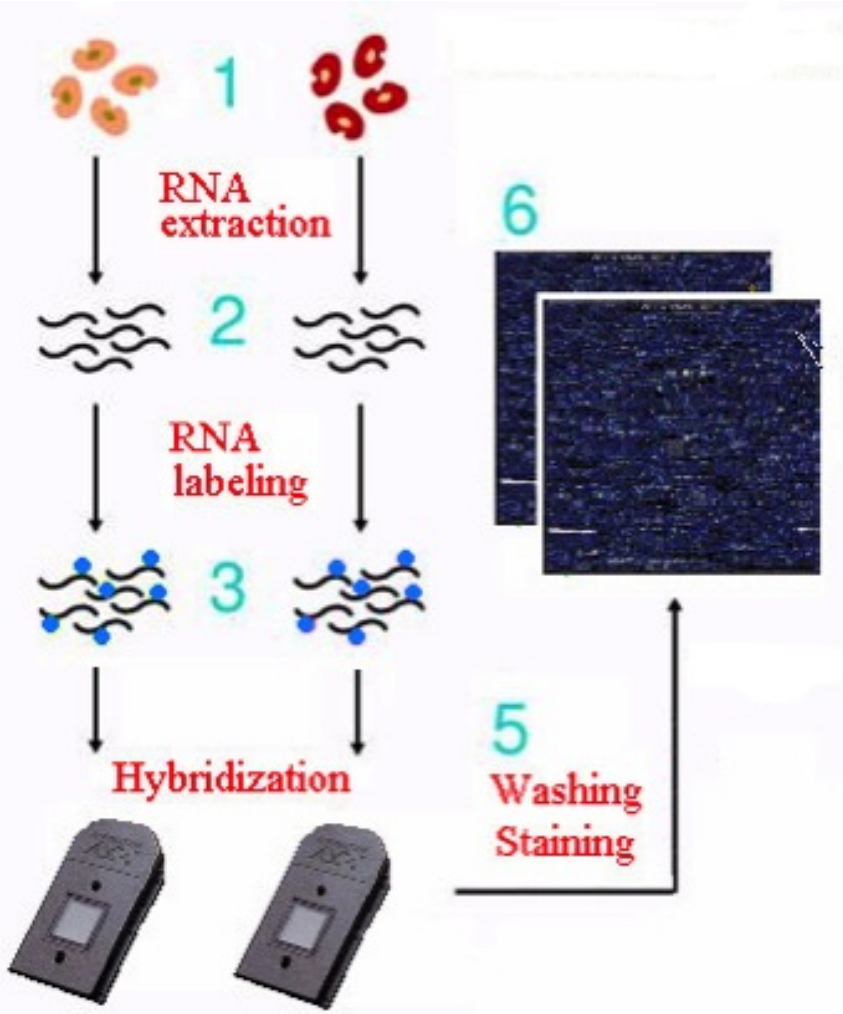
STAT588/BIOL588: Genomic Data Science
Lecture 12: Processing Microarray Data

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

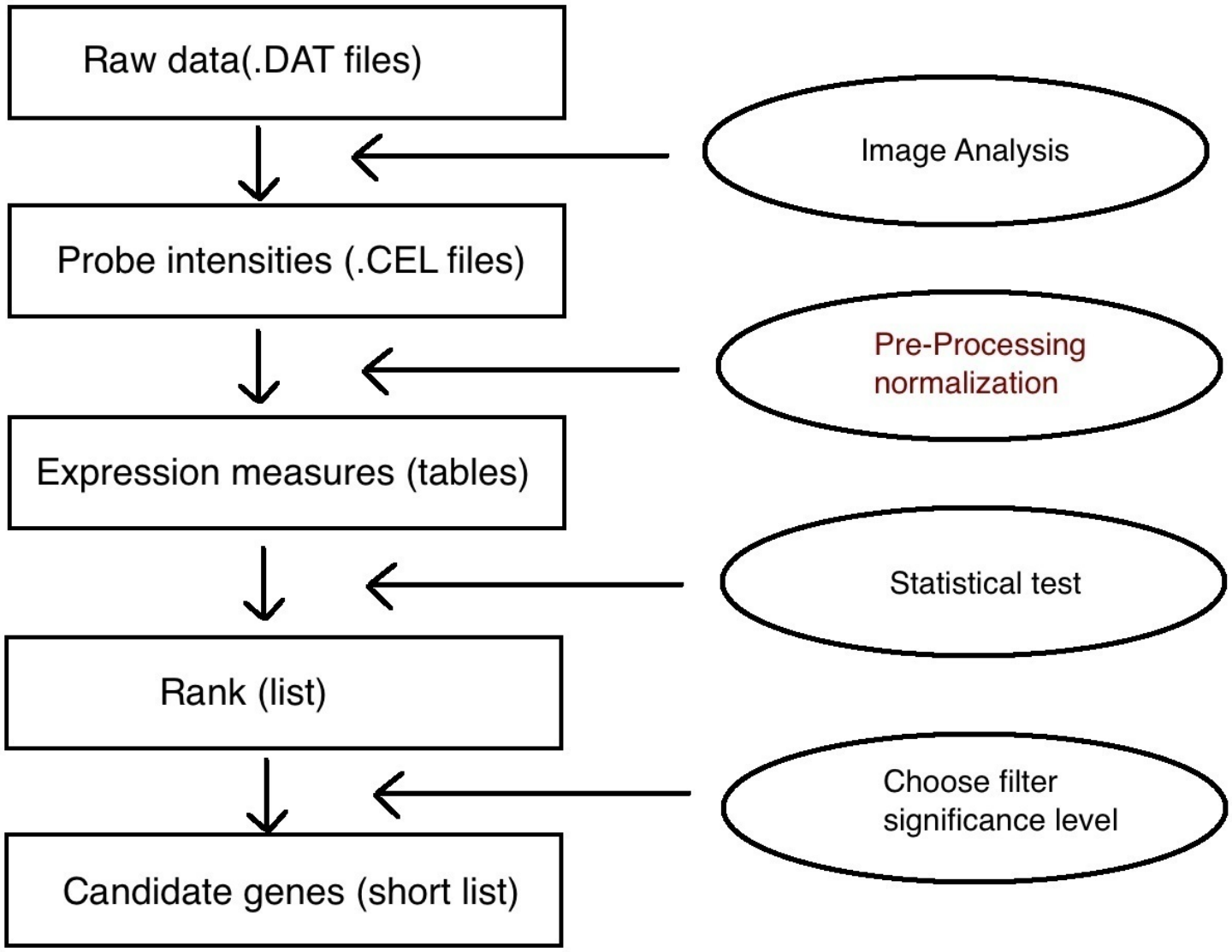
Objectives of Lecture 12

- Structures of Genomic Data
- Quality Assessment
  - Image Plot
- Preprocessing
  - Background Correction
  - Normalization
  - Probe Level Data Summarization

Affymetrix GeneChip® Experiment Protocol



Analysis Flow Chart



Affymetrix Files

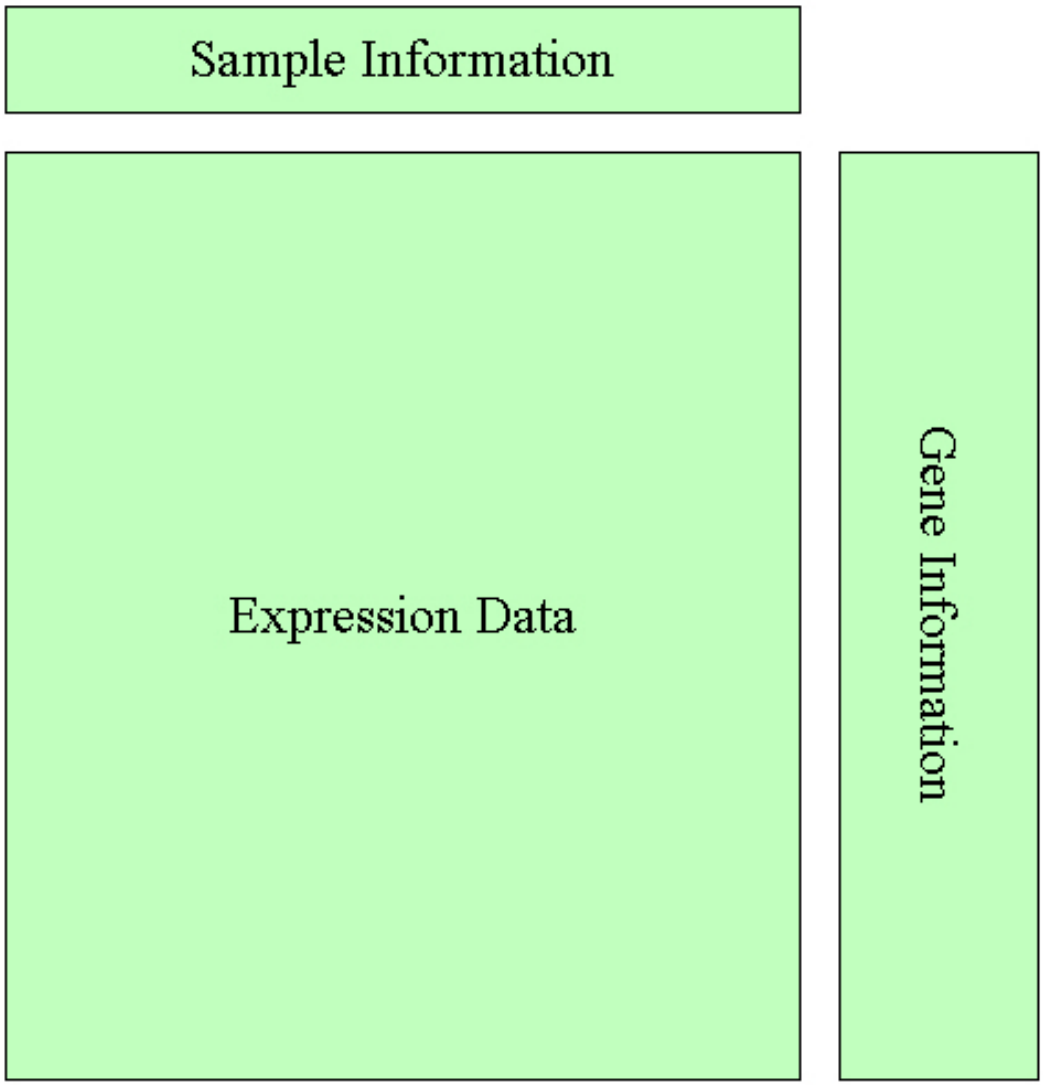
- DAT file: Scanned image.
- CEL file: Output fro image analysis software. Contains cell intensity file, probe-level values.
- CDF file: Chip description file. Describes which probes go in which probe-sets and the location of the probes on the chip.

MIAME

MIAME (Minimum Information About a Microarray Experiment)

- The raw data for each hybridization (e.g., CEL or GPR files)
- The final processed (normalized) data for the set of hybridizations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
- The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
- The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridizations are technical, which are biological replicates)
- Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
- The essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

Microarray Data Structure



Experiment/Sample Information

	Array	Age	Gender	Status
1	Array1.CEL	44	F	cancer
2	Array2.CEL	60	F	cancer
3	Array3.CEL	41	F	cancer
4	Array4.CEL	55	M	cancer

Affymetrix Data Structure

```
>library(affydata)
>data(Dilution)
>Dilution
AffyBatch object
size of arrays=640x640 features (35221 kb)
cdf=HG_U95Av2 (12625 affyids)
number of samples=4
number of genes=12625
annotation=hgu95av2
notes=
```

Look at the Experimental Design

```
>phenoData(Dilution)
An object of class "AnnotatedDataFrame"
sampleNames:  20A 20B 10A 10B
varLabels:  liver sn19 scanner
varMetadata:  labelDescription
>str(phenoData(Dilution))
>pData(Dilution)
      liver  sn19  scanner
 20A  20     0     1
 20B  20     0     2
 10A  10     0     1
 10B  10     0     2
```

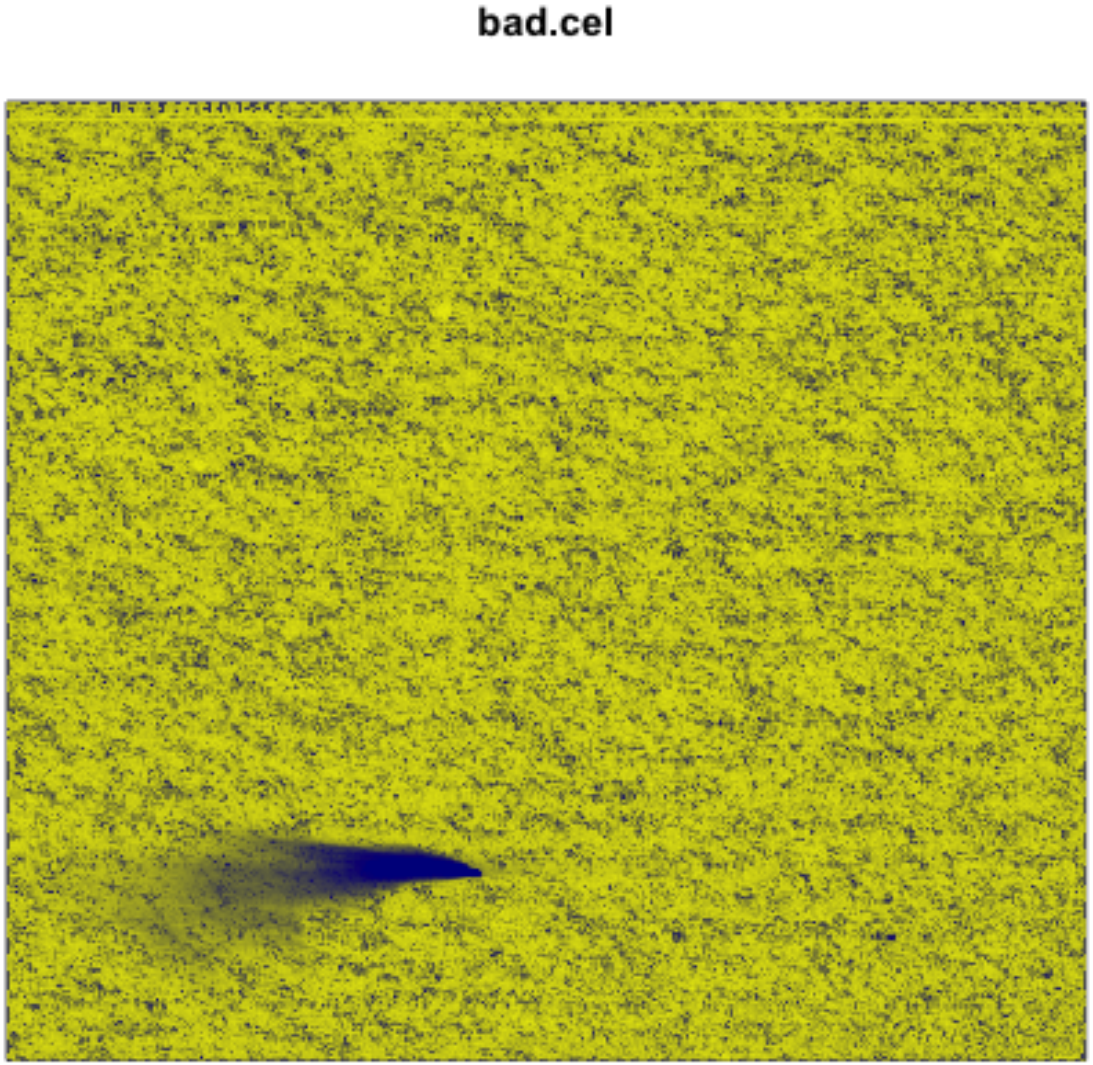
Estrogen Data

```
>library(affy)
>library("estrogen")
>library("limma")
>datadir<-file.path(.find.package("estrogen"), "extdata")
>dir(datadir)
>targets<-readTargets("phenoData.txt",
>path=datadir, sep="", row.names="filename")
## same as targets<-read.table("phenoData.txt", sep="")
>targets
>ab<-ReadAffy(filenames=targets$filename,
>celfile.path=datadir)
```

Quality Assessment

- Image plot
- simpleaffy
- affyPLM

Image Plot



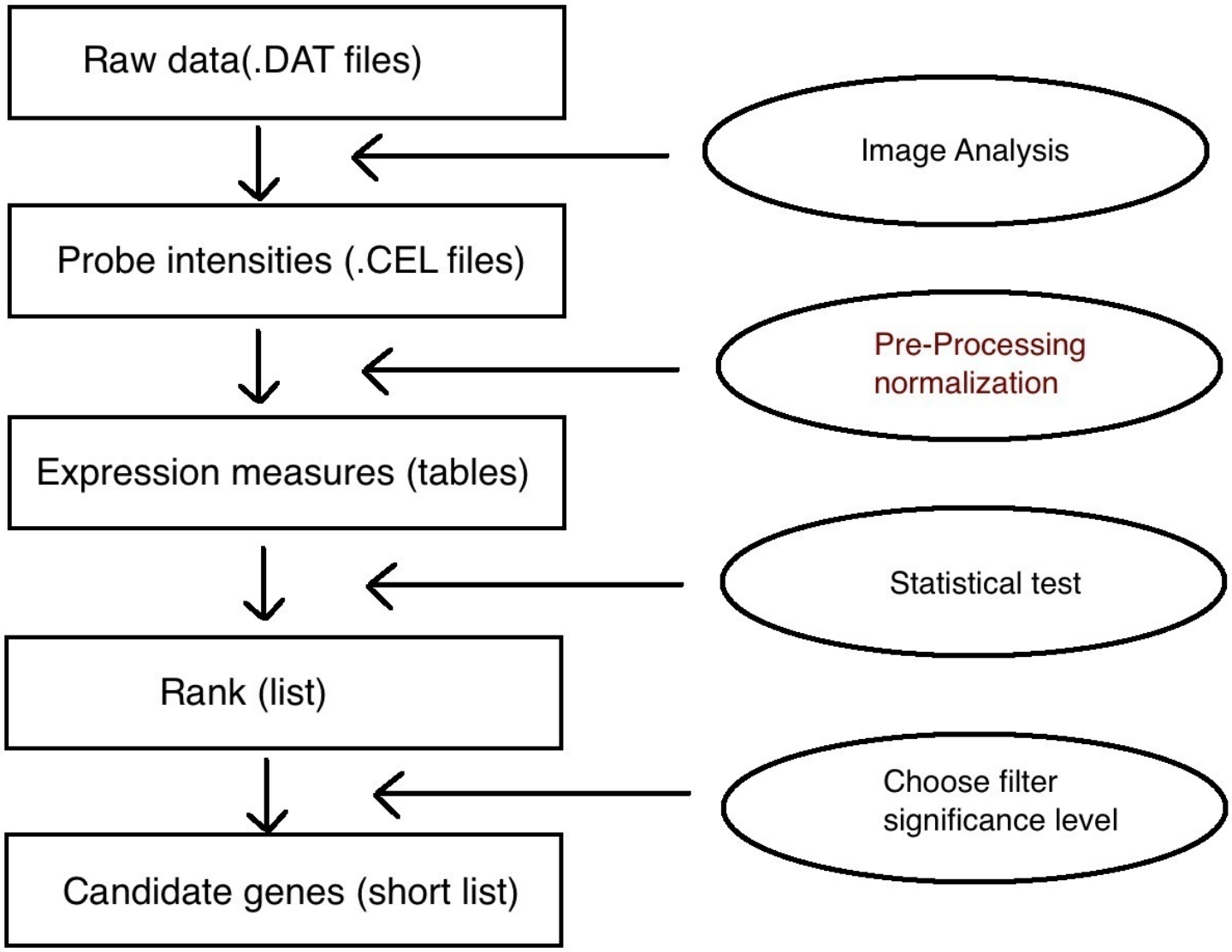
R Code

```
>image(ab[,1])
>image(ab[,8])
>badc = ReadAffy("bad.cel")
>image(badc)
>image(badc, col=heat.colors(12))
```

Objectives of Lecture 13

- Structures of Genomic Data
- Quality Assessment
  - Image Plot
- Preprocessing
  - Background Correction
  - Normalization
  - Probe Level Data Summarization

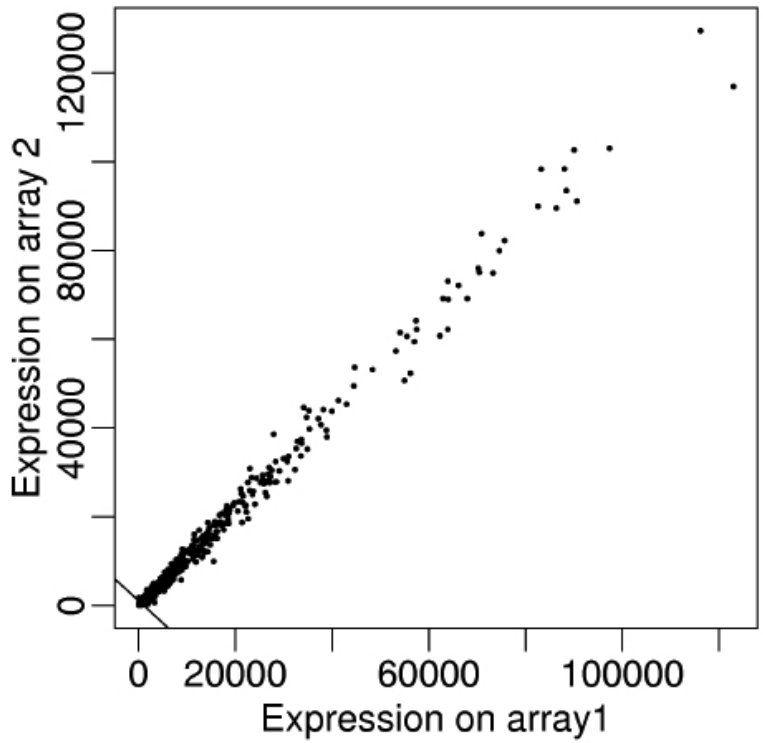
Analysis Flow Chart: Preprocessing



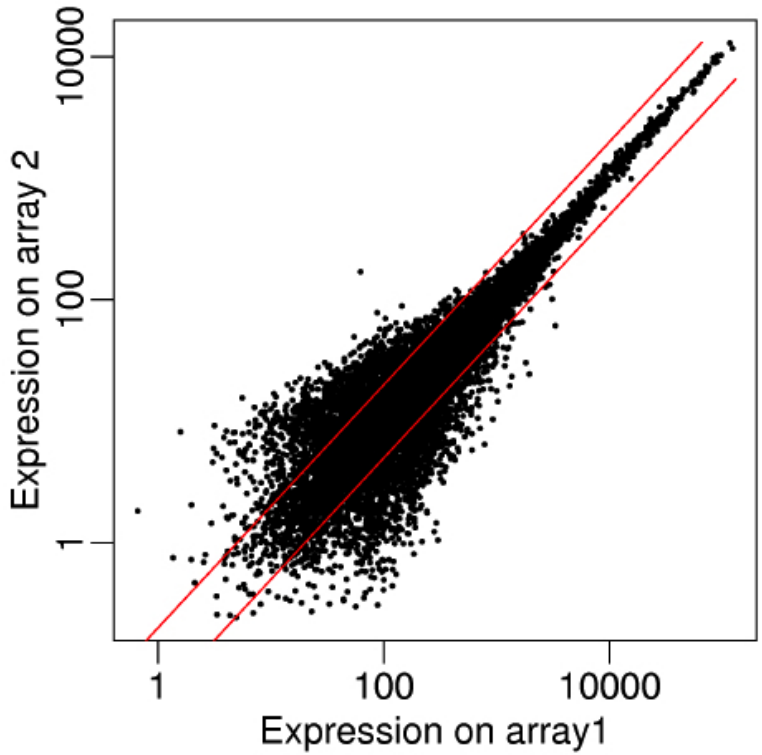
Why log

- Fold changes are the preferred quantification for differential gene expression. Fold changes are basically ratios.
- Ratios are not symmetric around 1. This makes it problematic to perform statistical operations with ratios. Hence we prefer logs.

Raw data from two arrays

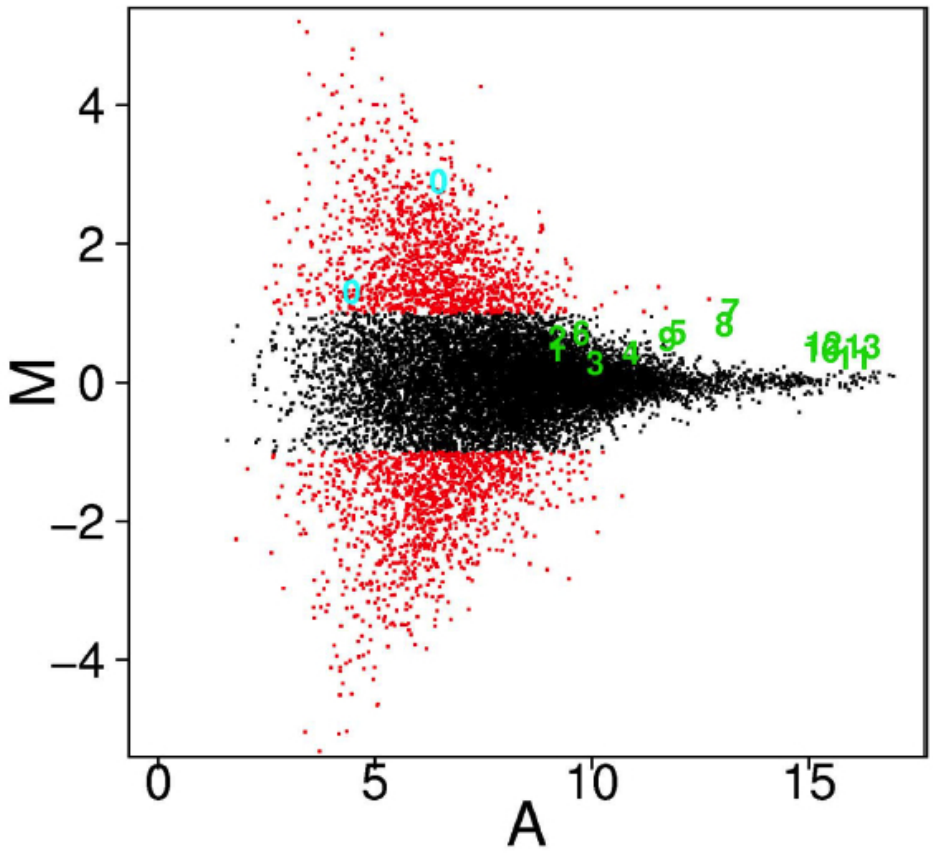


Same data in log scale



Background Noise

$M = \log_2 I_1 - \log_2 I_2, \qquad A = (\log_2 I_1 + \log_2 I_2)/2$



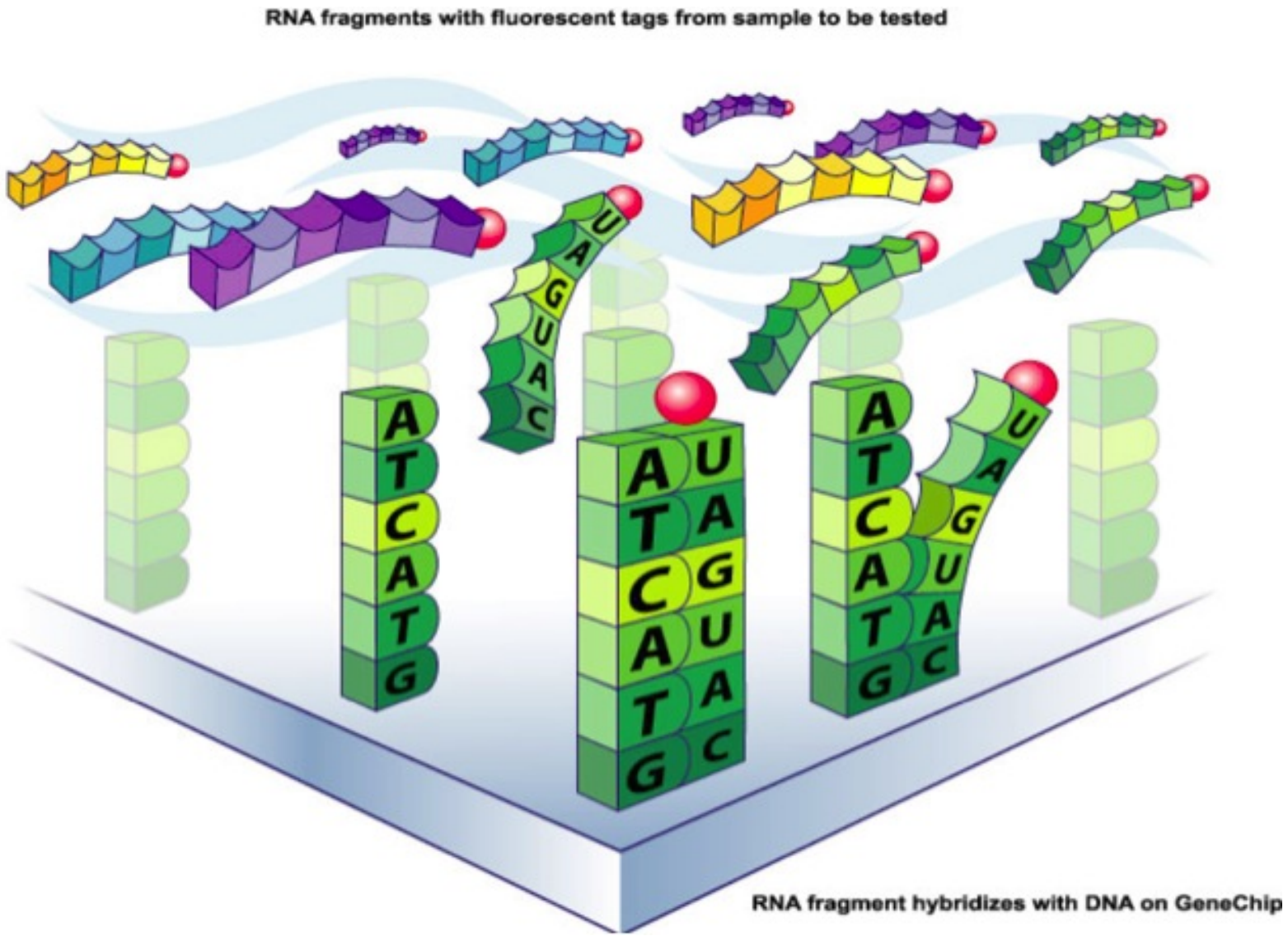
colored numbers are probes from spike-in experiment

Preprocessing: Three Steps Procedure

BioConductor breaks down the low-level processing of Affymetrix data into three steps. The design is highly modular, so you can choose different algorithms at each step. It is highly likely that the results of later (high-level) analyses will change depending on your choices at these steps.

- Background Correction: Adjust for Non-Specific Binding
- Normalization
- Probe Level Data Summarization

Affymetrix GeneChip®



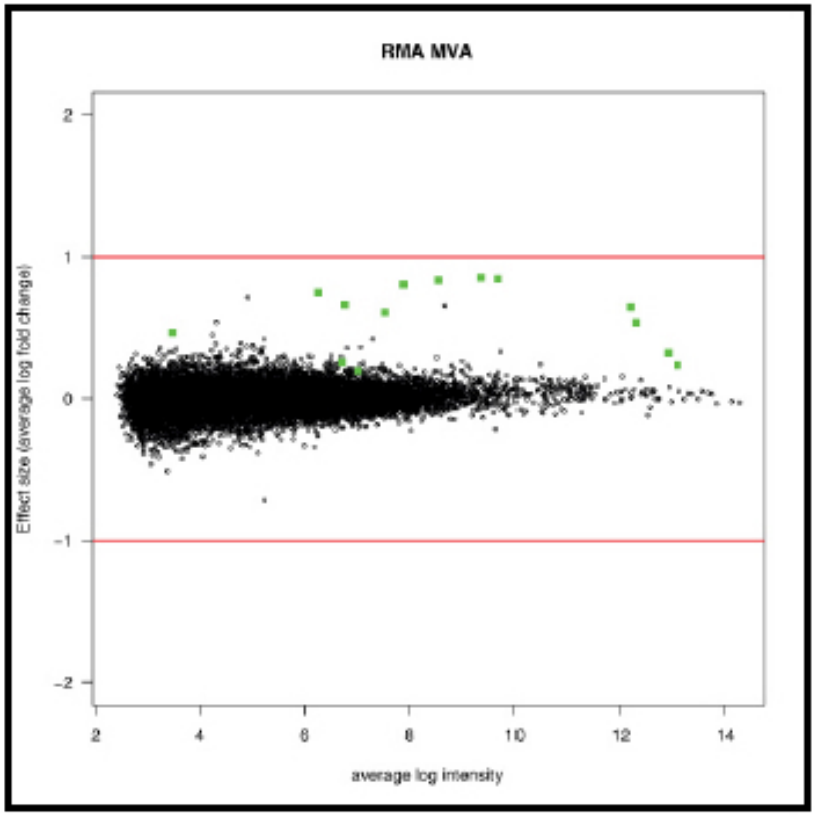
source: Affymatrix

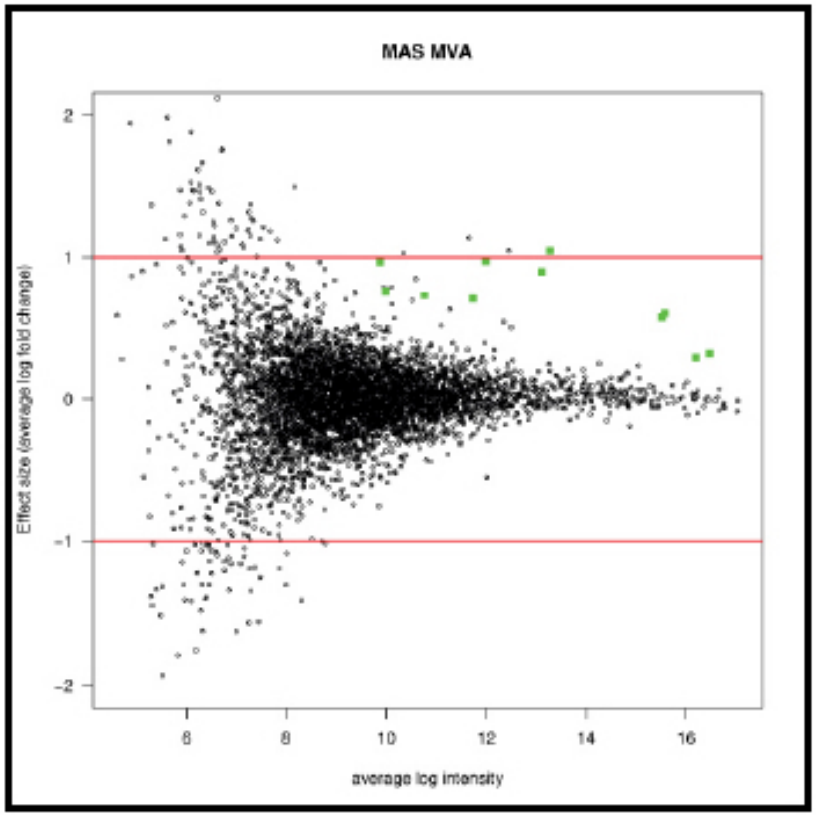
Background Adjustment

Purpose

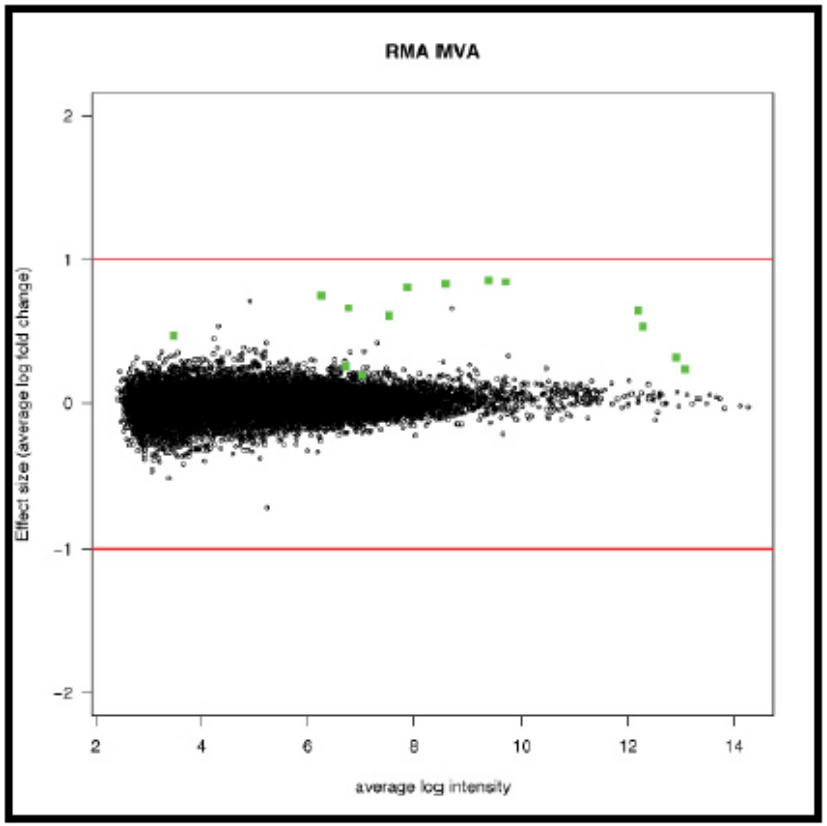
- Correct for background noise and processing effects
- Adjust for cross-hybridization, i.e. binding of non-specific DNA.
- Adjust expression measures so that they are linearly related to concentration

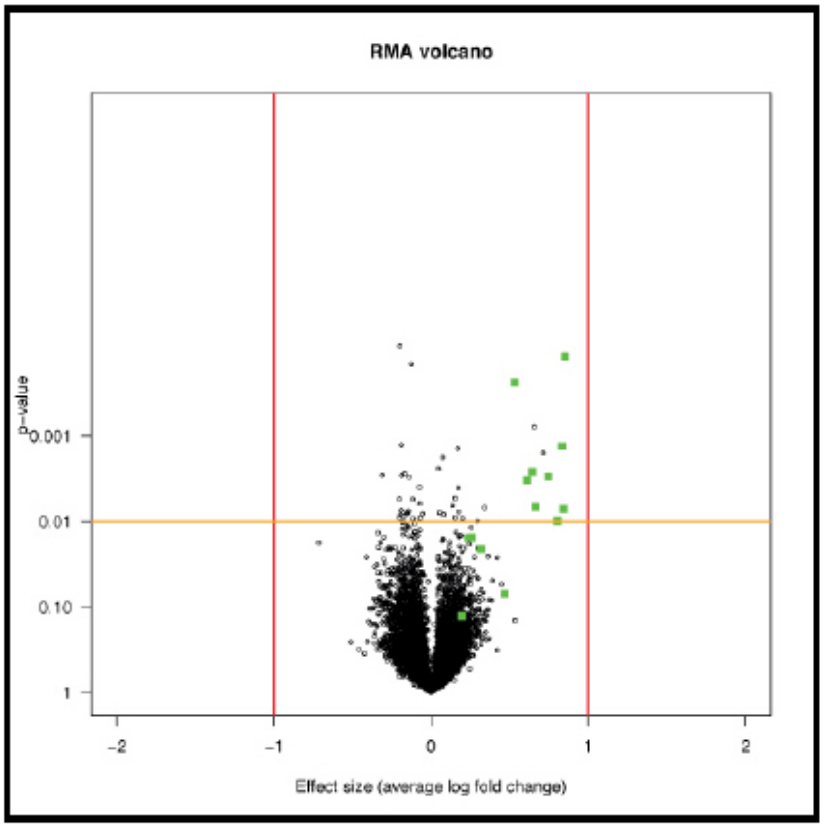
RMA versus MAS 5.0





Volcano plot





Summary

- Take logs: probe effect is additive on log scale
- Background correction reduces noise from non-specific binding
- RMA improves precision and power to detect differentially expressed genes