Affymetrix array Data Quality Assessment and Pre-Processing

Nolwenn Le Meur

Fred Hutchinson Cancer Research Center

November 2007

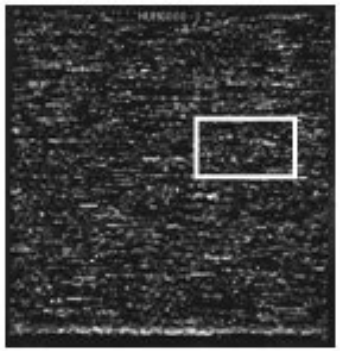
Introduction

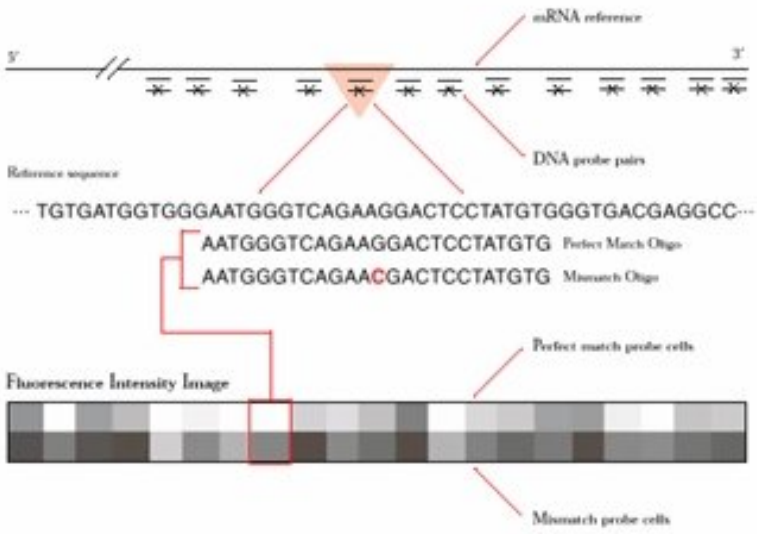
Lecture based on *Chapter 2 and 3 (Bolstad et al.)* from the book *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*

Outline

- Technology
- Quality Assessment and Quality Control
- Pre-processing
  - Background
  - Normalization
  - Summary

Affymetrix Microarray Technology





Affymetrix data

- CEL files
- CDF maping from the spot location to the probeset

Import Affymetrix data...

```
> library("affy")
> affyD <- ReadAffy()
```

- list.celfiles can be used to select the list CEL file in the directory

...into an AffyBatch object

```
> library("affydata")
> data(Dilution)
> Dilution
```

```
AffyBatch object
size of arrays=640x640 features (52588 kb)
cdf=HG_U95Av2 (12625 affyids)
number of samples=4
number of genes=12625
annotation=hgu95av2
notes=
```

...into an AffyBatch object

- pm() and mm() provide access to the probe level data

```
> pm(Dilution, '1001_at')[1:3,]
```

```
           20A   20B   10A   10B
1001_at1 128.8  93.8 129.5  73.8
1001_at2 223.0 129.0 174.0 112.8
1001_at3 194.0 146.8 155.0  93.0
```

```
> mm(Dilution, '1001_at')[1:3,]
```

```
           20A 20B   10A 10B
1001_at1 138.8  90 131.5  77
1001_at2 128.5  88 113.0  71
1001_at3 148.0 105 125.5  86
```

ProbeSet

```
> cdfName(Dilution)
[1] "HG_U95Av2"
> length(geneNames(Dilution))
[1] 12625
> length(probeNames(Dilution))
[1] 201800
```

Phenotypic data

- phenoData slot is where phenotypic data is stored.
- function pData() can be used to access this information.

```
> pData(Dilution)
```

```
    liver sn19 scanner
20A    20    0       1
20B    20    0       2
10A    10    0       1
10B    10    0       2
```

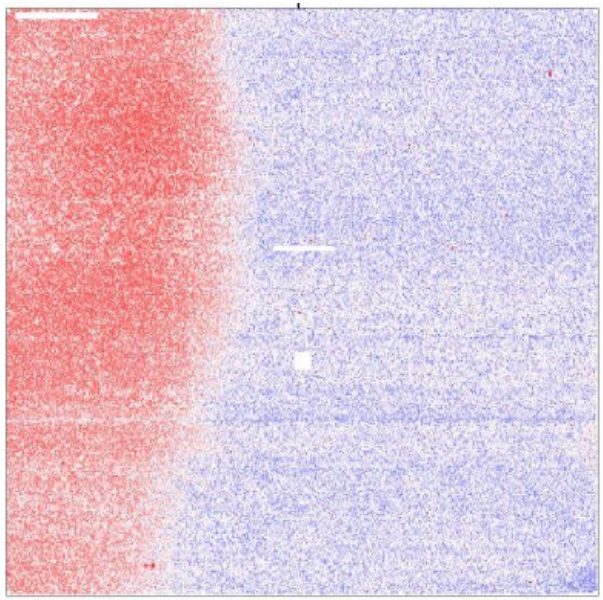
- phenotypic data consists of the concentrations of RNA from two different samples, obtained from liver and central nervous system total RNA, along with the ID of the scanner

QA/QC

- **Quality Assessment**: computation and interpretation of metrics that are intended to measure quality.
- **Quality Control**: possible subsequent actions, such as removing data from bad arrays or re-doing parts of an experiment.

Technical and experimental issues

- Systematic error
- Stochastic bias



QA/QC Tools

- Statistical summary
- Diagnostics plots

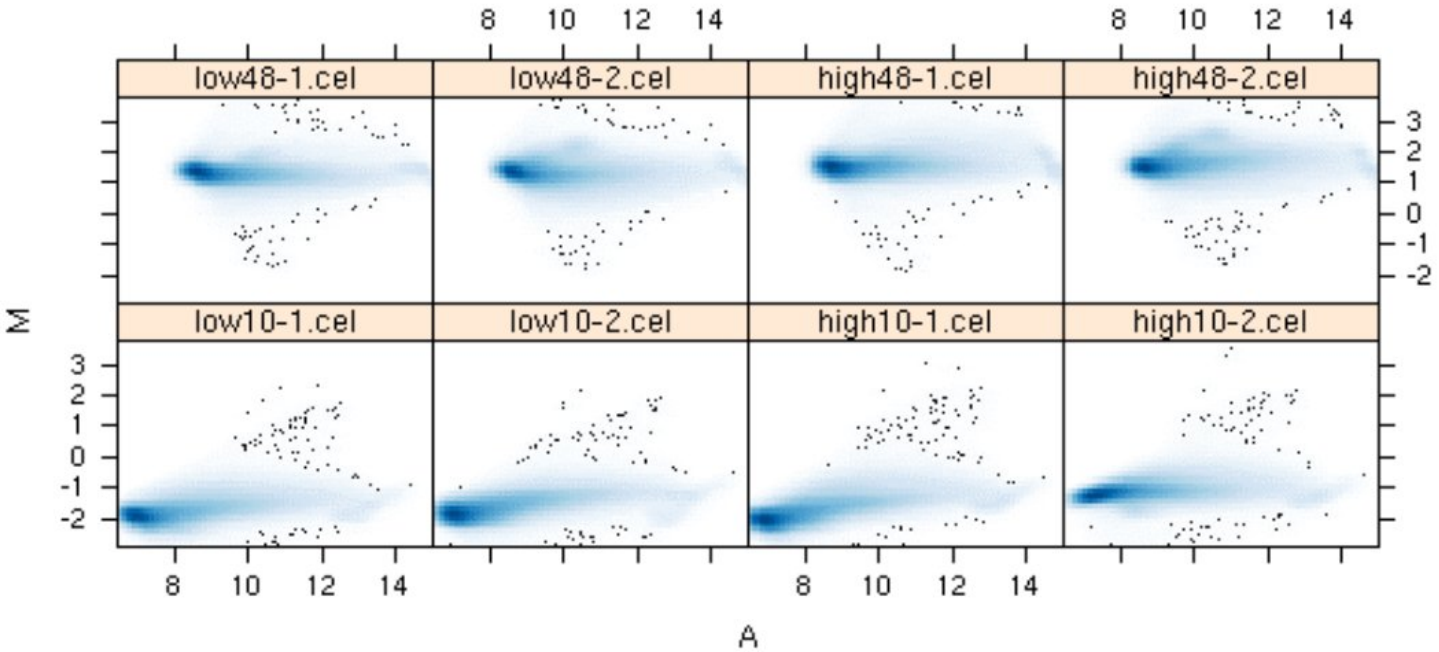
Affymetrix quality assessment metrics (simpleAffy)

- Average Background: the average of the 16 background values.
- Scale Factor: The constant i which is the ratio of the trimmed mean for array i to the trimmed mean of the reference array.
- Percent Present: the percentage of spots that are present according to Affymetrix detection algorithm.
- 3'/5' ratios: for different quality control probe sets, such as Actin and GAPDH, each represented by 3 probesets, one from the 5' end, one from the middle and one from the 3' end of the targeted transcript. The ratio of the 3' expression to the 5' expression for these genes serves as a measure of RNA quality.

Plot methods

- Individual array
- Homegeneity between array and experiments
- Stratify

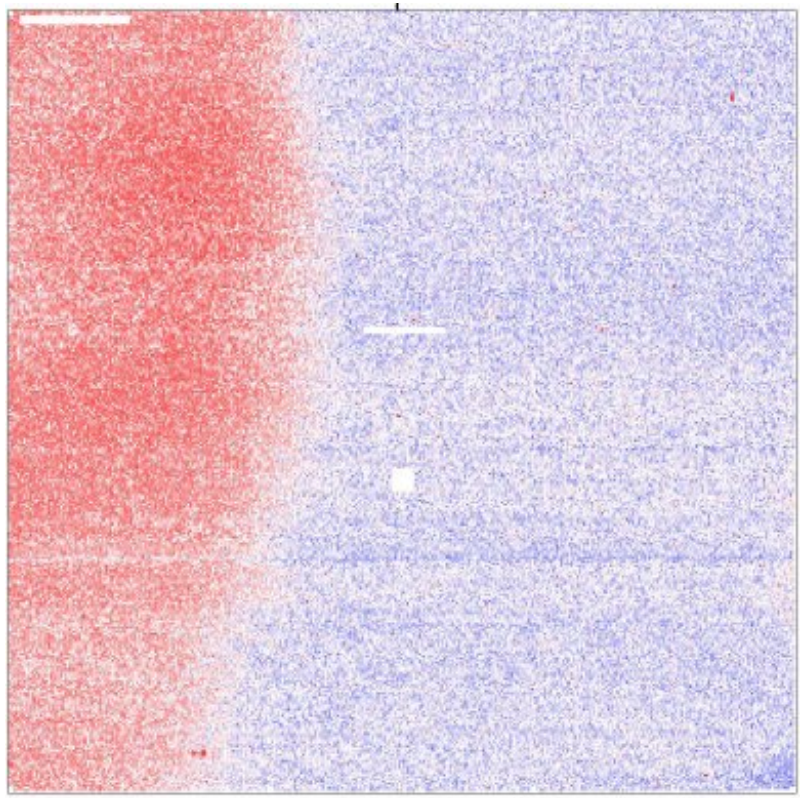
Individual array quality: MA plot



We compare to a median to 'pseudo'-array.

Individual array quality: image plot

- Spatial distribution of intensities (features, background)
- Color scale proportional to rank of intensity level

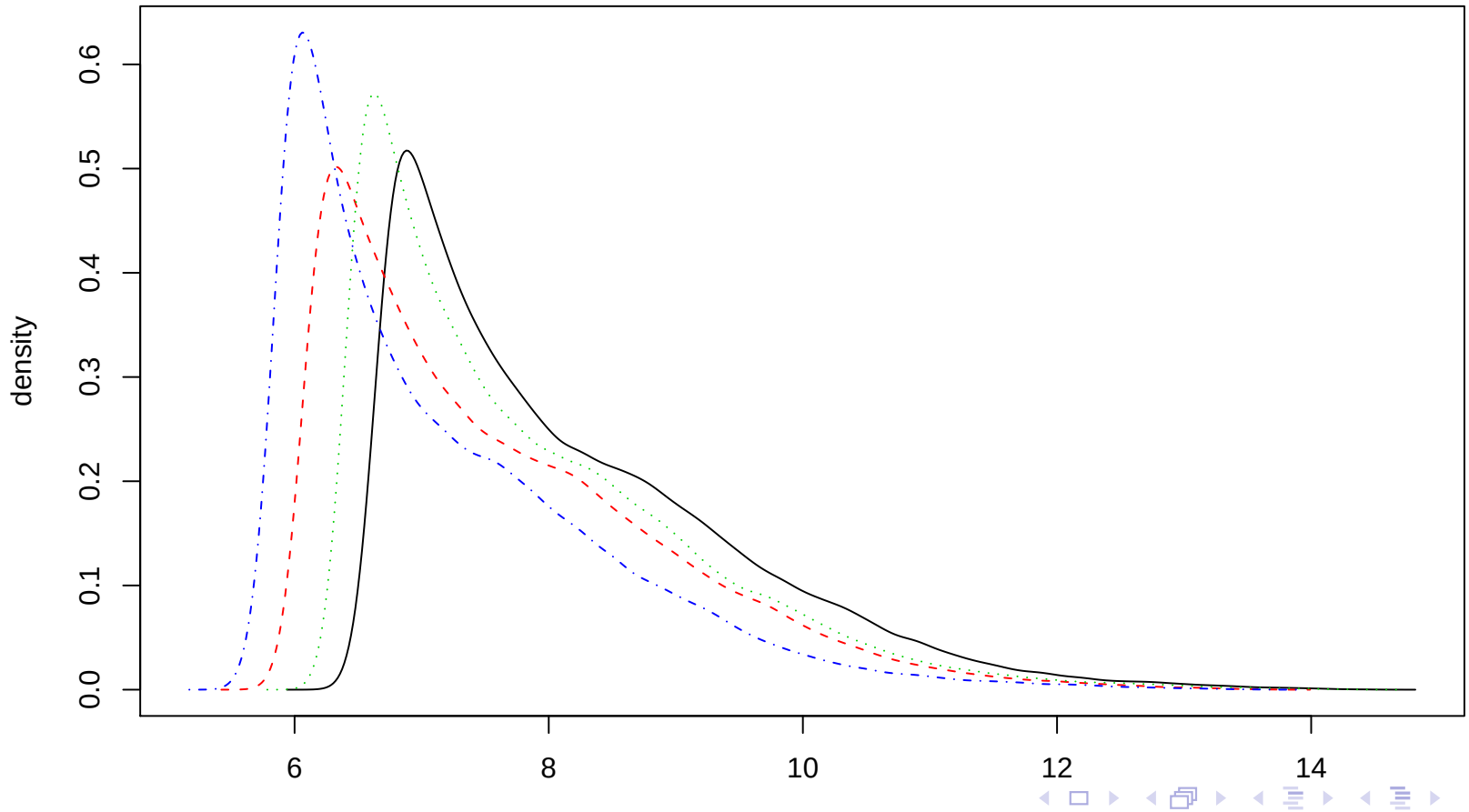


Probe Intensity Behavior

- Histogram or boxplot method: examine probe intensity behavior between arrays
- Boxplots are useful for identifying differences in the level of raw probe-intensities
- Differences between arrays in the shape or center of the distribution often highlight the need for normalization

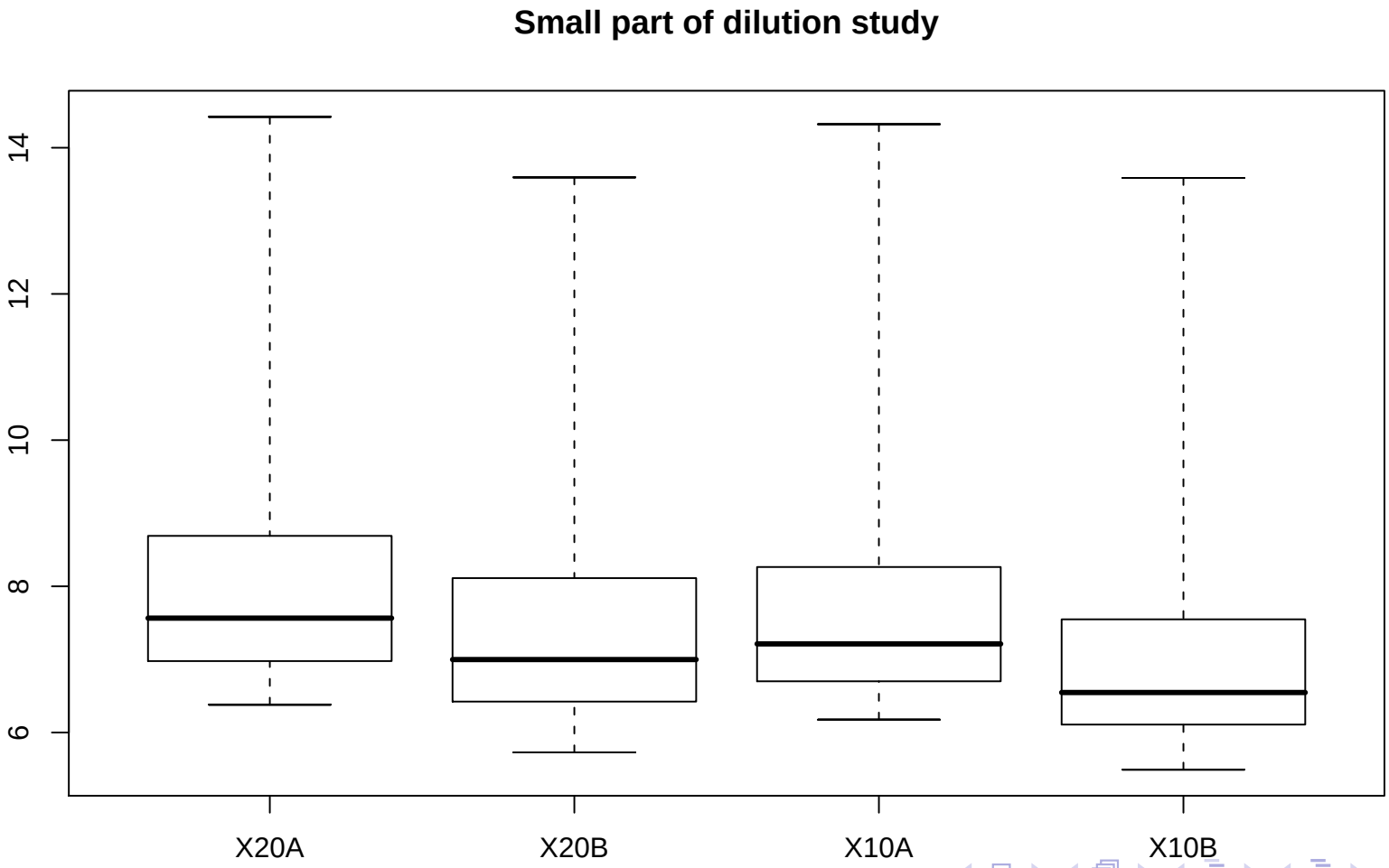
Histogram

```
> hist(Dilution)
```



Boxplot

```
> boxplot(Dilution)
```



QA report

- BioC packages:
  - simpleaffy,
  - affyQCReport
  - affyPLM
- Different output (pdf, html)

Quality Control

- Remove bad quality array
- Redo all or part of the experiment

Outline

- Technology
- Quality Assessment and Quality Control
- Pre-processing
  - Background
  - Normalization
  - Summary

Background adjustement and Normalization

- Wide variety of methods
- Take an *Affybatch* and return an *Affybatch*
- Summarization produce object of class *ExpressionSet* containing expression summary values

Why?

- Identify and remove the effects of systematic variation
- Closely related to quality assessment
- Needed to ensure that differences in intensities are indeed due to differential expression, and not some printing, hybridization, or scanning artifact
- Necessary before any analysis which involves between slide comparisons of intensities

An error model: additive background and multiplicative error

$$Y = B + S \tag{1}$$

- B: intensity due to the background noise
- S: specific binding

$$log(S) = \theta + \phi + \epsilon \tag{2}$$

- θ : logarithm of true abundance
- ϕ : probe effect
- ϵ : measurement error

Background adjustment

- Adjust intensities for non-specific signal
- Increase array sensitivity

Background adjustment for Affymetrix

- The suggested purpose of the MM probes was that they could be used to adjust the PM probes for probe-specific non-specific binding by subtracting the intensity of the MM probe from the intensity of the corresponding PM probe.
- This becomes problematic because, for data from a typical array, as many as 30% of MM probes have intensities higher than their corresponding PM probes
- Thus, when raw MM intensities are subtracted from the PM intensities many negative expression values result, which makes little sense

Background adjustment for Affymetrix

- RMA, GCRMA
- MAS 5.0
- ideal mismatch

Background adjustment for Affymetrix: RMA convolution

- The PM values are corrected, array by array, using a model for the probe intensities motivated by the empirical distribution of probe intensities.
- The observed PM probes are modeled as the sum of a noise component and a signal component.
- To avoid the possibility of negatives expression values, the Normal distribution is truncated at zero.

```
> bg.correct(Dilution, method='rma')
```

Background adjustment for Affymetrix: GCRMA

- Global RMA ignore the fact that probes undergo non-specific binding. As a result the background is often underestimated
- Characteristics of each probe are determined by its sequence.
- Using sequence information an affinity measure is computed

```
> bg.adjust.gcrma(Dilution)
```

Normalization

- Adjust intensities for technical variabilities between arrays
- Many methods
  - linear normalization (scale)
  - non-linear: cross-validated splines, running median lines, loess smoothers
  - quantile normalization: imposes the same empirical distribution of intensities to each array
- Generic function *normalize* may be used

```
>       library("affy")
>       normalize.AffyBatch.methods

[1] "constant"          "contrasts"         "invariantset"      "lo
[5] "qspline"           "quantiles"         "quantiles.robust"

> Dilution.scale <- normalize(Dilution, method='constant')
```

Variance Stabilizing Normalization (vsn)

- Combine background correction and normalization into a single procedure
- Information across arrays can be shared to estimate the background correction parameters

```
> library("vsn")
> Dilution.vsn <- normalize(Dilution, method='vsn')
```

Summarization

- Process of combining the multiple probe intensities of each probeset to produce an expression value
- Bioconductor packages provide a number of function:
  - *expresso*, *threestep* for wide variety of user specified preprocessing parameters
  - *rma*, *gcram*

expresso

- Easy to use
- Allows most background adjustment, normalization and summarization methods to be combine
- But can be slow

RMA

- convolution background corection
- quantile normalization
- median polish summarization

```
> eset <- rma(Dilution)
```

Assessing preprocessing methods

- Which methods is better?
- *affycomp* package