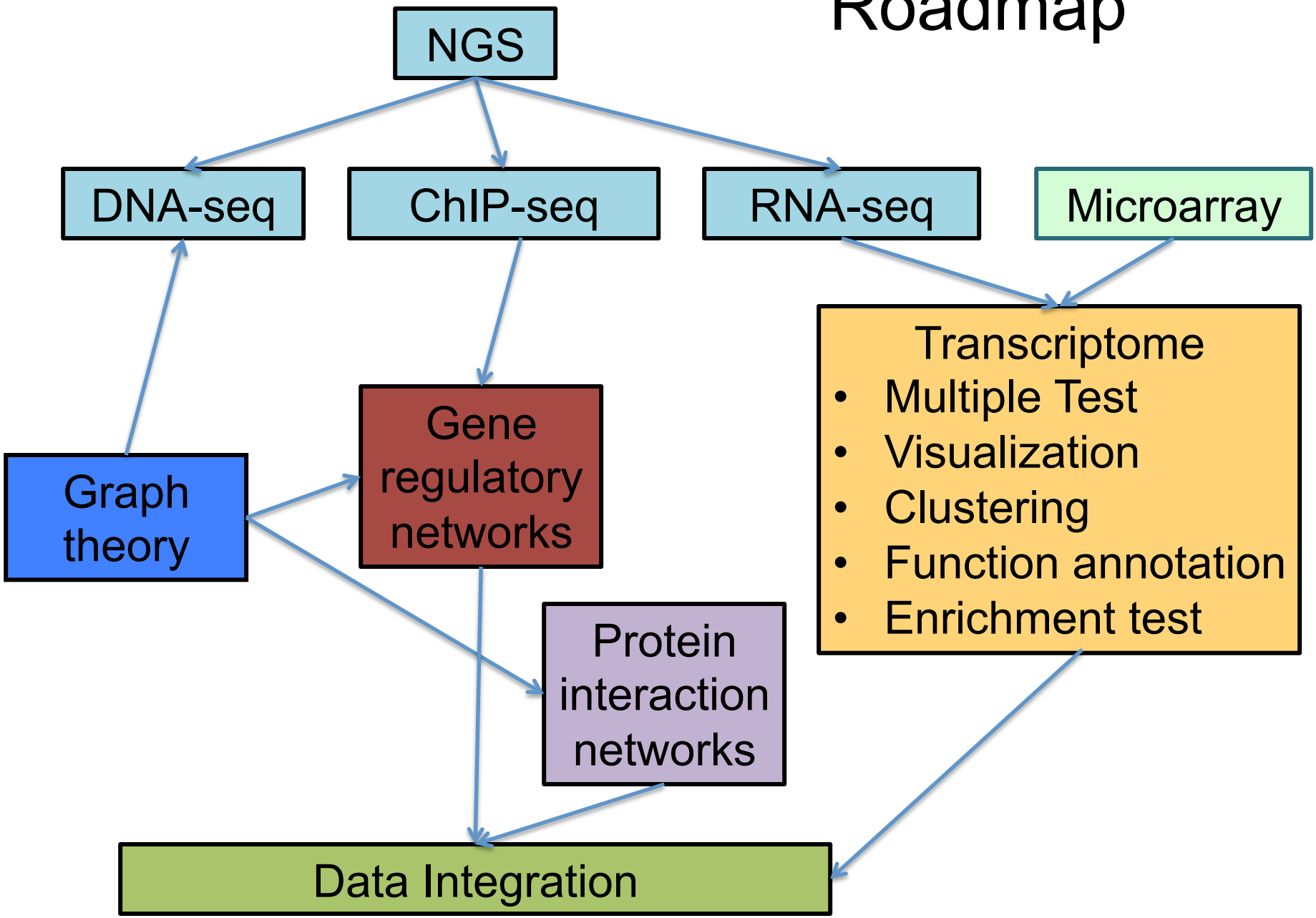
Roadmap


NGS
DNA-seq
ChIP-seq
RNA-seq
Microarray
Transcriptome
• Multiple Test
• Visualization
• Clustering
• Function annotation
• Enrichment test
Graph theory
Gene regulatory networks
Protein interaction networks
Data Integration

Microarray

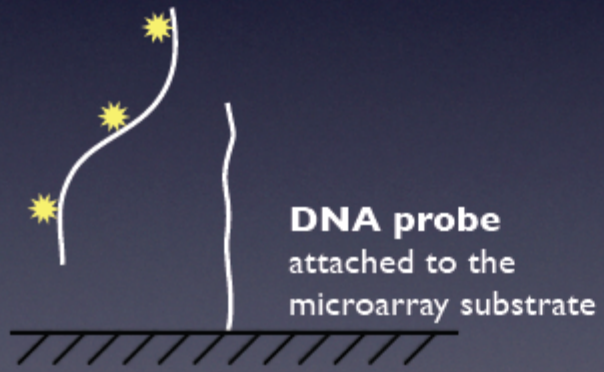
Lecture One

Outline

- Background
- Preprocessing of oligonucleotide microarray
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

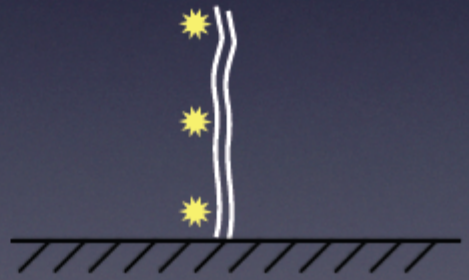
How DNA microarrays works

Labeled target
nucleic acid derived from
a biological sample



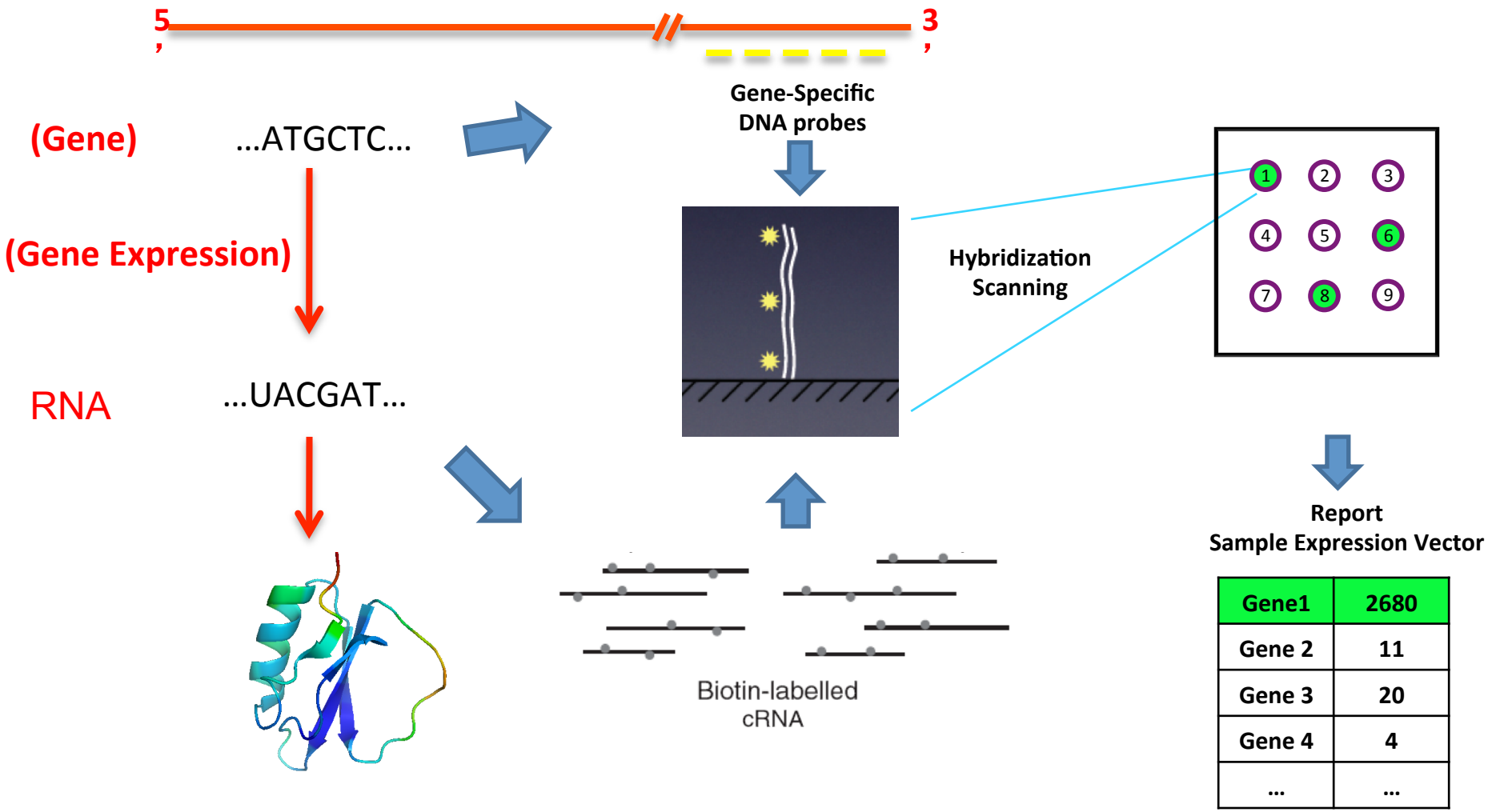
DNA probe
attached to the
microarray substrate

Specific **hybridization**
between complementary
probe and target detected
by fluorescence



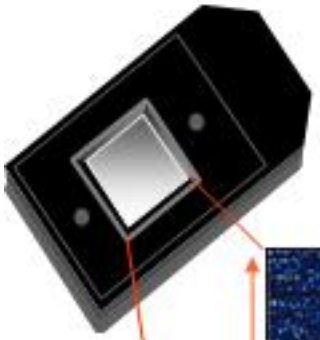
H. Jarmer 2011

Microarray

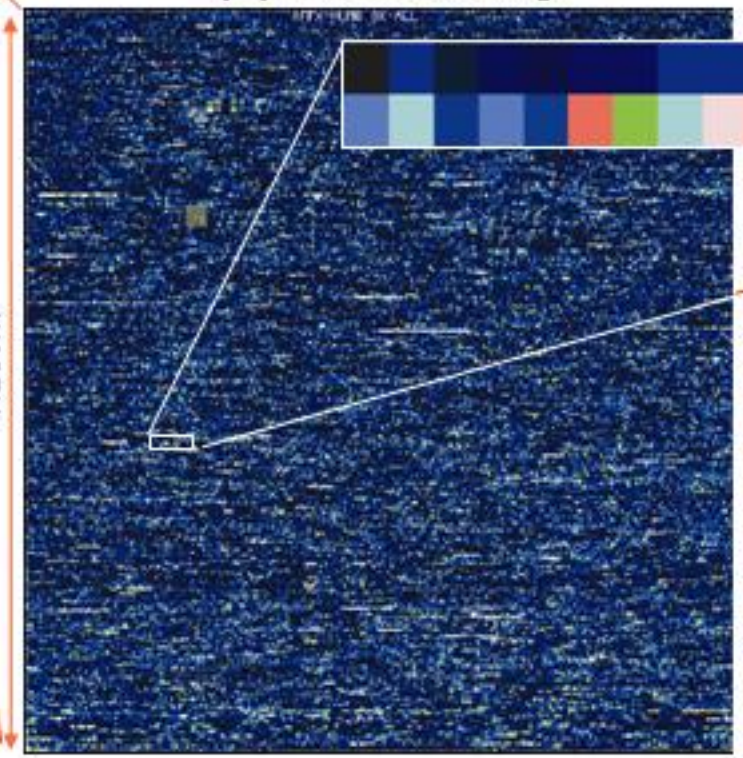


Gene1	2680
Gene 2	11
Gene 3	20
Gene 4	4
...	...

Human Genome U133A GeneChip® Array



(1) Probe Array



(2) Probe Set

Each Probe Set contains 11 Probe Pairs (PM:MM) of different probes

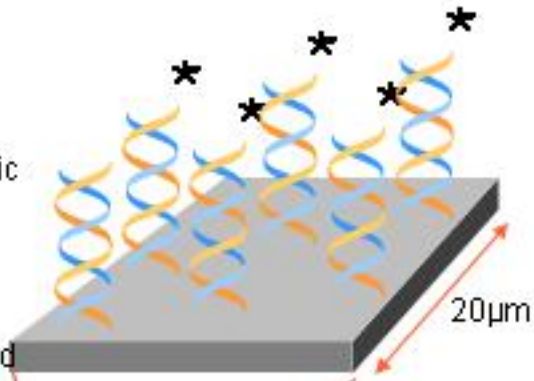
(3) Probe Pair

Each Perfect Match (PM) and MisMatch (MM) Probe Cells are associated by pairs

(4) Probe Cell

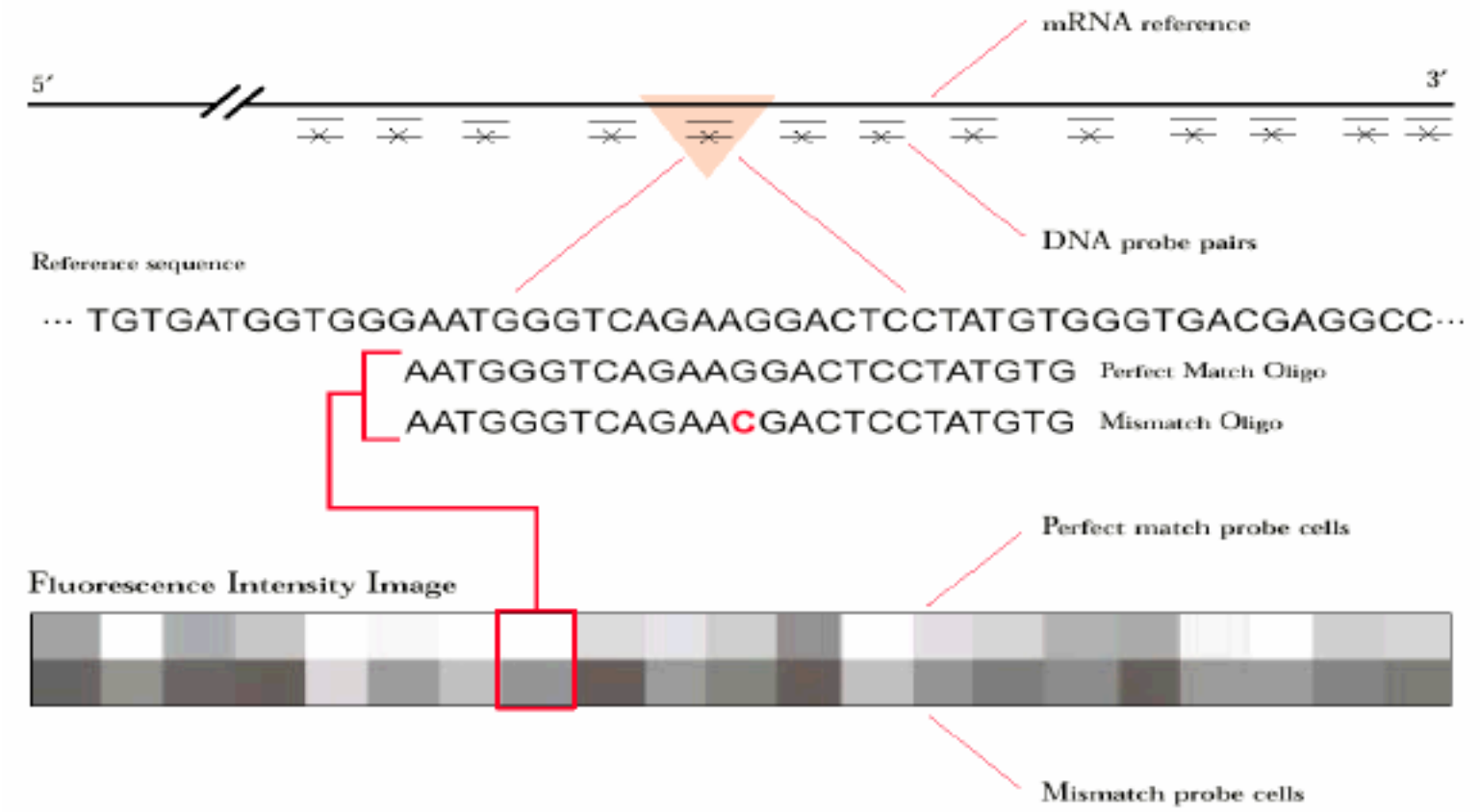
Each Probe Cell contains ~40×10^7 copies of a specific probe complementary to genetic information of interest

probe : single stranded, sense, fluorescently labeled oligonucleotide (25 mers)



The Human Genome U133 A GeneChip® array represents more than 22,000 full-length genes and EST clusters.

Probe set

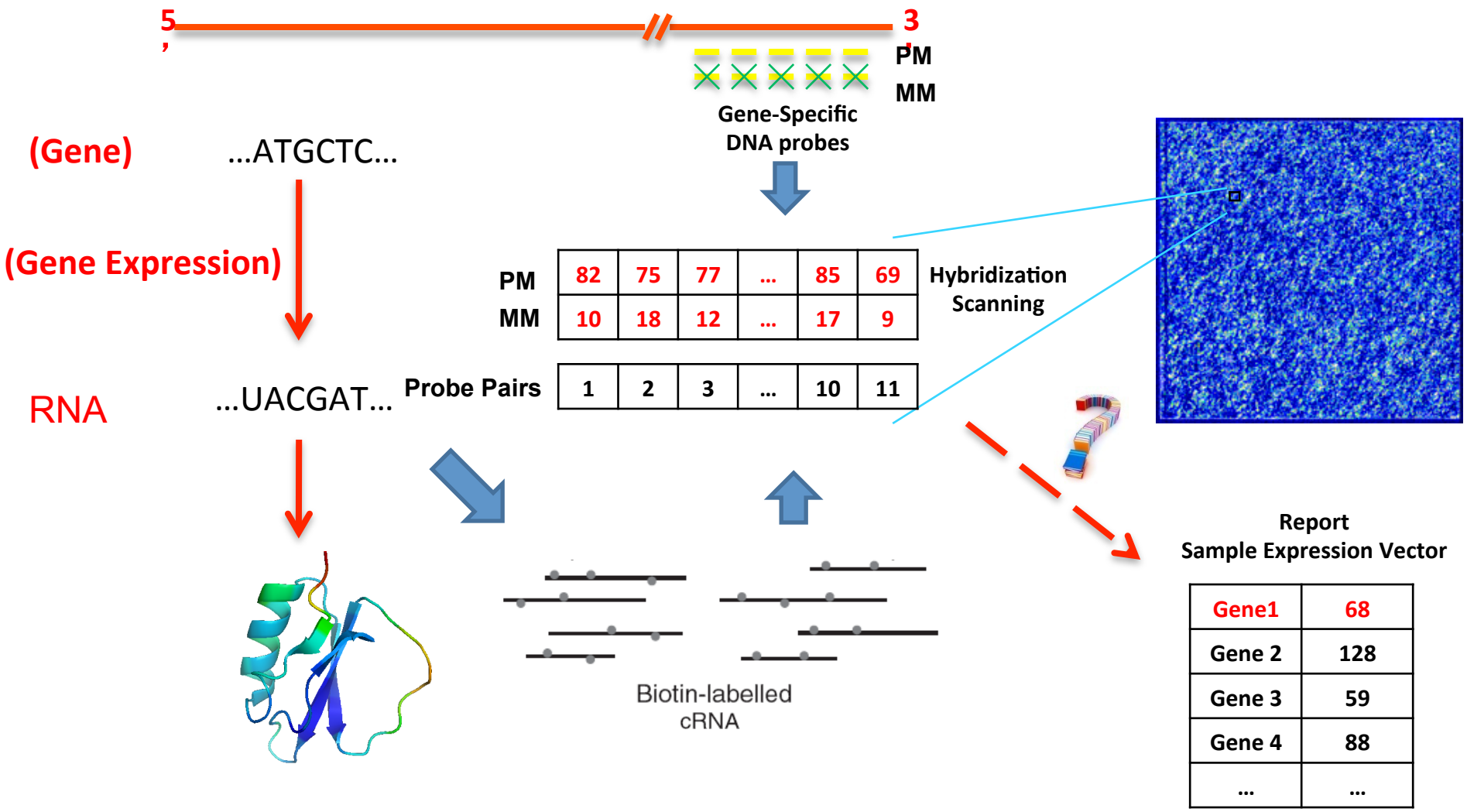


- **A probe set, consisting multiple (11-20) probe pairs, is used to measure mRNA levels of a single gene.**
- **Each probe pair contains a perfect match (PM) probe and a mismatch (MM) probe, each with 25 nucleotides in length.**

PM and MM

- **What is the difference between PM and MM probe?**
- **A PM probe perfectly matches part of a gene sequence – to maximize the hybridization**
- **A MM probe is identical to a PM probe except that the middle nucleotide (13th of 25) – to ascertain the degree of cross-hybridization**

Affymetrix Microarray



Affydata in Biconductor

- Installation

  > source(http://bioconductor.org/biocLite.R)

  > bioLite("affy")

  > bioLite("affydata")

Bioconductor packages for Affymetrix data

- affy: provides a number of statistical methods for the analysis of Affymetrix oligonucleotide arrays

  > library("affy")

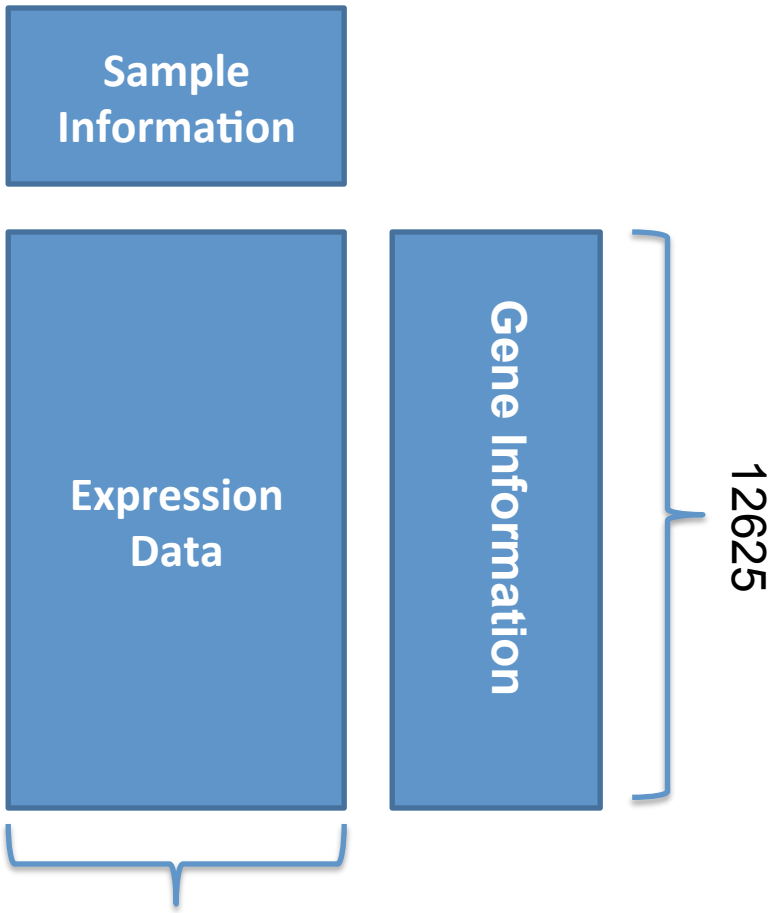
- affydata: Affymetrix data for demonstration purposes

  > library("affydata")

  > data(Dilution)

  - Function of "data" loads specified data sets, or list the available data sets.
  - The data in Dilution is a small sample of probe sets from 2 sets of duplicate arrays hybridized with different concentrations of the same RNA

Microarray Data Structure in Bioconductor:exprSet (affybatch)




```
> library("affy")
> 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=
```

Note: the number of probes is larger than the total number of genes

Sample information: pData()

```
> pData(Dilution)
      liver sn19 scanner
20A   20    0          1
20B   20    0          2
10A   10    0          1
10B   10    0          2
```

- The first two arrays: technical replicates (same RNA) from liver tissue, each array replicate was processed in a different scanner
- The second two arrays are different from the first two arrays

Expression Data: exprs()

> all_exprs_data=**exprs(Dilution)**

> **dim(exprs(Dilution))**

[1] 409600 4 # a matrix of 409600 (probes) x 4 (arrays)

> exprs(Dilution)[1,]

> all_exprs_data[1,]

20A 20B 10A 10B

149 112 129 60 # the first probe

> exprs(Dilution)[,1] # display or access the first sample

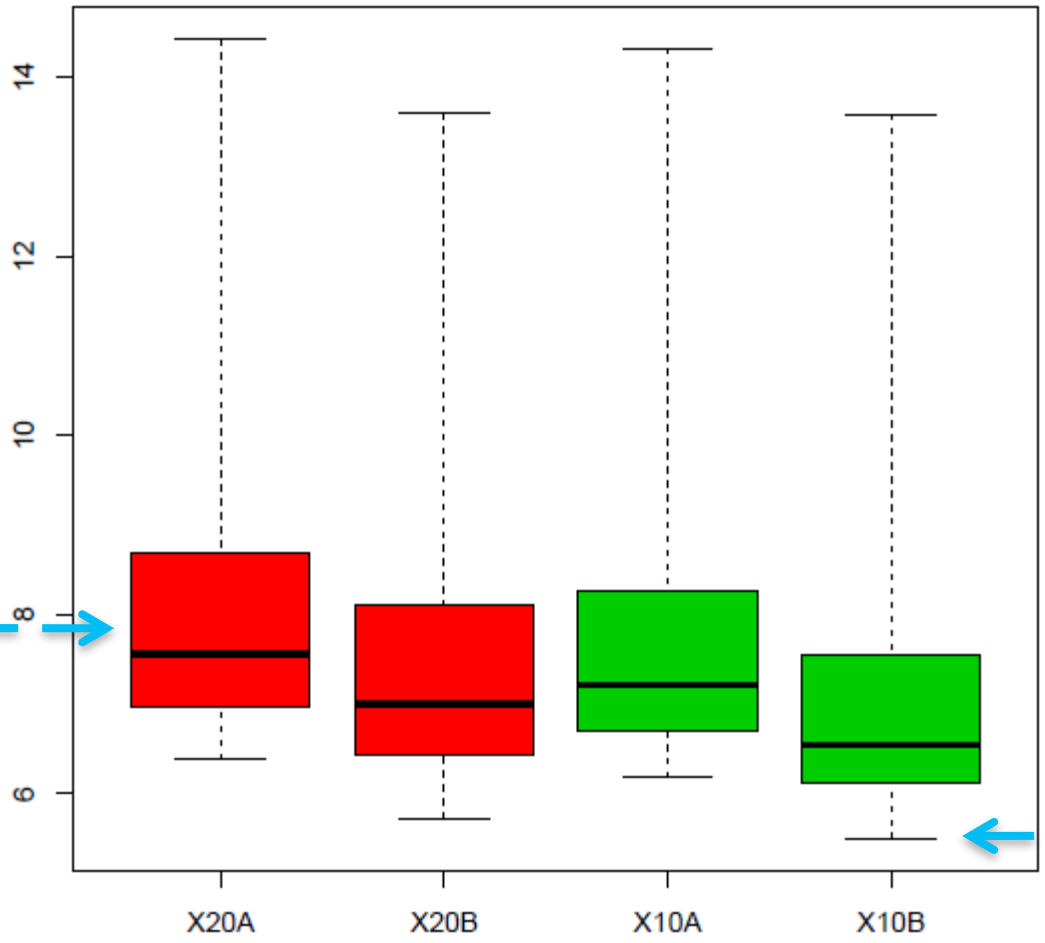
> all_exprs_data[,1]

Expression Data: pm() or mm()

```
# pm() can access the perfect match probes
# mm () can access the mismatch probes
> pm(Dilution)
> mm(Dilution)

> dim(pm(Dilution))
[1] 201800      4
> dim(mm(Dilution))
[1] 201800      4
```

Expression Data: a summary view of distribution

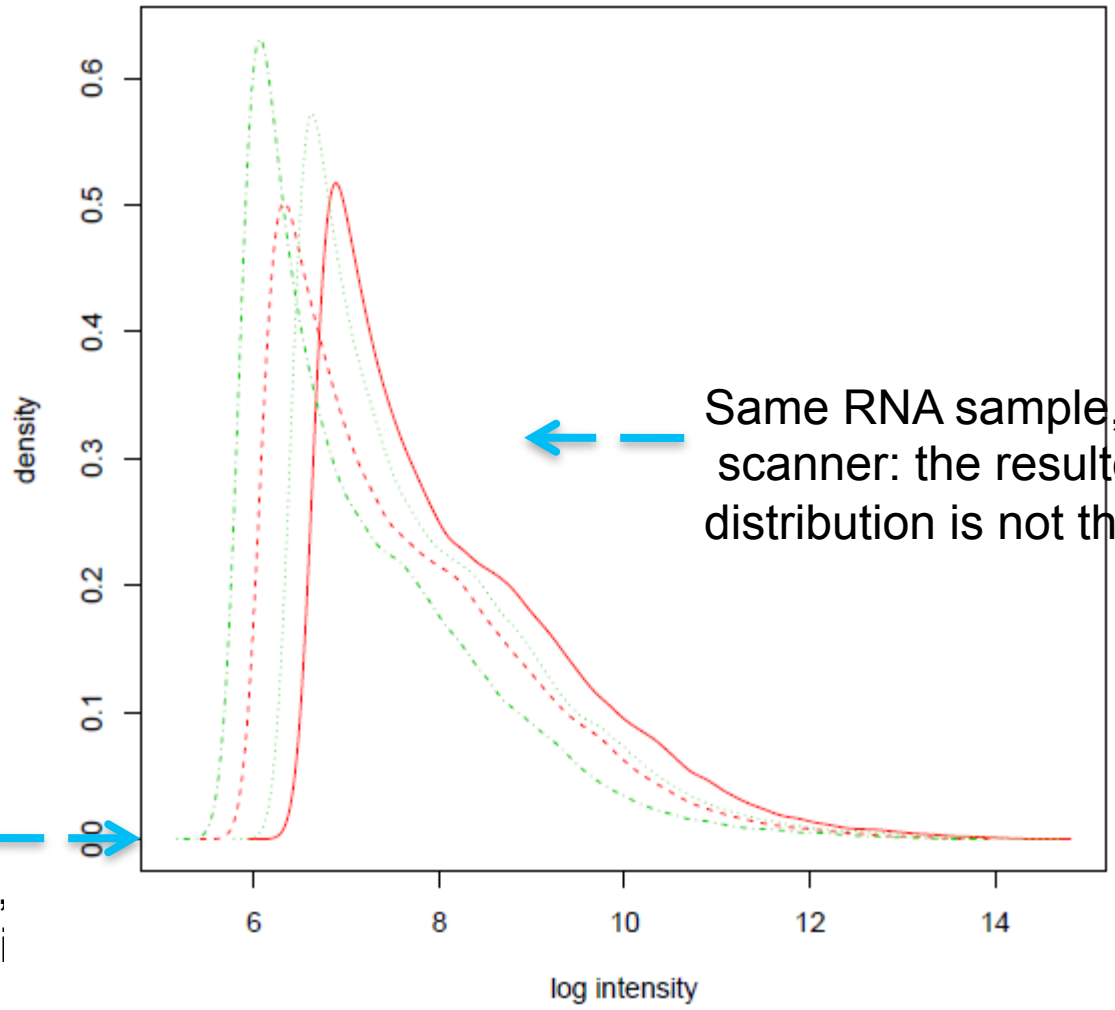


The distribution is not the same, why?

The minimum value is not zero, why? Shouldn't it be around 0?

> **boxplot(Dilution, col = c(2, 2, 3, 3))**

Expression Data: a summary view of distribution

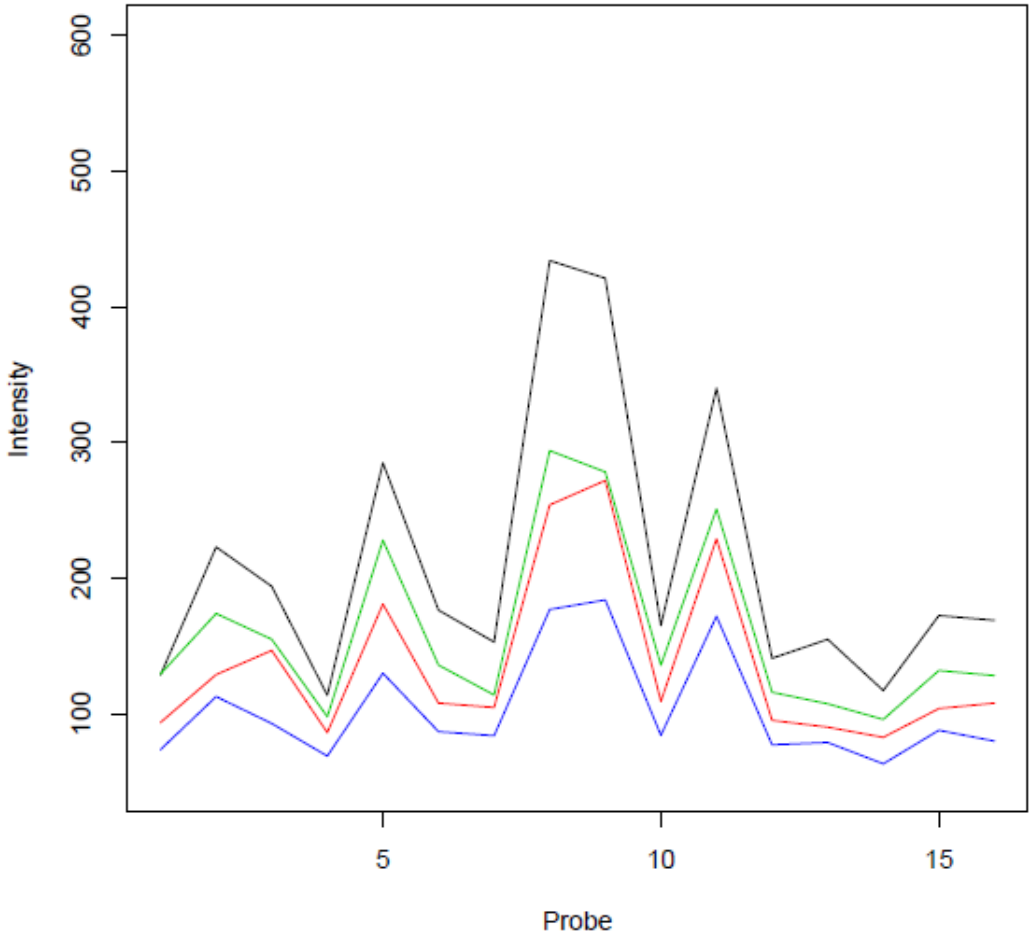


Same RNA sample, just different scanner: the resulted intensity distribution is not the same! Why?

The minimum value is not zero, why? Shouldn't i

> **hist(Dilution, col = c(2, 2, 3, 3))**

Expression Data: individual probeset



- The intensity profile for PM probes of probeset “1001_at” at the 4 different arrays

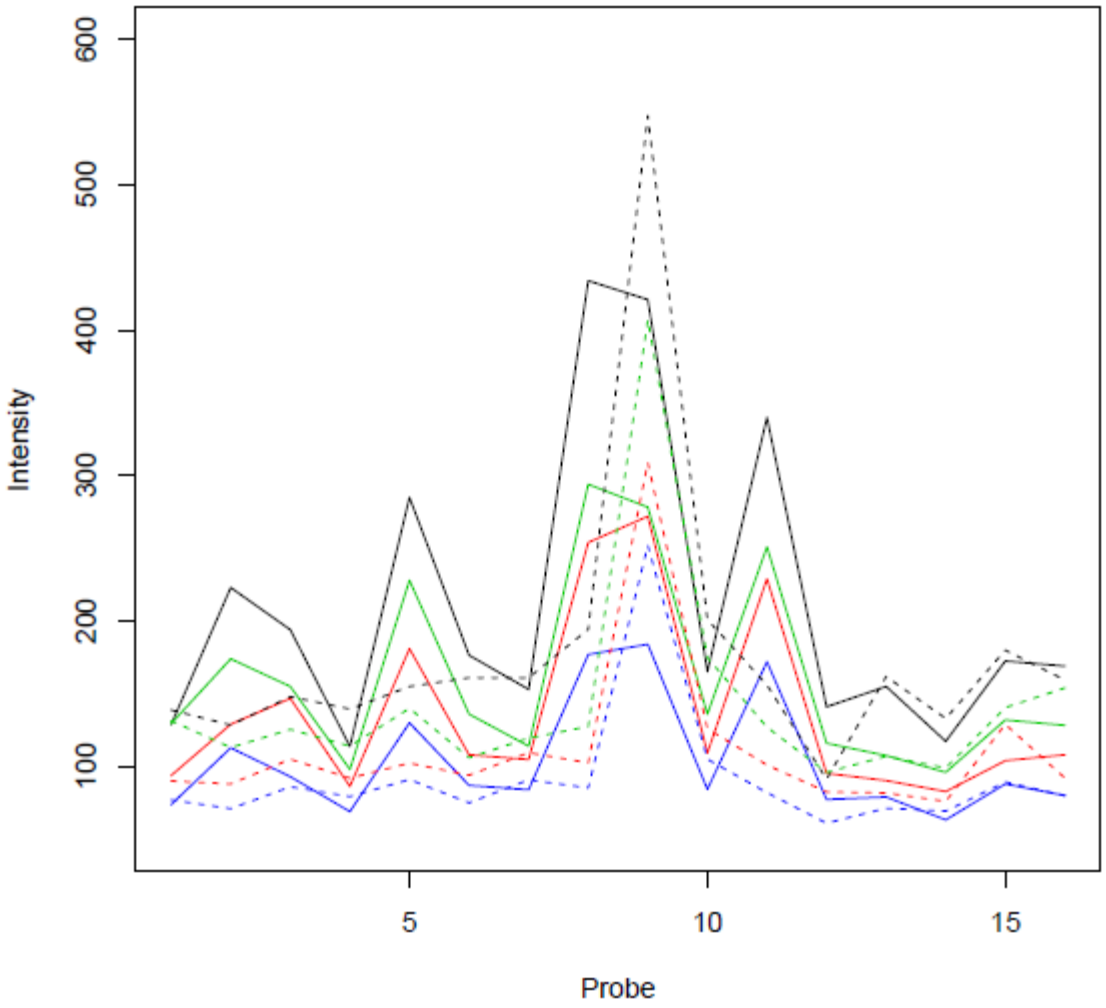
```
> plot(c(1,16), c(0, 800), type='n', xlab='Probe', ylab='Intensity')
> for (i in 1:4)   lines(pm(Dilution, "1001_at")[,i], col=i)
```

Expression Data: individual probeset

- The affy package includes tools for extracting individual probe set from a complete AffyBatch object.

```
> pm(Dilution, "1001_at")
> mm(Dilution, "1001_at")
              20A   20B   10A   10B
1001_at1  138.8  90.0 131.5  77.0
1001_at2  128.5  88.0 113.0  71.0
...
1001_at15 180.0 129.0 140.5  89.3
1001_at16 159.5  92.0 154.0  80.0
```

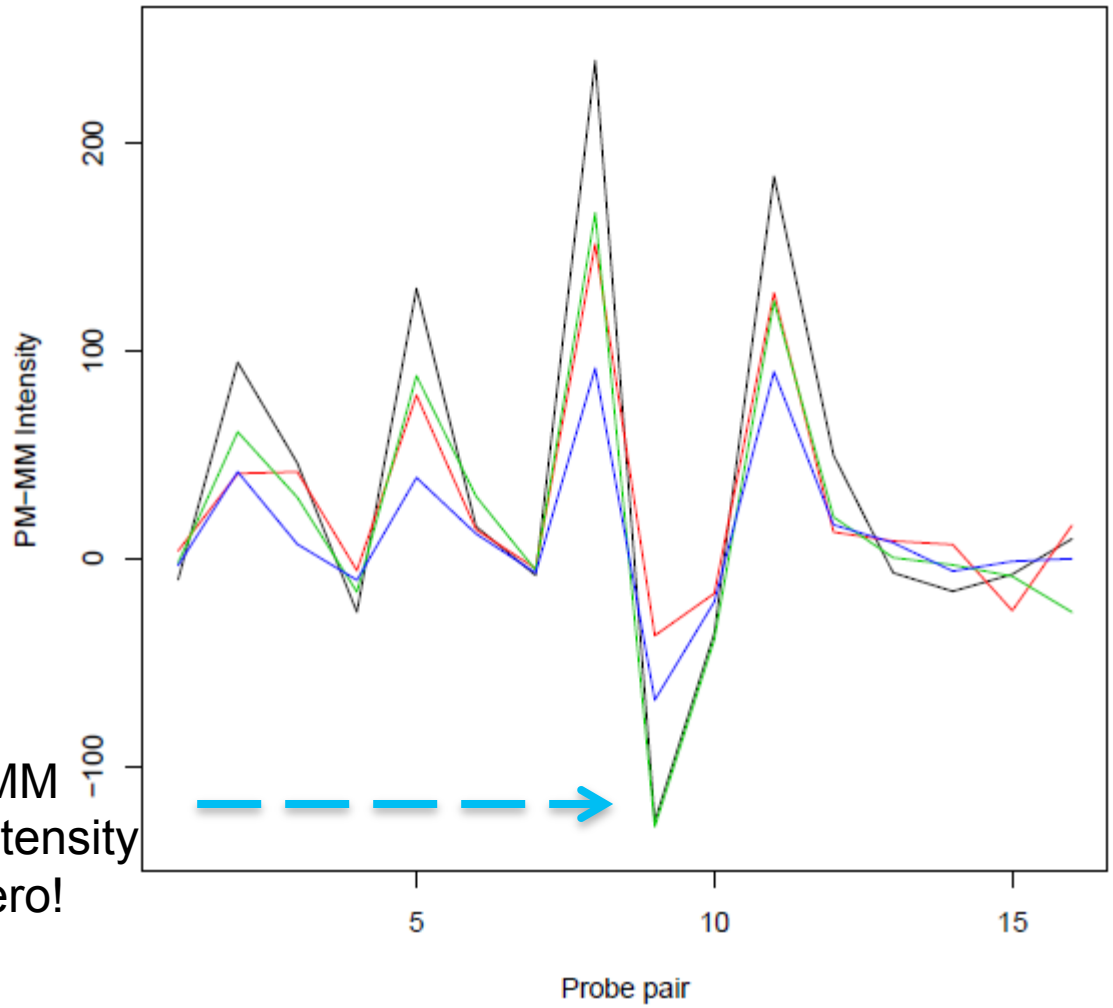
Expression Data: individual probeset



- Adding the intensity profile for MM probes of probset "1001_at" at the 4 arrays (PM: solid line, MM: dash line)

> for (i in 1:4) lines(mm(Dilution, "1001_at")[,i], col=i, lty=2)

PM – MM for individual probeSet



Some PM-MM pair have intensity less than zero! Why?

- The PM –MM intensity profile for probset "1001_at" at the 4 arrays

> for (i in 1:4) lines(pm(Dilution, "1001_at")[,i]-mm(Dilution, "1001_at")[,i], col=i)

Outline

- Background
  - Biology Background
  - Introduction to useful packages in Bioconductor
- Preprocessing of oligonucleotide microarray
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

Pre-processing affy microarray

BioConductor breaks down the pre-processing of Affy microarray into four steps. Different algorithms can be chosen at each step. It is highly likely that the pre-processing results will change depending on the choices at each steps.

1. Background correction
2. Normalization
3. PM-MM correction (optional)
4. Summarization

Preprocess methods in Bioconductor

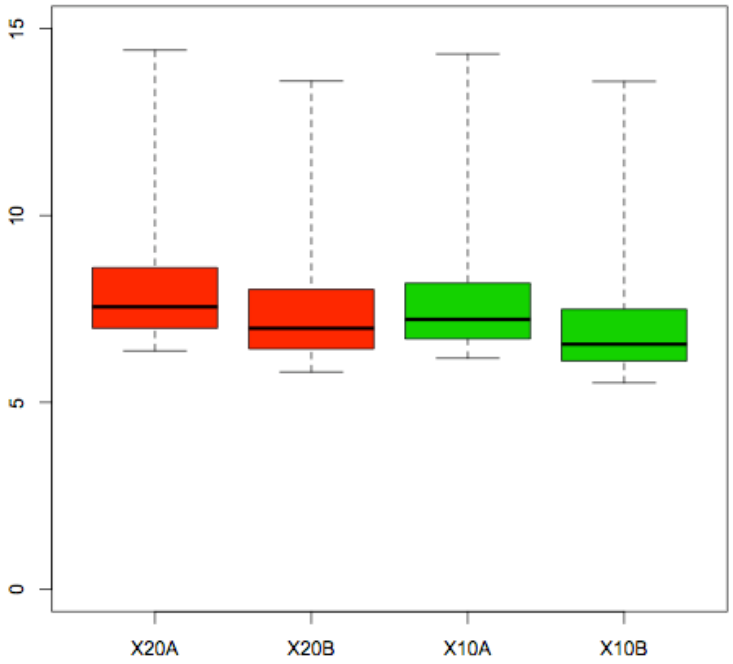
- MAS (Microarray Analysis Suite) 5.0
- RMA (Robust Multi-array Average)
- These two are the most popular methods for preprocessing Affymetrix data. Each method consists of different algorithm at each step of preprocessing.

MAS 5.0: Background correction

- Under R:

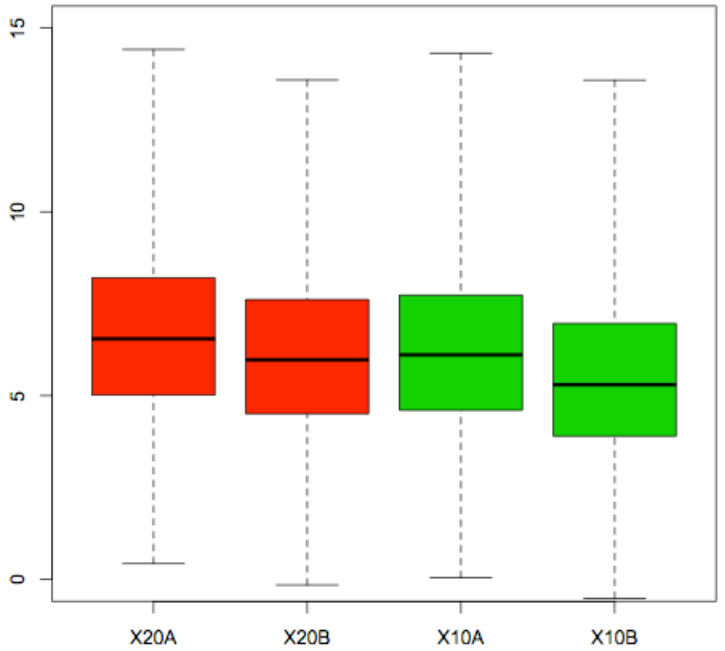
  > a=bg.correct (Dilution, method="mas")

> boxplot(Dilution, col = c(2, 2, 3, 3), ylim = c(0, 15))



Before

> boxplot(a, col = c(2, 2, 3, 3))



After

Normalization

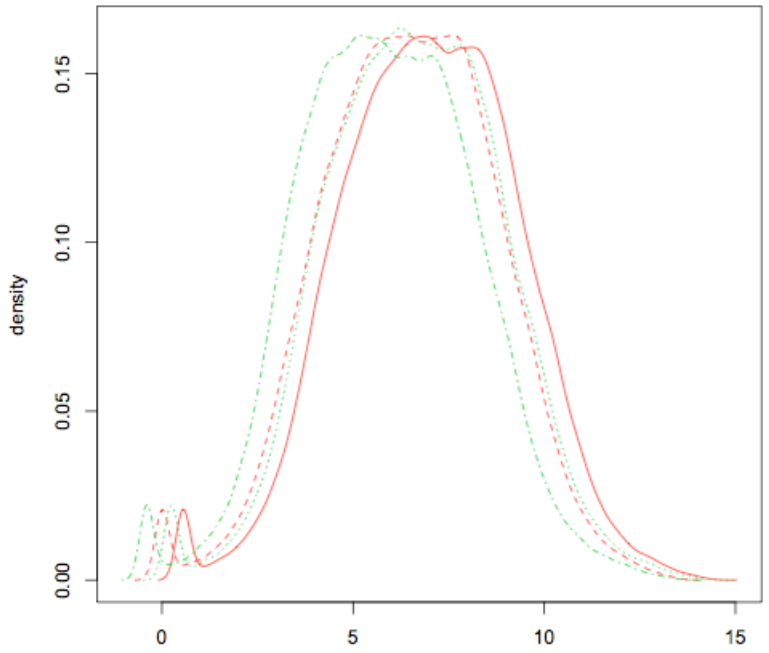
- Most approaches to normalizing expression levels assume that the overall distribution of RNA numbers doesn't change much between samples, and that most individual genes change very little across the conditions.

- If most genes are unchanged, then the mean expression intensity should be the same for each sample.
  - Scaling normalization
- An even stronger version of this idea is that the distributions of expression intensity must be similar.
  - Quantile normalization

MAS 5.0: normalization

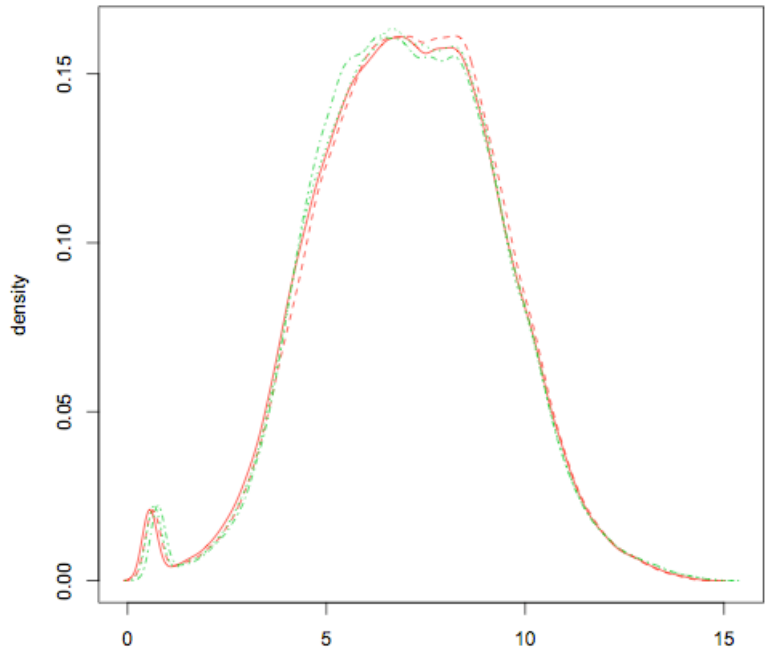
- Under R:

  > b=normalize (Dilution, method="constant")



Before

> hist(Dilution, col = c(2, 2, 3, 3))

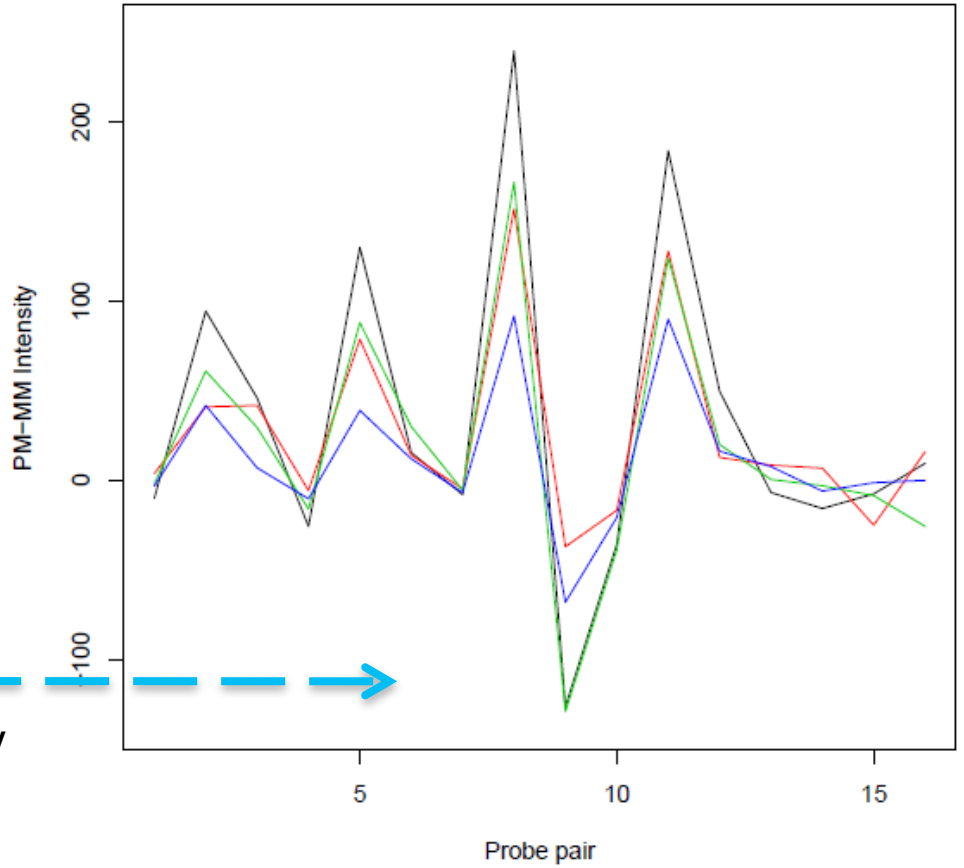


After

> hist(b, col = c(2, 2, 3, 3))

Why PM-MM correction is optional

After correcting (i.e., subtracting) MM intensity, some probe pairs have intensity less than zero!



As the intensity for probes not supposed to be hybridizing to anything (*i.e.*, not expressed) should be 0, what does negative intensity mean?

The "negative expression value" will introduce difficulty in data interpretation and should be avoided

MAS 5.0: PM-MM correction

- MAS introduces **Ideal Mismatch (IM)** computation to replace MM, so that it will remedy the negative impact of using raw MM values.

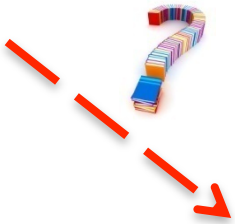
- The goal is to guaranteed the computed IM value to be smaller than the corresponding PM intensity so that it is usable.

- The principle of IM computation is to calculate a robust average of the log ratios of PM to MM for each probe pair in the probe-set k.

Why Summarization

PM	82	75	77	…	85	69
MM	10	18	12	…	17	9

Probe Pairs	1	2	3	…	10	11



Report
Sample Expression Vector

Gene1	68
Gene 2	128
Gene 3	59
Gene 4	88
…	…

Each gene will be measured by multiple (11-20) probes.

The vector of probe intensity need to be summarized into one expression value for its gene.

MAS 5.0: Summarization

- After background correction, normalization and PM-IM correction, we obtain a vector of "processed" (background corrected, normalized, and IM corrected) probe values for a given probeset.

- We need to summarize the vector of probe values into one expression value of the stuided probeset.

- MAS 5.0 use one-step turkey's biweight function to yield a robust weighted mean of probe value, which is relatively insensitive to outliers, to represent the probeset expression value.

expresso()

1. Background correction: weighted average of grid background
2. Normalization: trimmed mean scaling
3. PM-MM correction (optional): Ideal mismatch
4. Summarization: one step Tukey's Biweight function

```
>  set <- expresso(Dilution, bgcorrect.method = "mas",
normalize.method = "constant", pmcorrect.method = "mas",
summary.method = "mas")
```

MAS 5.0

```
> dim(exprs(Dilution))
409600     4
> expression_values <-   mas5(Dilution)
> dim(exprs(expression_values))
12625     4
> write.exprs(expression_values,
file="mymas5data.txt")
```

MAS5: Summary

- Good
  - Usable with single chips (though replicated preferable)
  - Gives a p-value for expression data
- Bad:
  - Lots of fudge factors in the algorithm
- Misc
  - Most commonly used processing method for Affy chips
  - Highly dependent on Mismatch probes

Preprocess methods in Bioconductor

- MAS (Microarray Analysis Suite) 5.0
- RMA (Robust Multi-array Average)
- These two are the most popular methods for preprocessing Affymetrix data. Each method consists of different algorithm at each step of preprocessing.

RMA

- Robust Multichip Analysis
- Used with groups of chips (>3), more chips are better
- Assumes all chips have same background, distribution of values.

1. Background correction: RMA convolution
2. Normalization: quantile normalization
3. PM-MM correction (optional): none
4. Summarization: Fitting probe level model

RMA: background correction

- Only PM values are corrected, array by array, using a global model for the distribution of probe intensities.

- Under R:

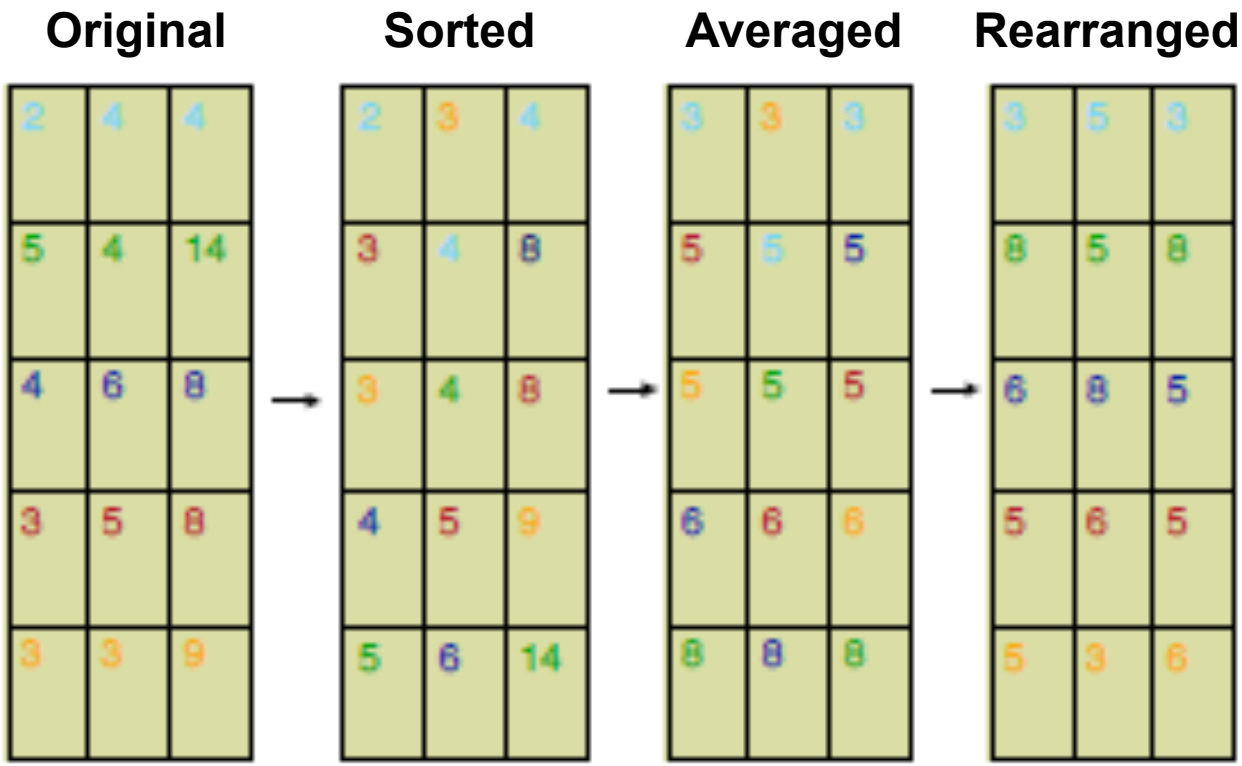
  > bg.correct (Dilution, method="rma")

RMA: Normalization

- Most approaches to normalizing expression levels assume that the overall distribution of RNA numbers does not change much between samples, and that most individual genes change very little across the conditions.

- If most genes are unchanged, then the mean expression intensity should be the same for each sample.
  - Scaling normalization
- An even stronger version of this idea is that the distributions of expression intensity must be similar.
  - Quantile normalization

RMA: Normalization



- 5 genes (A, B, C, D, E), 3 chips
- The final rearrangement step is to ensure we are comparing the expression values of the same gene on different chips.

RMA: normalization

- Under R:

  > normalize (Dilution, method="quantiles")

RMA

1. Background correction: RMA convolution
2. Normalization: quantile normalization
3. PM-MM correction (optional): none
4. Summarization: Fitting probe level model

- Under R

> set <- expresso(Dilution, bgcorrect.method = “rma", normalize.method = "quantiles", pmcorrect.method = “pmonly", summary.method =“medianpolish")

> expression<-rma(Dilution)

RMA: Summary

- Good:
  - Results are $\log_2$ scaled from the raw intensity values
  - Rigidly model based method: defines model then tries to fit experimental data to the model. Fewer fudge factors than MAS5
- Bad
  - RMA cannot be applied to single chip. (need replicates or comparison)
- Misc
  - The input is a group of samples that have same distribution of intensities.
  - Requires multiple samples

Outline

- Background
- Preprocessing of oligonucleotide microarray
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

Homework 6

- Get to know “SpikeInSubset” package
- Using some commands learned

  i.e. exprs(), pm(), boxplot()
- To plot curves for a probe set.
- Do preprocessing
- Using slides of this class, help of R, or other tutorials.