http://sysbiostor.unl.edu/Teaching/BIOS497897_2014/

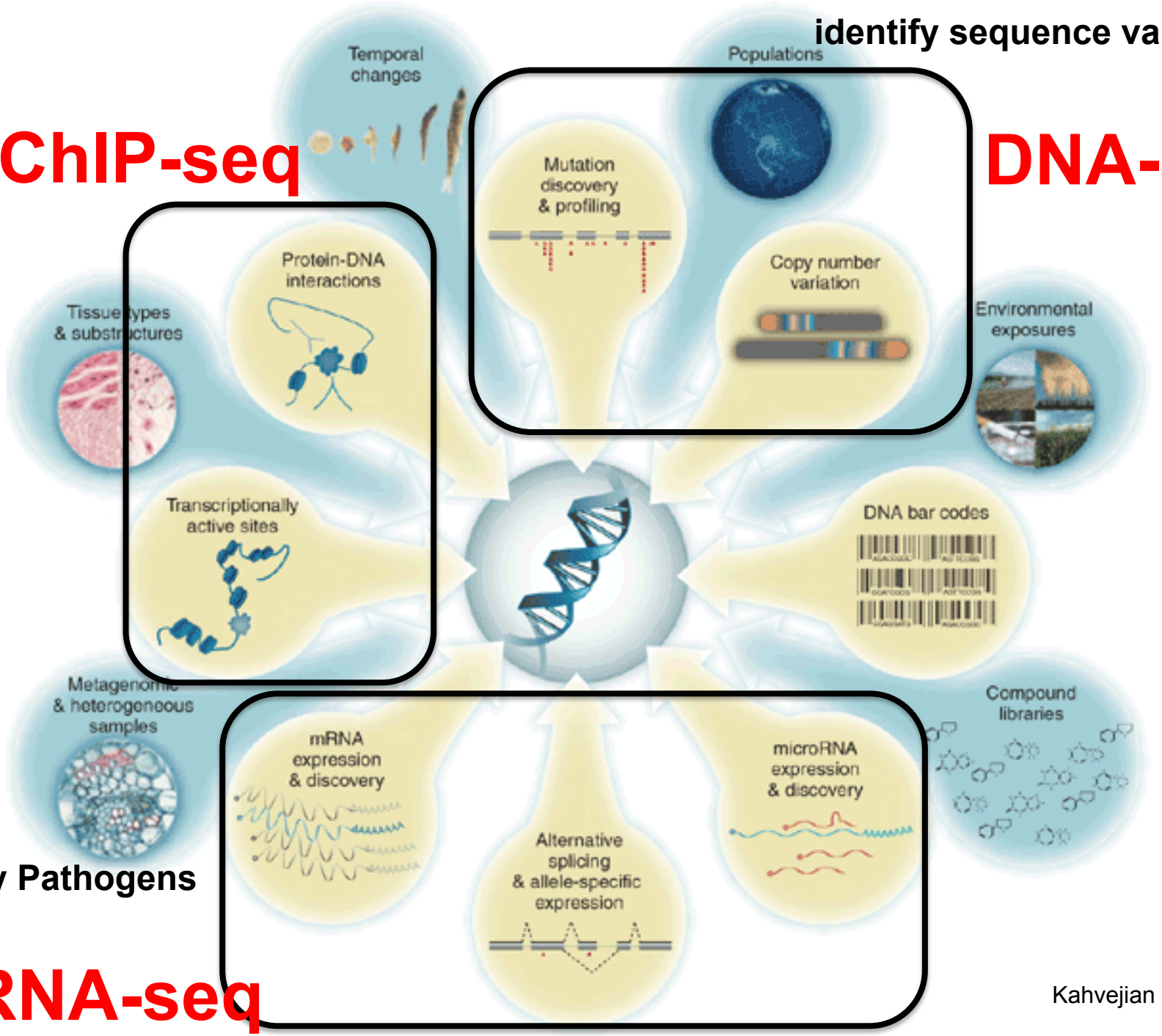
Microarray

Lecture One

identify sequence variations

ChIP-seq

DNA-seq



Identify Pathogens

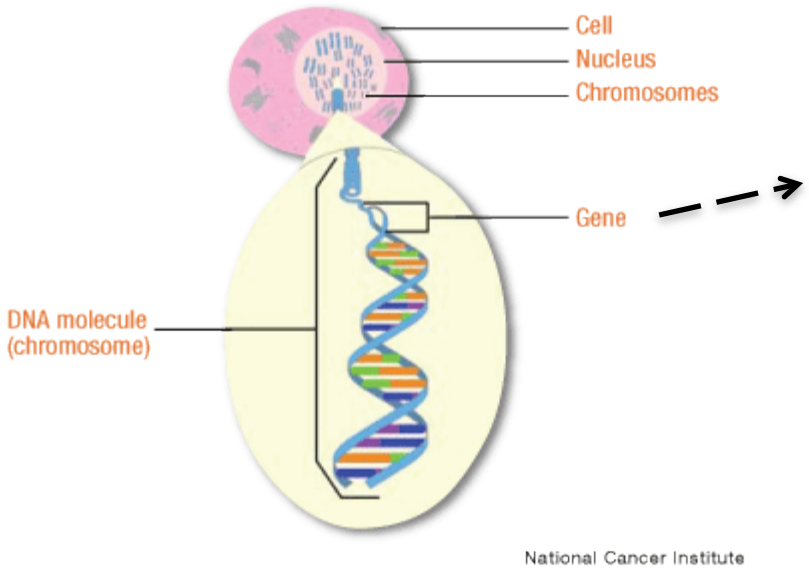
RNA-seq

Kahvejian *et al*, 2008

Outline

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

DNA: "Blueprints" for a cell

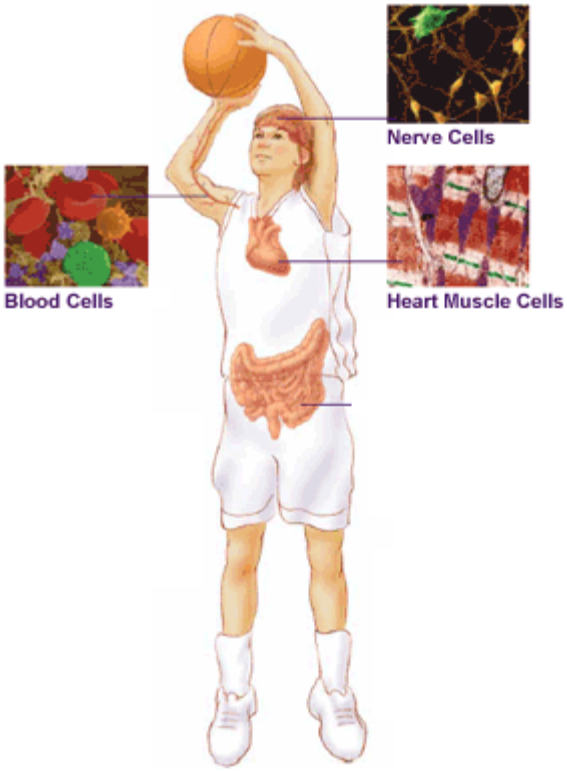


GAATTCTCTTTGGTATCCAATGAAGAAATCGAATCCATACCCATAGCTATAAAAAACAT
TTCAGGAGAAAATAAGACCGAAGCTGCTCAATTAGGCGCAATTGATTCGTTTCAAAAAT
GTGAAACTTGCCAGCTTACTTCGGCATGTCCTGGTCATTTTGGAAAATTTCATCTTACT
CAACCATTATTTAAAGTCGCATTTAAAAAACTTGTTGAAAATATTTTTAAATATACTTG
TTCTTTCTGTGGTGCTTTACAAAATCTTGAACTTCTGGAATTGATCAAGCAGATAGACG
AACGAAATACTGGAATAACAGTTAAAGATCGTGCTGCTTTTAAAAAAATTTTAGAAGCT
ACCAAACAAAGCAAATTCAAGTGTATTGCACCTAATTGCCAAAAACAAGTCTCTCCTTT
ACAATATTCGAAAAATAATAACTTTATATATAATTCGGGTACTACAAAGGGTATAGTTT
TGGATAACAGGCATGTGTTTAATATCTTACAAAATCTTCCACAAACGTTTAAATTATTG
TTAACCCCTTCGAATGCTCATCAAATCGTATCTCCCGAAAATGTCTTTTATGCTAATAG
TATCTTACTTCCACCACATAATCTACGAACTATCAATGTTTATGATGGTCAGGTTACGA
GTTTGTTAACAAGTGATTTGAATCTGATAATGCGAAGAGTTGCTAATAATGAGACAAAT
GCAAAAATACAAAAAATCTTGGATTCTATCGATAACAGCCGAGGTGCCAATCCATATGC
TACAAATAAAAAGCTTACTTTGGATACTTTGACAGGTGGACACTCAAAAGAATCTTATT
TGCGAAGTTATATTAATGGCAAACGTATTCCTGAGACTGCCAGAGCTGTAATCGAACCC
TCTATGAATAAAACTGGCTTTATTGAAGTACCATCTTACATTTTAAACAAGTTAAGAGA
TGTTGTCTTTTTATAATCACGTTACGAAAGATAACATACTCAAAAGTCTTCAAAACGAAC
AAGCTTTTCTAACATATATCAAAAGTGATCATAATTCTGAAAATCCTTATATGGTTTAT
GATTTAGCACAGAAGAATGGATATTTAACCTTGGCTCCTAATTTCGGTGATATTTTCGA
AAAAAGGAAAGAGGAAGGTGGTTTTGTAACTATTTGCAGACATCCATCTATCTGGTTAA
CTAATATCCAATCTGGTATAATAAAAAGATCAGAAGGGTTTACTATTAACATCCCAACC
ACAATTTGCACATCTTTTAATGCTGATTTTGATGGAGATGAGATGACAATATATTCTTT
CAAATCCCCATGTGCCAATCTCGAACAAGCTTTGATTATGAACTCACGAAATCTCTTCA
AAAATTCTATAACAAGCAATCCAATGTTCGGCTTGGTCCAAGATCAAATACCAGCCTTG
AATAAGTTATATAGACGACAAAATTATACATATAACGATGCGTTGGTGATTTTAGGACA
ATTCGGATTTCTGTTAACACCTGGAAAAGATAATTATACCGGAAAAGATATACTTTCTT
GTGTATTCCCAAAACATTATACACTCAAAGGAATTGTTGAAAATGGCGAACTTATTTTG
GAGAATTTTACAAATAAACTCGTTTCCGCAAATTCCTCAAAGTCCATCTTTGGGCATCT
TGTTTTATTTTATGGACAAGAGTATGGTTTGACTATATTGGATACAATGCGAGATATTG
TTCAAAATTTTATTACACATTTTGGTTTCAGTGTAAAAATCCGAGATATGATCCCAAGC
CCAAAAATTTTGGATATTCTAGAAAAGATCGTAGACCAAGAAGTGGATAAAATTGATAA
ACAAACAAAACTTCTATATGACGATATCGAACAAGGTAAGGTTATAATCAACTCTTATG
ATGATATTTCTGAGTTCAGATTAAAAAATGTGGCTATTATGAAAAAGAAACTAGAAAGC
AAACTTTTGGAACTTTTGGATGAATATTATGATGAAGACAATAATTTCCTAGAGATGTA
TAGAACGGGATATAAGGTCAACATTAACGAACTTCTCTCTATTATGTGTTTCTCGGGTT
TTAAAAATTATGGAAATATCGAAATGATTACACCGGGTCTTAATGGTAAAACATCTTTG
TTTAGCTTACCAGATTCTATAAACTTACAAGATTATGGGTTCATCAAAAGCTCTATTGC
CAAAGGGTTAACGTTTGAAGAATATGCTACAATCGTAAAACAAGAAGCTTTTCCACAAA
TTGTTAATGTTACAACTGGTACTTCACAAACAGGATTTTTGGGGAAAAAAATGGTTAAA
ATGGCTTCTGAATTC

- Each human cell has **identical** genetic information – a total of 3 billion DNA base pairs, including 25,000 genes

Why are cells different?

- The trillions of cells in human body are organized into >200 major tissue types, each customized for a particular role, for example
  - Red blood cells carry life-giving oxygen to every corner of your body.
  - Nerve cells sling chemical and electrical messages that allow you to think and move.
  - Heart cells constantly pump blood, enabling life itself.



Studying the Expression of Groups of Genes

- A major goal of biologists is to learn how genes act together to produce and maintain a functioning organism.
- Large groups of genes are studied by a systems approach.
- Such approaches allow networks of expression across a genome to be identified.

Transcriptome

- Transcriptome: How to genome-wide measure the expression of those genes? How to get the gene expression profiles.
- **gene expression profiling** is the measurement of the expression of thousands of genes at once, to create a global picture of cellular functions.
- These profiles can distinguish between cells that actively dividing, or show how the cells react to a particular treatment.
- Genome-wide expression studies can be carried out using **RNA-seq** or **microarray assay**.

DNA Microarray

- Gene (DNA sequence) and its expressed product (RNA sequence) containing complementary base pairs have a natural tendency to **specifically** hybridize together

  ...AAAACGCTTT...

  ...UUUUGCGAAA...

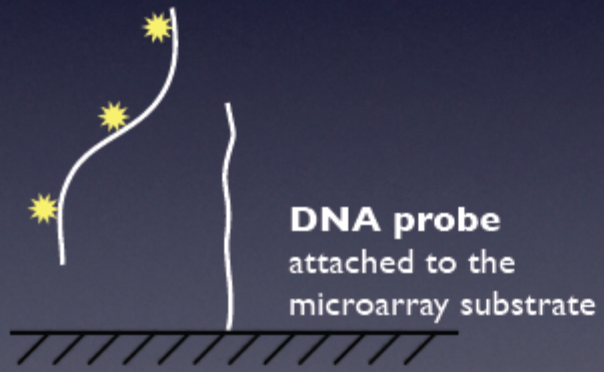
- To measure the expressed product (RNA) of a gene (DNA), we can build a specific probe to recognize it (RNA) using its complementary sequence (DNA).

DNA Microarray

- An oligonucleotide microarray is a microarray whose probes consist of synthetically created DNA oligonucleotides (short sequence of nucleotides).
- A probe (for a gene) is chosen to match a portion of its target mRNA transcript that is unique to that sequence.
- The dominant platform is affymetrix genechip
  - http://www.affymetrix.com

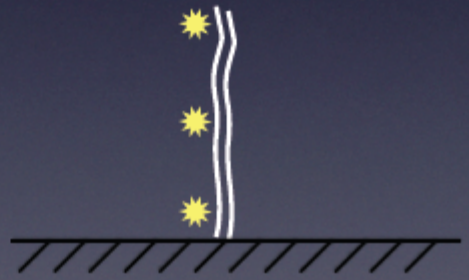
How DNA microarrays works

Labeled target
nucleic acid derived from
a biological sample



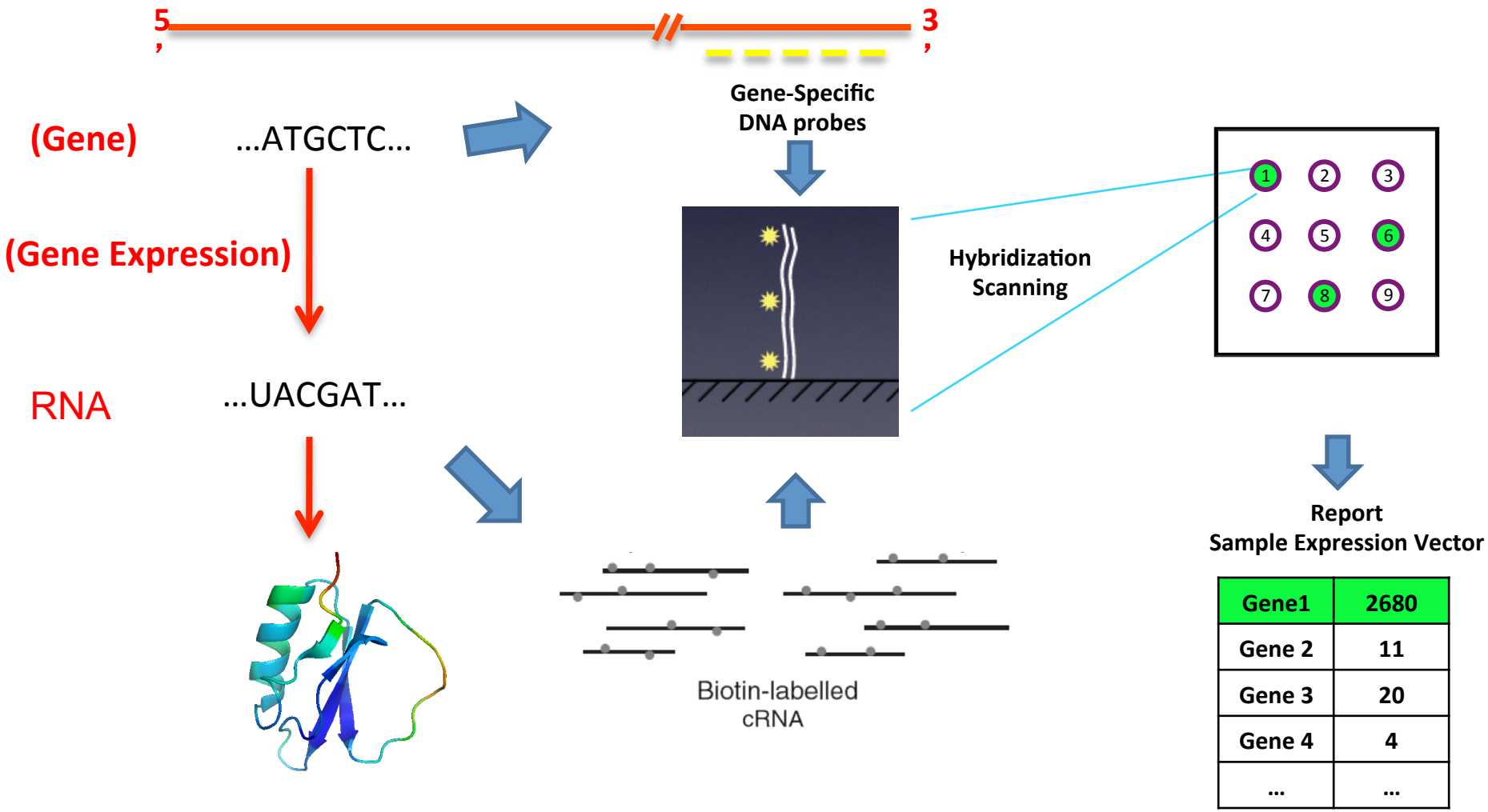
DNA probe
attached to the
microarray substrate

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



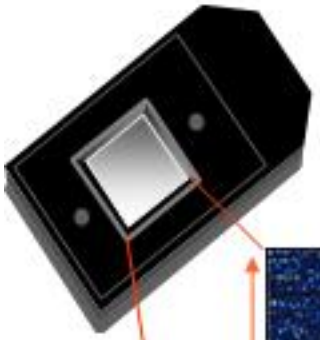
H. Jarmer 2011

Microarray

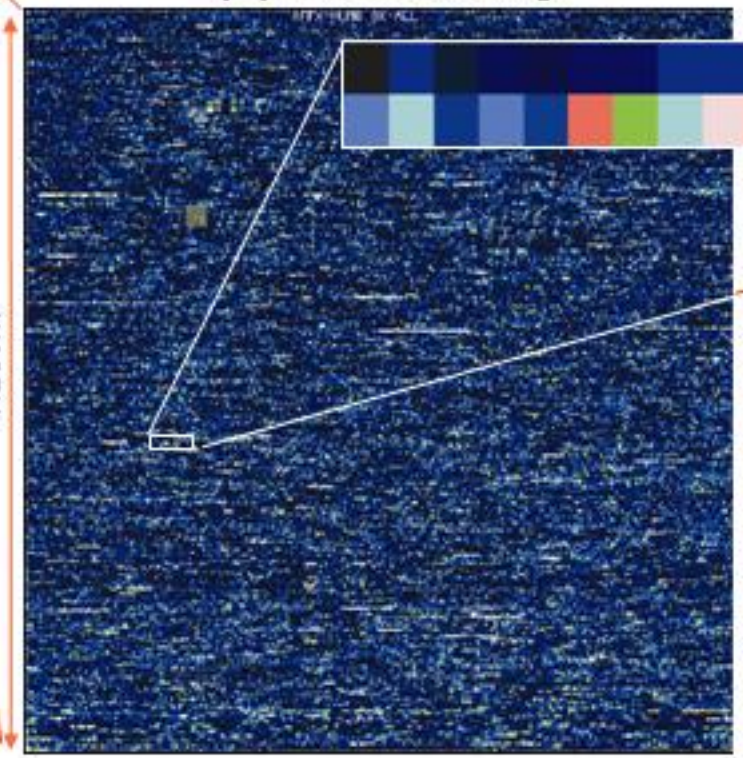


Gene1	2680
Gene 2	11
Gene 3	20
Gene 4	4
...	...

Human Genome U133A GeneChip® Array



(1) Probe Array



(2) Probe Set

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

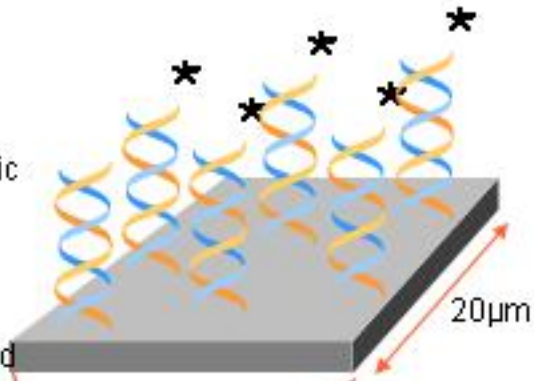
(3) Probe Pair

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

(4) Probe Cell

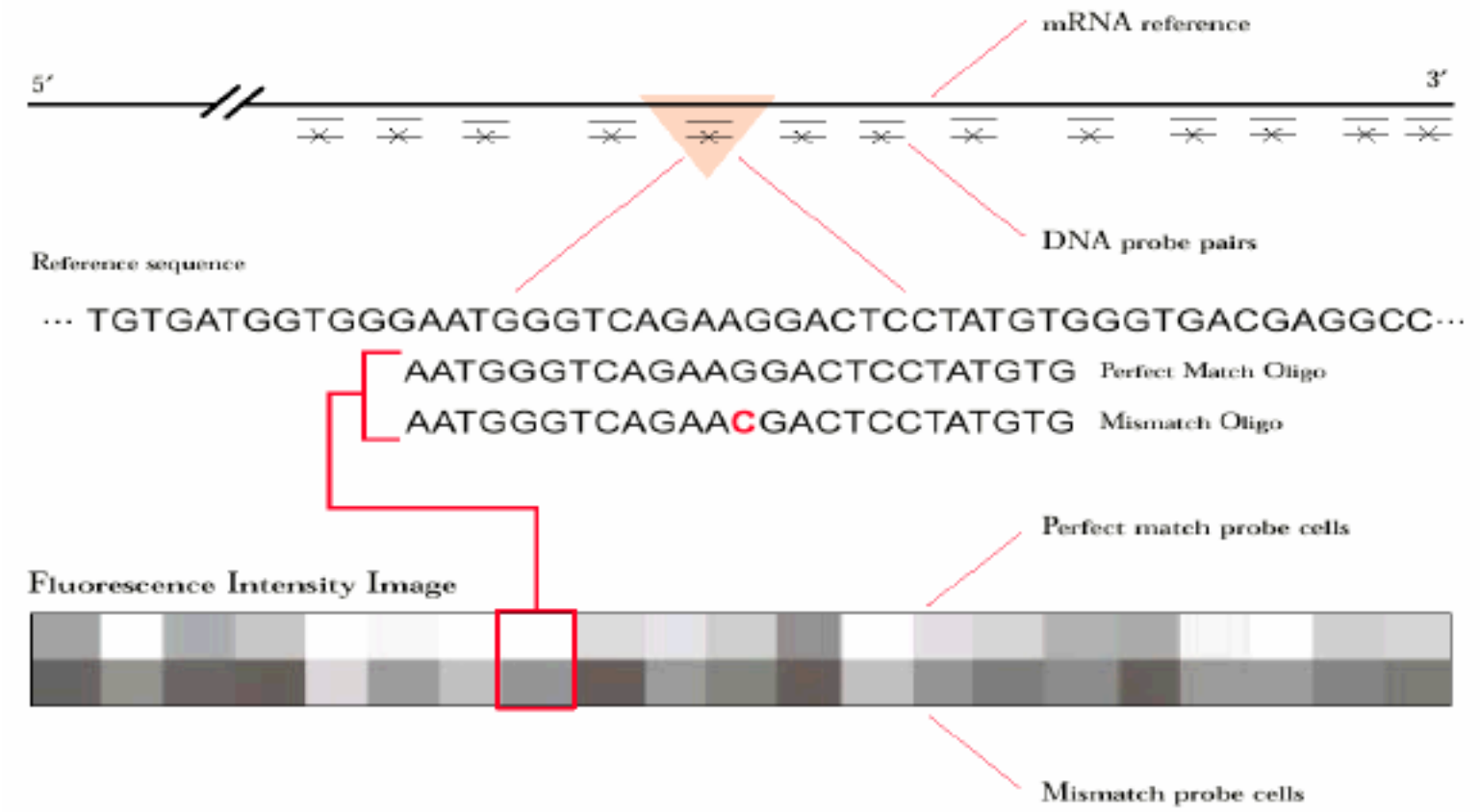
Each Probe Cell contains ~40×10^7 copies of a specific probe complementary to genetic information of interest

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

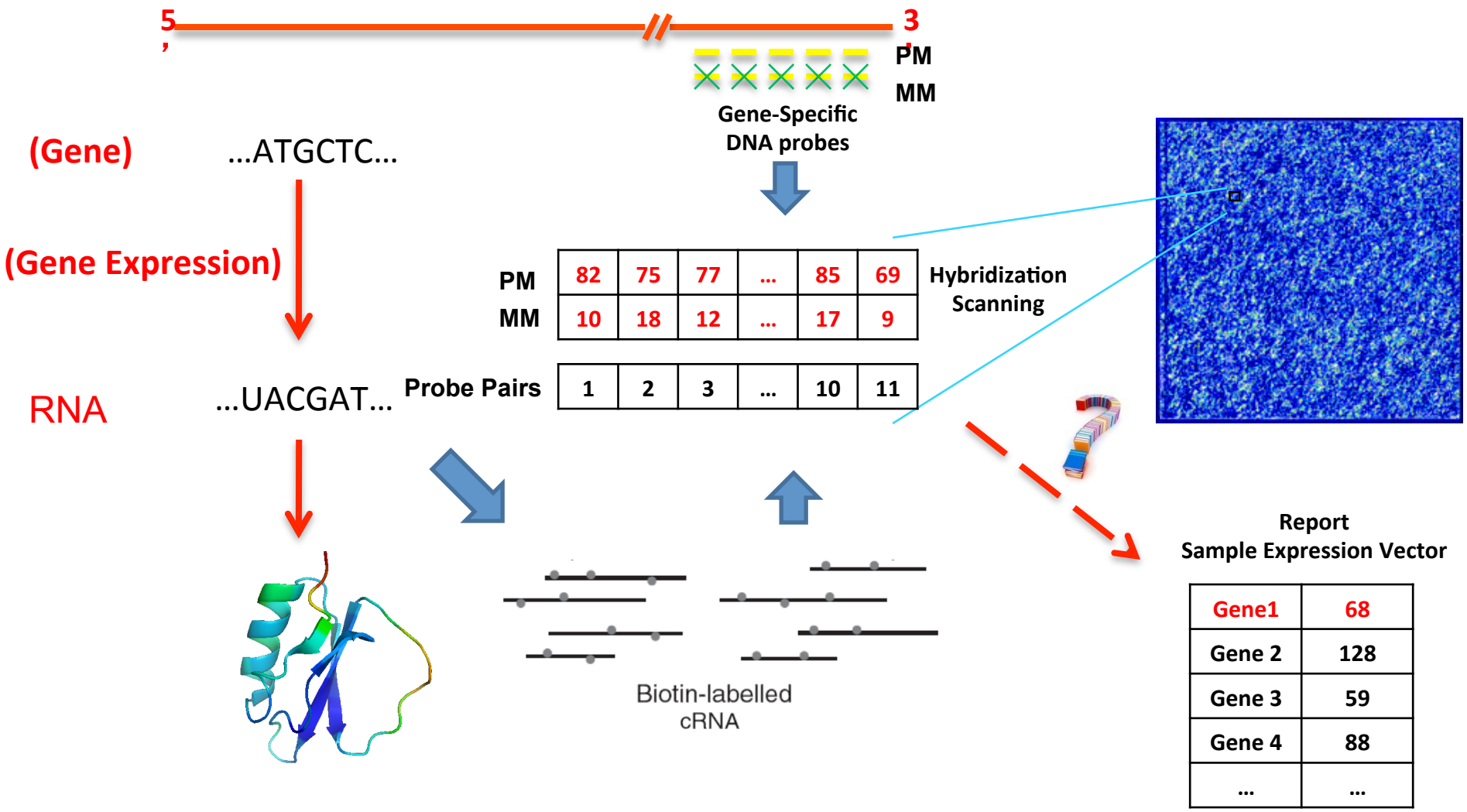


- **A probe set, consisting multiple (11-20) probe pairs, is used to measure mRNA levels of a single gene.**
- **Each probe pair contains a perfect match (PM) probe and a mismatch (MM) probe, each with 25 nucleotides in length.**

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



Biconductor

- There already exists an extensive package of microarray analysis tools, called BioConductor, written in R.
- R and BioConductor are open source and free.
- Where is it?

  http://www.bioconductor.org
- Installation

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

  > biocLite()

  > bioLite("affydata")

Bioconductor packages for Affymetrix data

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

  > library("affy")

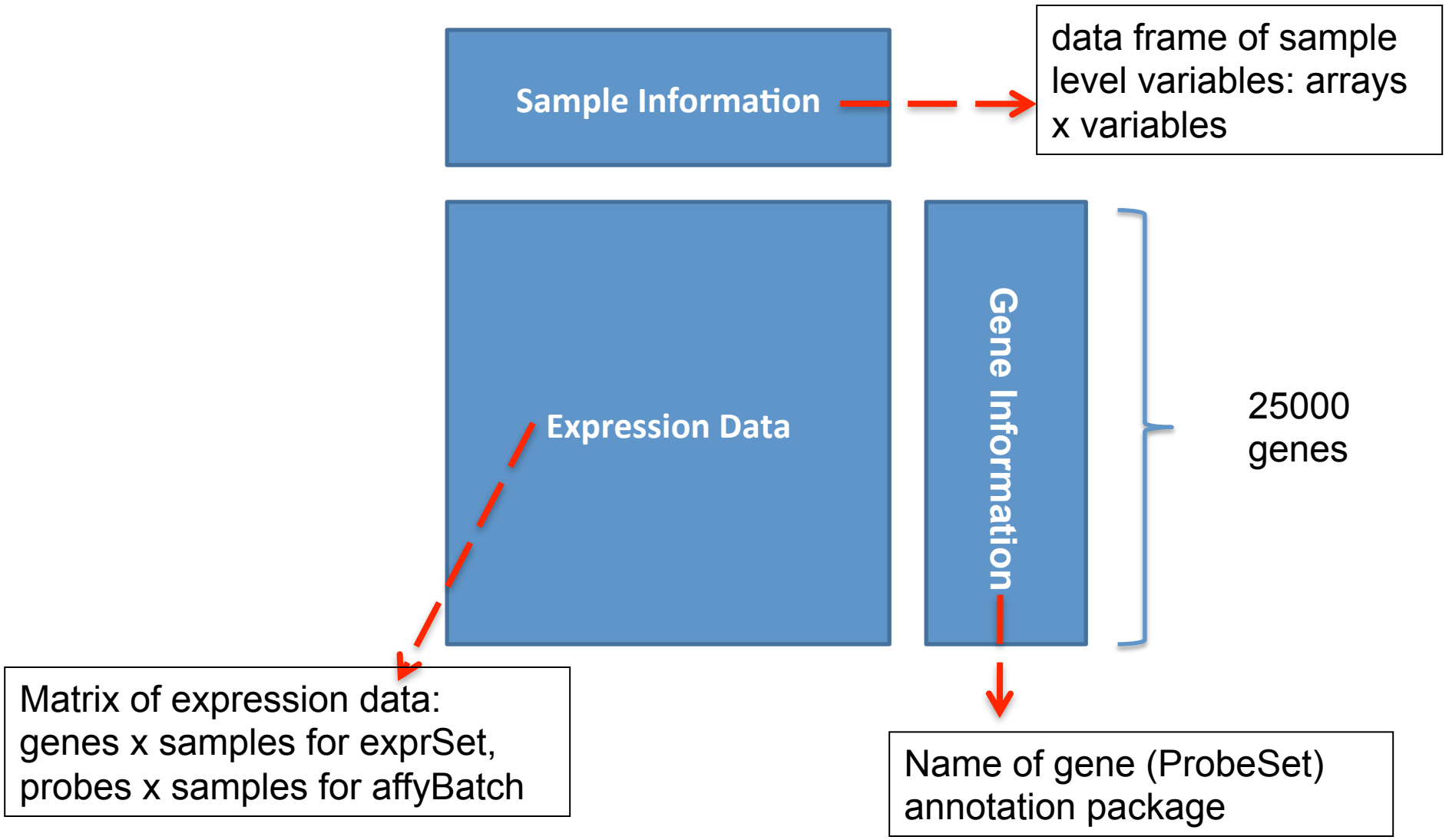
- affydata: Affymetrix data for demonstration purposes

  > library("affydata")

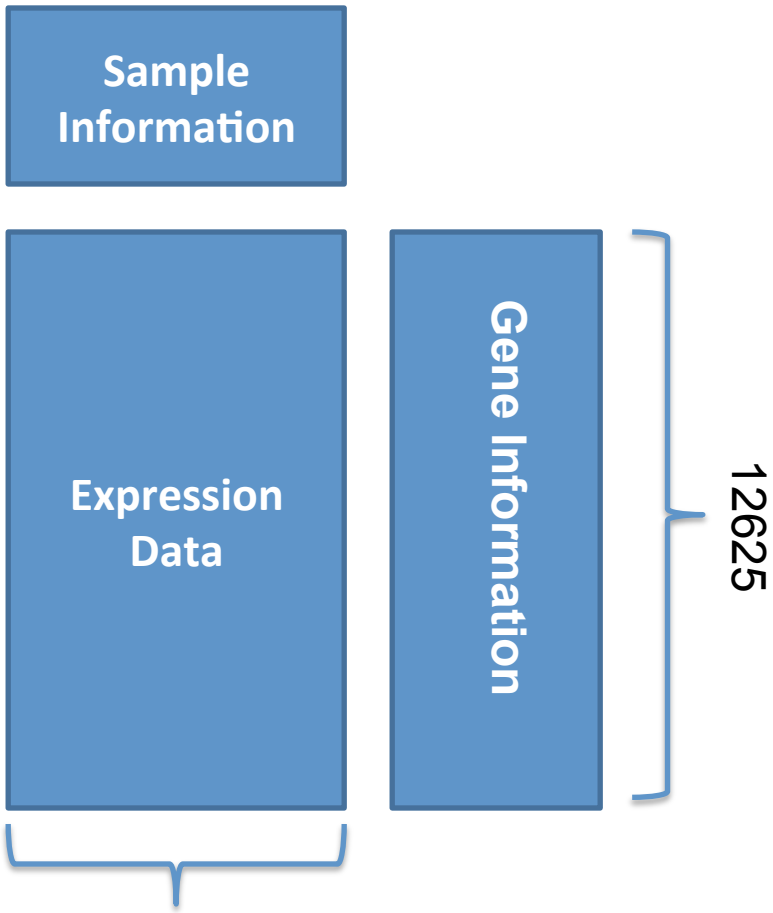
  > data(Dilution)

  - 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)



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

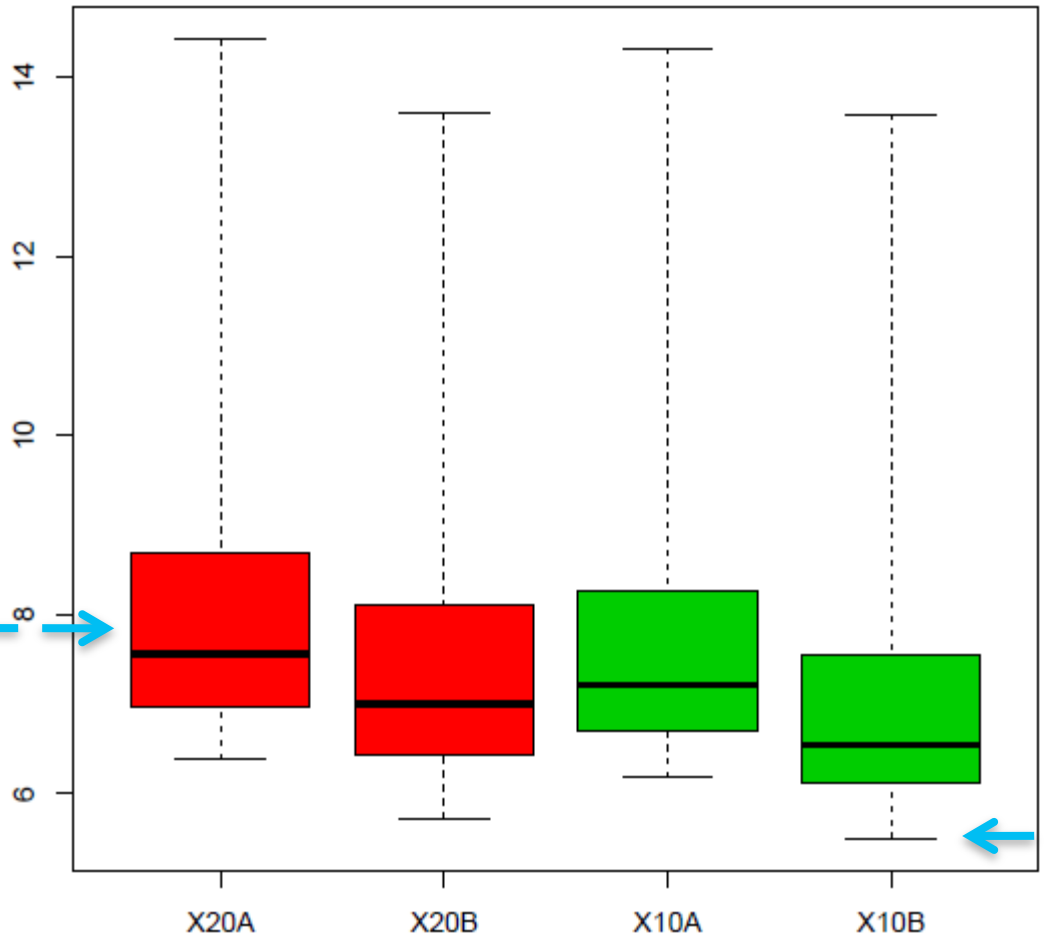
> all_exprs_data[,1]

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

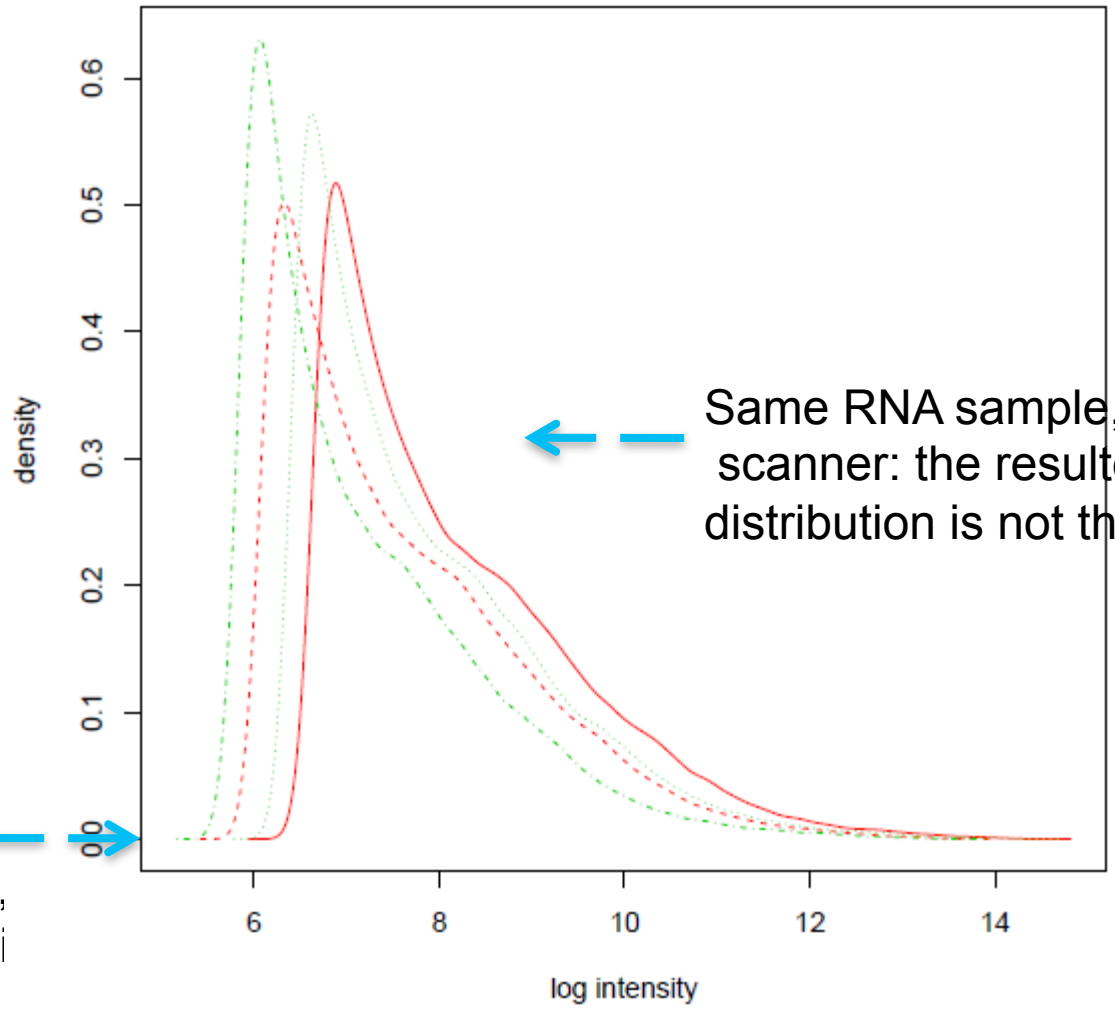


The distribution is not the same, why?

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

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

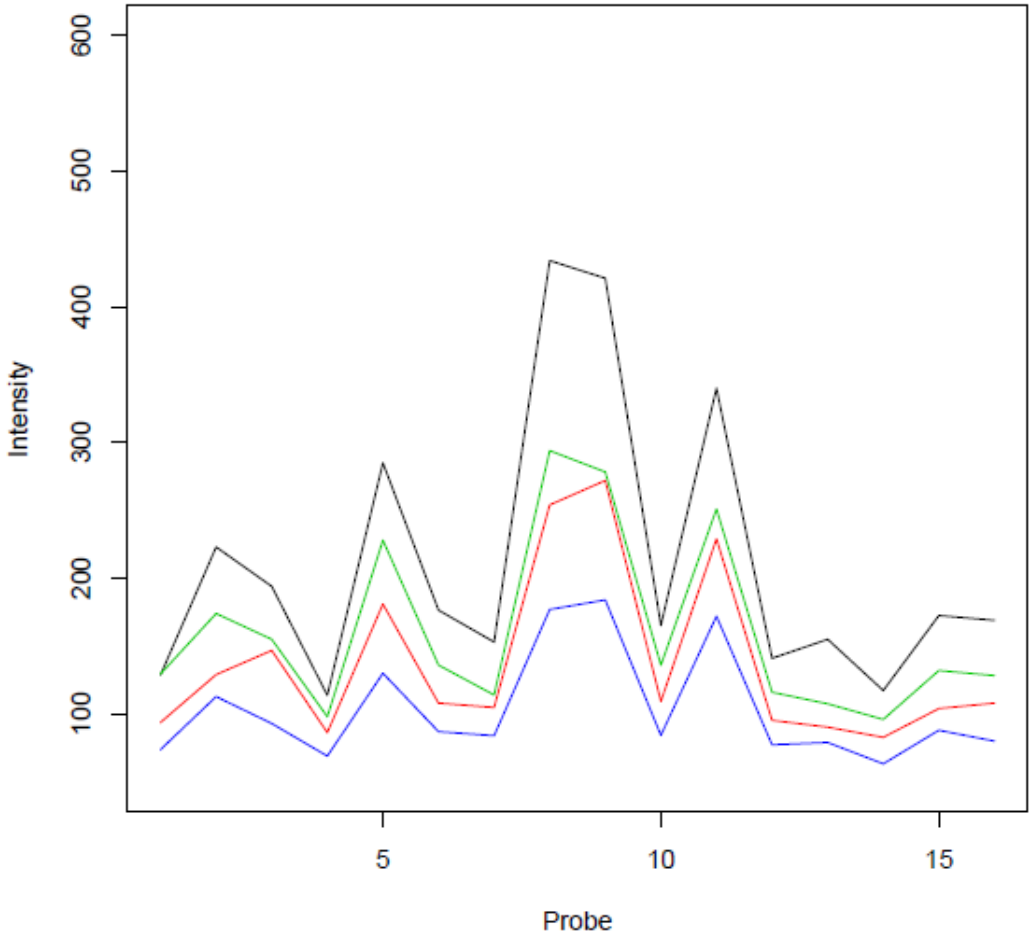
Individual probe set name and data

- The affy package can extract individual probe set name and data from a complete AffyBatch object.

> geneNames(Dilution)

```
> geneNames(Dilution)[1:5]
"1000_at"   "1001_at"   "1002_f_at" "1003_s_at" "1004_at"
> pm(Dilution, "1001_at")
             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_at15 172.5 104.0 132.0  88.0
1001_at16 169.0 108.0 128.3  80.0
```

Expression Data: individual probeset



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

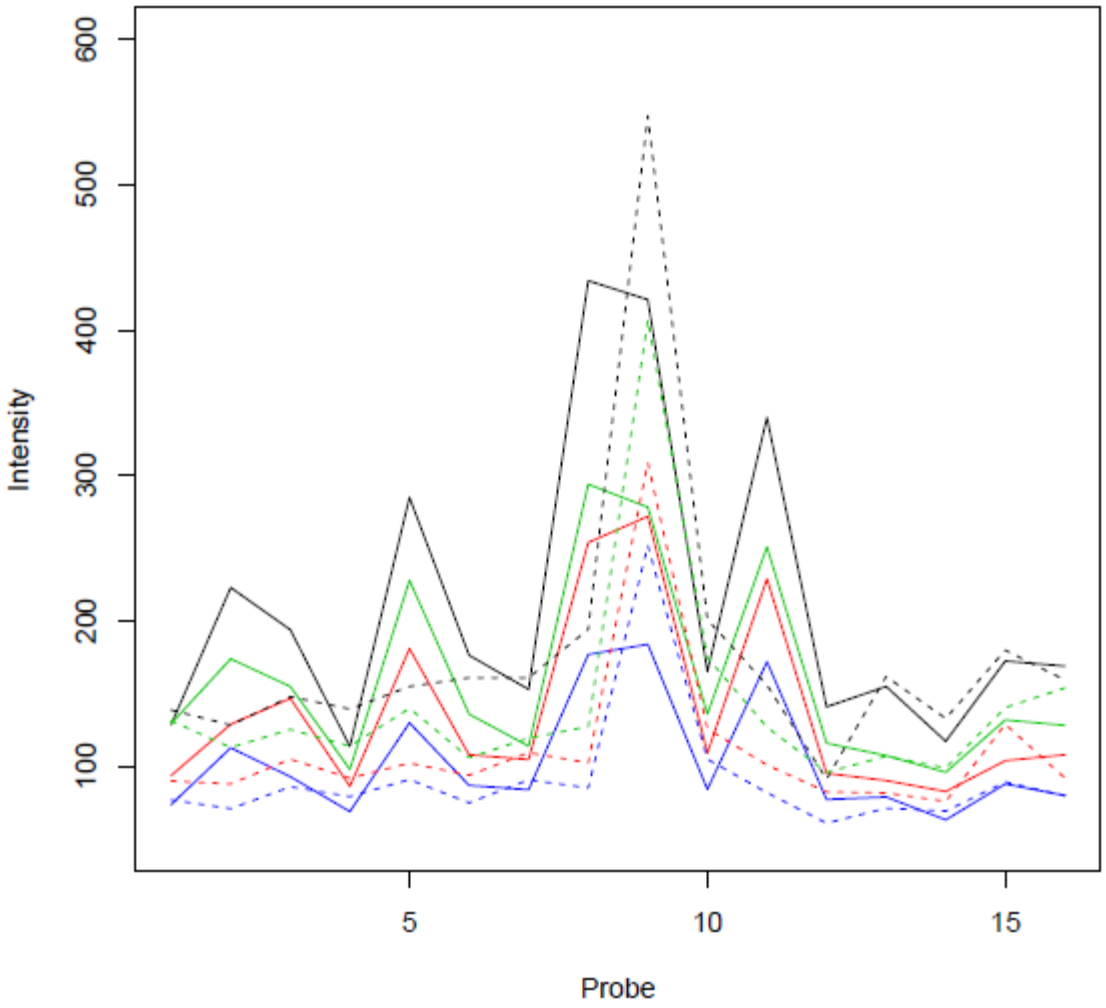
```
> plot(c(1,16), c(0, 800), type='n', xlab='Probe', ylab='Intensity')
> for (i in 1:4)   lines(pm(Dilution, "1001_at")[,i], col=i)
```

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
```

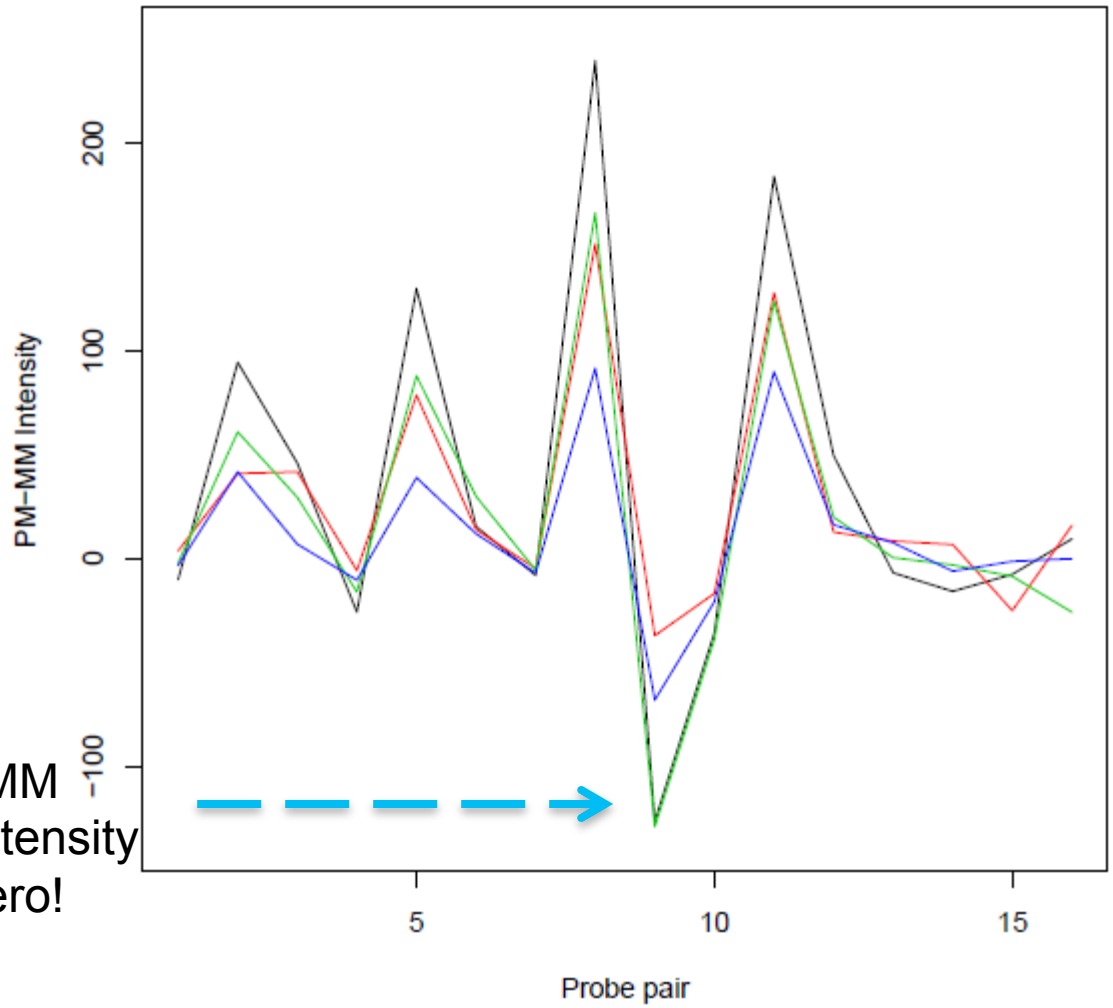
Expression Data: individual probeset



- 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

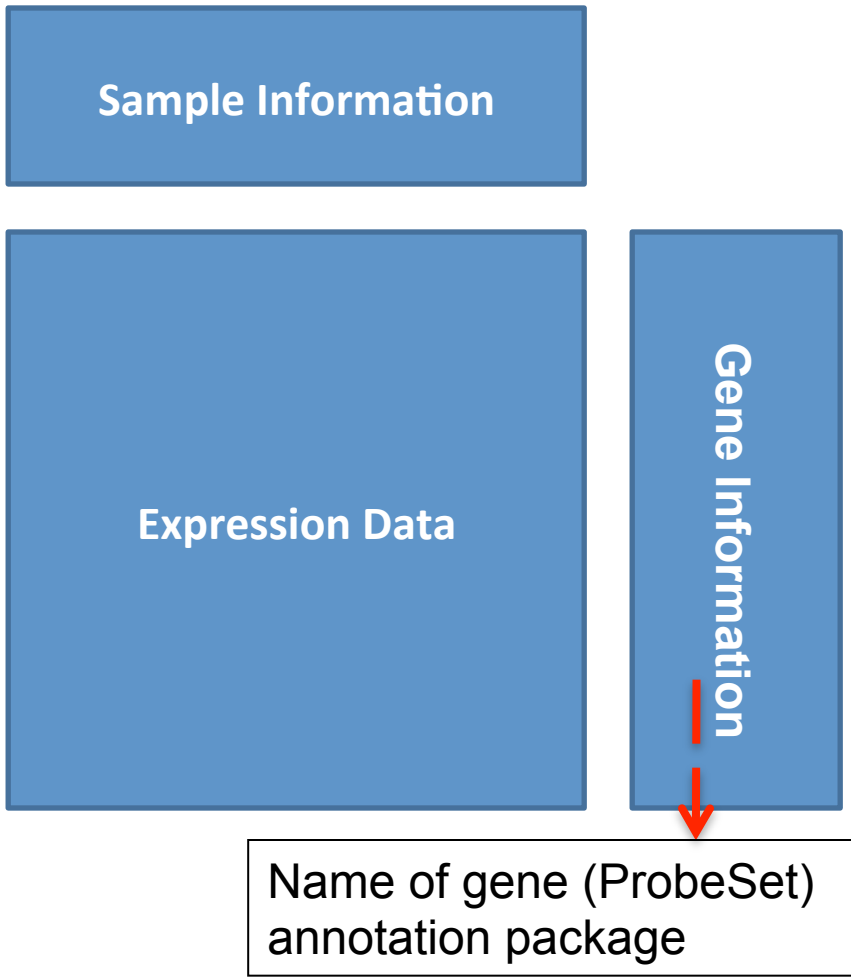


Some PM-MM pair have intensity less than zero! Why?

- 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)

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=
```

Annotation package

Installation:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("hgu95av2.db")
```

```
> library("hgu95av2.db")
> hgu95av2()
```

This package has the following mappings:

hgu95av2ACCNUM has 12625 mapped keys (of 12625 keys)

…

hgu95av2GENENAME has 11725 mapped keys (of 12625 keys)

…

Gene Key Map

```
> hgu95av2GENENAME
GENENAME map for chip hgu95av2 (object of class "ProbeAnnDbBimap")

 >  mappedkeys(hgu95av2GENENAME)[1:3]
"1000_at"   "1001_at"   "1002_f_at"
```

Methods for manipulating the keys of a Bimap object

```
 > as.list(hgu95av2GENENAME[1:3])
 $`1000_at`
 [1] "mitogen-activated protein kinase 3"
 $`1001_at`
 [1] "tyrosine kinase with immunoglobulin-like and EGF-like domains 1"
 $`1002_f_at`
 [1] "cytochrome P450, family 2, subfamily C, polypeptide 19"
```

Homework 6

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

  i.e. exprs(), pm(), geneNames(), boxplot()
- To plot curves for a probe set.
- Using slides of this class, help of R, or other tutorials.
- Due by March. 3rd, 11:59PM

Outline

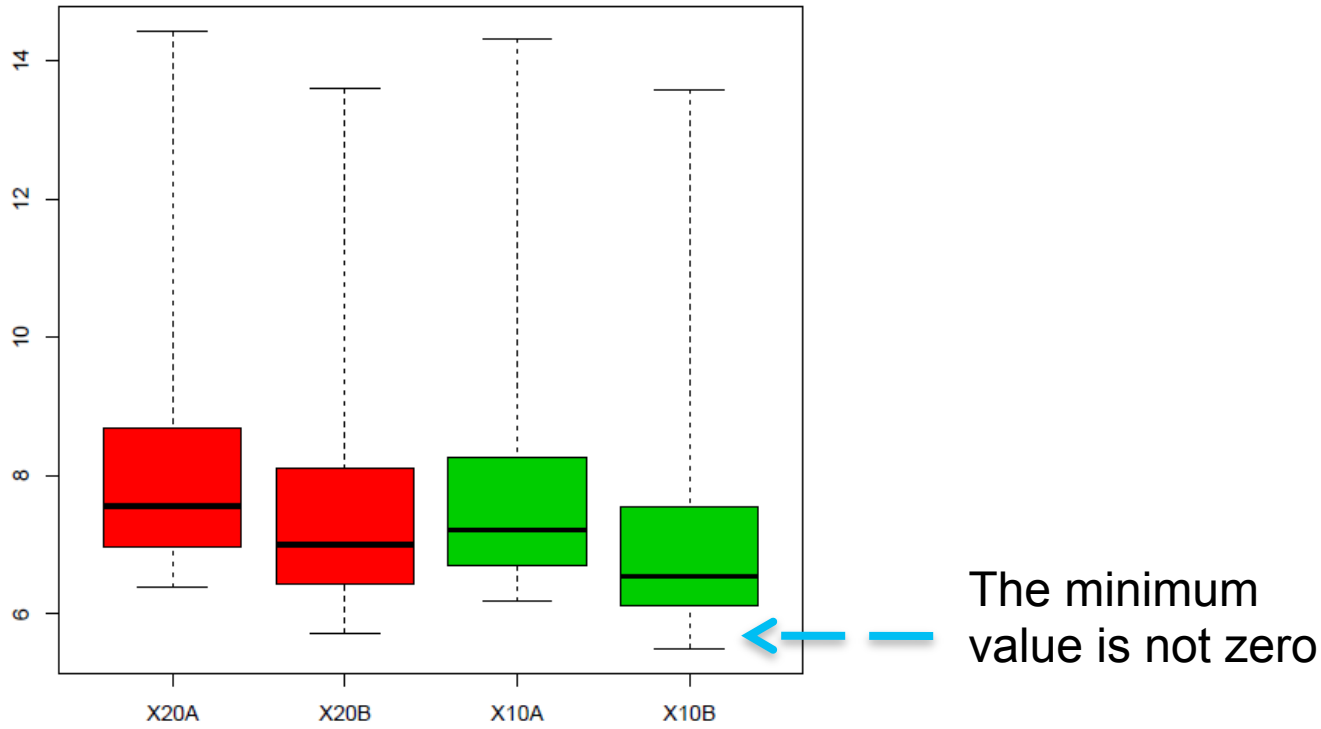
- Background
  - Biology Background
  - Introduction to useful packages in Bioconductor
- Preprocessing of oligonucleotide microarray
- Differential Expression Testing
- Multiple Testing Procedures
- Data Visualization

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

Why Background correction



As many of the probes are not supposed to be hybridizing to anything (*i.e.*, not all genes are expressed), many intensity measurements should be 0.

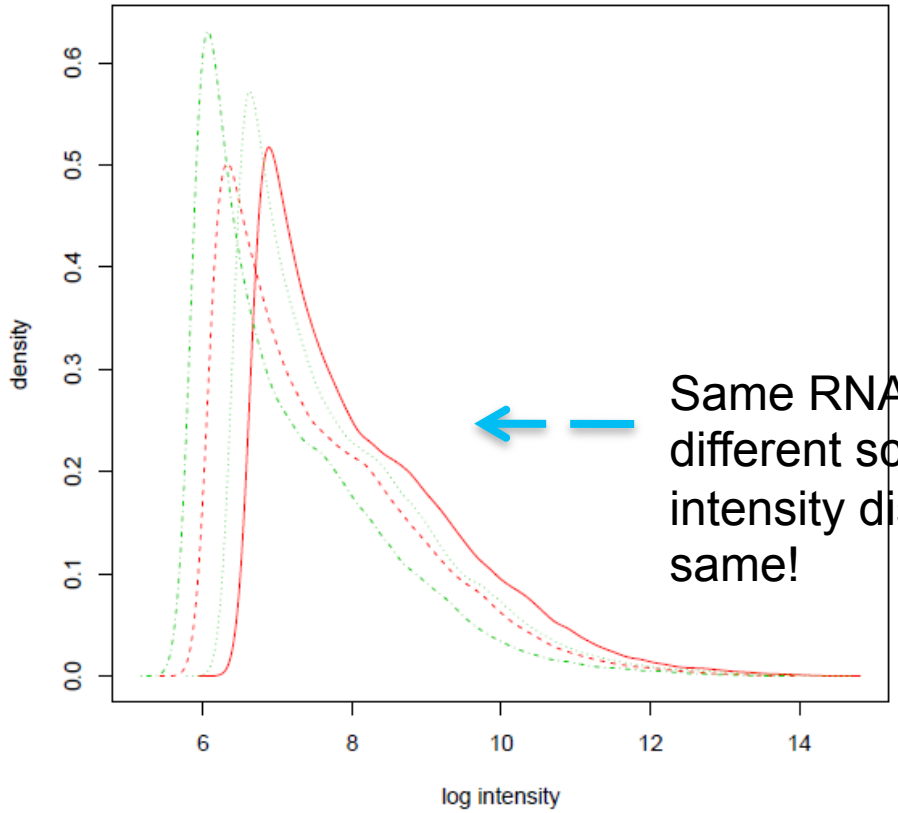
The plot shows the existence of background noise.

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

Why Normalization



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

The plot shows the distribution of raw intensity across different microarrays are not the same.

Normalization (to the same scale) is needed to ensure that differences in intensities are indeed due to differential expression, and not some printing, hybridization, or scanning artifact…

Why normalization

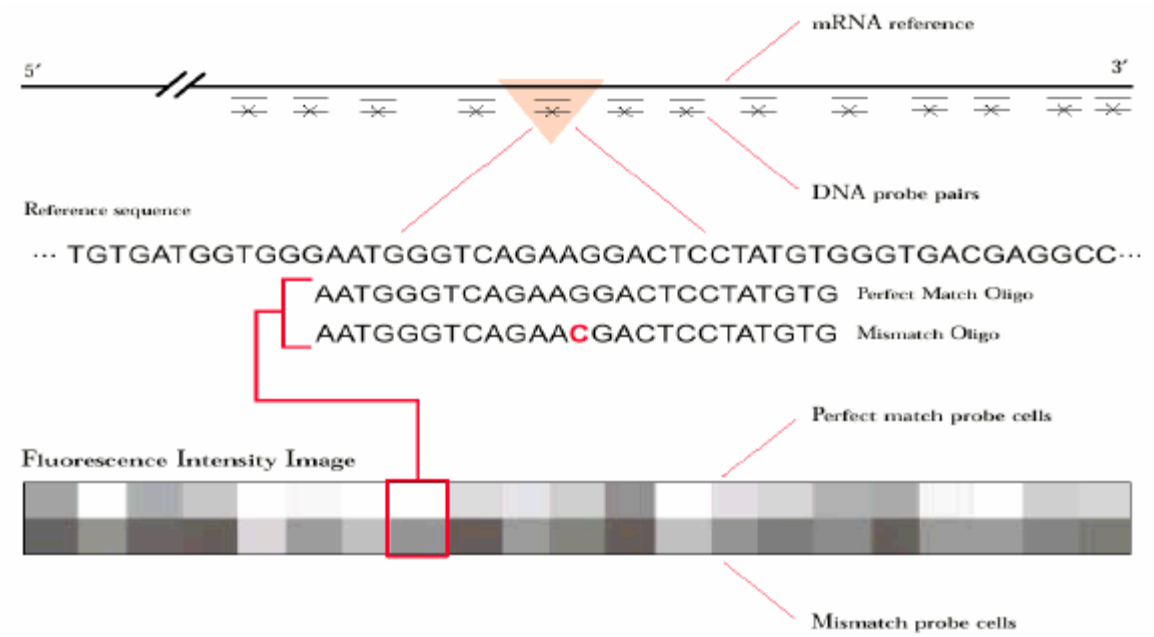
- Biologists have long experience coping with systematic variation between experimental conditions (technical variation) that is unrelated to the biological differences they seek.
- Differences in treatment of two samples, especially in labeling, in hybridization and in scanning, bias the relative measures on any two chips.
- Normalization is the attempt to compensate for systematic technical differences between chips, to see more clearly the systematic biological differences between samples.

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

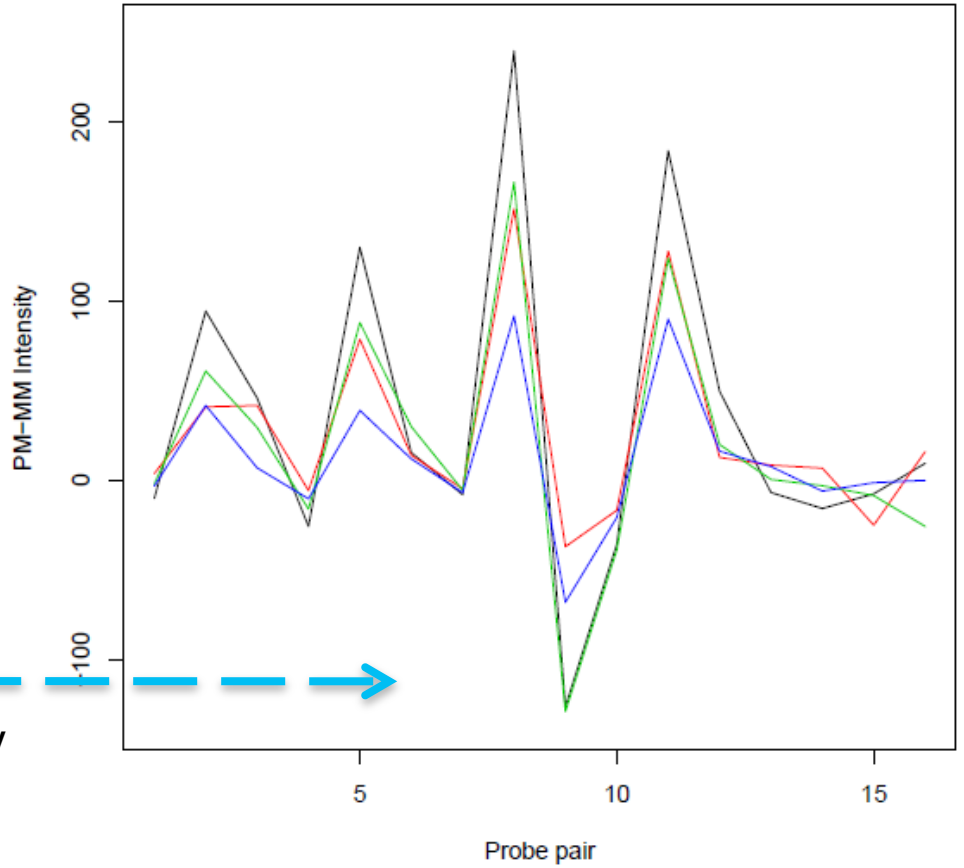
Why PM-MM correction



- **Each probe pair contains a perfect match (PM) probe and a mismatch (MM) probe, each with 25 nucleotides in length.**
- **The purpose of using MM probe is to remove non-specific hybridization**
  - **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**

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

Pre-processing affy microarray

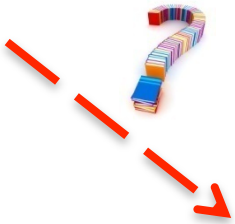
BioConductor breaks down the pre-processing of Affy microarray into four steps. Different algorithms can be chosen at each step. It is highly likely that the pre-processing results will change depending on the choices at each steps.

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

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.

Preprocess methods in Bioconductor

- The affy package provides a number of statistical methods for the preprocess of Affymetrix data

```
> library(“affy”)

> bgcorrect.methods()
"bg.correct" "mas"       "none"      "rma"

> normalize.methods(Dilution)
[1] "constant"        "contrasts"       "invariantset"    "loess"
[5] "methods"         "qspline"         "quantiles"       "quantiles.robust"

> pmcorrect.methods()
"mas"       "methods"   "pmonly"    "subtractmm“
```

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.