Statistics for Human Genetics and Molecular Biology
Lecture 23: Processing Microarray Data

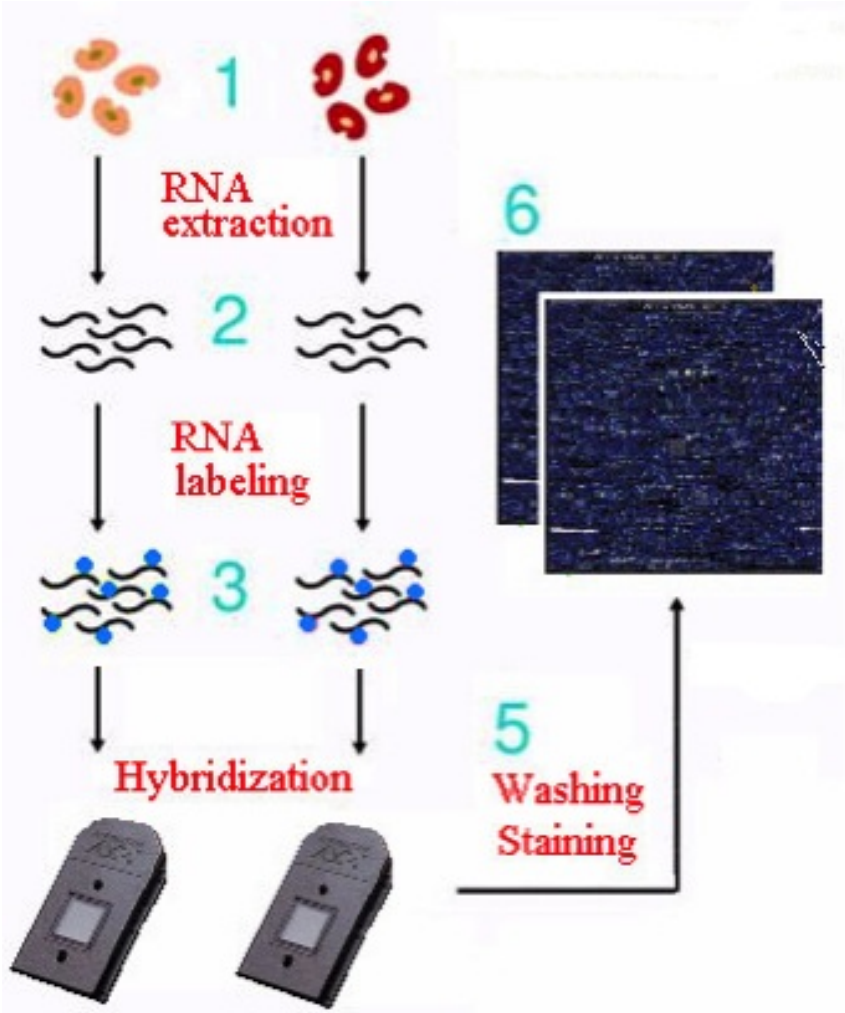
Dr. Yen-Yi Ho (yho@umn.edu)

Oct 30, 2015

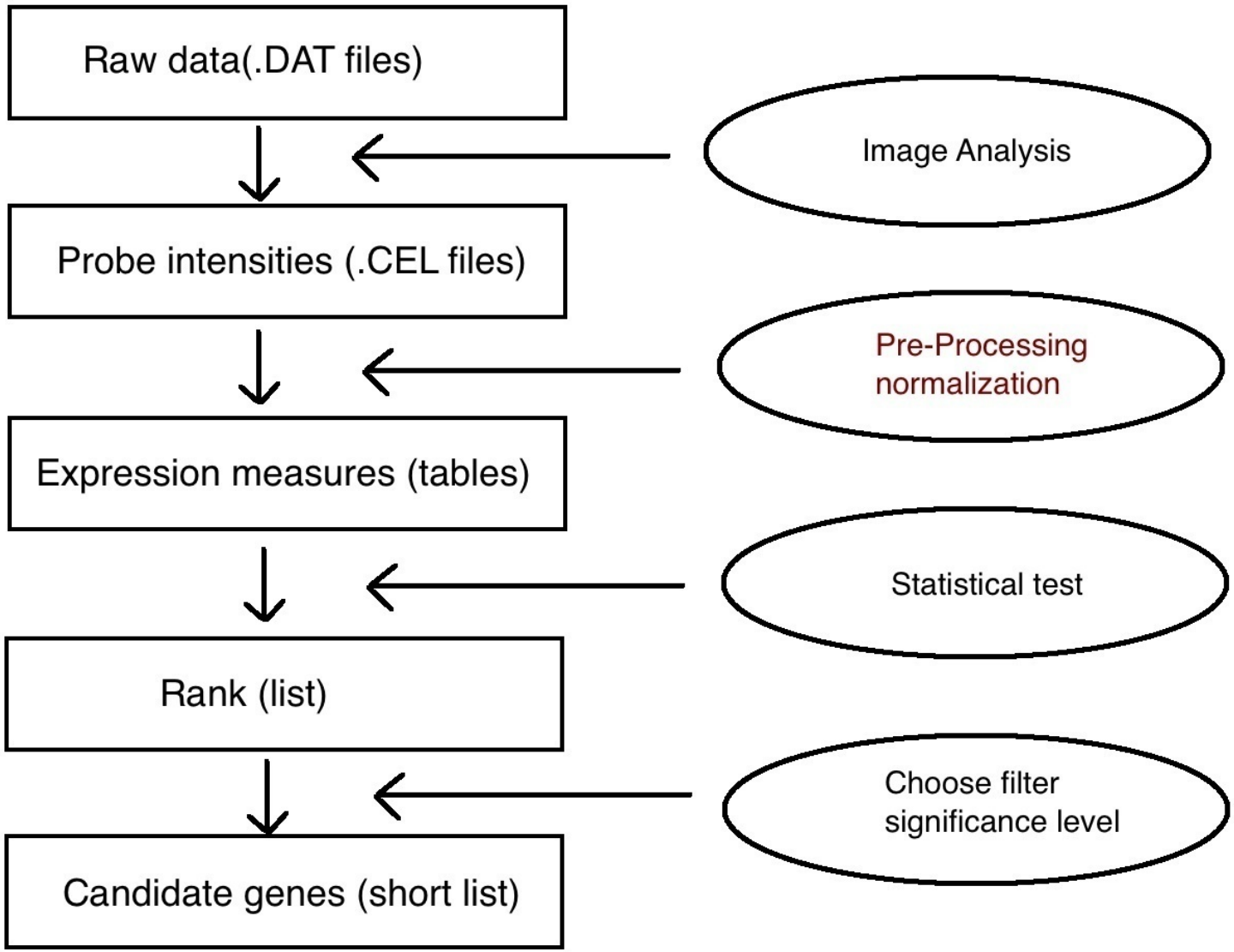
Objectives of Lecture 23

- Structures of Genomic Data
- Quality Assessment

Affymetrix GeneChip® Experiment Protocol



Analysis Flow Chart



Affymetrix Files

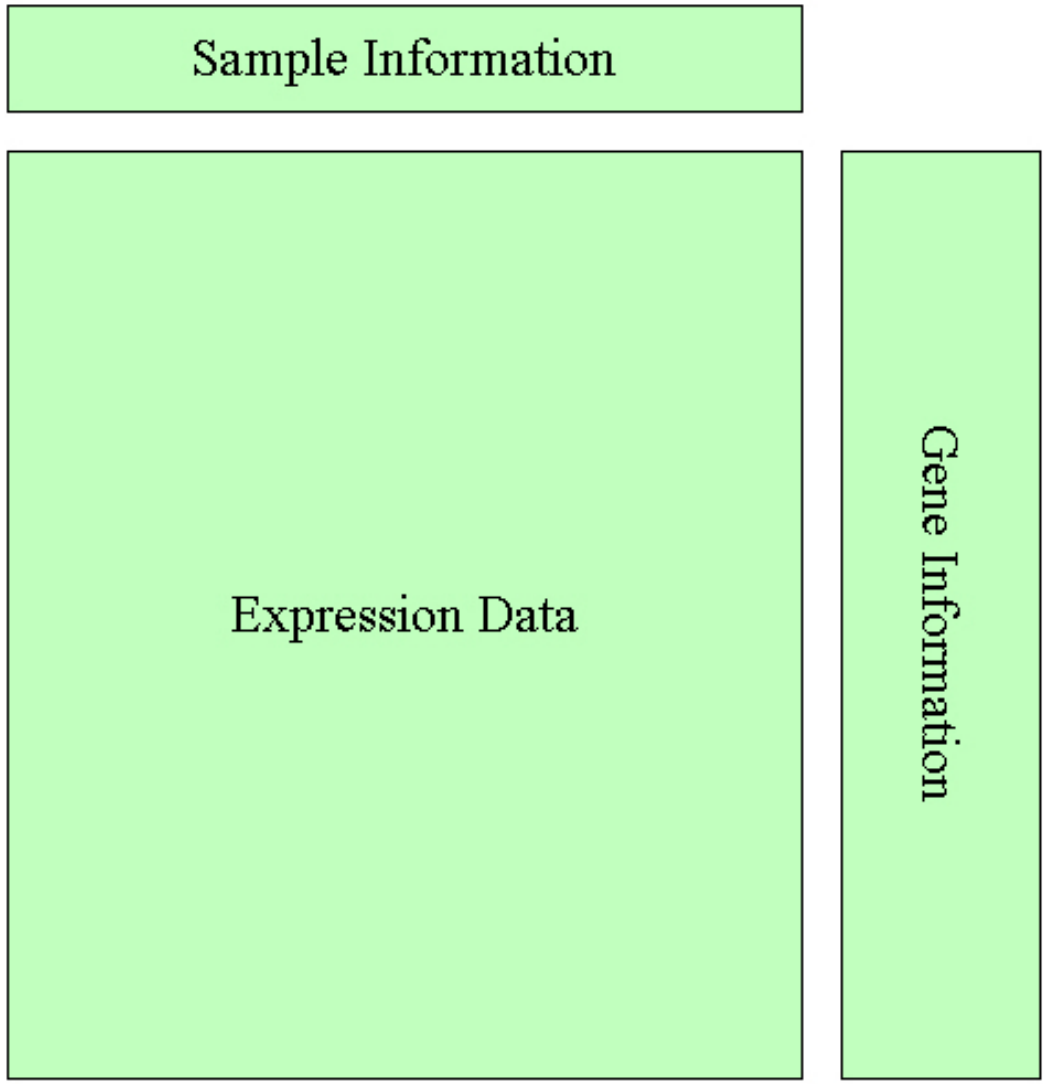
- DAT file: Scanned image.
- CEL file: Output fro image analysis software. Contains cell intensity file, probe-level values.
- CDF file: Chip description file. Describes which probes go in which probe-sets and the location of the probes on the chip.

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

Mock Data

- a data.frame with sample information
- a meta data.frame describing the variables in the data.frame
- Bioconductor uses "AnnotatedDataFrame" to describe phenotype data

```
>fake.data<-matrix(rnorm(8*200), ncol=8)
>sample.info<-data.frame(spl=paste("A", 1:8, sep=""),
stat=rep(c("cancer", "healthy"), each=4))
>meta.info<-data.frame(c("Sample Name"), "Cancer
Status")
>pheno<-new("AnnotatedDataFrame", data=sample.info,
varMetadata=meta.info)
>my.experiment<-new("ExpressionSet", exprs=fake.data,
phenoData=pheno)
```

Warning

- If you create a real ExpressionSet this way, YOU need to ensure that the column of the expression matrix are in exactly the same order as the rows of the sample information data frame.
- You'll also need to put together something that describe the genes used on the microarrays.

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=
```

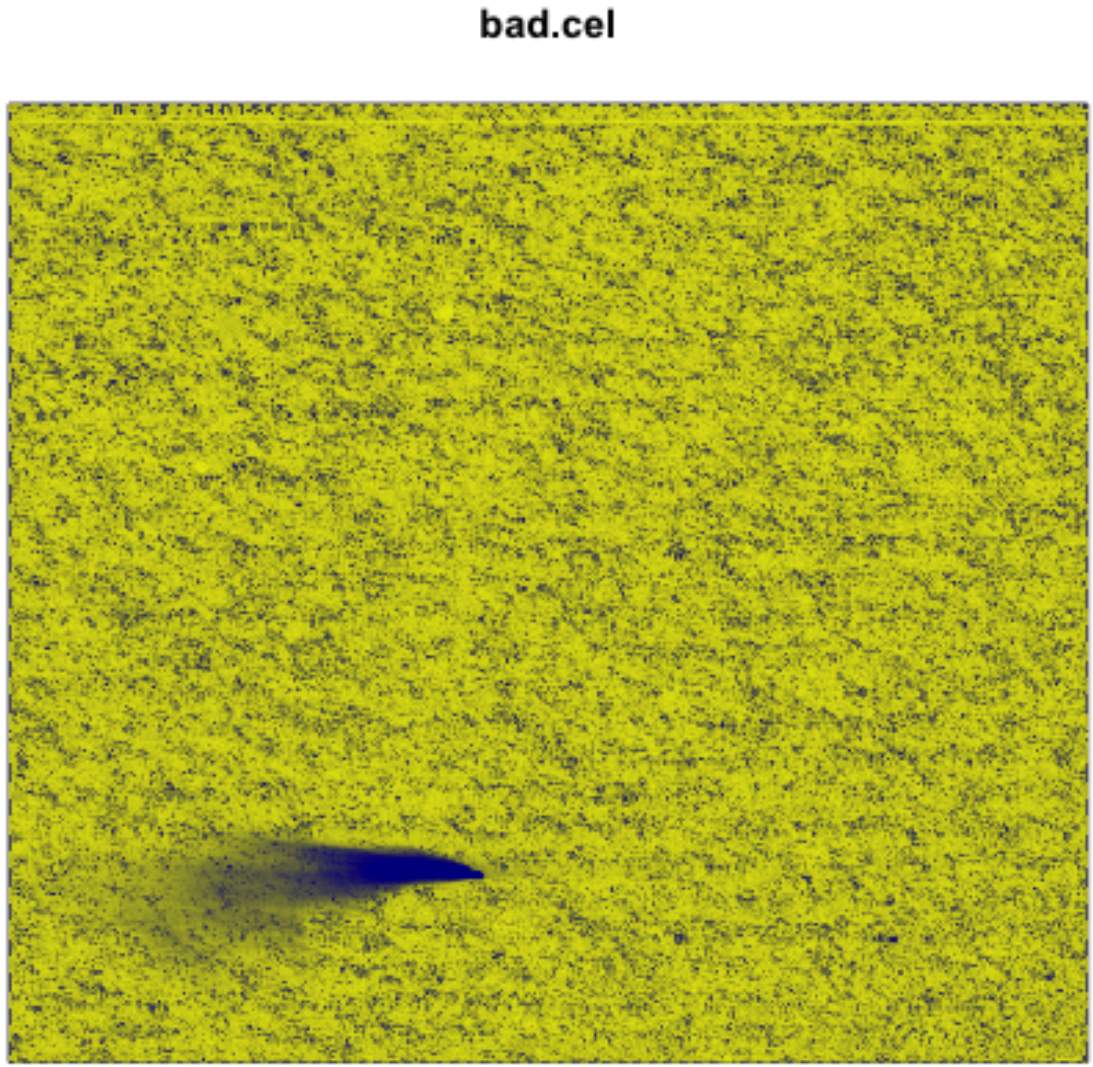
Look at the Experimental Design

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

Quality Assessment

- Image plot
- simpleaffy
- affyPLM

Image Plot

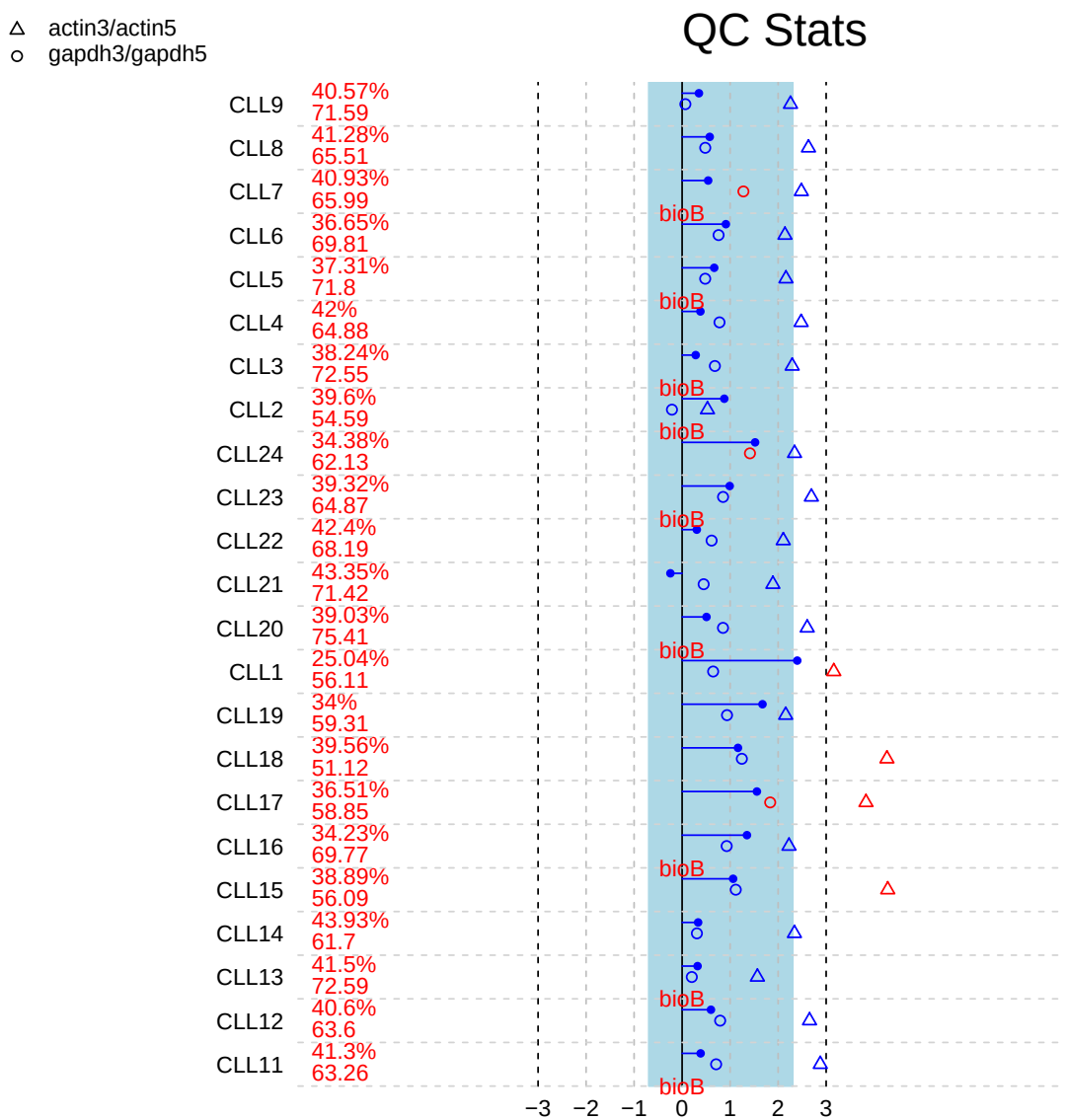


simpleaffy

Assess quality of samples

- average background
- % Present: percentage of probe pairs that have PM $>$ MM
- β-actin & GAPDH ratio: Use house-keeping genes to measure the quality of the sample hybridized to the chip

simpleaffy


actin3/actin5
gapdh3/gapdh5
QC Stats
CLL9 40.57% 71.59
CLL8 41.28% 65.51
CLL7 40.93% 65.99
CLL6 36.65% 69.81
CLL5 37.31% 71.8
CLL4 42% 64.88
CLL3 38.24% 72.55
CLL2 39.6% 54.59
CLL24 34.38% 62.13
CLL23 39.32% 64.87
CLL22 42.4% 68.19
CLL21 43.35% 71.42
CLL20 39.03% 75.41
CLL1 25.04% 56.11
CLL19 34% 59.31
CLL18 39.56% 51.12
CLL17 36.51% 58.85
CLL16 34.23% 69.77
CLL15 38.89% 56.09
CLL14 43.93% 61.7
CLL13 41.5% 72.59
CLL12 40.6% 63.6
CLL11 41.3% 63.26
bioB
−3 −2 −1 0 1 2 3

simpleaffy

```
>library("CLL")
>library("simpleaffy")
>data("CLLbatch")
>CLLbatch
>saqc<-qc(CLLbatch)
>plot(saqc)
```

affyPLM

- Chip Pseudo-Images
- Relative Log Expression (RLE)
- Normalized Unscaled Standard Error (NUSE)

RLE

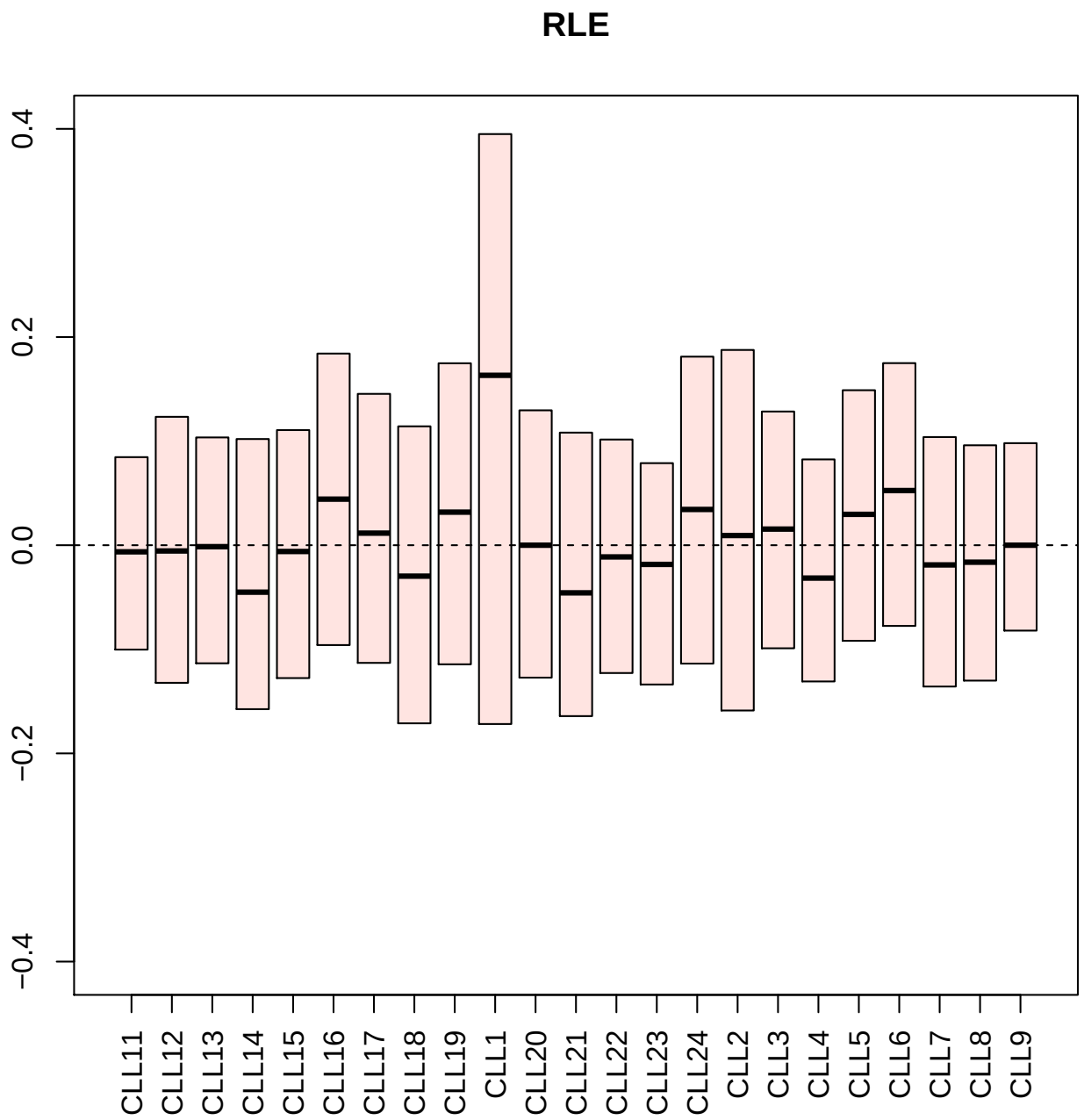
RLE values are computed for each gene by comparing the expression value on each array against the median expression value for that gene across all arrays.

$$RLE_{gi} = \hat{\theta}_{gi} - m_g$$

$\hat{\theta_{gi}} =$ expression of gene g on array i,
M_g: median of $\hat{\theta}_{gi}$.

Assuming that most genes are not changing in expression across arrays means ideally most of these RLE values will be near 0.

RLE


RLE
0.4
0.2
0.0
−0.2
−0.4
CLL11
CLL12
CLL13
CLL14
CLL15
CLL16
CLL17
CLL18
CLL19
CLL1
CLL20
CLL21
CLL22
CLL23
CLL24
CLL2
CLL3
CLL4
CLL5
CLL6
CLL7
CLL8
CLL9

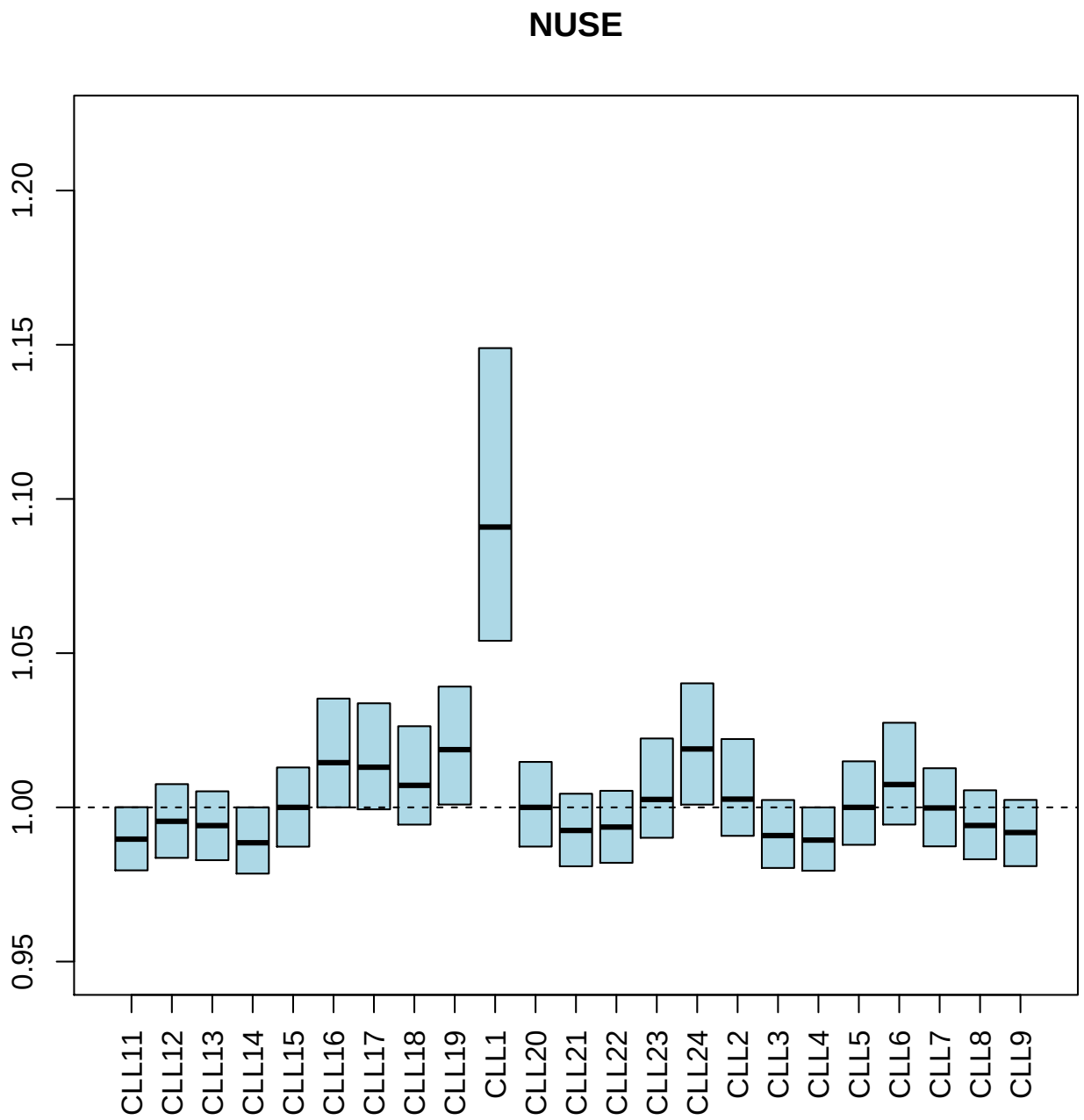
NUSE

The standard error estimates obtained for each gene on each array from fitPLM are taken and standardized across arrays.

$$NUSE_{gi} = \frac{SE(\hat{\theta_{gi}})}{median(SE(\hat{\theta_{gi}}))}$$

$\rightarrow$ the median of NUSE for each array should center around 1.

NUSE


NUSE
1.20
1.15
1.10
1.05
1.00
0.95
CLL11
CLL12
CLL13
CLL14
CLL15
CLL16
CLL17
CLL18
CLL19
CLL1
CLL20
CLL21
CLL22
CLL23
CLL24
CLL2
CLL3
CLL4
CLL5
CLL6
CLL7
CLL8
CLL9