STAT588/BIOL588: Genomic Data Science
Lecture 13: Processing Microarray Data

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

Objectives of Lecture 13

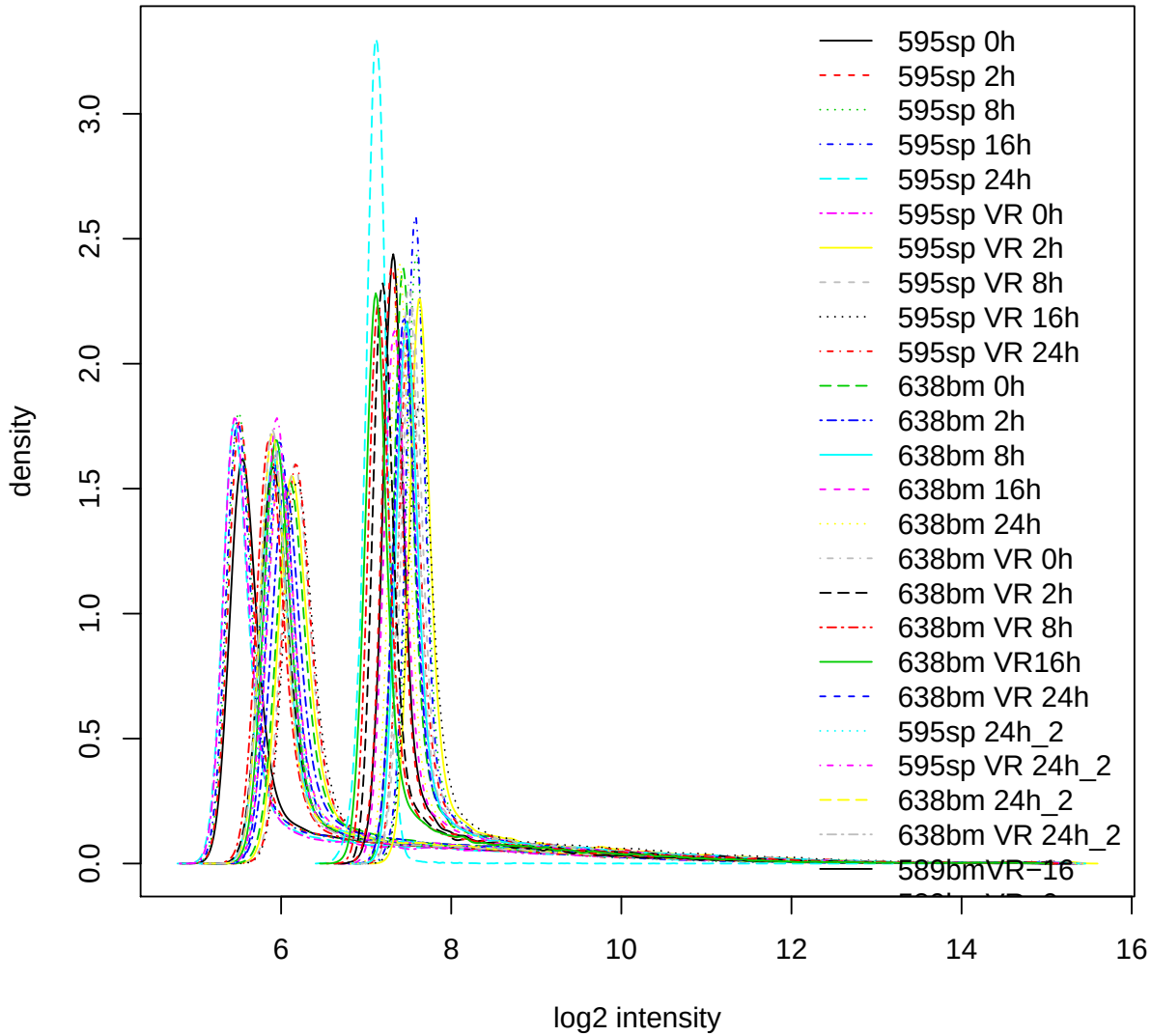
- Structures of Genomic Data
- Quality Assessment
  - Visualizations for Quality control
- Preprocessing
  - Background Correction
  - Normalization
  - Probe Level Data Summarization

After Background Correction: Normalization

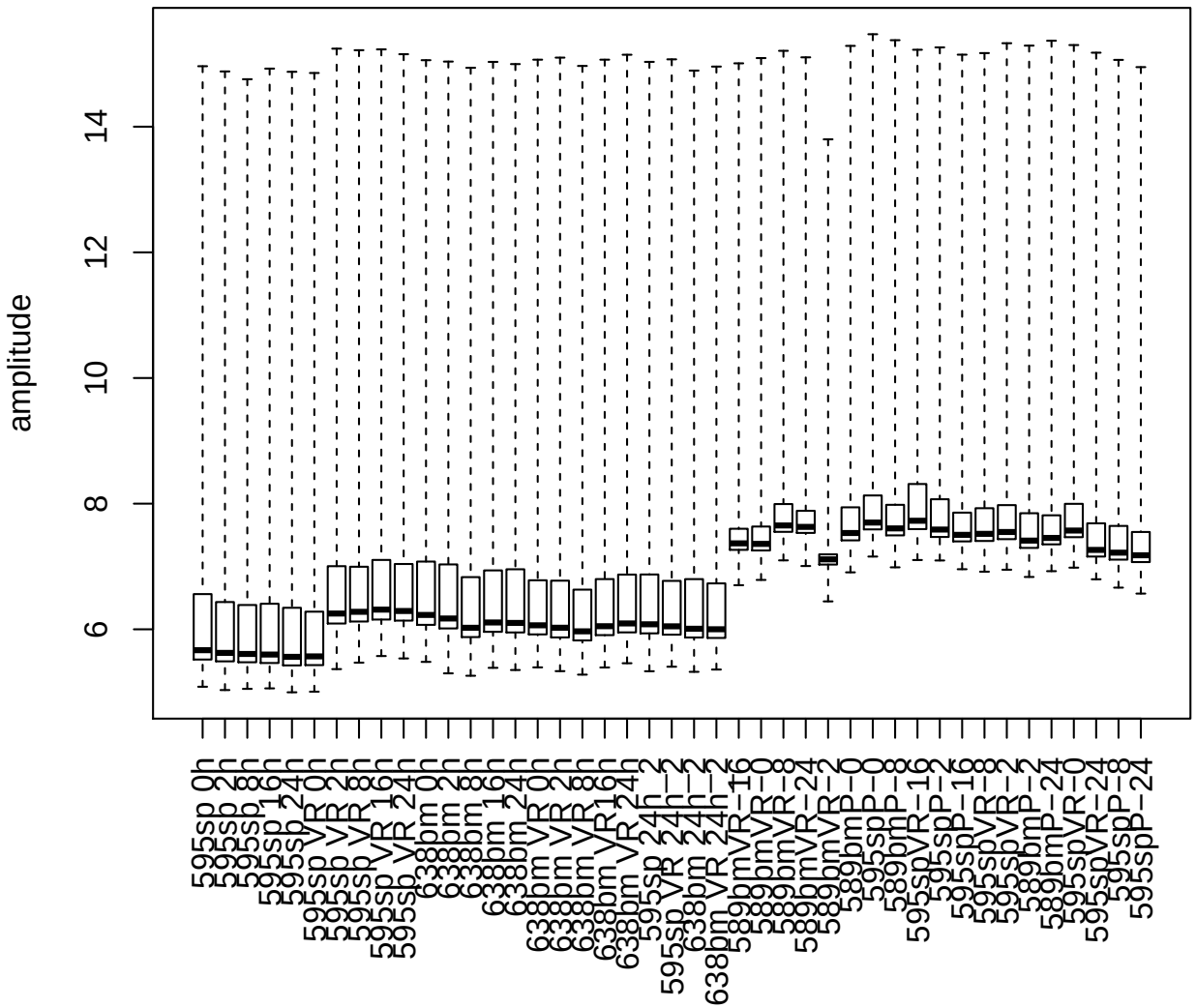
Purpose

- Remove the effect of technical variation across chips
- For example: scanner settings, amount of hybridized target mRNA

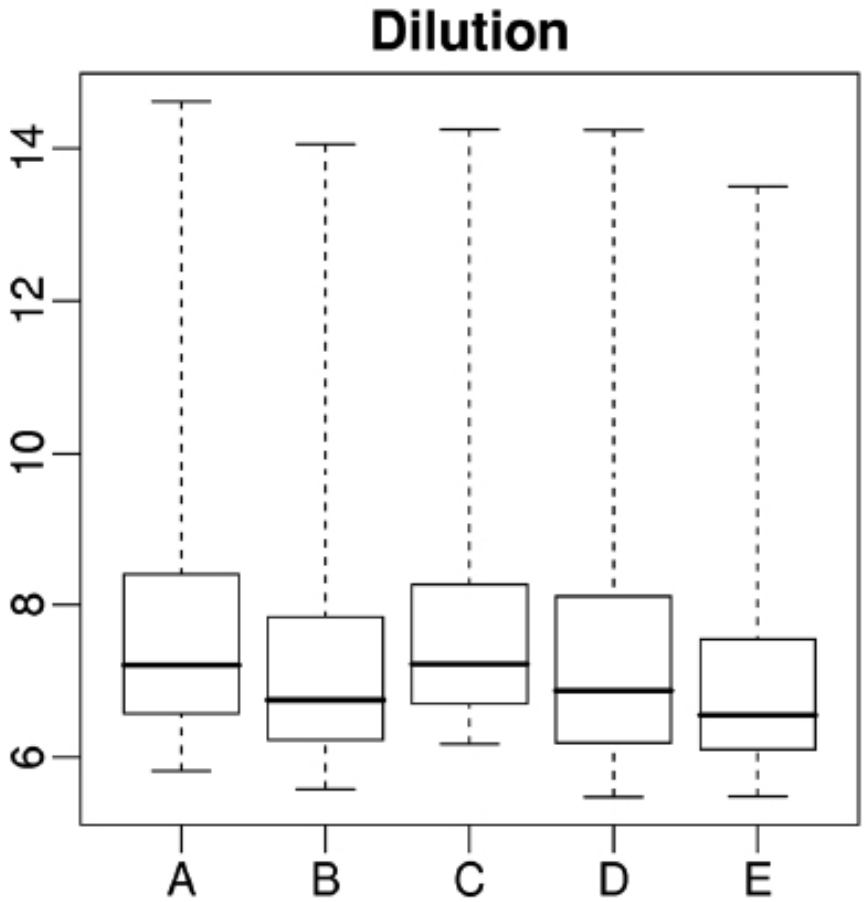
Batch Effect



Batch Effect

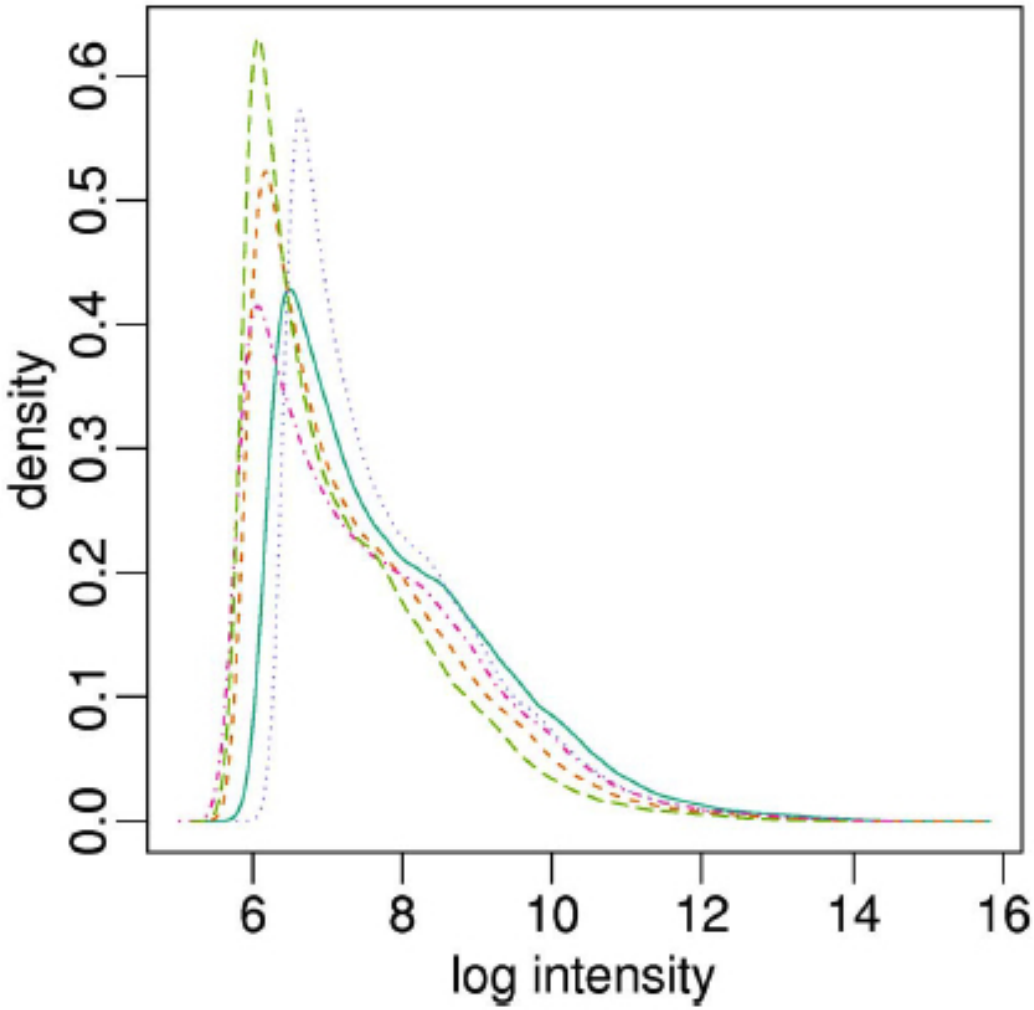


Replicate Data

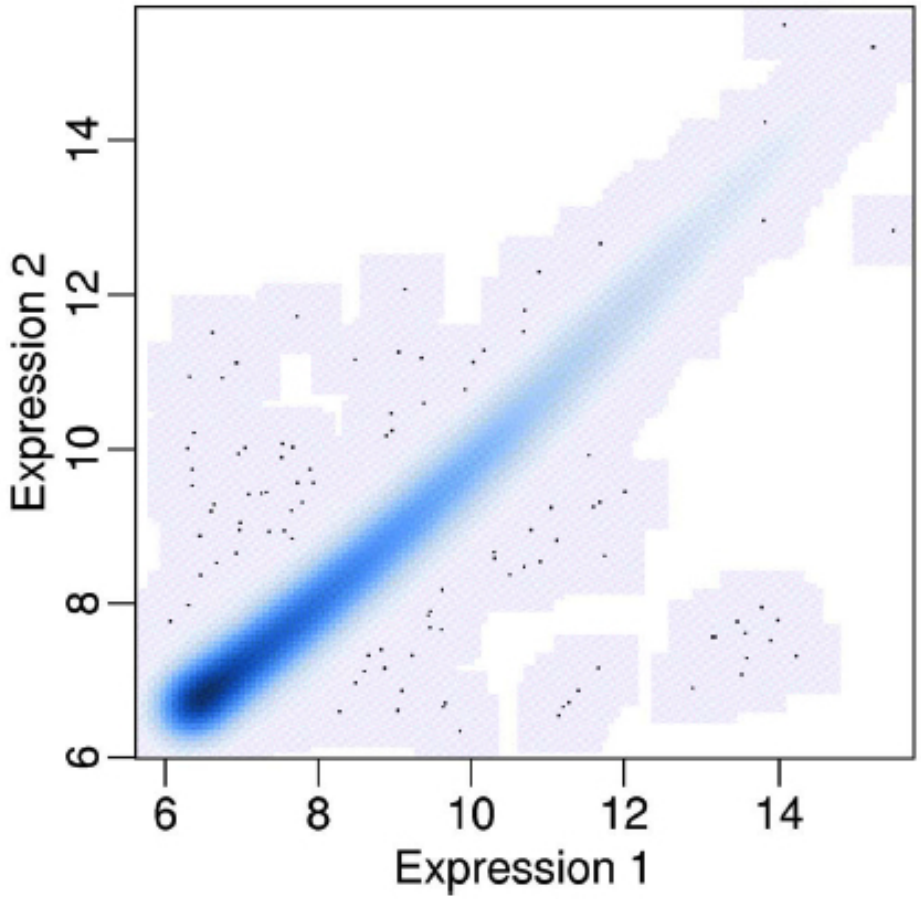


Here different scanner were used

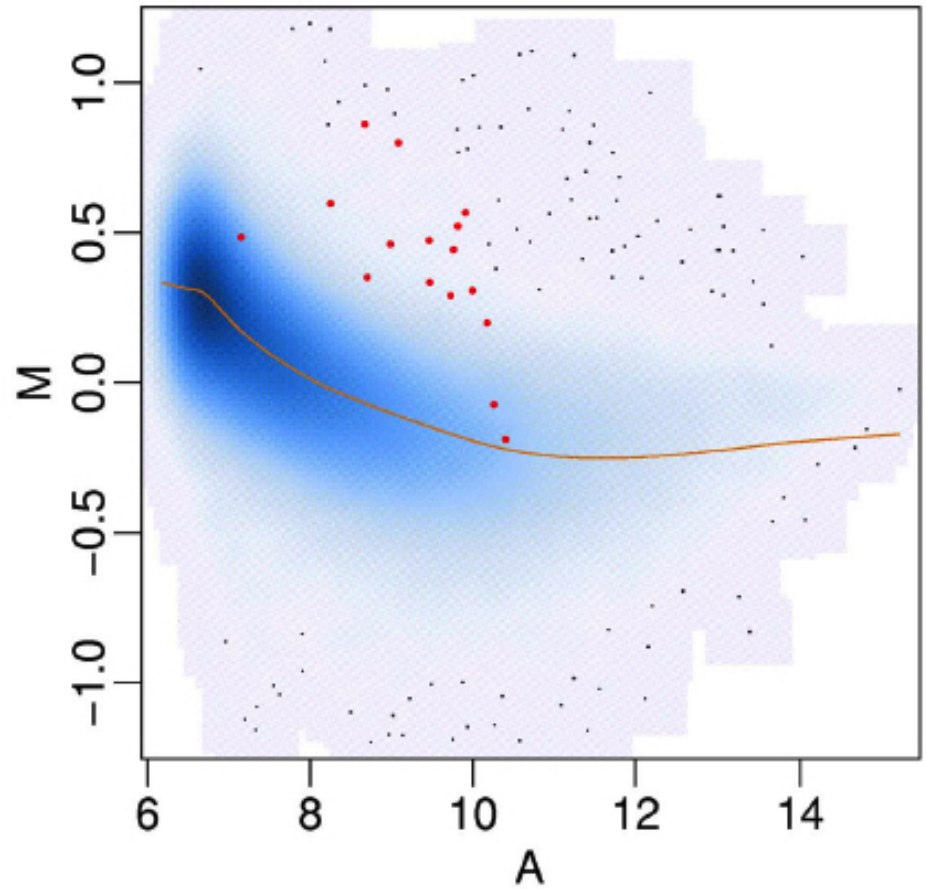
Replicate Data



Scatter Plot

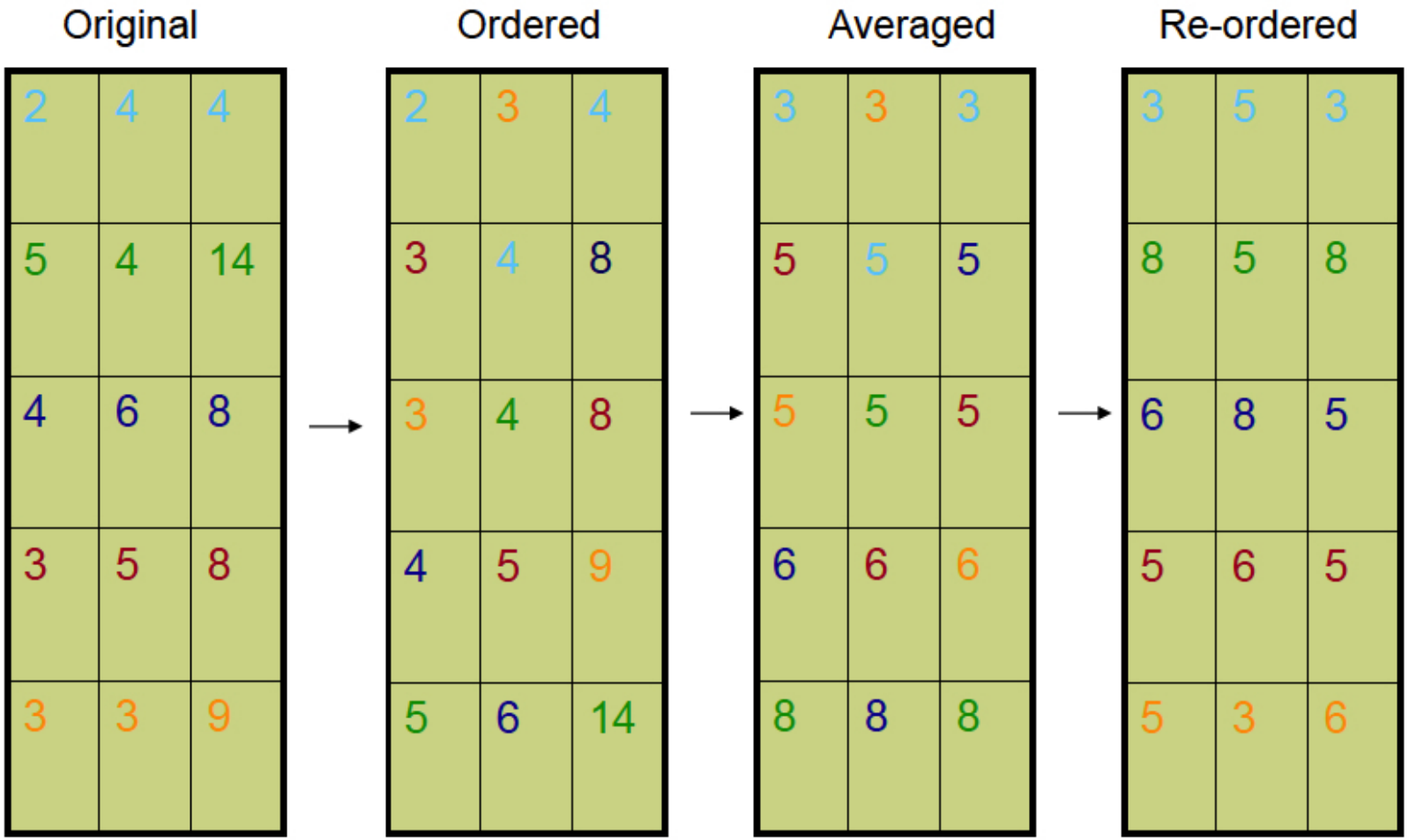


Intensity Effect



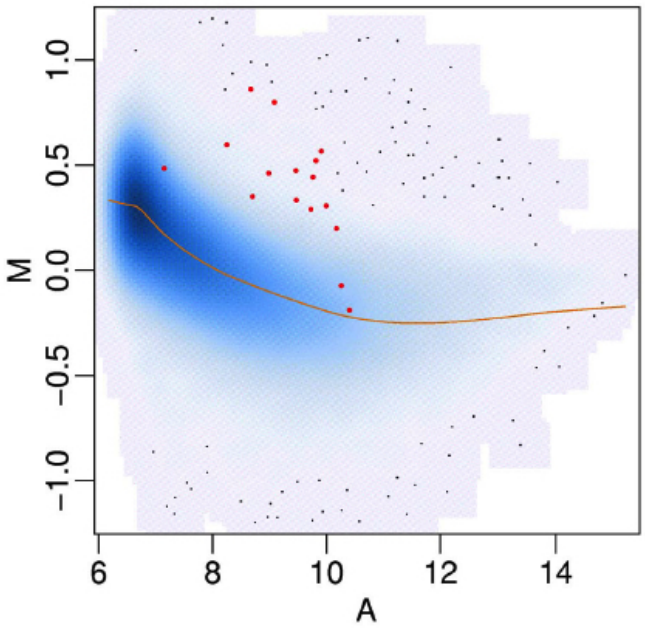
$$M = \log P_1 - \log P_2$$
$$A = (\log P_1 + \log P_2)/2$$

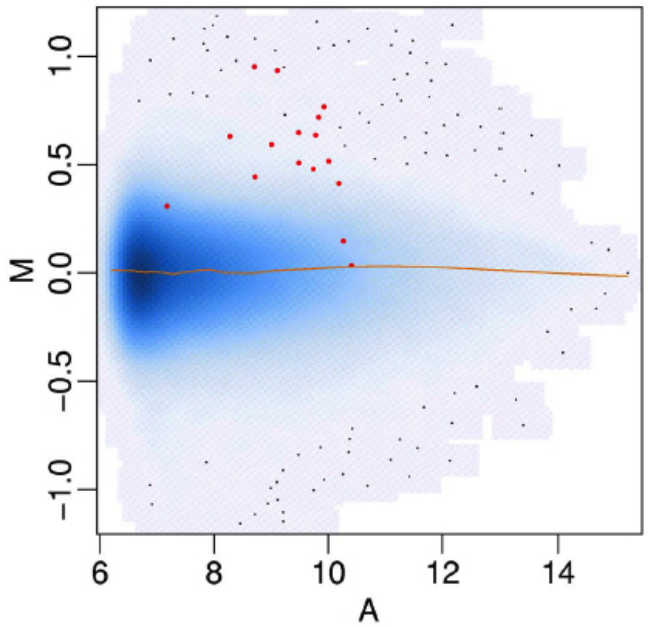
Quantile Normalization



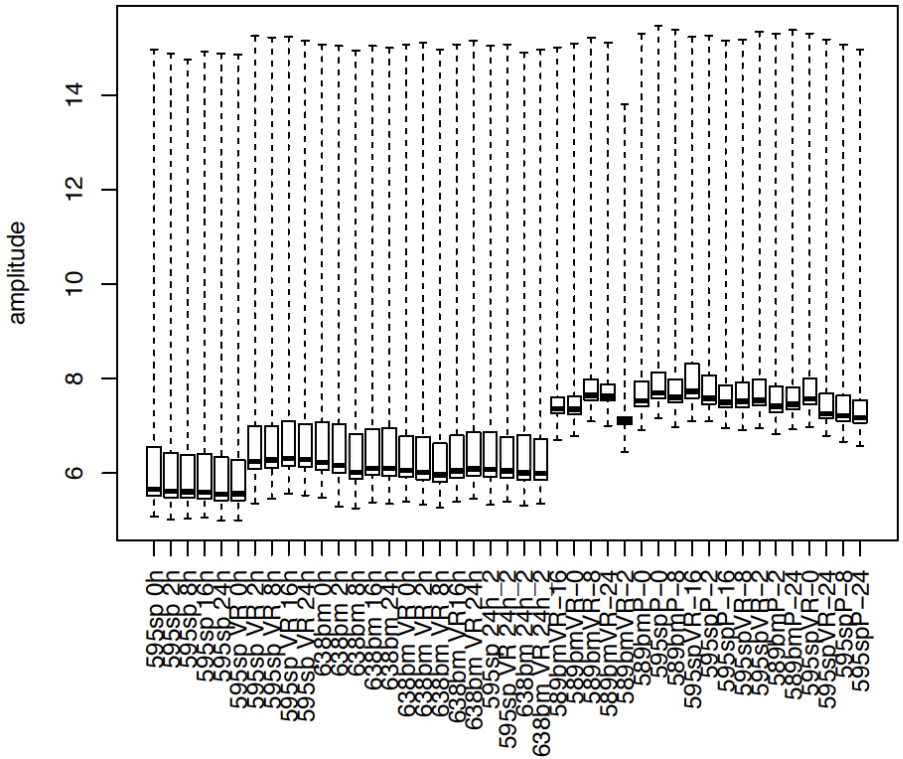
Assumption: same probe-level intensity distribution across chips

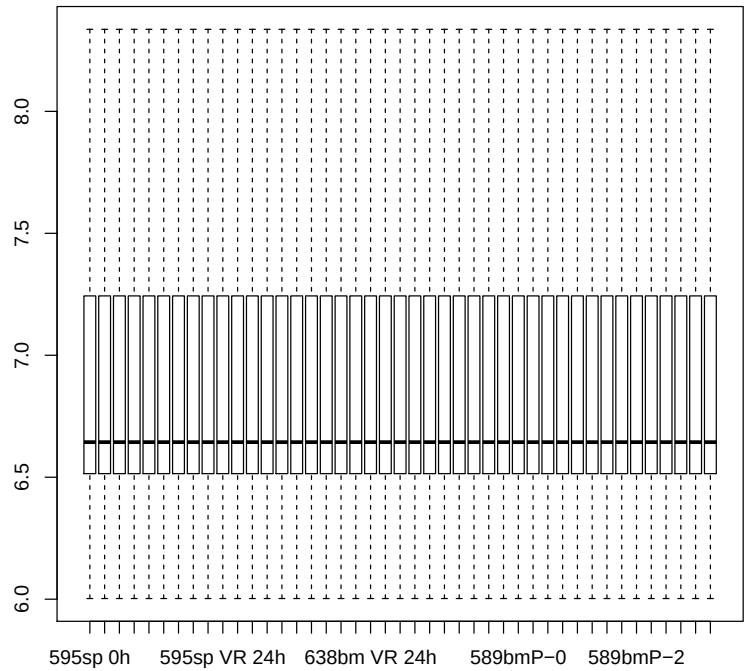
After Quantile Normalization





Batch Effect

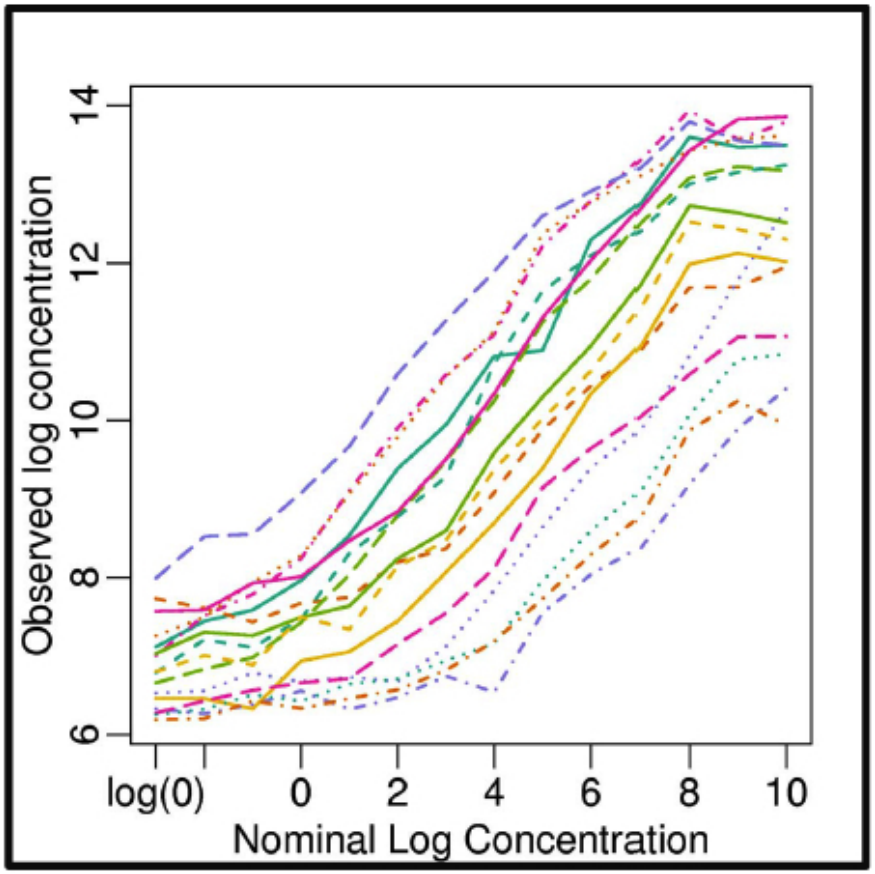




Outlines

- Structures of Genomic Data
- Quality Assessment
  - Visualizations for Quality control
- Preprocessing
  - Background Correction
  - Normalization
  - Probe Level Data Summarization

Probe specific background effect



Each probe has its own α !

RMA: Summarization

The Robust Multi-Array Average (RMA) model

$$Y_{ij} = \alpha_j + \beta_i + \epsilon_{ij},$$

and $\sum_{i=1}^{I} \alpha_j = 0$.
Where
Y_{ij} is $\log_2$ background-adjusted and normalized PM intensity,
β_i is the expression level of gene for chip i,
α_j is a probe effect.

- Then use robust regression method to estimate values

R Code

```
>library("CLL")
>CLLrma <- rma(CLLB)
>e <- exprs(CLLrma)
>dim(e)
>dim(CLLrma)
```

Exercise: Homework 7 (3) (4) (5)

1. Download CEL files from GSE18088 at gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/)
2. Normalize data in the study

Hint:

library(oligo)
library(siggenes)
library(limma)
library(pd.hg.u133.plus.2)
library(hgu133plus2.db)
library(hgu133a.db)
exprs
rma