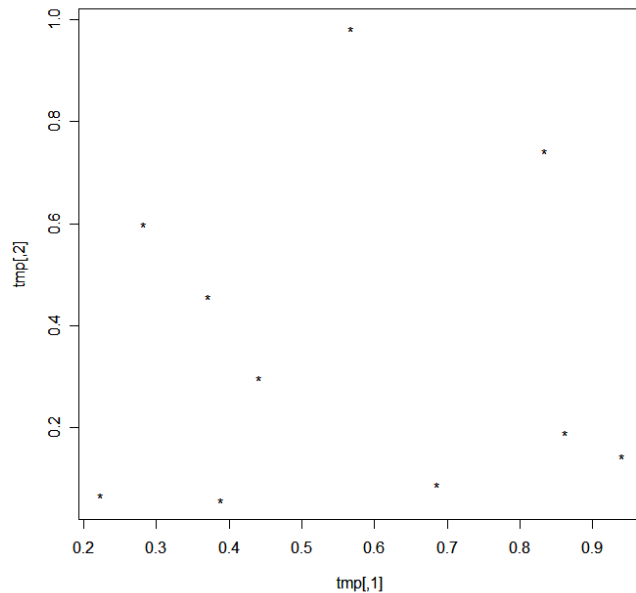


# Statistical Assessment and Quality Control of High Throughput Data

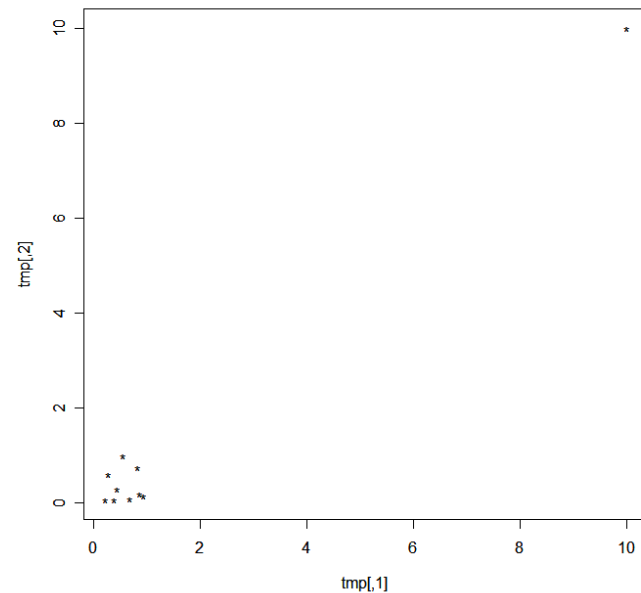
# Distributional Assumptions and Outliers

- Most statistical approaches rely on well-regulated statistical distribution
- Skewed distribution → Seek a transformation of data
- Heterogeneous variances → variance stabilization transformation or variance-adaptive statistical methods
- Outlier → Robust statistical methods or bootstrapping

Cor = 0.02



Cor = 0.99

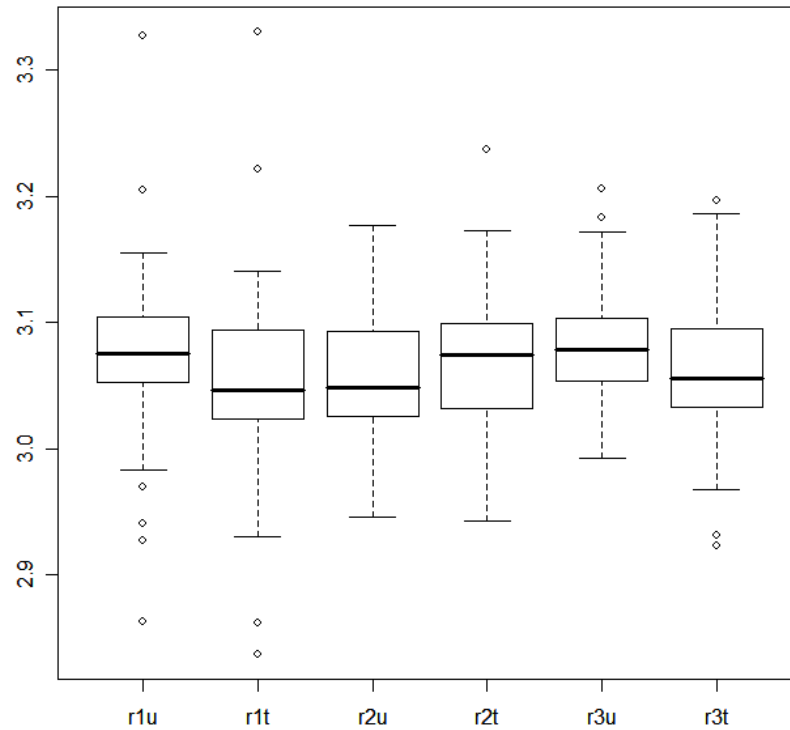


# Dispersion, correlation, and graphical assessment

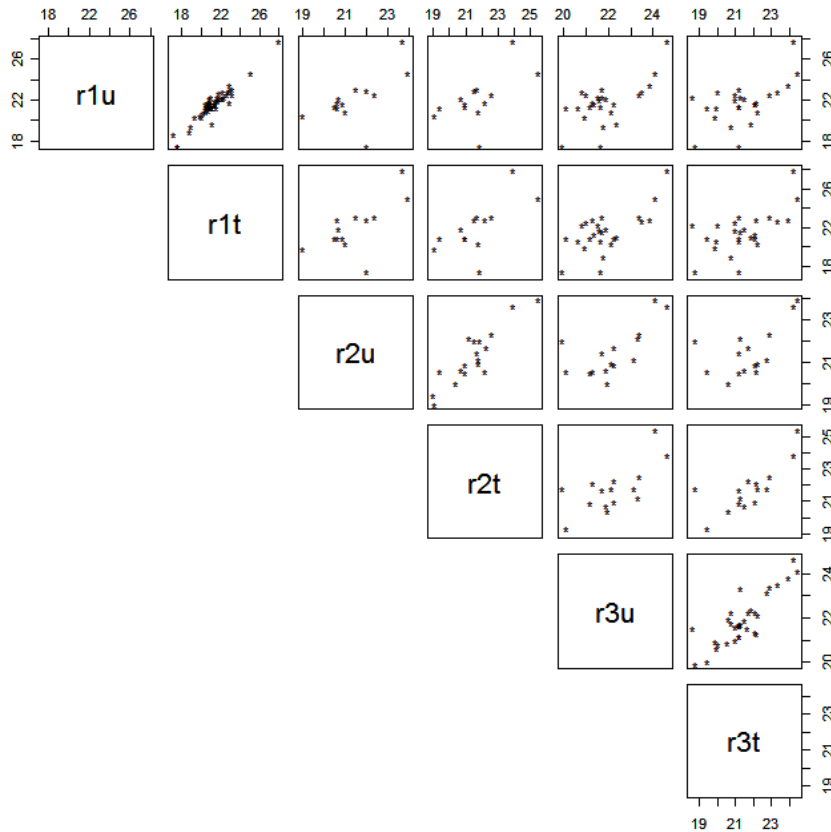
- Basic dispersion parameters: minimum, maximum, mean, median, quartiles, variance, missing values

	r1u	r1t	r2u	r2t	r3u	r3t
Min.	17.5	17.1	19.0	19.0	19.9	18.6
1st Qu.	21.2	20.6	20.6	20.8	21.2	20.8
Median	21.6	21.0	21.1	21.6	21.7	21.2
Mean	21.6	21.2	21.3	21.5	21.9	21.4
3rd Qu.	22.3	22.0	22.0	22.1	22.3	22.1
Max.	27.9	27.9	24.0	25.4	24.7	24.4
NA's	9	9	41	41	26	26

- Boxplot: inter-quartile ranges



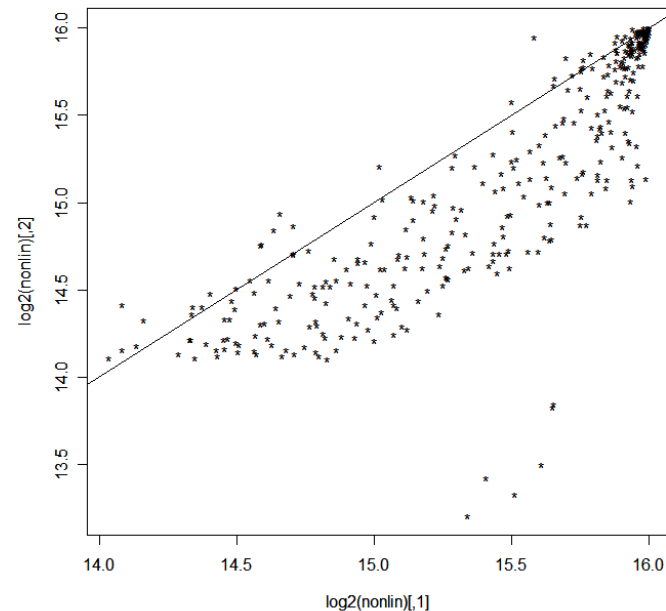
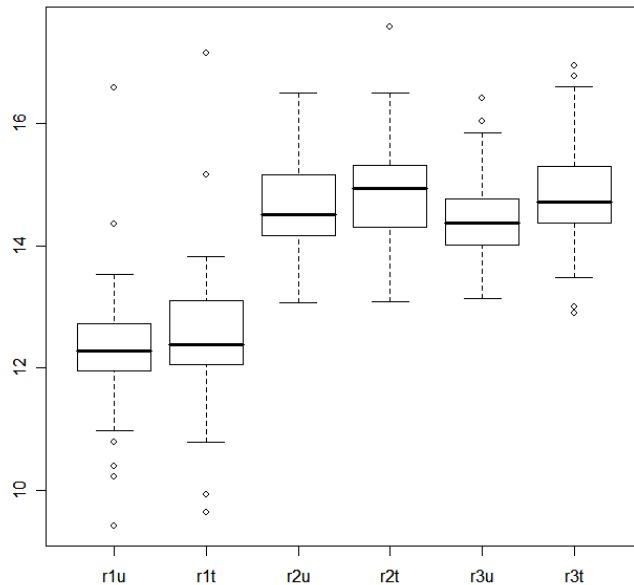
- Pairwise scatter plot and correlation analysis



	r1u	r1t	r2u	r2t	r3u	r3t
r1u	1.00	0.95	0.62	0.59	0.65	0.59
r1t	0.95	1.00	0.66	0.67	0.68	0.63
r2u	0.62	0.66	1.00	0.90	0.68	0.59
r2t	0.59	0.67	0.90	1.00	0.69	0.77
r3u	0.65	0.68	0.68	0.69	1.00	0.85
r3t	0.59	0.63	0.59	0.77	0.85	1.00

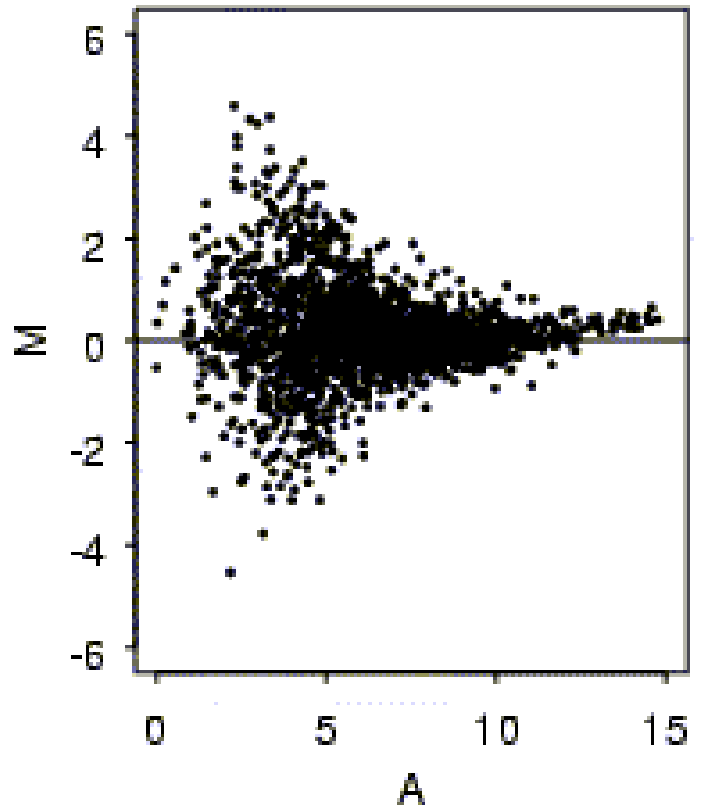
# Normalization

- Bring all the data from different experimental settings, e.g. different samples, hybridizations, scannings to the same baseline distribution
  - Additive or multiplicative constant correction
  - Non-linear regression fitting correction
- original or transformed scale?
  - Need to make a least mathematical manipulation



# AM transformation

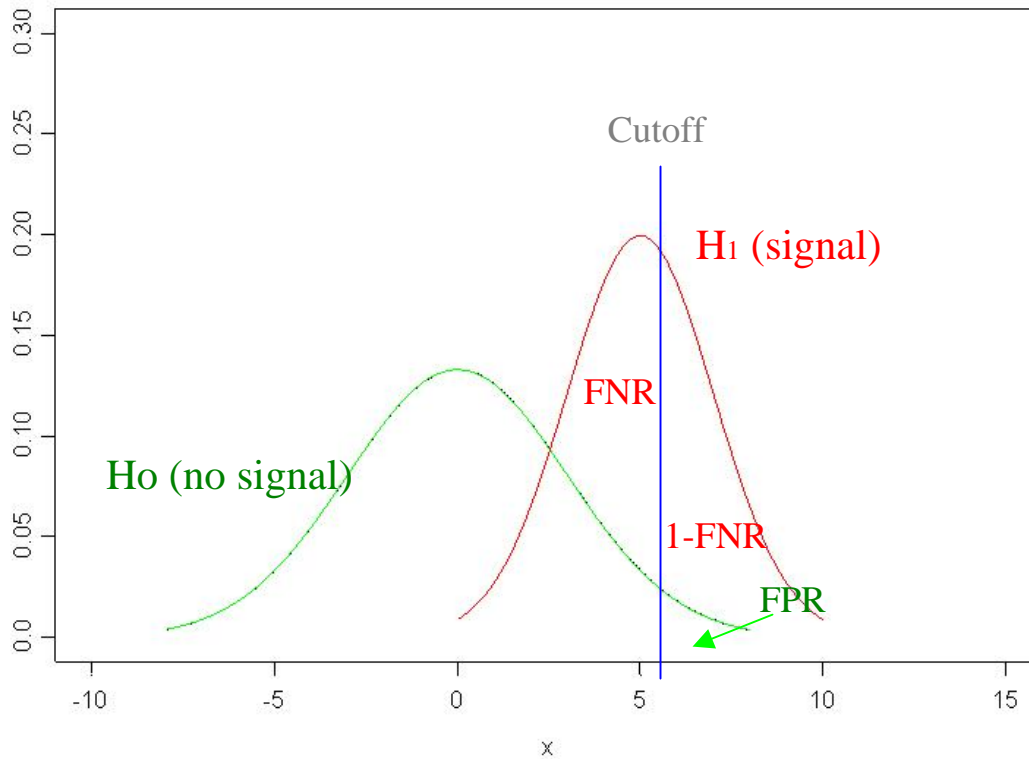
- XY scatter plot often leads to seeing biased error patterns
  - Mathematical bias when a regression-based normalization used
- AM transform:  $A = (X1+X2)/2$  and  $M = X1-X2$



# Sensitivity and Specificity

- Sensitivity: statistical power to detect the real positive ones (1-FNR), FNR=false negative error rate
- Specificity (or reproducibility): statistical power to detect false positive ones (1-FPR), FPR=false positive error rate

# Type I Error (FPR; False Positive Error Rate) and Type II Error (FNR; False Negative Error Rate)



**ROC Curve:**  
FPR vs. 1-FNR

# Receiver Operator Characteristic (ROC) curve

- Plot for FPR vs.1-FNR of a statistical test

