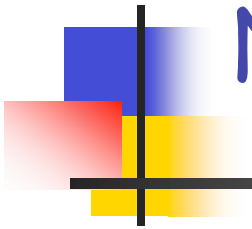
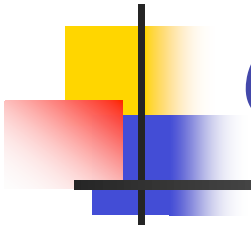


Normalization of Microarray Data

Paul Gauthier

Michael Ringenburg

CSE527 - 12/12/03



Outline

- Sources of error in microarray experiments
- cDNA array normalization
  - Global, linear and non-linear
  - Dye swapping, print tip effects
  - Evaluation of approaches
- Variance stabilization

Sources of Error

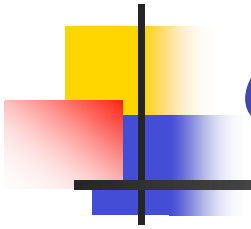
Fundamental

- Gene isoforms
- Probe specificity (3')
- MM probe masks legitimate signal
- Incorrect probes
- Inconsistent results: cDNA/Oligo/Northern

Normalization Applicable

- Dye color variation
- Print-tip effects
- Scanning variation
- Slide preparation
- Wet-lab variables
- Variance ~ expression

cDNA Microarrays



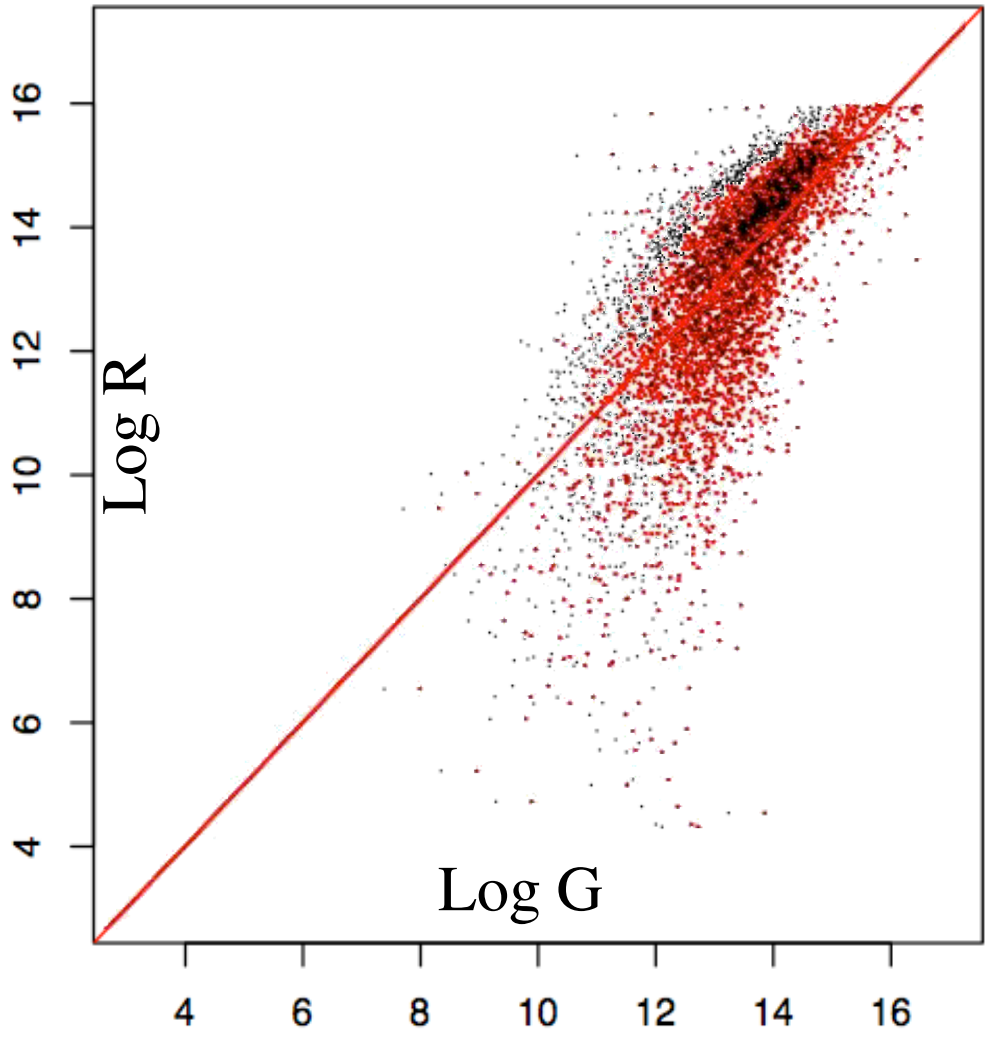
- cDNA array output:
- Per gene:
  (log R, log G)
- Fold change:
  M = log(R/G)
- Mean log-intensity:
  A = 1/2 log(R/G)

- Goal : correct for experiment differences
  - Dye specific issues, or
  - Sample related
- Control genes are constantly expressed :
  - You expect/want:
    M = log(R/G) ~ 0

Global Normalization

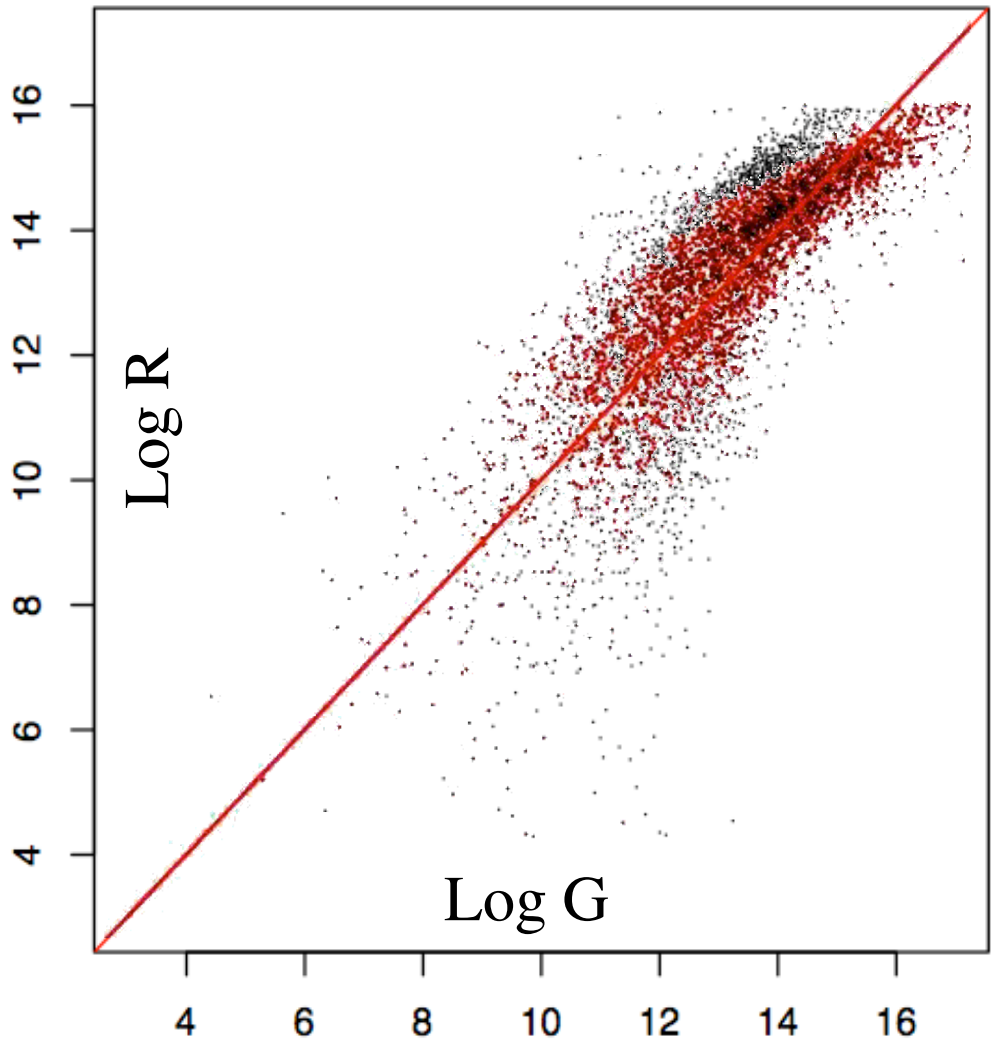
- $M^* = M + c = \log(kR/G)$
- c = median(M)

- Median is robust estimator if most genes are constantly expressed
- Yang, et al.; Park, et. al.



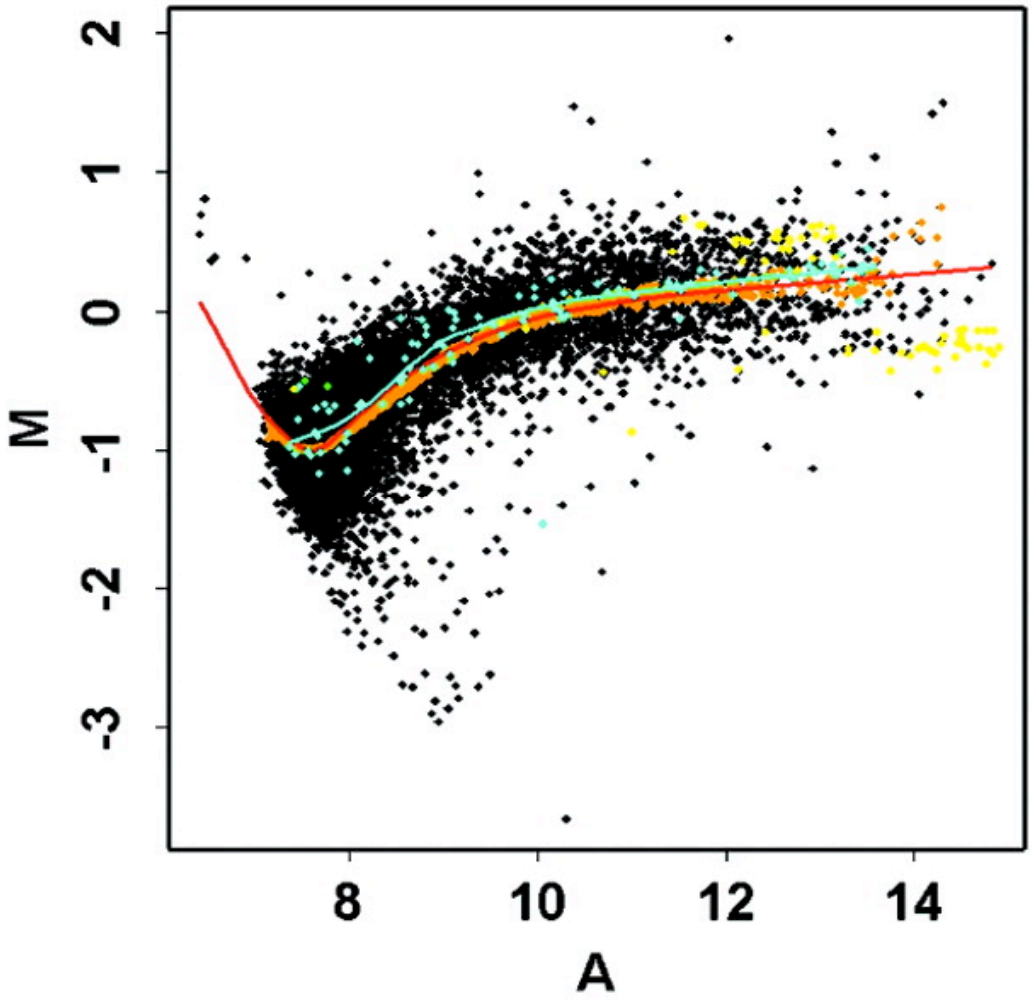
Linear Normalization

- $M^* = M + bA + c$
  $= \log(jA\ k\ R/G)$
- Compute b,c with b least-squares fit
  - Fit control genes or
  - Use robust fitter
- Park, et al.

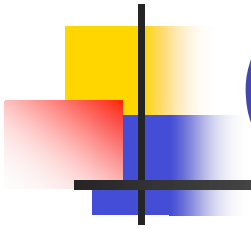


Non-Linear Normalization

- $M^* = M - c(A)$
  $= \log(k(a)\, R/G)$
- $c(A)$ fit by **lowess**
- Lowess:
  - Robust, locally line
    scatter-plot smoot
- Yang, et al.

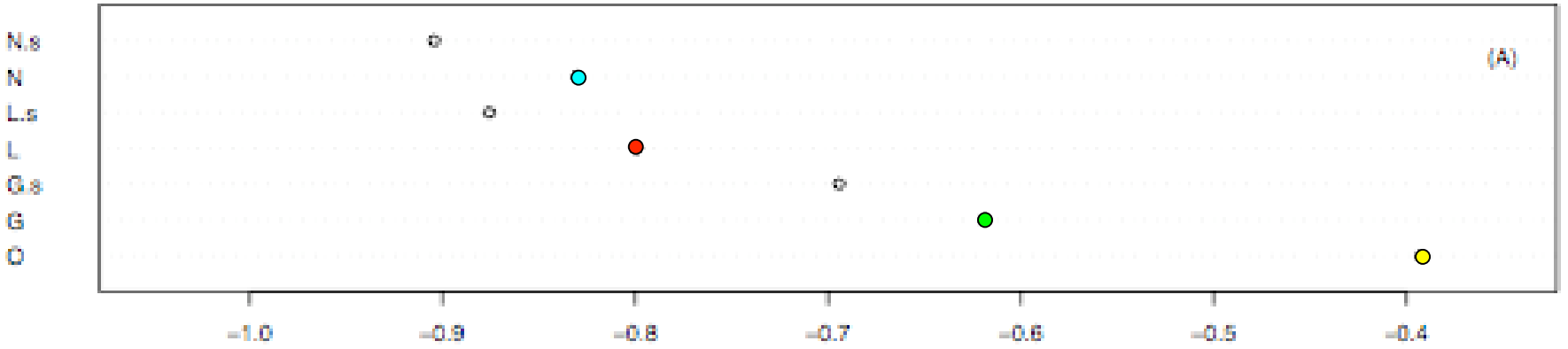


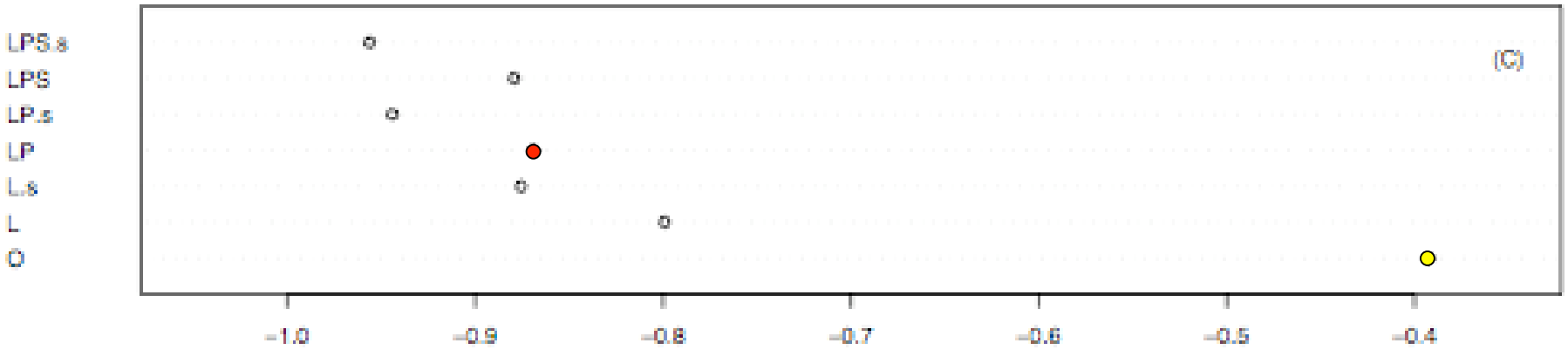
Special Cases (Yang, et. al.)

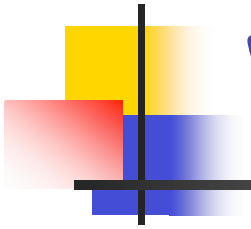


- Dye swap experiments
  - Duplicate experiments (M, A, M′, A′), dyes swapped
  - Can assume c ~ c′
    - Verify with control genes
    - Compute c using: M″ = 1/2(M + M′), A″= 1/2(A + A′)
- Print tip effects
  - Different slides sections use different print tips
  - Compute seperate c_i for each of the i=1..p print tips

Comparison of Approaches (Park, et al.)





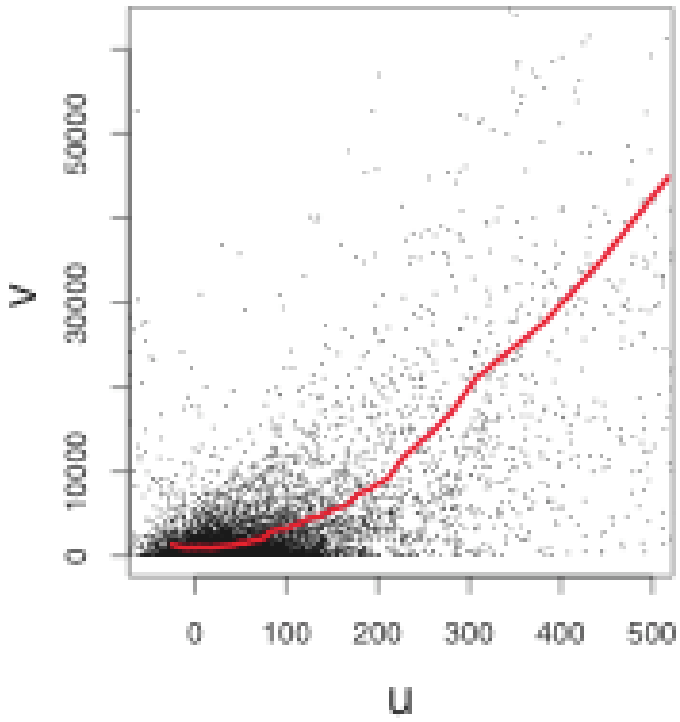


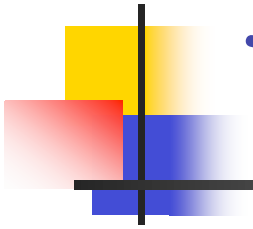
Variance Stabilization

- Previous methods discussed normalization.
- Huber et. al. and Geller et. al. add another goal — variance stabilization.
- Construct a difference statistic Δh whose variance does not depend on the mean.
  - Detecting differential expression: Let Δh replace M.
- Concentrate on the method of Huber et. al.

Motivation

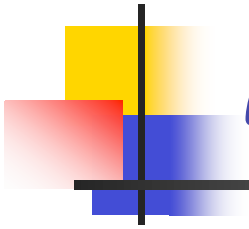
- In real microarray data, the variance depends on the mean intensity
- If variances equalized, can compare genes and decide which differences are most significant.





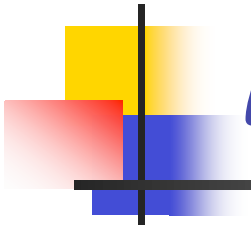
The Model

- Assume we can normalize with a linear model
  - $y_{ik} \rightarrow \ddot{y}_{ik} = o_i + s_i y_{ik}$
  - parameters $o_2,\ldots,o_d$ and $s_2,\ldots,s_d$
- Assume variance has quadratic dependence on mean.
  - $v(u_k) = (c_1 u_k + c_2)^2 + c_3$



Model

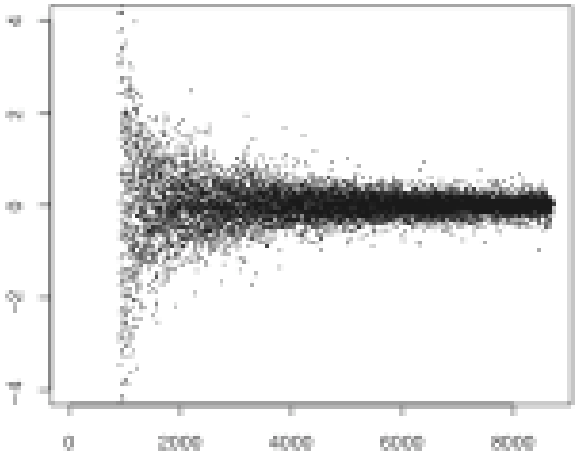
- Applying the variance stabilization technique from Tibshirani '88
  - $h(y) = g \operatorname{arsinh}(a + by)$
  - $g = c_1^{-1}$, $a = c_2 / \sqrt{c_3}$, $b = c_1 / \sqrt{c_3}$
- Combine with the normalization model
  - Omit scaling factor g
  - $y_{ik} \rightarrow h(\ddot{y}_{ik}) = \operatorname{arsinh}(a + b\,(o_i + s_i\, y_{ik}))$



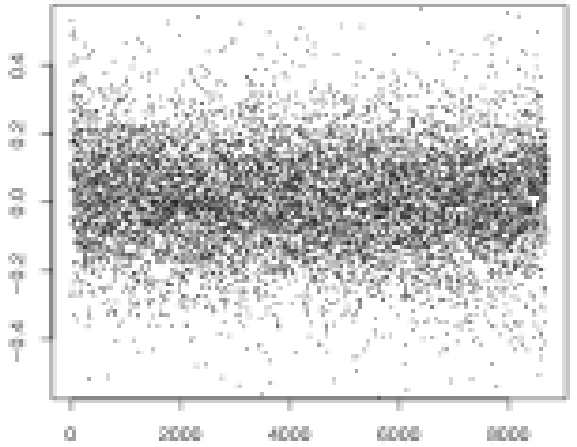
Model

- Set $a_i = a + bo_i$ and $b_i = bs_i$
  - Get $h(\ddot{y}_{ik}) = \text{arsinh}(a_i + b_i y_{ik})$
- $\Delta h_{k;ij}$ is our difference statistic
- Estimate parameters with EM/MLE
  - Estimate parameters from genes not differentially expressed
  - Estimate genes not differentially expressed from parameters
  - Iterate

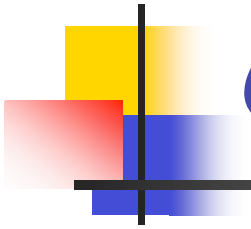
Results



Lowess Normalization



Variance Stabilization



Conclusions

- Microarray data has many sources of error.
- Some can be corrected by normalization and variance stabilization, some can not.
- Important question not addressed in these papers: how does the choice of normalization method effect the results of clustering, classification, et cetera?