

Transformaciones para datos de microarreglos

Diana M. Kelmansky
dkelman@ic.fcen.uba.ar

Instituto de Cálculo
FCEN-UBA

Transformaciones en datos de experimentos de microarreglos

Dos tipos de problemas

1. La variabilidad de las intensidades observadas depende de la media.
2. Los gráficos M-A presentan estructuras curvas no atribuibles a motivos biológicos.

Notación

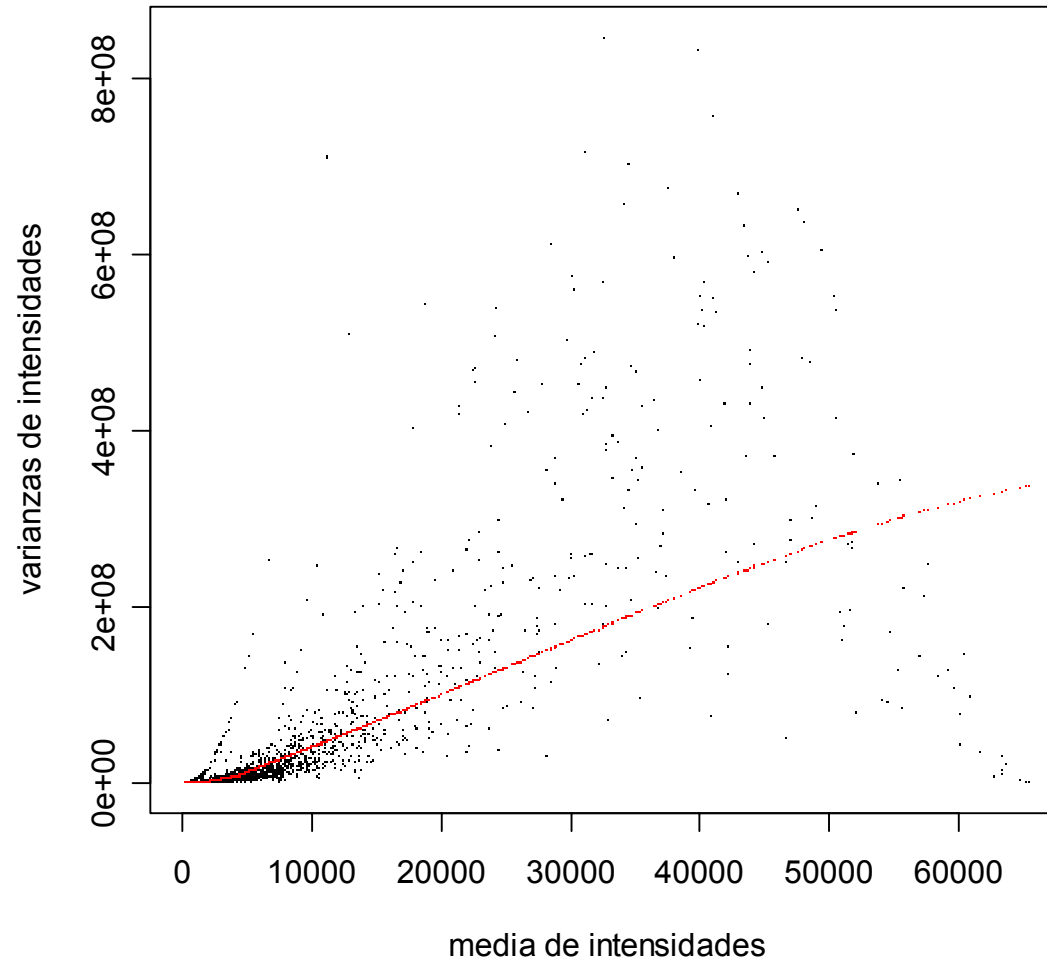
Y_{ik} intensidad observada para el gen k, el subíndice i indica el microarreglo

Media muestral, para las **intensidades** de cada gen entre microarreglos

Varianza muestral, para las **intensidades** de cada gen entre microarreglos

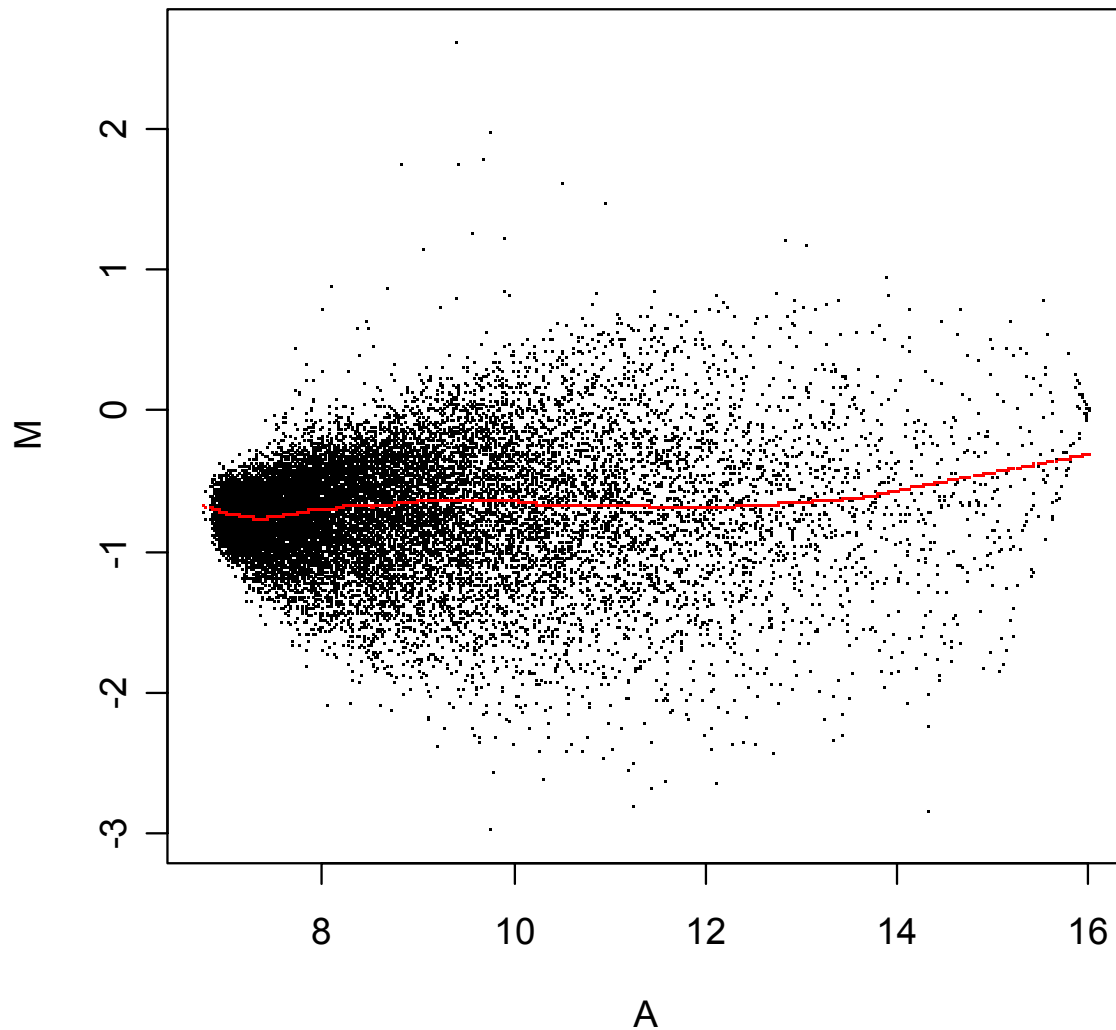
$$\bar{Y}_k = \sum_{i=1}^n Y_{ik} / n \qquad \sum_{i=1}^n (Y_{ik} - \bar{Y}_k)^2 / (n - 1)$$

beta7-, 6 microarreglos



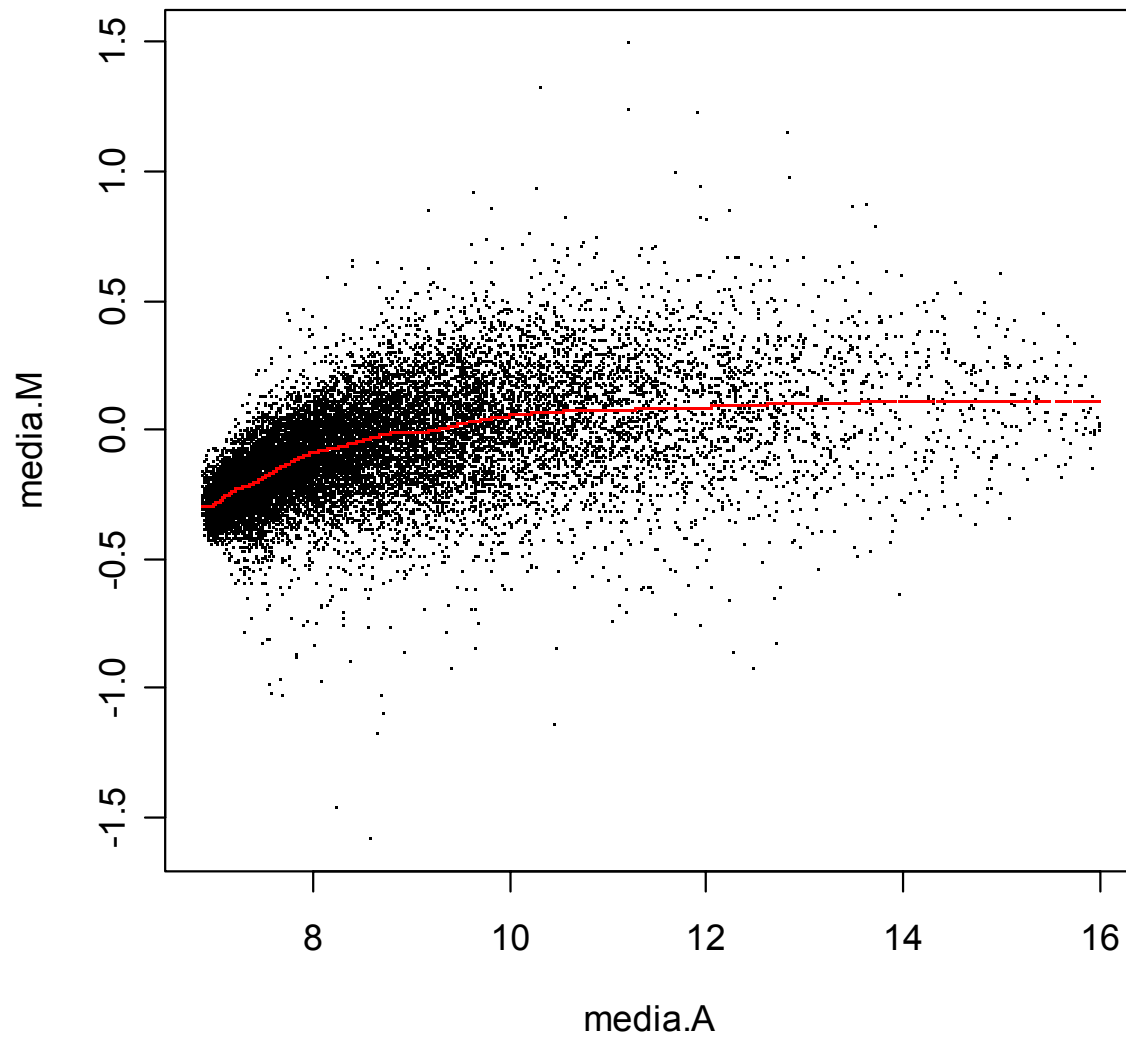
Las varianzas aumentan a medida que lo hacen los niveles de expresión de los genes

M-A plot para el primer microarreglo datos beta7



$$M = \log_2 (Y_{rk} / Y_{gk}) \quad A = 0.5 \log_2 (Y_{rk} * Y_{gk})$$

M-A plot sobre medias de 6 microarreglos datos beta7



Intensidades medias de beta7 + y beta7-

Transformación logaritmo en base 2

Es la transformación preferida por los biólogos

R: intensidad de un gen en la condición experimental

•G: intensidad del mismo gen en la condición de control

▪ $R/G = 1$ *no hubo cambio* $\log_2 (R/G) = 0$

▪ $R/G = 2 / 1$ $\log_2 (R/G) = 1$

el gen en la condición experimental *duplicó su intensidad* respecto de la condición de control.

▪ $R/G = 0.5 = 1 / 2$ $\log_2 (R/G) = -1$

el gen en la condición de control *duplica su intensidad* respecto de la condición experimental.

En general

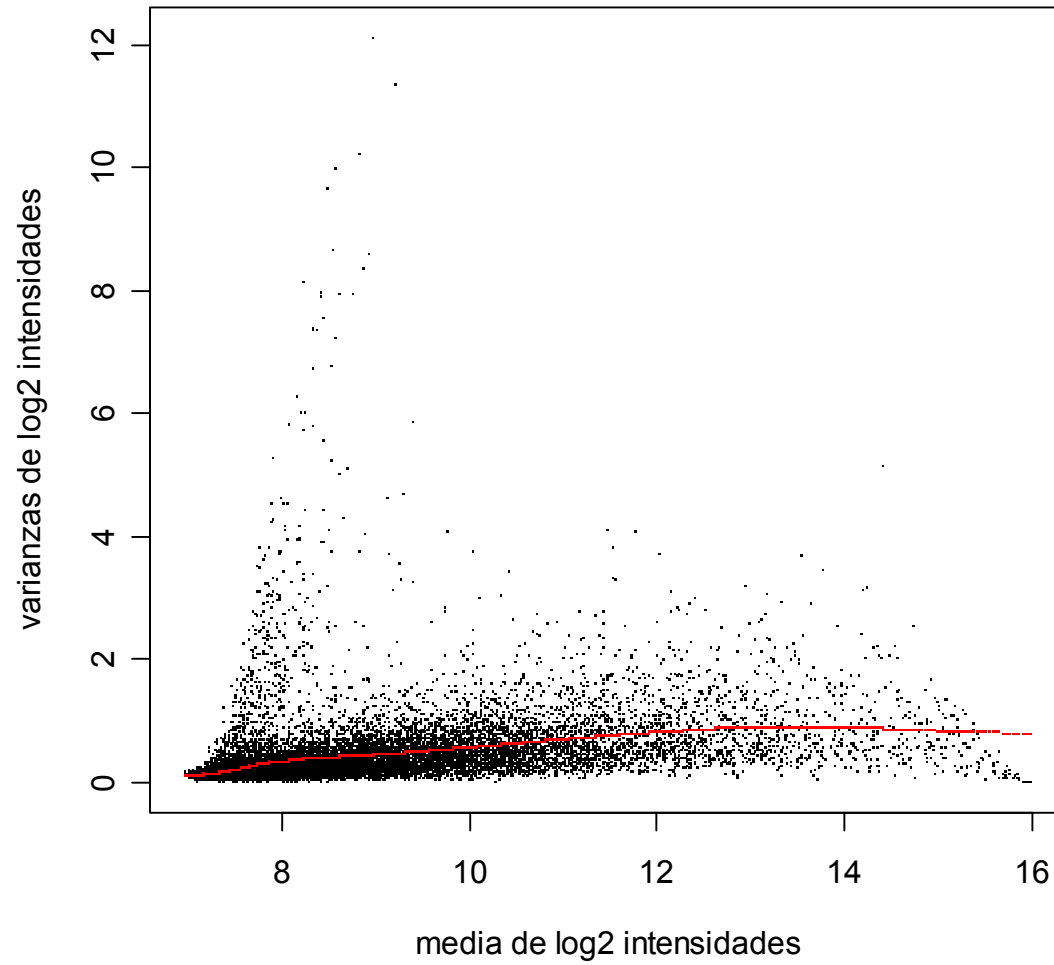
$$\begin{aligned}\log_2(2x) &= \log_2(x) + \log_2(2), \text{ doble} \Leftrightarrow \text{sumar } 1 \\ \log_2(x/2) &= \log_2(x) - \log_2(2), \text{ mitad} \Leftrightarrow \text{restar } 1\end{aligned}$$

- Aumentos y reducciones correspondientes al mismo “fold change” tienen el mismo tratamiento en la escala $\log_2$.

- Si un log ratio aumenta en una unidad esto corresponde duplicar el valor del numerador:

“one fold change = 2 times”

beta7- con log2, 6 microarreglos



La transformación log2 aumenta la escala de los transcriptos con intensidades bajas mientras que comprime la escala de los que tienen intensidades bajas.

Muchas veces revierte la dirección de la relación entre la varianza y la media de los niveles de expresión.

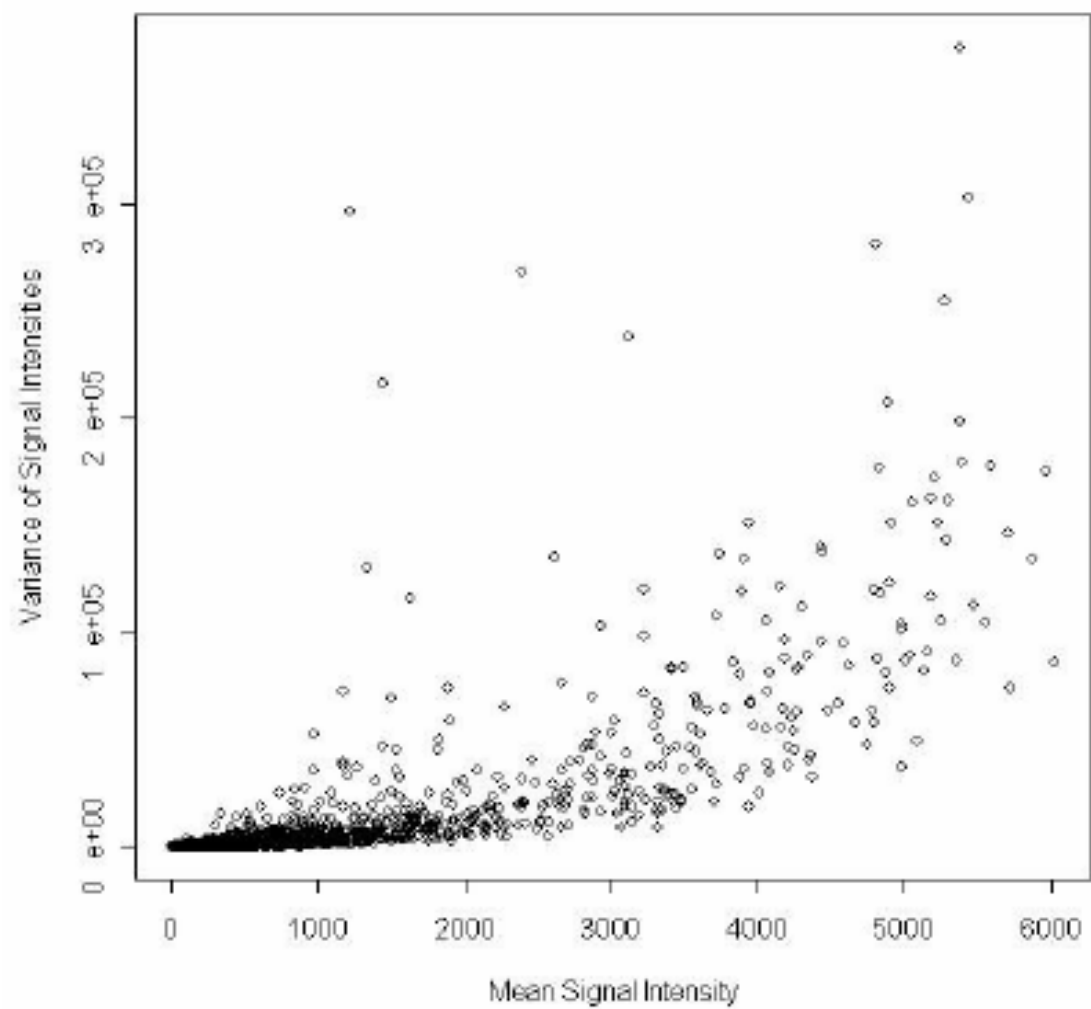
Luego de la transformación

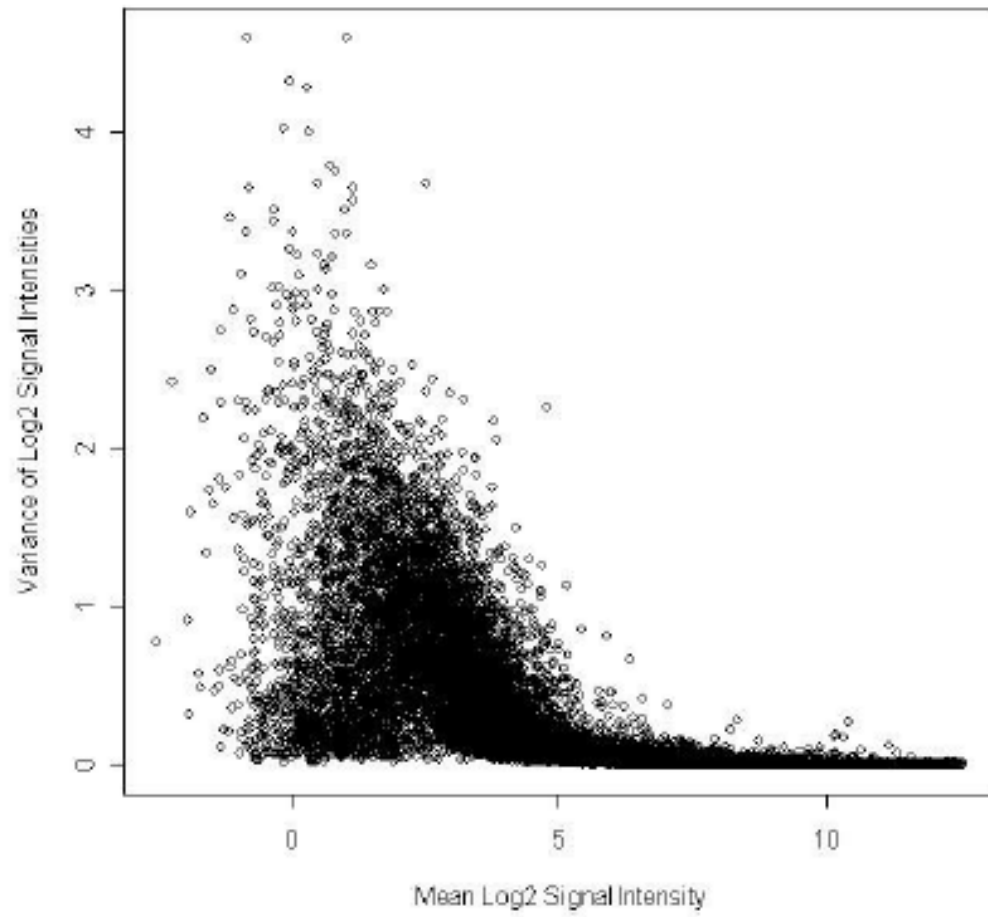
niveles bajos de expresión >--< varianza más alta
niveles altos de expresión >--< varianza más baja

Archer et al (2004) muestran esta situación.

Datos: 16 replicaciones técnicas

16 chips: Affymetrix GeneChip® HG-U133A .





Modelo aditivo multiplicativo

$$Y_{ik} = a_i + b_i X_{ik} e^{\eta_k + \varsigma_{ik}} + \varepsilon_k + \delta_{ik}$$

Y_{ik} intensidad medida

X_{ik} intensidad verdadera

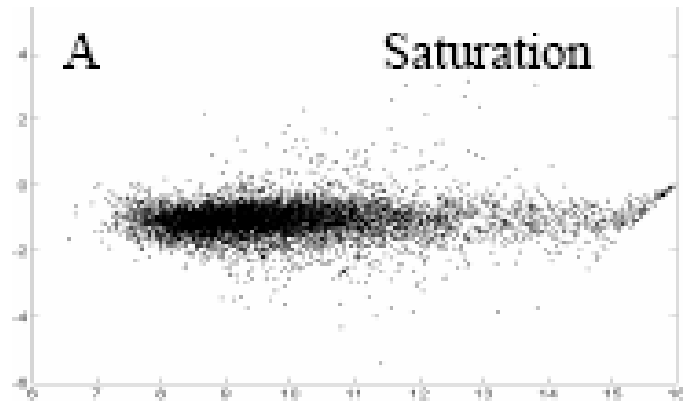
- **Componente multiplicativa:** etiquetado, escaneado y características del spot.
- **Componente aditiva:** relacionada con el background local.

Durbin et al. 2002, Rocke y Durbin (2001), Durbin et al. (2002) y Huber et al. (2002), Kerr-Churchill 2002

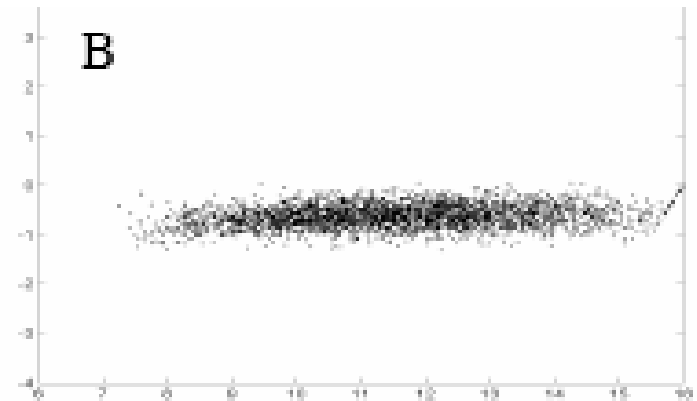
Cui et al 2003 simularon diferentes características observadas en los diagramas de dispersión M-A

utilizando diferentes valores de los parámetros del modelo

Real



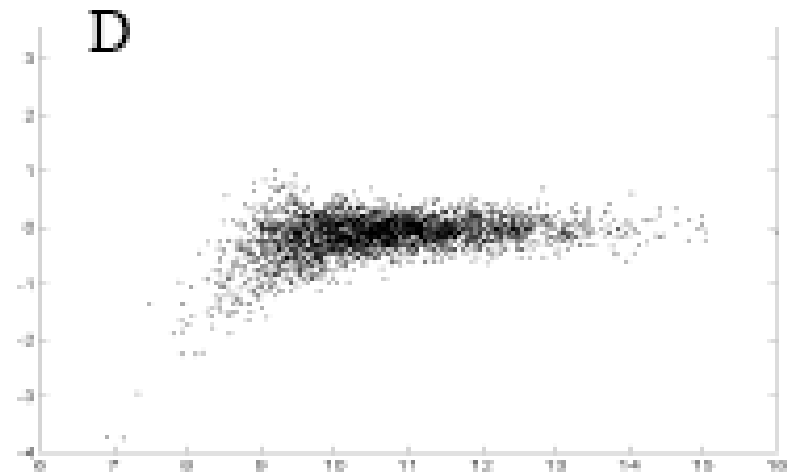
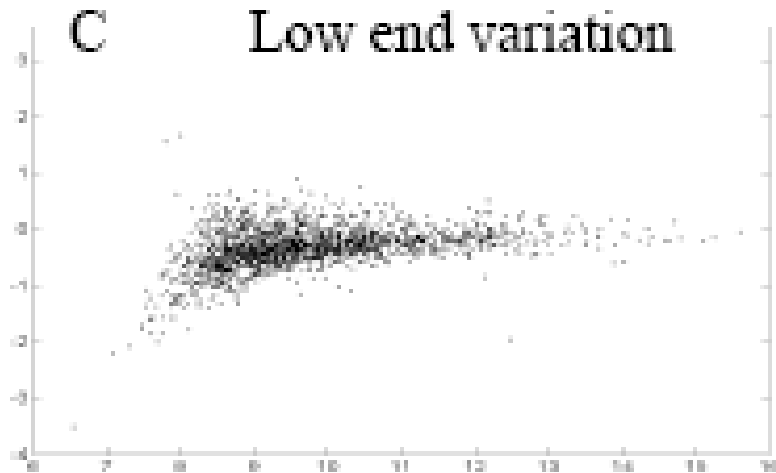
Simulado



El truncamiento en el lado de las intensidades altas es debido a que una cantidad moderada spots ha alcanzado la saturación del escaner

Real

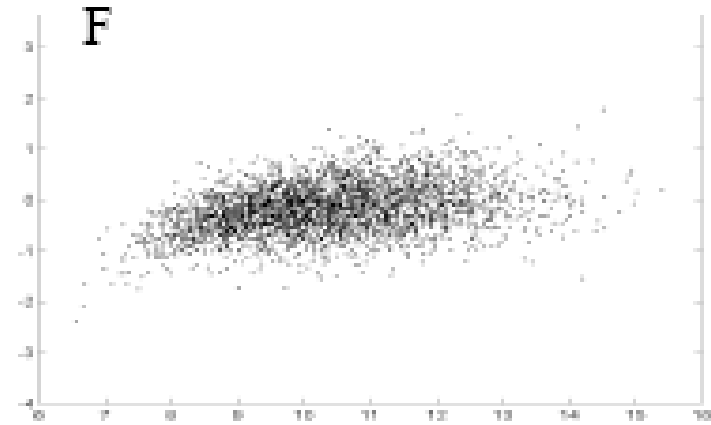
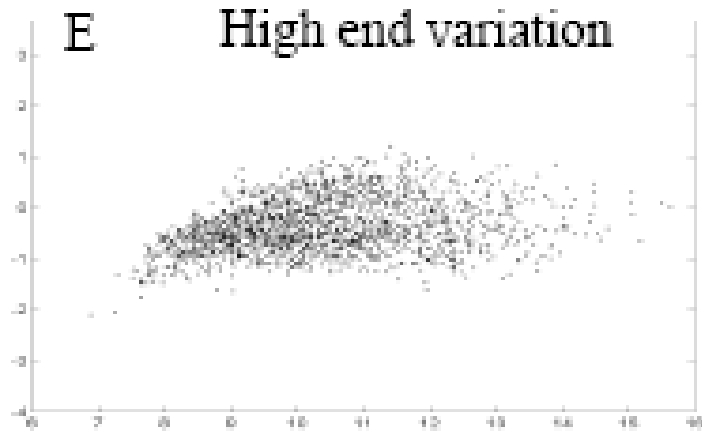
Simulado



La variación más alta en intensidades bajas puede simularse utilizando **errores aditivos grandes** específicos para cada canal

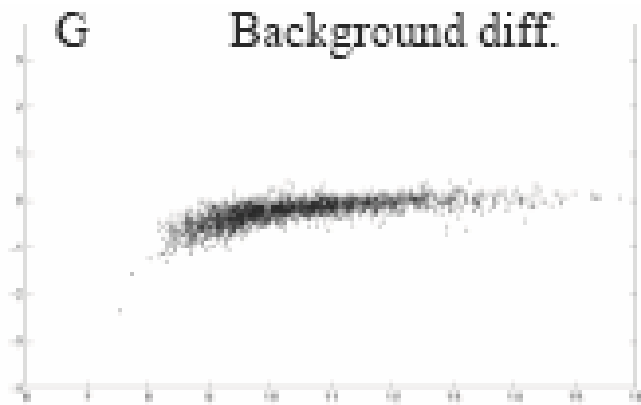
Real

Simulado

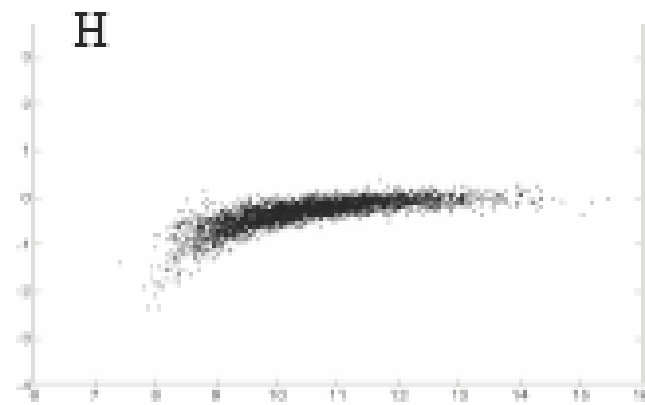


La variación mayor de los log ratios en intensidades altas se controlan por los **errores multiplicativos** específicos de cada canal con **varianzas altas**

Real



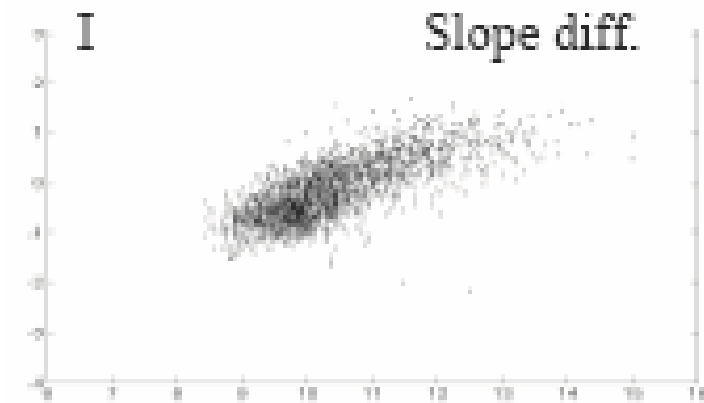
Simulado



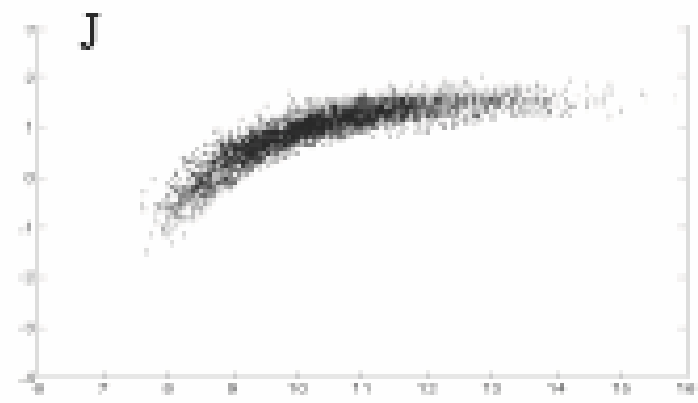
Background medio diferente entre los dos canales.

La curvatura es más pronunciada en la zona de intensidades bajas.

Real



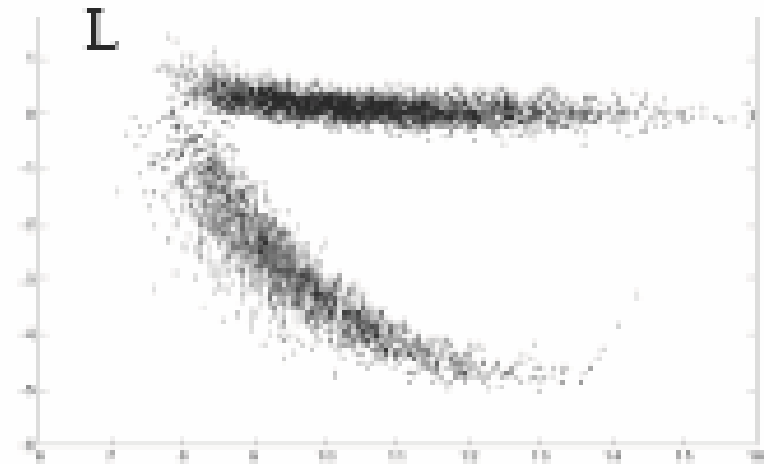
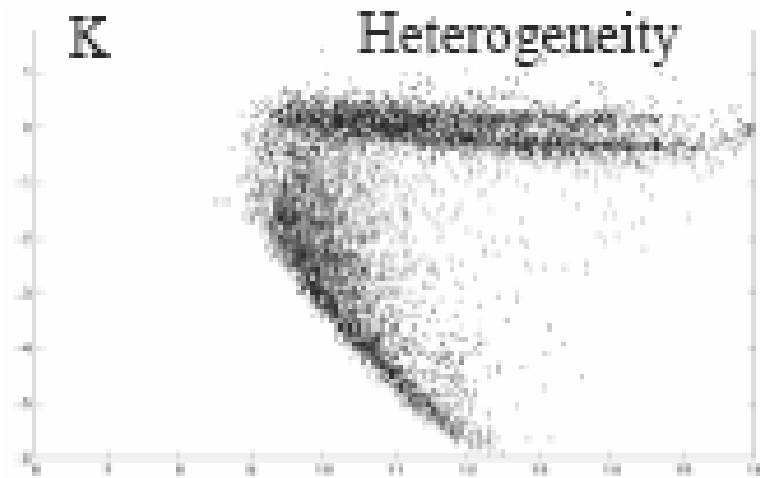
Simulado



La curvatura generada por **pendientes diferentes** se mantiene a lo largo de todo el rango de intensidades

Real

Simulado



Diferencias espaciales en las pendientes

¿Porqué es necesario modificar la transformación logarítmica?

Y_{ik} la intensidad medida en escala original
 Z_{ik} dato transformado.

Generalmente se utiliza la transformación logaritmo en base 2, de manera que

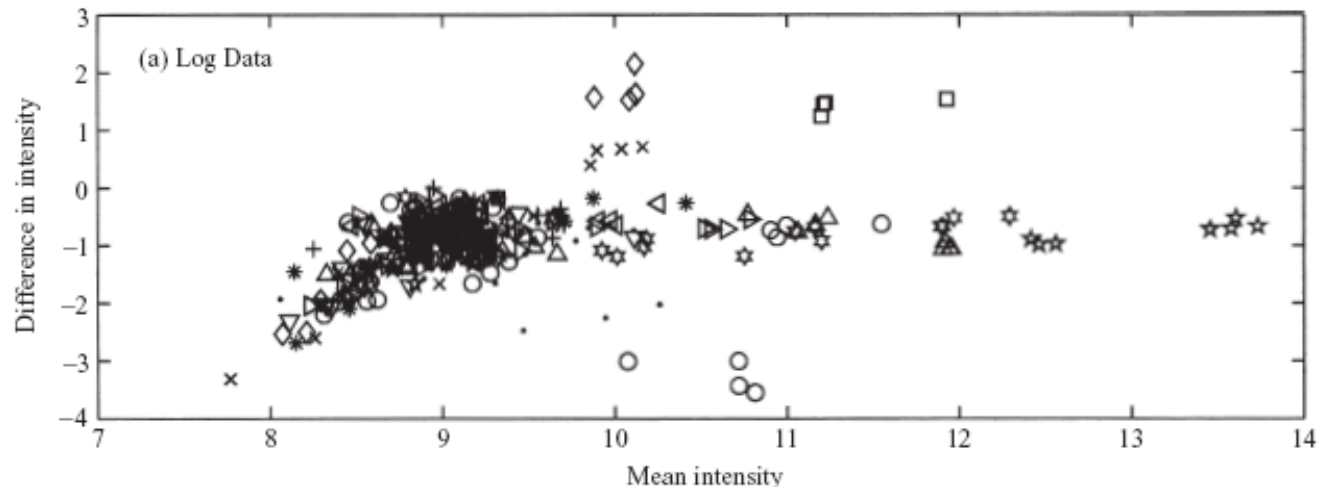
$$Z_{ik} = \log_2 (Y_{ik})$$

errores proporcionales a la señal $\rightarrow$ errores aditivos en la
escala logarítmica.

errores aditivos $\rightarrow$ un problema en escala logarítmica

Logaritmos corridos - shift log

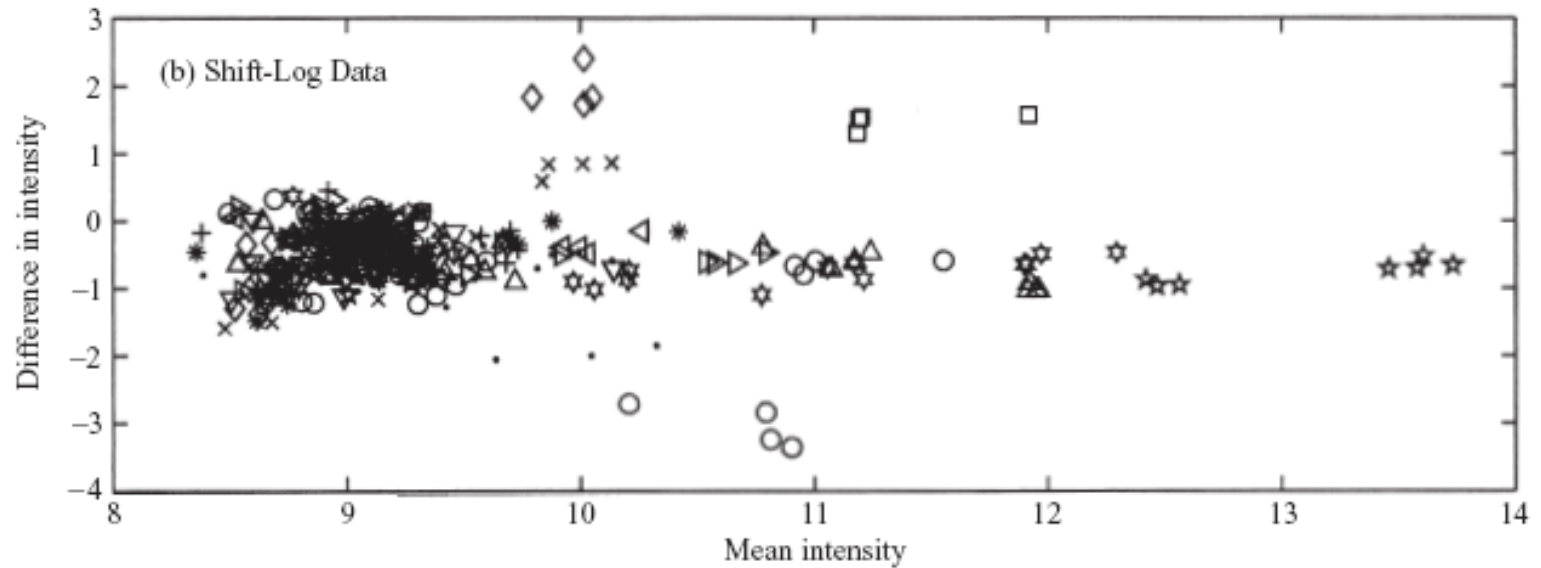
Kerr y Churchill (2001) microarreglo de 2 colores
curvatura en un M-A plot



Lo corrigen tomando

$$Z_{rk} = \log 2(Y_{rk} + C)$$

$$Z_{gk} = \log 2(Y_{rk} - C)$$



Shift-log es útil cuando la curvatura es debida únicamente a diferencias en la intensidad del fondo

Transformaciones afines

Bengtsson H y Hössjer O. (2006) muestran que

una transformación afin,

$$Y = a + b X,$$

de las intensidades permite

explicar y eliminar (normalizar) las dependencias no lineales dependientes de la intensidad observadas en los M-A plots.

Transformaciones por ajuste de curvas

- Es una de las técnicas más utilizadas.
- Permite eliminar las estructuras no atribuibles a factores biológicos en un M-A plot.

Consiste en ajustar una ó más curvas al gráfico M-A .
Los datos sobre esa curva $\leftrightarrow$ genes que no están expresados diferencialmente (Yang et al., 2002).

$$Z_{rk} = \log_2(Y_{rk}) + C_k / 2$$

$$Z_{gk} = \log_2(Y_{rk}) + C_k / 2$$

C_k es una constante que depende del spot.

Está determinada por un suavizado local (loess – lineal)

Transformación Linlog

$$Z_{ik} = \text{linlog}(Y_{ik}) = \begin{cases} \log_2(d_i) - 1/\ln 2 + Y_{ik}/(d_i \ln 2) & Y_{ik} < d_i \\ \log_2(Y_{ik}) & Y_{ik} \geq d_i \end{cases}$$

Cui et al 2003 - transformación estabilizadora de las varianzas basada en el modelo que sugiere que:

- el error aditivo es dominante para intensidades bajas
- el error multiplicativo es dominante para intensidades altas.

No corrige las curvaturas en los gráficos M-A.

Si se combina **linlog** con

- **shift-log** $\begin{cases} Z_{rk} = \text{linlog}(Y_{rk} + C) \\ Z_{gk} = \text{linlog}(Y_{gk} - C) \end{cases}$ ó
- una transformación loess

se estabilizará la varianza y minimizará la curvatura.

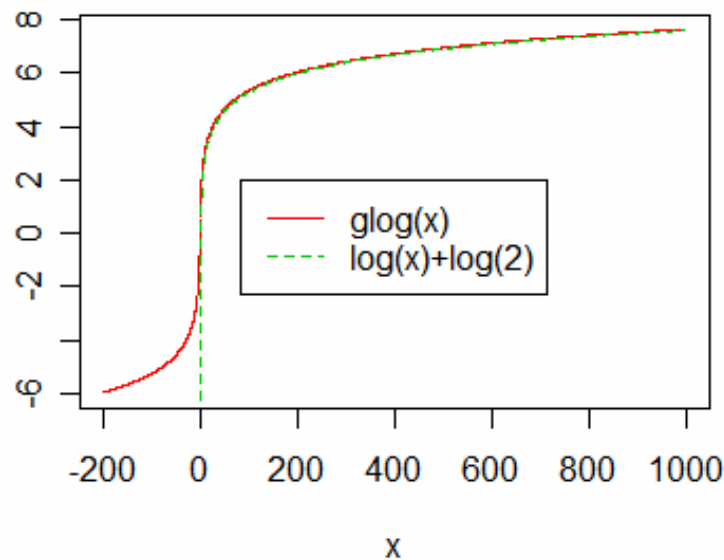
Transformación arcoseno hiperbólico - arsinh

Munson (2001), Durbin *et al.* (2002) y Huber *et al.* (2002) propusieron en forma independiente la transformación (Rocke and Durbin, 2003 la indican glog) para estabilizar la varianza de datos de micoarreglos que satisfacen el modelo aditivo multiplicativo.

La función arcoseno hiperbólico también llamada logaritmo generalizado está definida por

$$\text{glog}(x) = \text{arsinh}(x) = \log(x + \sqrt{x^2 + 1})$$

glog es una función similar al logaritmo natural para valores grandes, $\text{glog}(x) \approx \log(x) + \log(2)$



pero es menos empinada para valores pequeños

La transformación propuesta es

$$Z_{ik} = g \log(b_i Y_{ik} + C_i) = \log(b_i Y_{ik} + C_i + \sqrt{(b_i Y_{ik} + C_i)^2 + 1})$$

y está basada en una relación cuadrática entre la varianza y la intensidad de la señal en la escala original (Huber *et al.* (2002)).

Los parámetros son estimados por máxima verosimilitud mediante un procedimiento robusto.

La corrección de la curvatura de esta transformación introduce 4 parámetros para cada arreglo de 2 colores, no es una manipulación muy fuerte.

Para los valores altos de intensidad $g \log$ es similar al $\log$.

Esta propuesta ha sido criticada (Speed 2003) por su comportamiento en los valores altos, debido a que en ese rango es una transformación muy severa, **invirtiendo la tendencia** de la heterogeneidad de las varianzas

Transformación basada en el gráfico de dispersión nivel.

Archer *et al.* (2004) proponen, para datos de microchips de Affymetrix, elegir la transformación dentro de la familia de funciones de potencia de Box- Cox utilizando el

Spread-versus-Level (SVL) Plot:

Abscisas - log (mediana) de las intensidades replicadas

Ordenadas - distancia inter cuartos que es una variante de la distancia intercuartil

La potencia que da la transformación de Box-Cox adecuada está indicada por 1- pendiente de la recta ajustada a dicho gráfico (Emerson JD 1987).

Potencias más usadas, correspondientes a los miembros principales de la “escalera de potencias” de Tukey.

Transformación	Potencia	Pendiente del gráfico dispersión - nivel
Cúbica	3	-2
Cuadrada	2	-1
No cambio	1	0
Raíz cuadrada	1/2	1/2
Logaritmo	0	1
Inversa de la raíz cuadrada	-1/2	3/2
Inversa	-1	2

Funciones de potencia de Box- Cox

$$Z(p) = \begin{cases} \frac{Y^p - 1}{p} & p \neq 0 \\ \log Y & p = 0 \end{cases}$$

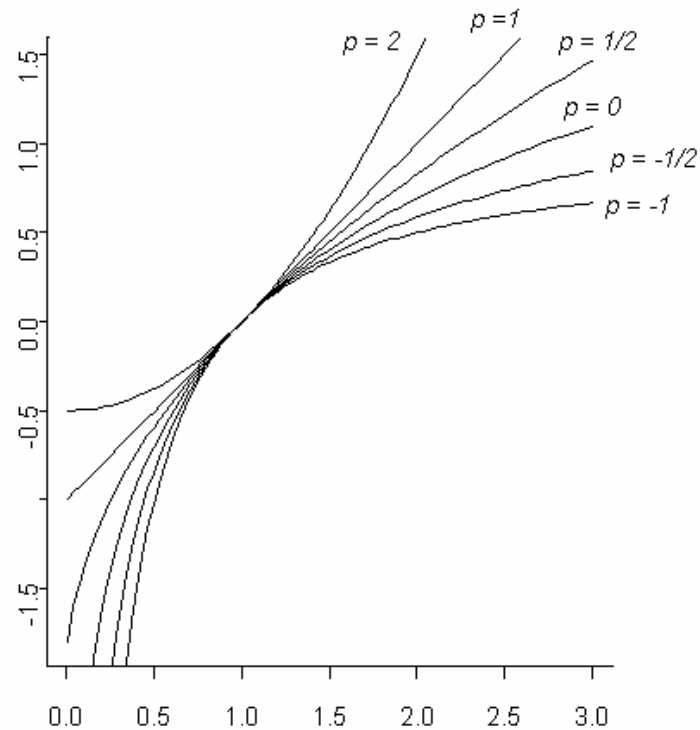
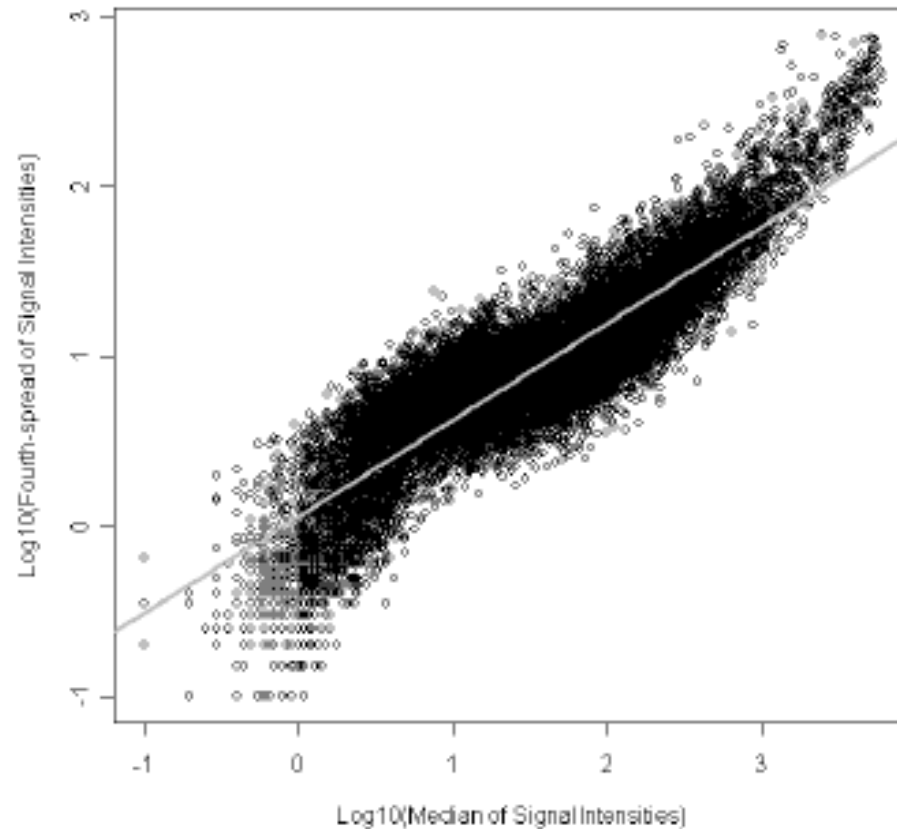
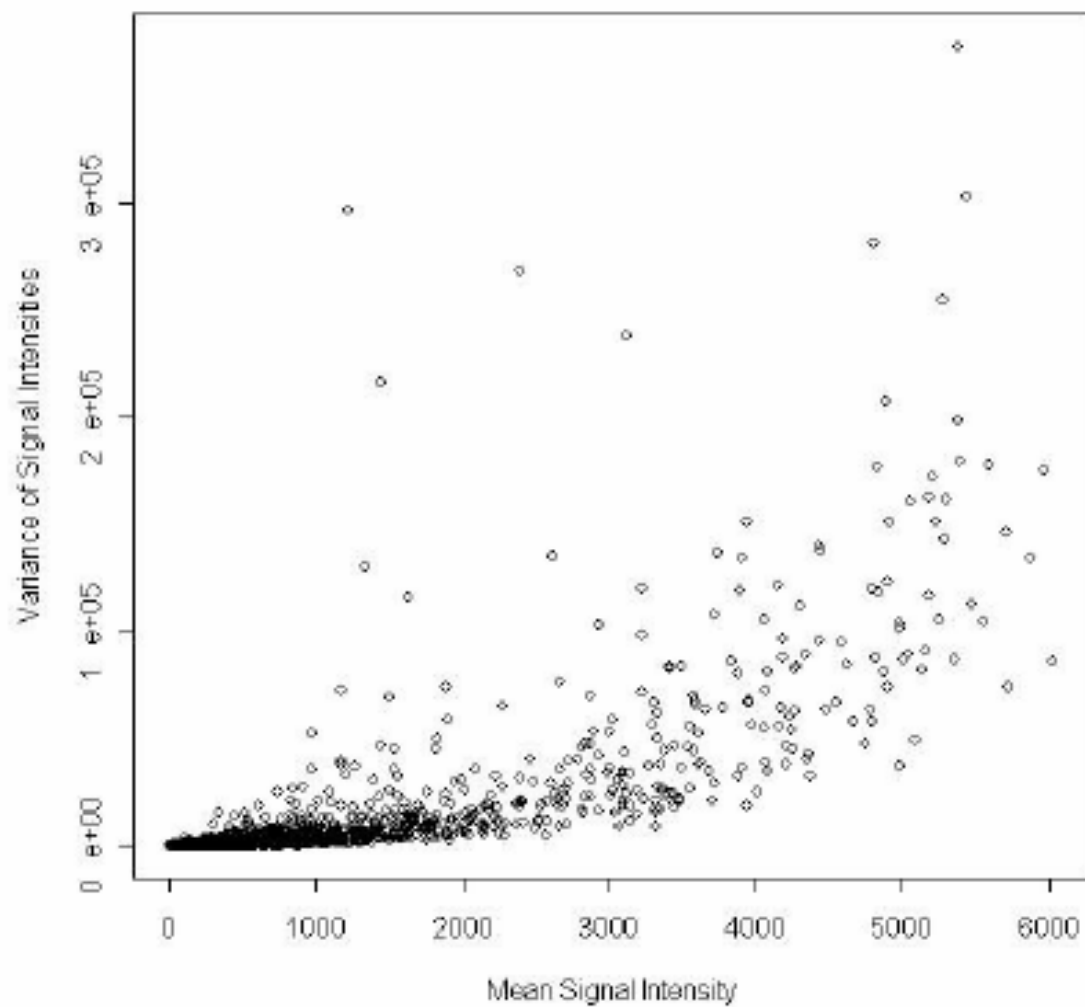


Gráfico de dispersión-nivel



Archer *et al.* (2004). Pendiente estimada por cuadrados mínimos 0.569

Datos originales



Datos transformados

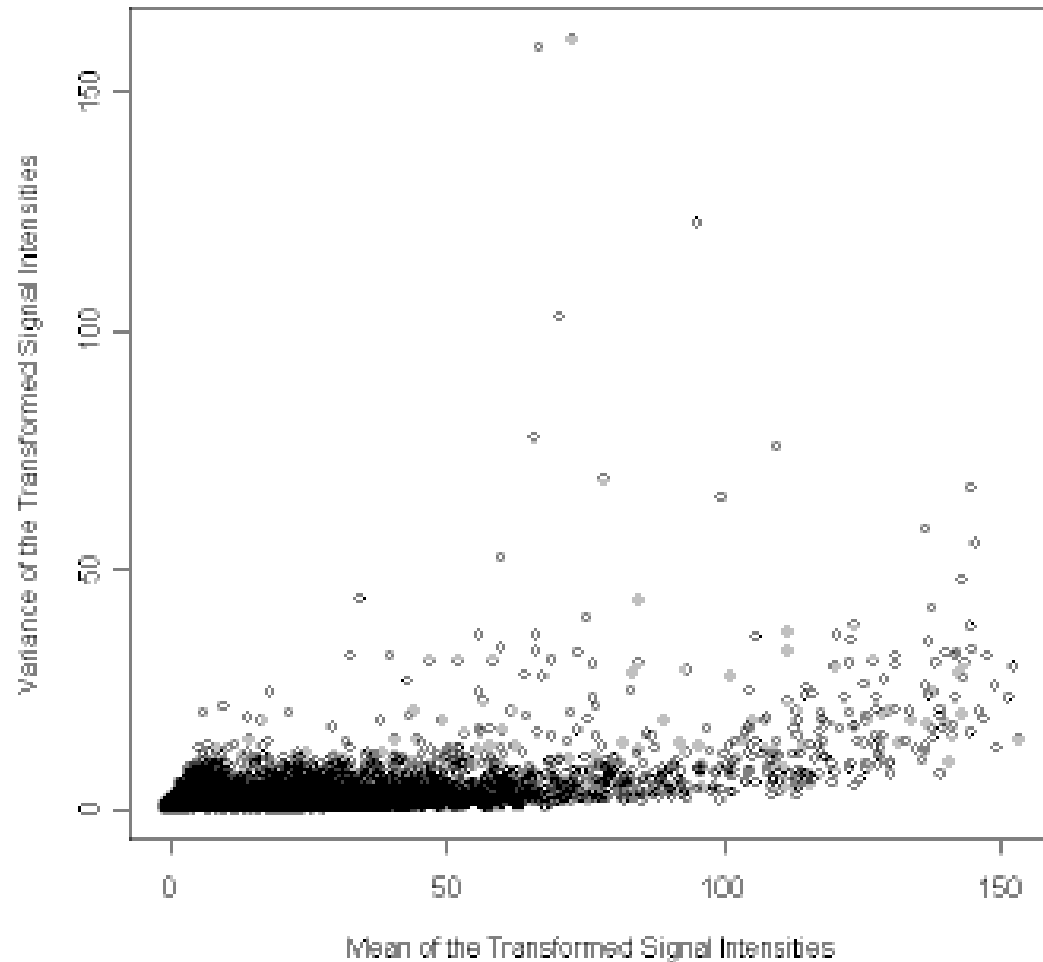


Gráfico de la varianza en función de la media para los datos transformados por $2\sqrt{x} - 2$

Datos transformados

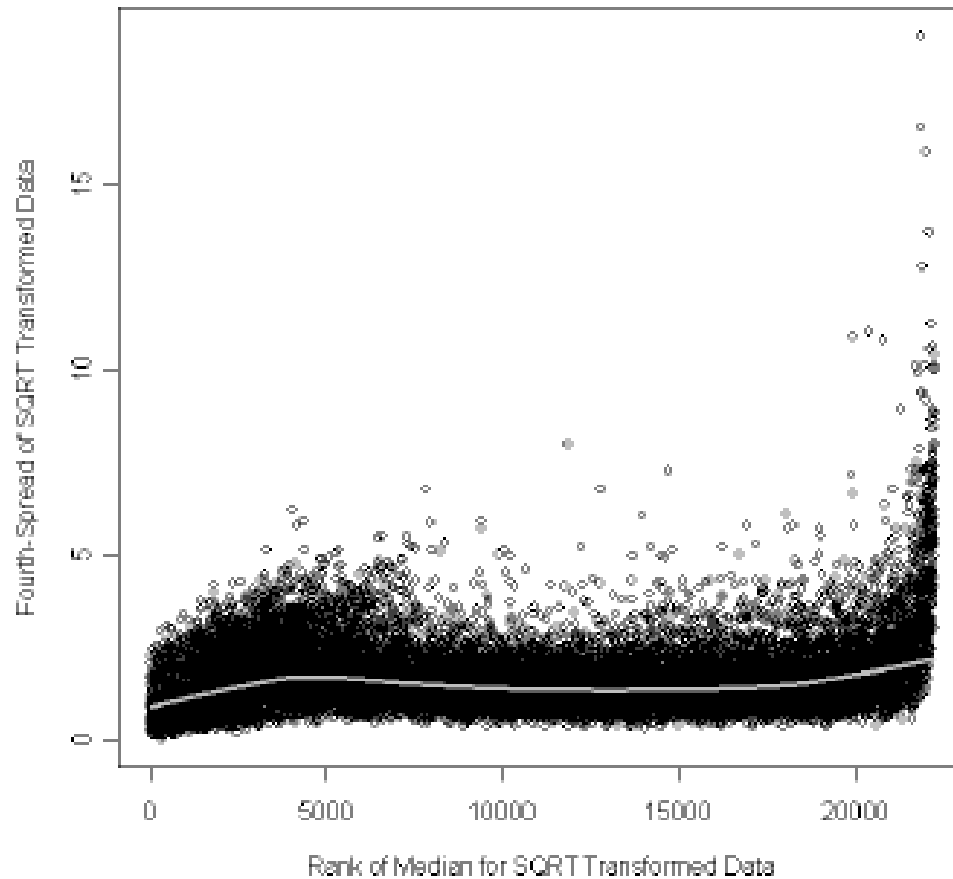
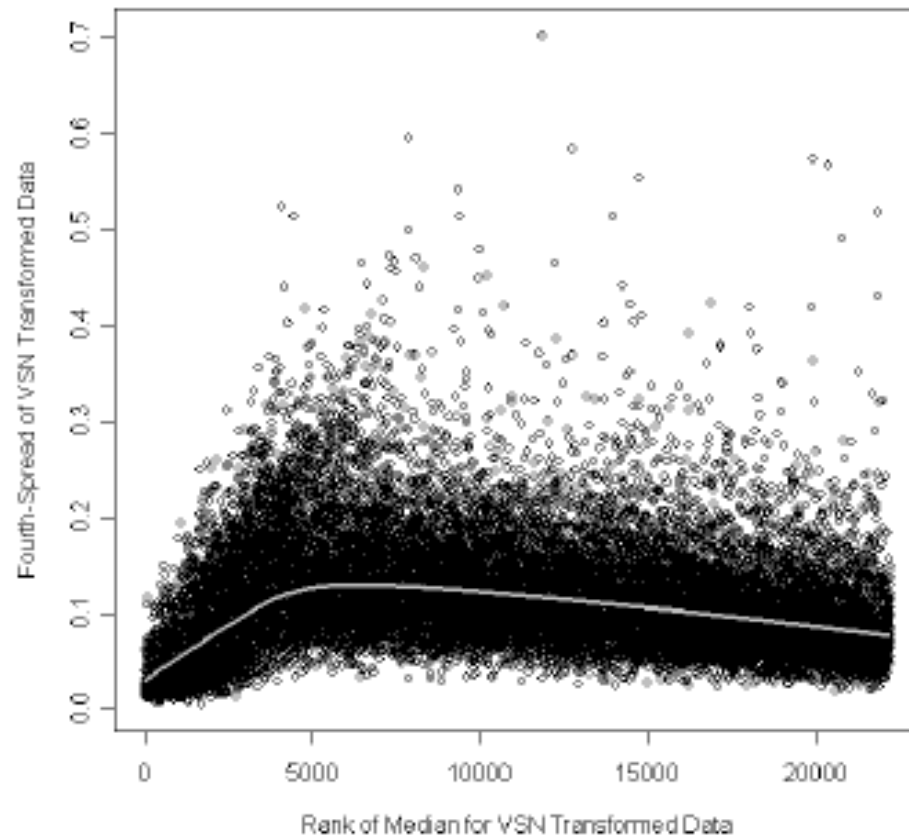


Gráfico de las distancias intercuartos en función del rango de la mediana, para los datos transformados por la raíz cuadrada

Datos transformados por glog - vsn



La transformación glog es demasiado fuerte para los valores con intensidades altas.

Referencias

Archer, K.J., Dumur, C.I. and Ramakrishnan, V. (2004) Graphical technique for identifying a monotonic variance stabilizing transformation for absolute gene intensity signals. *BMC Bioinformatics*, 5:60.

Bengtsson H, Hössjer O. (2006) Methodological study of affine transformations of gene expression data with proposed robust non-parametric multi-dimensional normalization method. *BMC Bioinformatics*. 2006; 7: 100. Published online 2006 March 1. doi: 10.1186/1471-2105-7-100.

Cui, Xiangqin; Kerr, M. Kathleen; and Churchill, Gary A. (2003) "Transformations for cDNA Microarray Data," *Statistical Applications in Genetics and Molecular Biology*: Vol. 2 : Iss. 1, Article 4.

Available at: <http://www.bepress.com/sagmb/vol2/iss1/art4>

Referencias

Durbin, B.P., Hardin, J.S., Hawkins, D.M. and Rocke, D.M. (2002) A variance stabilizing transformation for gene expression microarray data. *Bioinformatics*, **18**, S105–S110.

Durbin, B.P. and Rocke, D.M. (2003) Estimation of transformation parameters for microarray data, *Bioinformatics*, **19**, 1360–1367.

Emerson JD: **Mathematical Aspects of Transformation.** *Understanding Robust and Exploratory Data Analysis* . Edited by Hoaglin DC, Mosteller F, Tukey JW. New York: John Wiley & Sons;; 1987:247-282.

Kerr, M. K., Martin, M. & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819-37.

Kerr, M.K., Martin, M. and Churchill, G.A. (2001). Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, 77, 123--128.

Referencias

Kerr MK, Afshari CA, Bennett L, Bushel B, Martinez J, Walker NJ, Churchill GA (2002). Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sinica* 12:203– 217

Holder, D., Raubertas, R.F., Pikounis, V.B., Svetnik, V. and Soper, K. (2001) Statistical analysis of high density oligonucleotide arrays: a SAFER approach. GeneLogic Workshop on low level analysis of Affymetrix GeneChip data, Nov. 19, Bethesda, Maryland.

[http://128.32.135.2/users/terry/zarray/Affy/GL Workshop/SAFERv04.pdf](http://128.32.135.2/users/terry/zarray/Affy/GL%20Workshop/SAFERv04.pdf)

Hoyle DC, Rattray M, Jupp R, Brass A (2002). Making sense of microarray data distributions. *Bioinformatics* 18:576–584

Huber, W., Von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96-S104

Referencias

Huber, W., Von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Statist. App. Gen. Mol. Biol.*, **2**, Issue 1, Article 3.

Munson, P. (2001) A "consistency" test for determining the significance of gene expression changes on replicate samples and two-convenient variance-stabilizing transformations. GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data, Nov. 19, Bethesda, Maryland.

T. Speed - Editor. "Statistical Analysis of Gene Expression Data". 2003 . Chapman&Hall

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30:e15