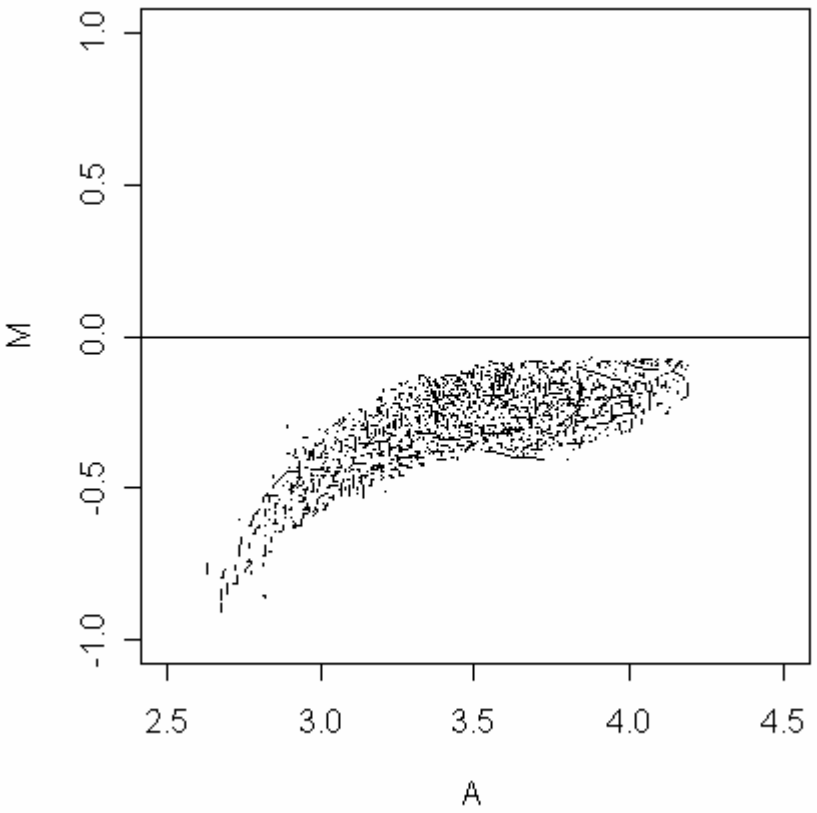
7. Normalización

El objetivo de la normalización es identificar y eliminar variaciones sistemáticas conservando la señal biológica. Las principales fuentes de estas variaciones se encuentran en diferencias entre los flúores Cy3 y Cy5 respecto de sus eficiencias de etiquetado y propiedades de escaneado, diferencias en los parámetros de escaneado, efectos de la aguja (print-tip) y del plato. Mediante la normalización se pretende asegurar que las diferencias en intensidad realmente reflejan la expresión diferencial de los genes y que no haya sesgos artificiales debido a factores técnicos.

Recordemos la notación M A:

intensidad del log–ratio $M = \log_2 R/G$
intensidad total $A = 0.5 \log_2 (RG)$,

Para visualizar la necesidad de la normalización nos remitiremos al trabajo Tseng et al. (2001). En él se muestra un MA plot de un experimento de calibración en el cual la misma muestra ha sido hibridada al canal rojo y al canal verde de un mismo microarray. La lectura del escaner debería ser la misma para ambos canales si no hubiese efecto de la tintura. En este caso ideal se espera que la nube de puntos en un MA plot no presente tendencias y sea pareja alrededor de la recta M=0. La figura siguiente es un esquema de lo encontrado en dicho trabajo. Es habitual que los experimentos self-self presenten estructuras en un MA plot



En la figura se pone de manifiesto el sesgo del tinte hacia el canal verde Cy3 para intensidades bajas. Esto es que para intensidades bajas los valores de Cy3 son mayores de lo esperado.

Una característica para tener en cuenta es que este sesgo depende del vidrio, esto es que, cuando un mismo lote de RNA es utilizado como muestra para hibridar distintos vidrios se obtuvieron otro patrones.

Los procedimientos de normalización más complejos tienden a eliminar sesgos técnicos en mayor medida que procedimientos simples. Estos podrían también eliminar más señal e introducir nuevas fuentes de variación debidas, por ejemplo, a la incerteza con que los parámetros óptimos de normalización han sido estimados. La elección entre los distintos procedimientos de normalización no es sencilla y actualmente se realiza ad-hoc. Una propuesta para alcanzar algún tipo de automatización está dada en Xiao et. al 2005

En un experimento ideal las normalizaciones no deberían ser necesarias. Sin embargo los sesgos técnicos son inevitables y pueden ser corregidos utilizando un procedimiento de normalización. La decisión respecto de corregir un vidrio vía normalizaciones o descartarlo no es simple.

Consideraremos dos tipos de métodos de normalización:

- *Métodos de normalización de dos canales*: se ajustan los valores de M utilizando los valores de A y otros factores como print-tip, plato, posición espacial como covariables.

- *Métodos de normalización para cada canal por separado*: se aplican a las intensidades originales utilizando los mismos factores que antes como covariables pero no A.

7.1 Normalización conjunta de dos canales

El proceso de normalización de dos canales puede separarse en dos componentes: posición ó locación (l) y escala (s). En general se estiman esas dos componentes de la distribución de los log-ratios (M) en función de las intensidades (A) de los dos canales o en función de las coordenadas espaciales del microarreglo.

Los log-ratios normalizados M_{norm} en general están dados por

$$M_{norm} = (M - l) / s$$

Los métodos difieren en cómo se obtienen l y s. Por ejemplo en la normalización global por la mediana el parámetro de localización se asume constante, es decir que es el mismo para todos los spots, mientras que en una normalización global dependiente de la intensidad (A) el parámetro de localización se supone que es una función suave de A y la función es estimada utilizando la opción robusta de la función **`loess`** para el suavizado en diagramas de dispersión.

La tabla siguiente muestra los procedimientos básicos de normalización para posición en microarreglos de dos canales.

Tabla: Procedimientos de normalización de localización

Nombre	**Descripción**
None	No se realiza normalización
Median	Normalización global por mediana
Loess	Normalización global dependiente de A utilizando el scatter plot smoother loess
Print-tip-loess	Normalización A-dependiente, dentro del grupo que determina cada aguja utilizando el scatter plot smoother loess
twoD loess	Normalización espacial, dentro del grupo que determina cada aguja, utilizando la función loess
scalePrintTipMAD	Normalización A-dependiente, dentro del grupo de cada aguja utilizando el scatter plot smoother loess, seguida de una normalización de escala dentro de cada grupo utilizando la función MAD

Dado un conjunto de números x1, . . . , xn, el MAD es la mediana de sus desvíos absolutos respecto de la mediana, es decir que si

$$m = \text{mediana}\{x1, \ldots, xn\}$$

entonces

$$MAD = \text{mediana}\{|x1 - m|, \ldots, |xn - m|\}.$$

La función de R para MAD es mad.

La función **maNorm** del paquete **marray** realiza los procedimientos de normalización descriptos en la tabla.

Del help del R tenemos

Uso:

maNorm(**mbatch**, **norm**=c("printTipLoess", "none", "median", "loess", "twoD", "scalePrintTipMAD"), **subse**t=TRUE, **span**=0.4, **Mloc**=TRUE, **Mscale**=TRUE, **echo**=FALSE, ...)

Veamos los argumentos principales:

mbatch: Objecto de clase 'marrayRaw', con los datos lote de arreglos para normalizar. Un objeto de clase "'marrayNorm"' también puede ser pasado si la normalización se realiza en varios pasos.

norm: Una cadena de caracteres especificando el procedimiento. Puede especificarse mediante la primera letra de cada procedimiento: “p”,”n”, “m”, “l”, “t”, “s”.

subset: Un vector “lógico” o “numérico” indicando el subconjunto de puntos para calcular los valores normalizados.

span: Este argumento controla el grado de suavizado en la función loess.

Mloc: Si vale 'TRUE', los valores de la normalización de localización (l) que devuelve la función son guardados en el **slot 'maMloc**' del objecto de clase **"marrayNorm**", si vale 'FALSE', estos valores no son guardados

Mscale: Si vale 'TRUE', los valores de la normalización de escala (s) son guardados en el **slot** '**maMscale**' del objecto de clase “**marrayNorm**", si vale 'FALSE', estos valores no son guardados.

echo: Si vale 'TRUE", el índice del arreglo siendo normalizado aparece en pantalla

La función **maNorm** devuelve un objeto de clase “**marrayNorm”** conteniendo los datos de intensidad normalizados.

La clase “**marrayNorm”** se utiliza para almacenar los datos pos-normalización de un lote de microarreglos de cDNA. Contiene slots para A, los log-ratios normalizados que seguiremos llamando M, y los valores de normalización de posición y escala, la geometría (layout) de los arreglos y la descripción de las muestras hibridadas y las secuencias spoteadas al arreglo.

La siguiente instrucción realiza una normalización Print-tip-loess

```
> beta7norm <- maNorm(beta7, norm= ”p”)

> summary(beta7norm)
```

```
Normalized intensity data:       Object of class marrayNorm.

Call to normalization function:
maNormMain(mbatch = mbatch, f.loc = list(maNormLoess(x = "maA",
    y = "maM", z = "maPrintTip", w = NULL, subset = subset, span =
span,
    ...)), Mloc = Mloc, Mscale = Mscale, echo = echo)

Number of arrays:       6 arrays.

A) Layout of spots on the array:
Array layout:    Object of class marrayLayout.

Total number of spots:                  23184
Dimensions of grid matrix:              12 rows by 4 cols
Dimensions of spot matrices:            23 rows by 21 cols

Currently working with a subset of 23184spots.
```

```
Control spots:
There are   5 types of controls :

  Buffer    Empty Negative Positive   probes
       3     1328      225      204    21424


Notes on layout:


B) Samples hybridized to the array:
Object of class marrayInfo.

       maLabels     FileNames SubjectID  Cy3  Cy5 Date of Blood Draw
1 6Hs.195.1.gpr 6Hs.195.1.gpr         1 b7 - b7 +         2002.10.11
2   6Hs.168.gpr   6Hs.168.gpr         3 b7 + b7 -         2003.01.16
3   6Hs.166.gpr   6Hs.166.gpr         4 b7 + b7 -         2003.01.16
4 6Hs.187.1.gpr 6Hs.187.1.gpr         6 b7 - b7 +         2002.09.16
5   6Hs.194.gpr   6Hs.194.gpr         8 b7 - b7 +         2002.09.18
6 6Hs.243.1.gpr 6Hs.243.1.gpr        11 b7 + b7 -         2003.01.13
  Date of Scan
1   2003.07.25
2   2003.08.07
3   2003.08.07
4   2003.07.18
5   2003.07.25
6   2003.08.06

Number of labels:  6
Dimensions of maInfo matrix:  6  rows by  6  columns

Notes:
 TargetBeta7.txt

C) Summary statistics for log-ratio distribution:
               Min. 1st Qu. Median  Mean 3rd Qu. Max. NA's
6Hs.195.1.gpr -6.47   -0.36  -0.01 -0.01    0.34 5.65 3415
6Hs.168.gpr   -6.21   -0.50  -0.01 -0.02    0.49 5.91 2839
6Hs.166.gpr   -6.68   -0.55  -0.02 -0.03    0.52 5.89 3440
6Hs.187.1.gpr -9.69   -0.28  -0.01 -0.01    0.26 5.20 2942
6Hs.194.gpr   -8.22   -0.36  -0.01 -0.03    0.33 5.73 6090
6Hs.243.1.gpr -5.58   -0.36  -0.01  0.08    0.41 8.39 2227

D) Notes on intensity data:
GenePix Data
>
```

Comparemos con los resultados del

```
> summary(beta7)
```

para los log-ratios de las intensides, de los datos sin normalizar

```
C) Summary statistics for log-ratio distribution:
               Min. 1st Qu. Median  Mean 3rd Qu. Max. NA's
6Hs.195.1.gpr -6.13   -1.00  -0.52 -0.50   -0.08 5.95 3415
6Hs.168.gpr   -7.08   -0.80  -0.21 -0.23    0.34 5.19 2839
6Hs.166.gpr   -7.07   -1.25  -0.64 -0.62   -0.02 6.15 3440
6Hs.187.1.gpr -9.81   -0.92  -0.60 -0.55   -0.25 5.00 2942
```

```
6Hs.194.gpr   -5.93    0.00   0.44  0.53    0.90 7.74 6090
6Hs.243.1.gpr -6.38   -1.13  -0.69 -0.64   -0.21 7.05 2227
```

La normalización por localización centra los log-ratios alrededor de cero al tener en cuenta los sesgos espaciales y de intensidad. ¿Por qué?

Los siguientes son los resultados de la normalización por mediana

```
>   beta7.normm <- maNorm(beta7, norm="median")
>  summary(beta7.normm)
Normalized intensity data:       Object of class marrayNorm.

Call to normalization function:
maNormMain(mbatch = mbatch, f.loc = list(maNormMed(x = NULL,
    y = "maM", subset = subset)), Mloc = Mloc, Mscale = Mscale,
    echo = echo)
```

.....

```
C) Summary statistics for log-ratio distribution:
               Min. 1st Qu. Median  Mean 3rd Qu. Max. NA's
6Hs.195.1.gpr -5.61   -0.48      0  0.02    0.44 6.47 3415
6Hs.168.gpr   -6.87   -0.59      0 -0.02    0.55 5.40 2839
6Hs.166.gpr   -6.43   -0.61      0  0.02    0.61 6.79 3440
6Hs.187.1.gpr -9.20   -0.32      0  0.05    0.35 5.60 2942
6Hs.194.gpr   -6.38   -0.44      0  0.09    0.46 7.30 6090
6Hs.243.1.gpr -5.68   -0.44      0  0.05    0.48 7.75 2227
```

Las normalizaciones anteriores, por localización, no ajustan por diferencia en escala entre microarreglos. La función que realiza una normalización por escala es la **maNormScale** del paquete **marray.** Se recomienda evaluar la necesidad de realizar esta normalización sobre la base de uncontrol caso por caso. En los casos en que la diferencia en escalas es pequeña puede ser preferible realizar unicamente una normalización por localización únicamente.

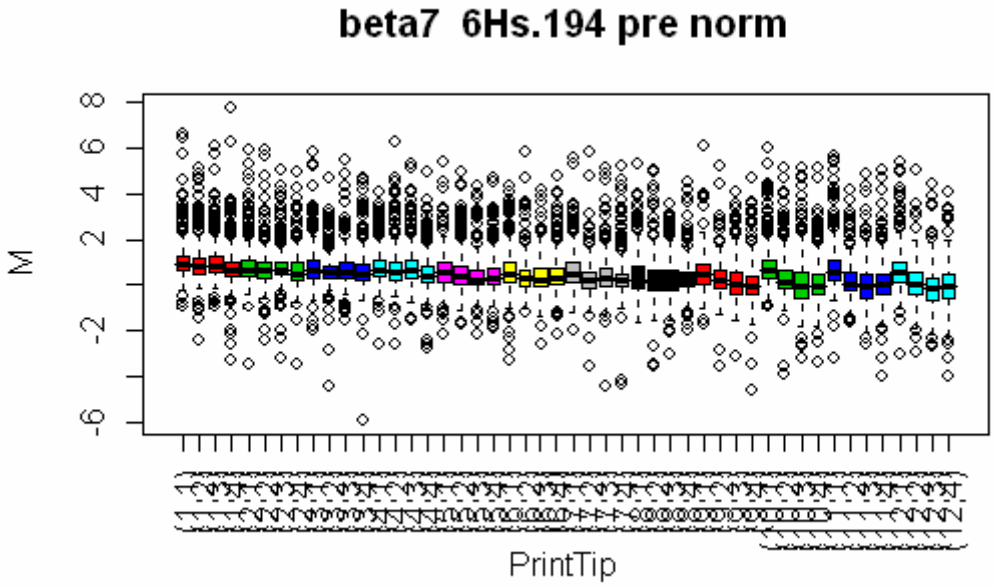
El siguiente código realiza una normalización por escala

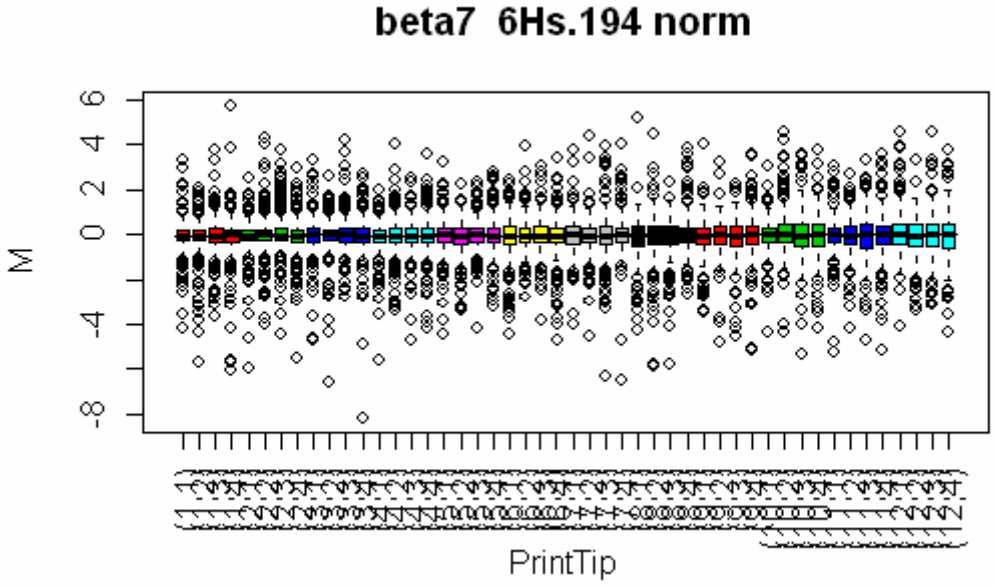
```
> beta7norm.escala <- maNormScale(beta7norm)
```

7.2 Visualización de los resultados de la normalización

Comparamos los datos pre y pos normalización

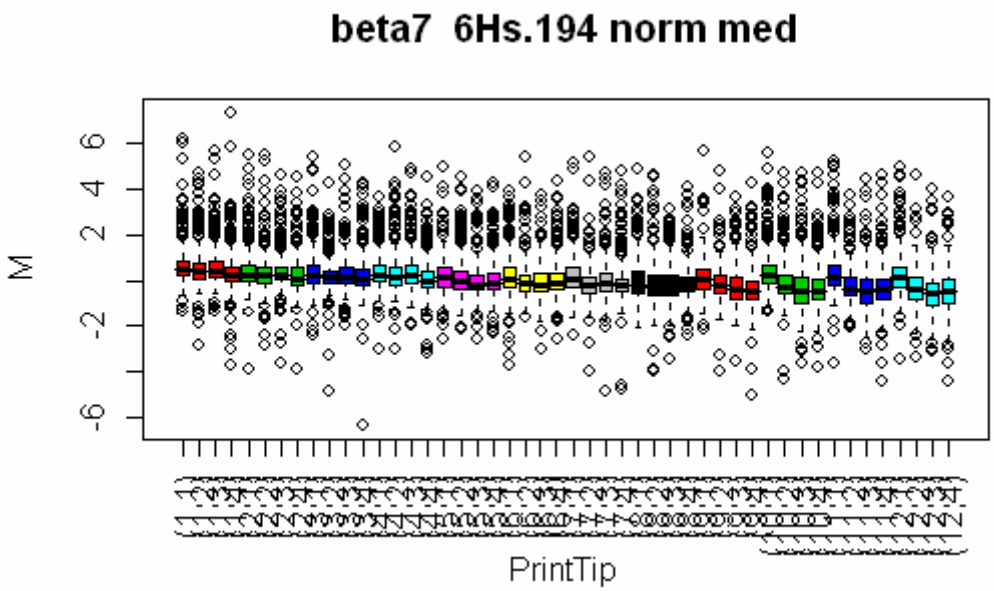
```
>  boxplot(beta7[, 5], xvar = "maPrintTip", yvar =
"maM",main="beta7  6Hs.194 pre norm")
```






```
> boxplot(beta7.norm[, 5], xvar = "maPrintTip", yvar =
"maM",main="beta7  6Hs.194 norm")
```

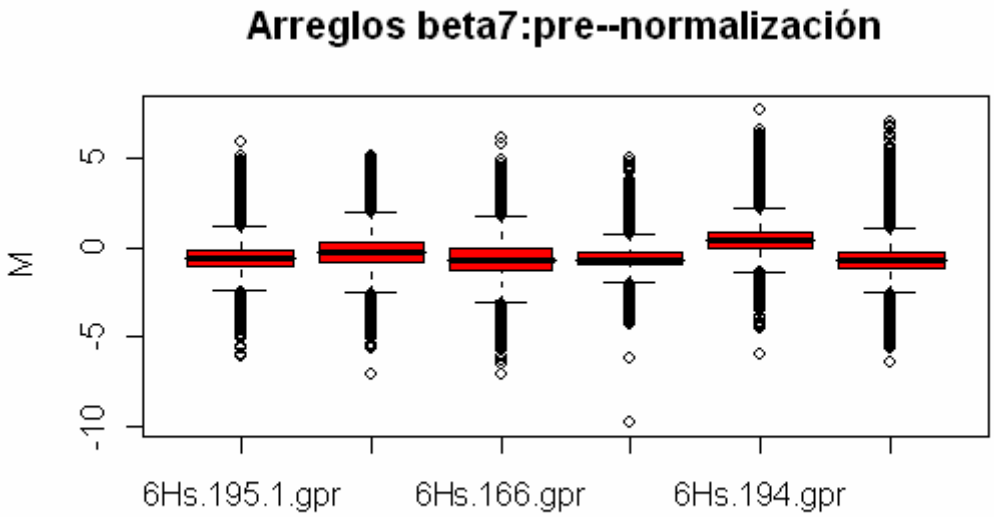
La normalización ha eliminado el sesgo por aguja.



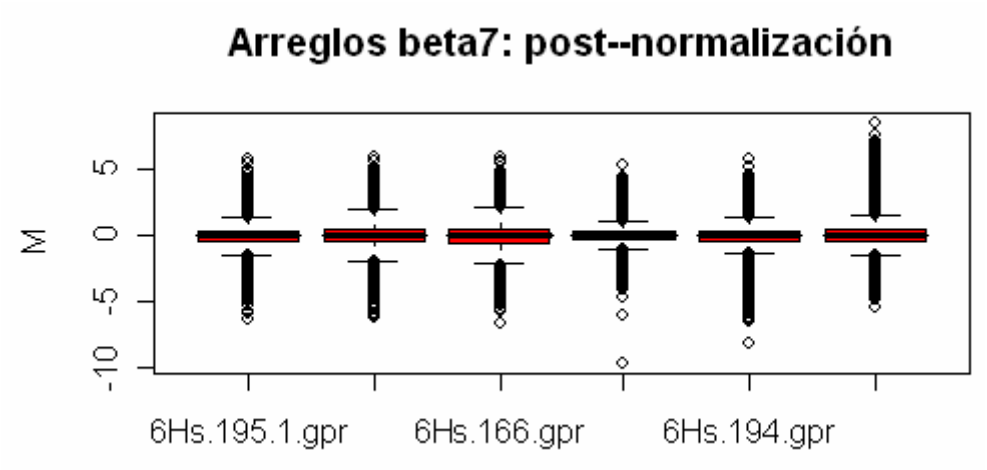
La normalización global por mediana no ha eliminado el sesgo por aguja.

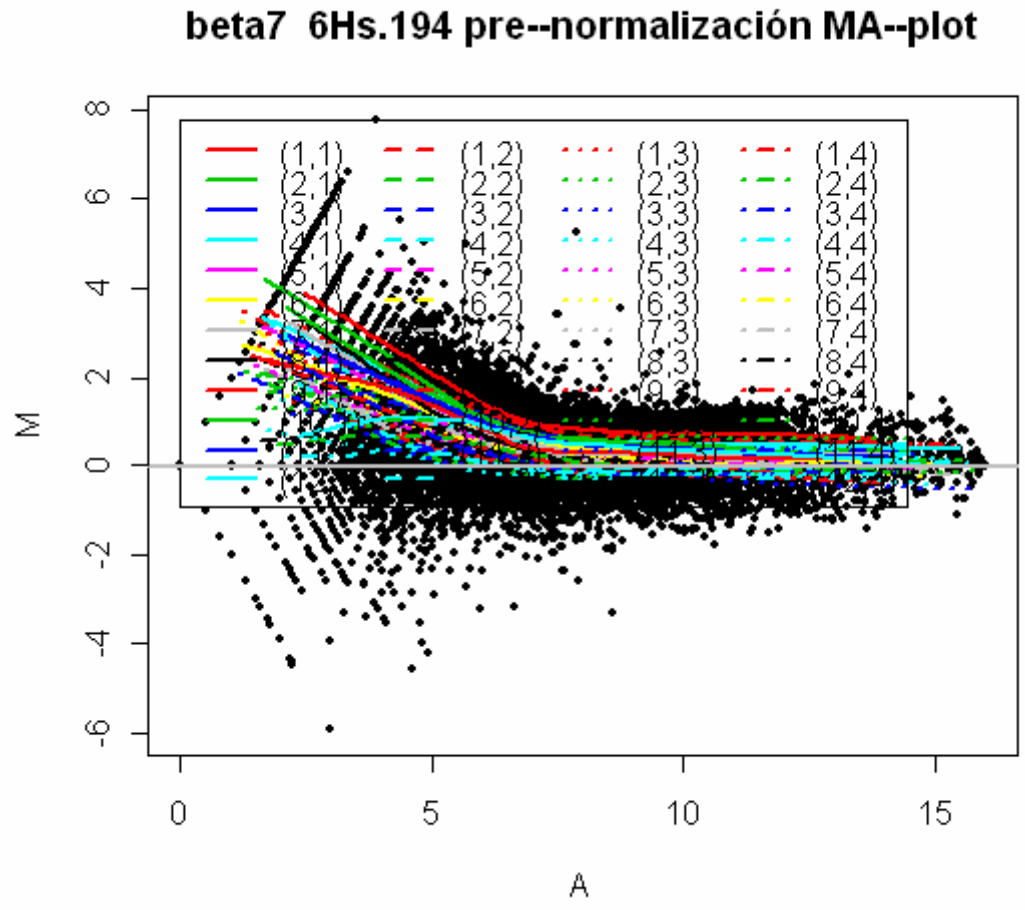
Veamos como se modifican las intensidades entre arreglos

```
>   boxplot(beta7, yvar = "maM", main = "Arreglos beta7:pre--
normalización")
```



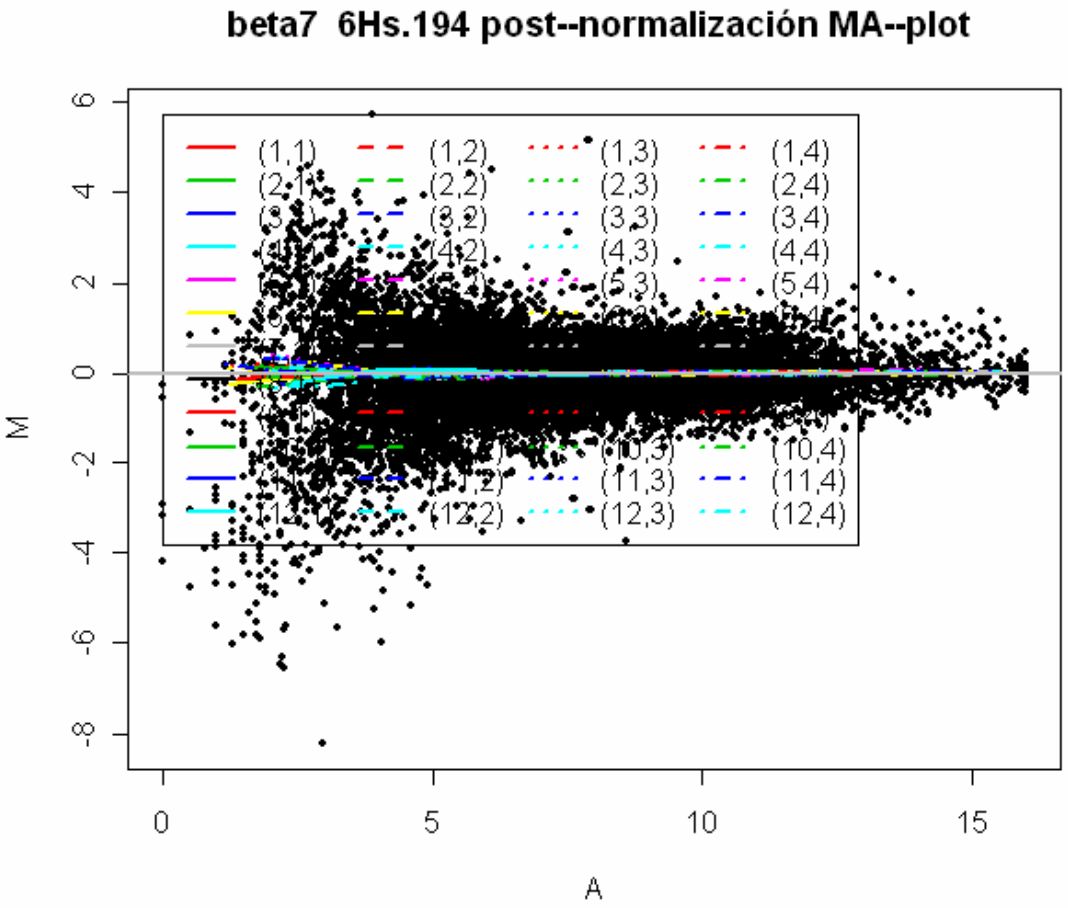

```
>  boxplot(beta7.norm, yvar = "maM", main = "Arreglos
beta7: post--normalización")
```






```
>   plot(beta7[, 5], main = "beta7  6Hs.194 pre--
normalización MA--plot")
```

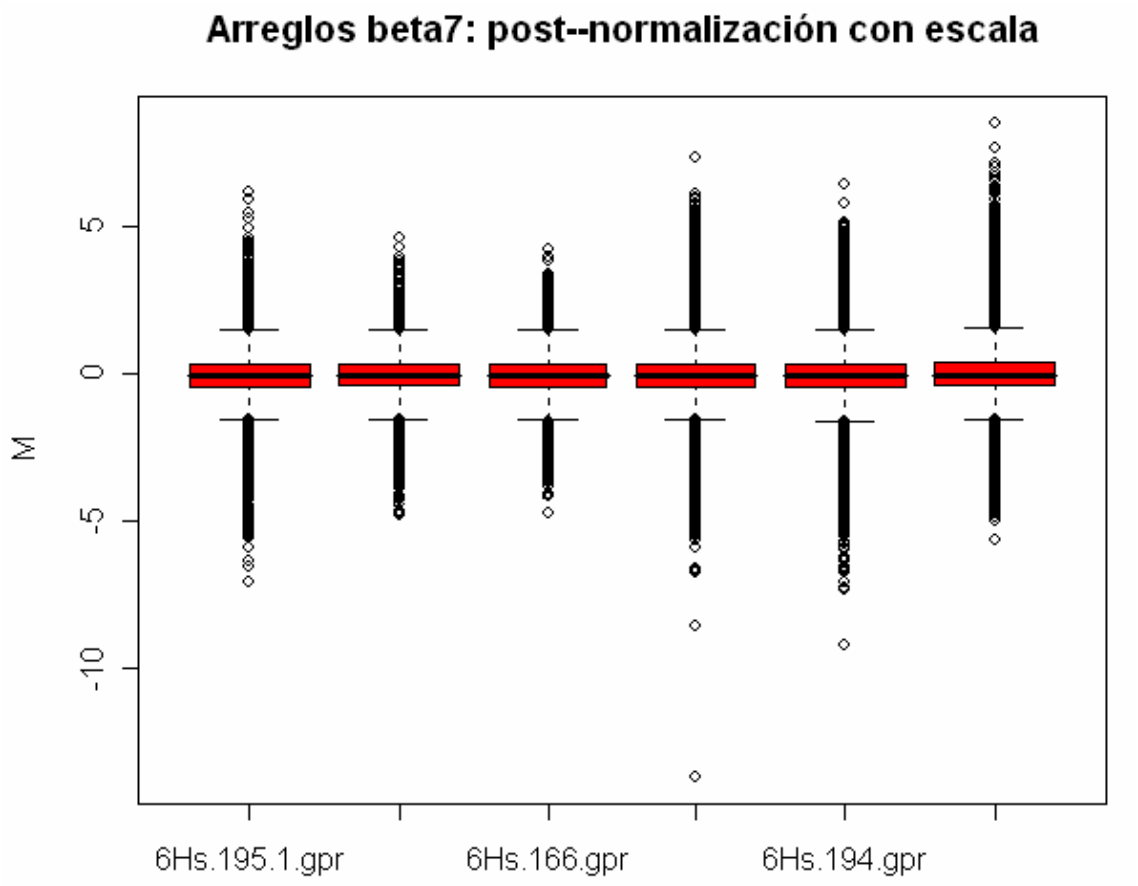
Se observa la tendencia de M en función de A, global y para cada print-tip



Se ha eliminado la tendencia de M en función de A, tanto globalmente como para cada print-tip

Veamos el efecto de la normalización por escala

```
> boxplot(beta7norm.escala, yvar = "maM", main = "Arreglos
beta7: post--normalización con escala")
```



Referencias

Cleveland, W.S. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, Vol. 74, pp. 829-836.

Cleveland, W.S. and Devlin, S.J. (1988) "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, Vol. 83, pp. 596-610.

Cleveland, W. and Loader, C. (1995) Smoothing by Local Regression: Principles and Methods (with discussion) Cleveland, W. and Loader, C. (1995): Computational Statistics

Maronna, R.A. , Martin, R.D. and Yohai, V.J. (2006). Robust Statistics: Theory and Methods, Wiley, West Sussex.

Yuanyuan Xiao, Mark R. Segal, and Yee Hwa Yang, "Stepwise Normalization of Two-Channel Spotted Microarrays" (November 5, 2004). *Center for Bioinformatics & Molecular Biostatistics.* Paper stepnorm5.
http://repositories.cdlib.org/cbmb/stepnorm5

Viñetas de Normalización del marray