3. Obtención y procesamiento de la imagen de un microarreglo

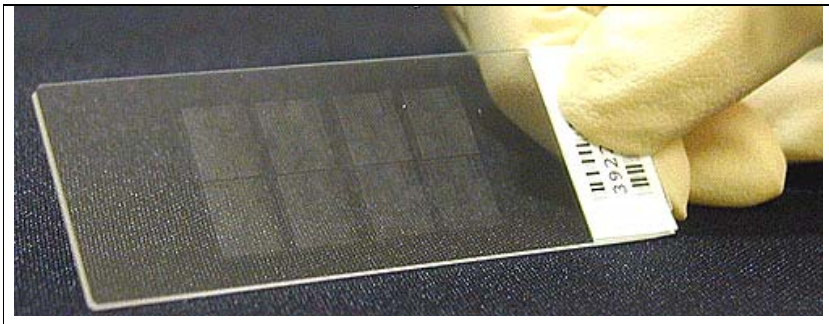
Una vez que se ha concluido la hibridación en un experimento de microarreglos el paso siguiente consiste en obtener una **imagen digital** que representa las intensidades de expresión de cada uno de los puntos (spots, manchas).

Obtención
Describiremos primero cómo es una imagen digital (sección 3.1), cómo se obtiene en general (sección 3.2) y luego veremos con más detalle la descripción para un microarreglo de dos canales (sección 3.3).

Procesamiento
Cada mancha con señal es representada por docenas de pixeles. Interesa obtener un valor global de la intensidad para cada spot. Sus correspondientes pixeles deben ser identificados (grillado y segmentación) y sus intensidades resumidas (cuantificación). Además a la intensidad global de cada punto con señal se calculan medidas auxiliares tales como una estimación de la intensidad aparentemente inespecífica (local background) y medidas de la calidad del spot (sección 3.4). En 3.5 veremos con más detalle la metodología propuesta por Yang et al en el 2002. En 3.6 describiremos algunas características específicas de las imágenes de microchips de alta densidad.

3.1 Imágenes digitales de un microarreglo

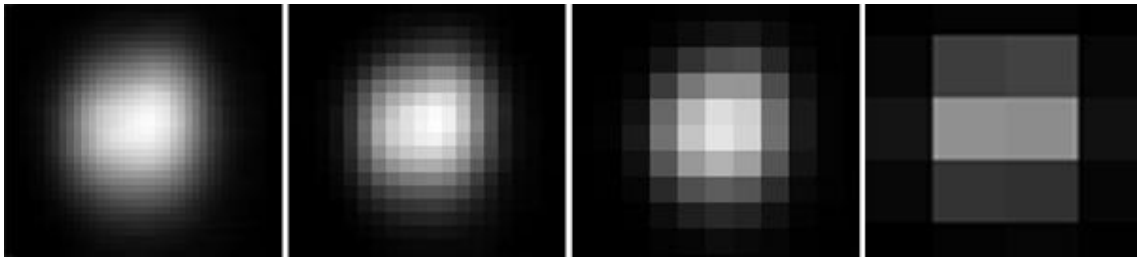
Los **spots** de los microarreglos son ***regiones fluorescentes*** localizadas sobre la superficie del sustrato.



http://www.kbrin.louisville.edu/archives/fellows/dobbins.html

gslc.genetics.utah.edu

Las *imágenes digitales* de los microarreglos permiten obtener *representaciones numéricas bidimensionales* del mismo a partir de intensidades de fluorescencia. Cada número corresponde a la intensidad de un *pixel*. Un pixel tiene una correspondencia biunívoca con una pequeña porción cuadrada de la imagen, es fácilmente visible si la imagen es ampliada en la pantalla de una computadora.



http://arrayit.com/Products/Microarray_Scanners/Microarray_Scanner_Colorimetri/microarray_scanner_colorimetri.html

Figura 16. Imágenes digitales escaneadas de **un spot** de 125 μm a distintas resoluciones

La cantidad de pixeles (o equivalentemente su tamaño) que representan al microarreglo depende de la *resolución* con la cual se ha obtenido esa imagen. A mayor cantidad de pixeles, menor su tamaño y mayor la resolución de la imagen. El tamaño de los pixeles se mide en micrones y corresponde a la resolución.

En un microarreglo de 2 canales se obtiene una imagen para cada tinte (Cy3, Cy5).

Si se utiliza una resolución de 10 μm (micrón), habrá 10 000 (100 x 100) pixeles para cada mm^2.

Se sugiere que un spot de un microarreglo no debería estar representado por menos de 8 pixeles por dimensión, o sea, como mínimo 64 pixeles por spot.

El archivo digital de un microarreglo se denomina *bit map*, provee una representación numérica 1:1 de los valores de fluorescencia de cada pixel. El formato estándar de estos archivos es TIFF (tagged image file format). En un archivo TIFF estándar de 16 bits ($2^{16} = 65536$) las intensidades están representadas por un número entre 1 y 65536.

3.2 Procedimientos de detección

Los instrumentos de detección (scanners o imagers) examinan los microarreglos excitando cada tinte fluorescente de su superficie mediante una luz monocromática y colectando la luz de emisión (fluorescencia) convirtiendo la corriente de fotones en valores digitales que pueden ser almacenados en una computadora. A cada tinte le corresponde una **longitud de onda de excitación** y una longitud de onda **de emisión** diferente.

Los **escaners** utilizan luz **láser**. Los **imagers** utilizan **luz blanca** y obtienen luz monocromática mediante un filtro.

El ***rango dinámico*** de un sistema de detección es la gama de intensidades que puede distinguir. Generalmente se representa como el cociente entre la intensidad de la señal más brillante y la más tenue que puede detectar por encima del fondo. Si los valores absolutos de ese rango se encuentran entre 66 y 65536 decimos que es de 1000 veces (1000 fold dynamic range). La imagen de un microarreglo con 1000-fold change tendría valores de intensidades distribuidos en una amplitud (span) de 3 órdenes de magnitud.

Reduciendo la intensidad del fondo por ejemplo a la mitad (de 66 a 33) el rango dinámico se duplicaría dando un rango dinámico de 2000 veces.

Tomando varias imágenes del arreglo con diferentes seteados del instrumento es posible aumentar el rango dinámico, entre todas ellas, a 100 000 veces es decir 5 órdenes de magnitud.

3.3 Obtención de la imagen en microarreglos de cDNA de dos canales.

Cada microarray es escaneado con un láser que produce para cada tinte un mapa digital o imagen con intensidades de fluorescencia para cada pixel. Un láser típico opera con las siguientes funciones: excitación, recolección de la luz emitida, evaluaciones espaciales, discriminación entre excitación/emisión y detección.

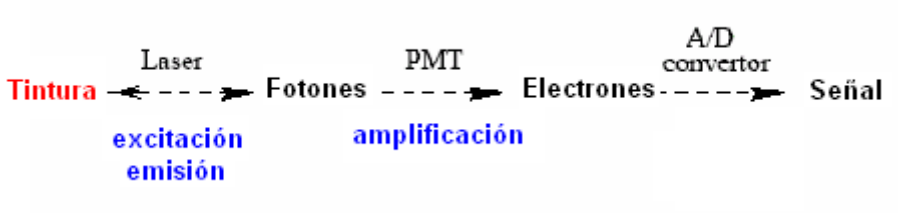


Figura 17. Diferentes procesos en el escaneado de un microarreglo (Yang et al 2000)

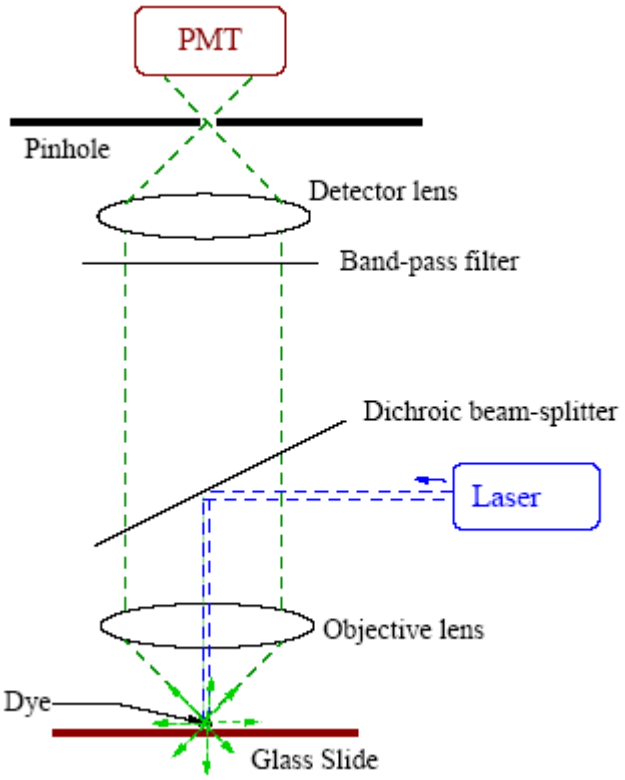


Figura 18. Diagrama mostrando los diferentes procesos de un escaneado de un migroarreglo (Yang et al 2000)

La **región** escaneada es **dividida en pixeles** de igual tamaño y el láser genera una **luz de excitación** que está enfocada sobre una porción pequeña del portaobjetos de vidrio. Las **moléculas fluorescentes** en ésta región **absorben** los fotones de excitación generados por el láser y **emiten** fotones de fluorescencia. Estos fotones pueden ir en cualquier dirección y una fracción de ellos es colectada por una lente. Estamos interesados en la cantidad de **fotones emitidos**, éstos son en general varios órdenes de magnitud menores que los fotones excitados.

Un **separador** dicroico y un filtro se colocan frente al detector para discriminar entre los **fotones de emisión y los fotones de excitación**. Esta discriminación es posible debido a que la luz de excitación o absorción tiene en general una **longitud de onda** levemente **menor** que la luz de emisión:

	Longitud de onda	
Tinte	**Exitación o absorción**	**Emisión**
Cy3	550 nm	570 nm
Cy5	649 nm	670 nm

nm $=10^{-9}$m

El detector en un scanner convierte los fotones de emisión en corriente eléctrica, comúnmente es un **tubo fotomultiplicador** (PMT, photomultiplier tube). Este convierte cada fotón en una cierta cantidad de electrones (alrededor de un millón).

Finalmente un **convertor analógico/digital** (A/D) se utiliza para convertir los electrones en una secuencia de señales digitales.

El proceso de digitalización promedia espacial y temporalmente y produce para cada píxel una señal que representa la fluorescencia en la región correspondiente a ese píxel. Para más detalles sobre la tecnología del scanner pueden consultarse los libros de Schena (1999, 2000, 2003). En un experimento de microarreglos de dos canales el scanner produce dos imágenes en blanco y negro, de 16-bits de formato TIFF, una para cada tinte fluorescente.

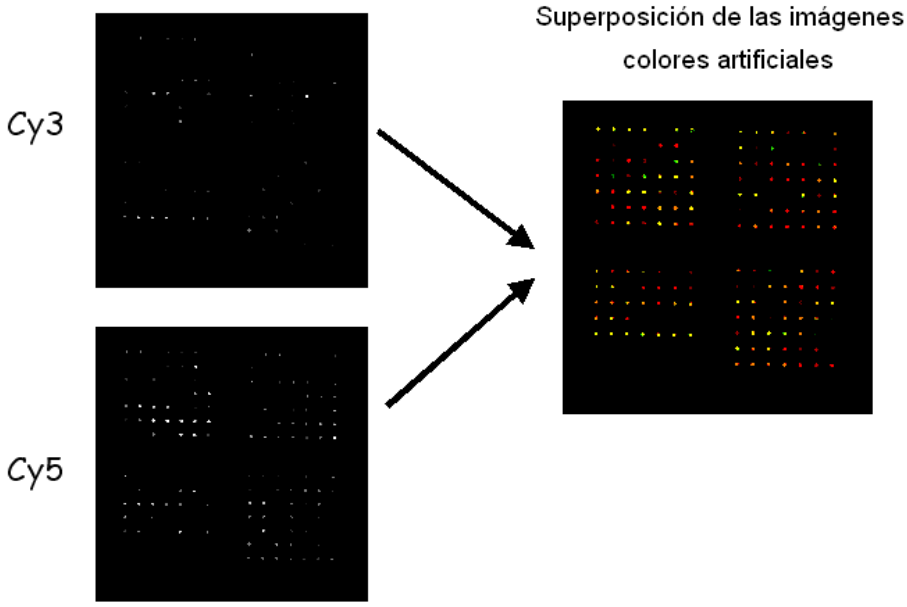


Figura 19a: Imágenes para tintes diferentes y su superposición

Un scanner secuencial escaneará el portaobjetos de vidrio con una longitud de onda primero y luego con la otra. Alternativamente un escaner dual, con dos lásers y dos detectores escaneará el vidrio con las dos longitudes de onda simultáneamente.

Veamos una superposición de imágenes con más detalle:

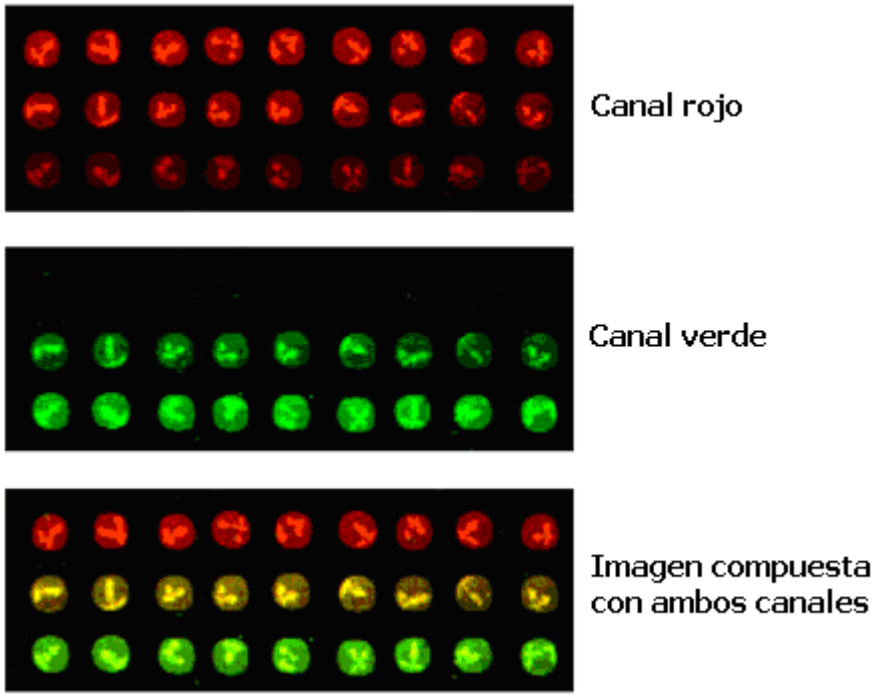


Figura 19 b.
http://arrayit.com/Microarray_Diagnostics/Diagnostic_Technology/arrayit_diagnostics_2color_600.jpg

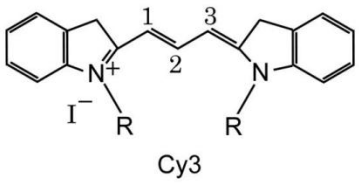
La figura 19 b muestra 3 **imágenes digitales** de un sector de un microarreglo. En la primera el escaneado es para el tinte rojo; se representan las intensidades mediante una

escala de rojos, se indica como "**canal rojo**", la segunda corresponde al tinte verde representando en verde las intensidades ("**canal verde**") y la tercera es una **imagen compuesta** de las dos anteriores. La primera fila correspondiente a la imagen del canal rojo es la más intensa y la tercera la más débil. Esto ocurre en forma inversa para el canal verde, siendo la tercera fila la de mayor intensidad. En la primera fila de **la imagen digital compuesta** las manchas (spots) son rojas, indicando que el en el canal rojo la intensidad de cada spot es mayor que en el verde, en la segunda fila el amarillo indica paridad de intensidades y la tercera que la intensidad mayor se encuentra en el canal verde.

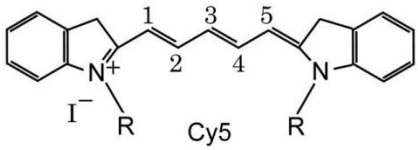
Hay muchos tipos de ruido que pueden afectar la señal final producida por el escaner que pueden ser clasificados en dos categorías: ruido de fuente y ruido de detección.

- Ruidos de la fuente: ruido de fotones, polvo en los vidrios, tratamiento de los vidrios.
- Ruidos de detección: están vinculados al proceso de amplificación y digitalización. Una imagen perfecta debería reflejar únicamente medidas de la intensidad de fluorescencia de los tintes de interés. Sin embargo en la práctica el sistema es imperfecto y las imágenes son combinaciones de señales no deseadas tales como ruido fotónico, ruido electrónico, luz láser reflejada y fluorescencia de fondo tanto como de las señales de fluorescencia deseadas.

Además del ruido, que es la componente aleatoria de la variabilidad, la señal está afectada por errores sistemáticos provenientes de las diferencias en ciertas propiedades de los tintes utilizados:



Cy3



Cy5

- El **tamaño diferente** de las moléculas de Cy3 y Cy5
- La eficiencia emisión de fotones (quantum yield) en el proceso de fluorescencia.
- El blanqueado por la luz (photo bleaching)…

Dependiendo del escaner es necesario que el usuario ajuste una cantidad de parámetros (tasa del escaneado (rate), potencia del láser, voltaje del PMT. A una mayor potencia se excitan más fotones y se genera mayor señal y mayor ruido de fuente. Un voltaje mayor del PMT amplifica más electrones por fotón y genera más señal y más ruido de detección. Podría ser preferible utilizar una potencia alta para el láser en vez de un alto voltaje para el PMT ya que esto excitaría mayor número de fotones para la señal en vez de producir más señal por fotón.

Sin embargo un **láser** más intenso puede **dañar las muestras** hibridizadas (**photobleaching**) y el escaner deberá ser calibrado dependiendo de la cantidad de

escaneos que deban ser realizados en cada muestra. Se recomienda utilizar la menor cantidad de iluminación posible para un experimento y minimizar la cantidad de lecturas. En un mismo microarray evitar la lectura de una porción de un chip más veces que otra porción. La mayoría de los escaners pueden leer un sustrato de 25 X 76 mm en un único paso.

En algunos escaners solo es posible ajustar el voltaje del PMT y no la potencia del láser. Fijar un nivel extremadamente alto del PMT puede saturar algunos pixeles, esto es que sobre un cierto nivel de electrones, el conversor análogo/digital A/D registrará la señal como 2^{16}-1 = 65535. En la práctica los usuarios ajustan el nivel del PMT de manera que los pixeles más brillantes estén justo debajo del nivel de saturación. Esto trae la pregunta de cuanto puede afectar los resultados finales utilizar niveles diferentes del PMT en los dos canales (Yang et al 2000, Bengstsson et al. 2004).

La figura 19 c) muestra en forma de esquema la construcción de un microarreglo, el experimento y la obtención de las imágenes digitales con colores artificiales.

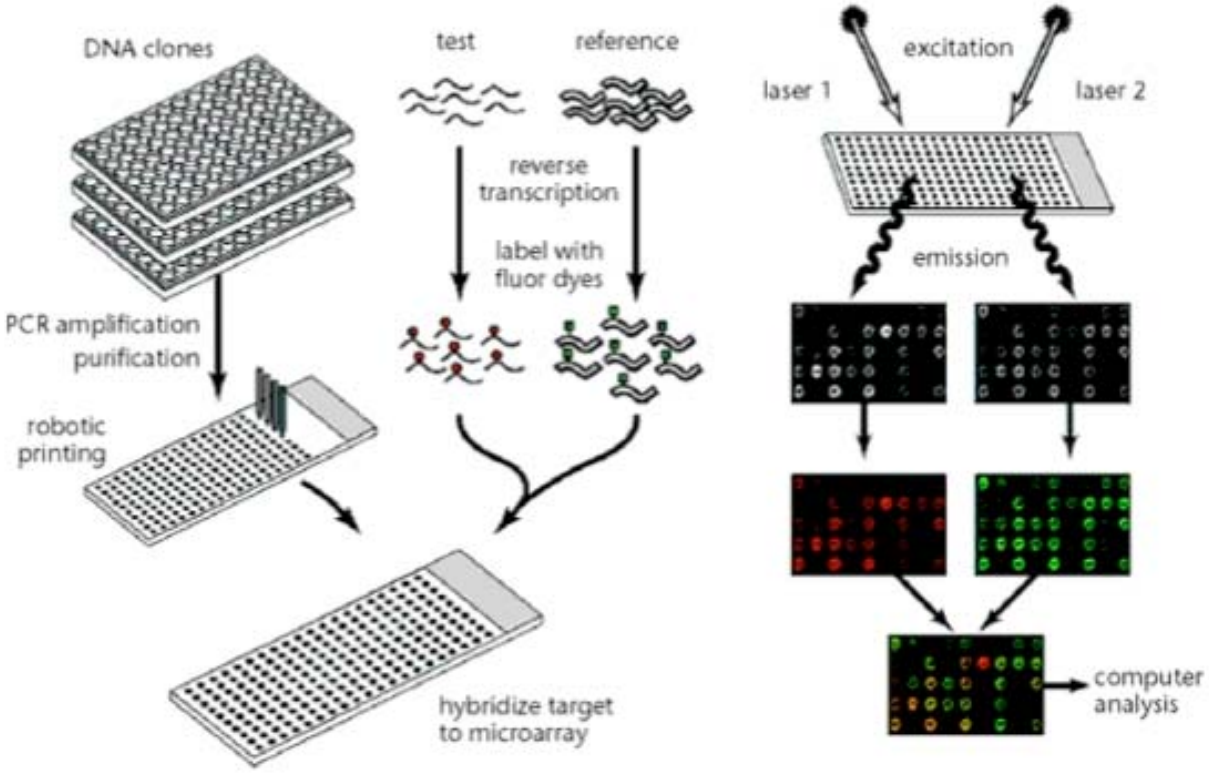


Figura 19 c.
http://www.medtrad.org/glosarios/bio_molecular/Glosario/DNArray.jpg

3.4 Métodos de procesamiento

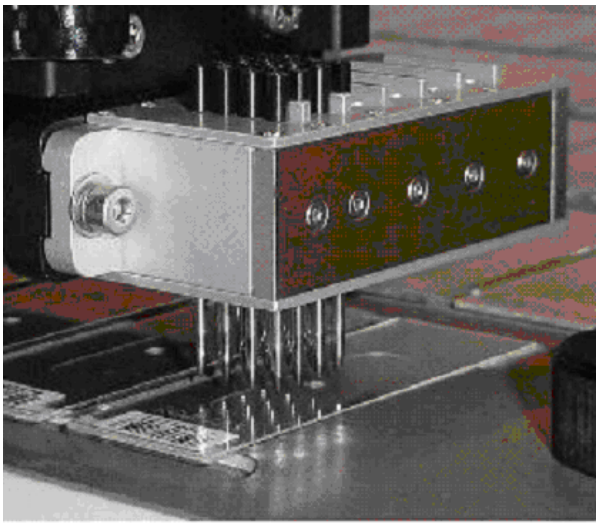
Como ya hemos visto el proceso de escaneado en un microarreglo de dos canales produce dos imágenes de 16-bits de formato TIFF (tagged image file format), una para cada tinte fluorescente. Estas imágenes son los *datos crudos* del experimento.

El objetivo es extraer, para cada secuencia de ADN fijada en el microarreglo, una medida de su abundancia en las muestras de mARN incógnitas.

El procesamiento de las imágenes escaneadas de microarreglos pueden ser divididos en tres tareas

- **Grillado**: es el proceso de asignación de coordenadas a cada uno de los spots
- **Segmentación**: clasificación de los pixeles en foreground (señal) y background (fondo)
- **Extracción de la intensidad**: calcular para cada spot del arreglo los pares de intensidades de fluorescencia (R, G) y medidas de calidad del spot.

3.4.1. Grillado



http://www.surrey.ac.uk/SBMS/Fgenomics/Microarrays/images/QArray2.jpg

Figura 20 : Cabezal y agujas imprimiendo un microarreglo

Los spots en un microarreglo están agrupados en una cuadrícula formada por subcuadrículas (grids, subgrids), con un espaciado mayor entre subcuadrículas que dentro de ellas. Las subcuadrículas surgen por que hay varias agujas que (pins) en el cassette del robot que realiza el spotting, **todos los spots** de una subcuadrícula han sido spoteados por **una misma aguja** (figura 20).

La estructura básica de la imagen está determinada por la construcción del microarreglo. Esto es cuántas filas y cuántas columnas tiene la grilla y dentro de cada subgrilla o sector cuántas filas y cuántas columnas hay.

La figura 21 muestra la superposición de las dos imágenes escaneadas de un microarray formado por una cuadrícula de 3 filas x 3 columnas = 9 sectores correspondientes a cada aguja (pin groups) con 18 filas x18 columnas = 324 spots cada subgrupo, dando un total de 2916 spots. Los colores son representaciones artificiales de las intensidades.

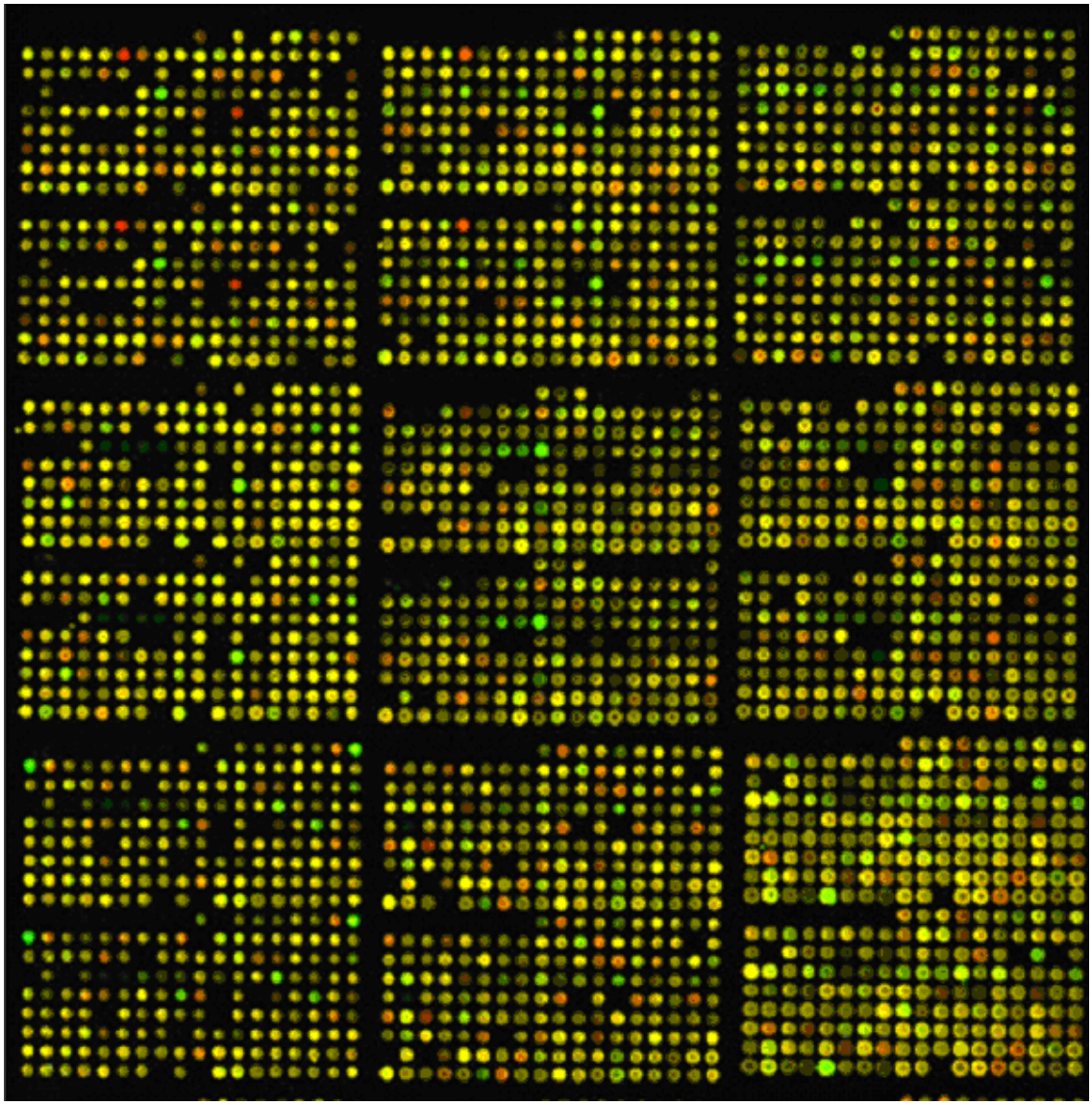


Figura 21. Imagen superpuesta de los dos canales de un chip de 3 x 3 grupos de agujas (pin groups)

- Rojo indica que el gen está **más expresado** en el **tejido patógeno** que en el sano
- Amarillo indica que el gen está expresado con **igual intensidad** en ambos tejidos
- Verde indica que el gen está **más expresado** en el **tejido sano.**

Para extraer las características de la grilla, el software requiere la información de:

- Cuántas subcuadrículas tiene el arreglo en cada dirección (x e y).
- Cuantos spots hay por subgrilla en cada dirección (x e y).
- Espaciado entre las filas y las columnas de la grilla.

Un programa de computadora debería realizar el procesamiento de la imagen superponiendo un arreglo de círculos de tamaño y espaciados predeterminados. Los pixeles dentro de los círculos podrían ser considerados señal y el resto fondo.

Desafortunadamente esto no es posible. Las posiciones exactas varían, en general el patrón no es completamente regular.

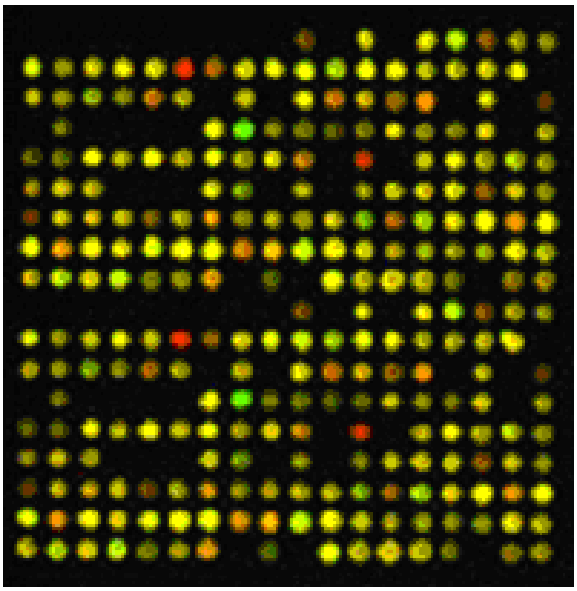


Figura 22. Uno de los 24 bloques del chip

El problema de identificar las posiciones de las características (features) de la imagen es que la posición y los tamaños de dichas características varían dentro de cada subgrilla y entre subgrillas. A continuación describimos los aspectos a tener en cuenta para la identificación de las posiciones de las características (features) de la imagen:

- **Posición general del arreglo en la imagen**: esto es lo más variable entre imágenes.
- **Posiciones desparejas de las subcuadrículas**: Las subcuadrículas no están alineadas. Esto ocurre cuando las agujas no están perfectamente alineadas en el cassette.
- **Curva dentro de la subgrilla**: El vidrio no es perfectamente plano o la aguja se ha corrido levemente en el cassette, de manera que los puntos son impresos en un patrón curvo sobre la superficie del vidrio.
- **Espaciado no parejo**: Las agujas se han movido levemente en el cassette o el vidrio no es completamente chato.
- **Tamaño desparejo de los spots**: Más ó menos fluido ha sido depositado sobre el vidrio durante la fabricación del arreglo.

La mayoría de los software proveen procedimientos automáticos y manuales.

La figura 22 muestra la imagen correspondiente a una de las subcuadrículas. Las imágenes resultantes del escaneado pueden no ser tan claras como las de las figuras 21 y 21. La figura 23 muestra las dos imágenes escaneadas superpuestas y coloreadas de un sector de un microarreglo en el que parecen **muchos artefactos**.

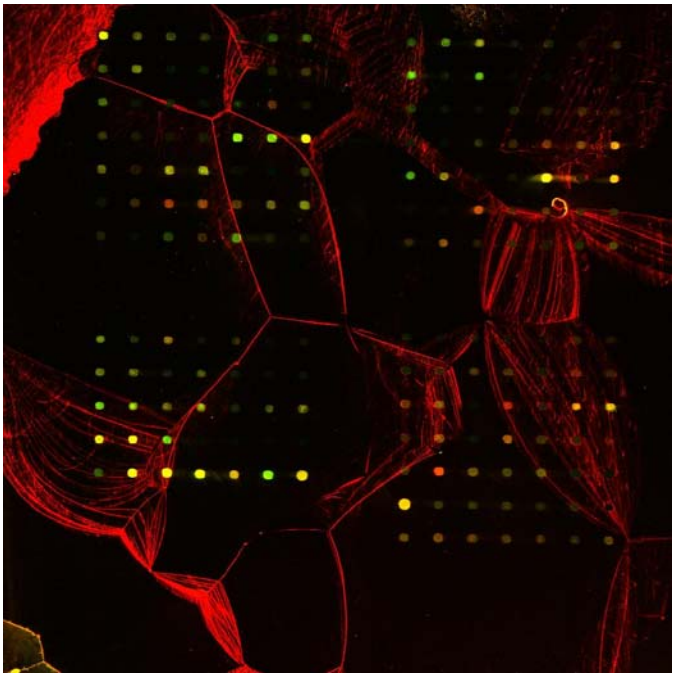


Figura 23a. Una imagen de un sector de un microarray que presenta artefactos
http://ludwig-sun2.unil.ch/~plangend/module8/microarray/image_analysis/

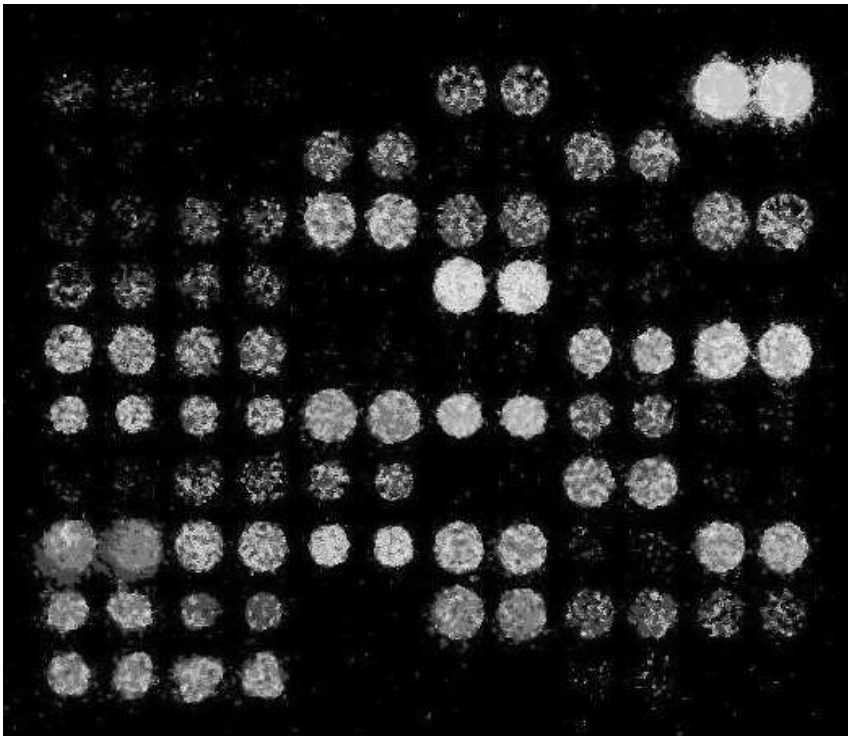


Figura 23b. Se observan los distintos tamaños de los spots y algunas chorreadas (**Qin** 2010)

3.4.2 Segmentación

Segmentación de una imagen puede definirse como el proceso de particionar la imagen en diferentes regiones con diferentes propiedades. Todos los procedimientos de segmentación de las imágenes digitales de los microarrays requieren seleccionar inicialmente el conjunto de los pixeles donde se supone se encuentra un spot. llamado *máscara del spot* (*spot mask*). Luego los pixeles de esa región son clasificados en *foreground* (primer plano, frente, es decir los que corresponden específicamente al spot de interés) o *backgound* (fondo), de manera que se puedan calcular intensidades como medidas de la abundancia de la secuencia transcripta para cada secuencia de DNA spoteada.

Los métodos de segmentación pueden clasificarse en

- Círculo fijo
- Círculo variable
- Histograma
- Forma variable

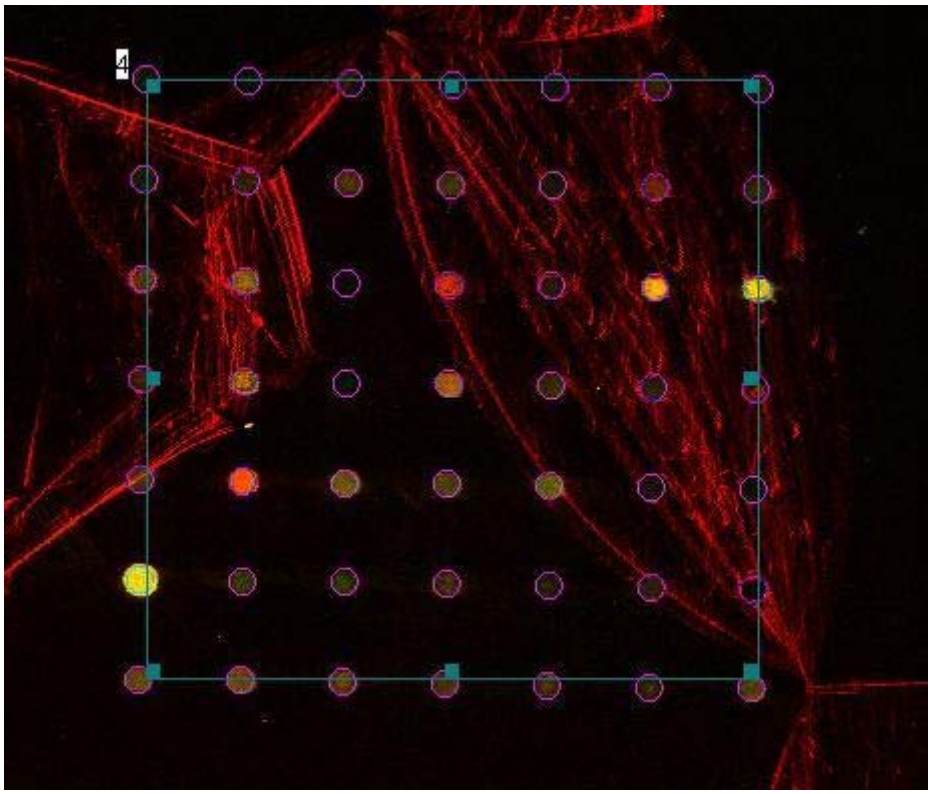


Figura 24. Grillado
http://ludwig-sun2.unil.ch/~plangend/module8/microarray/image_analysis/

Método	**Software que lo implementa**
Círculo fijo	ScanAnalyze, GenePix, ScanArray Express, QuantArray
Círculo variable	GenePix, QuantArray, Dapple, Agilent Feature Extraction

Histograma Forma adaptable	ImaGene, QuantArray Spot

Segmentación con círculos fijos

Superpone un círculo del mismo tamaño a todos los spots de la imagen. Este método es fácil de implementar y funciona bien cuando todos los spots son circulares y del mismo tamaño. Pero cuando los spots tienen forma variable, como ocurre en la mayoría de los microarreglos, tiende a dar resultados imprecisos.

Segmentación con círculos de tamaño variable

En este tipo de segmentación, el diámetro del círculo es estimado en forma separada para cada spot. Algunos software proveen la opción de ajustar a mano el diámetro de los círculos, spot por spot pero esto lleva mucho tiempo. Además en la práctica los spots rara vez son circulares y una máscara circular tendrá un mal ajuste. Las fuentes de no circularidad incluyen los procesos de la impresión (características de las agujas) o el post procesamiento de los portaobjetos luego de la impresión (tiempo insuficiente en la deshidratación)

Segmentación por histograma

Este tipo de método utiliza una *máscara objetivo* (target mask) que debe ser elegida más grande que cualquiera de los spots. Se coloca una máscara encima de cada spot. Para cada spot se seleccionan pixeles con señal (foreground) y de fondo (background) a partir del histograma de las intensidades de cada pixel. Idealmente esto producirá un histograma bimodal con los valores de las intensidades, con los más altos correpondientes a la señal y los más bajos al background.

QuantArray implementa un método utilizando una máscara rectangular y define dos intensidades una de foreground y otra de background como las intensidades medias entre valores predeterminados por percentiles. Estos son los percentiles del 5 y 20 % para el background y los del 80 y 90% para el foreground. Calcular las intensidades del foreground correspondiente a percentiles más altos este método lleva generalmente a intensidades mayores. La mayor ventaja de este método es su simplicidad. Sin embargo la mayor desventaja es que la cuantificación es inestable cuando se toma una máscara target grande para compensar la variación en los tamaños de los spots. Puede ocurrir, incluso, que la zona que se asigne al spot con el criterio anterior, resulte no conexa y reflejando intensidades de spots brillantes vecinos.

Estos métodos no utilizan ningún tipo de información local espacial.

Segmentación de Mann-Whitney

El método de Chen et al. (1997) implementado por QuantArray utiliza una máscara target circular interna (para la región del target) y una externa cuadrada (target patch)

como muestra la figura 25. Utiliza el test de Mann-Whitney para dos muestras independientes en forma iterativa.

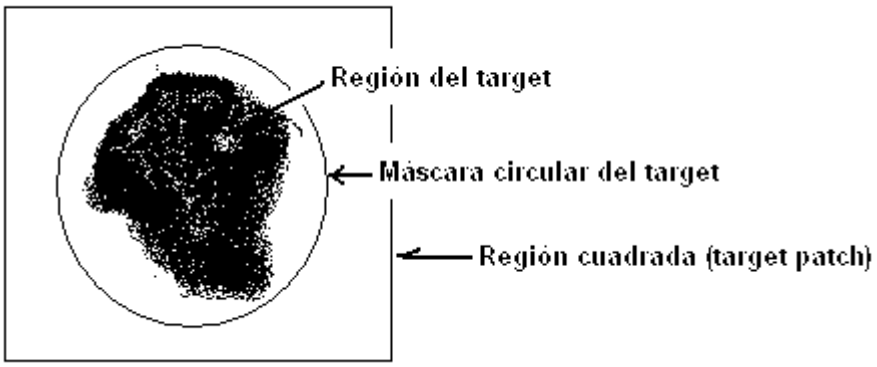


Figura 25. Regiones del método de segmentación por Mann-Whitney

Los valores iniciales para las dos muestras se eligen de la siguiente manera: 8 puntos de la región fuera de la mascara circular, donde se supone tiene sólo background y los 8 valores menores de la máscara circular. Para un nivel prefijado si el test no resulta en rechazo, se elimina al azar una cierta cantidad de puntos de la segunda muestra y se eligen los siguientes puntos con menor intensidad dentro de la máscara circular. Sigue hasta que el test resulta en rechazo. Se define como **spot** todos los **pixeles** que tienen una **intensidad mayor o igual** a la intensidad mínima de los 8 valores de la máscara circular para los cuales el test resultó en rechazo. Este método también puede dar como resultado una región no conexa.

Test de Mann-Whitney: para dos muestras independientes, es también conocido como Test de Wilcoxon (en `R wilcoxon.test`). Los primeros autores generalizaron el procedimiento que el segundo propuso para el problema de muestras independientes de igual tamaño.

Idea intuitiva: A cada dato se le asigna su rango en una muestra total ordenada de las dos muestras. Si los datos de las dos muestras provinieran de poblaciones con la misma distribución y ambas tuvieran **la misma cantidad de observaciones**, esperaríamos que la suma de rangos de la Muestra 1 fuera "similar" a la suma de rangos de la Muestra 2. Eso resultaría de datos de las dos muestras alternados en la muestra total ordenada. Una **suma de rangos** demasiado grande o demasiado pequeña sería indicativa de diferencias entre las dos poblaciones de las cuales fueron obtenidas las muestras. Por lo tanto, la hipótesis nula de que las dos poblaciones no difieren debería ser rechazada cuando la suma de rangos de una muestra tiende a ser notablemente mayor (o menor) que los de la otra muestra.

Consideramos a continuación dos modelos posibles para el problema. Cada modelo permite testear diferentes hipótesis respecto de las poblaciones de las cuales provienen los datos, pero el test es el mismo.
Modelo (1). Ambas muestras proviene de poblaciones cuyas distribuciones solamente difieren en la posición esto es:

$X_1, X_2, ..., X_n$ i.i.d ; distribución F(x)

$Y_1, Y_2, ..., Y_m$ i.i.d ; distribución G(x) = F(x+c)

Si hay una diferencia entre ellas se debe SÓLO a la posición de la distribución.

Llamamos θ_X a la mediana de las observaciones con distribución F y θ_Y a la mediana de las observaciones con distribución G, entonces el Test de Mann-Whitney permite decidir entre las siguientes hipótesis nula y alternativas:

Hipótesis nula (1): H_0: $\theta_X = \theta_Y$

Hipótesis alternativas posibles H_a:

a) $\theta_X \neq \theta_Y$	b) $\theta_X < \theta_Y$	c) $\theta_X > \theta_Y$

Modelo (2) Los datos de cada muestra provienen de diferentes distribuciones.

$X_1, X_2, ..., X_n$ i.i.d ; distribución F

$Y_1, Y_2, ..., Y_m$ i.i.d ; distribución G

Hipótesis (2): H_0: $F(x) = G(x)$ para todo x versus H_a: $F(x) \neq G(x)$ para algún x

La hipótesis nula afirma que las dos distribuciones poblacionales son iguales (es equivalente a la H_0 del Modelo (1)), pero la **alternativa** dice que **las dos distribuciones difieren** de algún modo, pero no dice de qué modo.

Estadístico del test es el mismo para ambos modelos

T = Suma de rangos de la muestra con menor cantidad de observaciones

La distribución de este estadístico está tabulada para tamaños de muestra pequeños. Para tamaños de muestras grandes se usa una aproximación normal a la distribución del estadístico. Cuando hay empates entre las observaciones se reemplaza el rango de cada empate por el promedio de los rangos de los empates, se modifica el estadístico anterior por una versión estandarizada y se utilizan valores críticos de la tabla de N(0,1) (Ver libro de Conover: "Practical Nonparametric Statistics", para más detalles)

Región sembrada creciente.

El método de región sembrada creciente *seeded region growing* (SRG) (Adams and Bischof [2]) está implementado en el software *Spot. Spot* es una versión especializada de otro paquete de *R* llamado *VOIR*, que está desarrollado por el "CSIRO Image Analysis Group", que provee un entorno más general para el análisis de imágenes. El SRG requiere la especificación de puntos iniciales o *semillas*, el punto débil de este tipo de métodos es la selección de la cantidad y posición de las semillas. El análisis de imágenes de microarreglos, sin embargo, tiene una situación inusual respecto de la cantidad de features (spots) es conocida exactamente *a priori* y las posiciones aproximadas de los centros de los spots están determinadas desde el inicio.
Veremos este procedimiento con detalle en la sección 3.5.

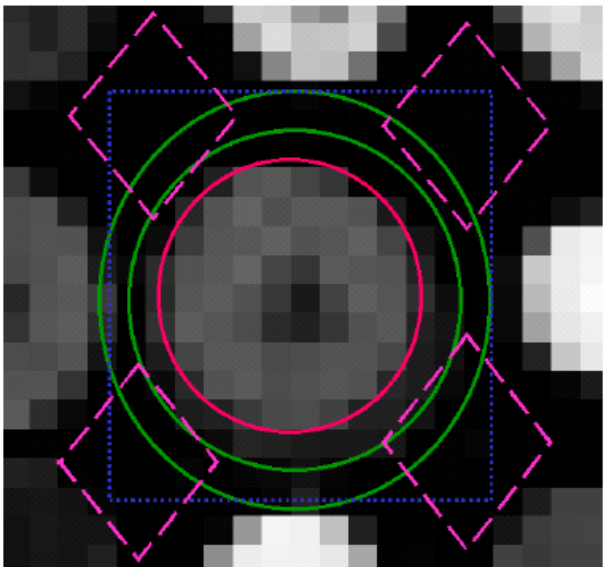


Figura 26: muestra las distintas zonas utilizadas por los distintos métodos de segmentación. Las regiones rojas y azul son dos máscaras del spot. La verde es una región de background. La rosa es la utilizada en *Spot.*

3.4.3 Extracción de la intensidad - Información numérica.

A partir de los pixeles que representan los spots (señal) y su fondo, los software de procesamiento de imágenes deben calcular medidas numéricas para cada parte. A continuación describimos las medidas incluidas en la mayoría de ellos.

- Media de la señal: media de las intensidades los pixeles que constituyen el spot
- Media del fondo: media de las intensidades de los pixeles determinados como background.
- Mediana de la señal: mediana de las intensidades de los pixeles que constituyen el spot
- Mediana del fondo: mediana de las intensidades de los pixeles determinados como background.
- Desvío estándar de la señal: de las intensidades de los pixeles que constituyen el spot
- Desvío estándar del background: de las intensidades de los pixeles que constituyen el background
- Medidas de la forma del spot: cantidad de pixeles que lo conforman. Medida de circularidad= $4\ \pi$ área / (perímetro ^{2}).
- Marca (flag), indicando la calidad del spot.

La medida más importante es la intensidad del spot que puede elegirse entre la media y **la mediana**, pero esta última es la más recomendable.

3.5 Una metodología con más detalle

Veremos con un poco más de detalle la metodología propuesta por: Yang, Buckley, Dudoit and Speed, "Comparison of methods for image analysis on cDNA microarray data", Journal of Computational and Graphical Statistics, January 2002, e implementada en *Spot*: http://experimental.act.cmis.csiro.au/Spot/index.php

3.5.1 Formación de una imagen combinada

La entrada al procedimiento de análisis de imagen consiste de un par de imágenes de 16-bits no signados, que son guardadas en archivos con formato TIFF. Nombraremos estas imágenes "$\mathcal{R}$" y "$\mathcal{G}$" para el "rojo" y el "verde", donde $\mathcal{R}$ corresponde al tinte Cy5 y $\mathcal{G}$ corresponde al Cy3. Tanto los pasos de direccionamiento (grillado) como la segmentación requieren de una única imagen. Otro requisito crucial es que valores muy altos de la imagen deben ser *moderados* en la imagen combinada. Esto es necesario para evitar que pixeles muy brillantes dominen las etapas de direccionamiento y segmentación. Computacionalmente, es conveniente que la imagen combinada sea de 8-bits. Los procedimientos de direccionamiento y segmentación automáticos se realizan sobre la imagen combinada de 8-bits. El método de segmentación producirá una *máscara del spot* que se utiliza junto con las imágenes originales de 16-bits para la extracción de las intensidades del spot (foreground) y del fondo (background).

El software *spot* utiliza los siguientes pasos para producir la imagen combinada $\mathcal{RG}$ de 8-bits:

1. Primero, se aplica una transformación raíz-cuadrada a las dos entradas $\mathcal{R}$ y $\mathcal{G}$ dando $\mathcal{R}'$ y $\mathcal{G}'$. Esto reduce la influencia de pixeles muy brillantes en las etapas de direccionamiento y segmentación.
2. Luego se calcula la mediana de cada una de esas imágenes, $m_{\mathcal{R}}$ y $m_{\mathcal{G}}$.
3. Se calcula una combinación inicial calculada como

$$\mathcal{G}' + (m_{\mathcal{G}} / m_{\mathcal{R}}) \mathcal{R}' .$$

4. Los valores mayores que 255 se fijan en 255.

El primero y el último de estos pasos producen moderación, mientras que el segundo y el tercero aseguran el balance entre las dos imágenes. El último paso asegura que el resultado $\mathcal{RG}$ pueda ser guardado en una imagen de 8-bits.

Hemos descripto un método para combinar los dos canales $\mathcal{R}$ y $\mathcal{G}$ en una única imagen con el propósito de realizar análisis automáticos. Este procedimiento es diferente al método estándar de *superposición* con el propósito de *visualización*.

3.5.2 Direccionamiento automático

El procedimiento de direccionamiento se basa en el concepto de *lote***.** Para el propósito del análisis de imagen, un lote es una **colección de imágenes** de microarreglos con la **misma estructura** geométrica global. Esto corresponde en general a portaobjetos impresos por el mismo robot y agujas del mismo cabezal, en tiempos cercanos y escaneados en forma similar.

La geometría de un microarreglo puede variar de diferentes maneras:

- Estructura básica: La estructura geométrica fundamental es la distribución de las grillas (por ejemplo 3 por 3) y el arreglo de los spots dentro de las grillas (por ejemplo 18 por 18).
  Las imágenes dentro de un lote deben ser idénticas en términos de la estructura básica.

- Configuración de las agujas: El segundo aspecto fundamental de la estructura geométrica es la configuración de las agujas. Las agujas en el cabezal del arrayer, no tienen una distribución perfectamente regular. Esto es que mientras las agujas están distribuidas nominalmente en forma de un arreglo rectangular (por ej. 4 por 4) pequeñas inclinaciones u otros efectos hacen que en la práctica se presenten pequeñas irregularidades en su distribución. Aún cuando las irregularidades de la configuración de las agujas sea muy pequeña pueden resultar en irregularidades importantes en las grillas del portaobjetos del microarreglo y por lo tanto en la imagen.

  Suponemos que los portaobjetos de un mismo lote contienen configuraciones de las agujas casi idénticas.

- Traslación global: varios factores, en particular el recorte de la imagen (image cropping), pueden llevar a un corrimiento general en las posiciones del spot entre una imagen y otra. *Sí* esperamos que ocurran este tipo de variaciones dentro de un mismo lote; de hecho un componente clave en el proceso de posicionamiento o grillado es la estimación de el corrimiento global entre la imagen que se está procesando y el molde (o template).
- No se esperan rotaciones y deformaciones: si se tienen cantidades importantes de ese tipo de distorsiones los resultados serán incorrectos.

Para comenzar el análisis de un lote de imágenes, el usuario elige una de las imágenes como molde para el lote completo (*template*). Luego especifica algunas características de esta imagen molde, tales como el extremo superior izquierdo y el inferior derecho de cada grilla. Esto capta la información respecto a la configuración de las agujas y la separación promedio entre filas y columnas de spots dentro de una grilla. El resto del grillado es automático. El software estima primero el corrimiento global de la grilla en procesamiento respecto de la imagen template. Esto es que las posiciones de la grilla en la nueva imagen se estiman como una traslación de la grilla del template. Luego se realizan pequeños ajustes para cada fila y columna de spots dentro de cada grilla. Esto permite pequeñas variaciones en la estructura entre grillas como también imprecisiones en la especificación del template. Hay dos representaciones de estas grillas estimadas que son esencialmente equivalentes. La primera que consiste en líneas verticales y horizontales que pasan por los centros estimados de los spots constituye las grillas

ajustadas del frente (*fitted foreground grids*). La segunda está definida como las **grillas ajustadas del fondo** (*fitted background grids*) (figura 27), consiste en líneas verticales y horizontales que pasan entre los centros de los espacios entre los spots. Las semillas del foreground se eligen hallando el píxel con intensidad máxima sobre una pequeña región alrededor del punto central de la grilla del foreground y luego se establece la semilla como el conjunto de pixeles de *nxn* centrados en ese punto. Las semillas del background se construyen como *cruces* basadas en la **grilla ajustada al background.**

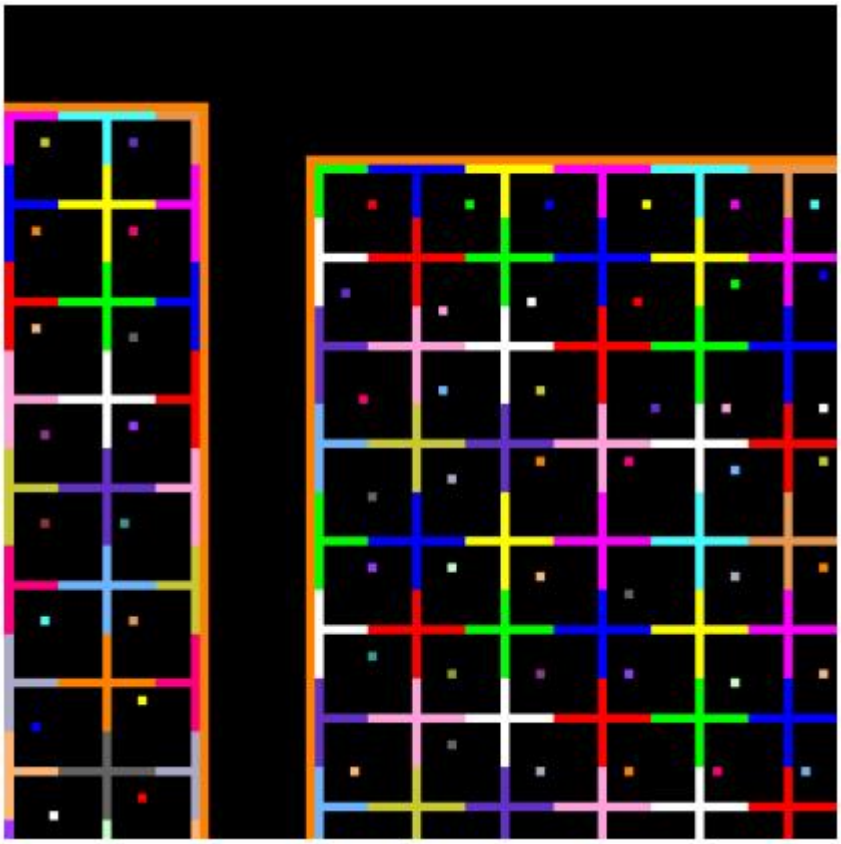


Figura 27: Imagen que ilustra la elección de las semillas del foreground y el background.

3.5.3 Segmentación método SRG

Al inicio del algoritmo se provee una cantidad de semillas que consisten en un píxel o un grupo de pixeles que sirven como puntos de partida del "region growing process". Luego el algoritmo hace crecer simultáneamente las regiones de fondo (background) y la imagen del spot propiamente dicha (foreground) hasta que todos los pixeles han sido asignados a una de las regiones. En cada paso, todos los pixeles que aún no han sido asignados pero que tienen un **vecino** que ha sido **asignado** a una región son considerados para su posicionamiento. El píxel no **asignado** se asigna a aquella **región** que tiene una **intensidad promedio** más cercana (en términos de diferencia absoluta de niveles de grises) a la del píxel de interés. El proceso se repite hasta que todos los pixeles han sido asignados.

Para la segmentación de microarreglos utilizando SRG, las semillas de foreground y background son elegidas utilizando las grillas calculadas en el proceso de direccionamiento (Sección 3.5.2) . Una forma obvia de elegir una **semilla** para cada **spot** es elegir un píxel en las intersecciones de las líneas de la grilla ajustada para el foreground. Podría ocurrir, en especial cuando el spot es pequeño que el píxel así seleccionado no pertenezca al interior del spot. Para evitar este problema se elige que tenga la **máxima intensidad** dentro de una **región pequeña centrada en el píxel de la intersección**. La **semilla del foreground** es fijada como un cuadrado de *n* por *n* pixeles centrado en este punto. El entero *n* es especificado por el usuario.

Para seleccionar las semillas del background un enfoque sencillo podría ser utilizar los puntos de intersección de las grillas del background, o utilizar todas las grillas juntas como una gran semilla de background que cubre la mayor parte de la imagen. Este procedimiento tiene la ventaja de separar las semillas del foreground entre si asegurando, por lo tanto, que los spots segmentados no se superpongan (merge or bleed into one another). Hay sin embargo, dos razones por las cuales utilizar semillas de background tan grandes es indeseable. La primera es que la intensidad del background varía localmente y el procedimiento de SRG tiene un funcionamiento pobre si las regiones no tienen una intensidad homogénea. Una segunda razón es que necesitamos una estimación local de la intensidad del background y esto se puede obtener teniendo regiones más pequeñas. Por estas razones se construyen semillas de background como cruces (figura 26). Con este diseño se logra separar los spots entre sí mientras que al mismo tiempo se producen regiones de background locales relativamente pequeñas. El SRG es aplicado utilizando estas semillas a la "imagen" combinada de la sección 3.5.1.

3.5.4 Extracción de la información

Intensidad del spot

La intensidad de cada píxel en una imagen escaneada representa el nivel de hibridación en una posición del vidrio. La cantidad total de hibridación para una secuencia spoteada particular de DNA es proporcional a la fluorescencia total del spot. Por lo tanto una medida natural de la intensidad del spot es la *media* o la *mediana* de las intensidades de los pixeles dentro del spot mask.

Otros estadísticos que pueden ser calculados dentro de cada spot mask y para cada uno de los dos canales son la *mediana* de los valores de los pixeles y algunas medidas de variabilidad.

Intensidad del background

El método de segmentación descripto en la sección anterior produce regiones de background locales y spots segmentados. Debido a la estructura de las semillas de foreground y background hay cuatro regiones de background alrededor de cada spot. Estas pueden ser utilizadas en diversas formas para calcular las estimaciones del background local. Uno de tales procedimientos, el método del *valle*, calcula la mediana de los valores en cada región del background y luego promedia dichos 4 valores para obtener el valor de background estimado para ese spot.

Morphological opening para estimar el background

Otro método para estimar la intensidad del fiondo sugerido por Yang, Buckley, Dudoit y Speed es el llamado *morphological opening* (Soille [19] para descripciones detalladas). Este método tiene dos etapas, primero reemplaza cada pixel por el valor mínimo de los pixeles en un cuadrado centrado alrededor de él. Una vez creada la imagen, se la vuelve a transformar tomando el máximo de los valores en la ventana cuadrada.

Se aplica a las imágenes R y G utilizando un elemento estructural cuadrado cuyo tamaño es al menos dos veces la distancia entre spots. Esta operación remueve los spots y genera una imagen para cada color que es un estimador del background para el vidrio completo. Para los spots individuales el background es estimado muestreando la imagen del background en el centro nominal del spot. Se muestrea esta imagen en vez de tomar promedio sobre una región porque se espera que ambos métodos produzcan resultados similares. Se utiliza una ventana muy grande para crear la imagen del "morphological background", y por lo tanto se espera que haya una variación espacial lenta.
Los estimadores de background resultantes del morphological opening son menores que los producidos por otros métodos más simples y menos variables ya que:

- se basan en valores de pixeles en una ventana local grande y
- no están afectados por pixeles más brillantes pertenecientes a los spots o a sus bordes.

3.6 Algunas características distintivas de los chips de alta densidad de oligos

La incorporación de los tintes en los microarreglos de dos colores se se realiza con anterioridad a la hibridación. En los chips de Affymetrix es posterior a la hibridación procedimiento que describimos a continuación.
Una vez que se han impreso los probes en el chip, se extrae el ARNm de una muestra de células y se producen copias complementarias (cARN) del ARN. Algunos de los nucleótidos que se utilizan para ensamblar esas copias han sido modificados con la incorporación de una pequeña molécula llamada biotina. Una vez que las moléculas de cARN etiquetadas con biotina han sido hibridadas al arreglo se tiñen con un compuesto streptavadina-phycoerythrina. La streptavadina tiene la afinidad no covalente más alta conocida con la biotina y la phycoerythrina es uno de los tintes fluorescentes disponibles más fuertes. El complejo final de probe impreso target biotinizado con el etiquetado indirecto streptavadina-phycoerythrina es escaneado para producir la imagen final.

Se obtiene una imagen para cada chip. Los spots son cuadrados ó rectangulares

Las imágenes resultantes del escaneado son en blanco y negro, nuevamente los colores de la figura anterior son artificiales. La figura 28 muestra la imagen resultante del escaneo de un chip

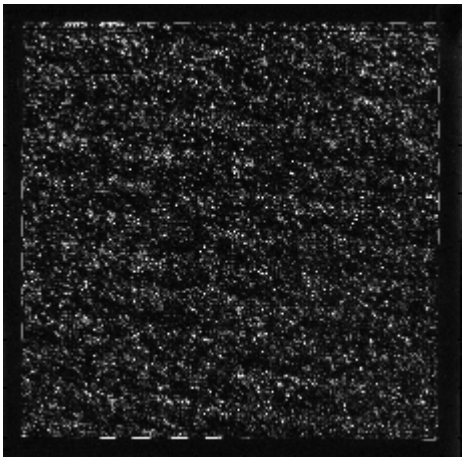


Figura 28: imagen resultante del escaneo de un chip.

Las imágenes son sometidas a un proceso de grillado y segmentación

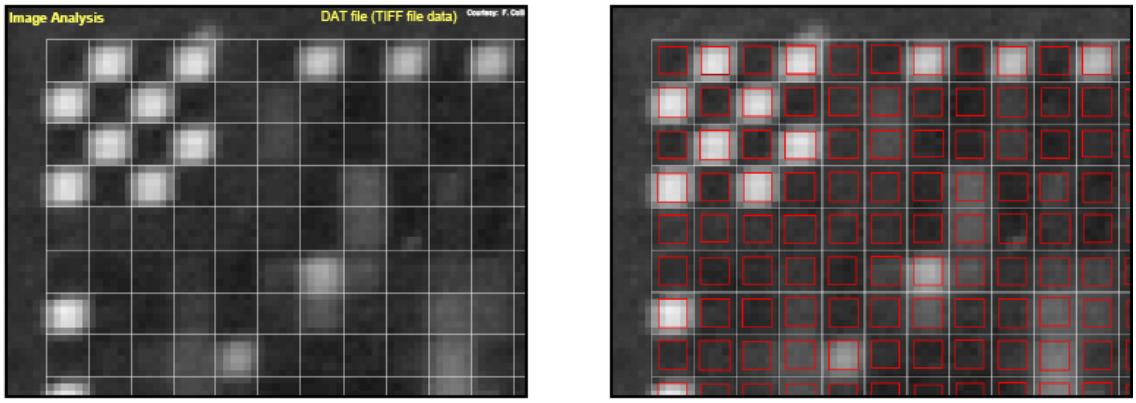


Figura 29: Grillado de la imagen digital de un chip de alta densidad

Intensidad de cada probe

Nos acercamos a uno de los probes, cada uno tiene aproximadamente 7 pixeles (20 micrones de lado),

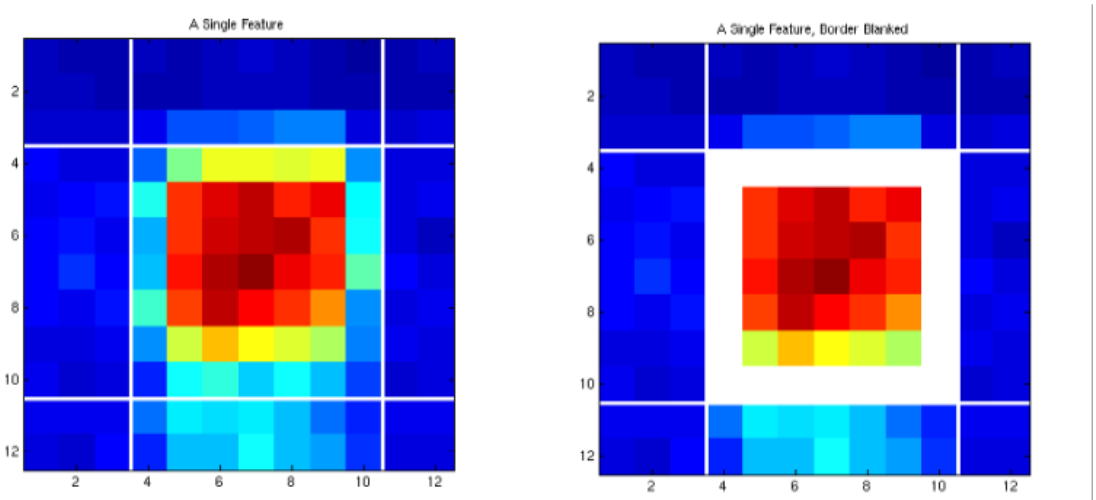


Figura 30: detalle del grillado y segmentación de un probe

blanqueando un pixel del borde, se toma la intensidad correspondiente al percentil del 75%.

Corrección por el fondo

MAS 5.0. Affymetrix. Estima un fondo para cada punto del chip de la siguiente manera.
i) se divide al chip en K regiones rectangulares (K=16, por defecto)
ii) se calcula el percentil del 2% de las intensidades en cada región
iii) para cada probe, con coordenadas (x, y) se calcula un fondo como el promedio pesado de los K valores del background. El peso depende de la distancia del punto al centroide del rectángulo $d_k^2(x, y)$ mediante la siguiente expresión

$$w_k(x, y) = \frac{1}{d_k^2(x, y) + S_0}, k = 1, \ldots, K$$

Procedimientos más utilizados para la extracción de intensidades.

MAS 5 or GCOS 1.0 algoritmos de Affymetrics

dChip http://www.dchip.org
Li and Wong (2001). Model-based analysis of
oligonucleotide arrays: expression index computation and outlier detection. *PNAS* 98, 31-36.

RMA (Robust Multichip Analysis) Irizarry *et al* (2003), Summaries of Affymetrix GeneChip probe level data. *NAR* 31(4); CGRMA-MLE; Wu Z, Irizarry R, Gentleman R, Murillo F, Spencer F., 2003, A Model Based Background Adjustment for Oligonucleotide Expression Arrays, Technical Report, John Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD

3.7 Paso final del procesamiento de imágenes

Microarreglos de dos colores

El análisis de las imágenes TIFF de los microarreglos concluyen con las intensidades de señal y background de los canales verde y rojo (Cy3 y Cy5), junto con otras medidas. Las intensidades son exportadas a una serie de archivos de texto plano (GenePix .gpr). En general se trata de un archivo para cada arreglo.

Para iniciar el análisis necesitamos además

i) un archivo que describa las muestras: qué muestra de RNA ha sido hibridizada a cada canal de cada arreglo (Targets File).

ii) un archivo que identifique los probes (Spot Types file (STF)).

Micrroarreglos de un color

Todos los chips de Affymetrix (GeneChips) son escaneados por un escaner de Affymetrix y las cuantificaciones iniciales de las señales de fluorescencia de las imágenes (initial quantification of features), es decir el pasaje de información pixel a pixel a la cuantificación para cada probe del chip, se realiza utilizando el software de Affymetrix. Las diferencias en opinión surgen en como combinar las cuantificaciones de las características (features) de un conjunto de probes (probe set) correspondientes a un gen.

El software genera varios archivos

.EXP Contiene información básica sobre el experimento.
.DAT Contiene la imagen cruda.
.CEL Contiene la cuantificación de las señales (features).
.CDF Vincula, mapea, las features, probes, probe-sets y genes.
.CHP Contiene los niveles de expresión de los genes tal como lo calcula el software de Affy.

LOS DATOS SON LAS IMÁGENES ESCANEADAS

Para seguir leyendo

Qunhua Li, Chris Fraley, Roger E. Bumgarner, Ka Yee Yeung, Adrian E. Raftery. Donuts, Scratches and Blanks: Robust Model-Based Segmentation of Microarray Images. Technical Report no. 473. Department of Statistics. University of Washington. January 2005

Jesús Angulo, Jean Serra, Automatic analysis of DNA microarray images using mathematical morphology. Bioinformatics Vol. 19 no. 5 2003, pages 553–562

Y. Chen, ER Dougherty, and ML Bittner
Ratio based decisions and the quantitative analysis of cDNA microarray images,
J. of Biomedical Optics 2(4), 364-374,1997.

Bengtsson, H., Jönsson, G., Vallon-Christensson, J.
Calibration and assesment of channel-specific biases in microarray data with extended dynamical range. *BMC Bioinformatics*. 2004.

Irizarry, R. A., B. Hobbs, F. C., Beaxer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003a).Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Li Qin, Luis Rueda, Adnan Ali and Alioune Ngom. Spot Detection and Image Segmentation in DNA Microarray Data (2010)
http://davinci.newcs.uwindsor.ca/~lrueda/papers/MicroSegmenAB.pdf

Qiu, P**.**, and Sun, J., "Using conventional edge detectors and
post-smoothing for segmentation of spotted microarray images,"
Journal of Computational and Graphical Statistics,, 18(1), 2009,
147--164.

Schena, M. editor. *DNA Microarrays: A practical approach*.
Oxford University Press, 1999.

Schena, M. editor. *Microarray Biochip Technology.* Eaton*, 2000.*

Schena, M. *Microarray Analysis.* Wiley.2003

Spot: cDNA Microarray Image Analysis Users Guide R. Beare and M. Buckley
October 11, 2004. http://spot.cmis.csiro.au/spot/doc/Spot.pdf

Yee Hwa Yang, Michael J Buckley, Sandrine Dudoit, Terence P Speed. Comparison of methods for image analysis on cDNA microarray data..
Technical report # 584. 2000. Journal of Computational & Graphical Statistics. March 1, 2002, 11(1): 108-136. doi:10.1198/106186002317375640

Yang Y., Dudoit S., Luu P., Speed T. Normalization for cDNA microarray data. Technical Report 2000

Wierling CK, Steinfath M, Elge T, Schulze-Kremer S, Aanstad P, Clark M, Lehrach H, Herwig R. Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. BMC Bioinformatics. 2002.

Wu Z, Irizarry R, Gentleman R, Murillo F, Spencer F., 2003, A Model Based Background Adjustment for Oligonucleotide Expression Arrays CGRMA-MLE Technical Report, John Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD