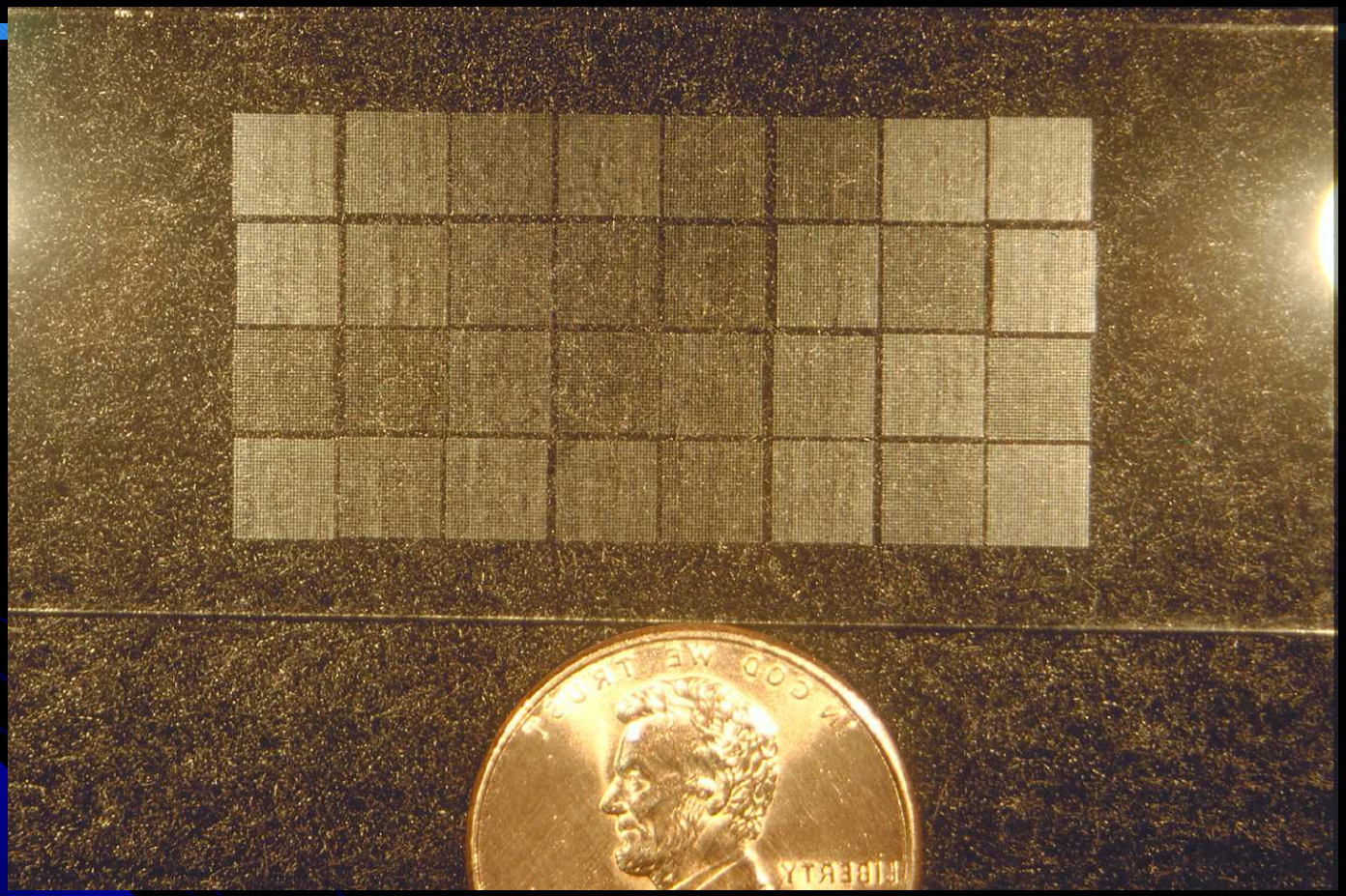
Aplicaciones Microarray

BT740

Ziomara P. Gerdtzen

Microarray de ADN

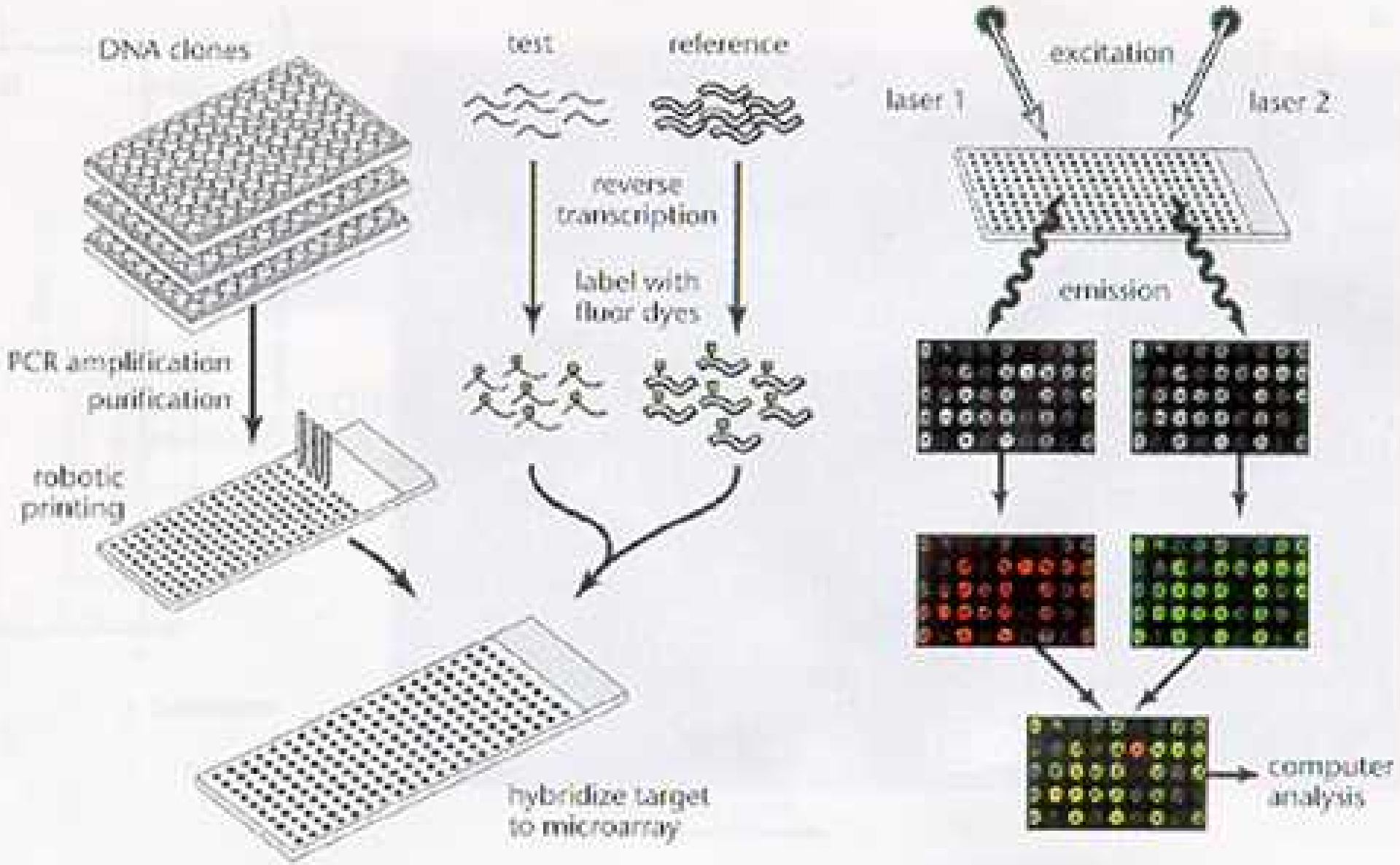


- 18,000 insertos clonales cADN amplificado por PCR

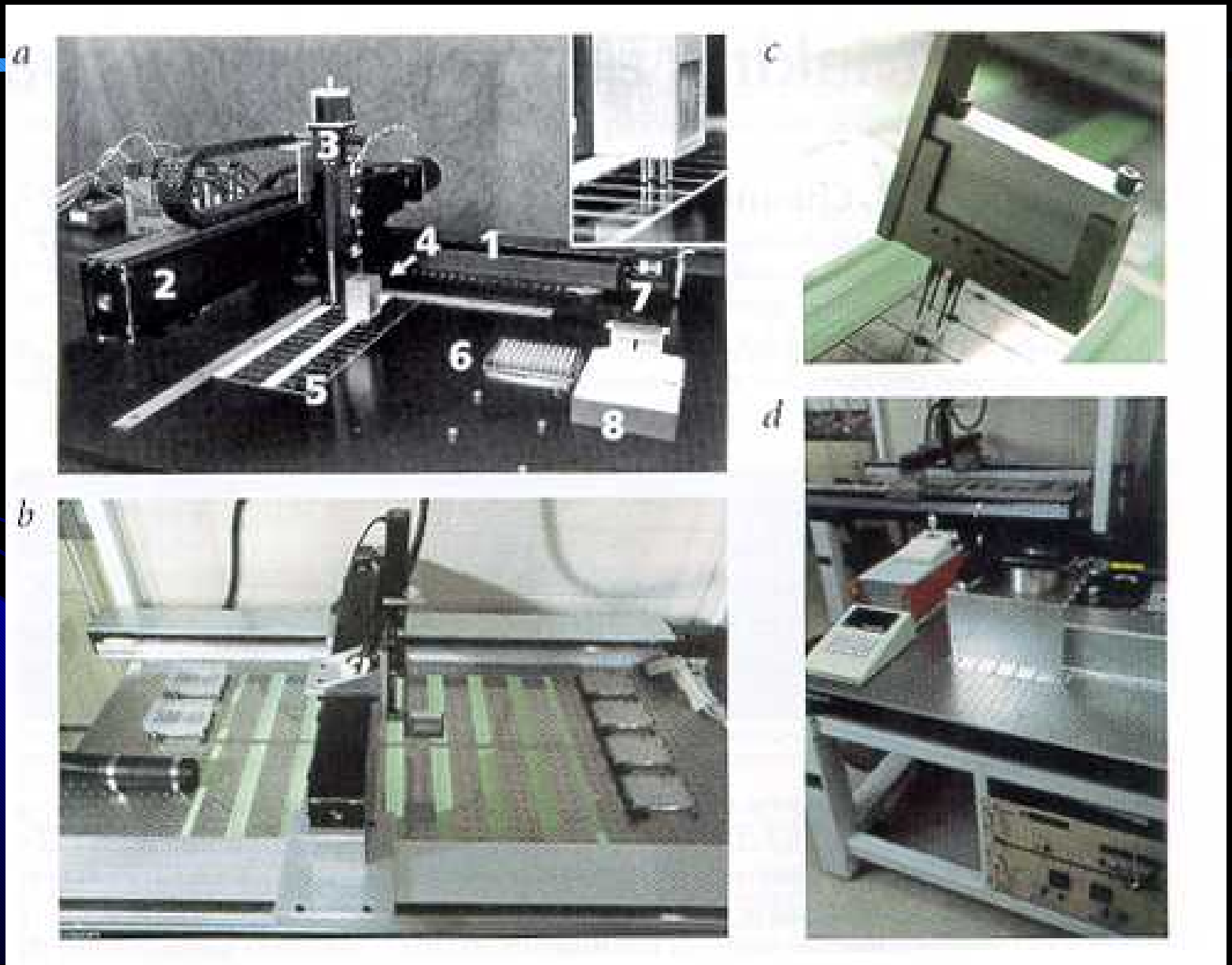
Tipos de Microarrays

- Muestras de ADN mecanicamente impresas/estampadas en un porta objeto
- Impresión Tipo Ink-jet de ADN (Agilent)
- Pequeños oligonucleotidos (~25 nt) sintetizados *in situ* usando fotolitografía (Affymetrix)

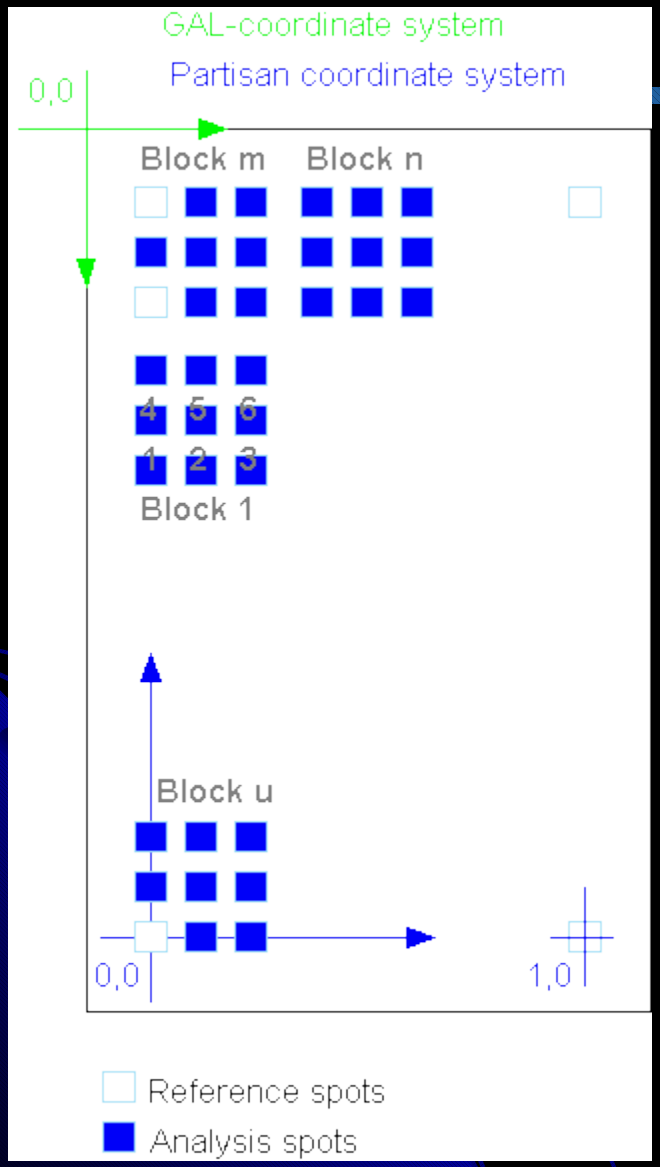
Microarray cDNA

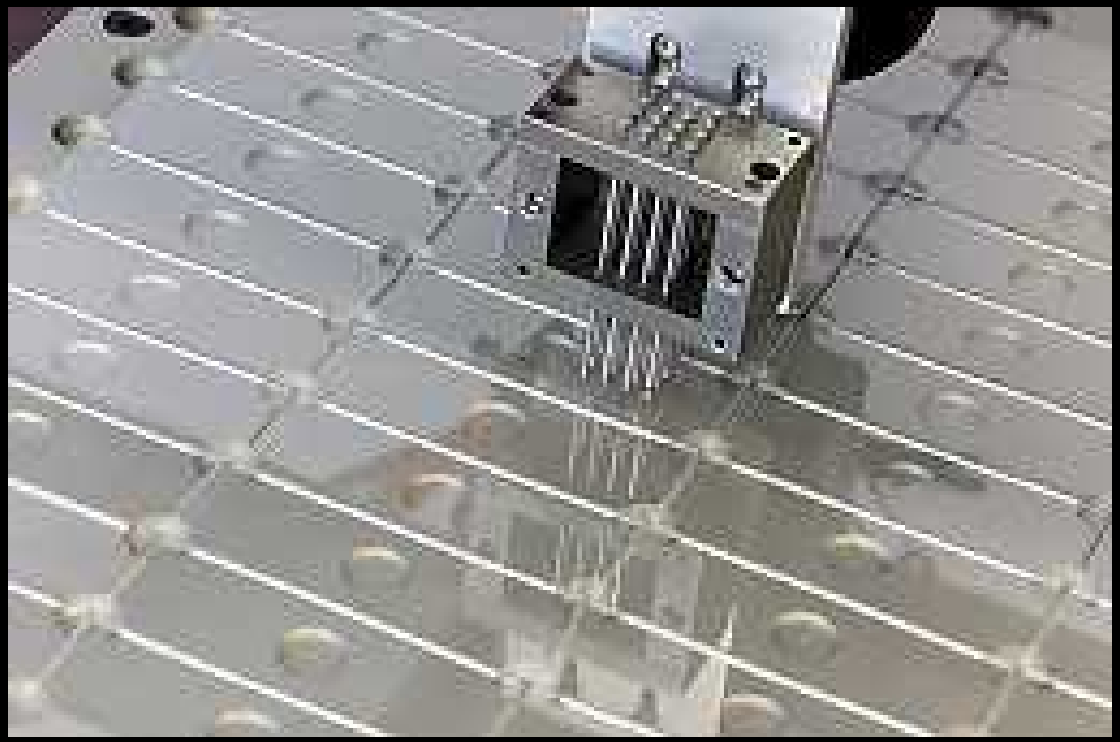


Robot Impresión de Microarrays

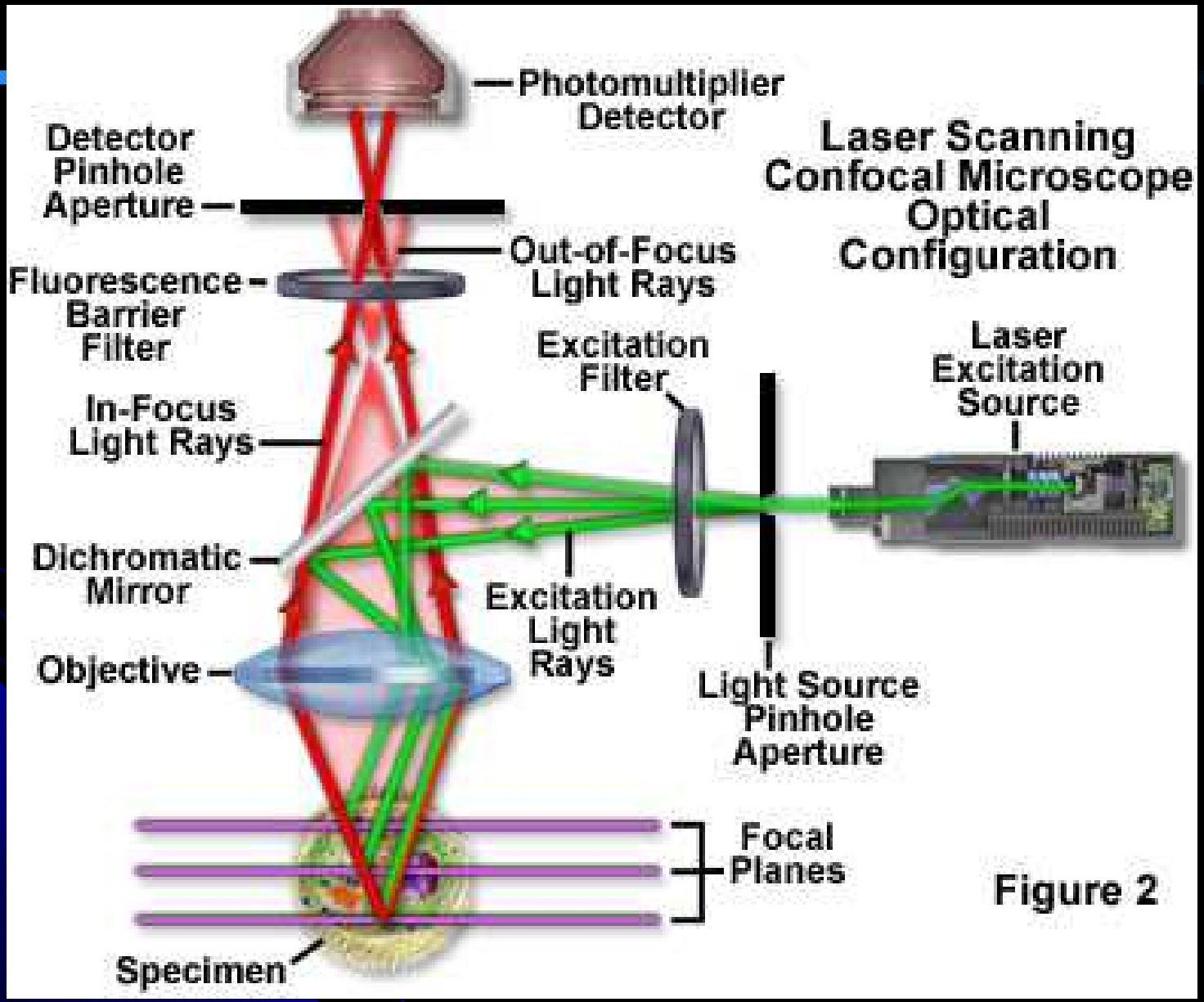


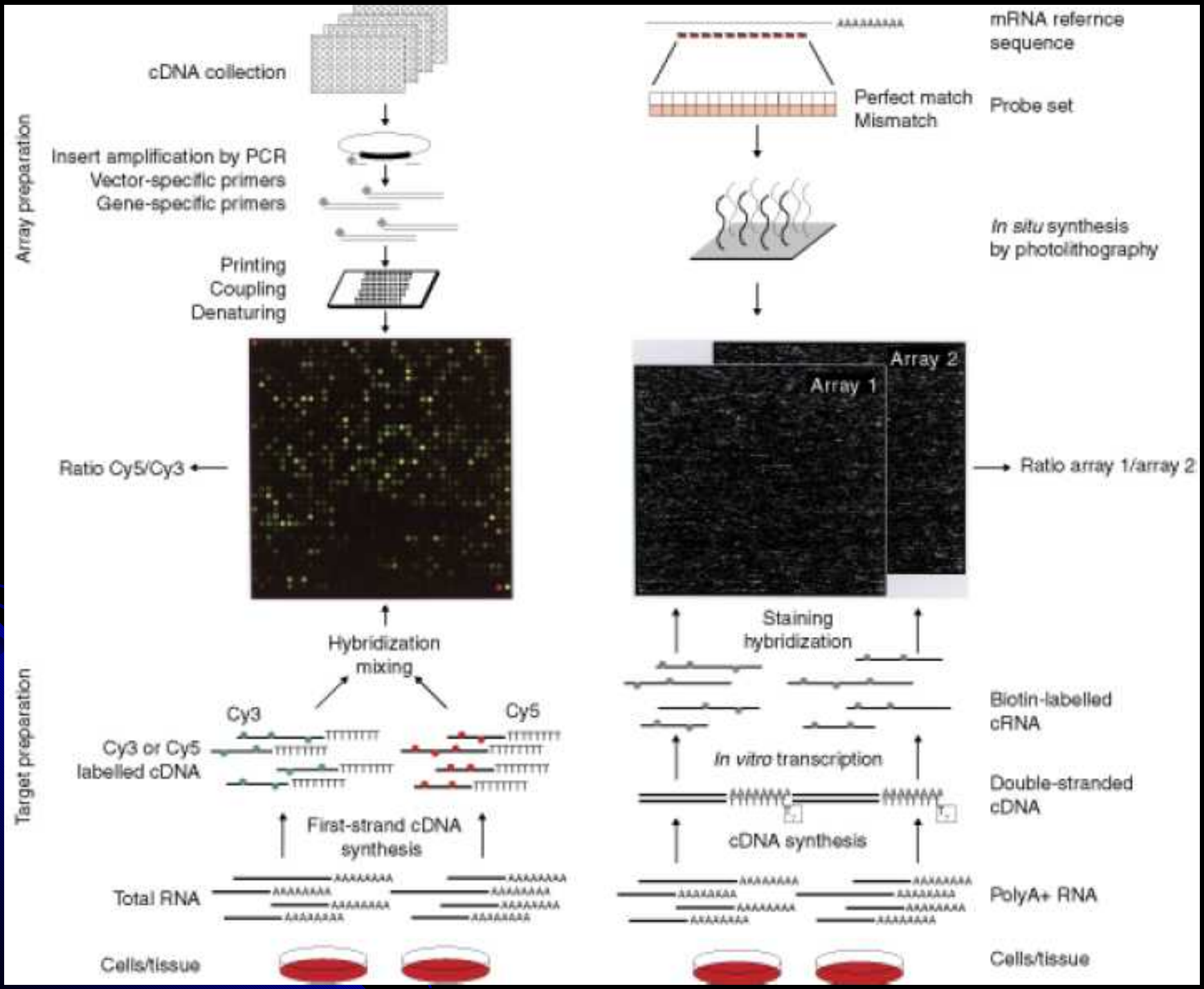
Robot Impresión de Microarrays



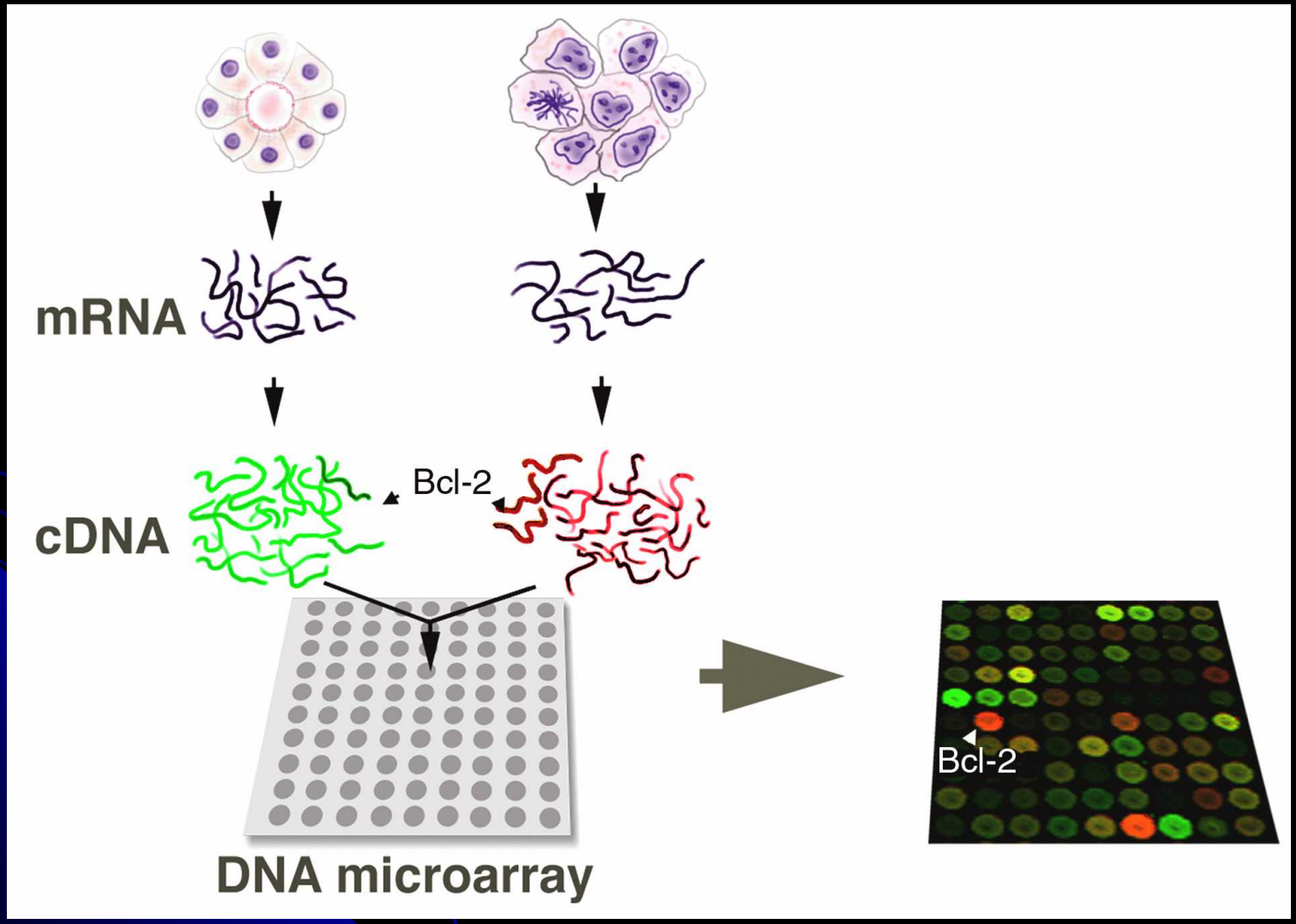


Detección de Fluorescencia




Array preparation
cDNA collection
Insert amplification by PCR
Vector-specific primers
Gene-specific primers
Printing
Coupling
Denaturing
Ratio Cy5/Cy3
Hybridization
mixing
Cy3
Cy5
Cy3 or Cy5
labelled cDNA
First-strand cDNA
synthesis
Total RNA
Cells/tissue
Target preparation
mRNA refernce
sequence
Perfect match
Mismatch
Probe set
In situ synthesis
by photolithography
Array 2
Array 1
Ratio array 1/array 2
Staining
hybridization
Biotin-labelled
cRNA
In vitro transcription
Double-stranded
cDNA
cDNA synthesis
PolyA+ RNA
Cells/tissue

Procedimiento Típico de Microarray



Uso Típico de Microarrays

- Comparación de de expresión de genes en una condición biológica con otra

- Por ejemplo
  - Tumor comparado con tejido normal
  - Celulas mutantes comparadas con wild type
  - Muestras tratadas comparadas con sin tratamiento

Otros Usos de Microarrays

- aCGH : Busca la supresión y amplificación de secciones de cromosomas
- ChIP-chip : Inmuno precipitación de cromatina para identificar regiones union de una proteina en el ADN
- GMS : Identificación de origen parental de cada gen
- Cualquier cosa que puedas pensar para separar dos poblaciones de ácidos nucleicos

Organismos Representados en Microarrays

- Metazoos: ser humano, ratón, rata, lombriz/gusano, insecto
- Hongo: levadura
- Plantas : Arabidopsis y otras
- Bacteria, virus

Ventajas de Experimentos con Microarray

Rápido	Datos de 0.5-30,000 genes en 1-4 semanas
Exhaustivo	Genoma de levadura o bacterial completo en un chip
Flexible	Mientras mas genomas son secuenciados, mas arreglos se pueden hacer
	Arreglos a medida para representar genes de interés
Fácil	Se pueden enviar muestras de ARN para análisis
Barato?	Chip representa 15.000 genes for USD350;
	Robot spotter/scanner cuesta USD100,000

Desventajas de Experimentos con Microarray

Costo	Muchos investigadores no pueden permitirse hacer controles apropiados y replicas.
ARN	El producto final de expresión de un gen es una proteina importante?
Calidad	Imposible para evaluar elementos en la superficie del array
Control	Artefactos con análisis de imagen
	Artefactos con análisis de datos

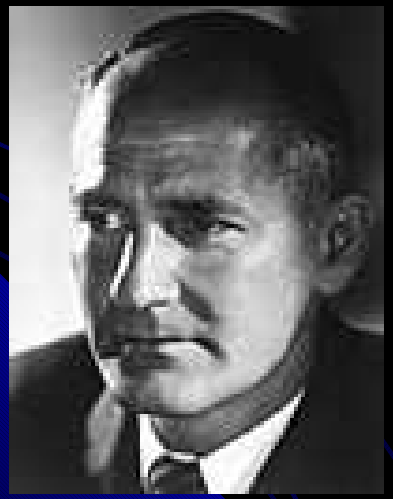
Un Gen en la Célula:

- Función
- Proceso
- Interacción
- Fenotipo

Reconocimiento Oficial de Genética Funcional

Premio Novel de Fisiología o Medicina 1958

"por el descubrimiento que los genes actúan regulando eventos químicos definidos"



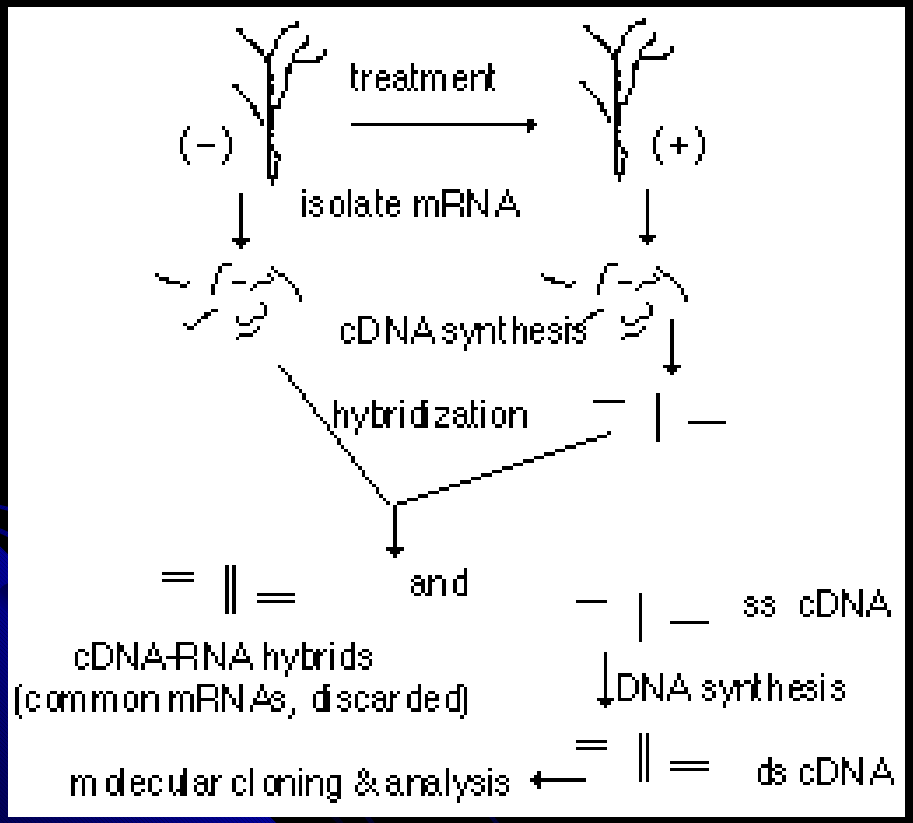
George Wells Beadle

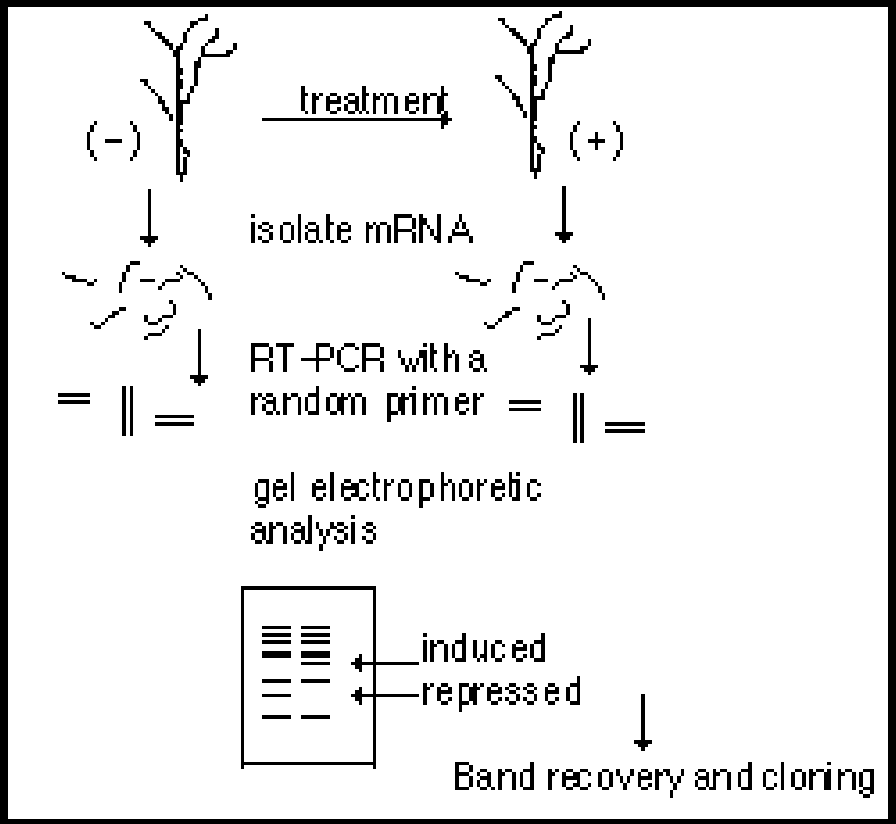
Edward Lawrie Tatum

"por el descubrimiento referente a recombinación genética y la organización del material genético de las bacterias"

Joshua Lederberg

Enfoque Inicial al Análisis Transcripcional Global



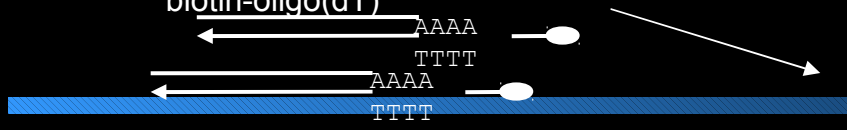


Hibridización Substractiva

Visualización Diferencial

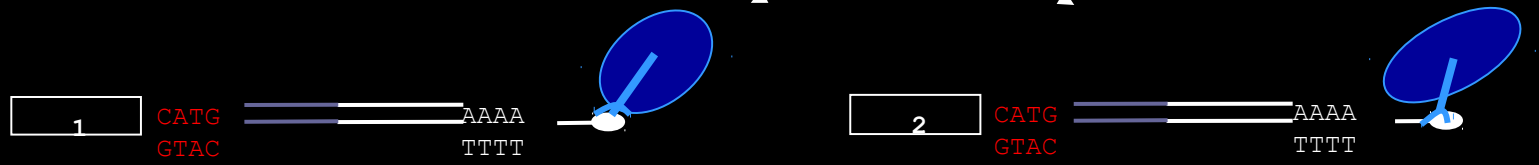
Representación Gráfica de SAGE

Generar cDNA primero con biotin-oligo(dT)

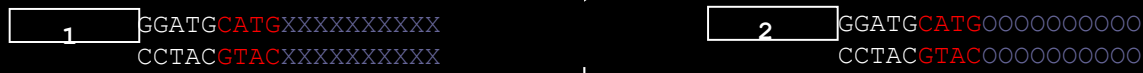


Digerir doble-hebra cDNA con un enzima de restriccion (NlaIII); unir a gránulos cubiertos de streptavidin

Divida la muestra por la mitad y ligúela a linkers (1 o 2), con un sitio de restricción para “enzima marcadora” (BsmFI)



Digerir con una enzima (BsmFI) que reconozca el linker y corta río abajo en una manera de la independiente de la secuencia; Rellena 5’ para extremos romos



Lige los extremos romos de los grupos 1 y 2 y amplifica via PCR con partidores específicas a las secuencias 1 y 2 del linker

Digerir con la misma enzima que ancla (arriba); aislar en gel “di-etiquetas” del gel; concatene los ditags y ligúelos al vector

“Serial analysis of gene expression.”
Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Science. 1995 Oct 20;270(5235):484-7.

Comienzos de la Genómica Funcional

- Nat Genet. 1993 May;4(1):11-8 Genomic mismatch scanning: a new approach to genetic linkage mapping. Nelson SF, McCusker JH, Sander MA, Kee Y, Modrich P, Brown PO.
- J Bacteriol. 1993 Apr;175(7):2026-3 Global regulation of gene expression in Escherichia coli. Chuang SE, Daniels DL, Blattner FR.
- Nature 1993 Aug 5;364(6437): 555-6 Multiplexed biochemical assays with biological chips. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL.
- Science 1995 Oct 20;270(5235):467-70 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Schena M, Shalon D, Davis RW, Brown PO.

Microarrays: una Herramienta para Monitorear Expresion de Genes

- **Un microarray es un soporte sólido (como una membrana o portaobjeto) en la cual ADN de secuencia conocida se deposita en arreglo cuadriculado.**
- **ARN esta aislado de las muestras de interés identificadas**

- El ARN es típicamente convertido a cDNA, etiquetado con fluorescia (o radioactividad), y luego hibridizado en el microarray para medir niveles de expresión de miles de genes.

Trabajo con Datos de Microarray

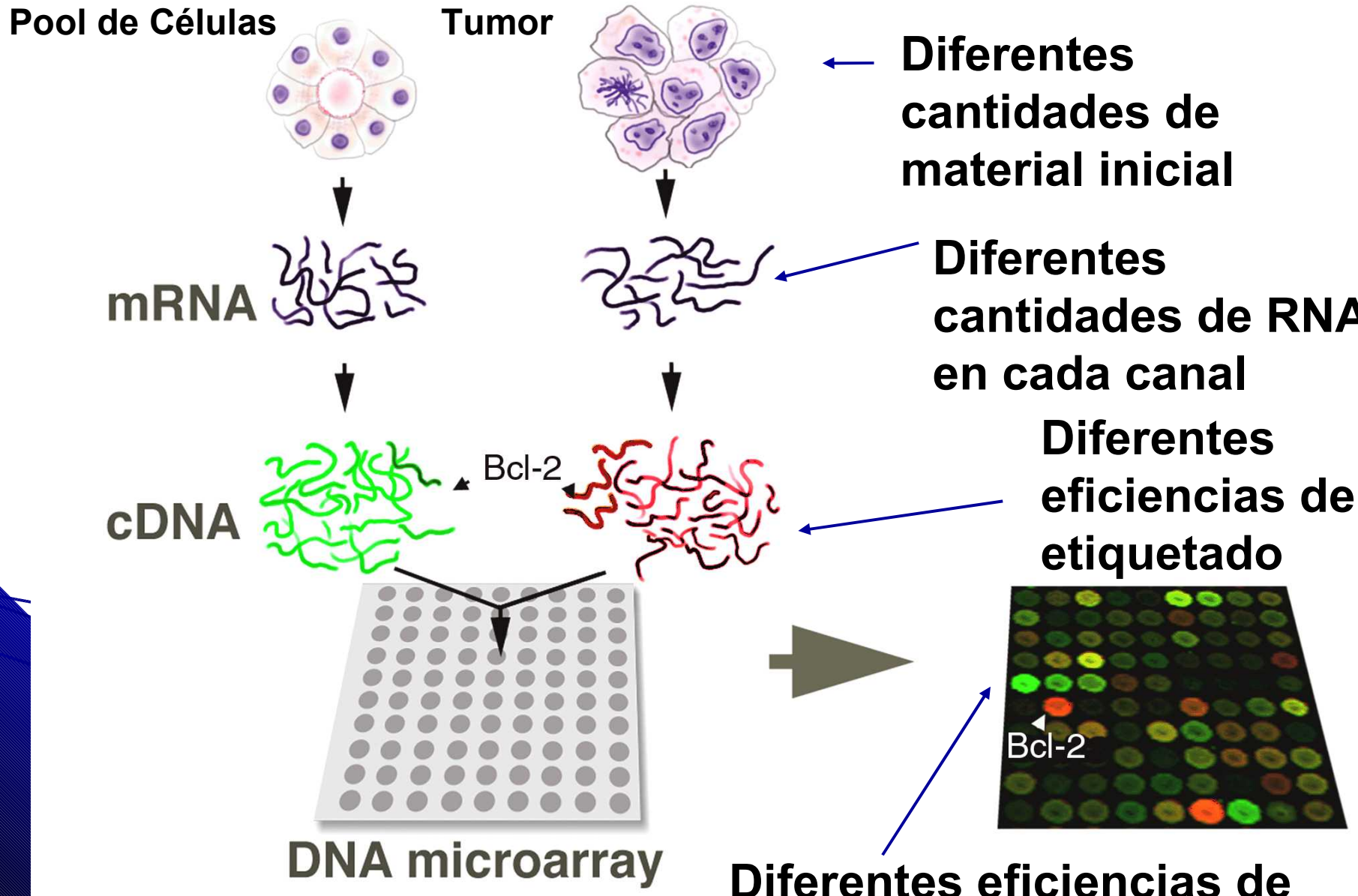
- Comienza con la matriz de datos

UID	NOMBRE	Condición uno	Condición 2
B0001	thrL	0.78	0.52
B0002	thrA	0.22	- 0.04
B0003	thrB	0.19	0.16
B0004	thrC	0.42	- 0.07
B0006	yaaA	- 0.01	0.3
B0008	talB	2.03	1.66
B0009	mog	0.57	0.42

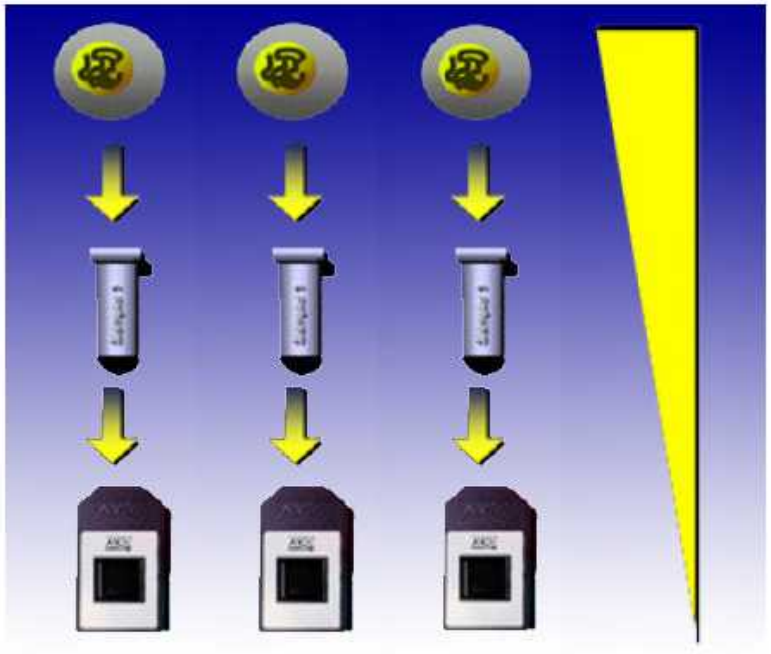
- Preprocesamiento
  - Estadística Inferencial
  - Estadística Descriptiva

Normalización de Datos

- Se puede introducir un sesgo sistemático en experimentos microarray en todas las etapas
- Se debe
  - Evitar (tanto cuanto sea posible)
  - Reconcer
  - Corregir
  - Descartar los datos irrecuperables


Pool de Células
Tumor
Diferentes cantidades de material inicial
mRNA
Diferentes cantidades de RNA en cada canal
cDNA
Bcl-2
Diferentes eficiencias de etiquetado
DNA microarray
Bcl-2
Diferentes eficiencias de hibridación en la superficie del array

Controlling the Sources of Variability



Biology

The main source of variability

Sample preparation

Depends on method and operator

Probe array processing

Depends on chips, instruments, reagents and operator

- Gender
- Age, cause of death, date/year
- Cell Cycle Patterns – time of day
- Tissue – cell type
- Diet – Eating habits, media types

Define processes and boundaries

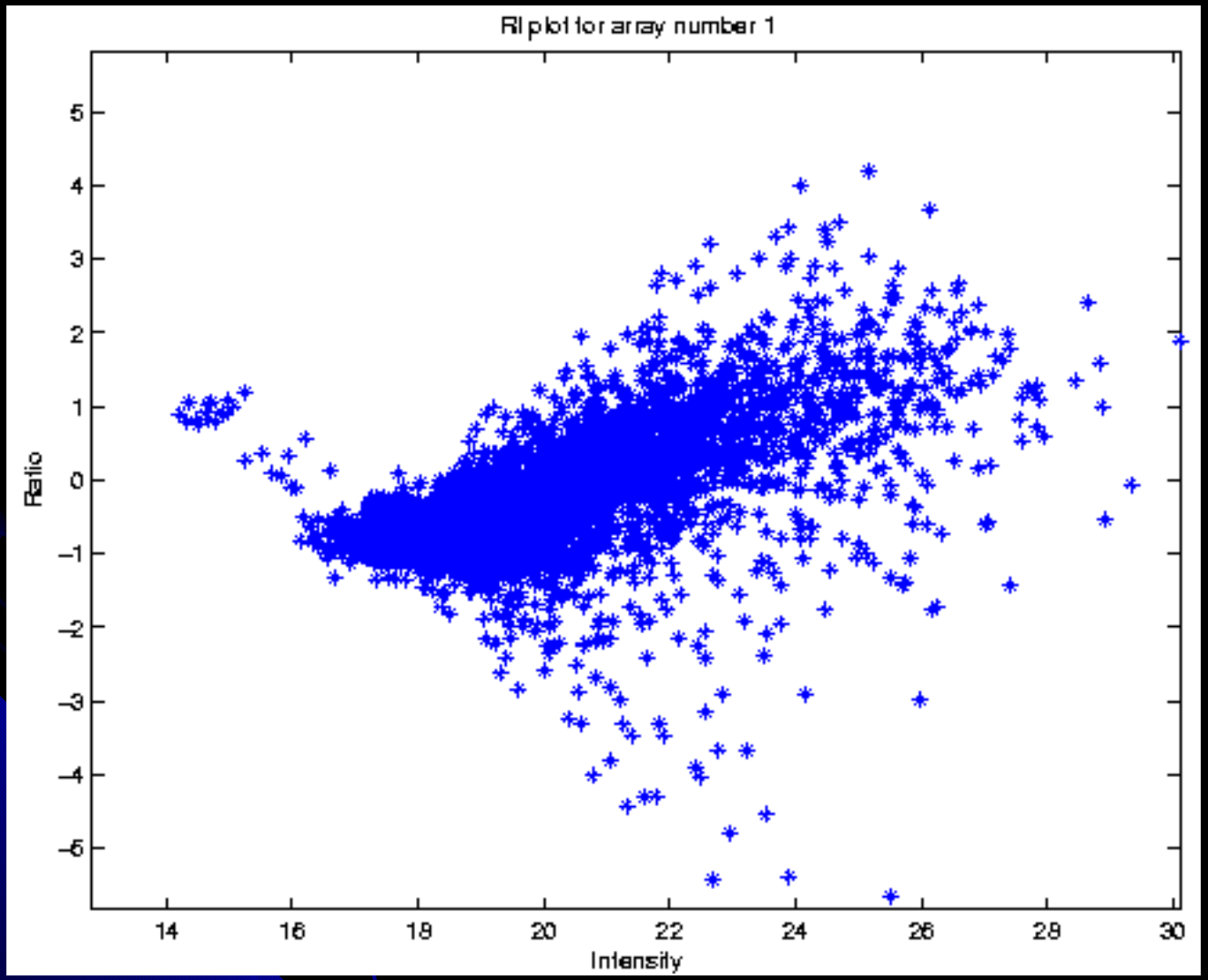
- Standard protocol, Sample quality control

Reduce process variability

- Assess technician-to-technician differences, Calibrate instrumentation, Control reagent variability

- Measured by hybridizing the same sample over two different arrays of the same type.
- Try to always use the same Core Facility to process the samples.

Tales Sesgos Tienen Consecuencias

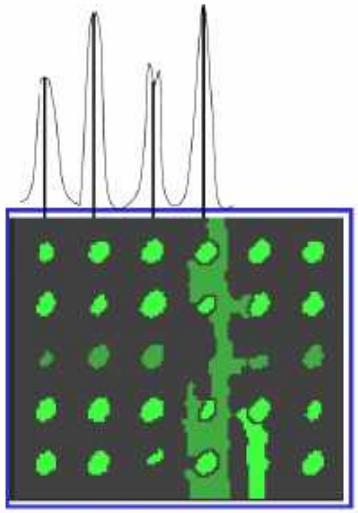


Cómo Lidiamos con las Consecuencias?

Normalización:

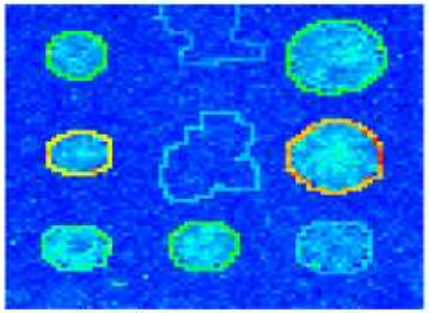
- **Se asume *generalmente* que el gen promedio no cambia**
- **Debe entender los datos, para saber si eso es una suposición apropiada o no**
- **El número de "reporteros" (clones o genes) que se está probando afectará esto**

cDNA Microarray Image Analysis

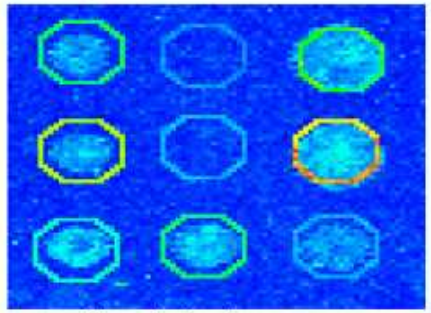


Addressing:
estimate location of the spot center.

Segmentation:
Classify pixels as foreground (signal) or background.



adaptive segmentation
seeded region growing

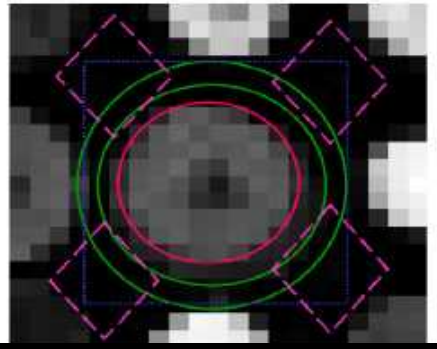


fixed circle
segmentation

Information extraction:
For each spot on the array and each dye

- foreground intensities;
- background intensities;
- quality measures.

---- GenePix
---- QuantArray
---- ScanAlyze



Qué es la Normalización de Datos?

- La normalización es una intento de corregir el sesgo sistemático en los datos
- Permite comparar datos de un array a otro
- En la práctica no entendemos siempre los datos – parte de la biología será eliminada inevitablemente (o por lo menos no revelada)

Métodos de Normalización

- Genes domésticos (Housekeeping genes)
- Intensidad total
- Corrección de Lowess
- Versiones locales de dos antedichos
- Controles

Housekeeping Genes

- Asume que los genes elegidos no cambian
- Definitivamente una mala opción - en bacterias no hemos visto ningún gene que no cambia bajo por lo menos alguna circunstancia

Normalización de la Intensidad Total (Promedio Global o Mediana)

- Para los puntos que se creen bien medidos, calcule la razon logaritmica promedio o mediana (log ratio)
- Utilice esto como factor de normalización para ajustar el resto de los log ratio
- Se debe definir “bien medida”
- Equivalente a asumir la misma intensidad total en ambos canales

Diagnóstico del Array: Diagrama Intensidad-Cuociente

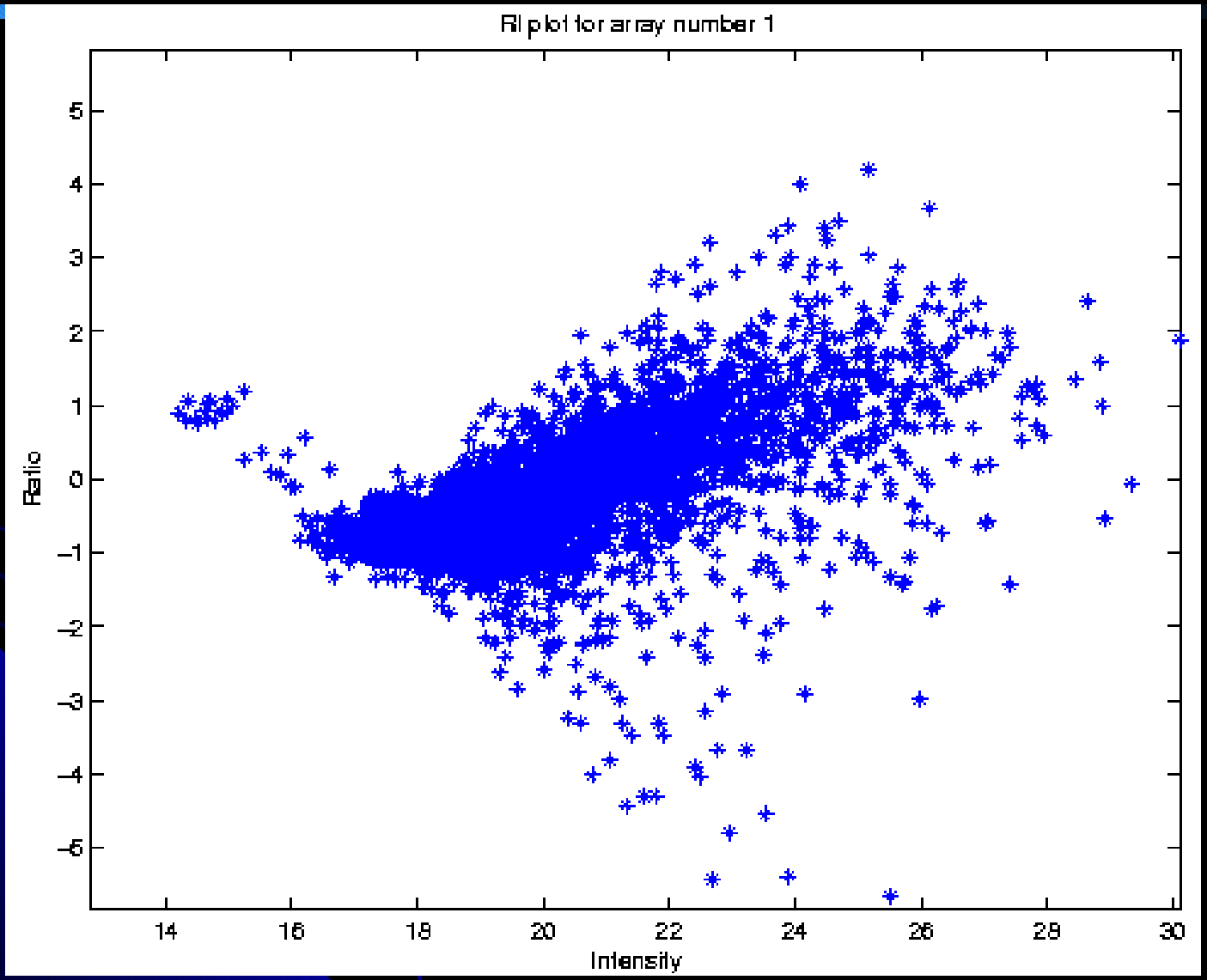
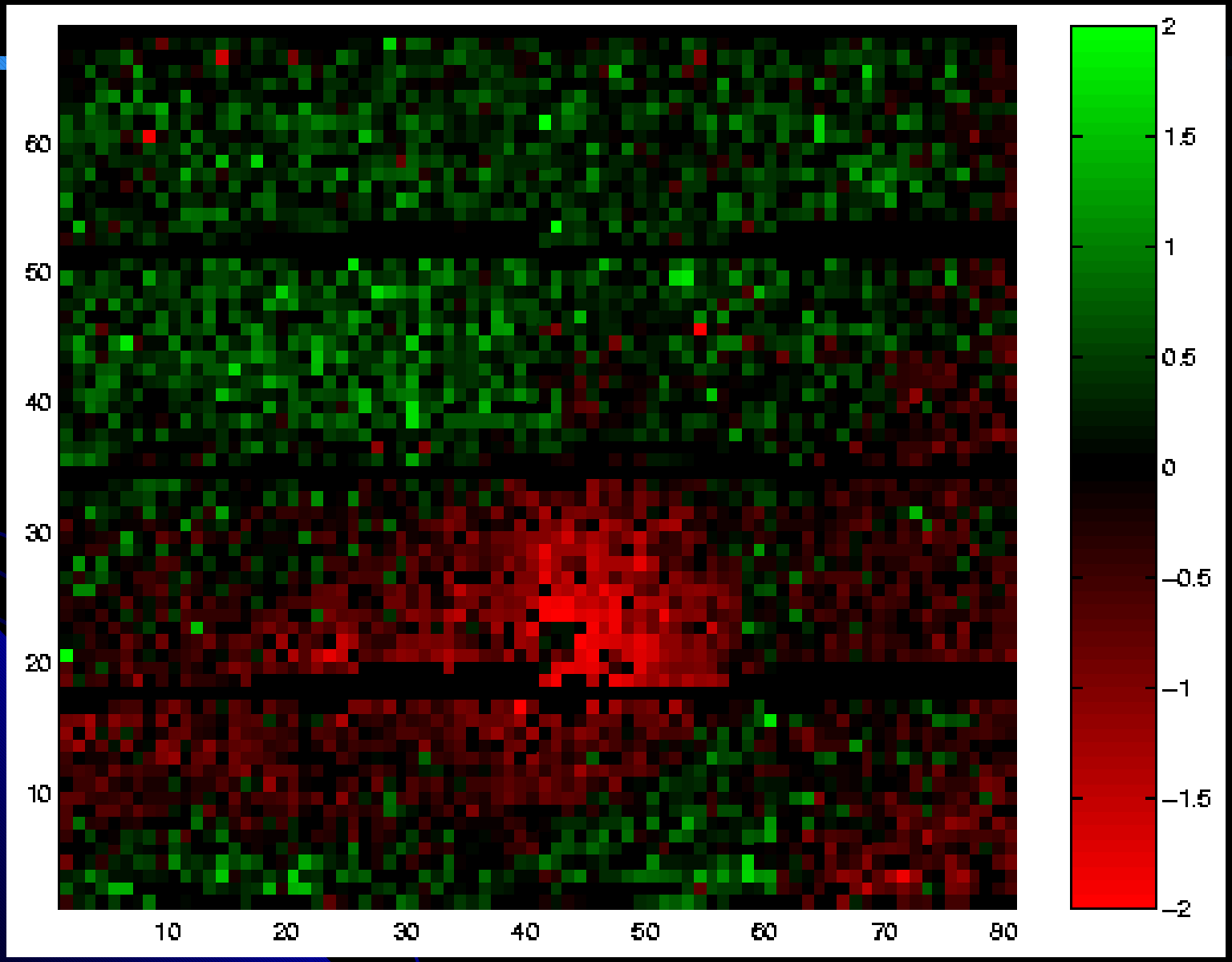
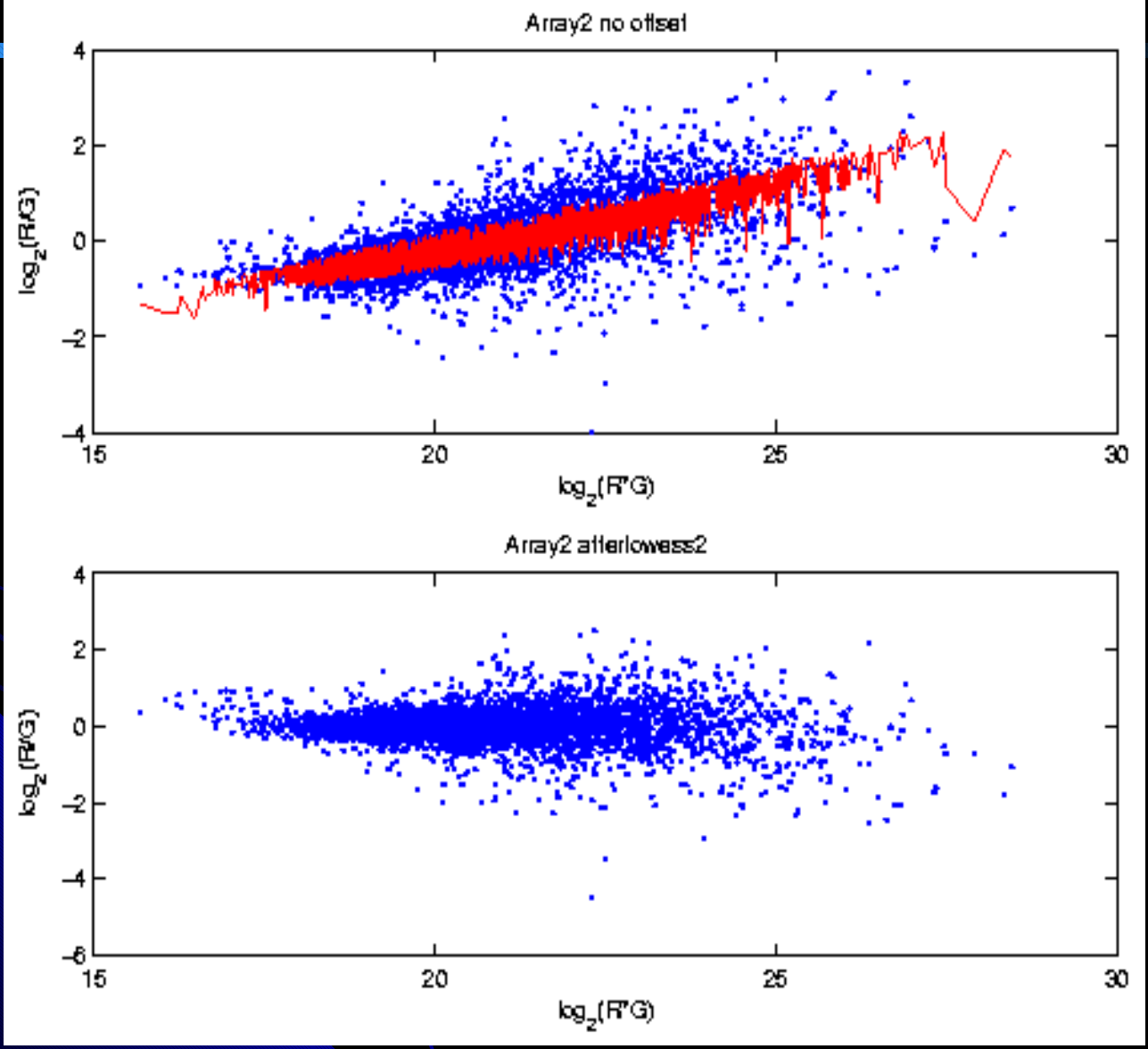


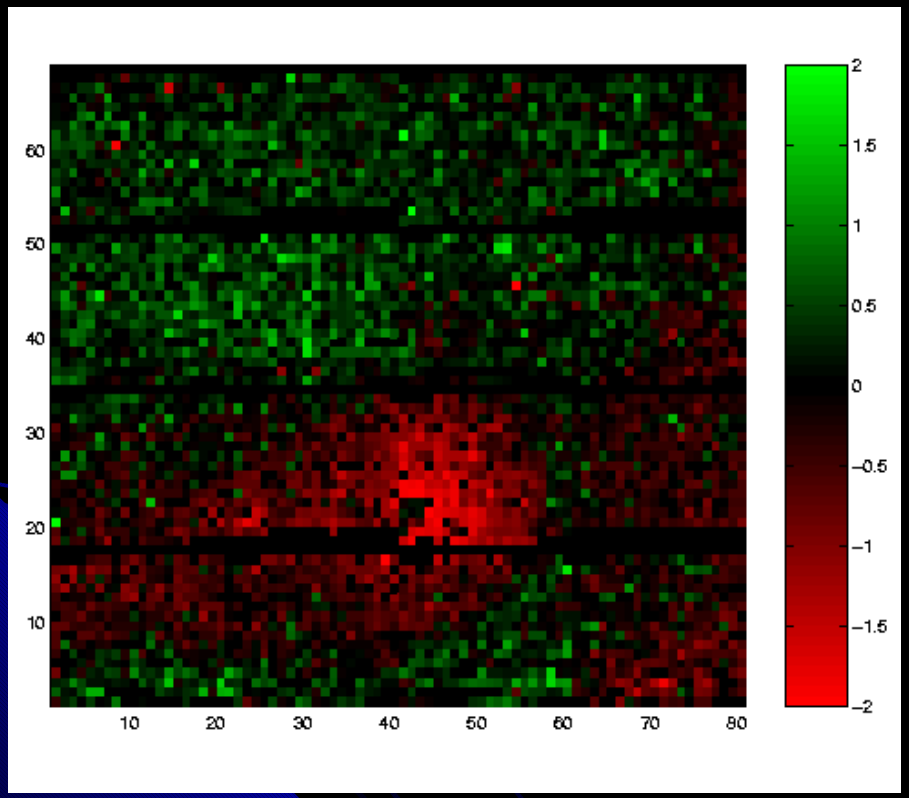
Imagen del Array

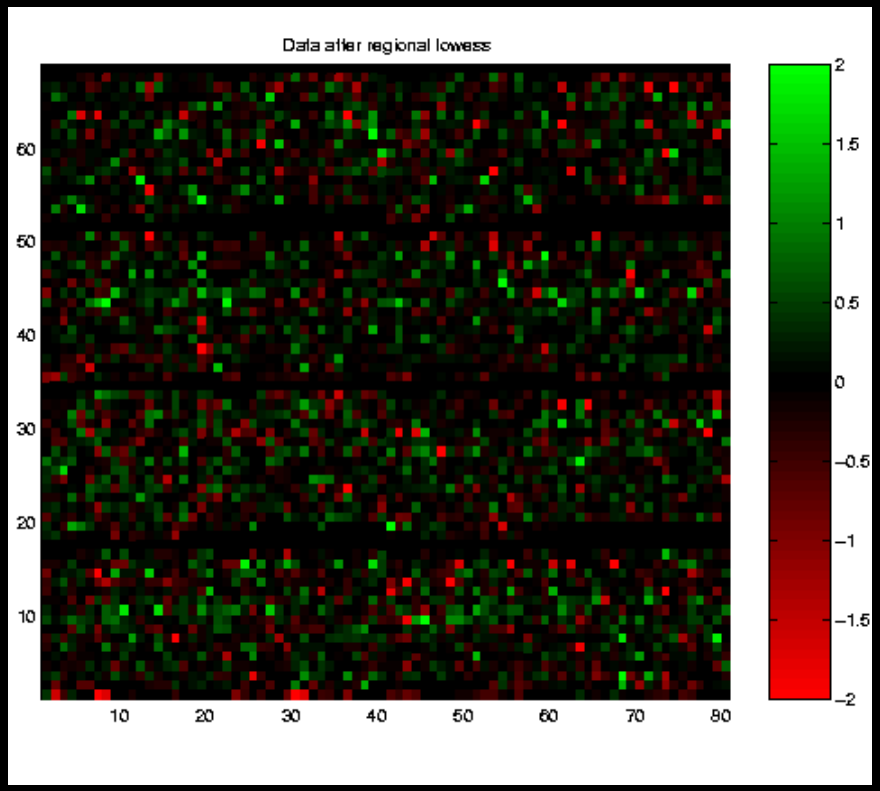


Suavizado de Lowess

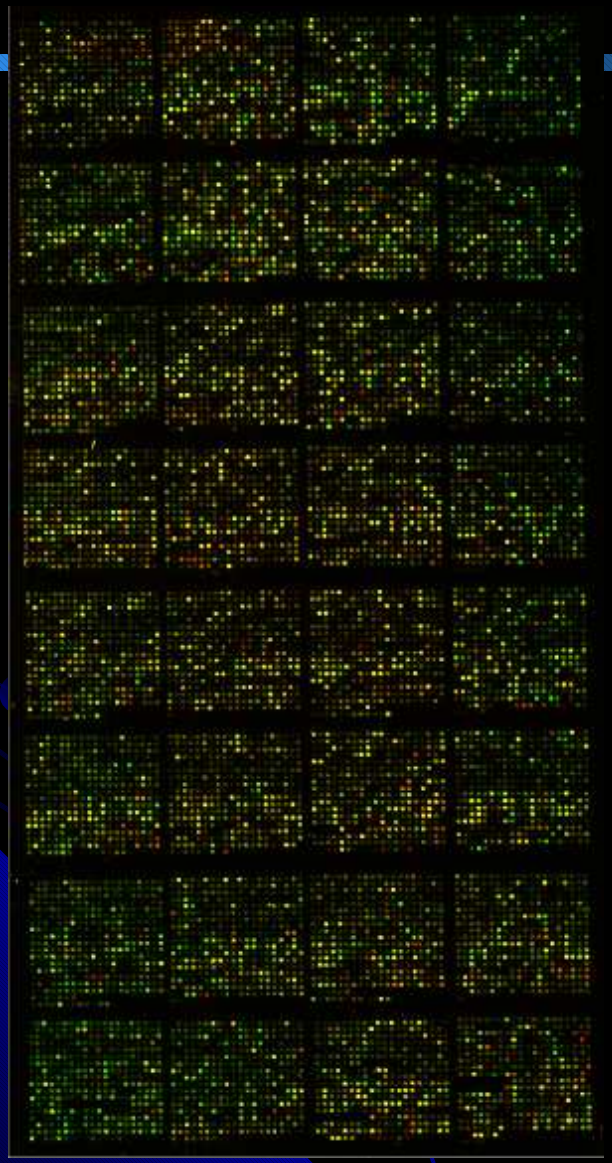


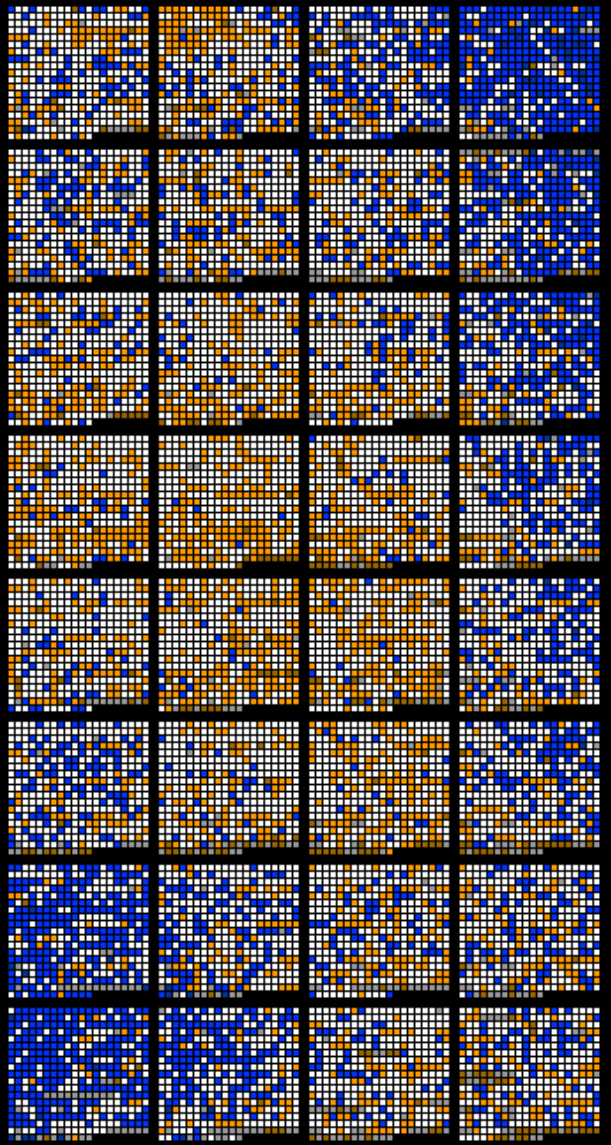
Resultados de Suavizado Local





Sesgo Espacial (Bias)





LOWESS Sector-Específico

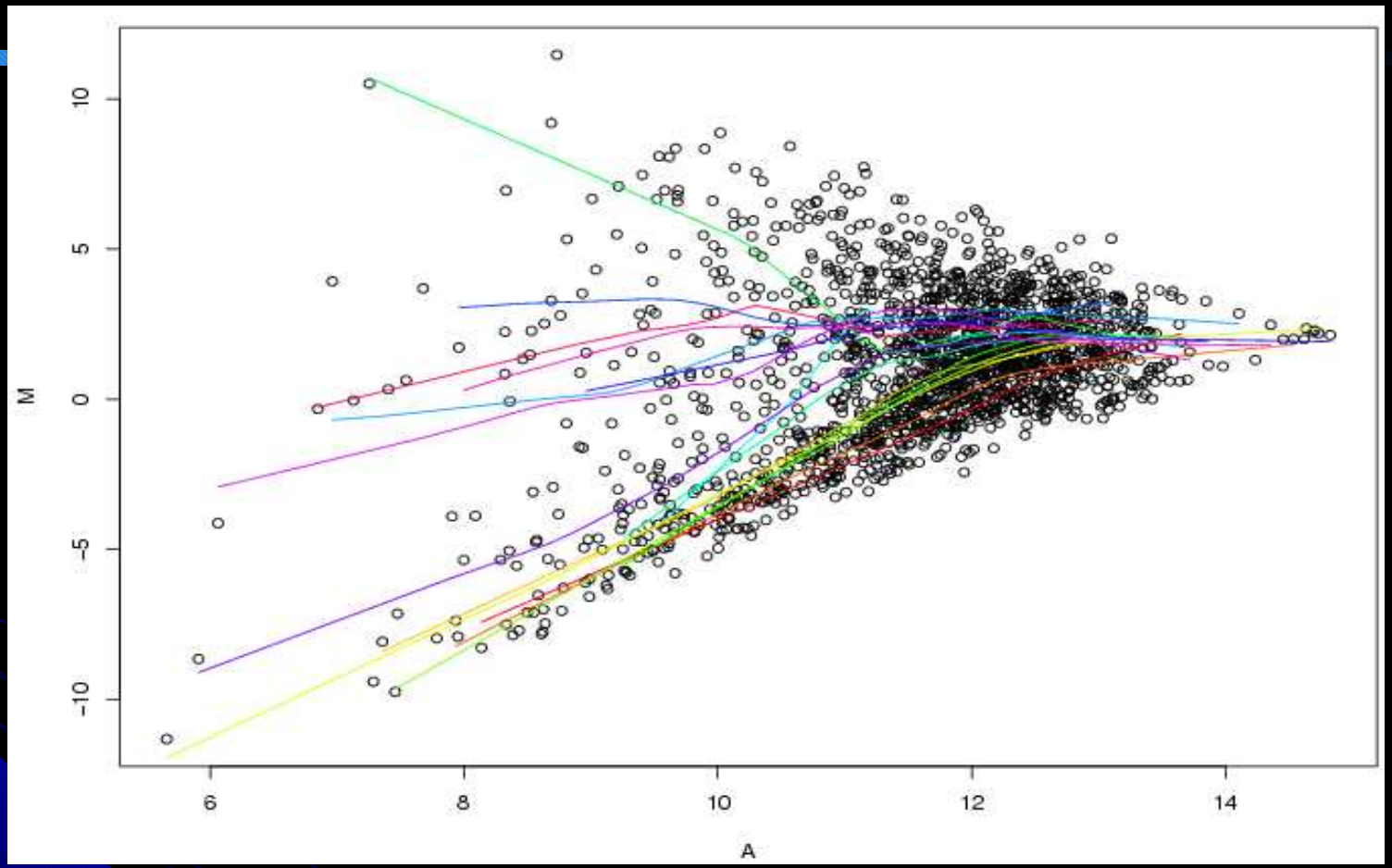
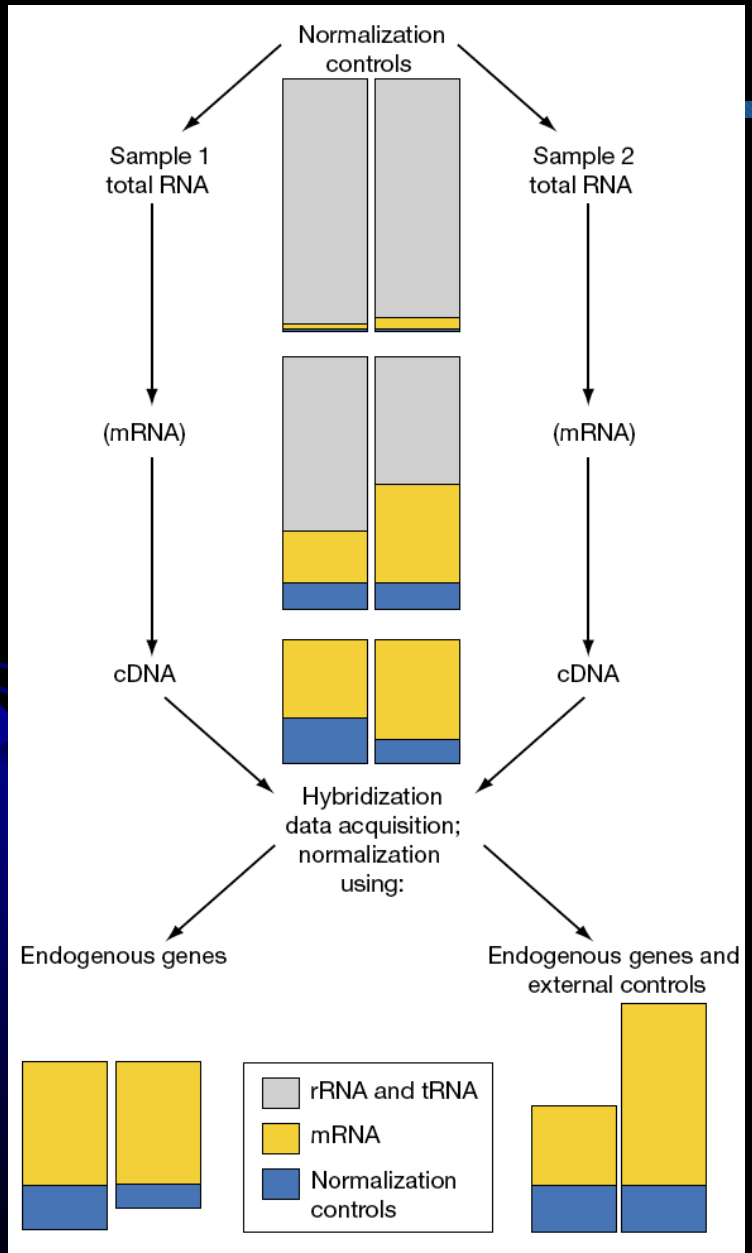


Diagrama M-A sobrepuesto para 16 lowess suavizados individuales, uno para cada grupo de impresión. M es la razón logarítmica, A es la media geométrico.

Controles Agregados

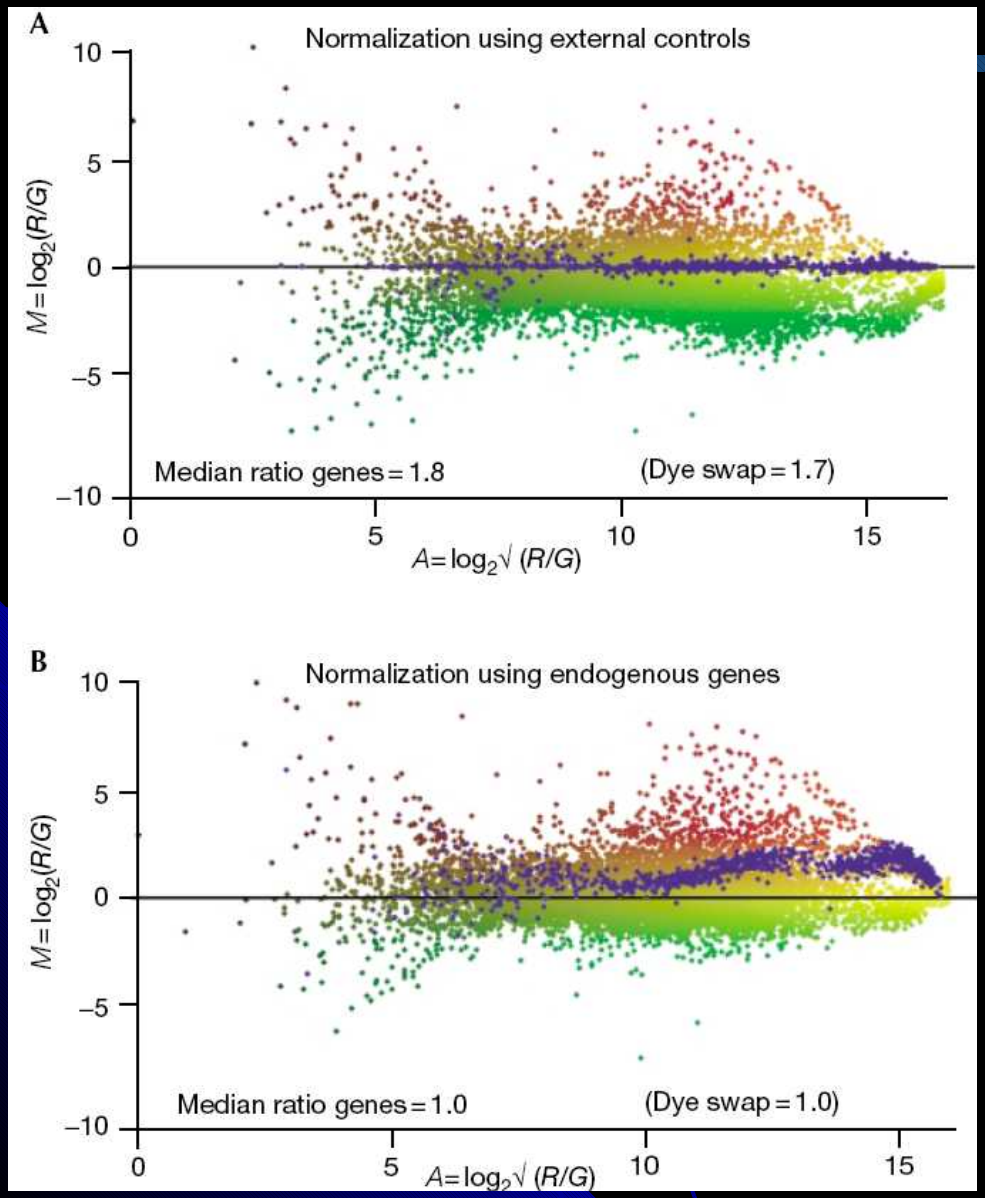
- Agregar a la mezcla de etiquetado moléculas conocidas a concentraciones conocidas
- Atraviese toda la gama de cuocientes y de intensidades.
- Pero debe usar cantidades iguales de material inicial para etiquetar
- Potencialmente más exacto, pero requiere gran cuidado
- van de Peppel *et al*, 2003

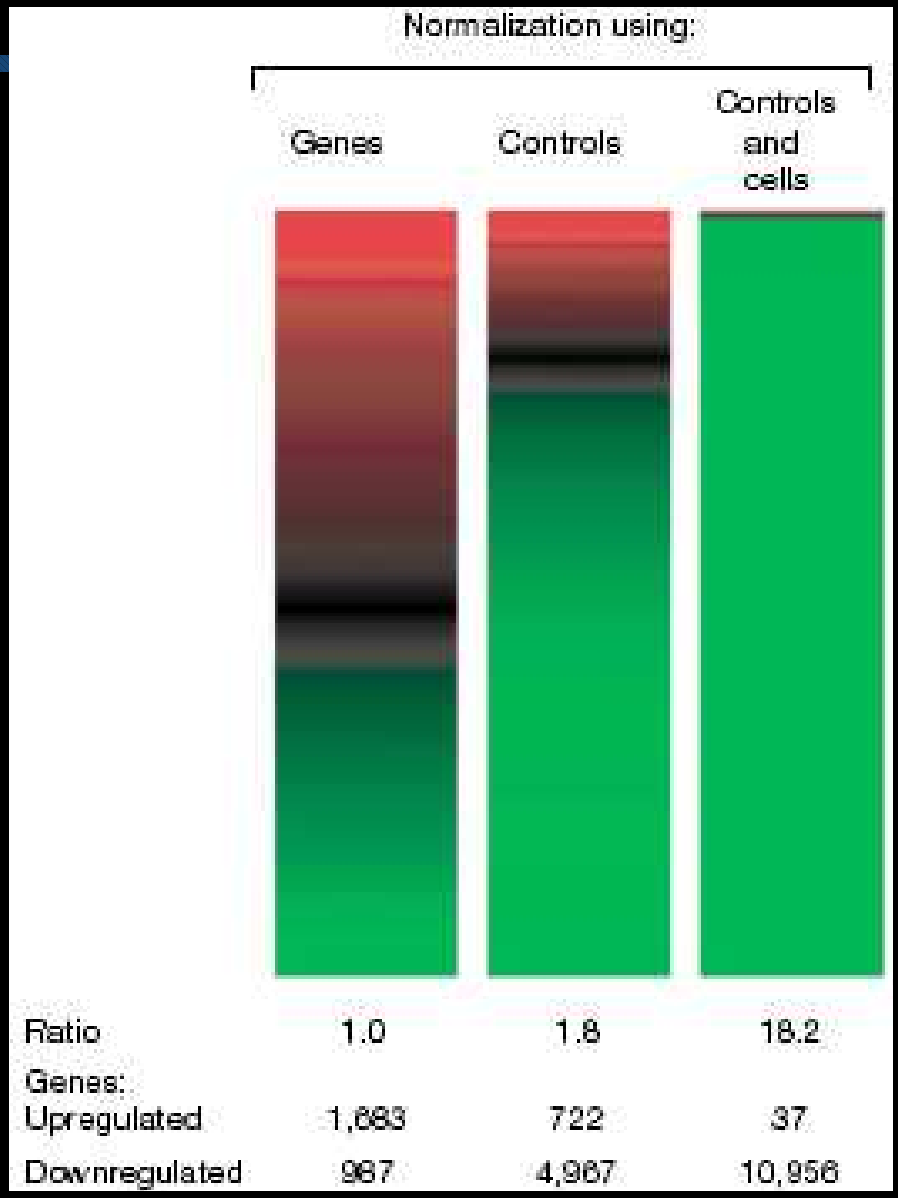
Razonamiento para Controles Externos



- La muestra 2 contiene más mRNA
- Los controles son agregados en proporción al RNA total
- Se obtienen respuestas muy diferentes con controles vs. sin control

Estrategia y Resultados

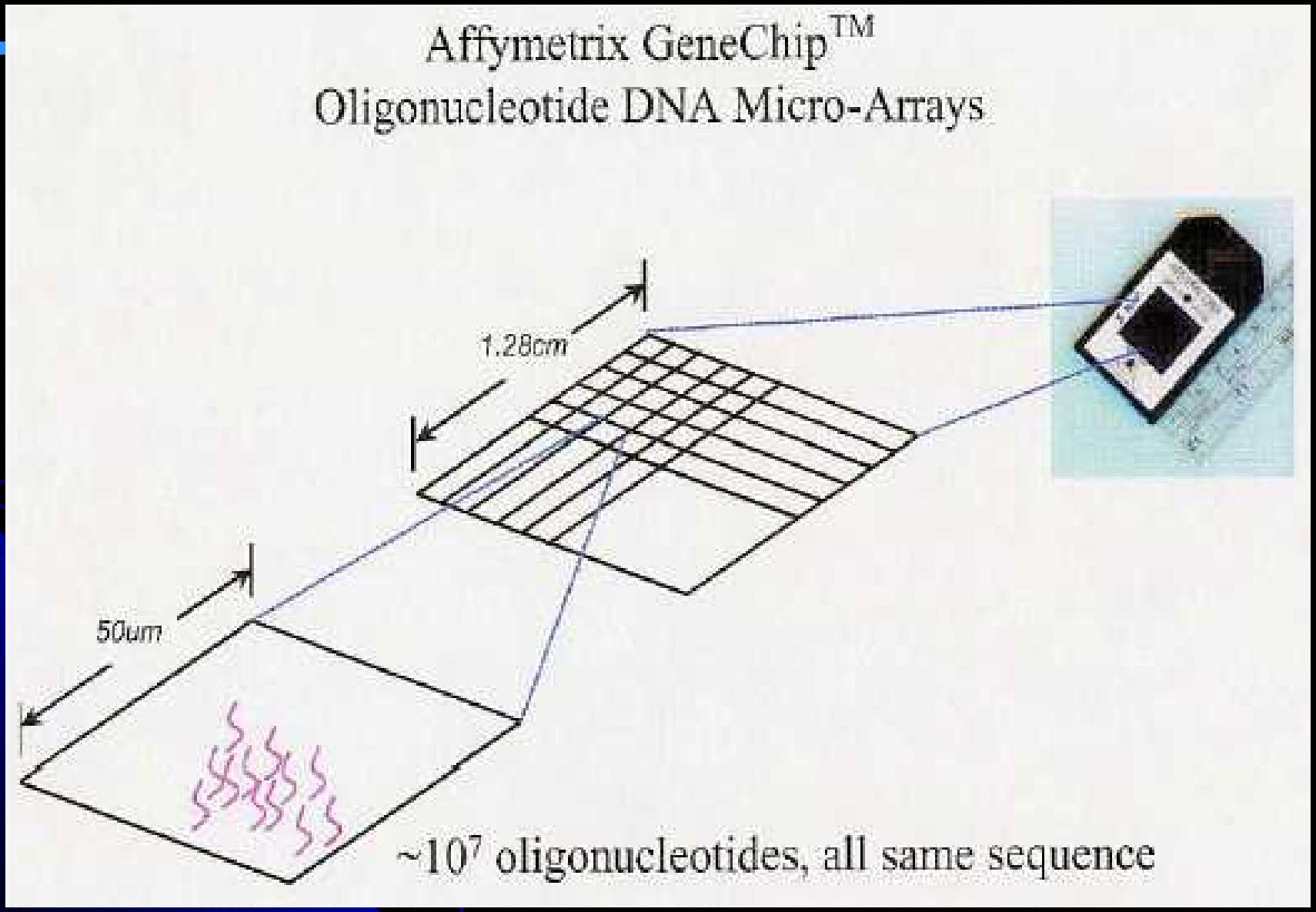




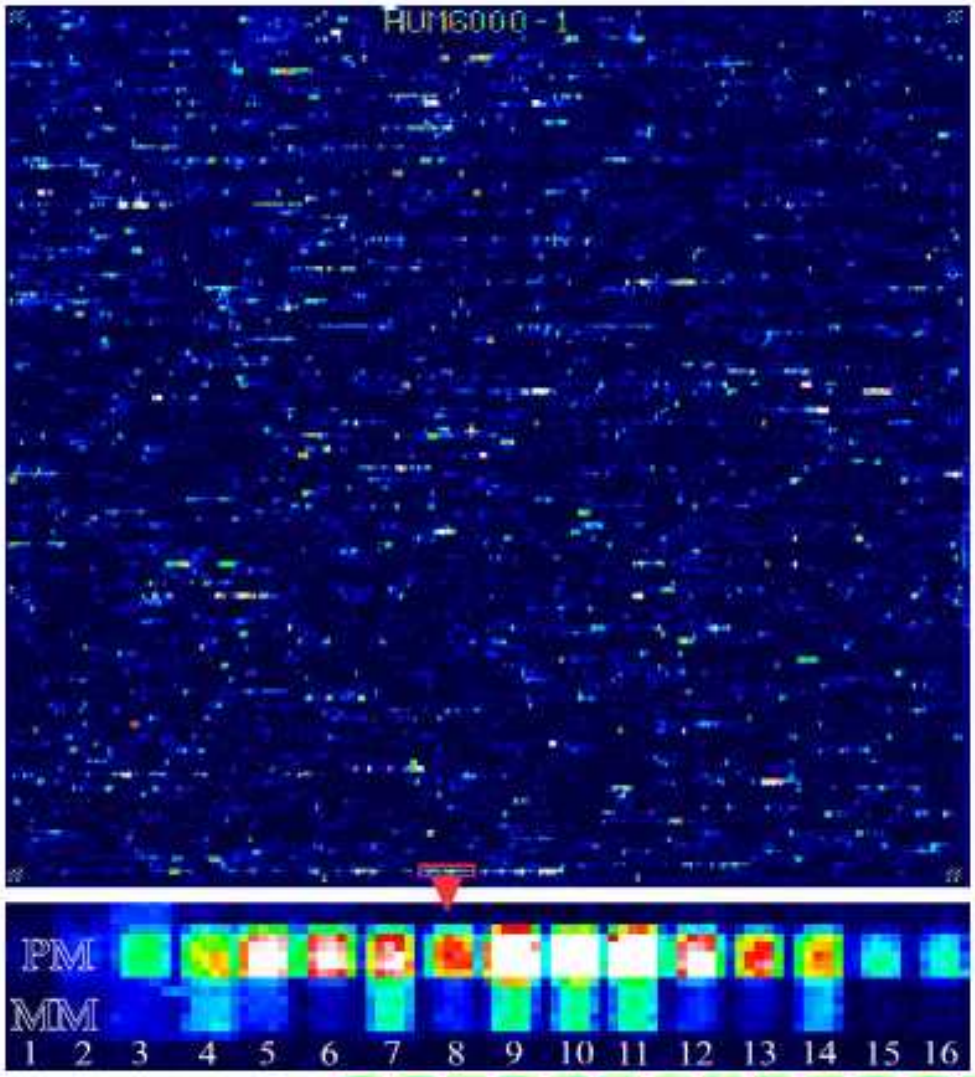
Resumen de Normalización

- Los sesgos sistemáticos existen en datos de microarray
- La normalización puede quitar estos sesgos, y dejar la biología detrás
- PERO, si los supuestos son incorrectos, se pueden introducir nuevos errores!
- La corrección de Lowess de cada sector del microarray basado en controles externos es probablemente el mejor método
- El mejor método es reducir al mínimo los errores, y utiliza buen diseño experimental

Affymetrix GeneChip



Affymetrix GeneChip



The Affymetrix GeneChip

Probe: 25 bases long single stranded DNA oligos

Probe Cell: Single square-shaped feature on an array containing one type of probe. 24-50μm.

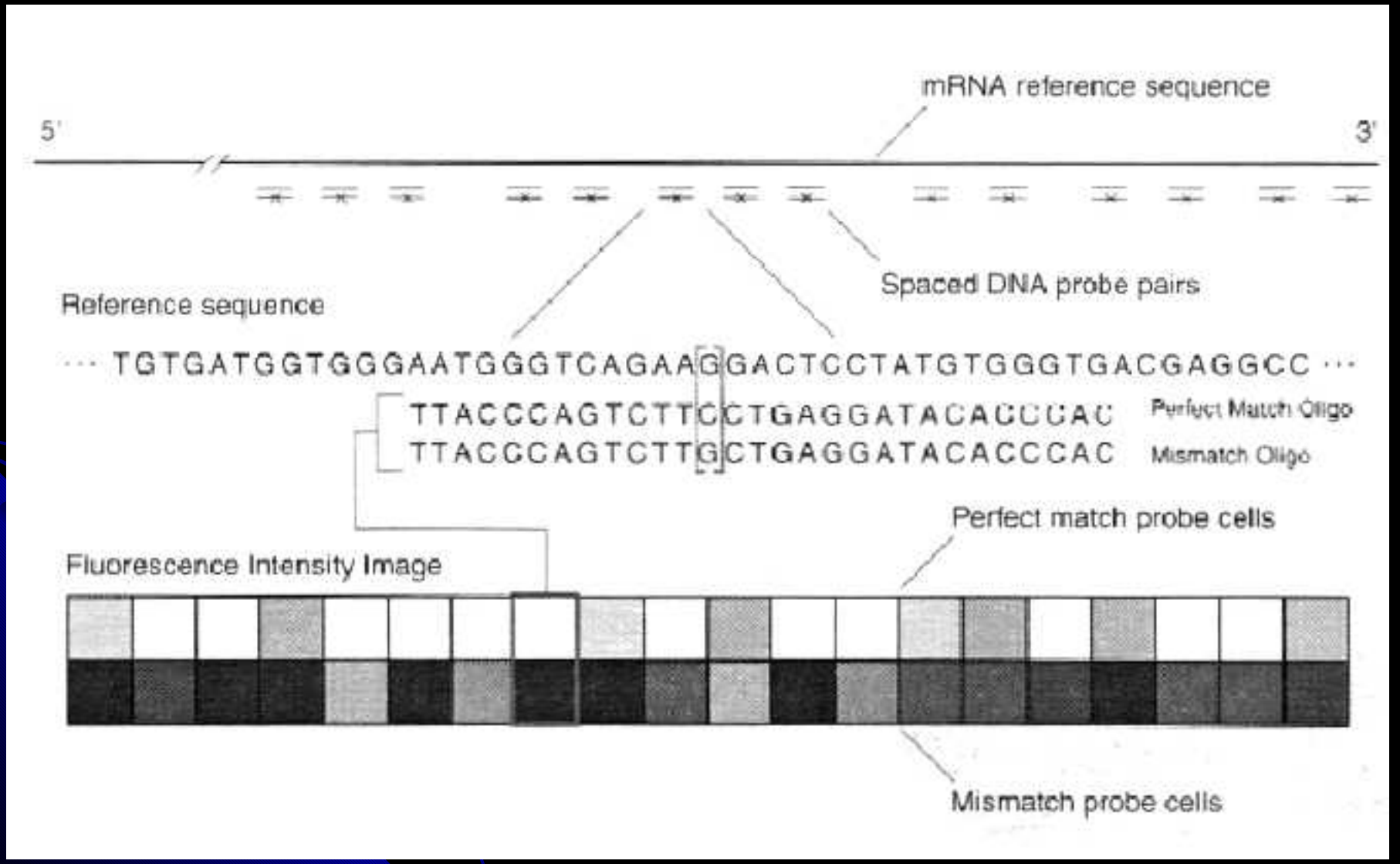
Contains millions of probe molecules

PM

MM

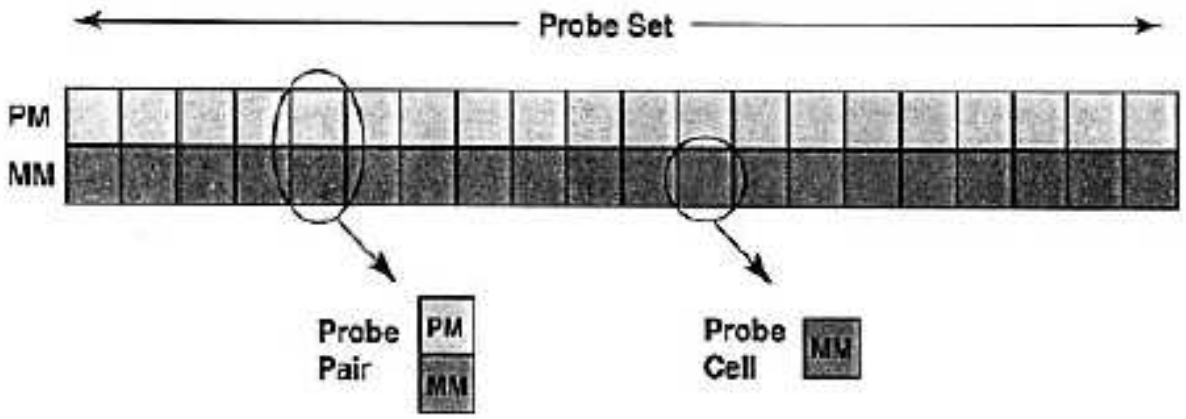
mRNA

Estrategia Expression Tiling

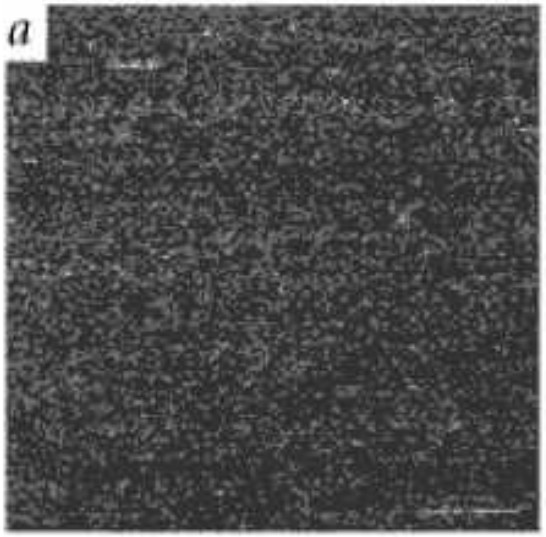


Estrategia Expression Tiling

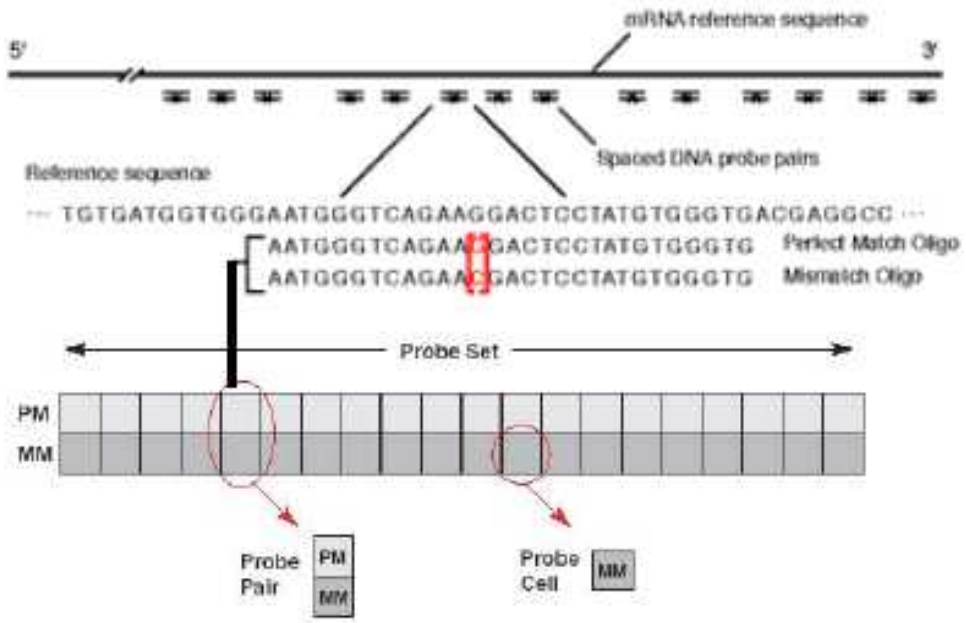
- Probe Set – 16-20 probe pairs that can uniquely identify a transcript
- Probe Pair – One PM cell above MM cell
  - Perfect match probe/mismatch probe
- Each Probe: 25-mer
- Probe cell : 1 million copies



Gene expression monitoring with oligonucleotide arrays.

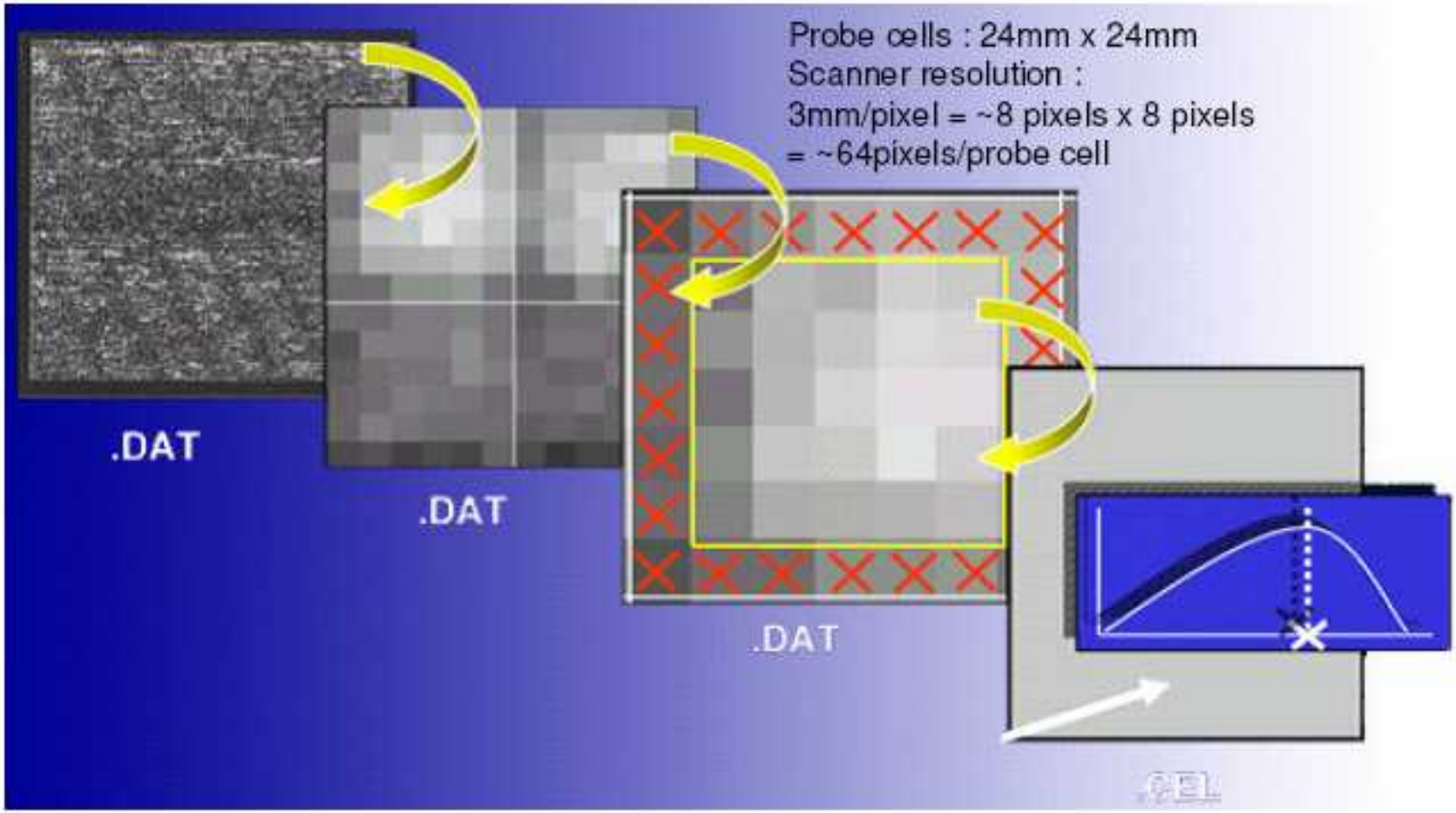


***a*, A single 1.28 X 1.28 cm array** containing 45,000 probes sets.



- **Probe**: a single stranded DNA oligonucleotide complementary to a specific sequence (25 bases long).
- **Probe Cell**: a single square-shaped feature on an array containing one type of probe (50 or 24mm).
- **Perfect Match**: (PM) probes that are designed to be complementary to a reference sequence.
- **Mismatch**: (MM) probes that are designed to be complementary to the reference sequence except for a homomeric base mismatch at the central (13th) position (control for crosshybridization).

The Probe Cell Average Intensity



Exclusion of the bordering pixels of the probe cell.
The remaining pixel intensity distribution is calculated, and the intensity value associated with 75% of the distribution is used as the **Average Intensity** of the probe cell.

GeneChip Operating Software - [Chip Report - 515D.RPT]

File Edit View Run Tools Window Help

Open Save Print Tree Shots Status Scans Expts Fluidics Start Add Resume Stop Analyze Image Help

Cut Copy Paste Find Find next

Data Source: Local

- Experiments
- Image Data
- Cell Intensities
- Analysis Results

GeneChip Software

Experiments

Batch Analysis

Publish

Sample History

Workflow Monitor

Instrument Control

Settings

Report Type: Expression Report
Date: 03:01 PM 04/09/2005

Filename: 515D.CHP
Probe Array Type: U133_2SP
Algorithm: Statistical
Probe Pair Thr: 8
Controls: Antisense

Alpha1: 0.05
Alpha2: 0.065
Tau: 0.015
Noise (RawQ): 2.670
Scale Factor (SF): 1.000
Norm. Factor (NF): 1.000

Background:	Avg: 70.56	Std: 1.47	Min: 67.60	Max: 74.70
Noise:	Avg: 3.26	Std: 0.11	Min: 3.10	Max: 3.70
Corner+	Avg: 183		Count: 32	
Corner-	Avg: 5743		Count: 32	
Control-	Avg: 6517		Count: 9	

The following data represents probe sets that exceed the probe pair threshold and are not called "No Call".

Total Probe Sets:	61958	
Number Present:	17209	28.0%
Number Absent:	43208	70.4%
Number Marginal:	941	1.5%
Average Signal (P):	275.5	
Average Signal (A):	10.3	
Average Signal (M):	42.1	
Average Signal (All):	85.2	

Housekeeping Controls:

Probe Set	Sig(5')	Det(5')	Sig(M')	Det(M')	Sig(3')	Det(3')	Sig(all)	Sig(3'/5')
AFFX-HUMISGF3A/M97935	1.9	A	4.5	A	141.8	P	49.43	73.11
AFFX-HUMRGE/M10098	385.1	P	1572.9	P	1462.3	P	1140.10	3.80
AFFX-HUMGAPDH/M33197	95.3	P	140.8	P	831.0	P	335.71	8.72
AFFX-HSAC07/X00351	12.1	A	3.4	A	306.8	P	107.43	25.25
AFFX-M27830	3621.4	P	3556	P	7.1	A	1327.36	0.00

Spike Controls:

Probe Set	Sig(5')	Det(5')	Sig(M')	Det(M')	Sig(3')	Det(3')	Sig(all)	Sig(3'/5')
AFFX-BioB	140.7	P	168.0	P	124.4	P	144.40	0.88
AFFX-BioC	358.7	P			485.4	P	422.04	1.35
AFFX-BioD	832.6	P			1381.1	P	1096.83	1.63
AFFX-Cre	3810.1	P			4341.1	P	4075.60	1.14

CAP

Lectura Recomendada

- Kepler TB, Crosby L, Morgan KT (2002). **Normalization and analysis of DNA microarray data by self-consistency and local regression.** *Genome Biol.* 3(7):RESEARCH0037.
- Hoffmann R, Seidl T, Dugas M. (2002). **Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis.** *Genome Biol.* 3(7):RESEARCH0033.
- Colantuoni C, Henry G, Zeger S, Pevsner J. (2002). **Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts.** *Biotechniques* 32(6):1316-20.
- Durbin BP, Hardin JS, Hawkins DM, Rocke DM. (2002). **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 18 Suppl 1:S105-10.
- Tran PH, Peiffer DA, Shin Y, Meek LM, Brody JP, Cho KW. **(2002). Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals.** *Nucleic Acids Res.* 30(12):e54.
- Bilban M, Buehler LK, Head S, Desoye G, Quaranta V. **(2002). Normalizing DNA microarray data.** *Curr Issues Mol Biol.* 4(2):57-64.
- Quackenbush, J. (2002). **Microarray data normalization and transformation.** *Nature Genetics Suppl.* 32, **496-501, in '***The Chipping Forecast II'.*
- van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D. and Holstege, F.C. (2003). **Monitoring global messenger RNA changes in externally controlled microarray experiments.** *EMBO Rep* 4, **387-393.**
- Huber W, Von Heydebreck A, Sultmann H, Poustka A, Vingron M. (2002). **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* **Jul;18 Suppl 1:S96-S104.**

The log transformation

- Provides values more easil meaningful from biological

  The idea of the log-ratio (bas

  0: no change

  +1: up by factor of $2^1 = 2$

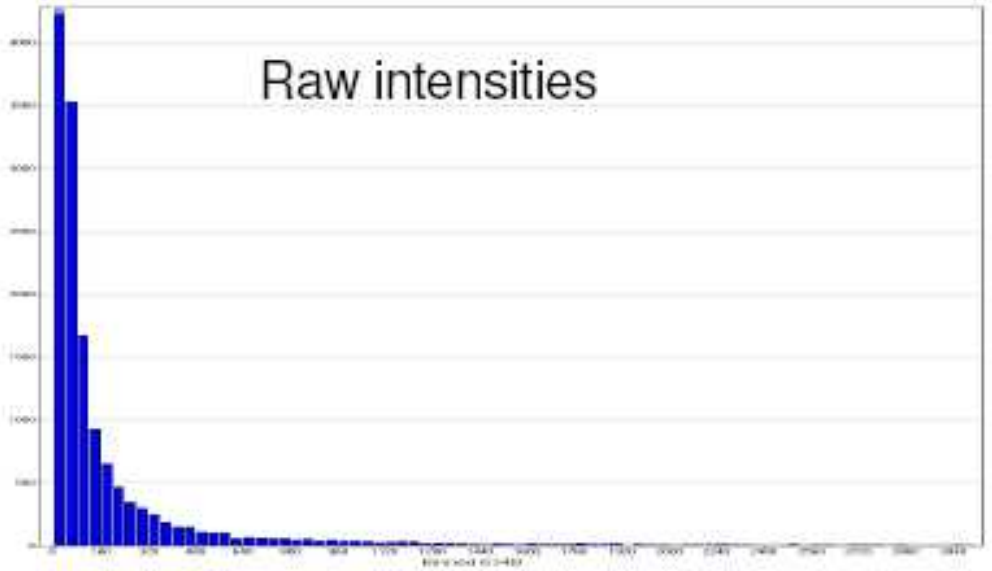
  +2: up by factor of $2^2 = 4$

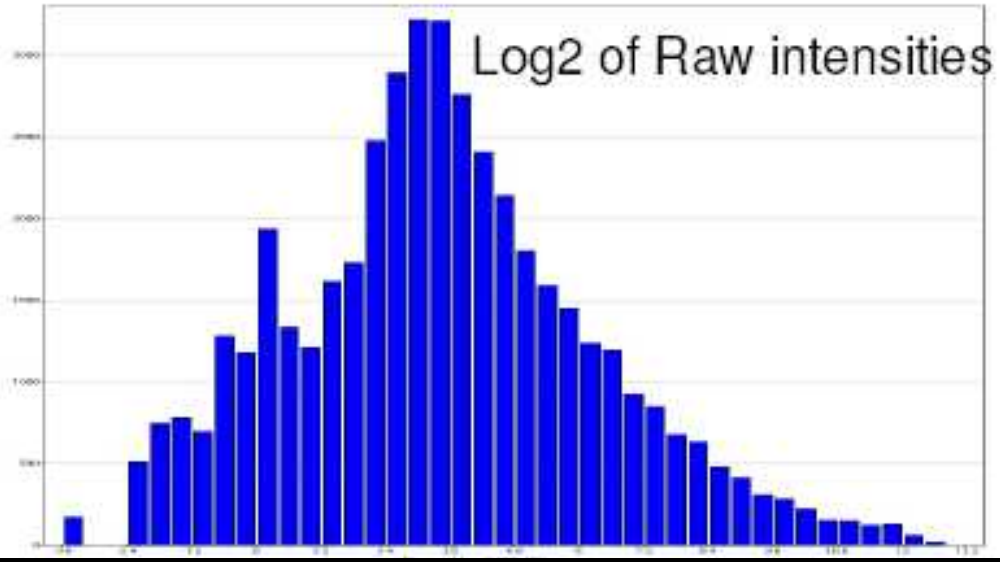
  -1: down by factor of $2^{-1} = 1/2$

  -2: down by factor of $2^{-2} = ¼$

  A unit for measuring changes in e 1000 to 2000 units has a similar b 10000.

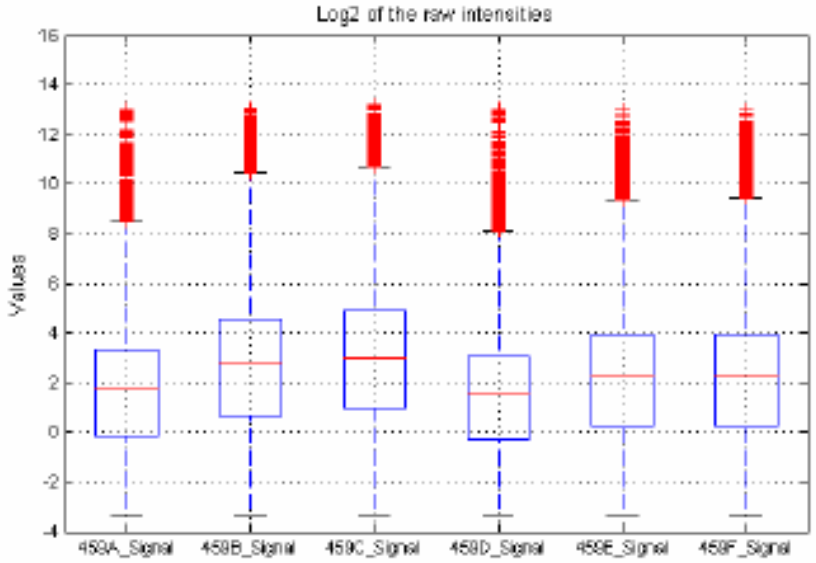
- Makes the shape of the distribution of the values symetrical and almost normal

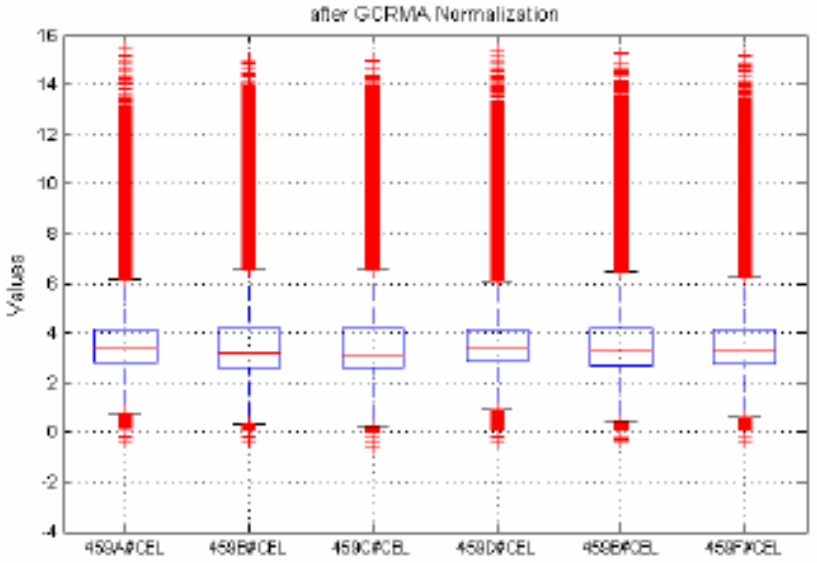




General Pre-Processing techniques

- Combining replicates and eliminating outliers
  - Depends on the number of values, standard deviation (10 replicates with a low variance is better than 3 replicates with a high variance)
  - Eliminating data situated outside $\pm 3\sigma$, then recalculating the parameters and looking for new outliers until no more are detected.
- Array normalization
  - Considered the diversity of the technologies, the best thing to do is be able to compare data obtained with a given technology. The difficulty is related to the fact that various arrays may have various overall intensities: for Affymetrix chips: different overall mean of each individual array, for cDNA arrays: difference between each individual channel (dye) on the same array.
  - The goal for both arrays (oligo and cDNA) is to normalize the data in such a way that values corresponding to individual genes can be compared directly from one array to another.





- Array normalization
  - Considered the diversity of the technologies, the best thing to do is be able to compare data obtained with a given technology. The difficulty is related to the fact that various arrays may have various overall intensities: for Affymetrix chips: different overall mean of each individual array, for cDNA arrays: difference between each individual channel (dye) on the same array.
  - The goal for both arrays (oligo and cDNA) is to normalize the data in such a way that values corresponding to individual genes can be compared directly from one array to another.

Normalization issues specific to cDNA data

- Background correction
  - local background correction, sub-grid background correction, group background correction, background correction using blank spots, using control spots.
- Other spot level pre-processing
  - unreliable spots in the image processing stage: Missing values are deleted or replaced by an estimate.
- Color normalization
  - The 2 dyes used may have different overall efficiencies: a non-linear dye effect with a stronger signal provided by one of the 2 dyes commonly found for cy3/cy5: then you have to perform a flip-dye experiment: using the same mRNA with both dyes.
  - Then the color non-linear distortion can be corrected by

    Curve fitting and correction

    Lowess normalization

    Piece-wise linear normalization

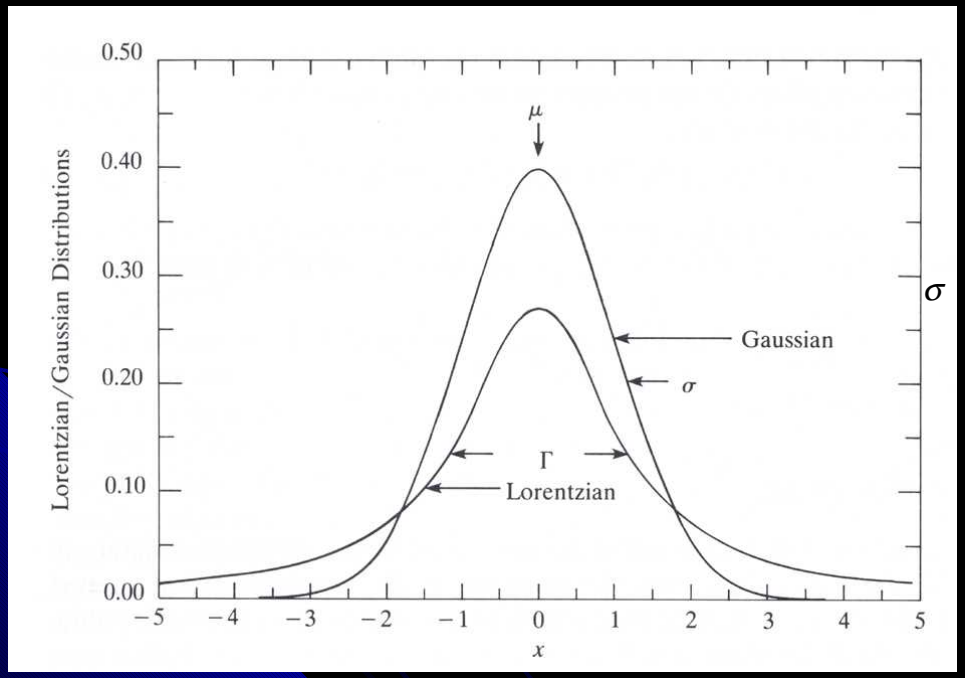
    And many other approaches are discussed in the litterature.

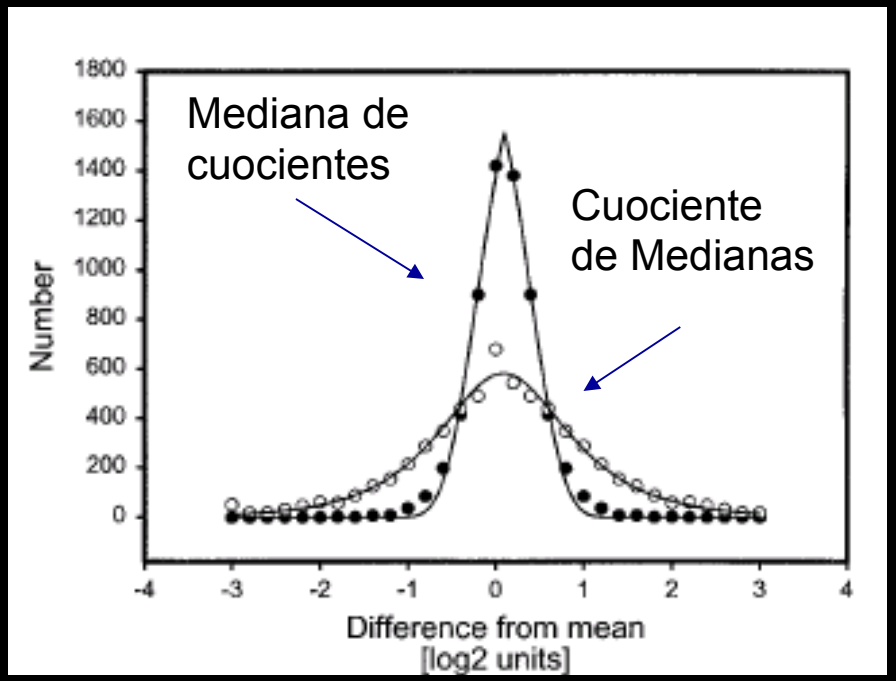
Selección de Genes Expresados Diferencialmente

- Cambio en niveles de expresión (Nº veces)
- Outlier
- Prueba de hipótesis
- ANOVA
- Estimación de la probabilidad máxima basada en modelo

Qué has estado midiendo?
Una nota de precaución

Comparación distribuciones de Gauss y Lorentz normalizados





(PNAS October 1, 2002 vol. 99 no. 20 pp.12975–

Diagramas de Dispersión

- Útil para representar valores de expresión de genes a partir de dos experimentos de microarray (control, experimento)
- Cada punto corresponde a un valor de la expresión del gen
- La mayoría de los puntos caen a lo largo de una línea
- Los outliers representan genes para sobre expresados o sub expresados

Methods for selecting differentially regulated genes

- Criteria:
  - Distinction of 4 categories (in a binary decision situation as changed/unchanged)
    - Truly changed that are reported as changed: **true positives**
    - Unchanged that are reported as changed: **false positives**
    - Truly changed that are reported as unchanged: **false negatives**
    - Truly unchanged that are reported as such: **true negatives**
  - Define the 4 quantitative criteria
    - Positive predicted value PPV=TP/(TP+FP)
    - Negative predicted value NPV=TN/(TN+FN)
    - **Specificity**=TN/(TN+FP)
    - **Sensitivity**=TP/(TP+FN)
      - accuracy=(TP+TN)/N

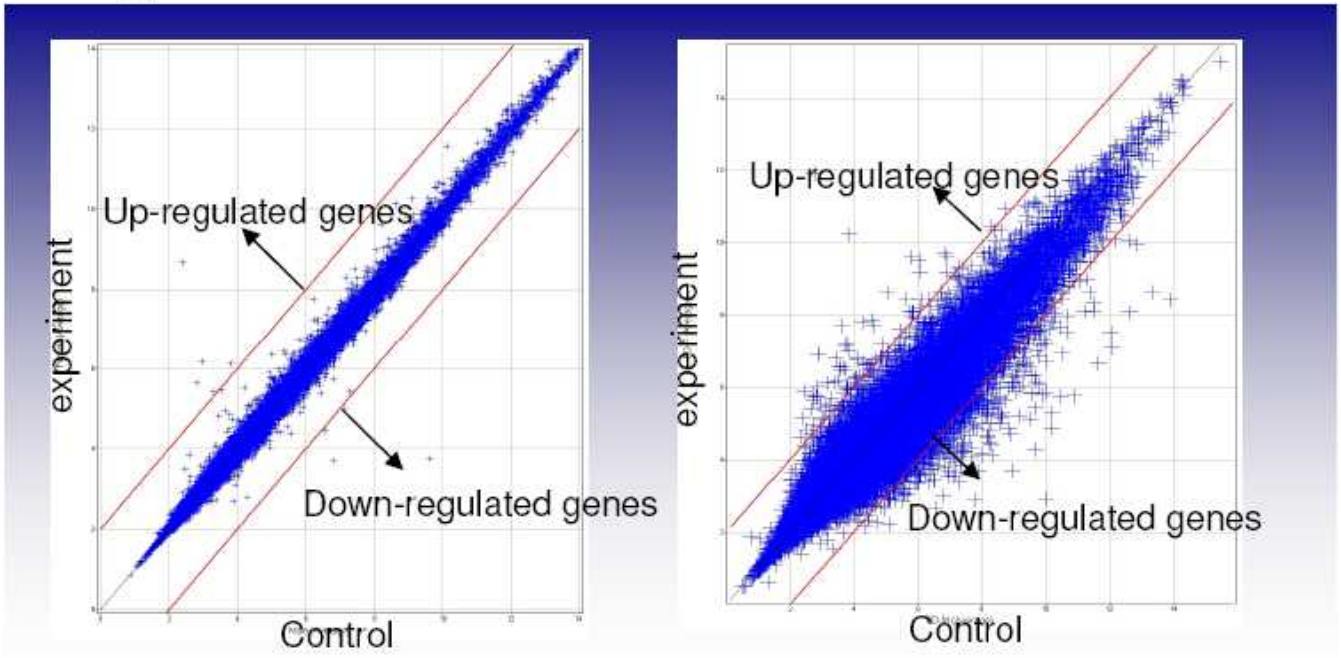
Methods for selecting differentially regulated genes

- Level of intensity: Whatever method you are using, consider the lower values as having the higher noise probability. If your normalization method didn't reduce the noise at the lower level, choose the statistical method in function of its efficacity at the lower intensities.

Methods for selecting differentially regulated genes

•Fold change : For cDNA: ratio between the 2 dyes expression levels for each gene on each array, for Affymetrix: ratio between the 2 conditions experiment/control for each gene.

–Simple and intuitive method but may often be inappropriate as the fold threshold is chosen arbitrarily: depending on the conditions of the experiments, a fold change of 2 can be too high (low sensitivity) or can be too low (low specificity)



Methods for selecting differentially regulated genes

•Fold change : For cDNA: ratio between the 2 dyes expression levels for each gene on each array, for Affymetrix: ratio between the 2 conditions experiment/control for each gene.

–Simple and intuitive method but may often be inappropriate as the fold threshold is chosen arbitrarily: depending on the conditions of the experiments, a fold change of 2 can be too high (low sensitivity) or can be too low (low specificity)

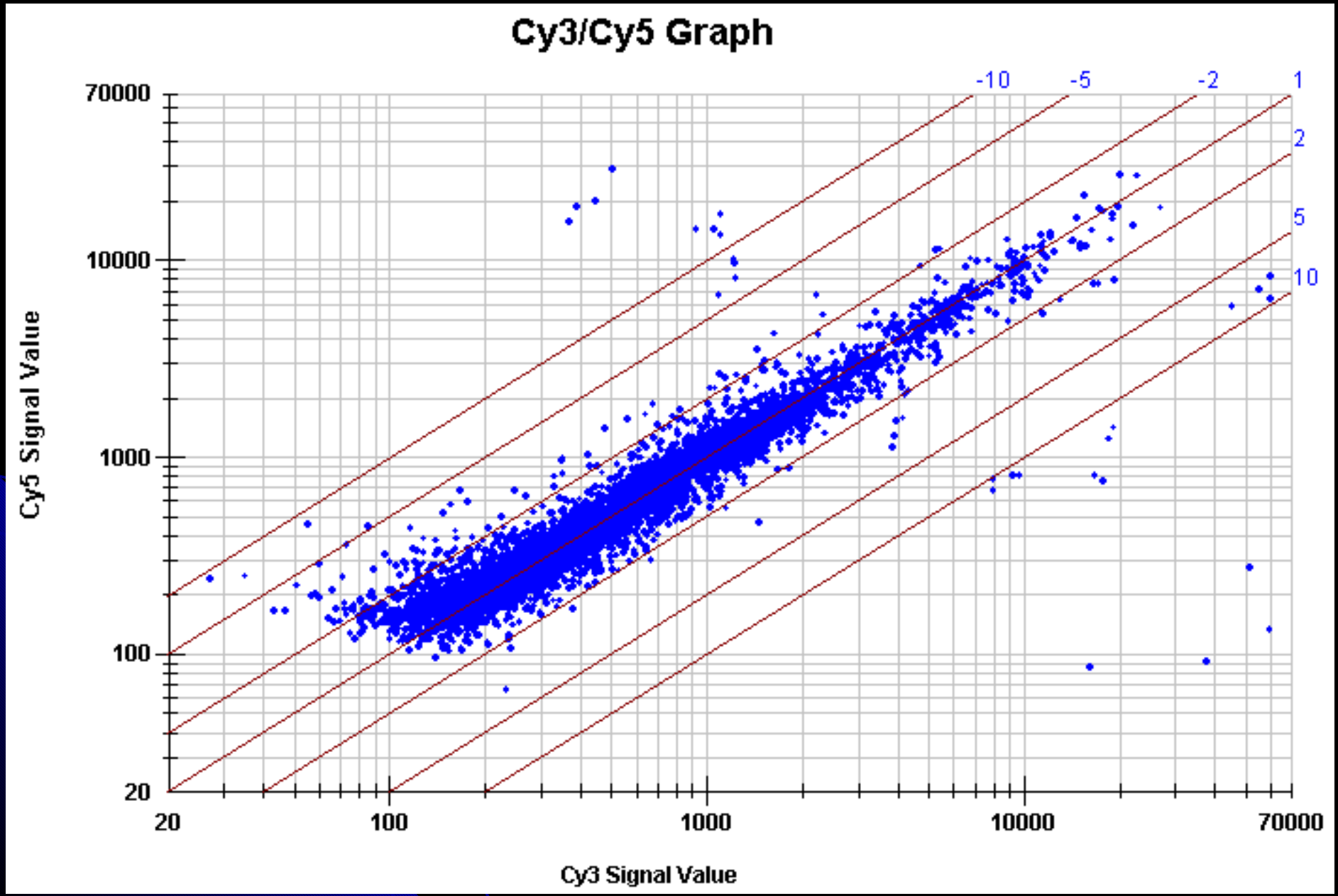
- Unusual ratio:
  - Uses the standard deviation of the ratio distribution as the unity. Generally one chooses $+2\sigma$ obtained by a Z transformation of the log of the ratio (subtract the mean and divide by the standard deviation): detects the 5% most regulated genes (even if there are more genes regulated than that).

Methods for selecting differentially regulated genes

- Hypothesis testing, corrections for multiple comparisons and resampling
  - Univariate statistical tests: t-test gives you a p-value which is the probability that the expression difference occurs by chance (error Type I).
  - Since a lot of genes are considered at the same time, with very different levels of expression, many different "corrections" and "approaches" can be used depending on the experimental design:
    - ✓Bonferroni, Sidak,
    - ✓the Holm step-down-group of methods, False Discovery Rate (FDR), permutation and significance analysis of microarray (PAM and SAM).

Choose the statistical method adapted with your normalization processes…

Diagrama de Dispersión



Estadística Inferencial

- La estadística deductiva se utiliza para hacer inferencias sobre una población de una muestra.
- La prueba de hipótesis es una forma común de estadística deductiva. Una hipótesis nula se indica, por ejemplo: “No hay diferencia en la intensidad de la señal para los niveles de expresión del gen en la cepa wild type y mutante.” La hipótesis alternativa es que hay una diferencia.
- Utilizamos una prueba estadística para decidir si aceptar o rechazar la hipótesis nula. Para muchos usos, fijamos el nivel de significancia a $p < 0.05$

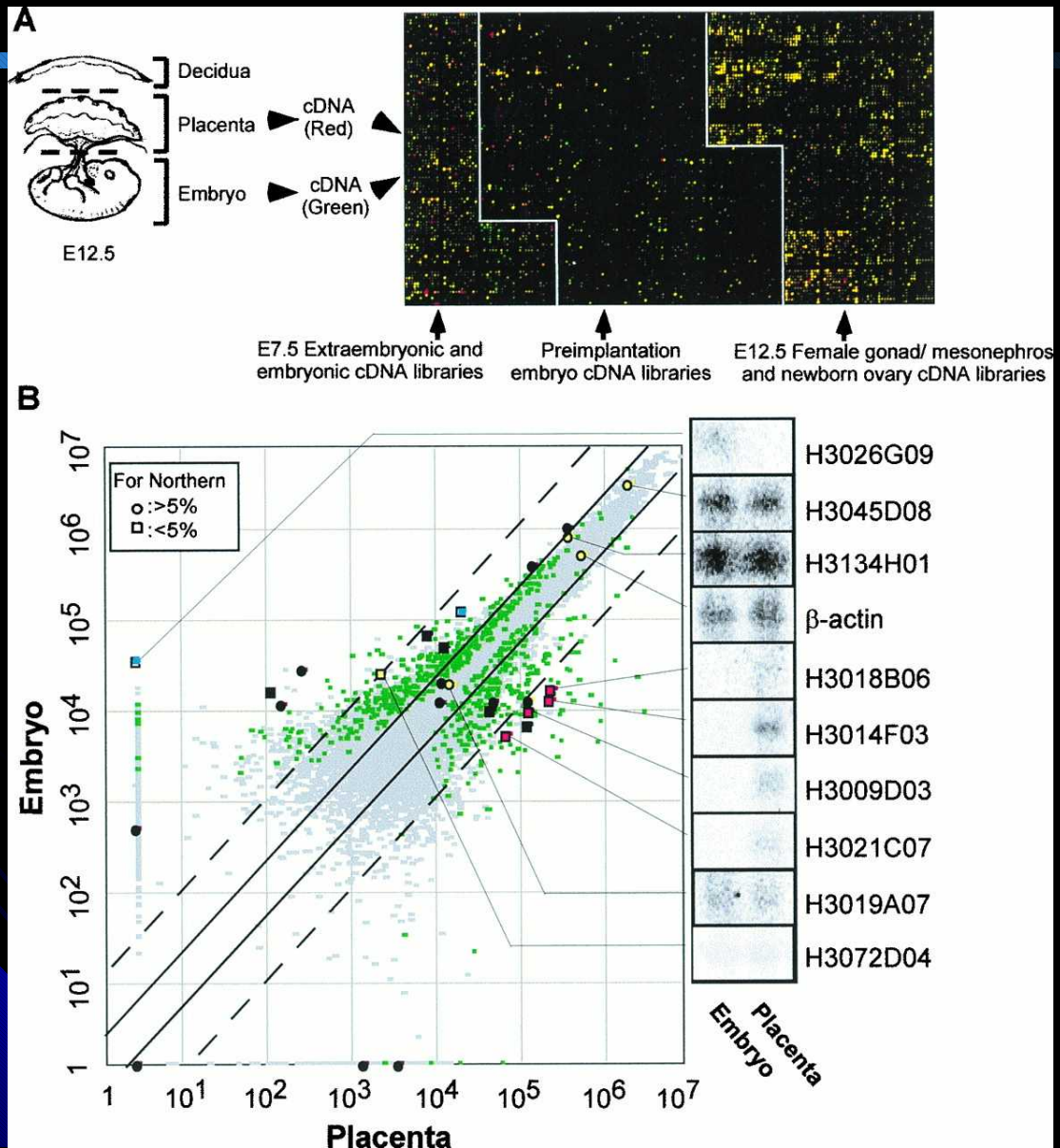
Estadística Estándar de t-test

- **Un t-test es una estadística de prueba común para determinar la diferencia en valores medios entre dos grupos.**

$$t = \frac{x_1 - x_2}{s} = \frac{\text{Diferencia entre valores promedio}}{\text{Variabilidad}}$$

- **Preguntas:**
  - **¿Es el tamaño de muestra (n) adecuado?**
  - **¿Los datos se distribuyen normalmente?**
  - **¿ Se conoce la varianza de los datos?**
  - **¿Es la varianza igual en los dos grupos?**
  - **¿Es apropiado fijar el nivel de significancia a $p < 0.05$?**

Revelación Estadística Sobre los Datos del Array

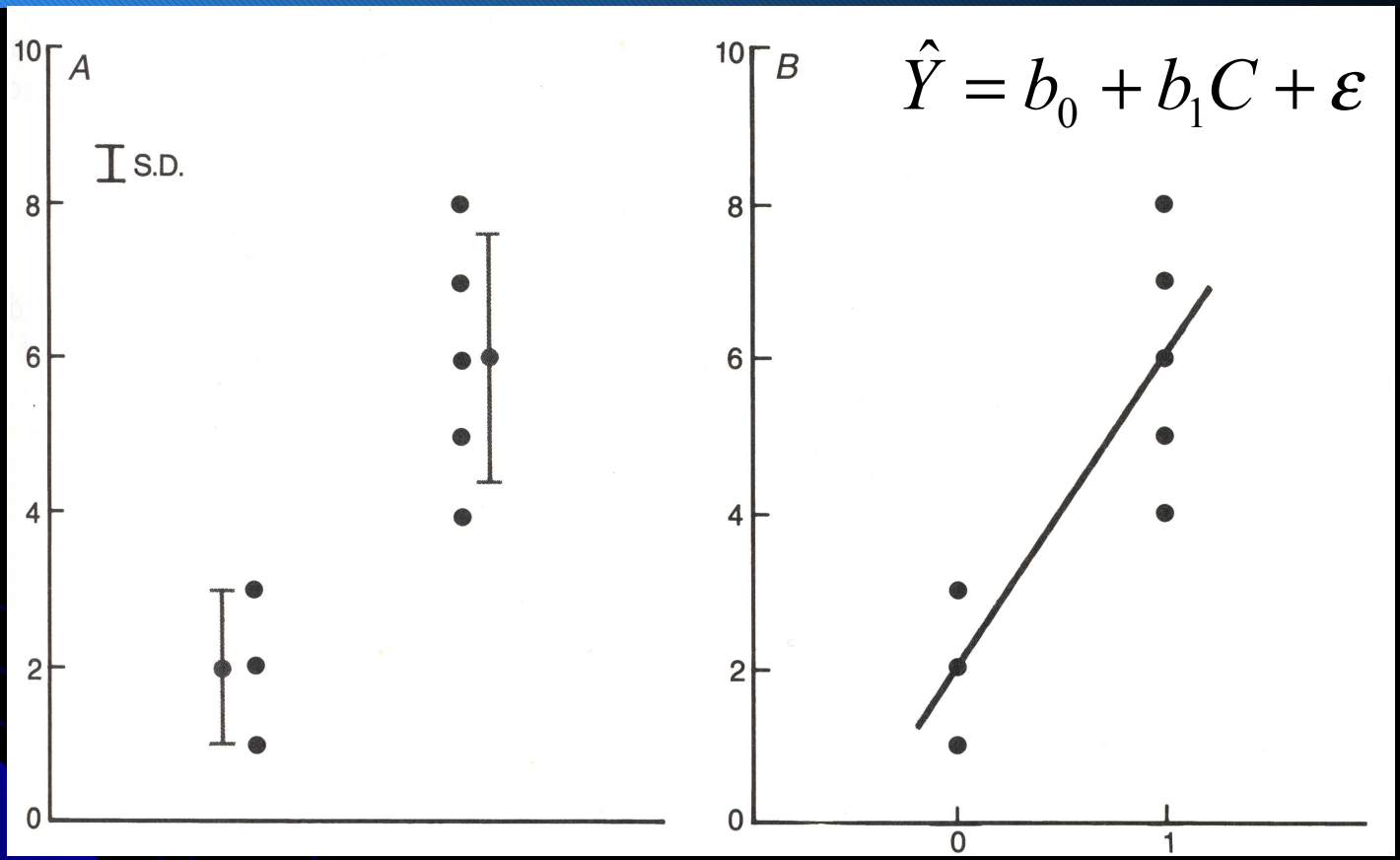


PNAS, 2000 Aug 1;97(16):9127-32.

Una Variedad de Soluciones Deductivas

Comparación	Prueba paramétrica	No paramétrico
Compare dos grupos desapareados	T-test desapareada	Prueba de Mann-Whitney
Compare dos grupos apareados	T-test apareada	Prueba de Wilcoxon
Compare 3 o más grupos	ANOVA	

t-test vs. Representación de la Regresión



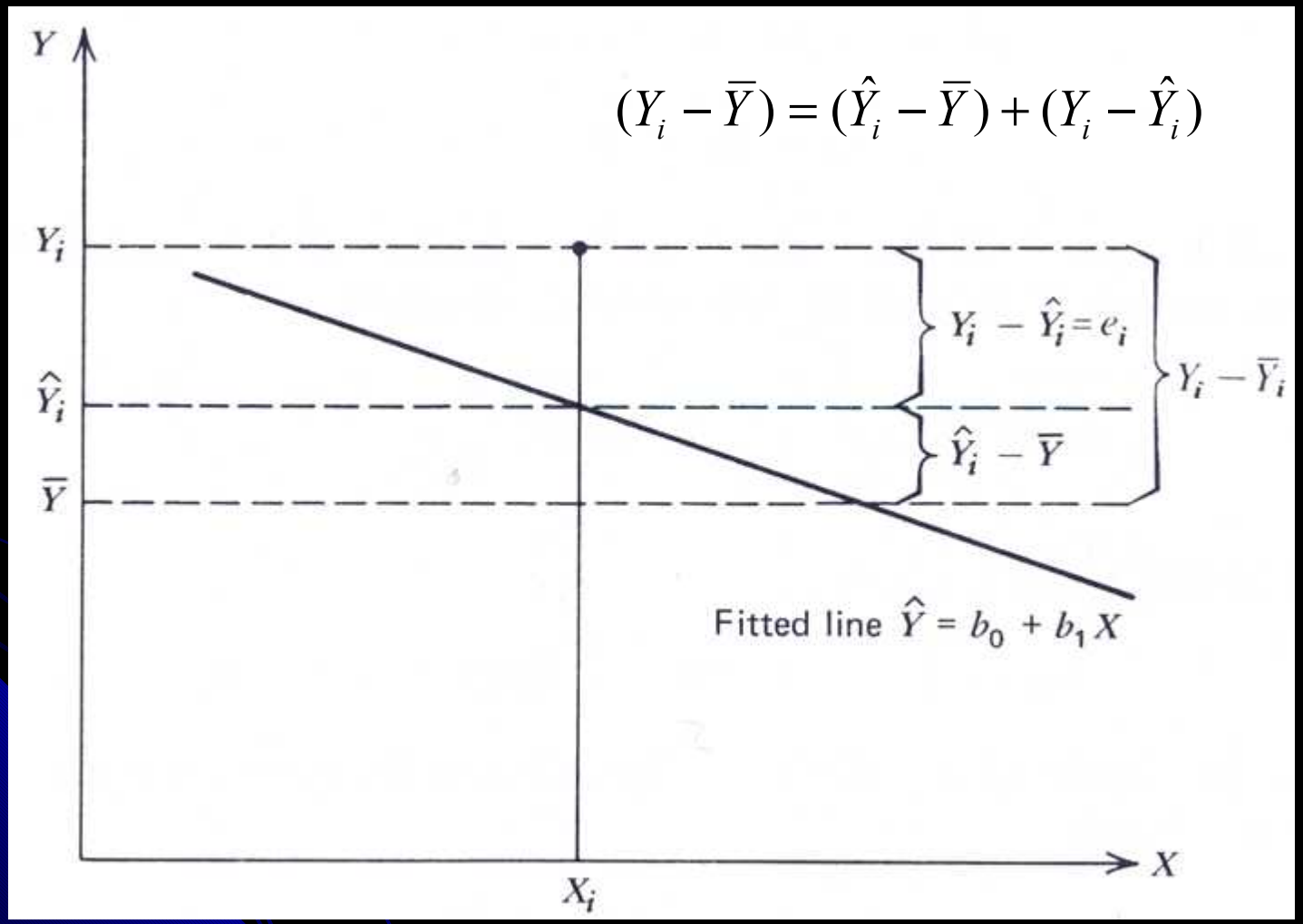
Dependent variable

A B

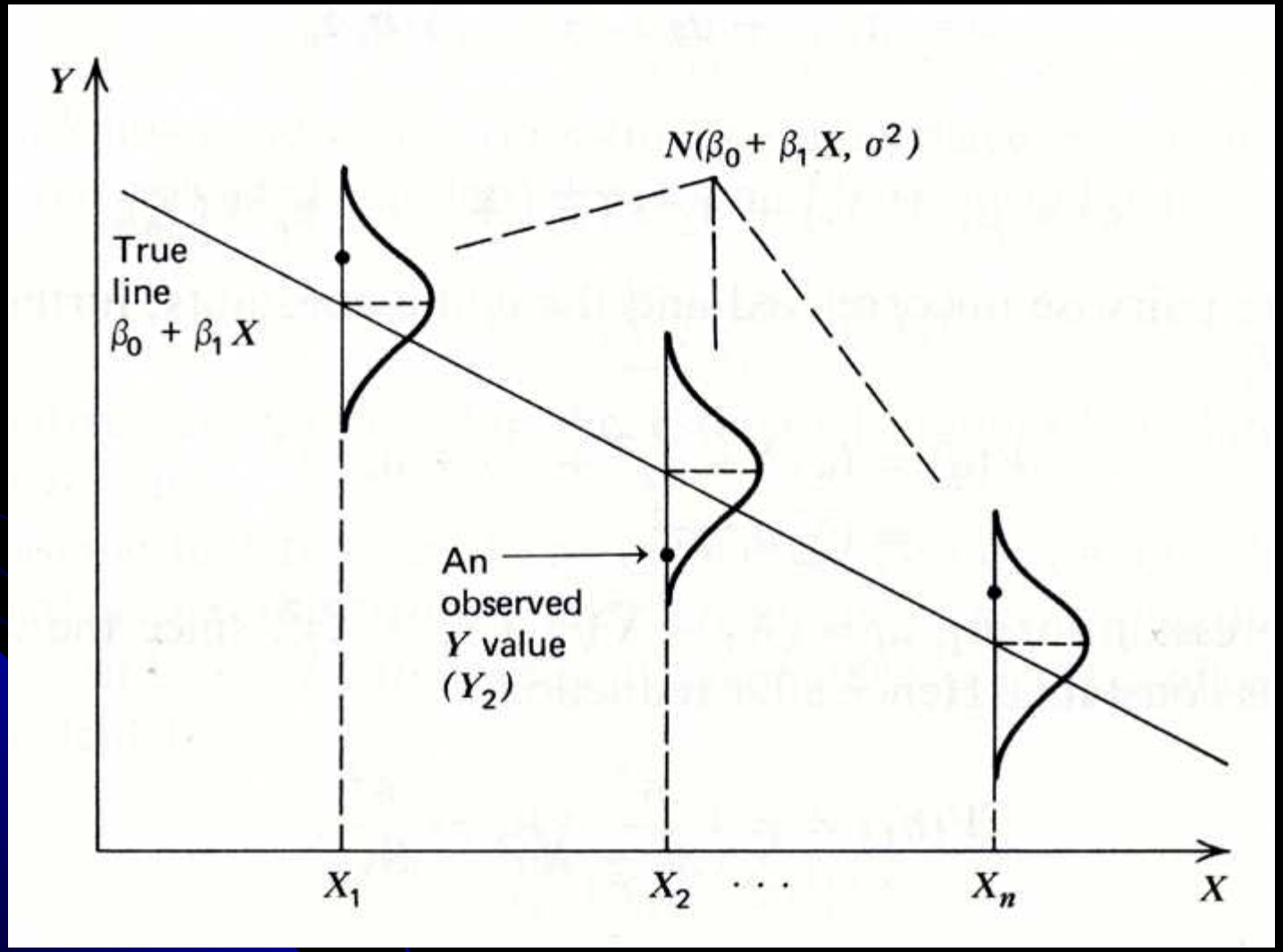
Categorical independent variables

Dummy variable, (C)

Representación Geométrica de la Regresión Lineal



La observación de la respuesta se asume proviene de una distribución normal centrada verticalmente en el nivel implicado por el modelo asumido



Análisis de la Varianza (ANOVA) para Microarrays

Statistical Model:

Spots!

$$y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{igr} + (DG)_{kg} + \varepsilon_{ijkg}$$

Array Dye Variety Gene Variety-by-Gene effects Gene-specific dye effects

We assume that there is independent, random error ε_{ijkg} with mean 0.

Quantities of interest are expression levels of gene specifically attributable to different varieties:

$$(VG)_{kg} - (VG)_{k'g}$$

Kerr and Churchill

Objetivos de ANOVA

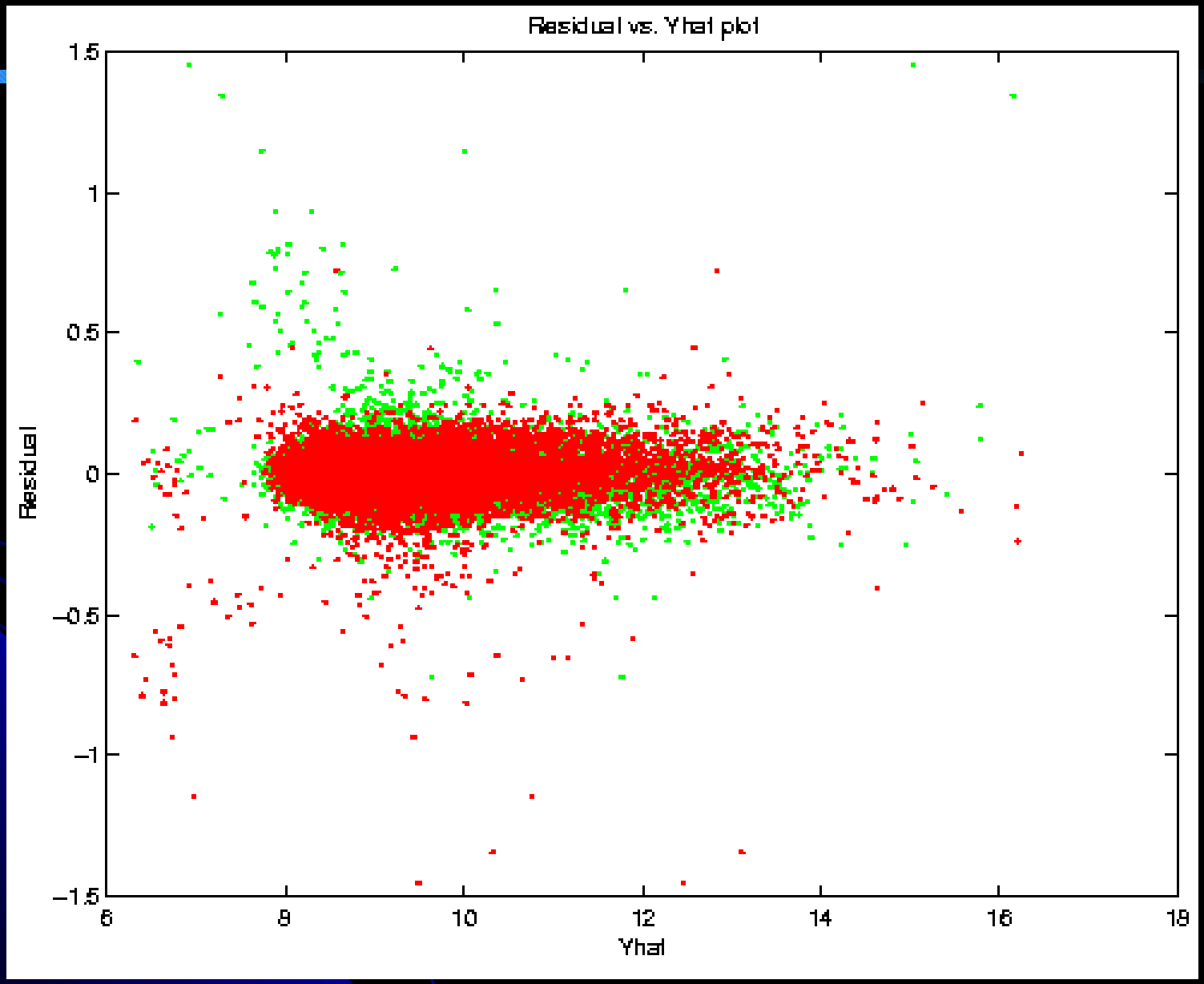
- Un objetivo es obtener estimaciones no sesgadas de los efectos de interes
- Un segundo objetivo es tener barras de error para esas estimaciones. Esto permite decidir si las diferencias observadas son significativas

Clave en el Análisis Microarray: Inferencia

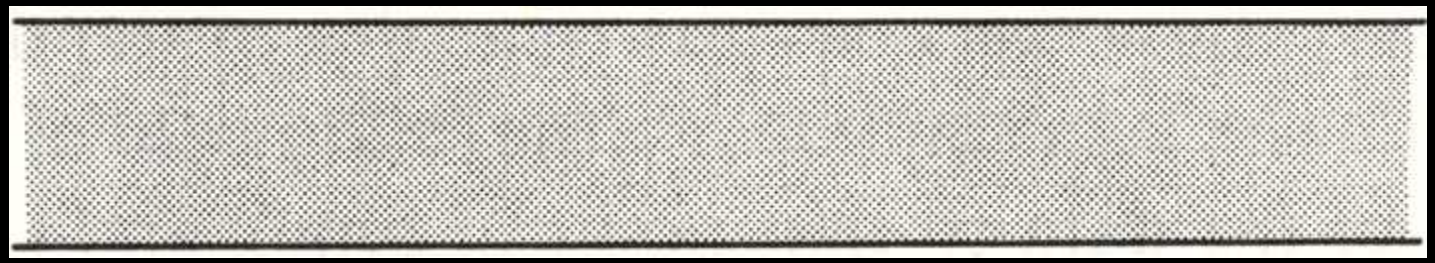
- Inferencia estadísitca: Como identificamos genes expresados diferencialmente entre muestras?

- Mediante la diferencia entre las caracteristicas de ambas muestras. Como agregar barras de error a estas estimaciones para determinar cuales son significativas?

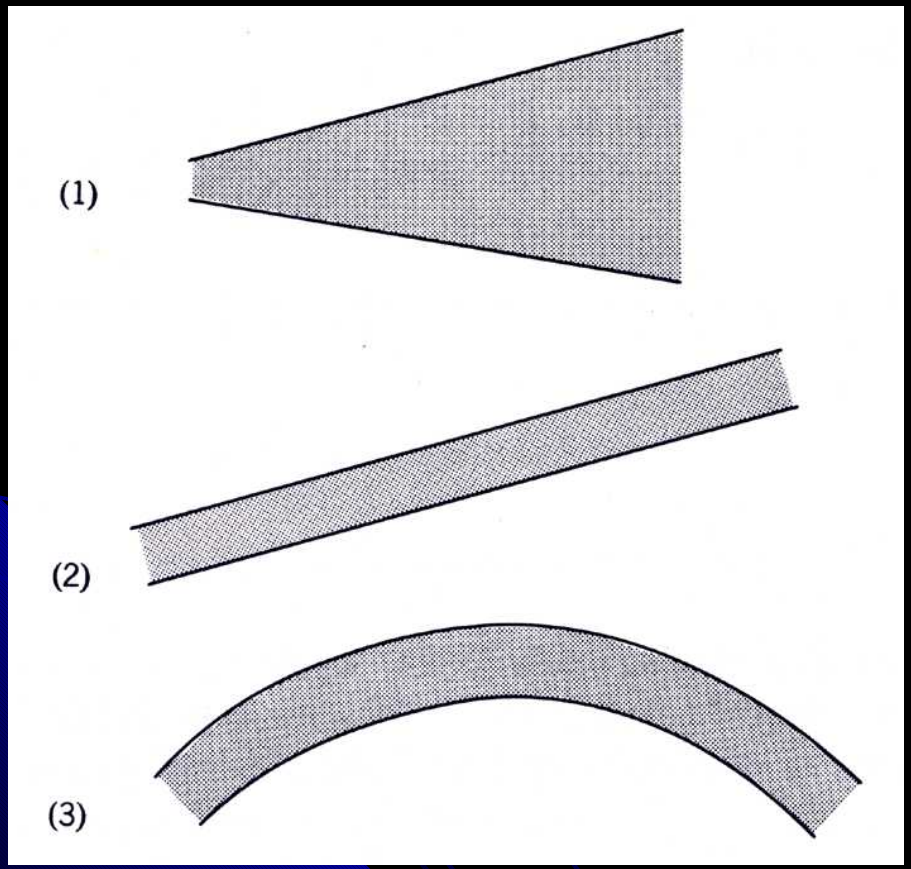
Diagrama de Residuales



Un Diagrama Residual Satisfactorio



Ejemplos de Comportamiento Residual Insatisfactorio



1. Varianza no constante; se necesita mínimos cuadrados o una transformación de la observación
2. Error en análisis; podía ser causado por omitir el término *b0*
3. Modelo inadecuado requiere términos adicionales

Intervalos de Confianza

Produce bootstrap simulated datasets

$$y^*_{ijkgr} = \hat{\mu} + \hat{A}_i + \hat{D}_j + \widehat{AD}_{ij} + \hat{G}_g + \widehat{(AG)}_{igr} + \widehat{(VG)}_{kg} + \widehat{(DG)}_{jg} + \varepsilon^*_{ijkgr}$$

The ε^*_{ijkgr} are drawn independently from residuals of the original model fit.

Obtain bootstrap estimates $\widehat{(VG)}^*_{1g} - \widehat{(VG)}^*_{2g}$ for each bootstrap dataset.

To form 99% confidence intervals for $(VG)_{1g} - (VG)_{2g}$, take the middle 99% of bootstrap estimates.

Por qué hacer las pruebas de *F* y *t-test*?

Aunque estas dos pruebas proveen información equivalente en regresión lineal simple con una variable independiente proporcionan información diferente en regresión múltiple:

- ***t-test de los* coeficientes individuales prueban si cada variable independiente, considerada una a la vez, contribuye a predecir la variable dependiente**
- **Prueba de *F* para el ajuste general prueba si todas las variables independientes, tomadas juntas, contribuyen a predecir la variable dependiente**

Prueba de Hipótesis

- Plantear el problema: Se sobreexpresa el gen?
- Hipótesis nula y alternativa

$$H_0 : \overline{X} = \mu \qquad H_a : \overline{X} \neq \mu$$

- Nivel de significancia: 5%?
- Encontrar y calcular estadísticos apropiados
- Determinar *p*-value del test estadistico
- Comparar *p*-value con nivel de significancia
- Rechazar o no la hipótesis nula

Resultados de Prueba de Hipótesis

Reported by the test	True (but unknown) situation	
	H_0 is true	H_0 is false
H_0 was not rejected	true negatives (correct decision) $1 - \alpha$	false negatives (Type II error) β
H_0 was rejected	false positives (Type I error) α	true positives (correct decision) $1 - \beta$

Power = 1 - β

Criterios para el Éxito Experimental

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

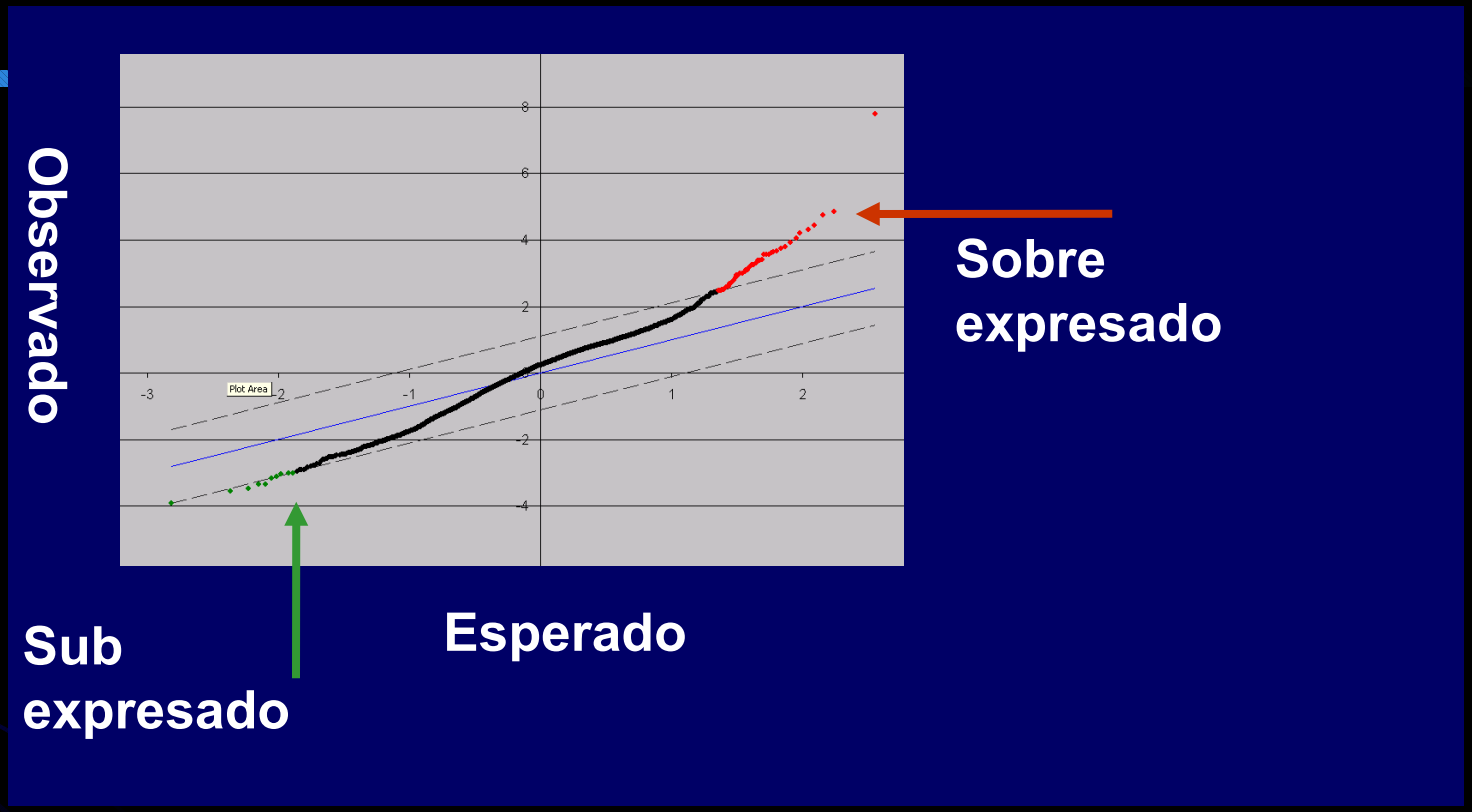
$$Accuracy = \frac{TP + TN}{N}$$

	True		
Reported	changed	unchanged	
changed	TP	FP	Positive predicted value $\frac{TP}{TP+FP}$
unchanged	FN	TN	Negative predicted value $\frac{TN}{TN+FN}$
	Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{TN+FP}$	

False Discovery Rate

PPV – positive predictive value
NPV –negative predictive value

Ejemplo del Diagrama SAM



$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

Relative difference in gene expression with gene-specific scatter:

$$s(i) = \sqrt{a\{\sum_m (x_m(i) - \bar{x}_I(i))^2 + \sum_n (x_n(i) - \bar{x}_U(i))^2}$$

Bayes's theorem

"An essay towards solving a problem in the doctrine of chances", By Rev. Thomas Bayes, 1763

Se han observado dos eventos, A y B

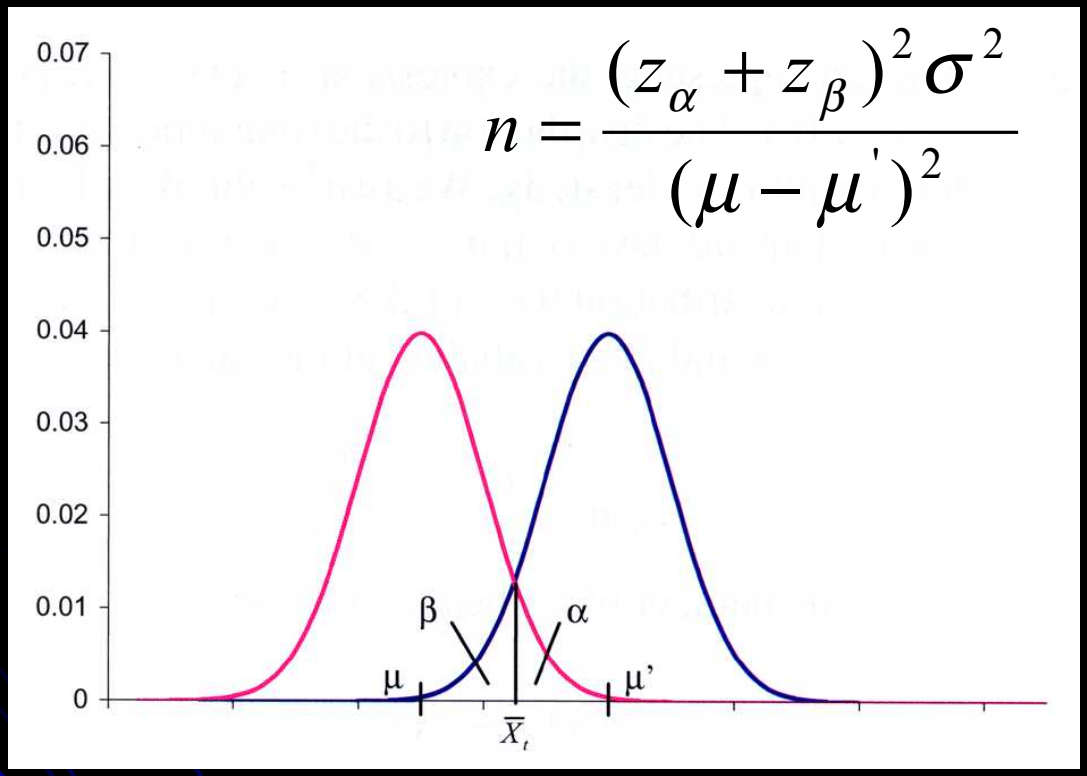
1. **A ocurre primero: B después de A** →
2. **B ocurre primero: A después de B** →

La probabilidad de que B ocurra, dado que A ya ocurrió es igual a la probabilidad de que A ocurra, dado que B ocurrió, multiplicado por la probabilidad de que B ocurra y dividido por la probabilidad de que A ocurra

Bayes y Test de Hipótesis

Cuántas Veces Tengo que Hacer el Experimento?

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu - \mu')^2}$$



Expression level/ratio for one gene

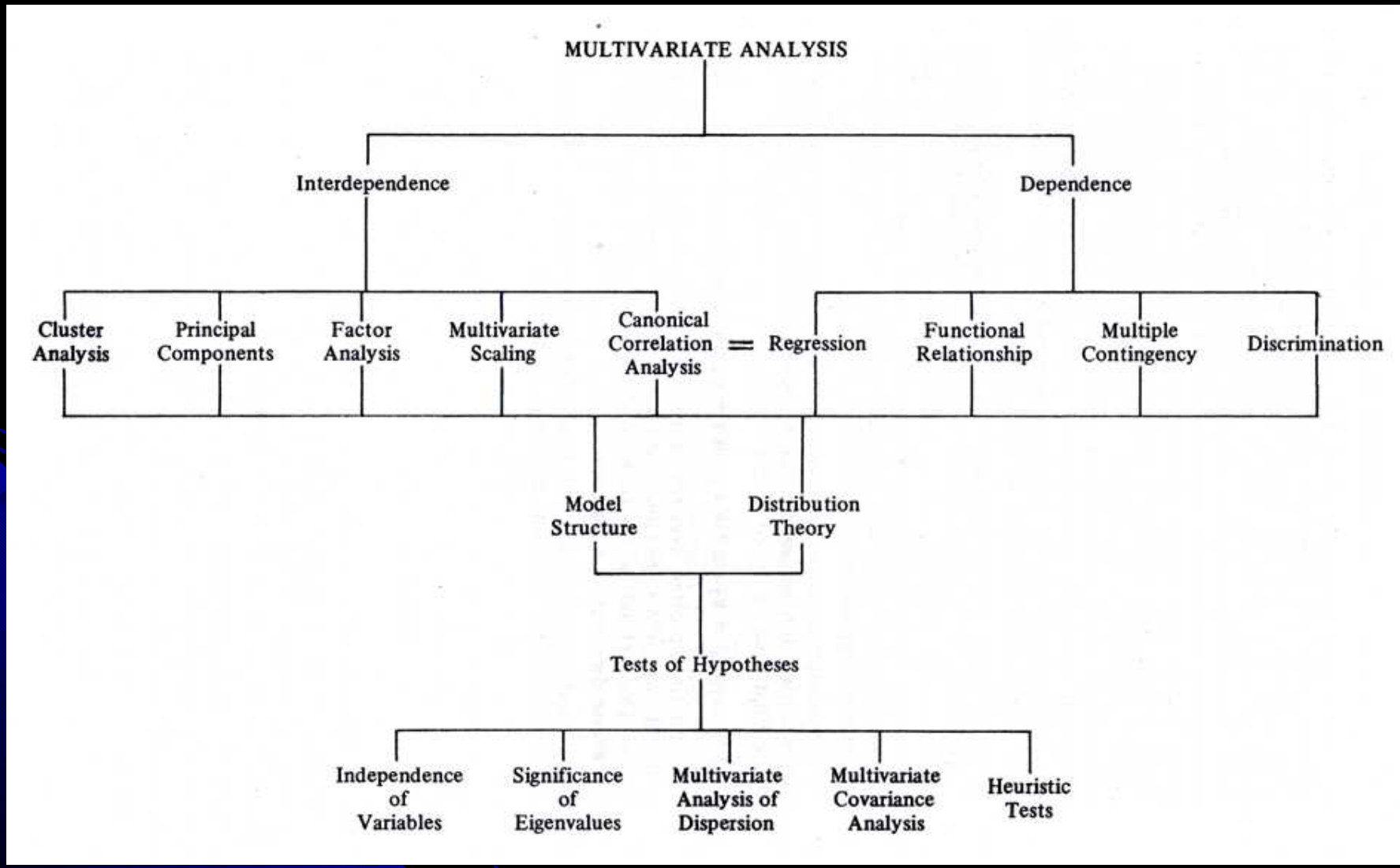
Puntos a Favor del Análisis Estadístico

- Microarrays estiman la expresión relativa del gen. Cualquier estimación requiere una barra del error
- El diseño experimental define el contenido de información de los datos, el modo de análisis, y la calidad de los resultados
- ANOVA es una herramienta natural para estudiar datos con fuentes múltiples de variación. Proporciona un marco global para el análisis de datos microarray, incluyendo normalización, control de calidad, y la detección de artefactos
- Los datos de microarray, como cualquier otro dato biológico, son ruidosos. Las herramientas analíticas más complejas deben considerar eso

Ventajas de Métodos Multivariados

- Están más cerca a cómo pensamos de los datos
- Permiten una visualización y una interpretación más fáciles de diseños experimentales complejos
- Permiten para analizar más datos simultaneamente (las pruebas son más sensibles y más poderosas)
- Los modelos de la regresión múltiple pueden dar más información respecto a relaciones estructurales subyacentes
- El análisis se centra en relaciones entre variables y tratamientos más bien que puntos individuales
- Permiten un manejo más fácil de diseños experimentales desequilibrados y de datos faltantes que el análisis de la varianza tradicional

Componentes del Análisis Multivariado (MVA)



Estadística Descriptiva

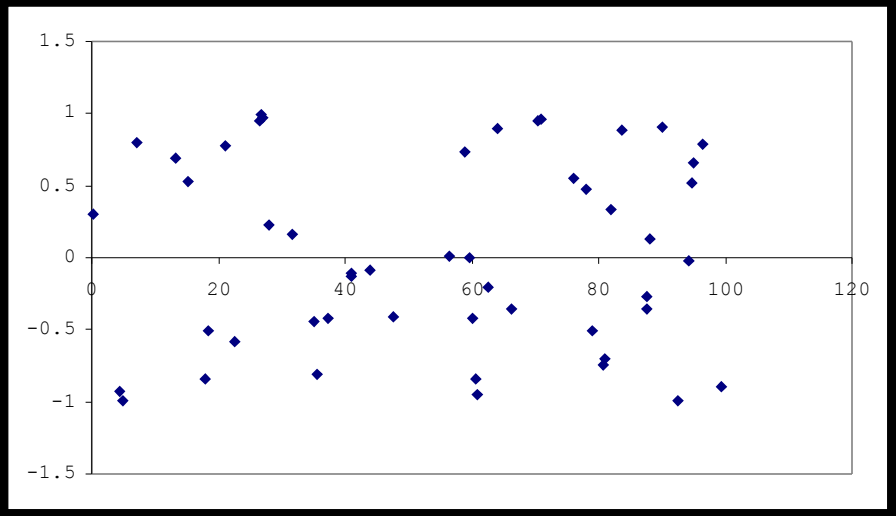
- Los datos de Microarray son altamente dimensionales: muchos millares de medidas hechas de un pequeño número de muestras
- La estadística (exploratoria) descriptiva le ayuda a encontrar patrones significativos en los datos

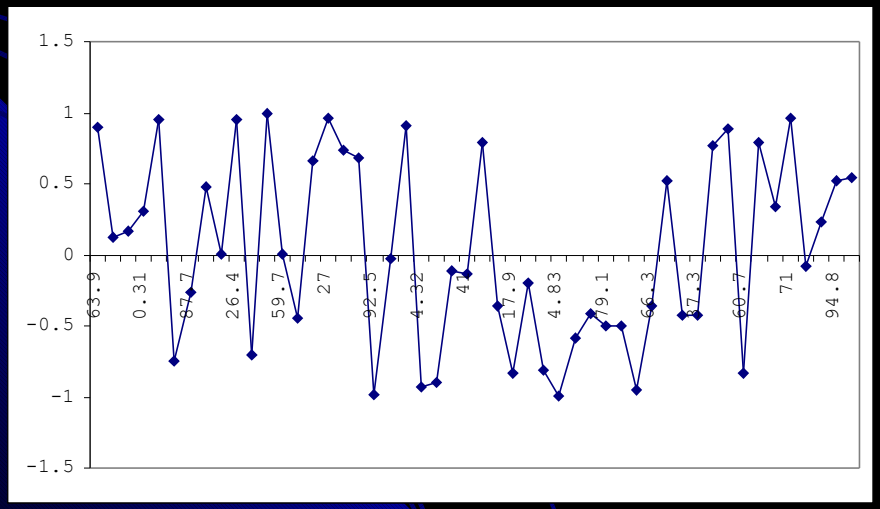
Por qué una Dimensionalidad más Alta de Vectores Correlacionados es Mejor?

	r_o									
N	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
3	100	94	87	81	74	67	59	51	41	29	0
6	100	85	70	56	43	31	21	12	6	1	0
10	100	78	58	40	25	14	7	2	0.5		0
20	100	67	40	20	8	2	0.5	0.1			0
50	100	49	16	3	0.4						0

La probabilidad $Prob_N(|r| \geq r_0)$ de que N medidas de dos variables no relacionadas produzcan un coeficiente de correlación $|r| \geq r_0$. Los valores son probabilidades porcentuales. Blanco indica valores bajo 0.05%

Correlación y Dependencia





Correlación Lineal = 0.03

Dependencia No-Lineal:
$Y = sin(x)$

Variables Independientes

Correlación Baja

PERO

Correlación Baja

Variables Independientes

Qué Técnica es Correcta?

- Clustering jerárquico
  - Single, Average, Complete, Centroid linkage, etc.
- Self Organizing Maps
- K-means clustering
- Otros métodos complejos

Usando Gene Ontology para Determinar Clusters

- Muchos análisis microarray dan lugar a una lista de genes interesantes
- Se puede componer una historia sobre cualquier lista al azar de genes
- Mirar todas las anotaciones de GO para los genes en una lista, y ver si el número de las anotaciones para cualesquiera nodo GO es significativo

Las Categorías de GO
(The Gene Ontology)

- Biological Process = goal or objective (Why)

(e.g. DNA replication, Cell Cycle Control, Cell adhesion)

- Molecular Function = elemental activity/task (What)

(e.g. Transcription factor, polymerase, protein kinase)

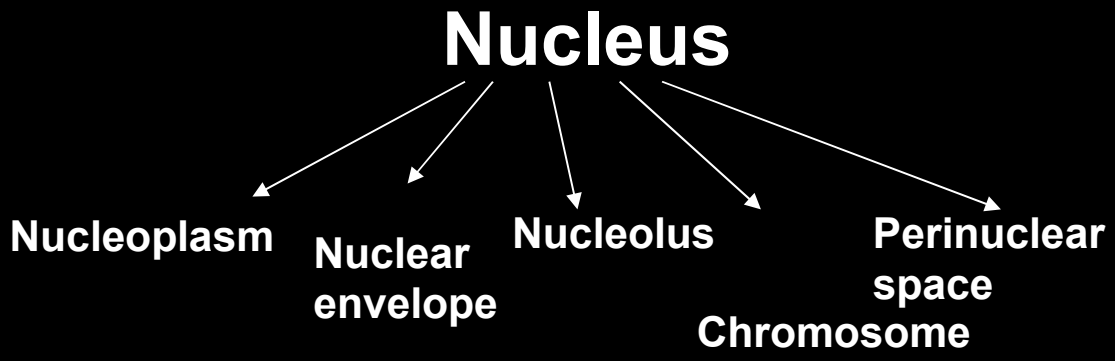
- Cellular Component = location or complex (Where)

(e.g. pre-replication complex, kinetochore, membrane)

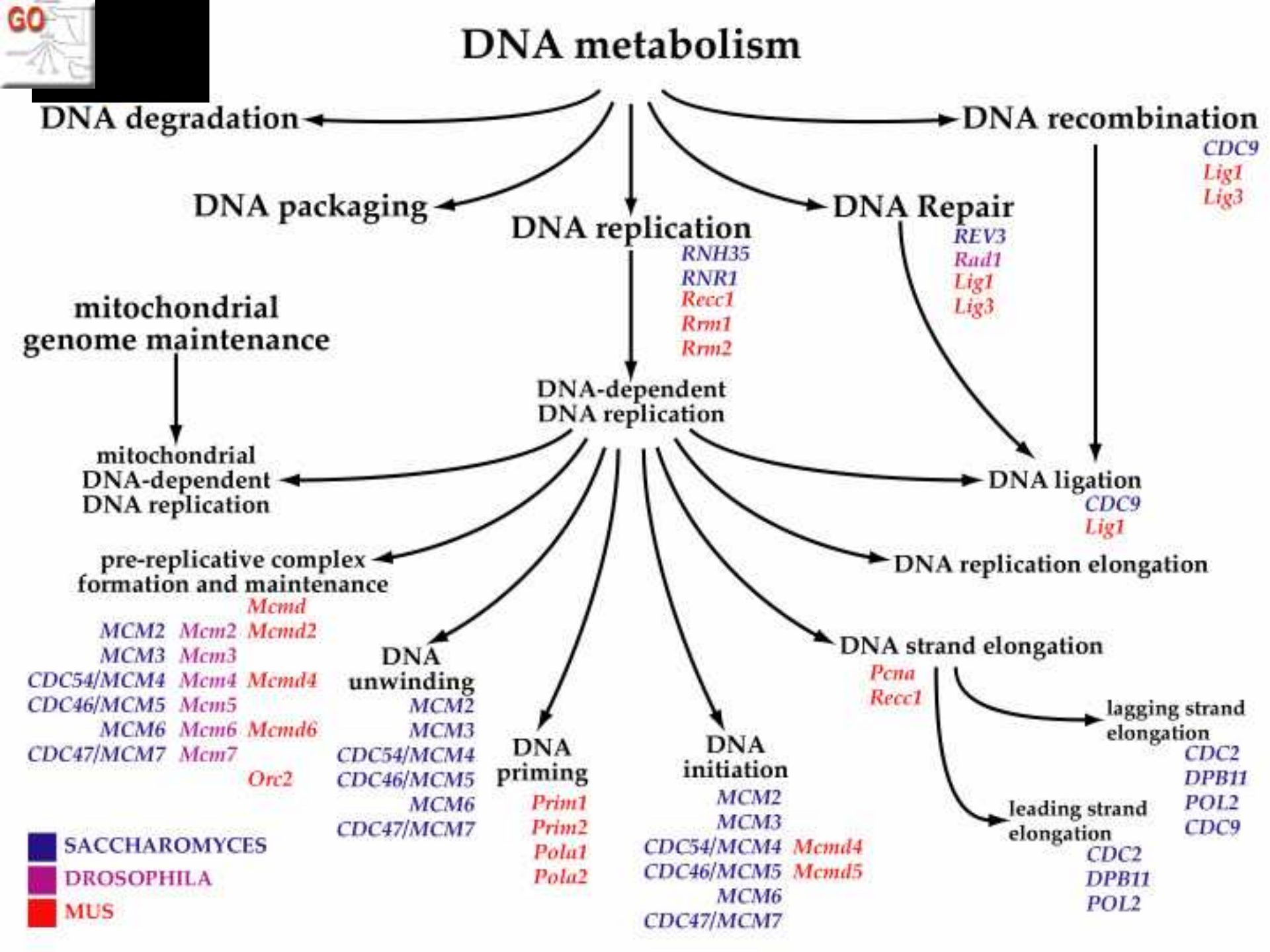
Cada categoría es un vocabulario estructurado, controlado

Relaciones Padre-Hijo

Un hijo es un subconjunto de elementos de un padre



El término del componente de la célula. *El núcleo* tiene 5 hijos


DNA metabolism
DNA degradation
DNA packaging
DNA replication
RNH35
RNR1
Recc1
Rrm1
Rrm2
DNA Repair
REV3
Rad1
Lig1
Lig3
DNA recombination
CDC9
Lig1
Lig3
mitochondrial genome maintenance
DNA-dependent DNA replication
mitochondrial DNA-dependent DNA replication
DNA ligation
CDC9
Lig1
pre-replicative complex formation and maintenance
Mcmd
MCM2 Mcm2 Mcmd2
MCM3 Mcm3
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7
Orc2
DNA replication elongation
DNA unwinding
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA priming
Prim1
Prim2
Pola1
Pola2
DNA initiation
MCM2
MCM3
CDC54/MCM4 Mcmd4
CDC46/MCM5 Mcmd5
MCM6
CDC47/MCM7
DNA strand elongation
Pcna
Recc1
lagging strand elongation
CDC2
DPB11
POL2
CDC9
leading strand elongation
CDC2
DPB11
POL2
SACCHAROMYCES
DROSOPHILA
MUS

P-values de GO para una Lista de Genes

Podemos calcular la probabilidad del tener *x de* de *n* genes con una anotación a un nodo GO, dado que en el genoma, *M de N* genes tienen esa anotación, usando la *distribución hipergeométrica*

$$P = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{x}}$$

Significancia de GO

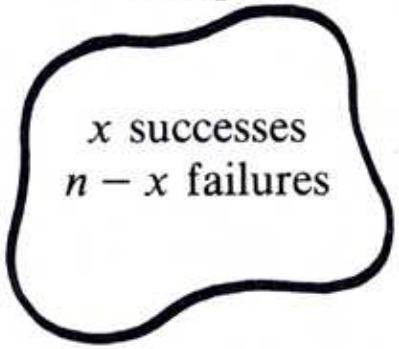
Para calcular un P-value, calculamos la probabilidad del tener *por lo menos x de las* anotaciones de *n*:

$$\text{P value} = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Fórmula para Probabilidad Hipergeométrica

N = Population size

n = Sample size



$$\binom{k}{x} = \text{ways of selecting } x \text{ successes from among } k \text{ available} \qquad (1)$$

$$\binom{N-k}{n-x} = \text{ways of selecting } n-x \text{ failures from among } N-k \text{ available} \qquad (2)$$

$$\binom{k}{x}\binom{N-k}{n-x} = \text{ways of selecting } x \text{ successes and } n-x \text{ failures from the respective numbers available} \qquad (1)\times(2)$$

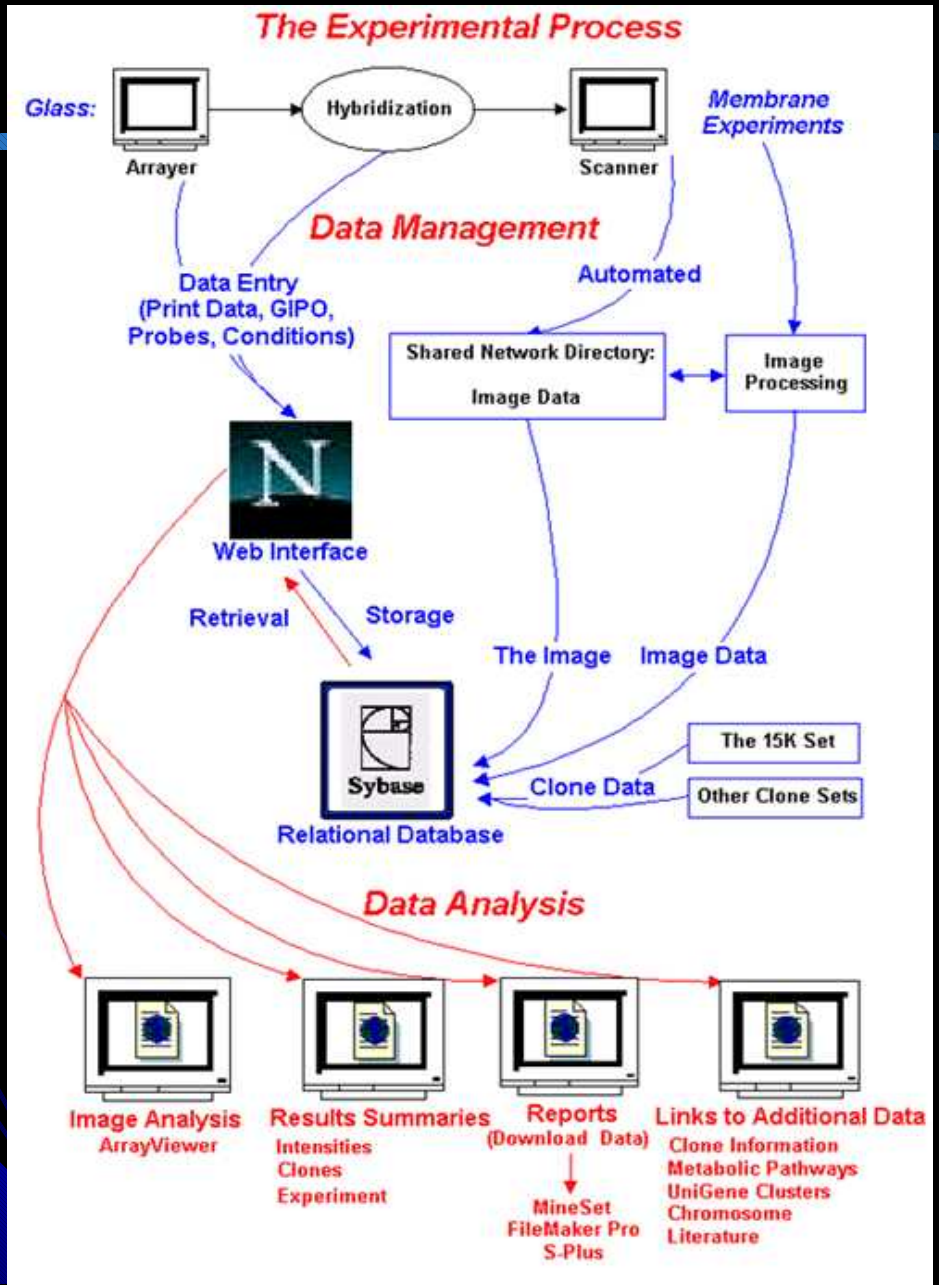
$$\binom{N}{n} = \text{number of samples of size } n \text{ that can be selected from the population of size } N \qquad (3)$$

$$h(x; n, k, N) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = \text{hypergeometric probability of getting } x \text{ successes (and } n-x \text{ failures) in a sample of size } n \text{ drawn from a population of size } N \text{ containing } k \text{ successes.} \qquad \frac{(1)\times(2)}{(3)}$$

Functional Analysis

- Identify over-represented Gene-Ontology terms
  - Biological Process
  - Molecular Function
  - Cellular Component
- Pathway Analysis:
  - Biocarta
  - GenMapp
  - Kegg
- Interactive Association Network : expression and / or annotations based
  - Graphic visualization that explain the relationship between genes
  - Expression profile based
  - Annotation based
  - Literature

Flujo de Datos

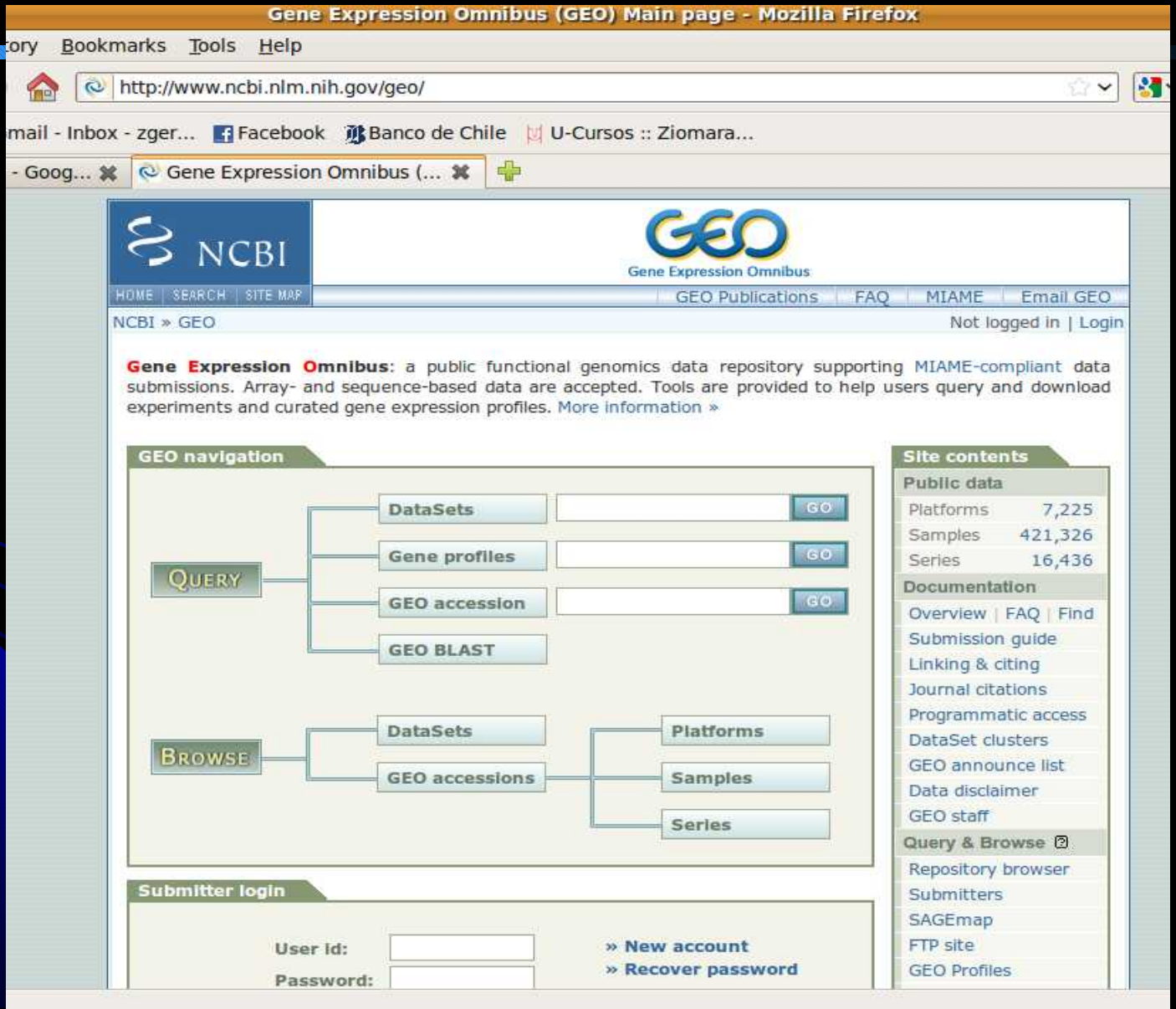


Bases de Datos de Microarray

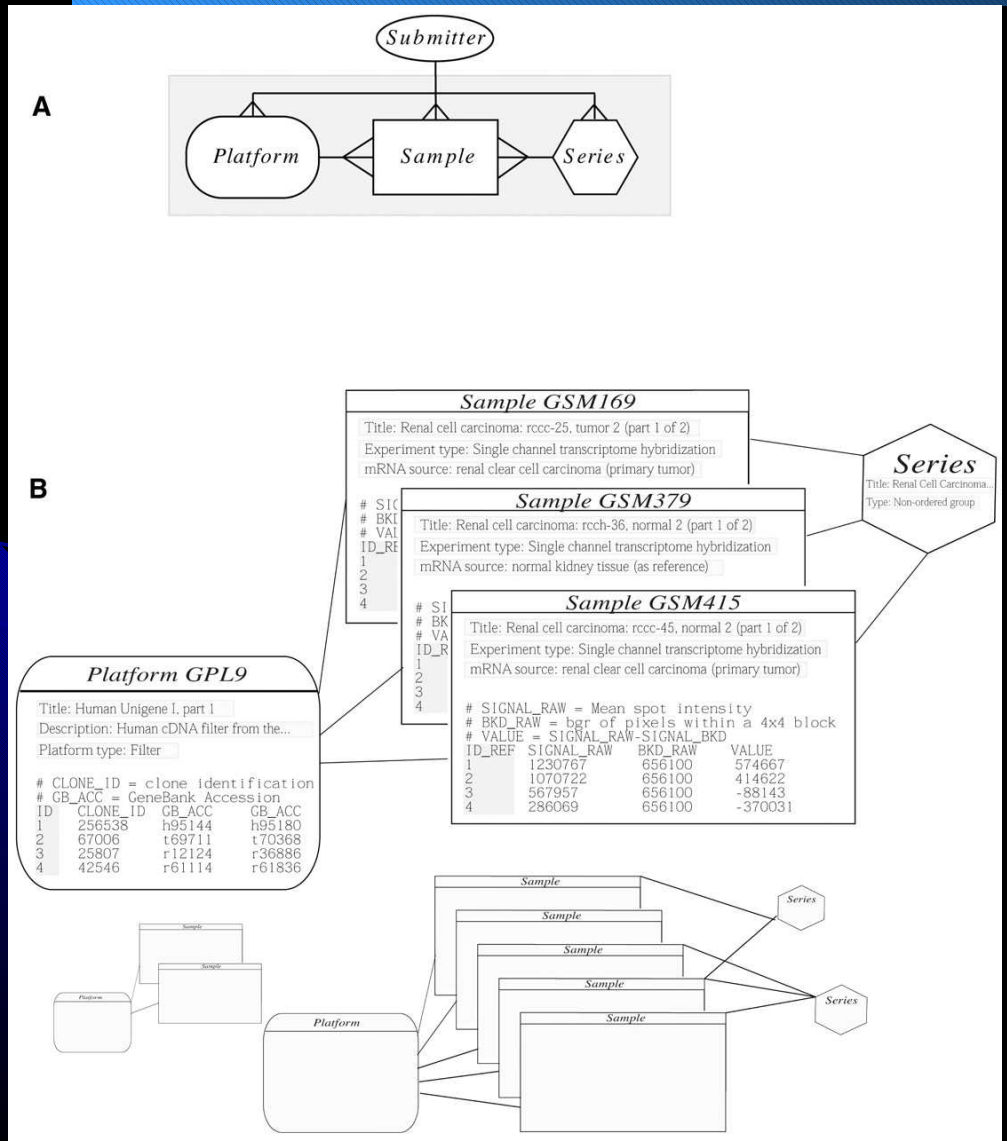
Hay dos depósitos principales:

- Gene expression omnibus (GEO) en NCBI
- ArrayExpress en el European Bioinformatics Institute (EBI)

Gene Expression Omnibus en NCBI



Gene Expression Omnibus: Esquema y Ejemplo



A. The entity-relationship diagram for the GEO database

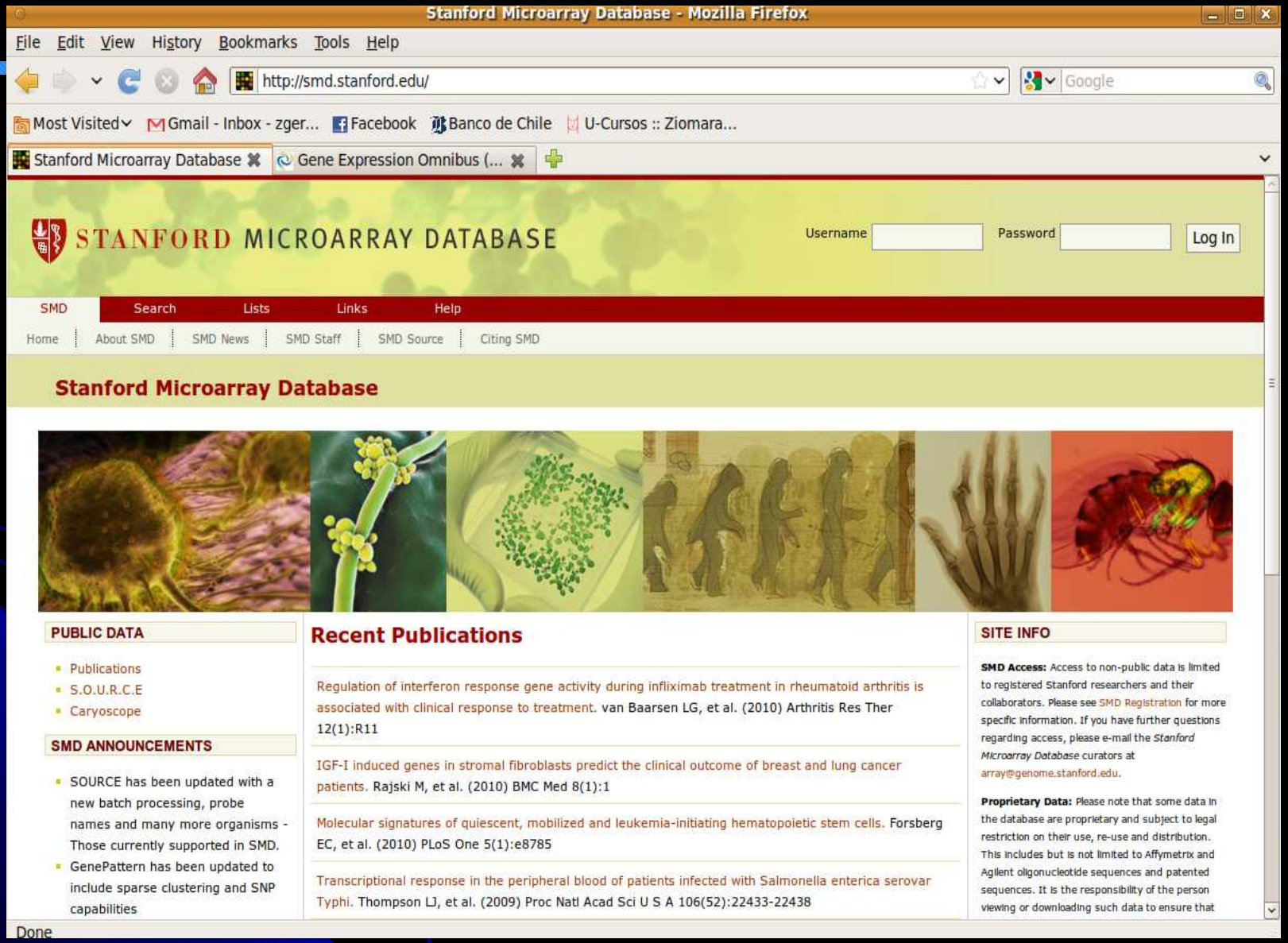
B. An actual example of three samples referencing one platform and contained in a single series

Nucleic Acids Research, 2002, Vol. 30, No. 1 207-210

Recursos HT Genómicos

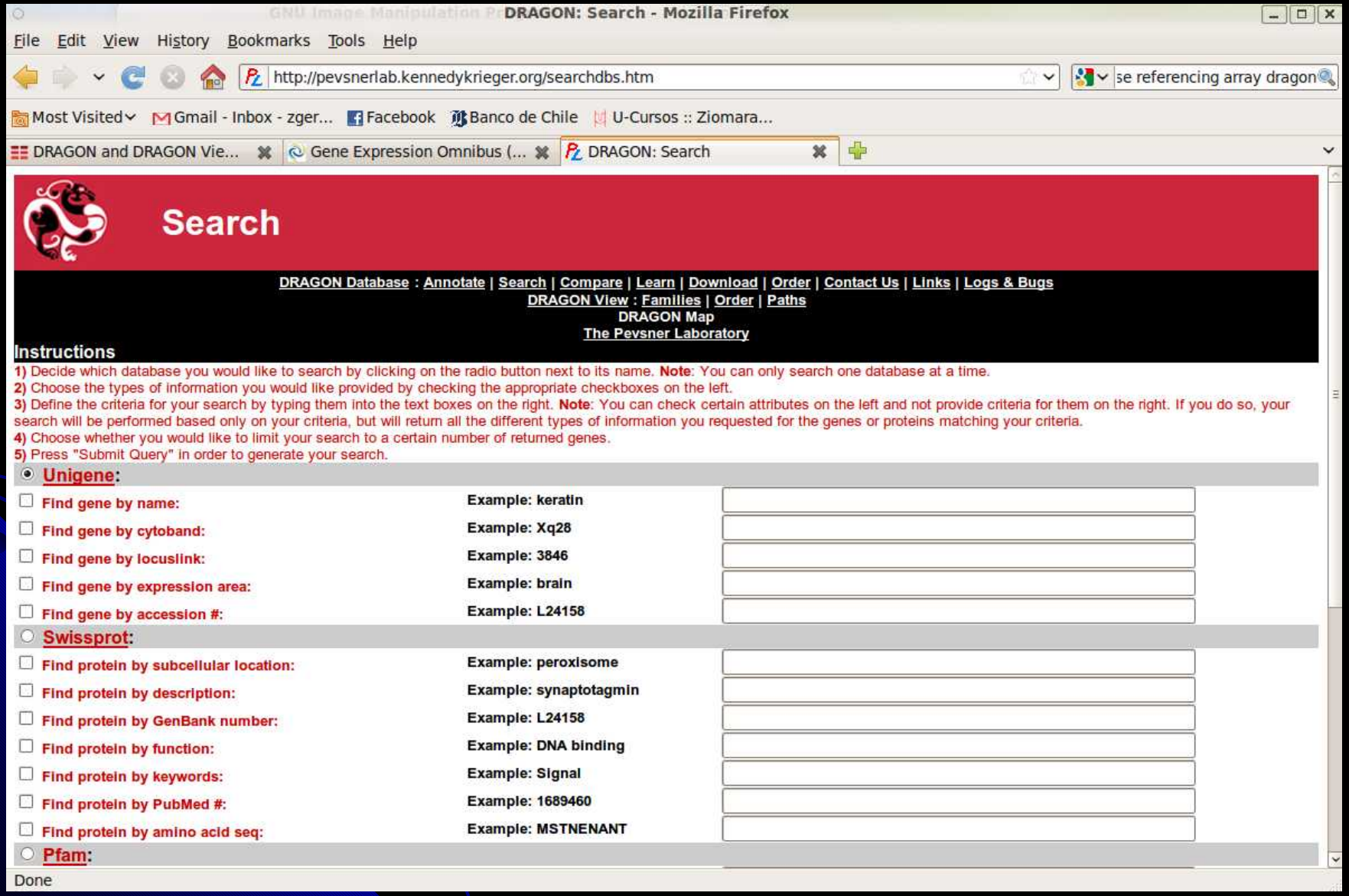
Resource name		Institution(s)	URL
Breast Cancer Cell Line Resource	1°	National Human Genome Research Institute, NIH	http://www.nhgri.nih.gov/DIR/CGB/CR2000
CGH Database	1°	Institute of Pathology, University Hospital Charité	http://amba.charite.de/~ksch/cghdatabase
Chip DB	1°	Whitehead Institute for Biomedical Research, MIT	http://young39.wi.mit.edu/chipdb_public
Drug & Alcohol Abuse Microarray Data Consortium	1°	Wake Forest University, Emory University, and Oregon Health and Science University	http://www.wfubmc.edu/microarray
ExpressDB	1°	Harvard–Lipper Center for Computational Genetics	http://arep.med.harvard.edu/ExpressDB
Global Gene Expression Group	1°	Science Park-Research Division, University of Texas M.D. Anderson Cancer Center	http://sciencepark.mdanderson.org/ggeg
MAExplorer	1°	National Cancer Institute, NIH	http://www-lecb.ncifcrf.gov/MAExplorer
Microarray center	1°	Children's National Medical Center	http://microarray.cnmcresearch.org/
Microarray project	1°	National Human Genome Research Institute, NIH	http://www.nhgri.nih.gov/DIR/Microarray
Rochester Muscle Database	1°	School of Medicine and Dentistry, University of Rochester Medical Center	http://www.urmc.rochester.edu/smd/crc/swindex.html
SADE	1°	Departement de Biologie Cellulaire et Moleculaire, CEA	http://www-dsv.cea.fr/thema/get/sade.html
SAGENET	1°	Johns Hopkins University School of Medicine	http://www.sagenet.org
Yeast Microarray Global Viewer	1°	Laboratoire de genetique moleculaire, Ecole Normale Superieure	http://transcriptome.ens.fr/ymgv
RNA Abundance Database	2°	Computational Biology and Informatics Laboratory, University of Pennsylvania	http://www.cbil.upenn.edu/RAD2
SAGEmap	2°	National Cancer Institute and National Center for Biotechnology Information, NIH	http://www.ncbi.nlm.nih.gov/sage
Stanford Microarray Database	2°	Dept. of Genetics, Stanford University School of Medicine	http://www.dnachip.org
Gene Expression Omnibus	3°	National Center for Biotechnology Information, NIH	http://www.ncbi.nlm.nih.gov/geo

Recursos Públicos para Microarrays

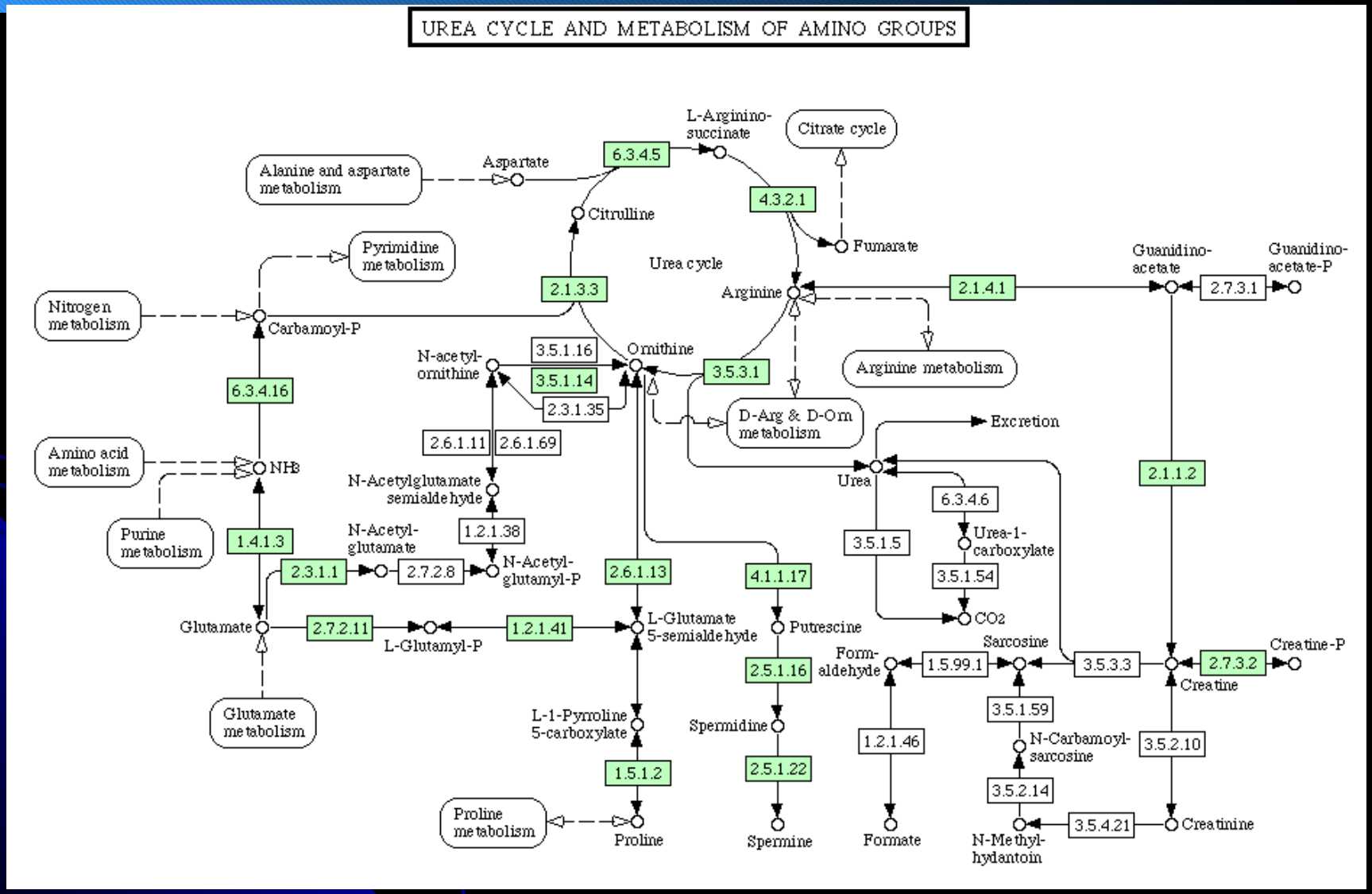


Database Referencing of Array Genes Online (DRAGON)

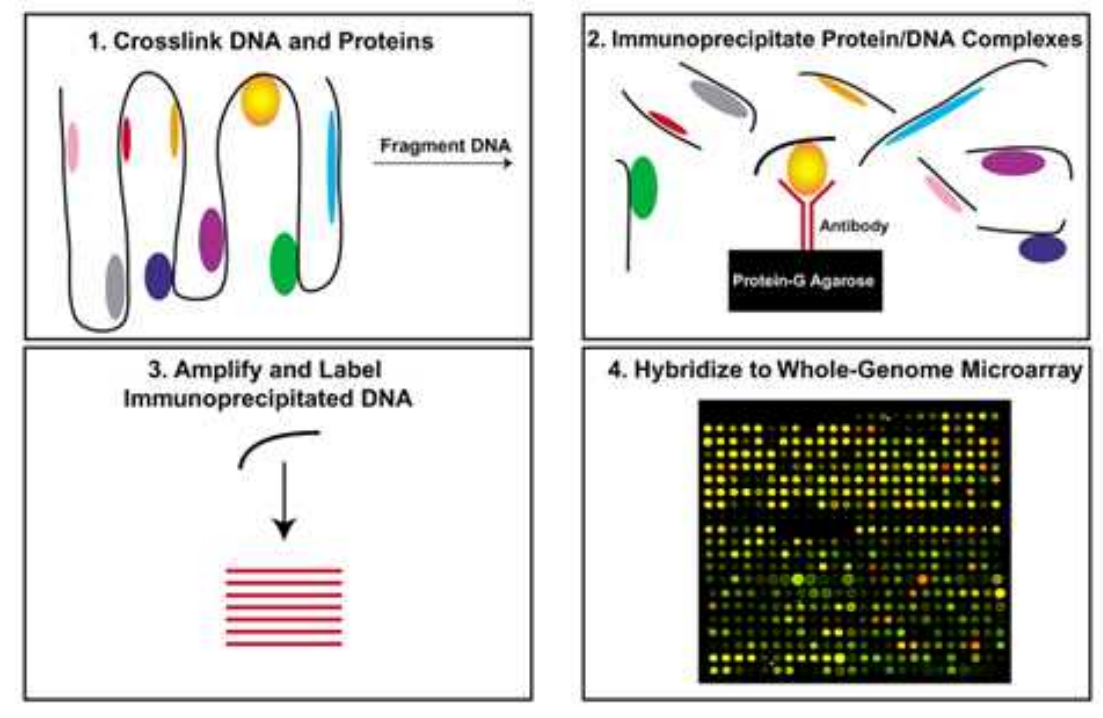
http://pevsnerlab.kennedykrieger.org/dragon.htm

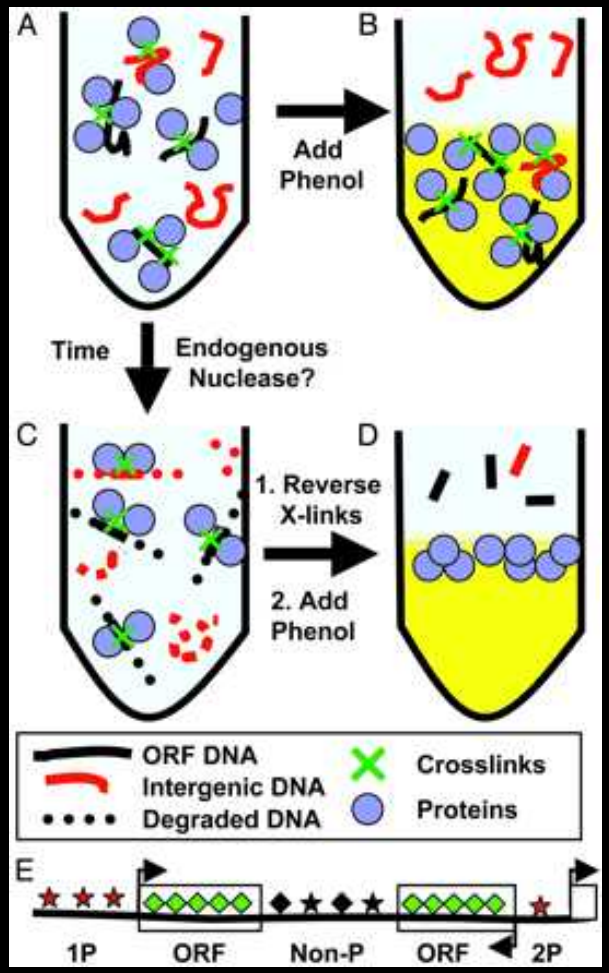

DRAGON: Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://pevsnerlab.kennedykrieger.org/searchdbs.htm
se referencing array dragon
Most Visited Gmail - Inbox - zger... Facebook Banco de Chile U-Cursos :: Ziomara...
DRAGON and DRAGON Vie...
Gene Expression Omnibus (...
DRAGON: Search
Search
DRAGON Database : Annotate | Search | Compare | Learn | Download | Order | Contact Us | Links | Logs & Bugs
DRAGON View : Families | Order | Paths
DRAGON Map
The Pevsner Laboratory
Instructions
1) Decide which database you would like to search by clicking on the radio button next to its name. Note: You can only search one database at a time.
2) Choose the types of information you would like provided by checking the appropriate checkboxes on the left.
3) Define the criteria for your search by typing them into the text boxes on the right. Note: You can check certain attributes on the left and not provide criteria for them on the right. If you do so, your search will be performed based only on your criteria, but will return all the different types of information you requested for the genes or proteins matching your criteria.
4) Choose whether you would like to limit your search to a certain number of returned genes.
5) Press "Submit Query" in order to generate your search.
Unigene:
Find gene by name: Example: keratin
Find gene by cytoband: Example: Xq28
Find gene by locuslink: Example: 3846
Find gene by expression area: Example: brain
Find gene by accession #: Example: L24158
Swissprot:
Find protein by subcellular location: Example: peroxisome
Find protein by description: Example: synaptotagmin
Find protein by GenBank number: Example: L24158
Find protein by function: Example: DNA binding
Find protein by keywords: Example: Signal
Find protein by PubMed #: Example: 1689460
Find protein by amino acid seq: Example: MSTNENANT
Pfam:
Done

DRAGON Relates Genes to KEGG Pathway



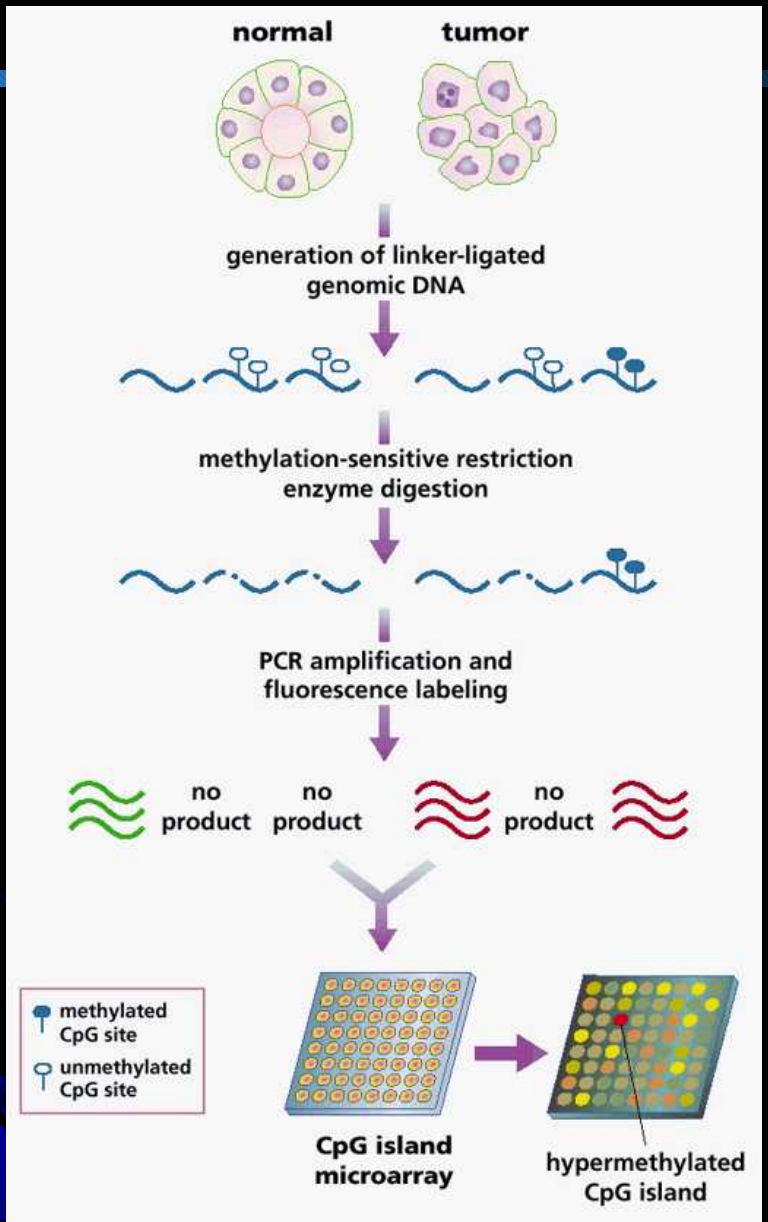
Mapeo de Interacciones DNA–proteina y ORF usando Microarrays





From *Nagy P.L. et. al*, PNAS vol. 100, 2003

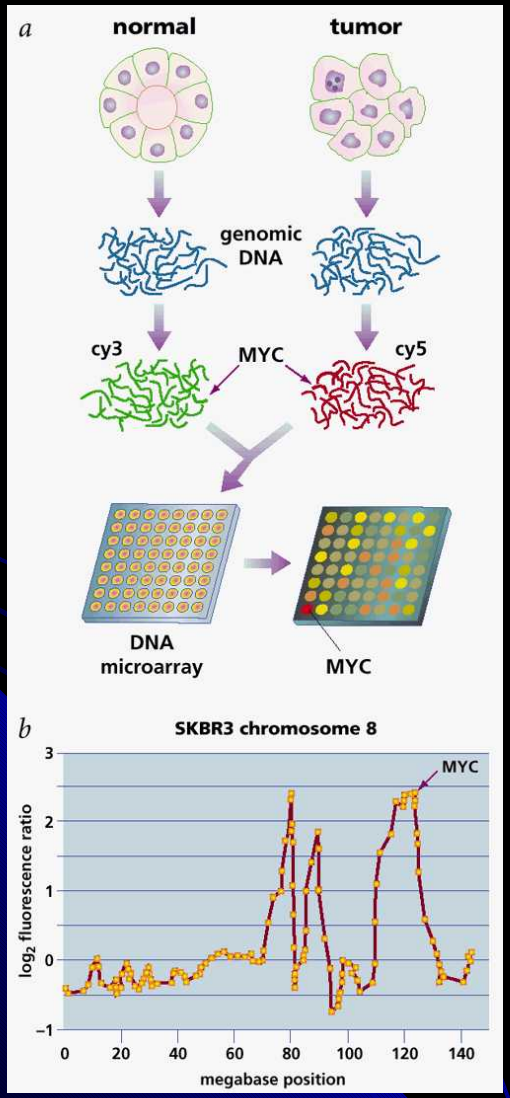
Differential Methylation Hybridization



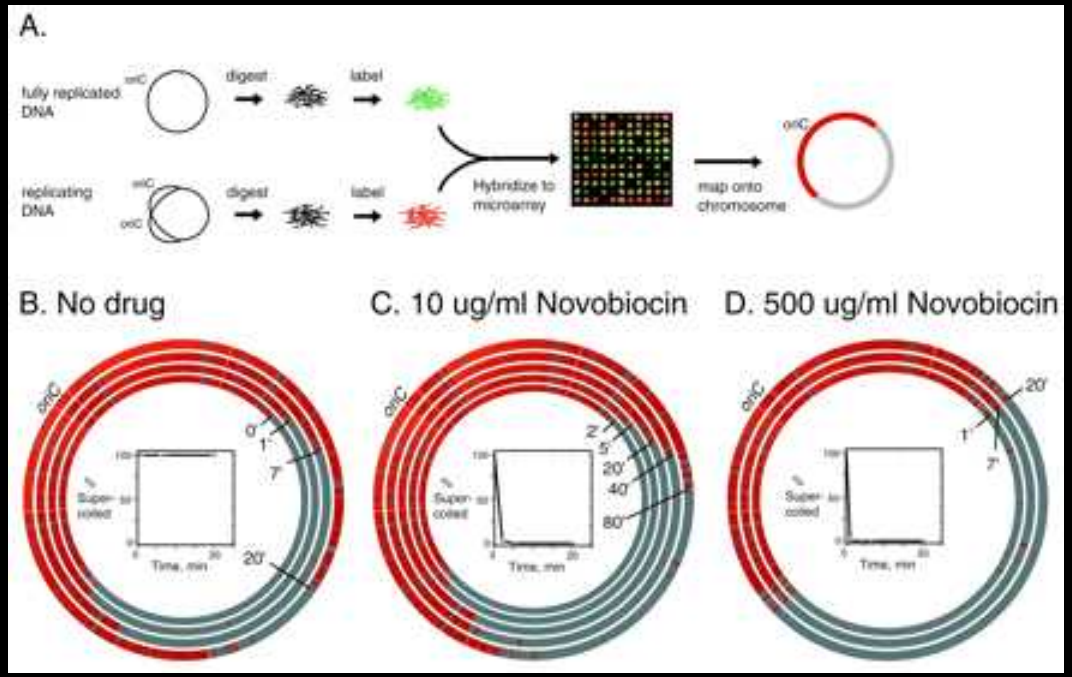
Pollack JR, Iyer VR.
Nat Genet. 2002 Dec;32 Suppl:515-21

Hibridación Comparativa en Microarrays

Esquema estático



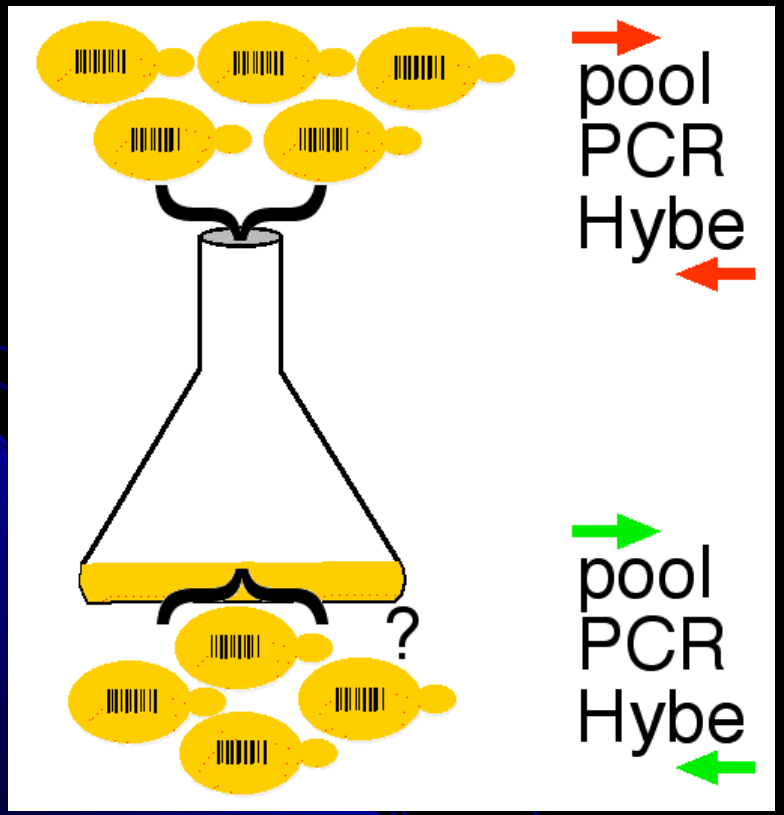
Esquema dinámico

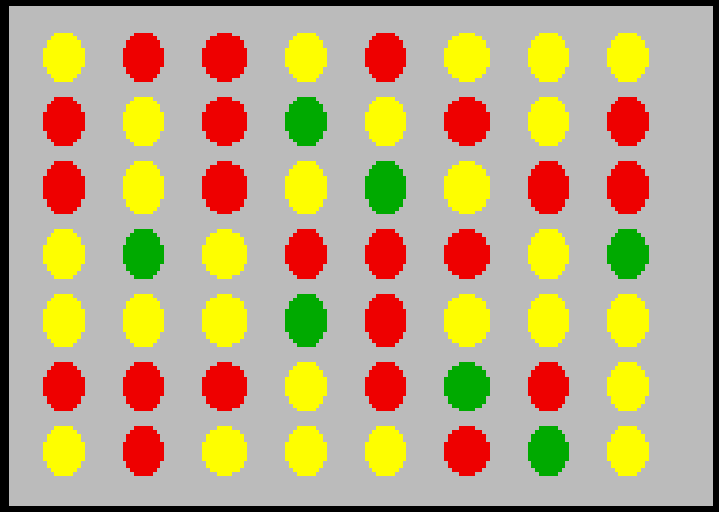


Khodursky AB et al., PNAS, PNAS 2000 vol. 97

Pollack JR, Iyer VR.
Nat Genet. 2002 Dec;32
Suppl:515-21

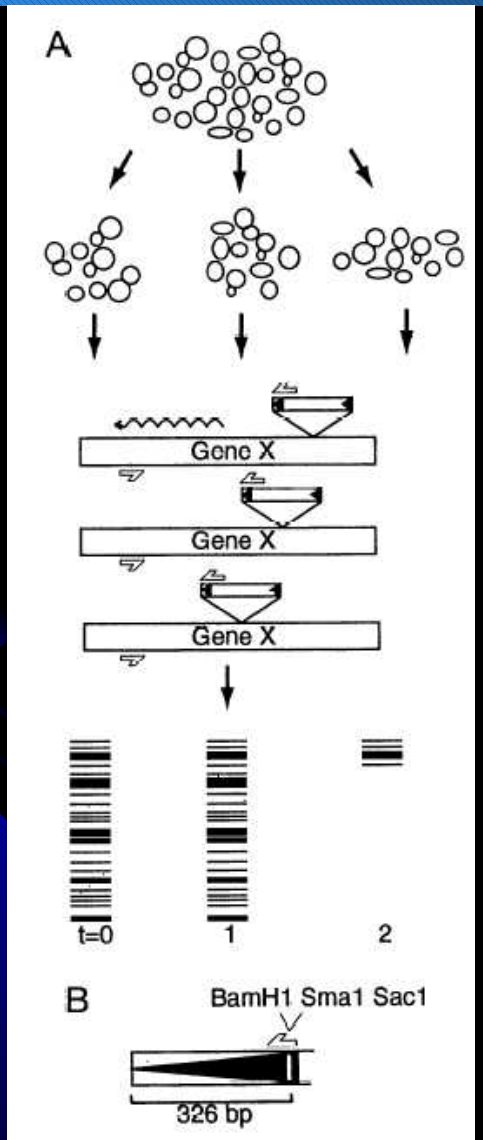
Microarrays de Código de Barras



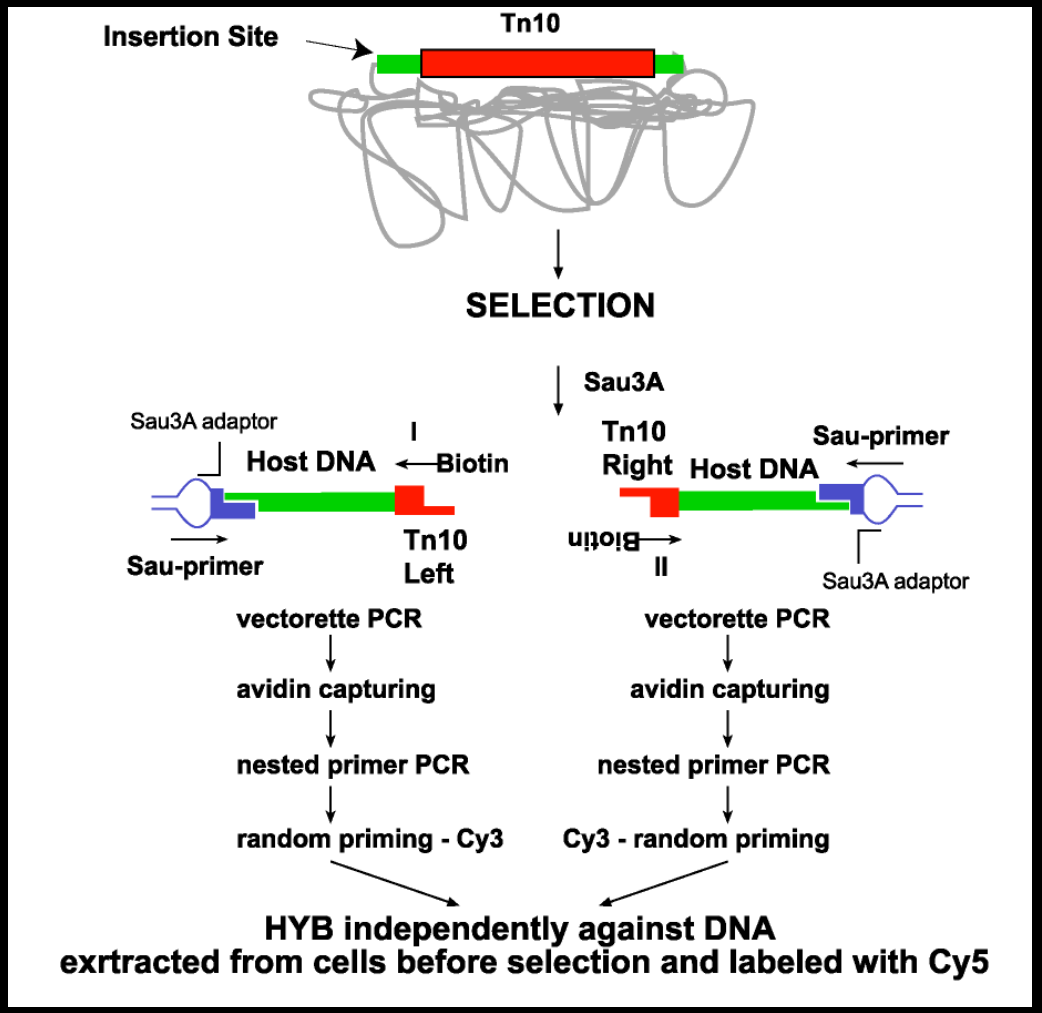


Footprinting Genético

Gel



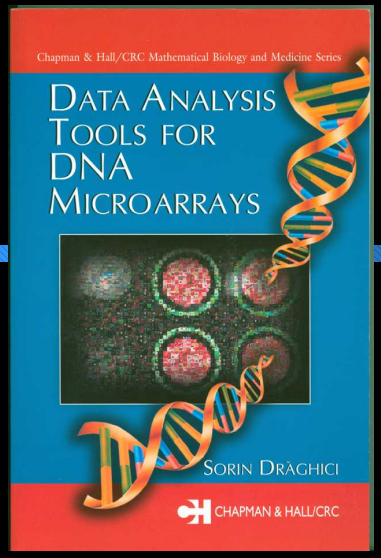
Array



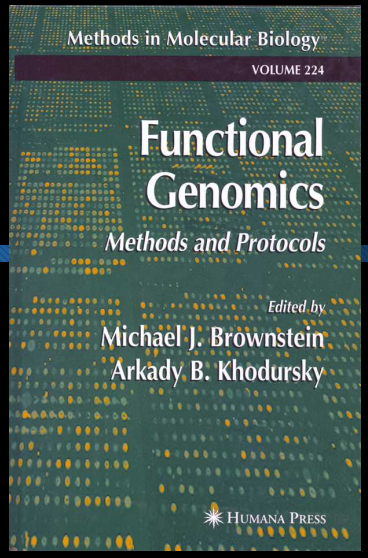
Proc Natl Acad Sci U S A. 1995 Jul 3;92(14):6479-83

Lectura Adicional: *Análisis Deductivo*

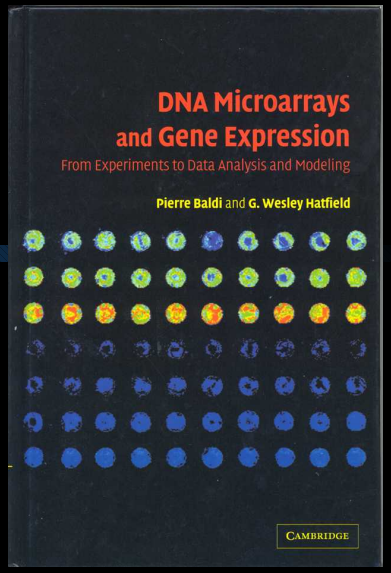
- Efron, B. and R. Tibshirani **(2002). "Empirical bayes methods and false discovery rates for microarrays."** Genet Epidemiol 23**(1): 70-86.**
- Ideker, T., V. Thorsson, et al. **(2000). "Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data."** J Comput Biol 7**(6): 805-17.**
- Kerr, M. K. and G. A. Churchill **(2001). "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments."** Proc Natl Acad Sci U S A 98**(16): 8961-5.**
- Kerr, M. K. and G. A. Churchill **(2001). "Statistical design and the analysis of gene expression microarray data."** Genet Res 77**(2): 123-8.**
- Kerr, M. K., M. Martin, et al. **(2000). "Analysis of variance for gene expression microarray data."** J Comput Biol 7**(6): 819-37.**
- Park, P. J., M. Pagano, et al. **(2001). "A nonparametric scoring algorithm for identifying informative genes from microarray data."** Pac Symp Biocomput**: 52-63.**
- Smyth, G. K., Y. H. Yang, et al. **(2003). "Statistical issues in cDNA microarray data analysis."** Methods Mol Biol 224**: 111-36.**
- Tusher, V. G., R. Tibshirani, et al. **(2001). "Significance analysis of microarrays applied to the ionizing radiation response."** Proc Natl Acad Sci U S A 98**(9): 5116-21.**
- Wolfinger, R. D., G. Gibson, et al. **(2001). "Assessing gene significance from cDNA microarray expression data via mixed models."** J Comput Biol 8**(6): 625-37.**
- Yang, Y. H., S. Dudoit, et al. **(2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation."** Nucleic Acids Res 30**(4): e15.**

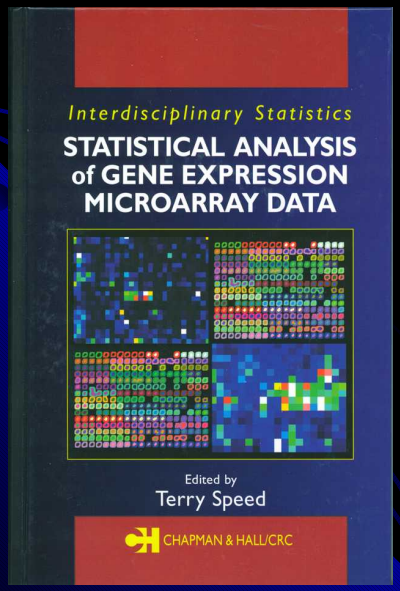


Accesible y bien pensado

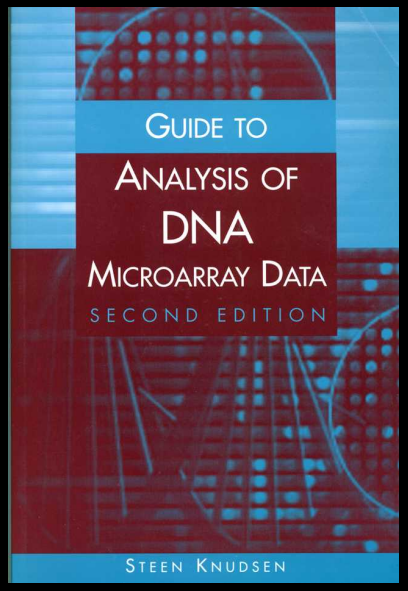


Coleccion de protocolos y contribuciones

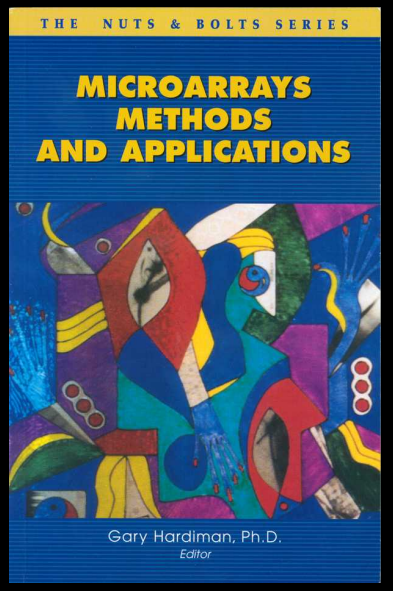




Tour de fuerza en estadística



Un poco de todo



Una guia para quienes saben poco o nada del tema

Lecturas Simples en Estadística y Clasificación