Genómica
y
Estadística

DIANA M. KELMANSKY
Instituto de Cálculo
FCEN-UBA
dkelman@ic.fcen.uba.ar

¿Por qué?

- Secuenciación del genoma humano hacia 2003.
- Microarreglos 1995.
- Secuenciadores de ultra velocidad 2008
- Explosión de datos y trabajos

Posibilidad de medir
miles de secuencias genómicas
simultáneamente

en una gran variedad de organismos y en cualquier momento de su desarrollo.

Recordemos

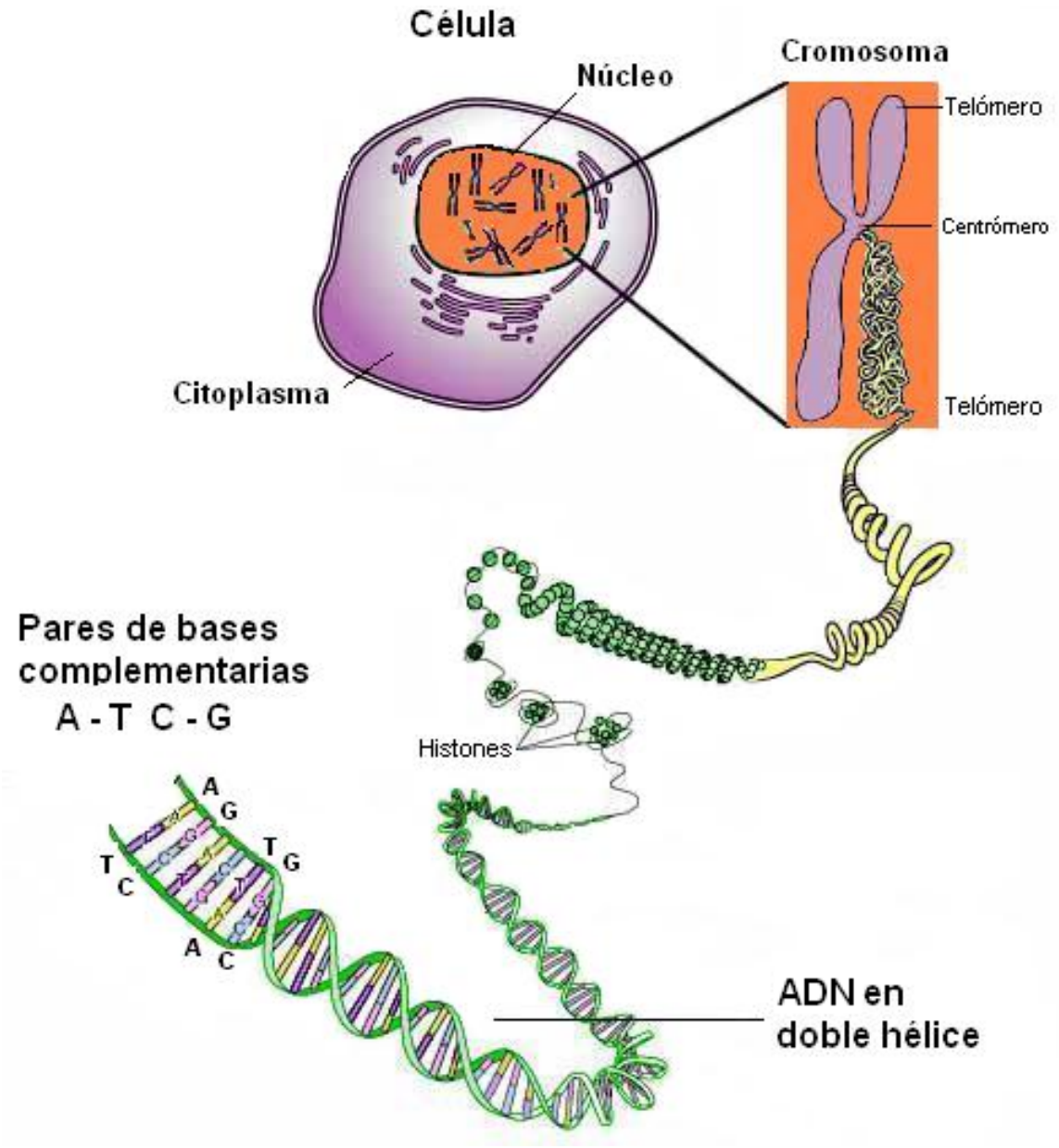
- El modelo del ADN 1953
- fue propuesto por
- Francis Crick, James Watson y Maurice Wilkins

- Premio Nobel de Medicina -1962-

Recordemos

en base al
trabajo de ROSALIND FRANKLIN como
bióloga molecular y cristalógrafa

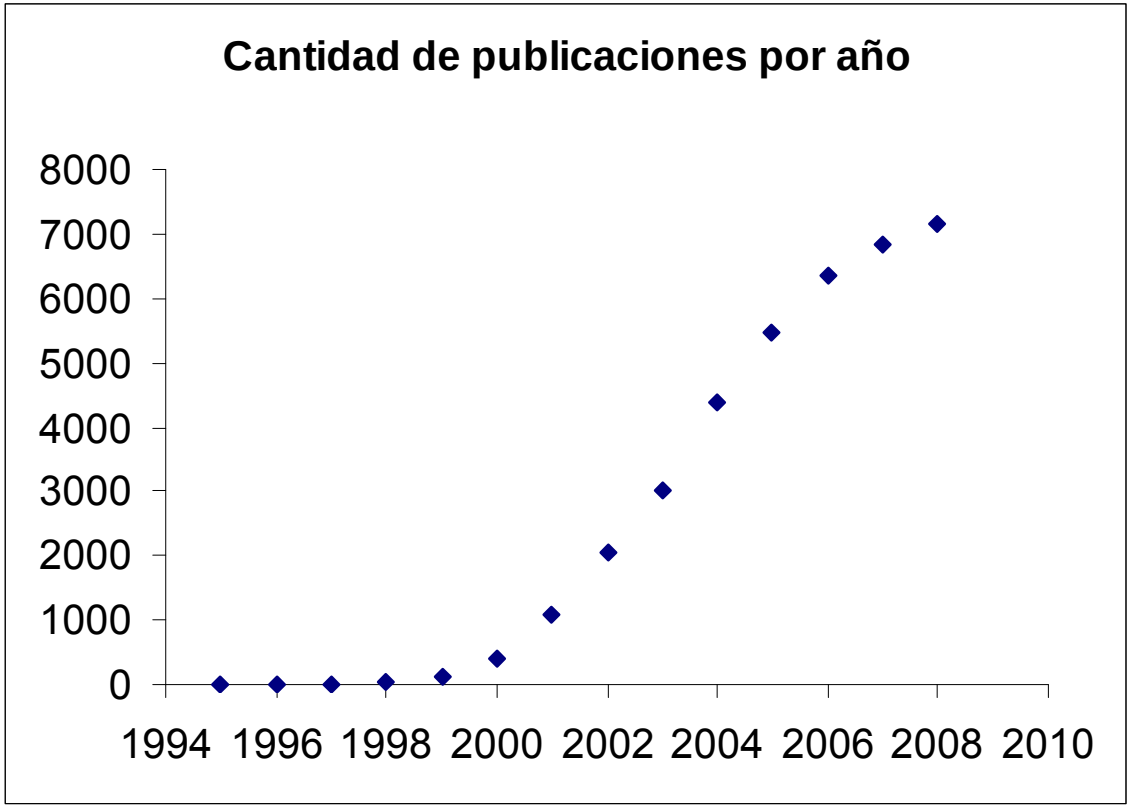
La fotógrafa del ADN
murió de cáncer en 1958 con 37 años


Célula
Núcleo
Cromosoma
Telómero
Centrómero
Telómero
Citoplasma
Pares de bases
complementarias
A - T C - G
Histones
A
G
T
G
T
C
A
C
ADN en
doble hélice

Parece simple

- Un código de 4 letras
- Toda la información está allí escrita
- Sólo hace falta conocerla
- Ideal para que la manejen las computadoras
- Nace la Bioinformática

¡Grandes Esperanzas!



PubMed palabra clave microarray

Schena M,et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (1995)

¡Grandes Expectativas!

Eric S. Lander 1999 *Nature Genetics* publica el trabajo titulado "**Array of Hope**"

Mark Schena *Microarray Analysis* 2003

Al final de la introducción:

"Fifty years from now, and long after human disease has been eradicated, we will look back incredulously at the start of this millennium and wonder how we ever endured cancer, heart disease, AIDS and thousands of other illnesses that compromise our well-being"

¡Grandes expectativas!

En el avance de los conocimientos sobre:

- procesos moleculares biológicos
- diagnóstico y pronóstico de enfermedades
- mecanismos acción de una droga
- mejoramiento de las estrategias terapéuticas – medicina personalizada

Objetivos: Identificar cambios

1) en la abundancia de ARNm **genes expresados (**transcriptomic array)
2) en la secuencia del ADN de algún sector de un cromosoma

entre condiciones diferentes

Todas nuestras células contienen *la misma información genética*.

¿Qué es lo que hace que, por ejemplo, las células de la piel sean diferentes de las del hígado?

Diferentes genes se expresan en diferentes niveles

Epigenética

- Cambios en la apariencia (fenotipos) o comportamiento que se transmiten de una generación a otra sin que se produzcan cambios en el ADN.

- Ocurren mecanismos por encima (epi) de la genética.

Epigenética - mecanismos

- Dos mecanismos epigenéticos predominantes son
- la metilación del ADN y
- la modificación de las histonas

.

1) ¿Qué es un gen?
2) ¿Qué significa que un gen se exprese?

Gen: segmento específico de la molécula de ADN que contiene toda la información necesaria para instruir a la célula que sintetice un producto específico.

Dogma central de la biología molecular

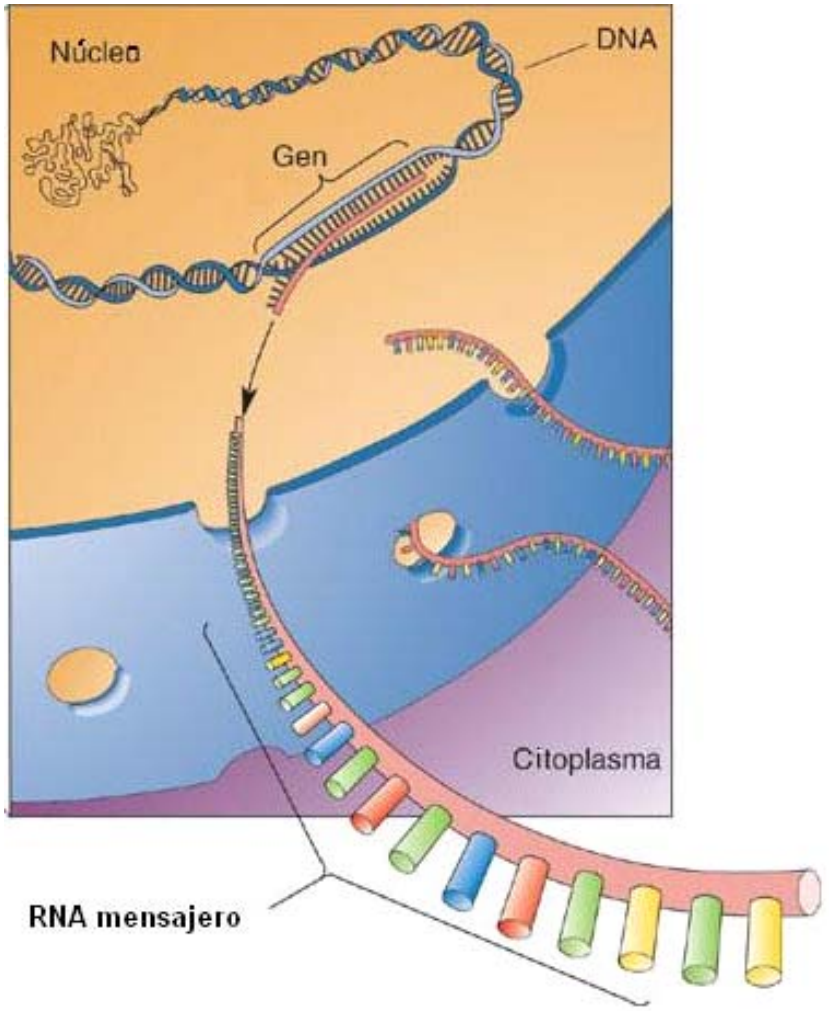
Doble cadena de ADN

↓ transcripción o expresión

Simple cadena de ARNmensajero

↓ traducción

Proteína



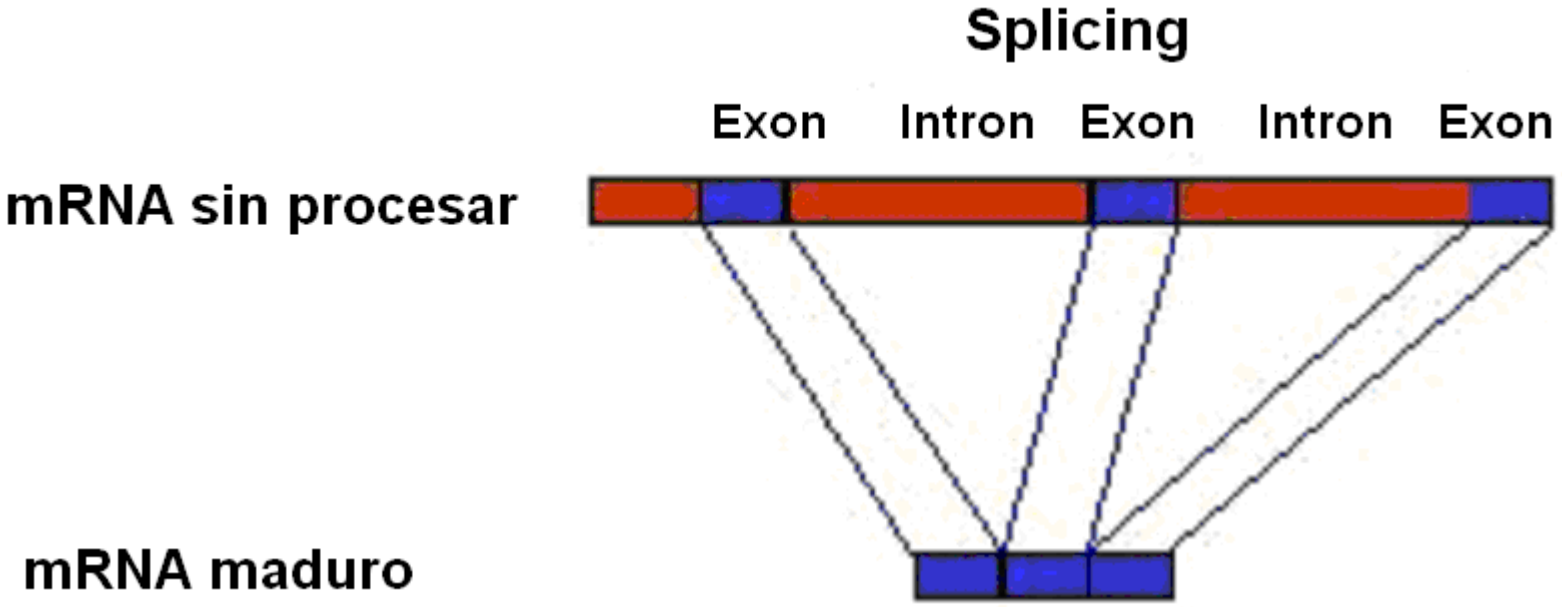
Transcripción de un gen

Dentro de cada gen del ADN, hay segmentos que tienen un papel activo en el proceso de codificación: **exones**, ARNm que sale del núcleo luego de la transcripción

y

también hay otros segmentos que no codifican: **intrones,** parte del ARNm que se transcribió pero que no sale del núcleo.

El ARN mensajero que sale del núcleo es el ARNm *maduro* que ha sufrido los procesos de capping (G), polyadenylation (AAAA...) y splicing.



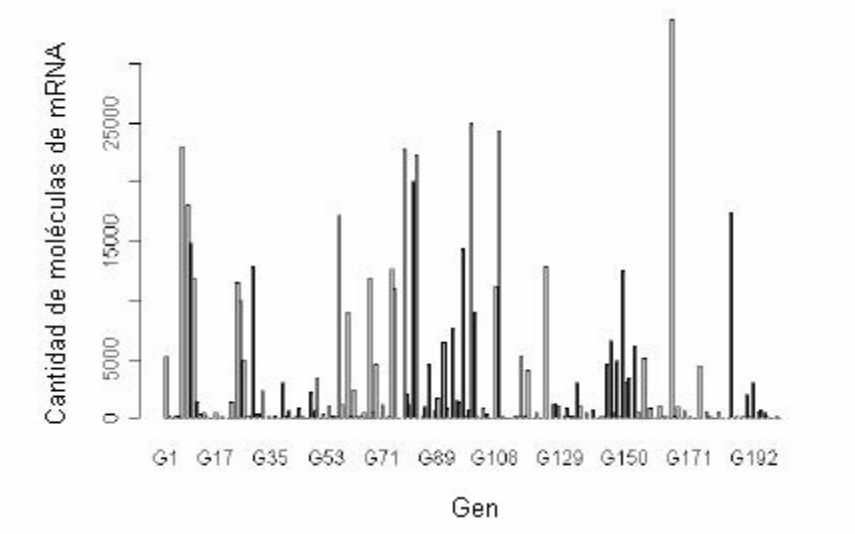
Esquema del proceso de splicing

Cualquier secuencia (cadena genómica, o gen) que esté activa de esta manera se dice que está ***expresada*,**

El *nivel* de expresión de un gen es la *cantidad de copias* de ARNm transcriptos presentes en la célula en un determinado momento

Perfil de expresión

Si pudiésemos contar la cantidad de moléculas de ARNm para cada gen en una única célula obtendríamos su perfil de expresión "verdadero".



Perfil de expresión "verdadero"

Microarreglos

El microarreglo actúa como un detector de, por ejemplo, la cantidad de ARN mensajero presente en el tejido.

Detector de ARN mensajero

Doble cadena de ADN

↓ transcripción o expresión

Simple cadena de ARNm

↓

Microarreglo → → → → → → →

↓ traducción

Proteína

Detector de ADN cromosómico

Doble cadena de ADN

↓

Microarreglo → → → → → → →

↓ transcripción o expresión

Simple cadena de ARNm

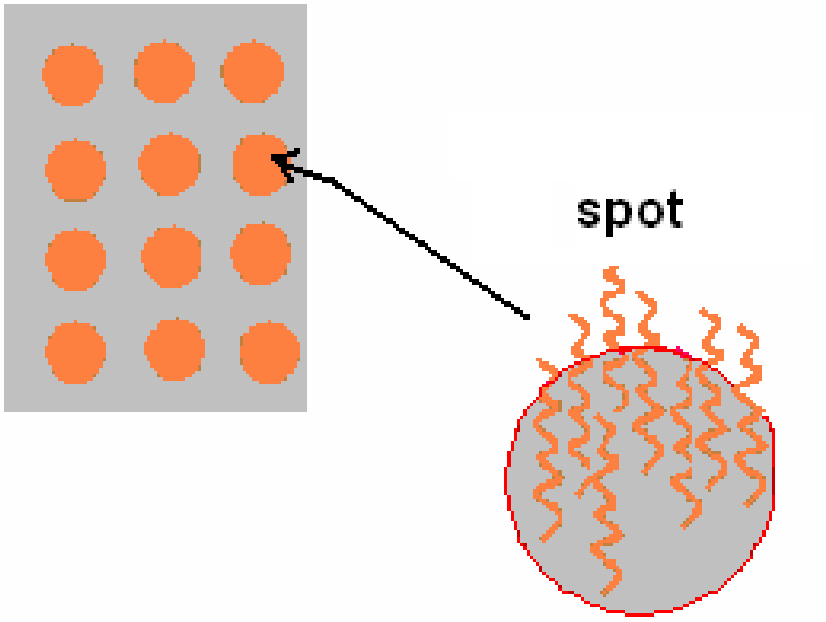
↓ traducción

Proteína

¿Qué es un microarreglo?

Sustrato sólido (vidrio, plástico o silicio)
tiene adheridas (por unión covalente)
sondas (probes) microscópicas
con clones de ADN, cADN, oligos
ordenadas en forma matriz de miles de puntos (10000 – 40000) equiespaciados, cubriendo parte o toda la secuencia de un genoma-transcriptoma de un organismo

Cada sonda (probe, spot) contiene millones de hebras “idénticas”



¿Cómo actúan las sondas de un microarray?

P*rincipio* biológico de *complementaridad*

A – T

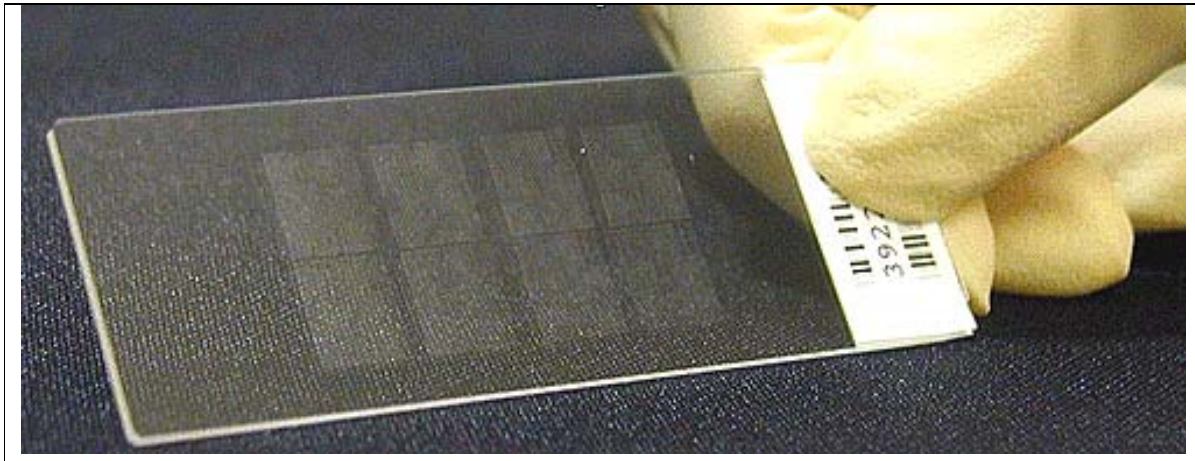
C - G

Es el mismo que el que determina que el ADN en las células tenga una estructura de doble cadena.

Cada sonda del microarreglo actúa a modo de tubo de ensayo.

Aquellas cadenas, en el material incógnita, que tienen una *secuencia complementaria* a las de esa sonda del arreglo se pegan por el principio de complementaridad, formando una doble cadena.

Dos tecnologías



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

gslc.genetics.utah.edu

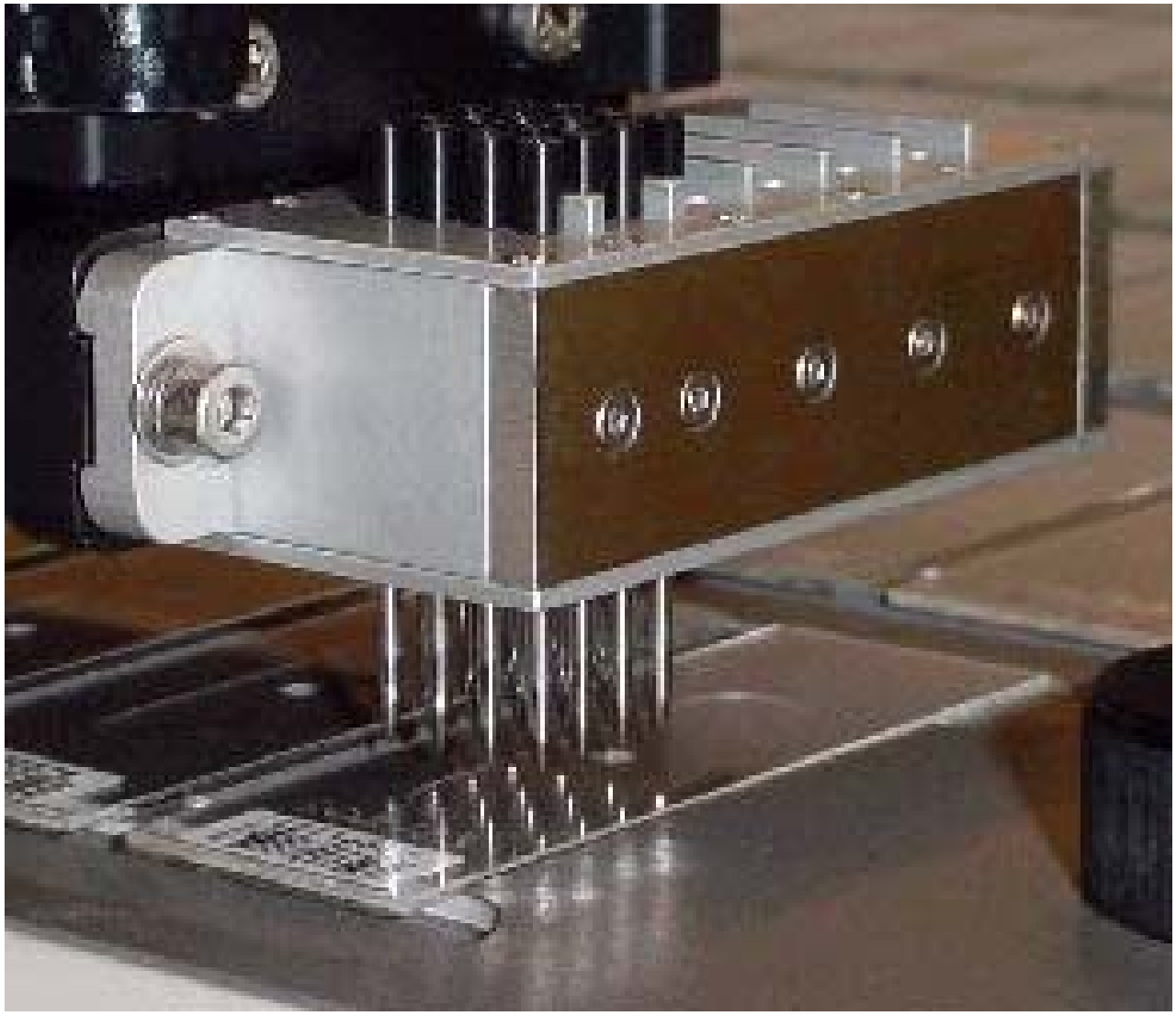
Depósito (delivery)

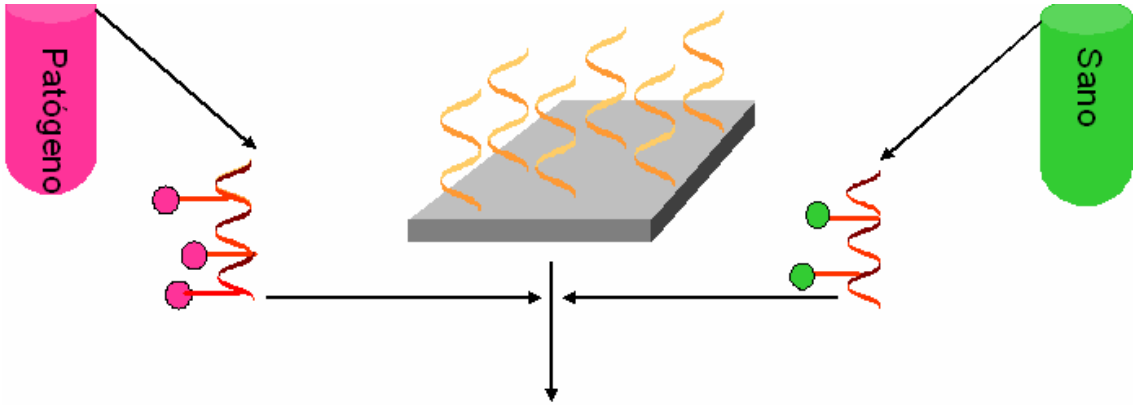
arrays

Síntesis

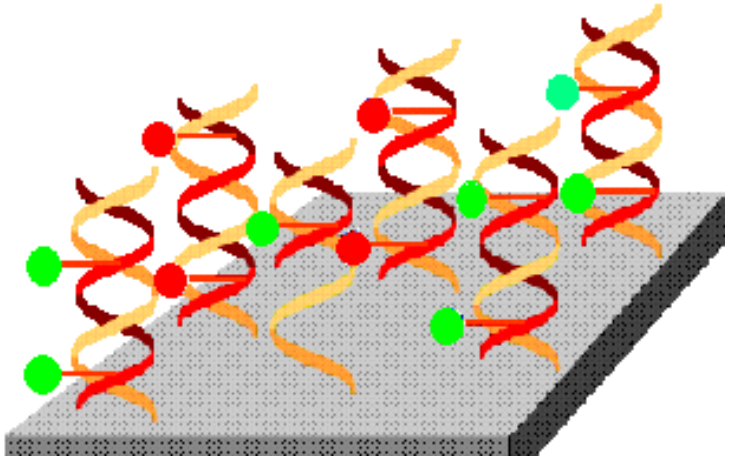
chips

Portaobjeto y cabezal de un robot



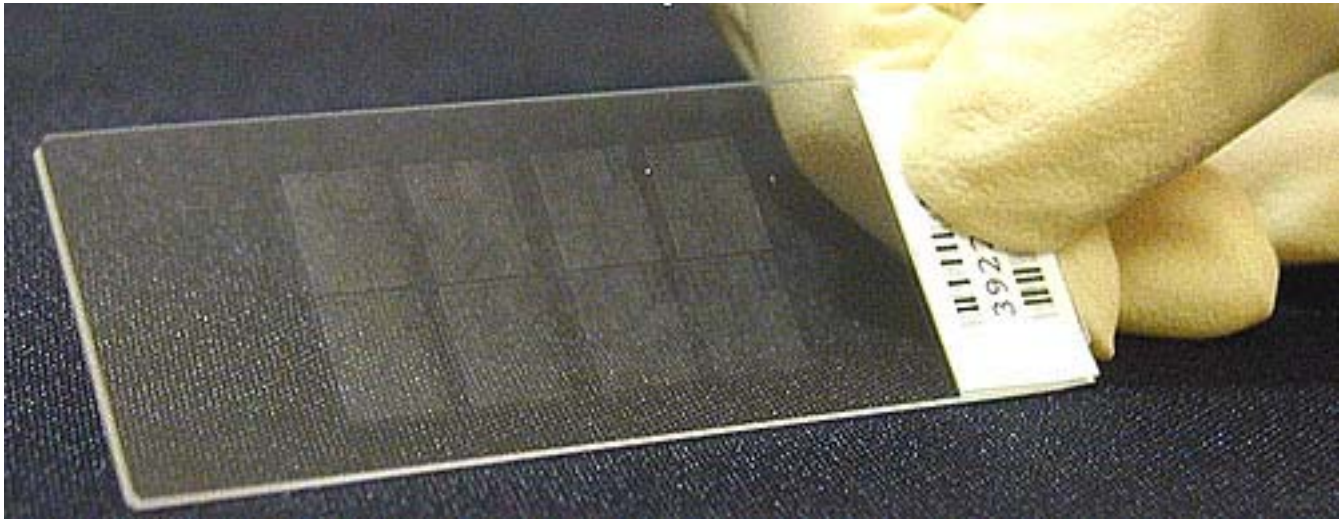


Etiquetado con tintes con fluor de las muestras e incorporación sobre el microarreglo



Al finalizar el experimento obtendremos

2 imágenes digitales de

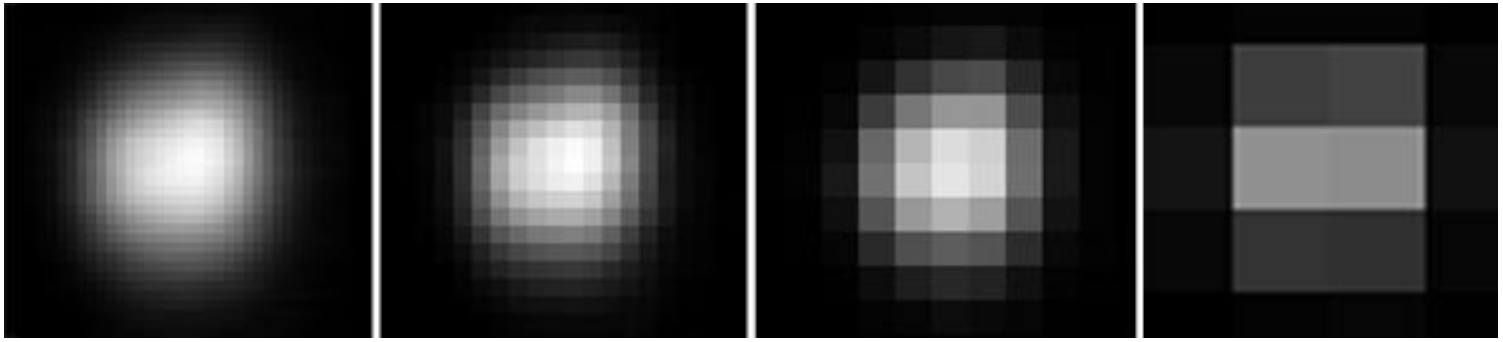


two color spotted microarray
un microarreglo de dos colores

Obtención de las imágenes:

Las moléculas de flúor son excitadas mediante una luz láser. Imagen digital para cada tinte.

La intensidad de cada píxel representa la abundancia del gen en un sector de una sonda del microarreglo.



Datos iniciales

- Pares de archivos de imágenes uno para cada tinte.
- Por ejemplo
  - ~ 43000 spots;
  - ~ cientos de pixeles por cada spot (depende de la resolución);

Datos iniciales

– un número para cada píxel

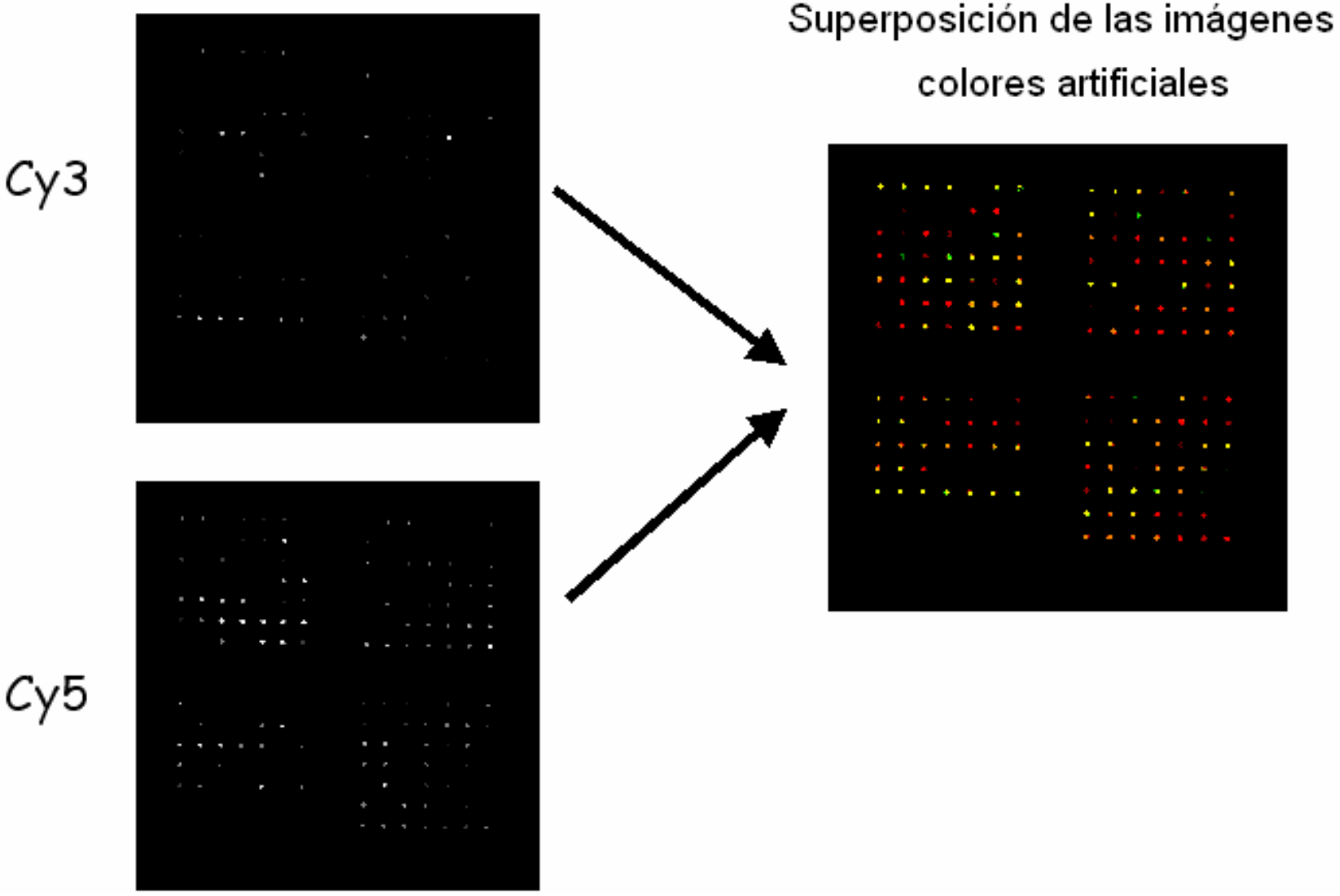
– ~ cientos de números por cada sonda - gen

Procesamiento de imágenes

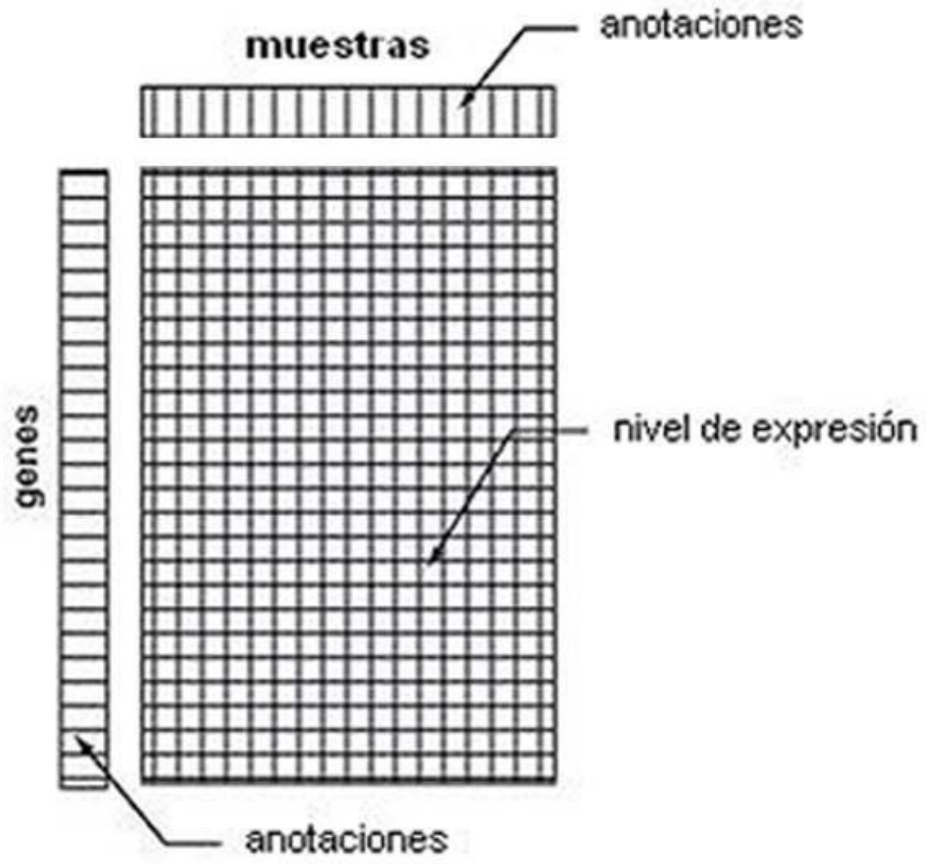
↓

Datos crudos: 1 número para cada sonda

== nivel de expresión de un gen


Cy3
Cy5
Superposición de las imágenes
colores artificiales

Matriz de expresiones



Aspectos estadísticos

Diseño

Diseño del arreglo: decidir qué sondas y donde, serán impresas al sustrato sólido.

Diseño de las muestras que se podrán sobre el arreglo:

- preparación
- replicaciones biológicas indispensables

Aspectos estadísticos - cont.

Preprocesamiento.

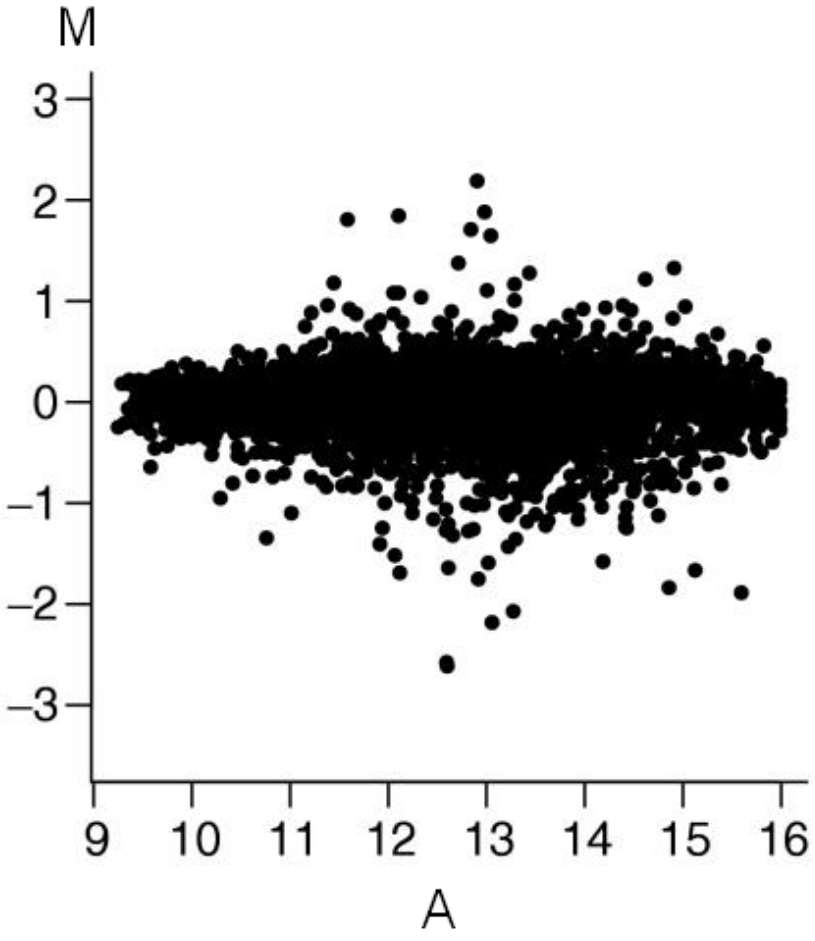
- **análisis de imágen** cuantificación de los "spots": distinguir las intensidades del foreground de las del background y los artifacts. Medidas resumen.
- **Mal llamada normalización**

  Para corregir las estructuras curvas no atribuibles a motivos biológicos en gráficos M-A .

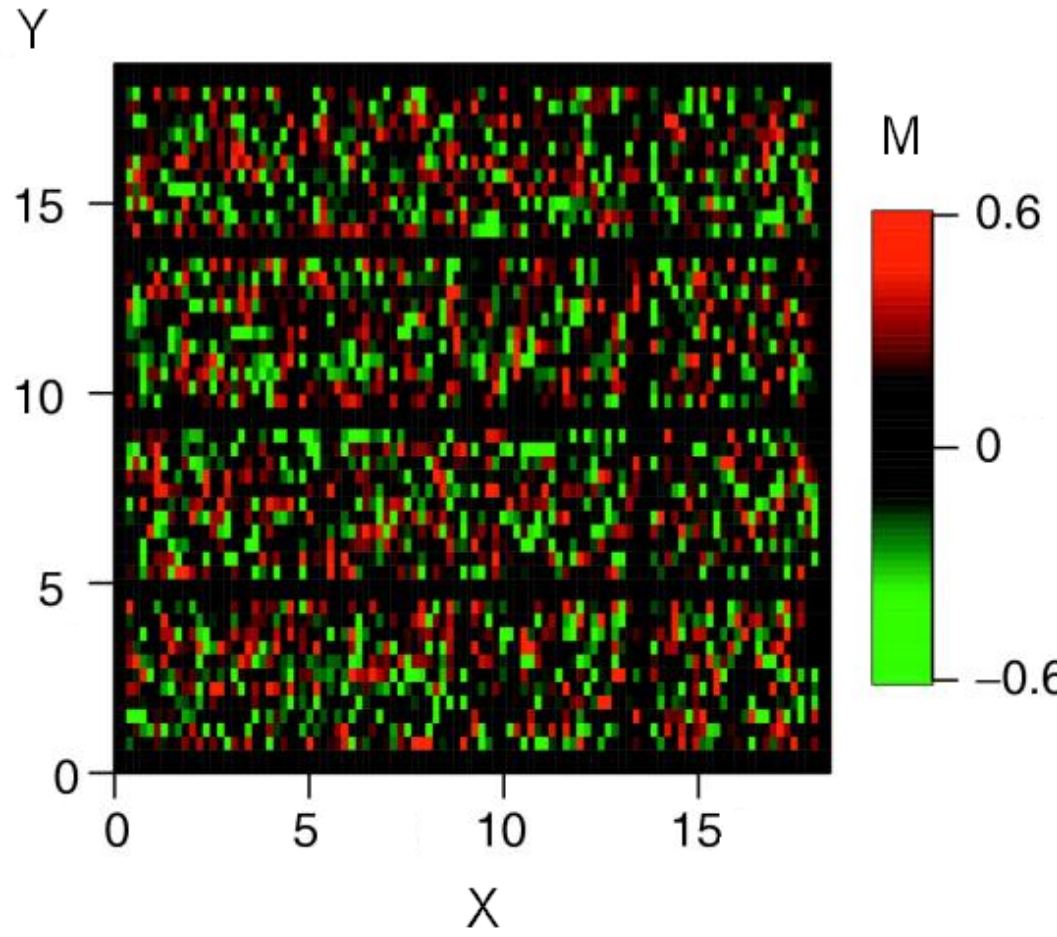
  Para eliminar dependencias media -varianza

Experimento SELF-SELF ideal

MA plot



MXY plot



MA-plot

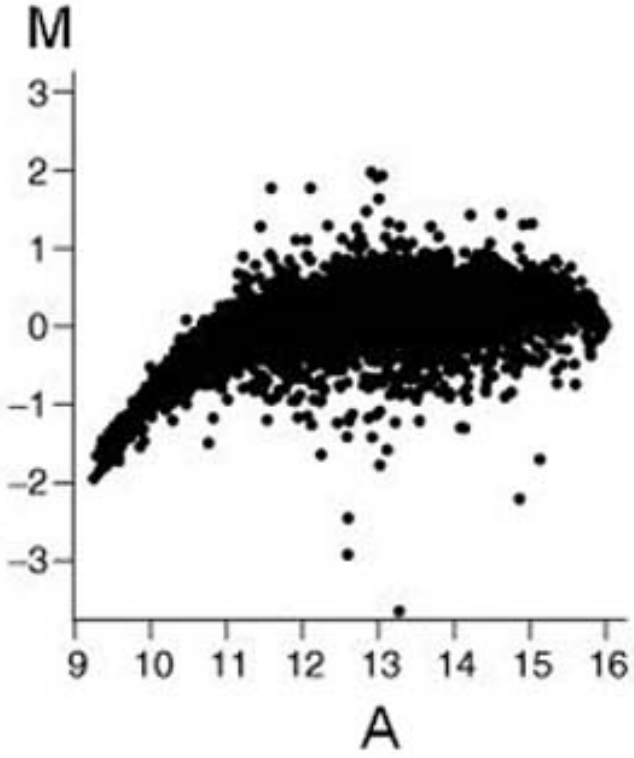
- Diagrama de dispersión (Scatter plot) de
  - **M** = $\log_2$ (Xred / Xgreen)
    = $\log_2$ (Xred) - $\log_2$ (Xgreen)

versus

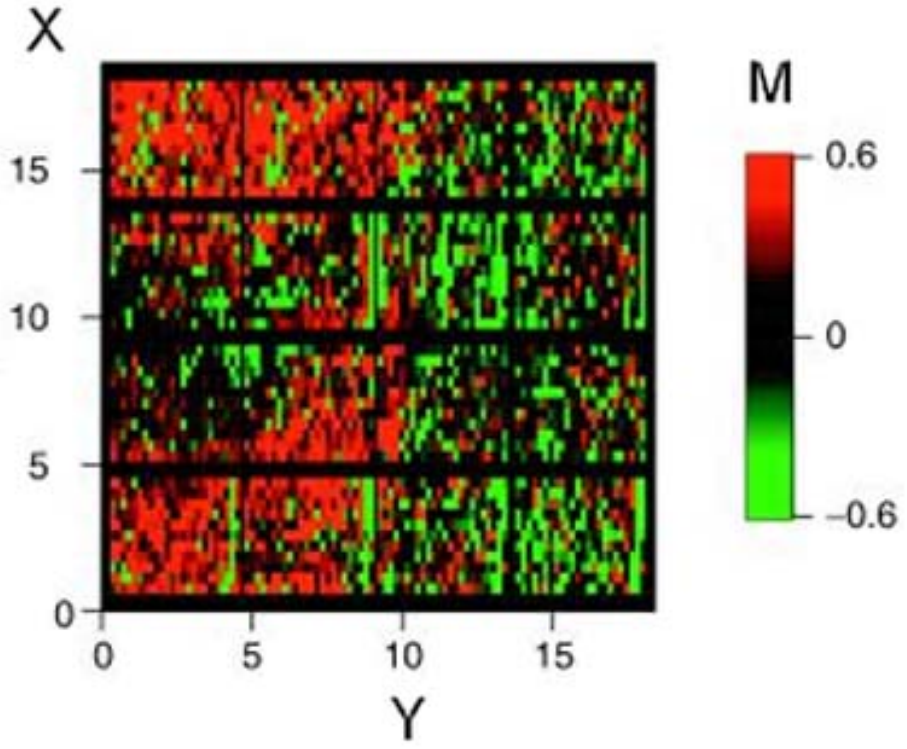
  - **A** = ($\log_2$ (Xred) + $\log_2$ (Xgreen)) / 2
    Intensidad

Experimento SELF-SELF real

MA plot

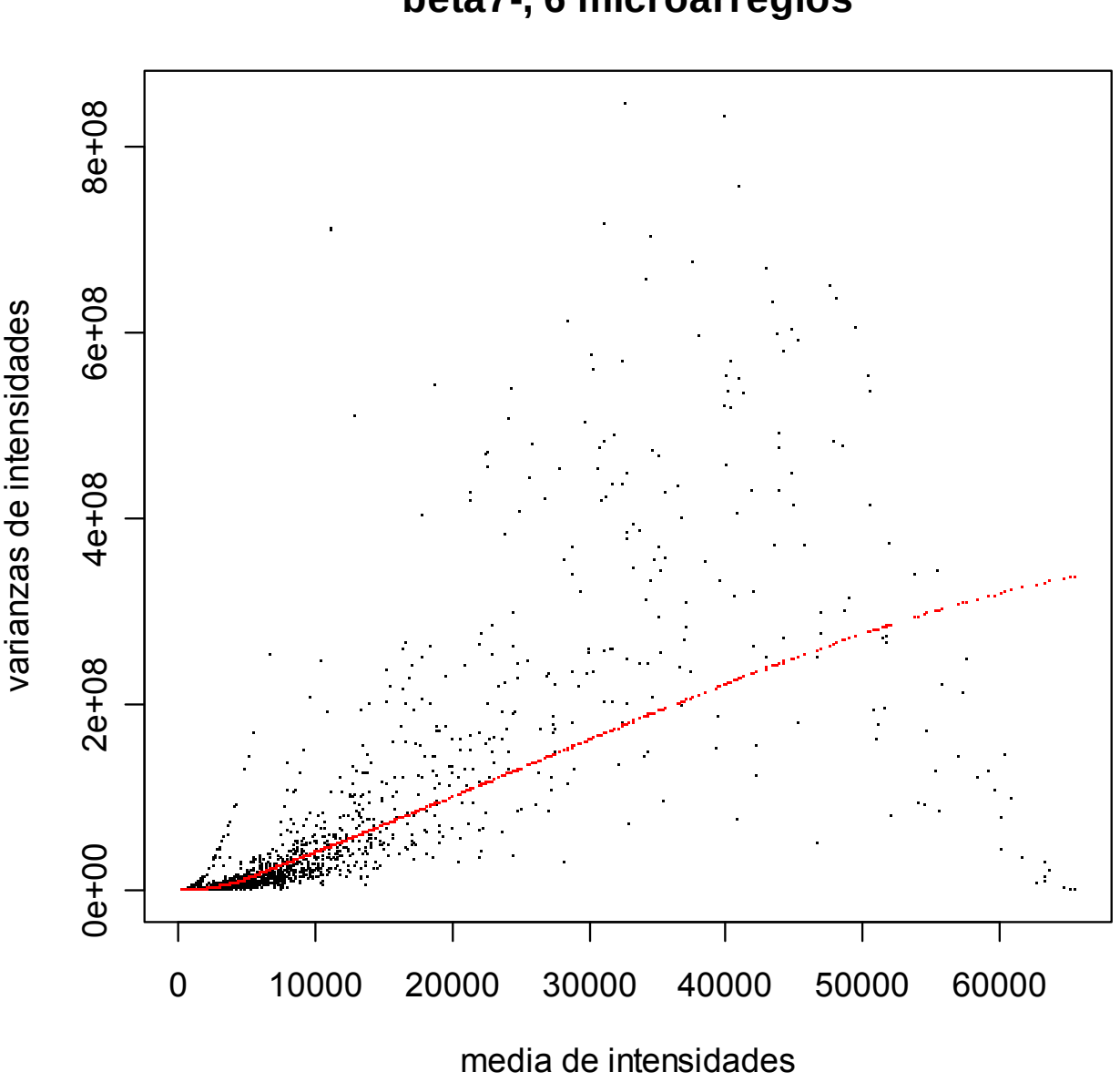


MXY plot



Sesgo dependiente de la intensidad

sesgo espacial


beta7-, 6 microarreglos
varianzas de intensidades
0e+00
2e+08
4e+08
6e+08
8e+08
0
10000
20000
30000
40000
50000
60000
media de intensidades

Transformaciones

Se utilizan para enfrentar los dos tipos de problemas

1. Las **estructuras curvas** no atribuibles a motivos biológicos de los gráficos M-A.
2. La dependencia de la **variabilidad** de las intensidades observadas con la **media**.

Modelo aditivo multiplicativo

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

Y_{ik} intensidad medida

X_{ik} intensidad verdadera

Para explicar y corregir

Modelo aditivo multiplicativo

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

Para explicar

- **Componente multiplicativa**: etiquetado, escaneado y características del spot.

- **Componente aditiva**: relacionada con el background local.

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

utilizando diferentes valores de los parámetros del modelo

Para corregir

Transformación arcoseno hiperbólico - arsinh

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

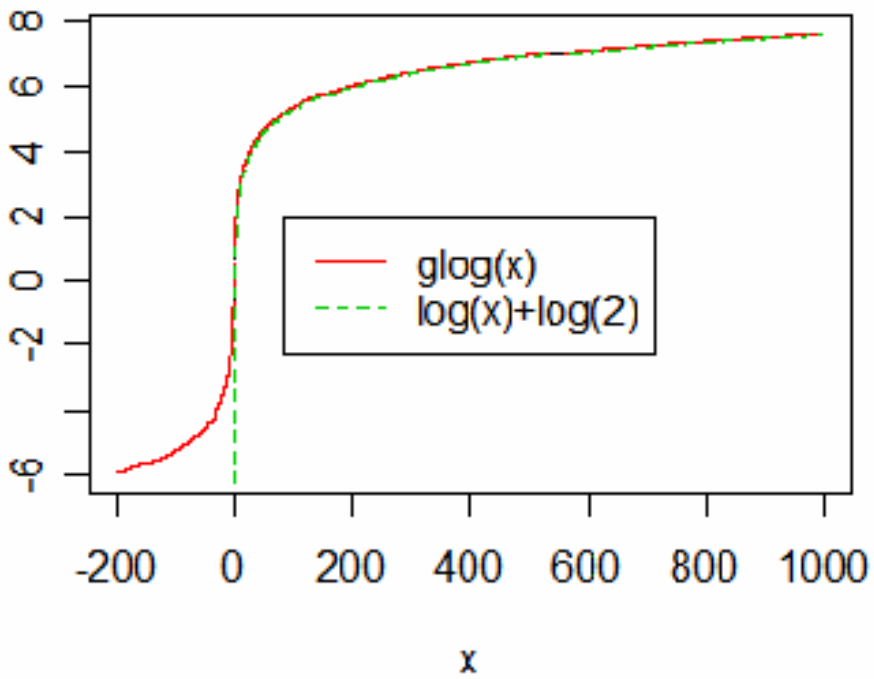
también llamada **logaritmo generalizado**

Munson (2001), Durbin *et al*. (2002) y Huber *et al*. (2002) propusieron en forma independiente la transformación (Rocke and Durbin, 2003 la indican glog) para

estabilizar la varianza de datos de micoarreglos que satisfacen el **modelo aditivo multiplicativo**.

glog es una función similar al logaritmo natural para valores grandes,

$$glog(x) \approx \log(x) + \log(2)$$



pero es menos empinada para valores pequeños

La transformación propuesta es la **composición** de una transformación **afín** (Bengtsson 2006) y el **glog**

$$Z_{ik} = g \log(\frac{Y_{ik} - a_i}{b_i})$$

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

- la **transformación afín** corrige las **curvaturas** de los gráficos MA

- la transformación **glog estabiliza las varianzas**

Aspectos estadísticos – cont.

Inferencia.

Procedimientos de tests simultáneos.

Generalmente respecto a qué genes están expresados diferencialmente.

Aspectos estadísticos -cont

Clustering y discriminación

(llamados **Clasificación** por “microarray biologists”**).**

Clases (categorias, etiquetas): pueden ser

muestras (1 - cientos)

o

genes . (10000 - 40000)

Clasificación

Es una de las primeras técnicas estadísticas utilizadas en el análisis de microarrays y es una de las preferidas.

El investigador tiene garantizada la obtención de un agrupamiento (clustering) de genes, sin importar

- el tamaño de la muestra,
- la calidad de los datos,
- el diseño del experimento o
- cualquier otra validez biológica que esté asociada con el agrupamiento.

Si es inevitable,
debería proveerse algún tipo de medida de reproducibilidad.

Aquellos procedimientos que **re-muestrean a nivel de caso** – más que a nivel de gen- todos tienen una performance razonable y ninguno es considerado el mejor.

¿Resultados?

¡Exitos!

- 1999 Science, Golub clasificó correctamente dos subtipos de cáncer utilizando únicamente solo niveles de expresión de genes.

Aunque estas formas de leucemia ya eran conocidas, la estrategia podría revelar subtipos no conocidos.

¡Exitos!

- 2001 Proceedings of theNational Academy of Sciences (PNAS)

Investigadores identifican cinco patrones de niveles de expresión de genes en cáncer de mama

y muestran que corresponden a diferentes tipos de enfermedades con diferente pronóstico.

¡Exitos!

Journal of Clinical Oncology 2006
Elaina Collie-Duguid, **PhD**, University of Aberdeen - Escocia,

halla un gen con un **fold-change de 50 veces** su nivel de expresión en pacientes de cáncer que

- **no respondían** a quimioterapia vs.
- **sí respondían**.

Ese gen codifica para una proteína que impide la muerte de una célula cancerígena.

¡Frustaciones!

Pero también, desde el advenimiento de la tecnología de microarreglos

- se ha acumulado una gran cantidad de frustración entre los biólogos que han dedicado sus esfuerzos en el seguimiento de direcciones falsas

- muchas publicaciones han sido desacreditadas

¡Frustaciones!

- Resulta difícil hallar estudios que apunten hacia algo concreto
- Falta de reproducibilidad de las listas de genes detectadas entre plataformas y laboratorios
- Dificultades en la validación de los resultados

¡Frustaciones!

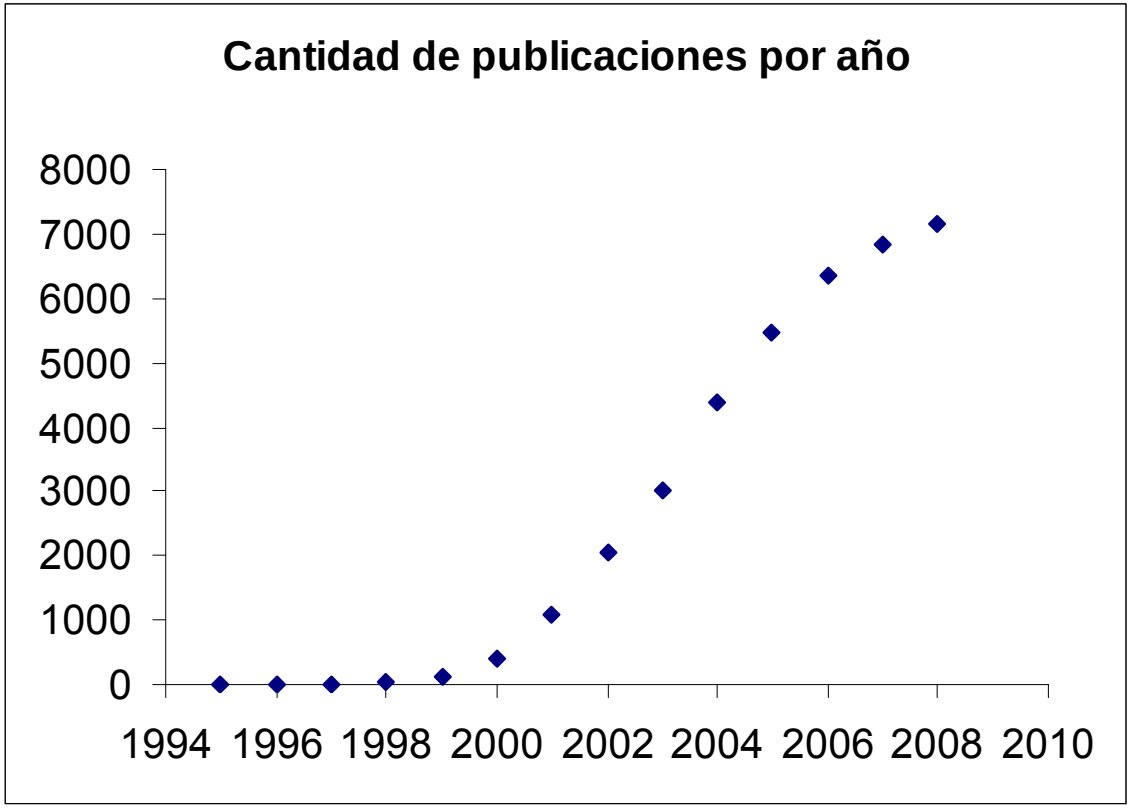
Como reflejo de esta frustración

- *Nature Reviews* 2005 "An Array of Problems"Frantz S
- *The Lancet* 2005 " Microarrays and molecular research: noise discovery?" John P.A. Ioannidis.
- Ruschhaupt M, et al., *Stat Appl Genet Mol Biol*, Jan 2004 llamó a los estudios de microarrays

"methodological wasteland"

"la tierra del desperdicio metodológico"

¡Frustaciones!



¿Causas?

- tecnología
- análisis de los datos

“In other contexts,
and possibly in these,

the results have been driven by study inadequacies rather than by biology.

Beware! (T. Speed 2005)”

- En Cobb (2006):
- Greg Engel, MD de Stanford: “la realización del experimento funciona bastante bien, cómo se analizan los datos sigue siendo muy confuso.”
- Golub: “El mayor desafío sigue siendo la interpretación de los datos.”
- Tibshirani: “Pienso que probablemente una buena proporción de los análisis de microarreglos están equivocados”

¿Alto nivel de ruido técnico en los datos?

Tendencia a explicar la notoria falta de potencia y la inestabilidad de los resultados del análisis de los datos,

por un alto nivel de ruido técnico en los datos.

¿Alto nivel de ruido en los datos?

Proyecto MAQC

Septiembre 2006 *Nature Biotechnology*

MicroArray Quality Control (MAQC) Consortium
Ha generado datos disponibles para el público.

Los datos permiten la evaluación de

- la repetibilidad dentro de un mismo sitio
- la reproductibilidad entre sitios
- la comparabilidad entre plataformas

Pero

¿Cuál es el nivel de ruido técnico en los datos?

"How high is the level of technical noise in microarray data?" Lev Klebanov and Andrei Yakovlev *Biology Direct* 2007

- reanalizan datos del estudio MAQC
- evalúan la magnitud de los errores de medición para la plataforma Affymetrix
- el **nivel de ruido** es estimado directamente a partir de **replicaciones técnicas** en ausencia de variabilidad biológica

Concluyen:

"al contrario de lo que habitualmente se cree, las fluctuaciones de los niveles de expresión causadas por el ruido técnico son bastante bajas y los resultados de la inferencia estadística para los datos de microarrays es despreciablemente pequeña"

Entonces, ¿cómo se explican los problemas?

Is there an alternative to increasing the sample size in microarray studies?
Lev Klebanov and Andrei Yakovlev

Bioinformation 1(10): 429-431 (**2007**)

Abstract:

Our answer to the question posed in the title is negative.

This intentionally provocative note discusses the issue of sample size in microarray studies from several angles.

We suggest that **the current view of microarrays as no more than a screening tool be changed** and **small sample studies no longer be considered appropriate.**

No es tan simple

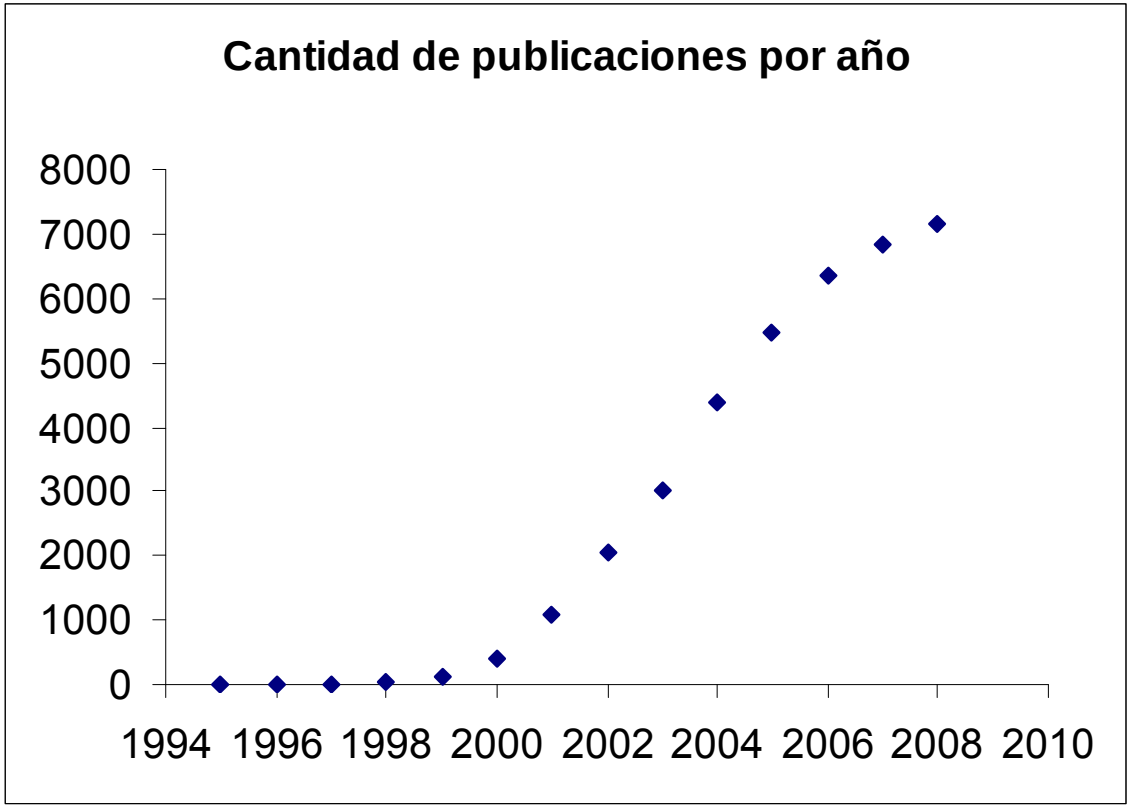
- no alcanza con conocer los genes para comprender su funcionamiento
- aumenta la evidencia en contra de la hipótesis que en general genes específicos controlan enfermedades o fenotipos específicos.

No es tan simple

Se pensaba que los genes actuaban en forma independiente
Pero
lo hacen en forma coordinada, comparten información e interactúan
Muchas enfermedades tienen componentes genéticos
Pero
no están causadas por un único gen

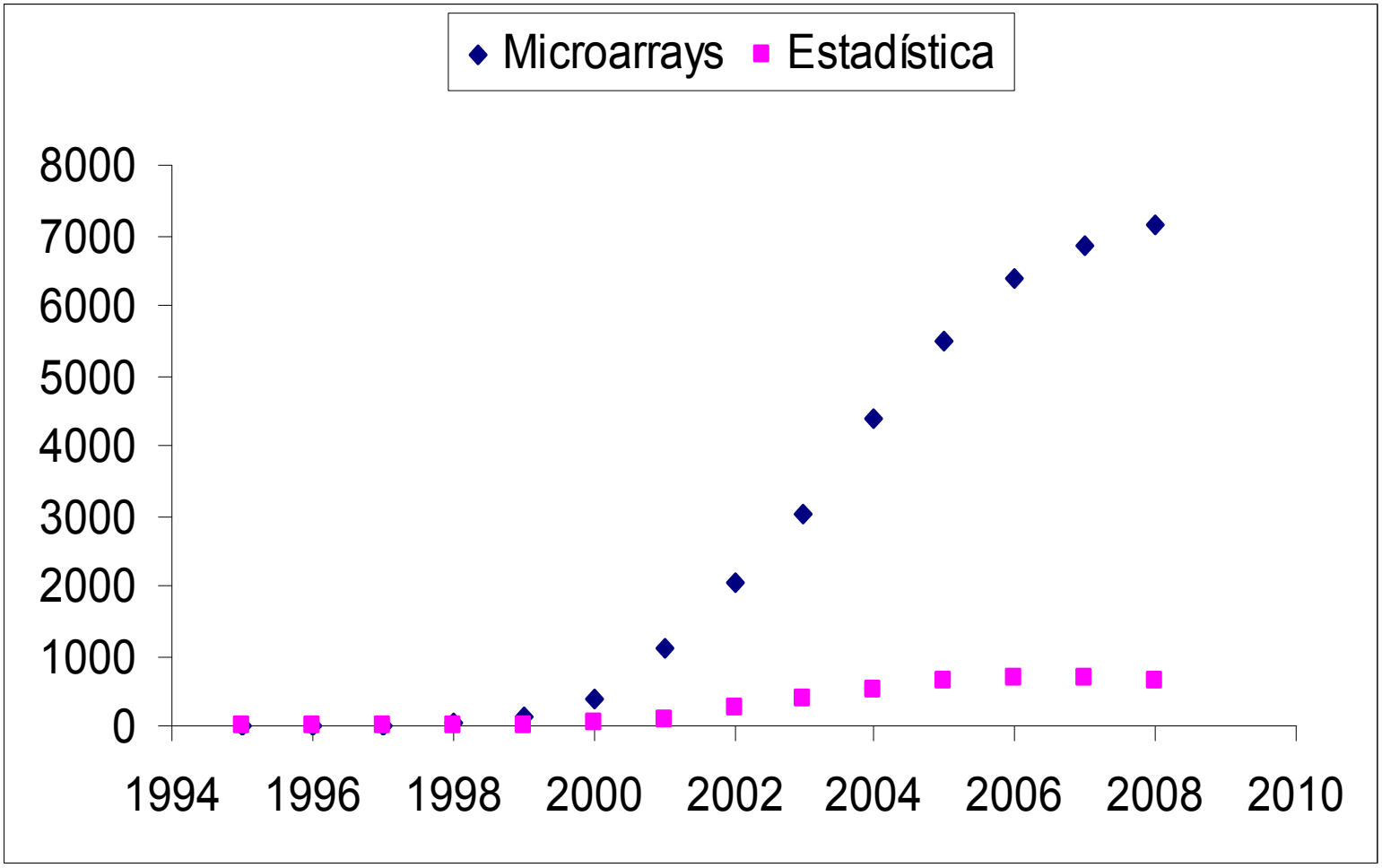
¡ No era tan fácil !

¿Estadística?



PubMed palabra clave microarray

Schena M,et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (1995)


Microarrays
Estadística
8000
7000
6000
5000
4000
3000
2000
1000
0
1994
1996
1998
2000
2002
2004
2006
2008
2010


¡MUCHAS GRACIAS!

BIOINFORMÁTICA

- ¿Una ciencia sin científicos?
- ¿Una gran cantidad de información procesada por las computadoras?

Recordemos

El genoma humano varía entre las personas

Muchas enfermedades están asociadas con alteraciones del genoma

Epigenética - ejemplo 1

- Diferenciación celular.
- Una única célula huevo fertilizada, a medida que se divide se transforma en muchos tipos de células: neuronas, músculos, piel, vasos sanguíneos, etc.
- Esto ocurre activando algunos genes y desactivando otros.

Epigenética – ejemplo 2

- Herencia epigenética
- Varones expuestos a hambrunas a edades cercanas a los 10 años tenían nietos (varones) por la línea paterna, con menor probabilidad de morir por enfermedades cardiovasculares.