

Buenas prácticas en investigación

Diseño & Análisis de datos

Dr. Cristóbal Fresno
cfresno@inmegen.gob.mx

Genómica Poblacional y Bioinformática
Instituto Nacional de Medicina Genómica

INMEGEN - Julio 2017

Agenda

Diseño de experimentos

- ① **Problemas típicos**
- ② **Buenas-Malas prácticas**

Análisis de datos

- ① **Control de calidad** - Búsqueda de sesgos
- ② **Análisis típicos**
- ③ **Validación de resultados**

Diseño de Experimentos

¿Para qué queremos un diseño?

- ① Porque es **NECESARIO** para abordar la pregunta!!
- ② ¿Cuál es **LA** pregunta?
- ③ ¿Qué efectos que pueden afectar al experimento?
 - ¿Qué **se puede controlar**?
 - ¿Qué **NO se puede controlar**?



Planificar!!!

- Explicar/discutir las ideas con los involucrados...
 - Actuales y futuras ...
- Explicar/discutir las ideas con los técnicos ...
 - ¿Qué tecnología? → Posibilidades
 - ¿Qué análisis? → Lo que **SI** y lo que **NO** se puede.
- **Hecho el experimento no hay vuelta atrás**



Identificando el problema tipo

Comparación de clases:

Caso vs Control



Predictión de clases:

¿Dónde lo/la asigno?



Descubrimiento de clases: ¿Cuántas hay?



Diseño experimental

Fuentes de variación

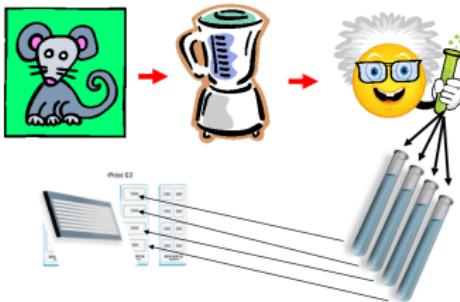
- **TODAS** las posibles: técnico, día, lote ...
- **Técnica!!!** Es imposible escapar de ella. Tener en cuenta.
- **Biológica!!!** Es justamente lo que buscamos.
- *Datos con variabilidad Técnica + Biológica.*
- Un buen diseño:
 - Disminuir lo técnico
 - Estimar razonablemente lo biológico.

Mitigación de las fuentes de variación técnicas

- 1 Aleatorización del experimento
- 2 Réplicas ¿Técnicas o biológicas?
 - Estimación adecuada del parámetro de interés
 - Inferencia poblacional de nuestra hipótesis

Réplicas ...

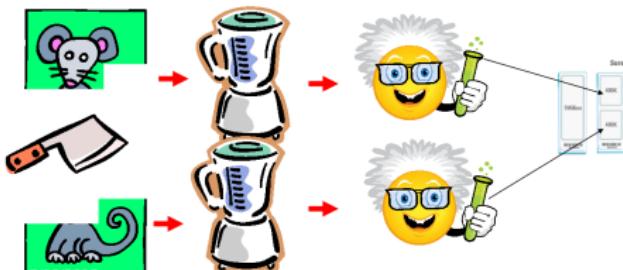
Técnica



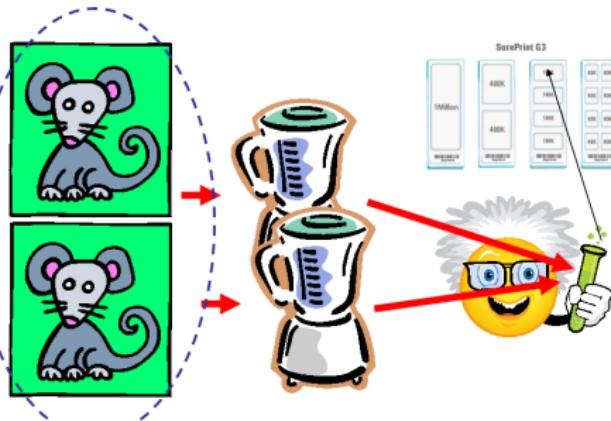
Biológica



Biológica pareada

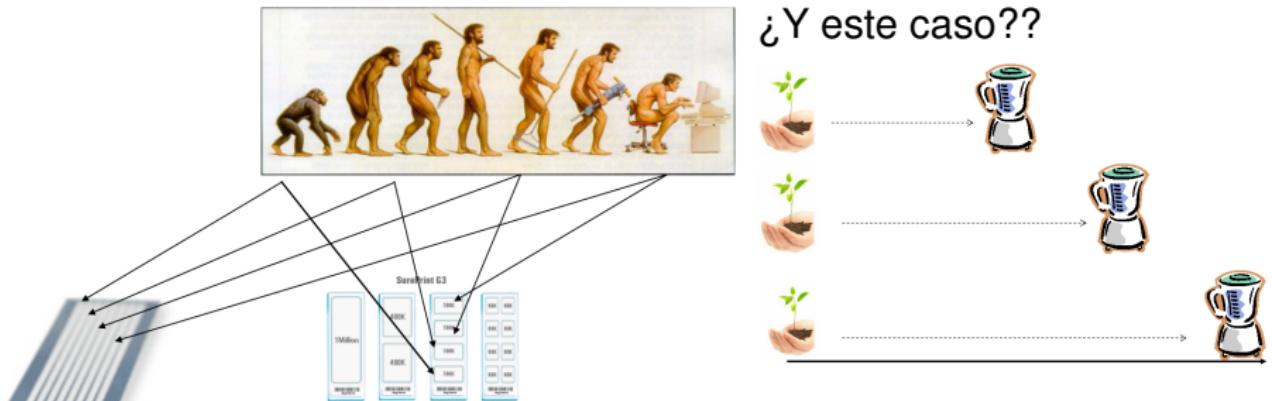


Biológica con pool



Experimentos longitudinales

Longitudinal → mismo individuo a lo largo del tiempo...

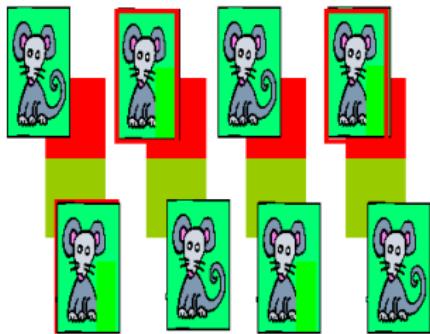


Las muestras se disponen para evitar confundimiento ...

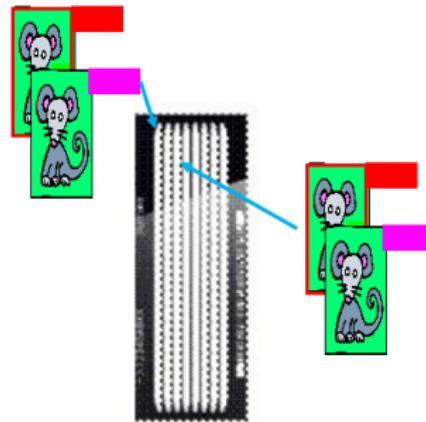
Dye-swap

Bloques completos

Boques incompletos

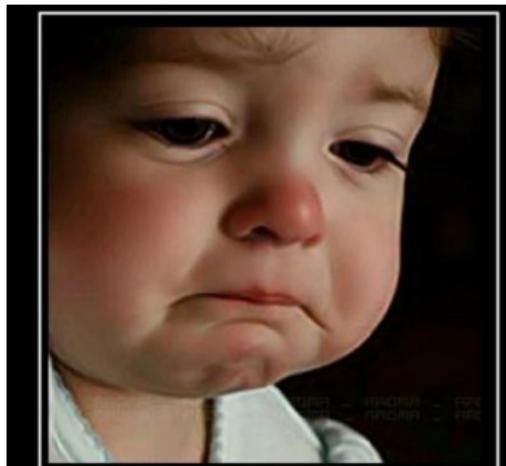


Multiplexado

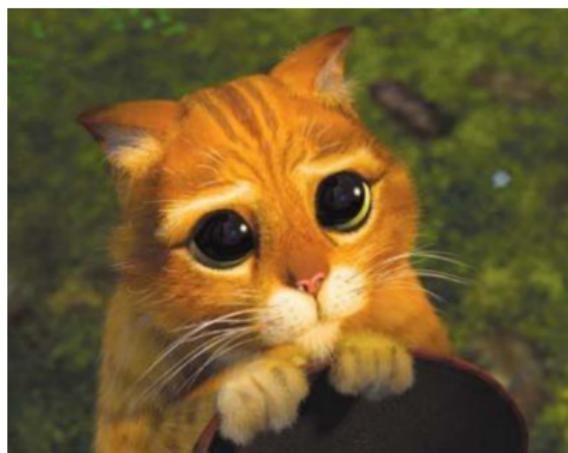


¿Cuántas Muestras?

- Considerar ...
 - Se está estimando un parámetro de una distribución!!!
 - Variación poblacional de los genes/proteínas??
 - Diferencia esperada?? $\log_2 FC \geq 1$?? $\log_2 FC \geq 2$??
- ↑ Muestras → ↑ Potencia de la prueba estadística



Pero solo puedo 2 muestras...



¿Dinero para más muestras?

Practice time!!!!



Buenas-Malas prácticas

Discusión...

En todos los ejercicios de al menos un **mal** y **buen** ejemplo por grupo:

- ① Diseñe un experimento caso vs control con una (1) réplica por condición.
- ② Diseñe un experimento caso vs control con cuatro (4) réplicas por condición en microarrays de cuatro (4) chips por vidrio.
- ③ 95 muestras a ser secuenciadas con 6-8 alicuotas pair end, en placas de ocho (8) canales y utilizando cinco (5) secuenciadores.

Buenas-Malas prácticas: ¿Qué opina de este diseño?

Discusión...



Algunas verdades...

- ① En **ausencia de un diseño apropiado**, es esencialmente **imposible discriminar la variación biológica de la técnica**.
- ② Cuando estas dos **fuentes de variación son confundidas**, no existe forma de conocer **cuál es la fuente de la cual provienen los resultados observados**.
- ③ **No existe algoritmo** por sofisticado que sea estadísticamente que permita separar factores confundidos una vez que se recolectaron los datos.
- ④ Un **diseño apropiado** debe asegurar que la *pregunta* puede ser respondida de forma *precisa*, dada las *restricciones experimentales*: costos de experimento, disponibilidad de muestra...

Diseño experimental

Discusión ...

Algunas verdades...

- ⑤ **Hecho el experimento no hay vuelta atrás**
- ⑥ Busque un **analista!!!** No busque un **forense/sepulturero!!**



Agenda

Diseño de experimentos

- ① **Problemas típicos**
- ② Buenas-Malas prácticas

Análisis de datos

- ① **Control de calidad** - Búsqueda de sesgos
- ② **Análisis típicos**
- ③ **Validación de resultados**

¿Qué estamos haciendo?



Muy sencillo...

Tenemos datos que procesamos y de los resultados publicamos!!!

Si tan sólo fuera así de fácil → SNI XXXX

¿Qué estamos haciendo?



Datos

Donde?
De qué?
Cómo qué?
Por quién?
Para qué?
Integridad?
Consistentes?

...

Proceso

Entradas?
Datos?
Supuestos?
Paramétrico?
Método/s?
Algoritmo?
Salidas?

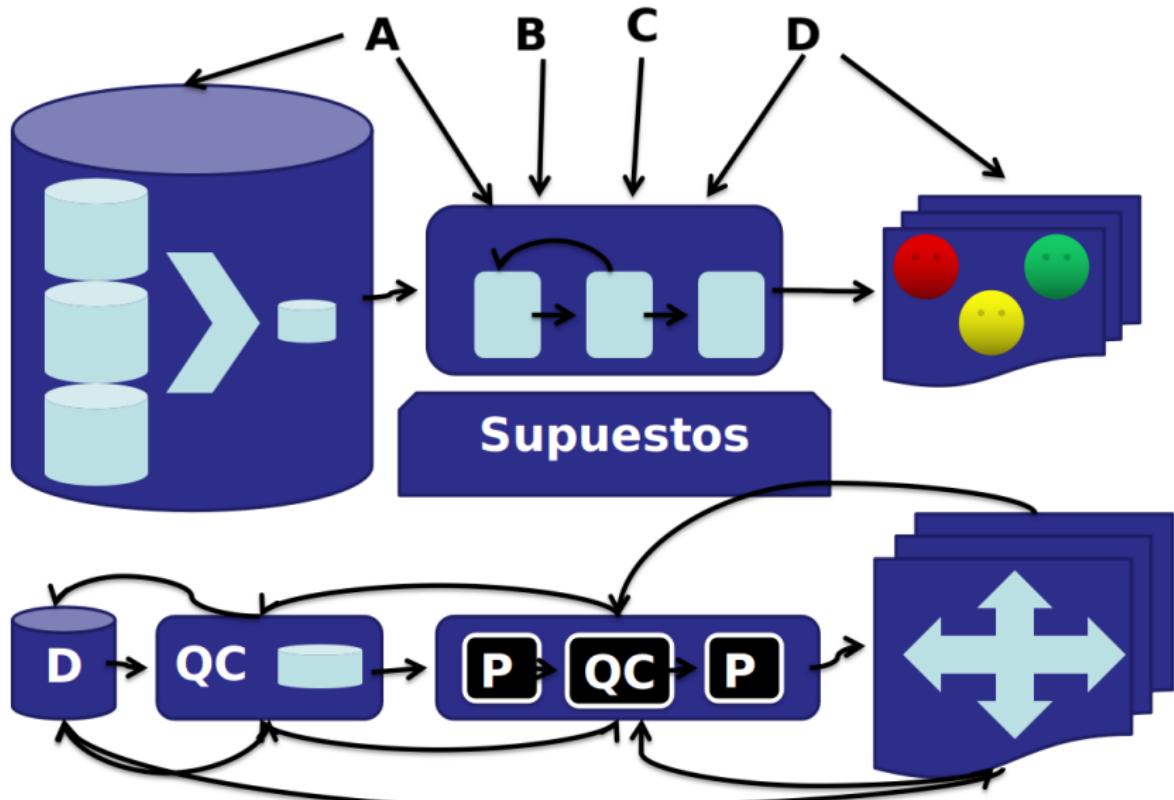
...

Resultados

Ahora que?
Sirven??
OK?
Integridad?
Hipótesis?
Congreso?
Publico?

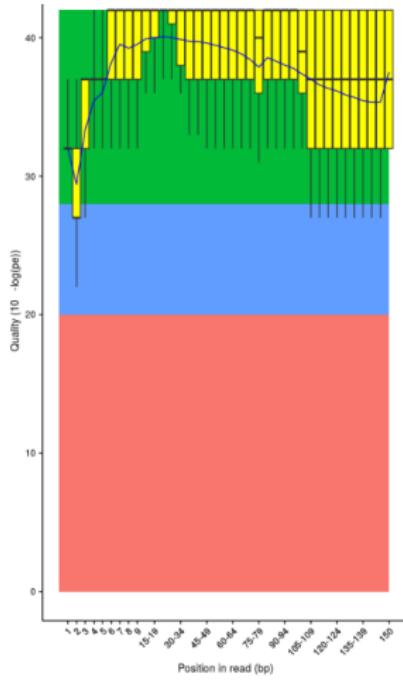
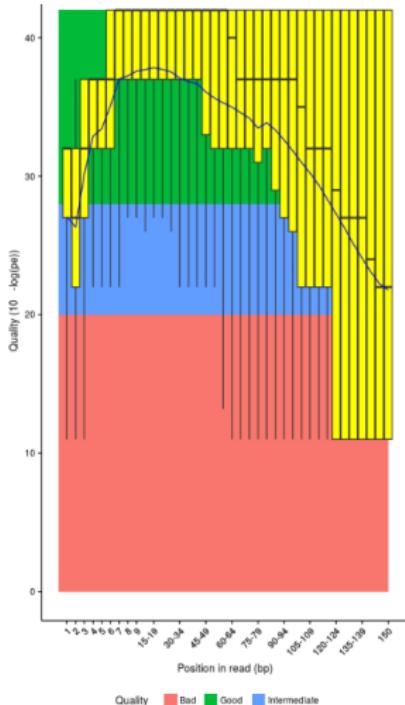
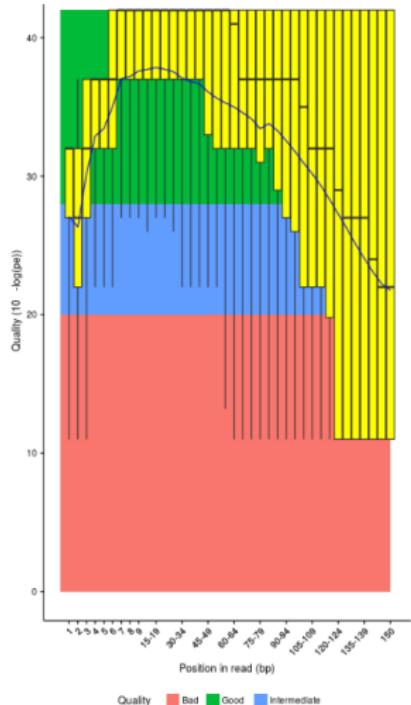
...

En realidad...



Ejemplo de calidad de secuencias

Datos crudos, procesados, re-procesados ¿Cuáles usamos?



Ejemplo de calidad en búsqueda de variantes

Contexto:

Se realizó llamado de variantes y por esas cosas de la vida se controló la integridad luego de tener los VCFs.

Control de integridad de referencias

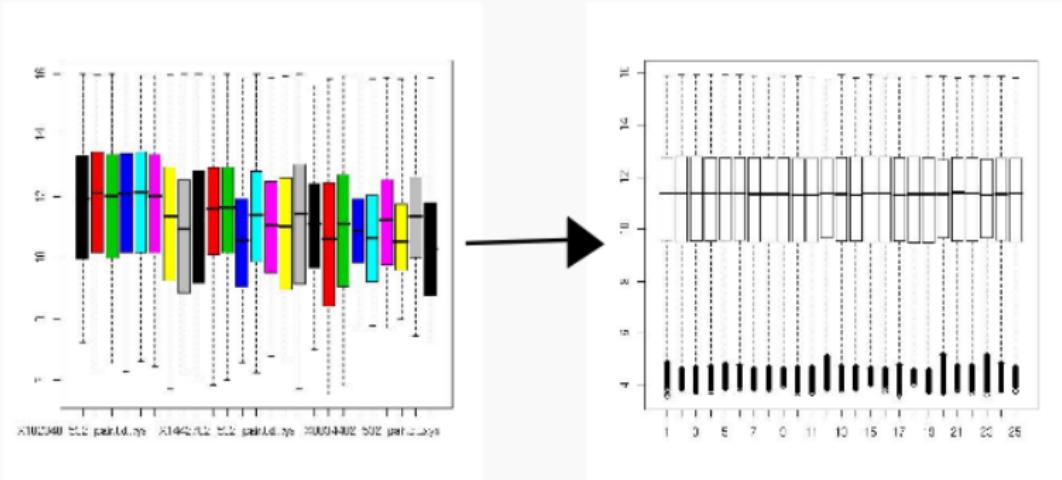
```
$ sha256sum -c referenciaHG38.sha256
hg38/Homo_sapiens_assembly38.dict: OK
hg38/Homo_sapiens_assembly38.fasta.64.alt: OK
hg38/Homo_sapiens_assembly38.fasta.fai: FAILED
hg38/Homo_sapiens_assembly38.fasta.gz: FAILED
```

¿Qué hacemos al respecto? ...

¿Sugerencias?

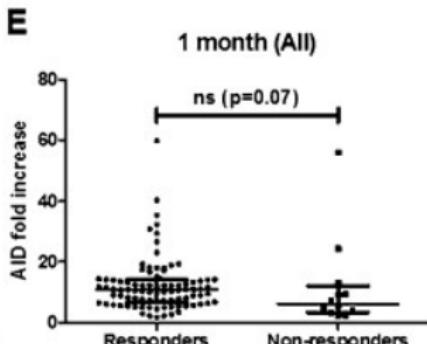
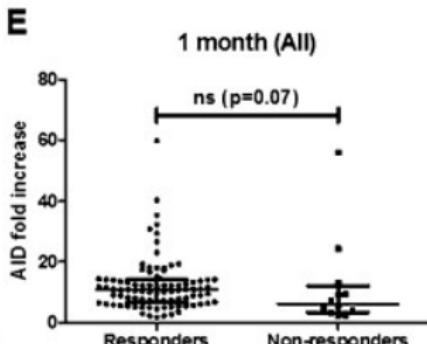
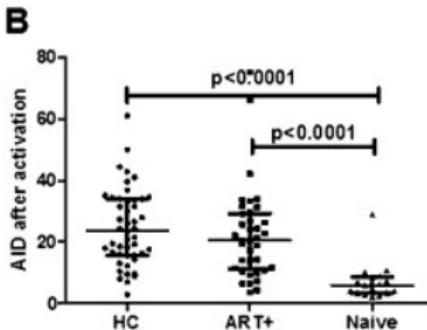
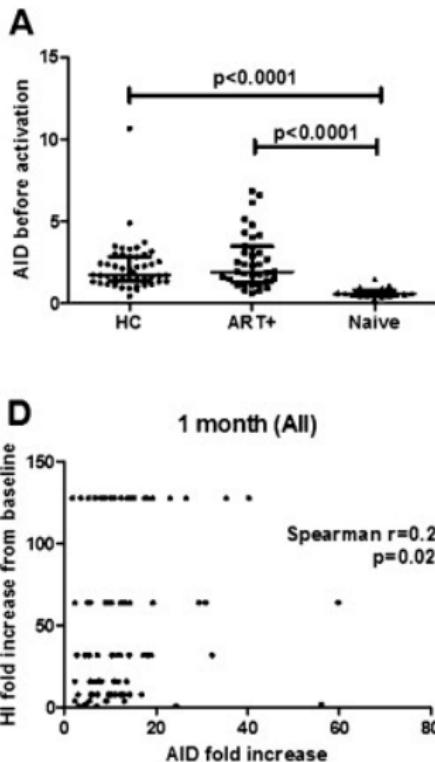
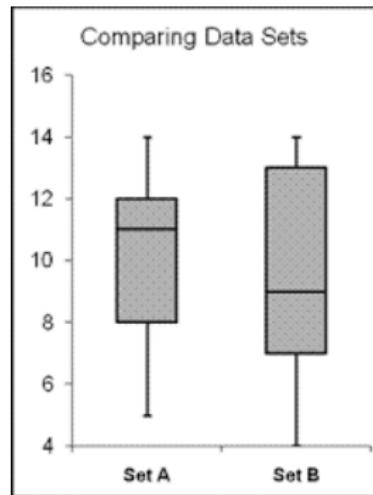
Ejemplo de expresión de transcriptos

Datos original y luego de normalizar ¿Para qué?



Ejemplo de expresión diferencial

¿Qué prueba es adecuada? T-test, Welch, ANOVA, Mann Whitney...



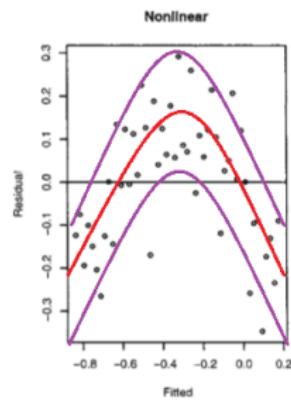
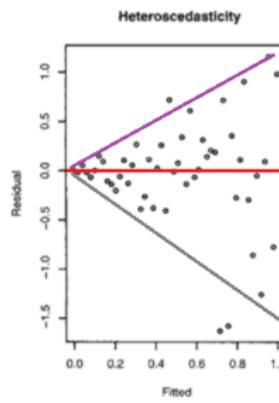
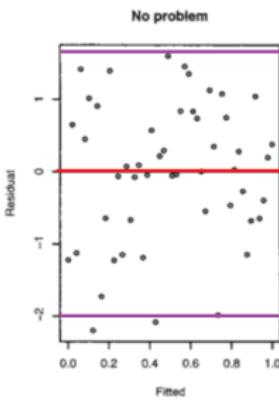
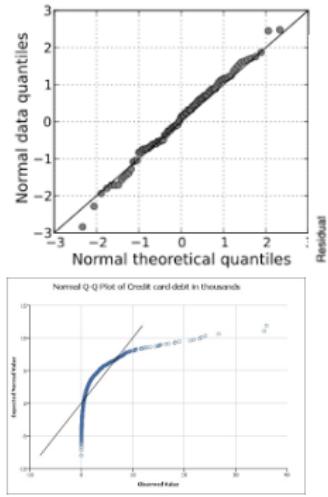
Ejemplo de Análisis de la Varianza (ANOVA)

Evaluación de supuestos

El modelo lineal

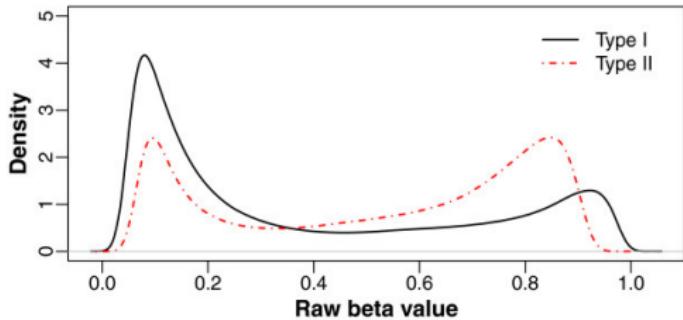
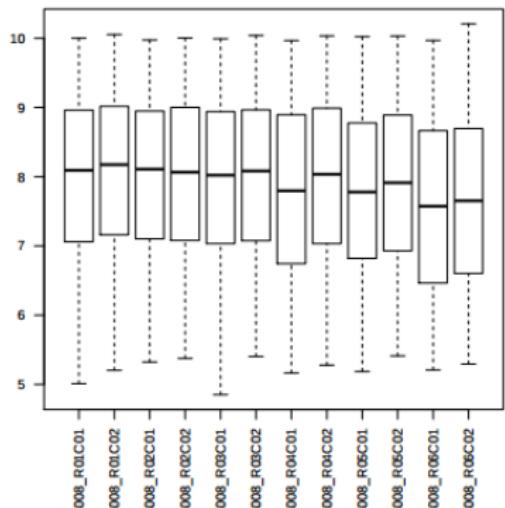
$$y_{ij} = X\beta = \beta_0 + \beta_1 x_{1j} + \dots + x_{ij}\beta_i + \dots + x_N\beta_N + \varepsilon_{ij}, \varepsilon \sim N(0, I\sigma^2)$$

donde el i ésimo gen de la j ésima muestra posee expresión y_{ij} ; matriz de diseño x_j , coeficientes β_i y error aleatorio ε_{ij} .



Ejemplo de metilación

Supongamos 6 casos vs 6 controles : ¿Se puede usar modelo lineal o prueba-t?



Boxplots de las 12 muestras. Distribución de muestra promedio de cada condición experimental.

① Control de Calidad de los datos de TCGA

① Global:

- ① Biotipos y expresión
- ② Profundidad

② Sesgos

- ① Gene Length
- ② GC content
- ③ RNA content

③ Hagamos algo!!! → Normalización

② Expresión diferencial.

③ Análisis funcional.

Datos de cancer de mamas

- **Platform:** UNC (IlluminaHiSeq_RNASeq)
- **Muestras:** Tumor (780), Sanas (101)
- **Partida:** level 3 raw counts (20.532 genes).

Archivo ejemplo:

gene	raw_counts	MNL	RPKM
? 100130426	0	0.0000	0.00
? 100133144	74	3.0604	0.61

Symbol | EntrezID (date, version)?

Anotación asociada a las muestras

Ensembl: Genes 80, Homo sapiens genes (GRCh38.p2)
(Chromosome Name, Gene start, Gene end, %GC content,
Gene type, Entrez Gene ID, HGNC Symbol, HGNC ID(s))

	Chr	Start	End	GC	Type	EntrezID	HGNCID	Symbol	Length
1	MT	577	647	40.85	Mt_tRNA	NA	HGNC:7481	MT-TF	70
2	MT	648	1601	45.49	Mt_rRNA	NA	HGNC:7470	MT-RNR1	953
3	MT	1602	1670	42.03	Mt_tRNA	NA	HGNC:7500	MT-TV	68

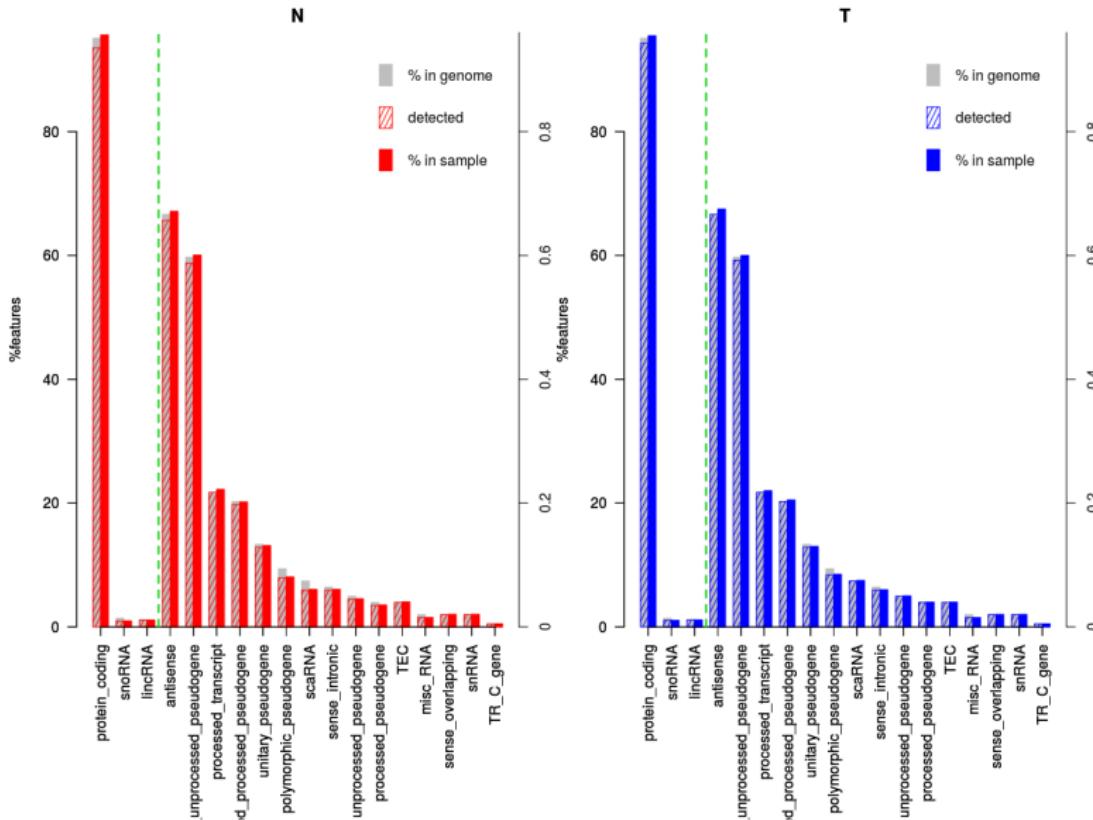
Filter criteria

- Biomart Filter:** Conventional Chr + EntrezID & Symbol
- Merge criteria:** Symbol TCGA = Ensembl & ↓ GC.

19.449 final genes

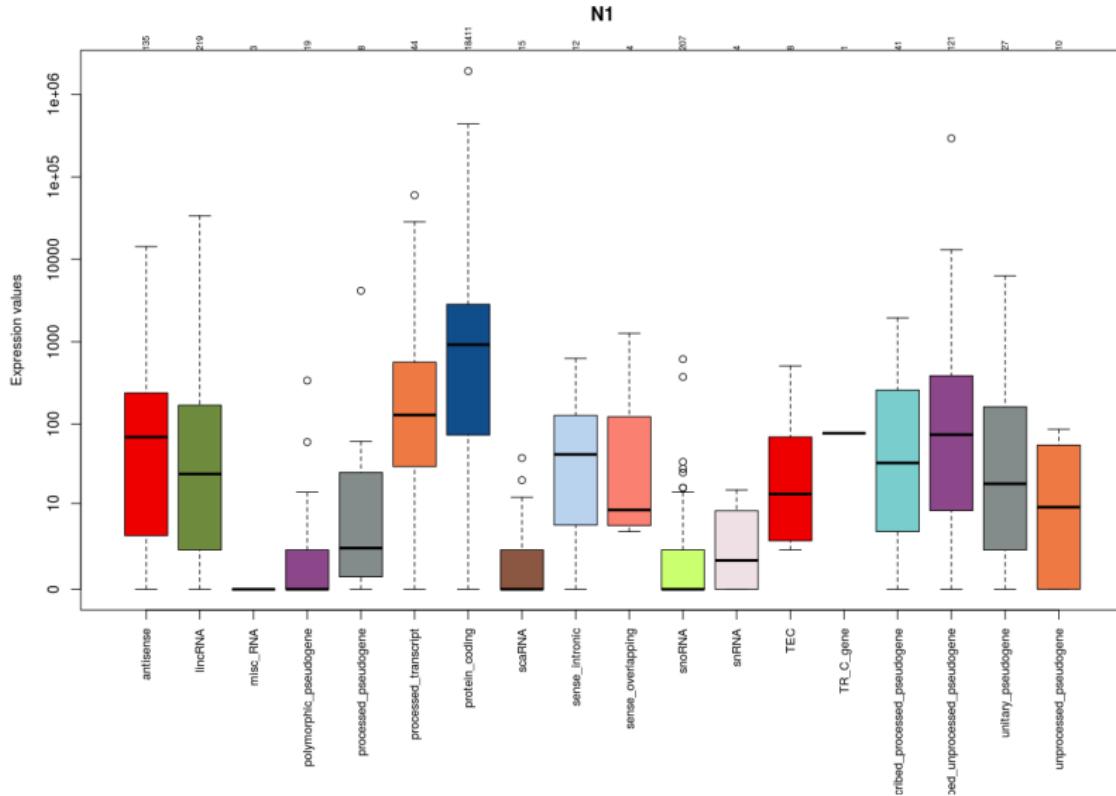
QC global: Biotipos I

Protein coding?



QC global: Biotipos II

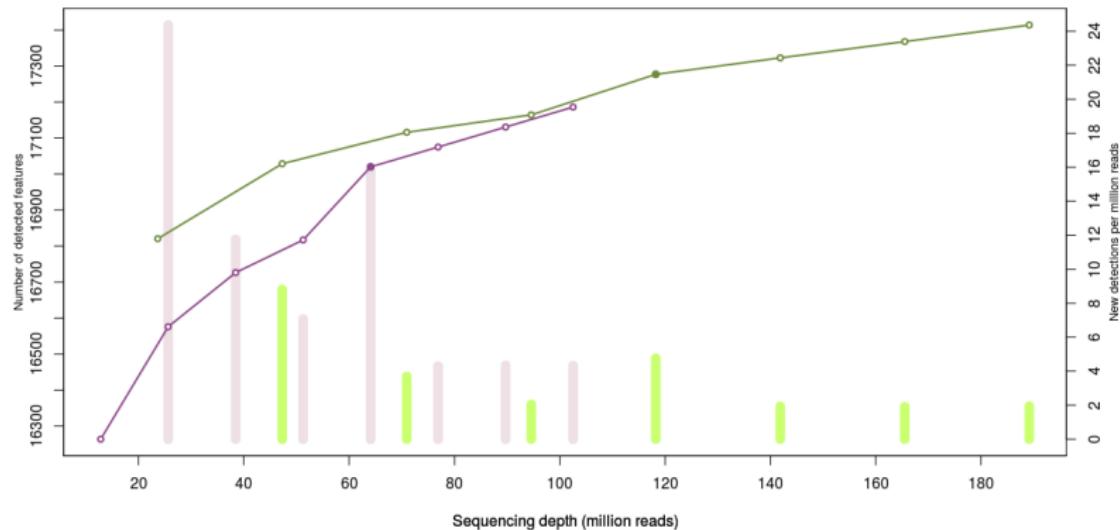
Counts per biotype. Qué hay de la expresión?



QC global: Saturation plot

Cuánto vemos?

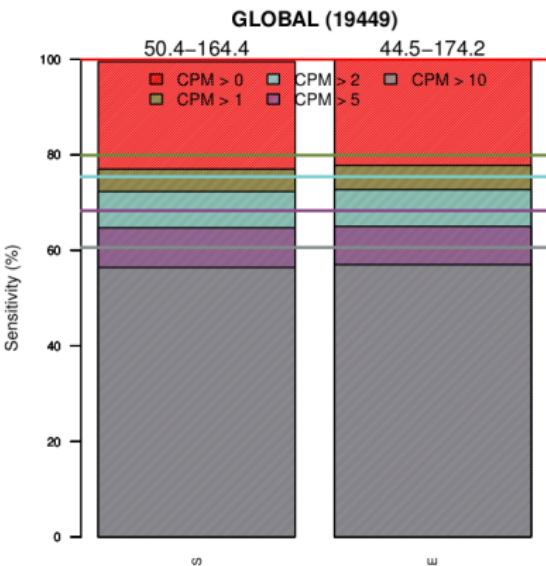
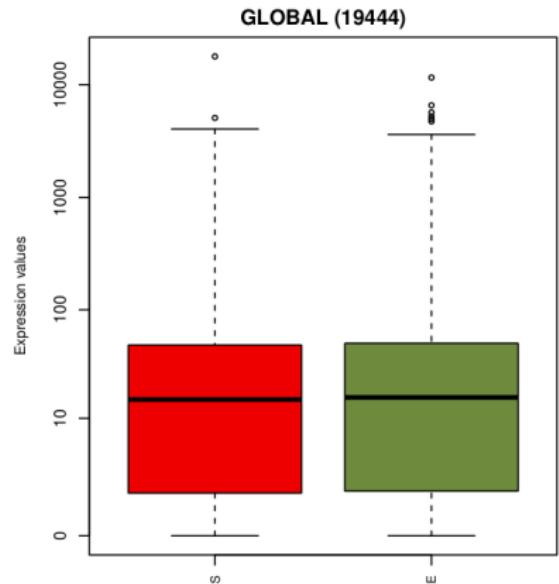
PROTEIN_CODING (18494)



	Left axis	Right axis	%detected
N1	•		92
N12	•	█	93.4

QC global: Biotipos II

Expresión global buena o mala?

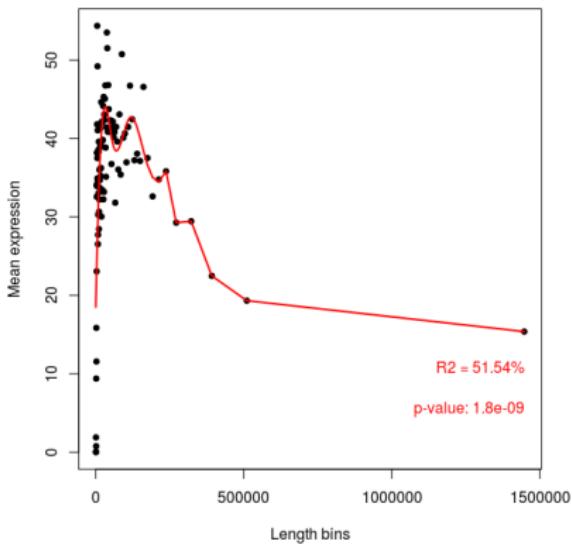
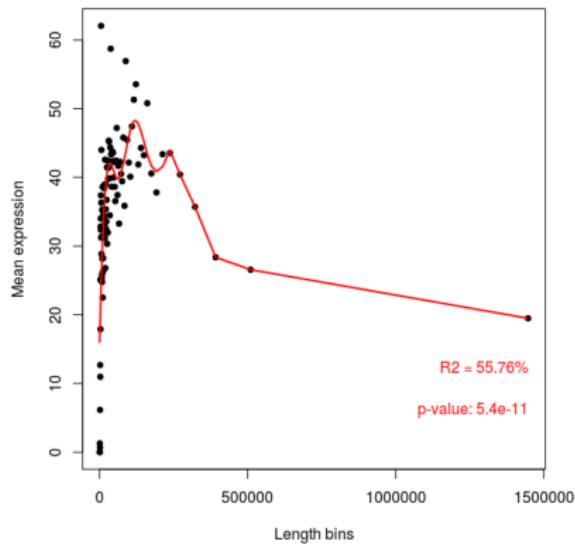


Sesgos: Gene length

Hay patrón?

N

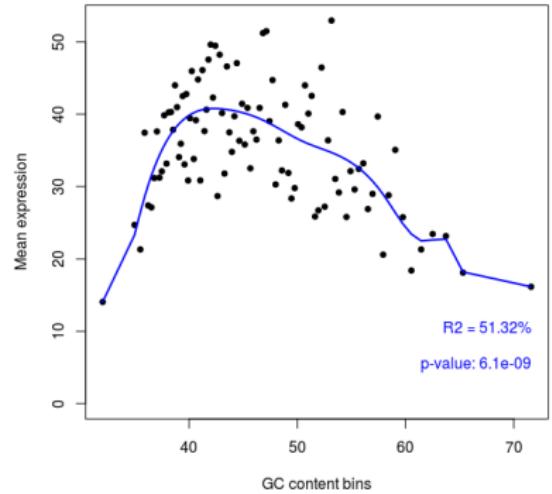
T



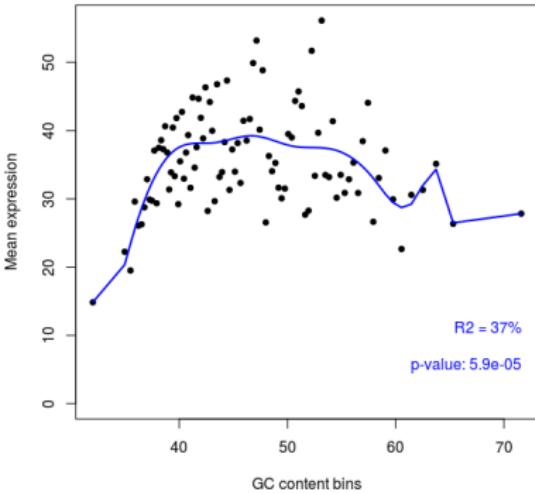
Sesgos: Contenido de GC

Hay patrón?

N

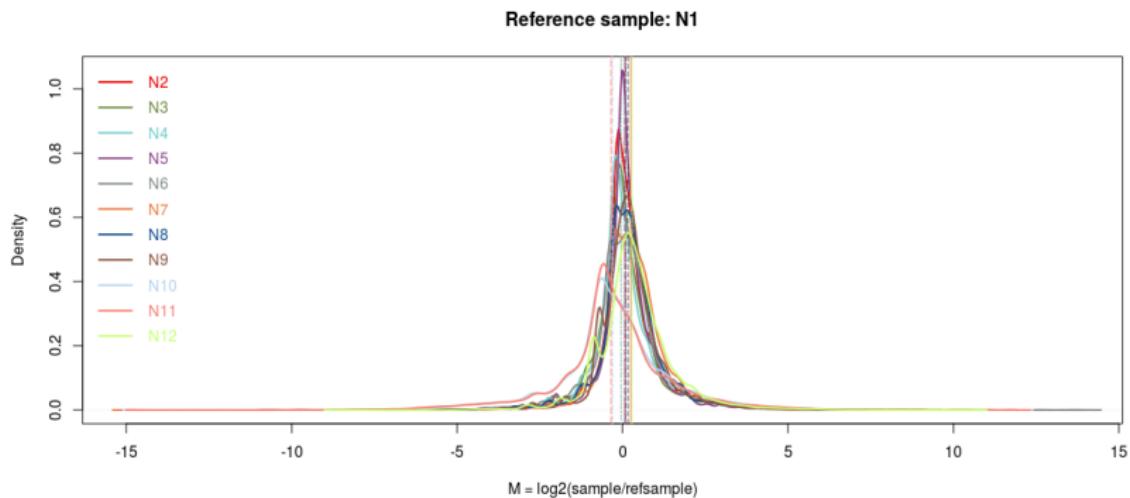


T



Sesgos: Contenido de RNA

Son comparables las muestras?



Normalización

Intra-muestra

Corregir efectos específicos como longitud de gen o contenido de GC

- RPKM
- Loess regression (EDASEq)
- Upper quantile (EDASEq)
- Median (EDASEq)
- Full quantile (EDASEq)
- Conditional Quantile Normalization (cqn)
- RNASeqBias
- NOISeq

Entre-muestras

Corrección de error sistemático por profundidad, composición, etc.

- TC
- RPKM
- TMM
- Upper Quantile
- Median
- Quantile Normalization
- cqn
- RLE (DESeq)
- NOISeq

Dillies MA et al. **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Briefings in bioinformatics*, 2013, 14(6), 671-683.

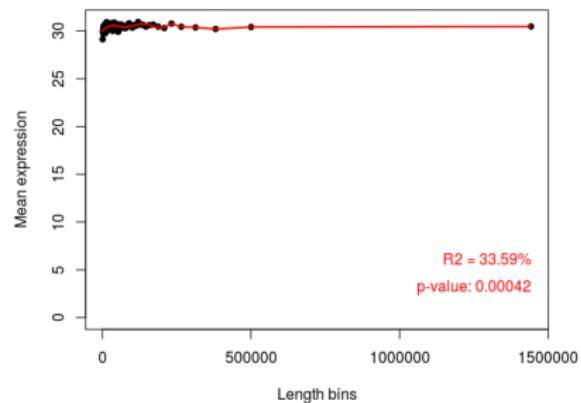
Paquetes de Bioconductor

- **edgeR** (**Robinson et al., 2010**): Trimmed Mean of M values (TMM).
- **DESeq2** (**Anders & Huber, 2010**): DESeq normalization method.
- **limma** (**Smith, 2004**): Quantile normalization.
- **NOISEq** (**Tarazona et al., 2011**): TC, RPKM, Upper Quartile and Trimmed Mean of M values (TMM).
- **EDASeq** (**Risso et al., 2011**): Loess robust local regression (within lane normalization, e.g. for GC-content) and global-scaling (median, upper quartile, full-quantile) for both within and between lane normalization.
- **CQN** (**Hansen and Irizarry, 2012**): Based on Poisson model where length or GC effects are incorporated as smooth functions using natural cubic splines and estimated using robust quantile regression, together with full-quantile between lanes normalization.
- **RNASeqBias** (**Zheng et al., 2011**): Generalized Additive Model to correct for gene length or GC content.

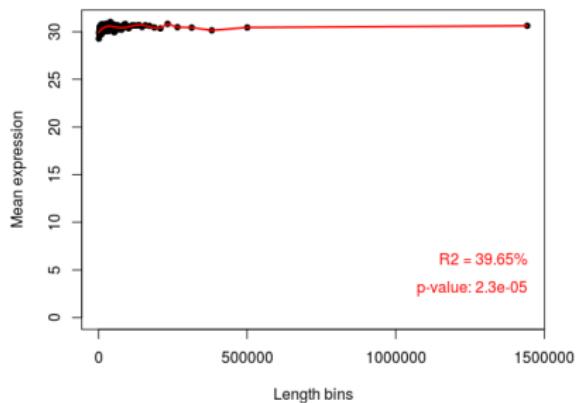
Control de remoción de sesgo de longitud de gen

Expresión normalizada en escala \log_2

N



T

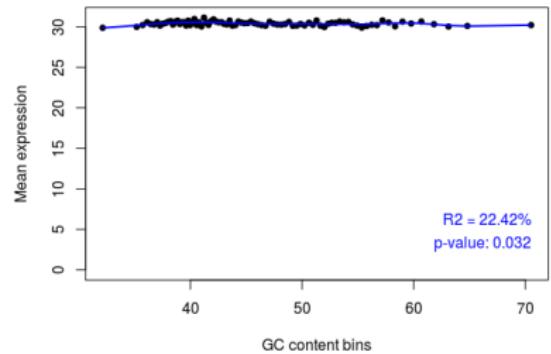


Se removió el efecto?

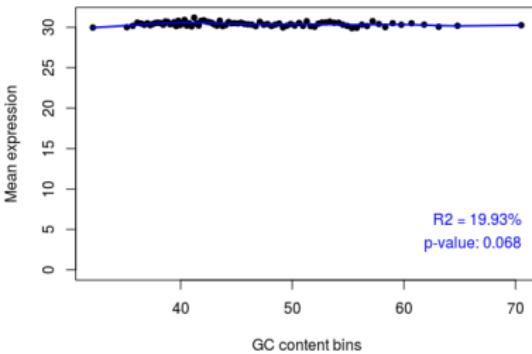
Control de remoción de sesgo de contenido de GC

Expresión normalizada en escala \log_2

N



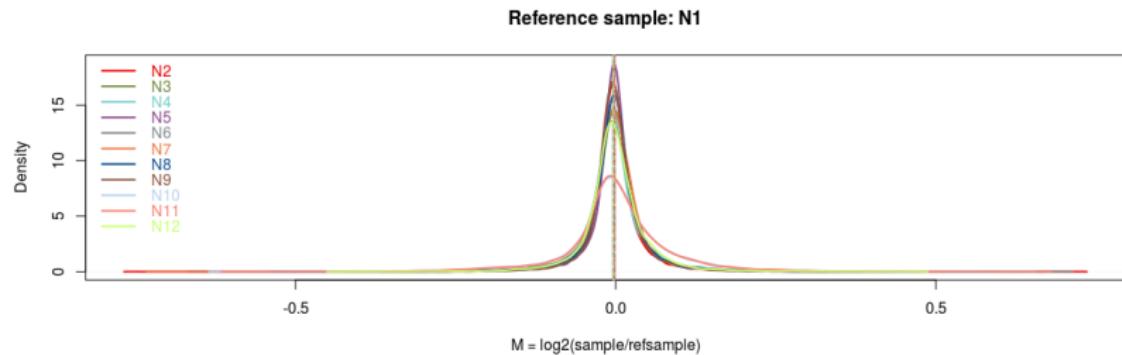
T



Se removió el efecto?

Control de remoción de sesgo de contenido de RNA

Expresión normalizada en escala \log_2



Se removió el efecto?

Removiendo Artefactos I

Genes/transcriptos ruidosos

Low Count Filter with NOISeq

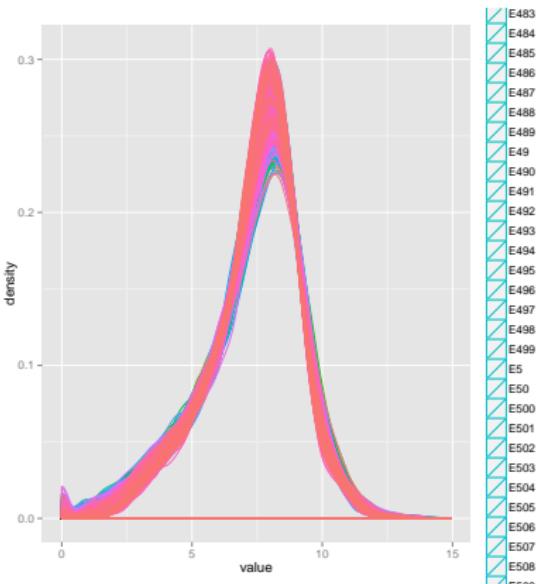
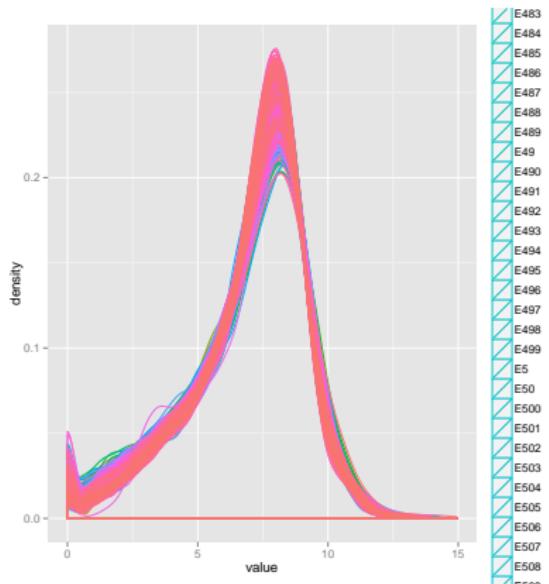
Remove low count features across all experimental conditions

CPM A cpm (counts per million) threshold is chosen by the user. By default, $\text{cpm} = 1$. A gene g is filtered out if the average CPM per condition is lower than cpm threshold for all the conditions.

Wilcoxon test The hypothesis to test for each gene and condition are $H_0 : m = 0$ versus $H_1 : m > 0$, where m is the median of CPM values per condition. Genes with p-value > 0.05 in all the conditions are filtered out. Only recommended when the number of replicates per condition is at least 5.

Proportion test Alternative to Wilcoxon test when few replicates per condition are available. $H_0 : p = p_0$ is tested versus $H_1 : p > p_0$, where $p_0 = \text{cpm}/10^6$. Genes with p-value > 0.05 in all conditions are filtered out.

Removiendo Artefactos I



15.281 genes con cpm > 10

Removiendo Artefactos II

Efecto de lote

Existe un efecto de lote?

Una exploración multivariada con Análisis de Componentes Principales (**PCA**) or **clustering** permite la detección de efectos de lote y estudiar cuán bien nuestras muestras se adecuan al diseño experimental.

Cómo remover/reducir el efecto de lote?

- **ASCA**: Nueda et al. Bioinformatics, 2007, 23(14), 1792-1800.
- **Imdme**: Fresno et al. Journal of Statistical Software, 2014, 56(7), 1-16
- **ComBat**: Johnson et al. Biostatistics, 2007, 8(1), 118-127.
- **ARSyN**: Ferrer & Conesa. Biostatistics, 2011, kxr042.

Basados en microarrays → **transformación log**

Reducción de ruido sistemático con ARSyN

La idea por detrás... se conoce el diseño experimental

Tenemos un modelo lineal para cada gen

$$x_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\alpha\beta\gamma)_{ijr}$$

En forma matricial podemos escribir

$$\mathbf{X} = \mathbf{1m}^T + \mathbf{X}_a + \mathbf{X}_b + \mathbf{X}_{ab} + \mathbf{X}_{abg}$$

Así descomponemos con PCA en

$$\mathbf{X} = \underbrace{\mathbf{1m}^t + \mathbf{T}_a \mathbf{P}_a^t + \mathbf{E}_a}_{\dots} + \underbrace{\mathbf{T}_b \mathbf{P}_b^t + \mathbf{E}_b}_{\text{PART I: Signal of interest}} + \underbrace{\mathbf{T}_{ab} \mathbf{P}_{ab}^t + \mathbf{E}_{ab}}_{\text{PART II: Residuals}} + \underbrace{\mathbf{T}_{abg} \mathbf{P}_{abg}^t + \mathbf{E}_{abg}}$$

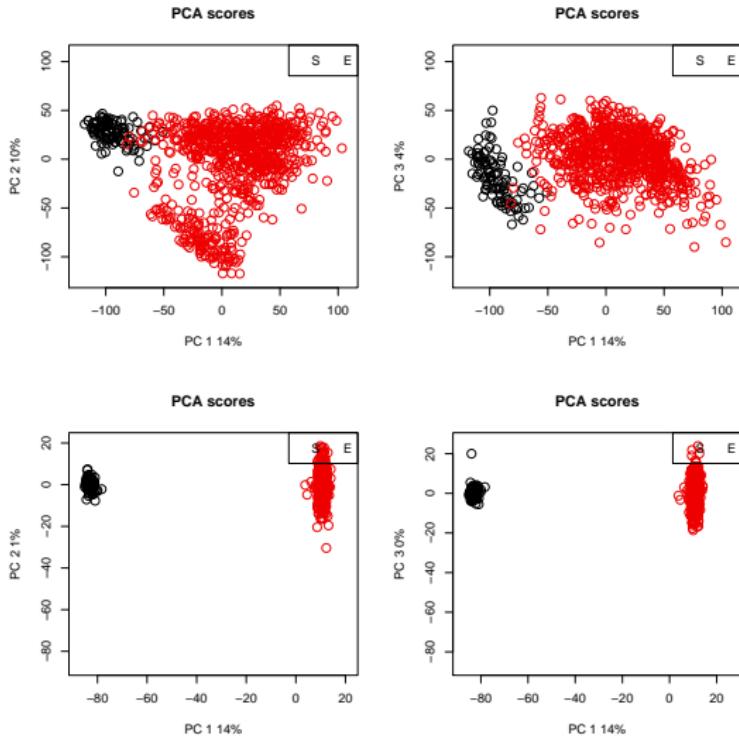
Finalmente nos quedamos con

$$\tilde{\mathbf{X}} = \mathbf{X} - \underbrace{\mathbf{E}_a + \mathbf{E}_b + \mathbf{E}_{ab}}_{\text{Noise of the signal}} - \underbrace{\mathbf{T}_{abg} \mathbf{P}_{abg}^t}_{\text{Signal of the noise}}$$

Ferrer & Conesa 2011

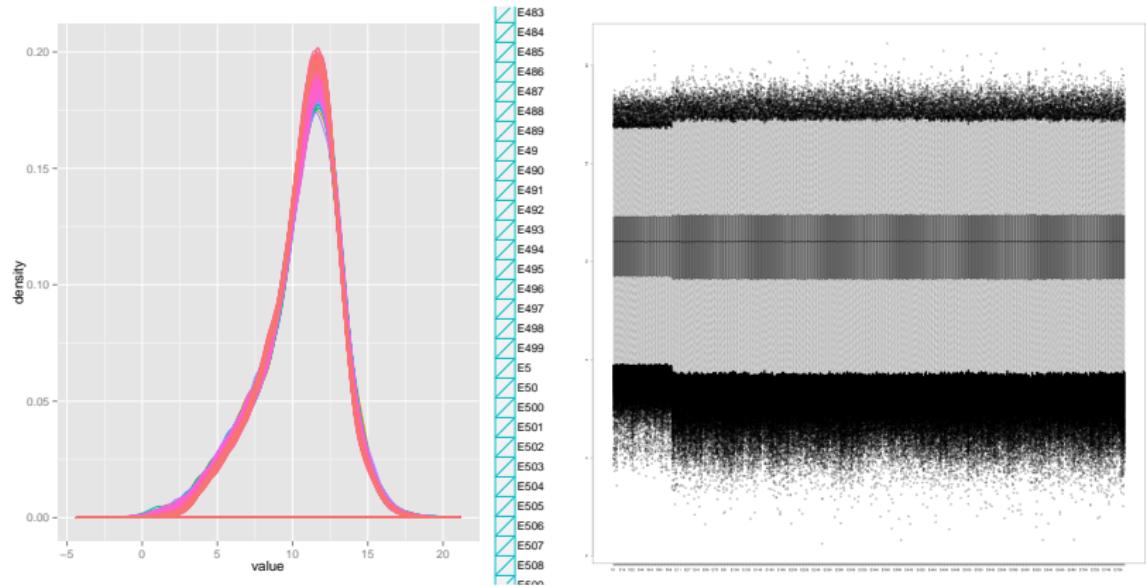
Reducción de ruido sistemático con ARSyN

PCA original y filtrado...



Reducción de ruido sistemático con ARSyN

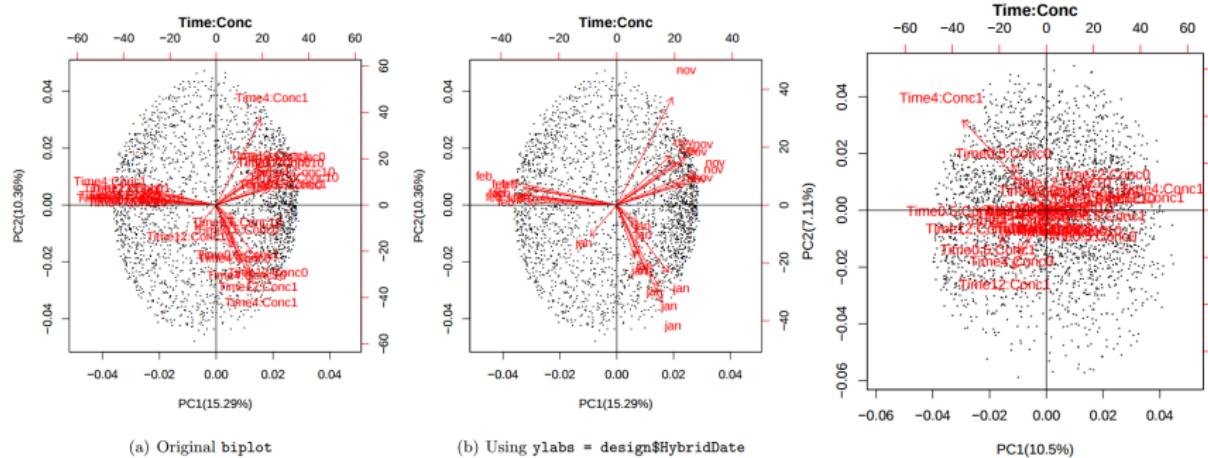
Densidades finales



Finalmente ya podemos trabajar

Ejemplo con paquete Imdme

Caso tiempo x concentración: Efecto de lote por mes de hibridización



Tres (3) grupos de Tiempo X Concentración asociado a:
NOV, FEB y JAN Fresno et al. 2014

Expresión diferencial con DESEQ2

Distribución Binomial Negativa

El modelo lineal generalizado

Distribución

- $Y_{ij} = Bi^-(\mu_{ij}\alpha_i)$, $\mu_{ij} = s_j q_{ij}$, $\sigma_{ij}^2 = \mu_{ij}(1 + \mu_{ij}\alpha_i)$

Predictor lineal

- $\eta = x_j \beta_i = \beta_{0i} + \beta_{1i} x_{1j} + \dots + x_j \beta_i + \dots + x_N \beta_N$

Función de enlace

- $link(\mu_{ij}) \approx link(q_{ij}) = \log_2(q_{ij}) = \eta$

donde el i ésimo gen de la j ésima muestra posee Y_{ij} conteos; μ_{ij} es el número esperado de conteos; α_i el parámetro de dispersión; s_j el coeficiente de profundidad de cada muestra y $q_{ij} = \hat{\mu}_{ij}$ es la estimación poblacional; η es el predictor lineal con matriz de diseño x_j y vector de coeficientes β_i .

Expresión diferencial con DESEQ2

Resultados de la prueba de Wald

```
head(DESeqDEResults, n=2)
```

log2 fold change (MAP): Group T vs N

Wald test p-value: Group T vs N

DataFrame with 6 rows and 6 columns

	baseMean	log2FoldChange	lfcSE
	<numeric>	<numeric>	<numeric>
388795	13.39395	0.5950653	0.4862802
8225	3160.31835	-0.1962587	0.1255709

	stat	pvalue	padj
	<numeric>	<numeric>	<numeric>
388795	1.223709	2.210622e-01	3.175138e-01
8225	-1.562931	1.180688e-01	1.903220e-01

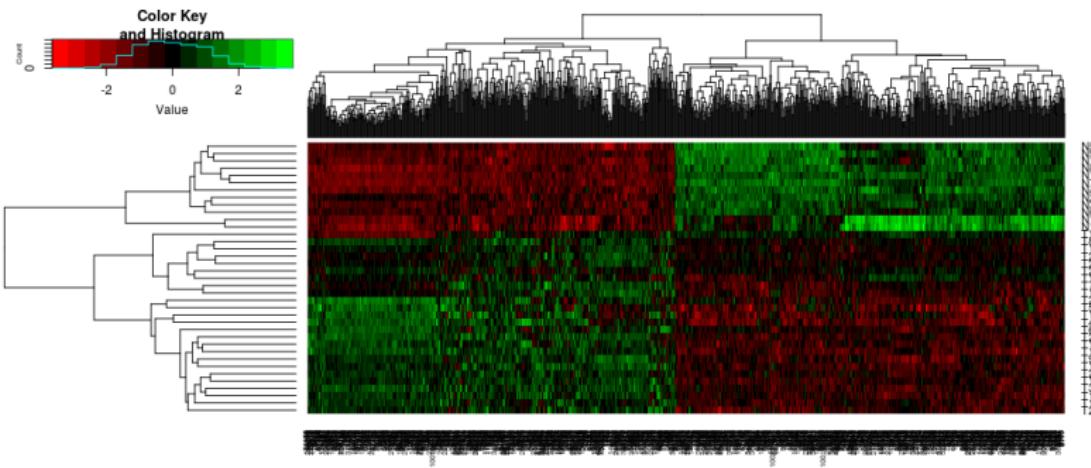
Expresión diferencial con DESEQ2

Resultados de la prueba de Wald

```
summary(DESeqDEResults)
```

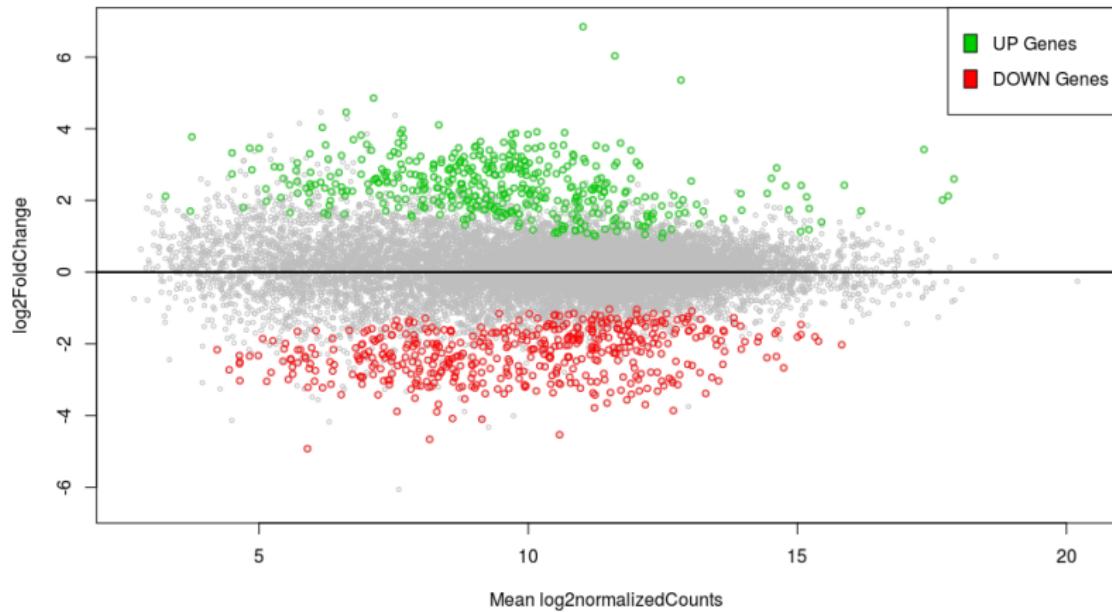
```
out of 17018 with nonzero total read count
adjusted p-value < 0.001
LFC > 0 (up)      : 2165, 13%
LFC < 0 (down)    : 1674, 9.8%
outliers [1]       : 950, 5.6%
low counts [2]     : 0, 0%
(mean count < 6)
```

Control de calidad de genes candidatos: Heatmap



Las muestras se agrupan por el diseño experimental??

Control de calidad de genes candidatos: MA plot



Hay algún patrón en los datos?

Qué términos están **enriquecidos** en nuestro experimento?

- **Set Enrichment Analysis** (paramétrico):
GOstats, DAVID, etc.
- **Gene Set Enrichment Analysis** (no-paramétrico):
GSEA, mGSZ, etc.

Table I

Two-by-two contingency table for flagged and unflagged genes in a GO category

	Flagged genes	Non-flagged genes	Total
In category	n_f	$n - n_f$	n
Not in category	$N_f - n_f$	$(N - n) - (N_f - n_f)$	$N - n$
Total	N_f	$N - N_f$	N

n_f is the number of flagged genes in the category, n is the total number of genes in the category, N_f is the number of flagged genes on the microarray, and N is the total number of genes on the microarray. All numbers are those obtained after dereplicating multiple instances of the same gene.



Zeeberg et al. 2003.

Análisis Funcional: GSEA

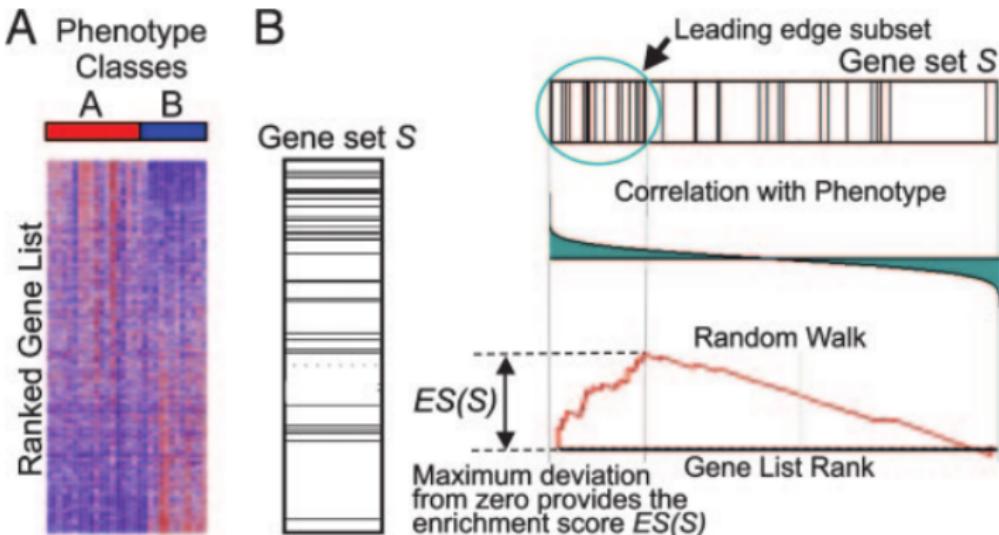


Fig. 1. A GSEA overview illustrating the method. (A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the “gene tags,” i.e., location of genes from a set S within the sorted list. (B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset.

Subramanian et al. 2005.