Paquete de R "alteredExpression": algoritmo para localizar genes con perfil de expresión alterado por una enfermedad

José Manuel Sánchez Santos, USAL
María Jesús Rivas López, USAL
Carlos Prieto Sánchez, CIC
Jesús López Fidalgo, UCLM
Javier de Las Rivas Sanz, CIC

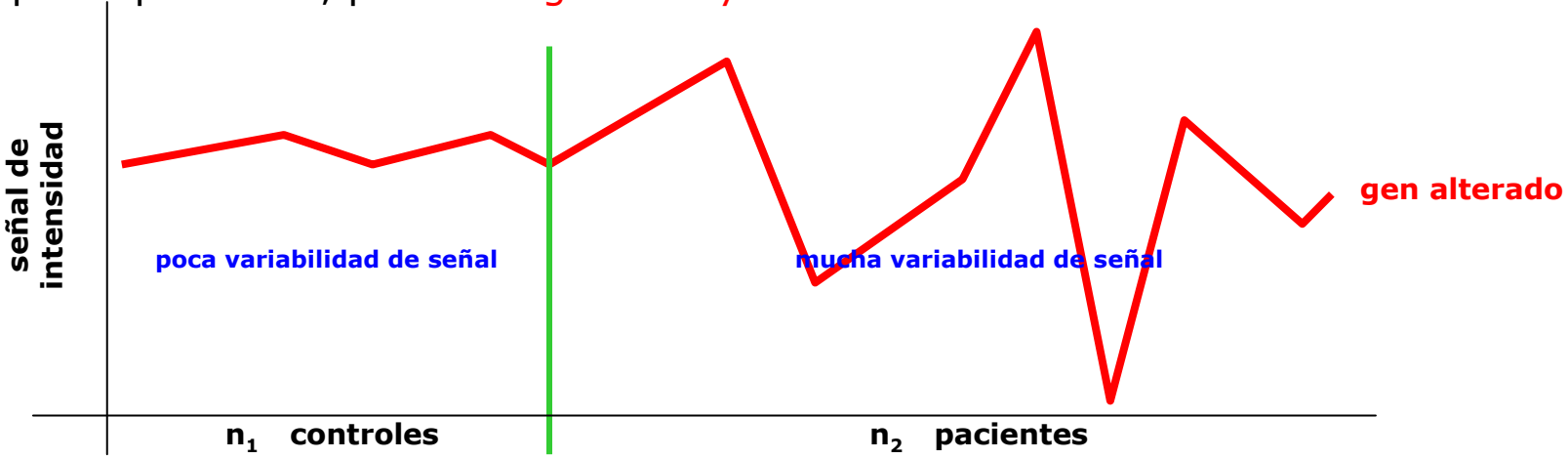
VNiVERSiDAD D SALAMANCA

	Individuos (microarrays)					
Genes	**1**	**…**	**n_1**	**n_1+1**	**…**	**$n_1+n_2=n$**
1	X_{11}	…	X_{1n1}	X_{1n1+1}	…	X_{1n}
2	X_{21}	…	X_{2n1}	X_{2n1+1}	…	X_{2n}
…	…	…	…	…	…	…
N	X_{N1}	…	X_{Nn1}	X_{Nn1+1}	…	X_{Nn}
Y_j	**1**	**1**	**1**	**2**	**2**	**2**

Herramienta para 2 muestras independientes (control/pacientes), con n_1 controles y n_2 pacientes.

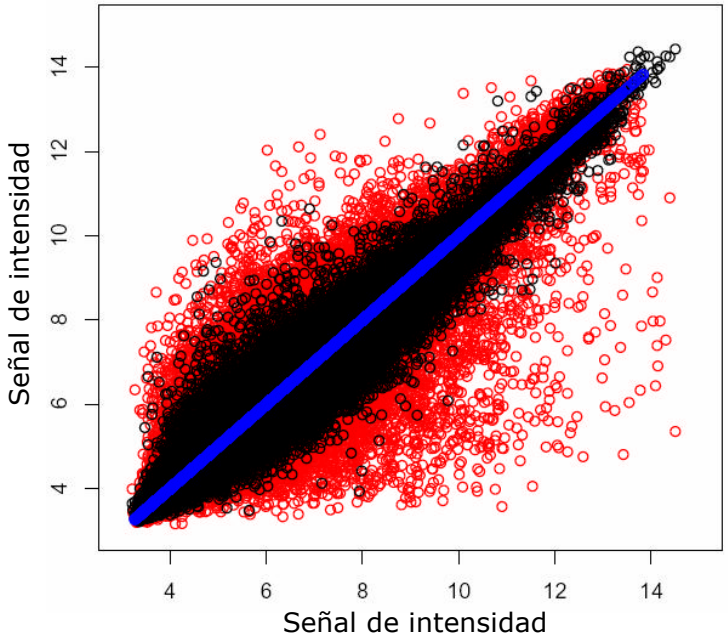
X_{ij} = señal de intensidad del gen i en el paciente j

OBJETIVO → Encontrar genes "desregulados" o con "expresión alterada", es decir, genes con señal casi constante en el grupo de control pero variable en el grupo de pacientes, por tanto genes cuya señal se altera con la enfermedad



Justificación → Variabilidad observada en la señal de intensidad:

Comparación de la señal de intensidad entre
3 Pacientes vs 3 Controles, 3 Controles vs 3 Controles, 1 Control vs si mismo



Variabilidad observada en los controles <
< Variabilidad observada en los pacientes

↓

Algoritmo basado en la comparación
de la ***Varianza Residual VR*** entre grupos

↓

Prieto, Rivas, Sánchez, Fidalgo
& de las Rivas (Bioinformatics2006)

Estadístico → Dado un grupo de genes G para un grupo de individuos S, se define la ***varianza residual***:

$$\mathbf{VR\ (G,\ S) = \sum_{G,S}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2 / (|\ G\ | - 1) \cdot (|\ S\ | - 1)}$$

Procedimiento:

Calcular, para un grupo de genes G, la VR del grupo de control S_C y la VR del grupo de pacientes S_P

$$VR(G,S_C) = \sum_{G,Sc}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2 / (|G| - 1) \cdot (|S_C| - 1)$$

$$VR(G,S_P) = \sum_{G,Sp}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2 / (|G| - 1) \cdot (|S_P| - 1)$$

Comparar ambas VR con la Varianza Residual Relativa VRR

$$VRR = VR(G,S_C) / VR(G,S_P)$$

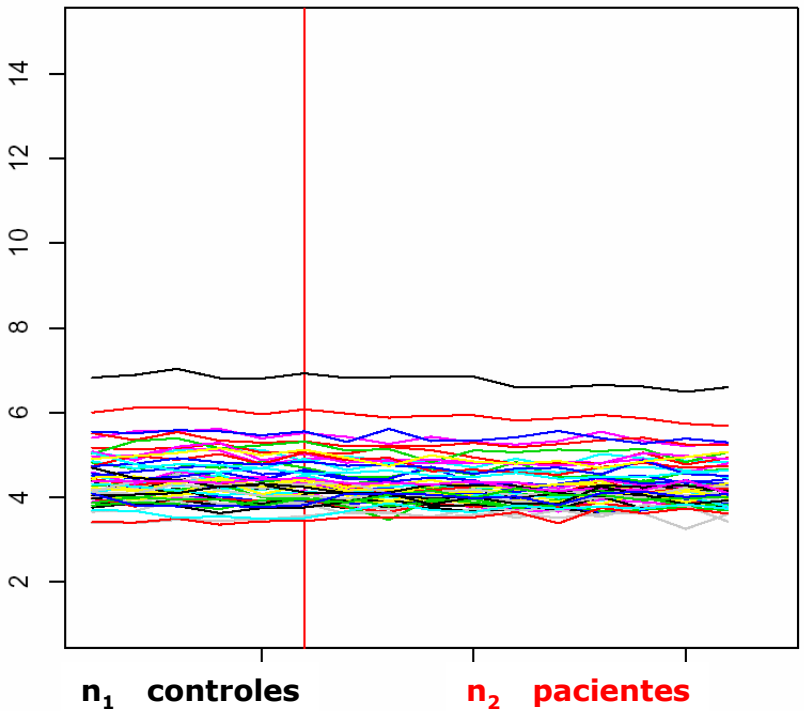
Baja variabilidad en el grupo de control → $VR(G,S_C)$ pequeña
Alta variabilidad en el grupo de pacientes → $VR(G,S_P)$ grande

Objetivo del algoritmo:

Encontrar un grupo de genes G que minimice VRR

Problema:

Genes con baja señal en los pacientes (genes planos) producen una VRR baja.



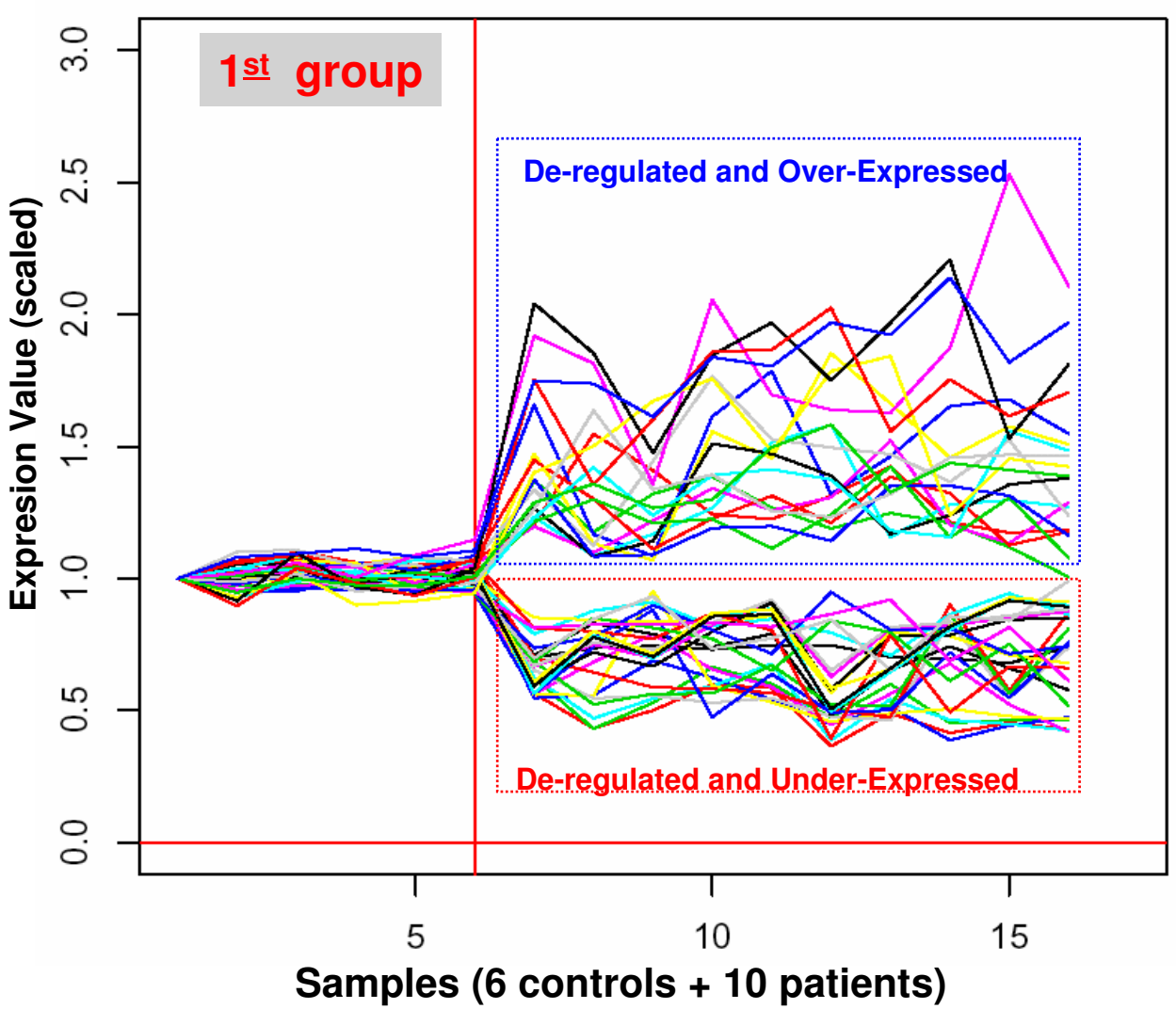
Variabilidad diferente en los controles y en los pacientes, pero señal de intensidad baja

Solución:

Filtro de diferencia de medias (ΔMF) de señal de intensidad

Para cada gen i: $\left|\overline{x}_{iC} - \overline{x}_{iP}\right| < \Delta \rightarrow$ gen plano y se elimina

Resultado 1 → Grupos de genes alterados en pacientes (el primer grupo contiene a los genes más alterados).



Resultado 2 → Puntuación F-score para medir la significación de cada grupo de genes alterados

La VR sigue una distribución Chi-cuadrado de Pearson, por tanto, la VRR es un cociente de Chi-cuadrados y se puede aproximar a una distribución F de Snedecor.

Utilizando los cuantiles de la F podemos asignar a cada grupo de genes alterados una puntuación llamada F-score

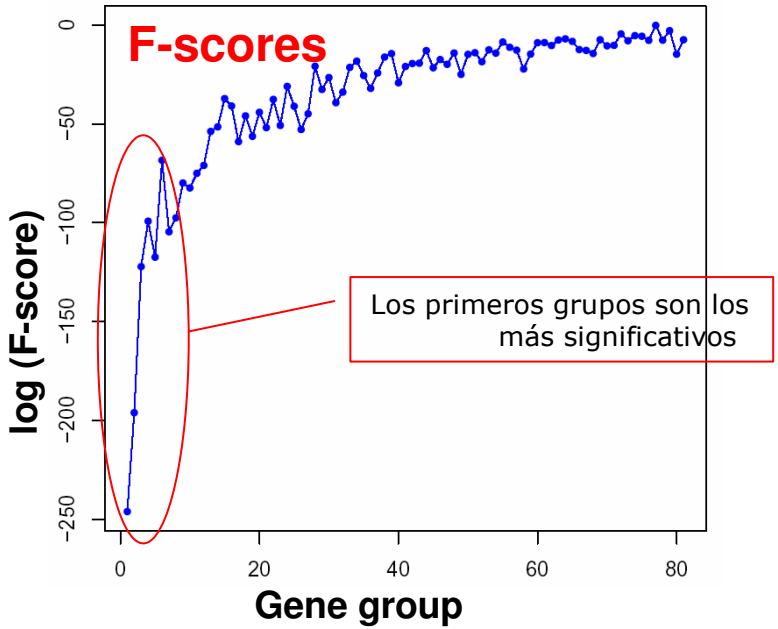
Grupo de genes

↓

F_{obs}

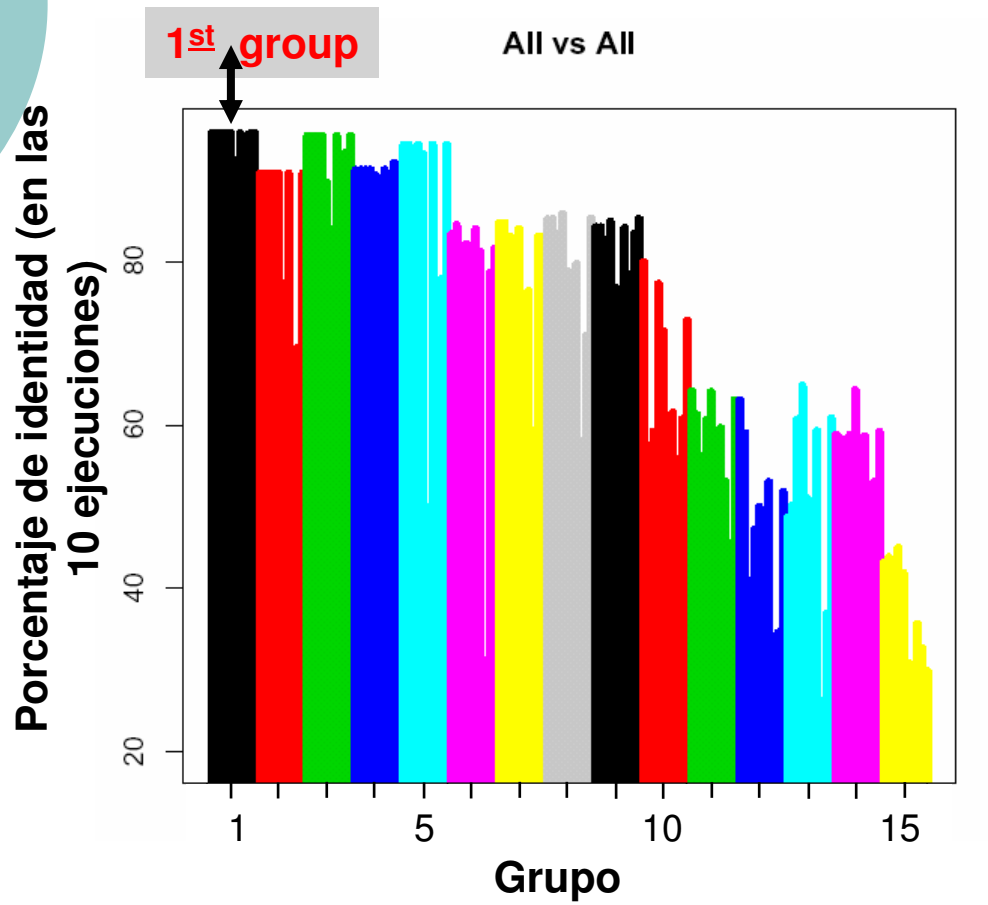
↓

$F\text{-score} = P\{ F \leq F_{obs} \}$



Estabilidad del algoritmo

Un grupo es estable cuando está formado siempre por los mismos genes alterados. Ejecutamos 10 veces el algoritmo para obtener 15 grupos cada vez:



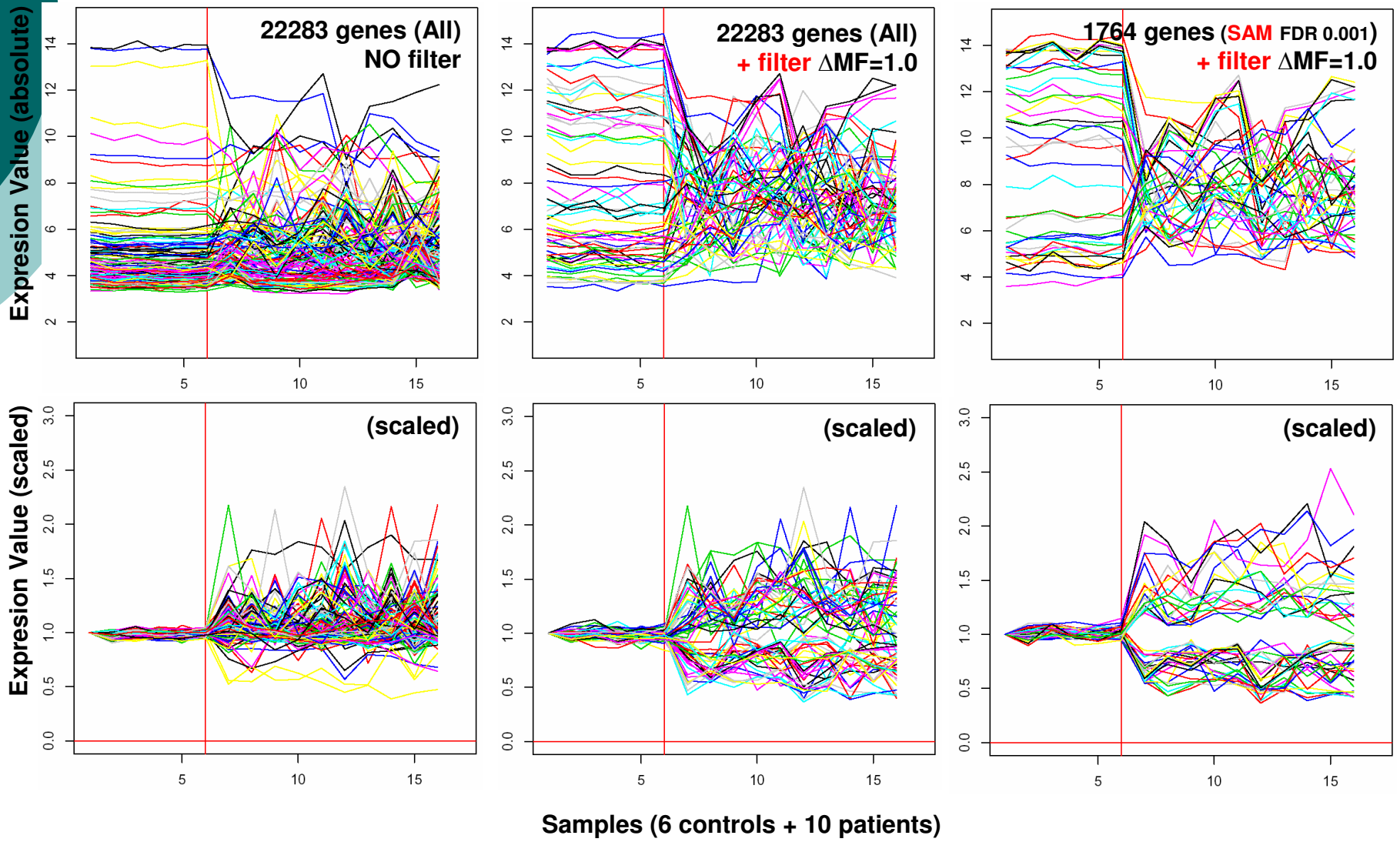
Los grupos del 1º al 5º presentan una estabilidad > 85%

Contienen los mismos genes alterados en más del 85% de las ocasiones

El grupo 1 presenta una estabilidad > 95%

AlteredExpression + SAM

(16 microarrays 6controles+10pacientes)



Referencias

- Prieto, Rivas, Sánchez, Fidalgo & de las Rivas (2006): Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics*, **22**, 9, 1103–1110.

http://bioinfow.dep.usal.es/AlteredExpression

Bolstad et al (2003): A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* **19**, 2, 185–193.

Benjamini & Hochberg (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS*, **57**, 1, 289-300.

Tusher, Tibshirani & Chu (2001): Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 9, 5116–5121.