

Protocolo de análisis de datos genéticos en Shiny




Antonio Rodríguez Gómez

Supervisado por Mercè Farré

Universidad Autònoma de Barcelona

Septiembre 13, 2018

Contexto experimental

- 1 : Científicos experimentales.
- 2 : Especialistas en genética obtienen la matriz de expresión genética.
- 3 GEA: Creación de un protocolo de análisis, interpretación de resultados, creación de una aplicación.
- 4  + GEA: Interpretación en el contexto experimental, mejoras . . .

Datos de expresión genética

- Pocos casos.
- Datos faltantes (NA).
- Variables: funcionalidad del gen, tratamientos...

$$\begin{bmatrix} y_{1,1} & y_{2,1} & \cdots & y_{M,1} & l_1 \\ y_{1,2} & y_{2,2} & \cdots & y_{M,2} & l_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{1,N} & y_{2,N} & \cdots & y_{M,N} & l_N \end{bmatrix}$$

Figure 1: Matriz de expresión genética

Protocolo de análisis

Objetivos del análisis

- 1 Analizar y comparar los diferentes tratamientos y como afectan a la expresión de los genes seleccionados.
- 2 Encontrar relaciones entre la funcionalidad del gen y su expresión bajo diferentes tratamientos.

Protocolo de análisis

Hipótesis planteadas

- 1 En algún gen (y cuál o cuáles) hay diferencias significativas entre los niveles de expresión entre tratamientos?

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

- 2 Entre cuáles tratamientos (parejas) hay diferencias significativas?

$$H_0 : \mu_i = \mu_j \text{ para cada pareja de medias } i \neq j$$

Análisis inferencial

- ANOVA para comparar medias entre diversos grupos/tratamientos.
- Normalidad, varianza igual e independencia.
- Logaritmo en los datos para normalizarlos.

Gen	$F_{k-1, N-k}$	P-valor
TFF3	5.29	0.005

- Suponemos que tenemos 50 genes, aplicamos el ANOVA para cada gen y listo?

Análisis inferencial

- Cuando realizamos un test podemos cometer el error de rechazar la hipótesis nula cuando realmente es cierta. (Error de tipo I)

Cálculo de probabilidades

- Si H_0 es cierta: $P(FP) = \alpha$ y $P(VP) = 1 - \alpha$
- $P(\text{Almenos un } FP \text{ en } m \text{ tests}) = 1 - (1 - \alpha)^m$
- Para $m = 1$, $P(\text{Almenos...}) = 1 - (1 - 0.05)^1 = 0.05$
- Para $m = 50$, $P(\text{Almenos...}) = 1 - (1 - 0.05)^{50} = 0.92$

Análisis inferencial

- Corregir los p-valores para controlar el error de tipo I

False Discovery Rate (FDR)

- Proporción esperada de falsos positivos entre todos los tests considerados como significativos.
- El método para controlar el FDR: Benjamini&Hochberg.

	Gen	P-valor	P-valor.BH
1	Gen1	0.003	0.009
2	Gen2	0.020	0.030
3	Gen3	0.450	0.450

Análisis inferencial

- De los genes que han salido significativos en el ANOVA, hacemos comparaciones 2 a 2 entre las medias de los grupos.

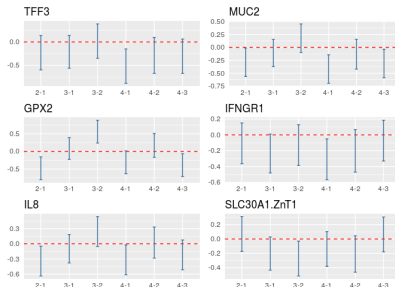


Figure 2: Comparaciones 2 a 2

Análisis exploratorio

- 1 Heatmap : explorar la doble agrupación (*clustering*) de genes y muestras.

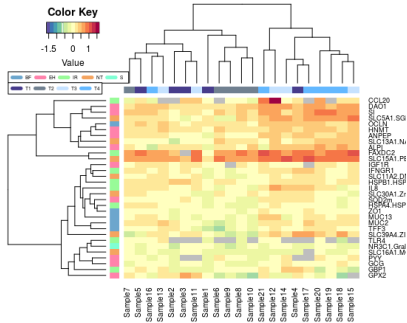


Figure 3: Heatmap

Análisis exploratorio

2 Representación ACP : explorar las correlaciones entre las medias de los tratamientos.

PCA: principal components analysis (Visual analysis)

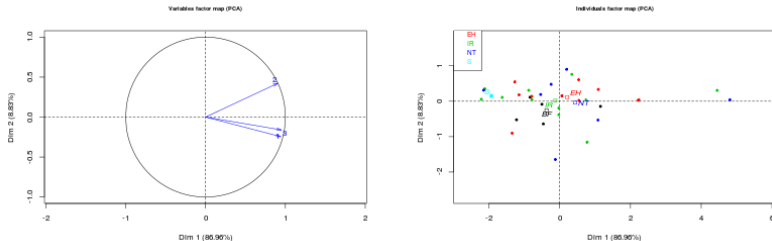


Figure 4: ACP

Análisis exploratorio

- 3 LinePlot : comparar las medias de los tratamientos a lo largo de los genes.

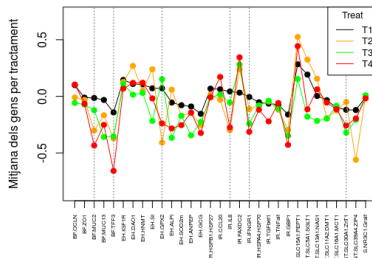


Figure 5: LinePlot

Aplicación en Shiny

La idea principal era crear una herramienta de análisis para los grupos de investigación.

- Desarrollada en R con el paquete Shiny.
- Interactiva, sencilla y accesible.
- Repositorio público.

<https://github.com/djangosee/TFGShinyApp>

Conclusiones

- Se ha descrito un protocolo de análisis capaz de analizar y explorar diversos aspectos de una matriz de expresión genética.
- La implementación del protocolo en una aplicación va a facilitar futuros análisis a los investigadores.
- La posibilidad de añadir mejoras en la aplicación. (Futuras versiones)
- Carácter multidisciplinario del trabajo: estadística, biomedicina y bioinformática.