Actividad. Representación gráfica de datos y análisis de resultados

Captura de pantalla de un celular con letras

Descripción generada automáticamente

Objetivos

En esta actividad, aplicarás los conceptos aprendidos sobre **visualización de datos biológicos y análisis estadístico**. Utilizarás un dataset que contiene información de la expresión de 46 genes en 65 pacientes, cada uno con distintos tipos de tratamiento y características tumorales. El objetivo principal es que seas capaz de interpretar los datos mediante la generación de gráficos adecuados y la respuesta a preguntas específicas que se detallan más adelante.

Pautas de elaboración

Esta actividad consistirá en dos partes principales relacionado con datos de diferentes genes: 1) elaboración de código para que generes gráficos y los reflejes en el documento HTML; y 2) breve interpretación de los gráficos del *dataset* en el documento HTML.

A continuación verás: un apartado que explica el *dataset*, mientras que el siguiente son las preguntas que deberás de contestar.

***Dataset* de expresión de genes**: para la realización de esta parte de la actividad, deberás cargar el *dataset* de interés «Dataset expresión genes.csv», el cual se trata de una base de datos de 65 pacientes que contiene información de la expresión de 46 genes con diferentes funciones (para más información, ver el apartado de «Información de interés del *dataset*» después de la rúbrica). Además de estas variables, contiene otras variables de interés como el tratamiento (A o B) que siguen cada paciente, tipo de tumor que tienen (colorrectal, pulmón y mama) y la extensión tumoral (localizado, metastásico o regional). Por último, se recoge información de variables bioquímicas, síntomas y otras variables sociodemográficas.

Tras importarlo, deberás **responder a las siguientes cuestiones**, teniendo en cuenta todos los gráficos generados mediante ggplot2, pheatmap o ComplexHeatmap. R se caracteriza por su amplia gama de funciones para la creación de gráficos originales. Los gráficos profesionales deben tener: título y subtítulo, etiquetas de ejes, leyenda (si aplica), etiquetas de datos, escalas apropiadas, colores y estilos uniformes, ajustes de líneas y puntos similar en todos los gráficos, temas consistentes, etc.

**1.** Teniendo en cuenta los siguientes genes: AQ\_ALOX5, AQ\_CD274, AQ\_CHKA, AQ\_CSF2, AQ\_FOXO3, AQ\_IL6, AQ\_LDHA, AQ\_LIF, AQ\_MAPK1, AQ\_NOS2, AQ\_IFNG, AQ\_PDCD1, AQ\_PPARG, AQ\_TGFB1, AQ\_TNF:

* Para responder este ejercicio, apóyate en un **diagrama de cajas** en el que visualices por cada expresión de gen 2 cajas: 1 para trata y otro para tratB.
* ¿Qué interpretación sacas de la distribución de la expresión de los genes cuando los comparamos por tipo de tratamiento? (mínimo 150 palabras de extensión).
* Consejos:
  + Crea los gráficos de forma individual y luego para visualizarlos mejor en 1 gráfica conjunta, utiliza la librería *library(patchwork)* (<https://patchwork.data-imaginist.com>):

combined\_plot <- plot\_A + plot\_B + … + plot\_layout(ncol = 3) # ncol es el número de columnas que quieres que haya en tu gráfico, juega con este dato para poder encajar bien los gráficos.

* + Para ayudarte en la interpretación de la expresión de genes (valores AQ), ver el apartado de Información de interés del *dataset* después de la rúbrica).

**2.** Teniendo en cuenta los siguientes parámetros bioquímicos para toda la población: glucosa, leucocitos, linfocitos, neutrofilos, chol, hdl, hierro, igA, igE, igG, igN, ldl, pcr, transferrina, trigliceridos, cpk:

* Para responder este ejercicio, apóyate en un **histograma** en el que se visualice la frecuencia de cada variable utilizando aproximadamente 30 bins (a mayor bins, mayor número de barras).
* ¿Qué interpretación sacas de la distribución de los datos de las variables bioquímicas en toda la población? (mínimo 100 palabras de extensión).
* ¿Crees que sigue una distribución normal o simétrica en la que se visualiza una distribución parecida a una campana de Gauss? (mínimo 100 palabras de extensión).
* Consejos: puedes utilizar la función geom\_density (<https://r-charts.com/distribution/histogram-density-ggplot2/>) vista en clase para visualizar la densidad.

**3.** Mapea todos los valores de expresión de genes para poder visualizar posibles patrones entre los datos de los pacientes:

* Para responder esta pregunta, apóyate en un *heatmap* en el que visualices los datos crudos de todas las variables AQ (expresión de genes).
* ¿Hay algún tipo de patrón de pacientes que tiene la expresión de genes similar o diferenciada? (mínimo 150 palabras de extensión).
* ¿Hay grupos de genes con patrones de expresión similares o diferenciadas?
* Consejos:
  + Puedes utilizar la librería pheatmap o ComplexHeatmap (instalación previa) vista en clase para visualizar patrones de los datos jerárquicamente.
  + Establece antes del gráfico siempre una semilla de aleatorización, para que pueda ser reproducible para todos de una forma similar, utilizaremos siempre set.seed(1995). Si no pones esta semilla de aleatorización, tu gráfico cambiará cada vez que ejecutas la sintaxis y no será igual al resto de tus compañeros.
  + Escala siempre los datos para que sean comparables entre sí usando la función scale(dataframe), lo que normaliza las variables centrándolas alrededor de una media de 0 y una desviación estándar de 1, asegurando que todas las variables contribuyan equitativamente al análisis.
  + No se te pide que hagas correlaciones, simplemente visualiza de forma cruda los datos, en donde el eje X se localicen los pacientes, mientras que en el eje Y las variables de los genes.
  + Para visualizar patrones, recuerda que hay funciones específicas dentro de las librerías de pheatmap o ComplexHeatmap. Por ejemplo, en pheatmap (https://r-charts.com/es/correlacion/pheatmap/ y <https://davetang.org/muse/2018/05/15/making-a-heatmap-in-r-with-the-pheatmap-package/>) está las funciones row\_km, cutree\_rows, column\_km, cutree\_cols. Por ejemplo, en ComplexHeatmap (<https://jokergoo.github.io/ComplexHeatmap-reference/book/a-single-heatmap.html#heatmap-split>) están las funciones row\_km y column\_km. Estas funciones podéis encontrarlas en el help de R o en Google para ver cómo funcionan.

Extensión y formato

Para la resolución de esta actividad, deberás entregar dos archivos:

* Un único fichero **R Markdown** (*.Rmd*) con todo el código y texto Markdown que hayas generado. No existe un límite de extensión para este fichero.
* Un fichero **HTML** generado a partir de dicho archivo R Markdown. Asegúrate de que aparecen todas las figuras, todo el texto y todos los cuadros de código R que hayas introducido en el archivo *.Rmd* anterior. Este fichero es el evaluable.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Representación gráfica de datos y análisis de resultados | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Gráficos | Construcción correcta de los gráficos tal y como se solicitan en el enunciado de la actividad | 3 | 30% |
| Cuestiones | Resolución de las cuestiones planteadas de forma correcta | 2 | 20% |
| Originalidad | Generación de un informe R Markdown original, empleando temas o estilos CSS, índices, etc. | 1 | 10% |
| Orden | Exposición de lo que se pide en el enunciado de forma clara y ordenada | 1 | 10% |
| Justificación | Justificación detallada en la respuesta a las preguntas planteadas y en el análisis de los gráficos generados | 3 | 30% |
|  |  | **10** | **100 %** |

Información de interés del dataset

**Conceptos de expresión de genes e interpretación de valores de AQ en la expresión de genes**

La PCR en tiempo real o qPCR (quantitative PCR) es una técnica ampliamente utilizada para medir y comparar la expresión génica. Este método permite la cuantificación de la cantidad de ADN presente en una muestra en tiempo real, a medida que se amplifica. A continuación, se explica cómo se relacionan los conceptos mencionados con la qPCR:

* **Ct (Threshold Cycle)**: En qPCR, el Ct es el ciclo en el que la cantidad de ADN amplificado supera un umbral detectable. Esto se utiliza para determinar la cantidad inicial de ADN en la muestra.
* **DCt (Delta Ct):** DCt se calcula restando el Ct del gen de referencia (normalizador, para este ejericio se ha utilizado el 18S rRNA o ARN ribosómico 18S, que es un gen comúnmente utilizado como control interno o gen normalizador debido a su expresión estable en diferentes tipos de células y condiciones experimentales) del Ct del gen de interés. Este valor nos permite normalizar la expresión del gen de interés respecto a un gen cuya expresión es constante.
* **AQ (Absolute Quantification):** AQ se calcula como 2(-DCt), y representa la expresión del gen en una condición específica sin comparación directa con otra condición. Los valores de AQ en la expresión de genes representan cuánto están activos o produciendo proteínas esos genes en respuesta a diferentes condiciones o factores en un organismo. En otras palabras, estos valores indican la cantidad de actividad o expresión de cada gen en el contexto específico de estudio, como tratamientos médicos, tipos de tumor, o la extensión de la enfermedad. Cuando **el valor de AQ es 0**, significa que **no se detecta ninguna expresión o actividad del gen** en las muestras analizadas bajo las condiciones específicas del estudio. Esto puede indicar que el gen no está siendo trascrito o que su expresión es extremadamente baja y no detectable por los métodos utilizados.

**Genes relacionados con el metabolismo y la adipogénesis:**

* **AQ\_ADIPOQ**: Adiponectina, implicada en la regulación del metabolismo de la glucosa y la descomposición de ácidos grasos.
* **AQ\_FASN**: Ácido graso sintasa, clave en la síntesis de ácidos grasos.
* **AQ\_PPARG**: Peroxisome proliferator-activated receptor gamma, regula la adipogénesis.
* **AQ\_SREBF1**: Sterol regulatory element-binding protein 1, regula la síntesis de lípidos.

**Genes relacionados con la respuesta inmune e inflamación:**

* **AQ\_CCL2 (MCP-1)**: Quimiocina que atrae monocitos.
* **AQ\_CCL5 (RANTES)**: Quimiocina implicada en la respuesta inflamatoria.
* **AQ\_CCR5**: Receptor de quimiocinas, implicado en la respuesta inmune.
* **AQ\_CD274 (PD-L1)**: Inhibe la respuesta inmune, relacionado con la evasión inmune en cáncer.
* **AQ\_IFNG**: Interferón gamma, juega un papel crucial en la inmunidad innata y adaptativa.
* **AQ\_IL10**: Interleucina 10, antiinflamatoria.
* **AQ\_IL1B**: Interleucina 1 beta, proinflamatoria.
* **AQ\_IL6**: Interleucina 6, proinflamatoria.
* **AQ\_TNF**: Factor de necrosis tumoral alfa, proinflamatorio.

**Genes relacionados con la señalización celular y el ciclo celular:**

* **AQ\_JAK1**: Janus kinase 1, involucrado en la señalización de citoquinas.
* **AQ\_JAK3**: Janus kinase 3, similar a JAK1, también en señalización de citoquinas.
* **AQ\_STAT3**: Signal transducer and activator of transcription 3, mediador de señales de citoquinas.
* AQ\_MAPK1: Mitogen-activated protein kinase 1, implicada en la proliferación y diferenciación celular.
* **AQ\_NFE2L2**: Nuclear factor erythroid 2-related factor 2, regula la respuesta antioxidante.
* **AQ\_NFKB1**: Nuclear factor kappa B subunit 1, regula la respuesta inmune y la inflamación.
* **AQ\_FOXO3**: Forkhead box O3, regulador de la apoptosis y la resistencia al estrés.
* **AQ\_FOXP3**: Forkhead box P3, regula el desarrollo y función de células T reguladoras.
* **AQ\_PDCD1 (PD-1)**: Programmed cell death protein 1, regula la inmunidad adaptativa.

**Genes relacionados con el estrés oxidativo y la apoptosis:**

* **AQ\_G6PD**: Glucose-6-phosphate dehydrogenase, protege contra el estrés oxidativo.
* **AQ\_GPX1**: Glutathione peroxidase 1, protege contra el daño oxidativo.
* **AQ\_SOD1**: Superoxide dismutase 1, elimina radicales libres.
* **AQ\_NOS2 (iNOS)**: Inducible nitric oxide synthase, produce óxido nítrico en respuesta a infecciones.
* **AQ\_NOX5**: NADPH oxidase 5, produce especies reactivas de oxígeno.

**Genes relacionados con la inflamación y la señalización inflamatoria:**

* **AQ\_ALOX5**: Arachidonate 5-lipoxygenase, produce leucotrienos inflamatorios.
* **AQ\_PTGS2 (COX-2)**: Cyclooxygenase 2, produce prostaglandinas inflamatorias.
* **AQ\_NLRP3**: NOD-like receptor family pyrin domain containing 3, componente del inflamasoma.
* **AQ\_TLR3**: Toll-like receptor 3, detecta patógenos virales.
* **AQ\_TLR4**: Toll-like receptor 4, detecta lipopolisacáridos bacterianos.

**Genes relacionados con el transporte y metabolismo energético:**

* **AQ\_CPT1A**: Carnitine palmitoyltransferase 1A, esencial para la oxidación de ácidos grasos.
* **AQ\_GPD2**: Glycerol-3-phosphate dehydrogenase 2, involucrado en el metabolismo energético.
* **AQ\_SLC2A4 (GLUT4)**: Facilita el transporte de glucosa en células musculares y adiposas.

**Genes relacionados con el desarrollo y la diferenciación celular:**

* **AQ\_BMP2**: Bone morphogenetic protein 2, implicado en la formación ósea.
* **AQ\_LIF**: Leukemia inhibitory factor, afecta la diferenciación celular y la respuesta inmune.

**Genes relacionados con la función y desarrollo de la célula inmunitaria:**

* AQ\_CSF2: Colony stimulating factor 2, estimula el crecimiento y diferenciación de células inmunitarias.
* **AQ\_CXCR1**: Receptor para IL-8, implicado en la migración de neutrófilos.