CLASE 09 - ANÁLISIS EXPLORATORIO DE DATOS DE EXPRESIÓN DE GENES.

Curso Análisis de expresión diferencial de genes e investigación reproducible.

Dr. José Gallardo Matus | https://genomics.pucv.cl/

Pontificia Universidad Católica de Valparaíso

29 October 2022

PLAN DE LA CLASE

1.- Introducción

- Análisis exploratorio de datos de expresión de genes.
- Diseño balanceado o desbalanceado.
- Clasificación de variables aleatorias.
- Variabilidad de datos de expresión de genes.
- Correlación de datos de expresión de genes.
- Funciones avanzadas de Tidyr.

2.- Práctica con R y Rstudio cloud

- Realizar un análisis exploratorio de datos de expresión de genes.
- Realizar gráficas avanzadas con ggplot2.

ANÁLISIS EXPLORATORIO DE DATOS EXPRESIÓN DE GENES

Preguntas clave para un exploratorio de expresión de genes.

- ¿Tengo un diseño balanceado de observaciones por factor?
- ¿Existen errores, datos faltantes, valores atípicos?
- ¿Cómo varian los datos de Ct entre genes y entre tratamientos?
- ¿La expresión de mis genes varia en el tiempo, o en el espacio o en función de algún tratamiento?
- ¿La expresión de los genes está correlacionada entre si y con otras variables?

EDA: IMPORTANCIA DE LA ESTRUCTURA DE LOS DATOS

Diseño equilibrado o balanceado: Todos los tratamientos son asignados a un número equivalente de unidades experimentales (observaciones).

¿Datos son balanceados o desbalanceados?

Table 1: Observaciones por sexo y dieta.

	diet_1	diet_2	diet_3	diet_4	diet_5	diet_6
Male	3	3	4	2	3	0
Female	9	7	8	9	11	12

note que la media de la dieta 6 podría ser distinta al resto como consecuencia de la falta de machos y quizás no del real efecto de la dieta en la variable de interes.

CONCEPTOS Y DEFINICIONES

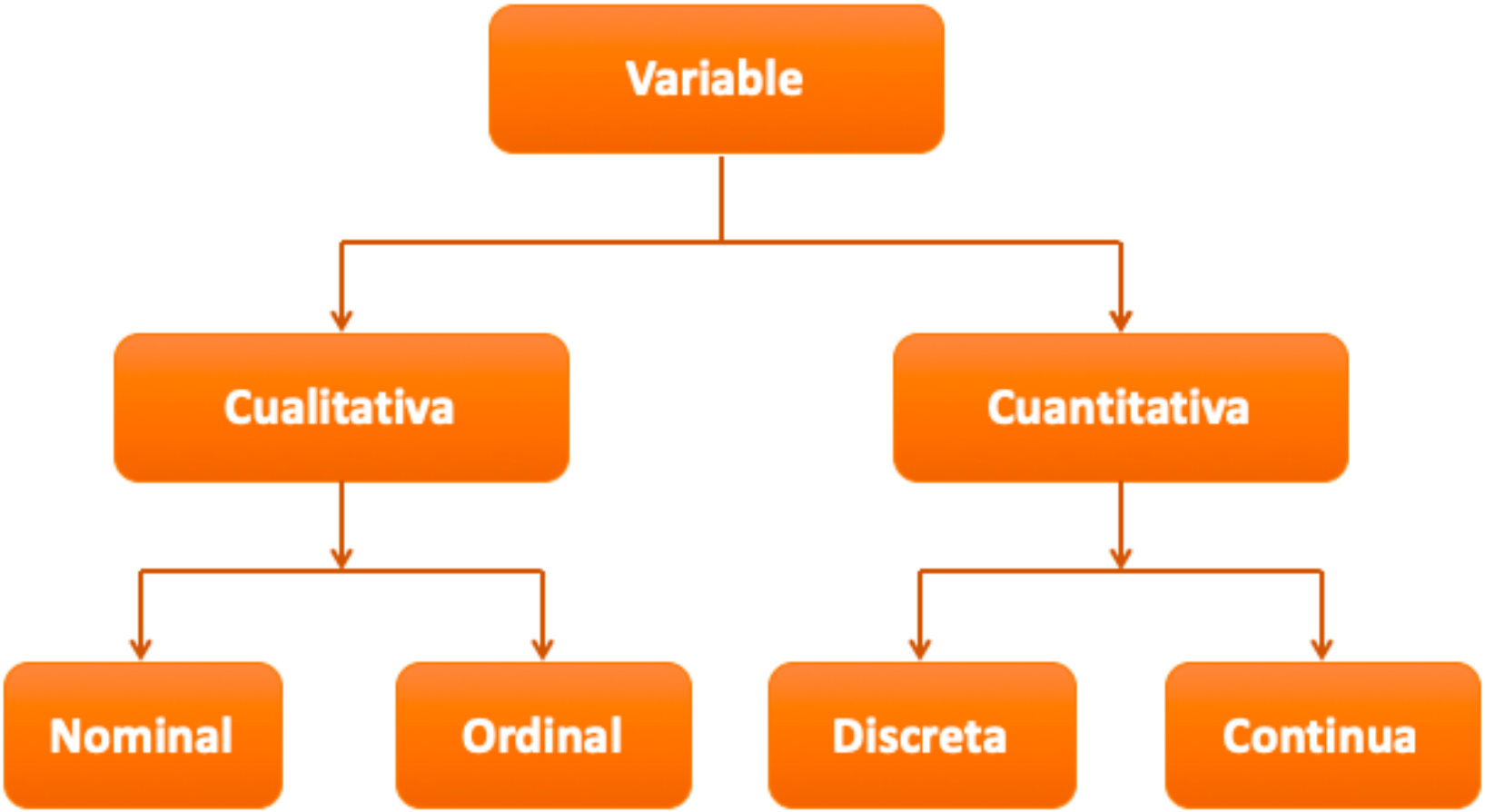
1. **Variable**: Características que se pueden medir u observar en un individuo o en un ambiente: peso, temperatura, Sexo, pH, Tipo de bacteria, abundancia de organismos, número de alelos, valor de Ct, expresión relativa de un gen.

2. **Variable aleatoria**: es un número que representa el resultado de un experimento aleatorio. Depende entonces de función matemática o distribución de probabilidad.

3. **Datos u observaciones**: Son los valores que puede tomar una variable aletoria. 25 gramos, 55 mm, 13ºC, 7 unidades de pH, 25 bacterias, 2 alelos, 32 ct, 1,5 fold change.

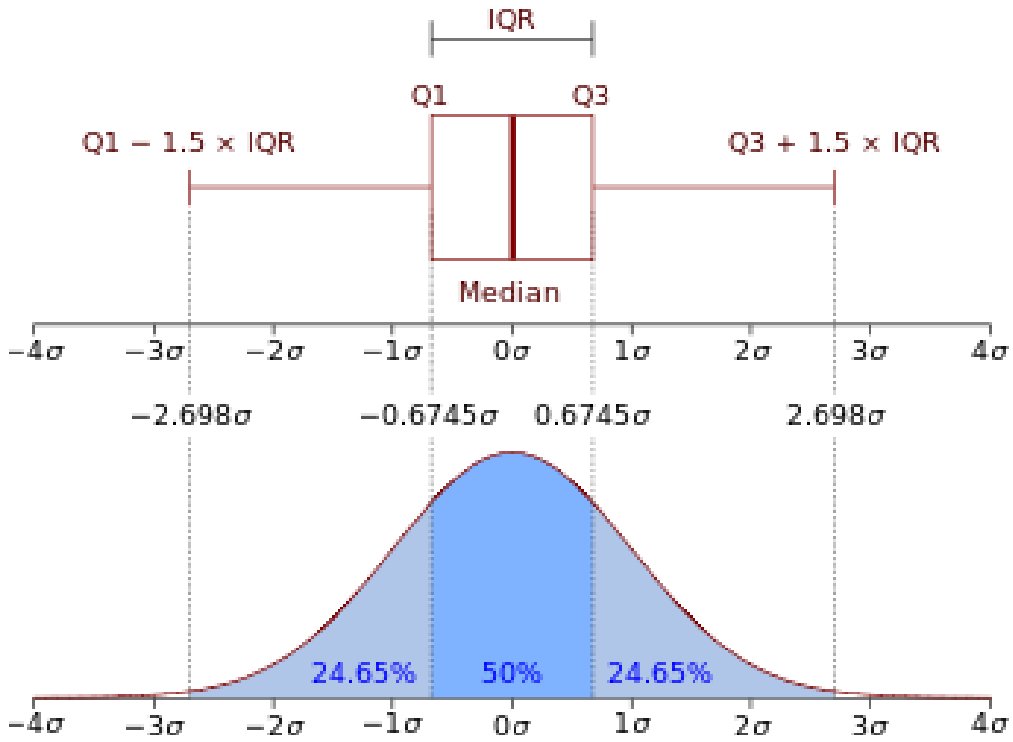
4. **Factor**: Usado para identificar tratamientos de un experimento o variables de clasificación. Se usan como *variables independientes o predictoras*, es decir tienen un efecto sobre una *variable respuesta o dependiente*. Ej. Sexo (niveles: macho o hembra) tiene un efecto sobre la expresión de un gen.

CLASIFICACIÓN DE VARIABLES



VARIABLE ALEATORIA CUANTITATIVA CONTINUA

Definición: Puede tomar cualquier valor dentro de un intervalo (a,b), (a,Inf), (-Inf,b),(-Inf,Inf). Tienen una distribución normal. Las gráficas de cajas y bigotes son muy adecuadas para observar variables aleatorias continuas.



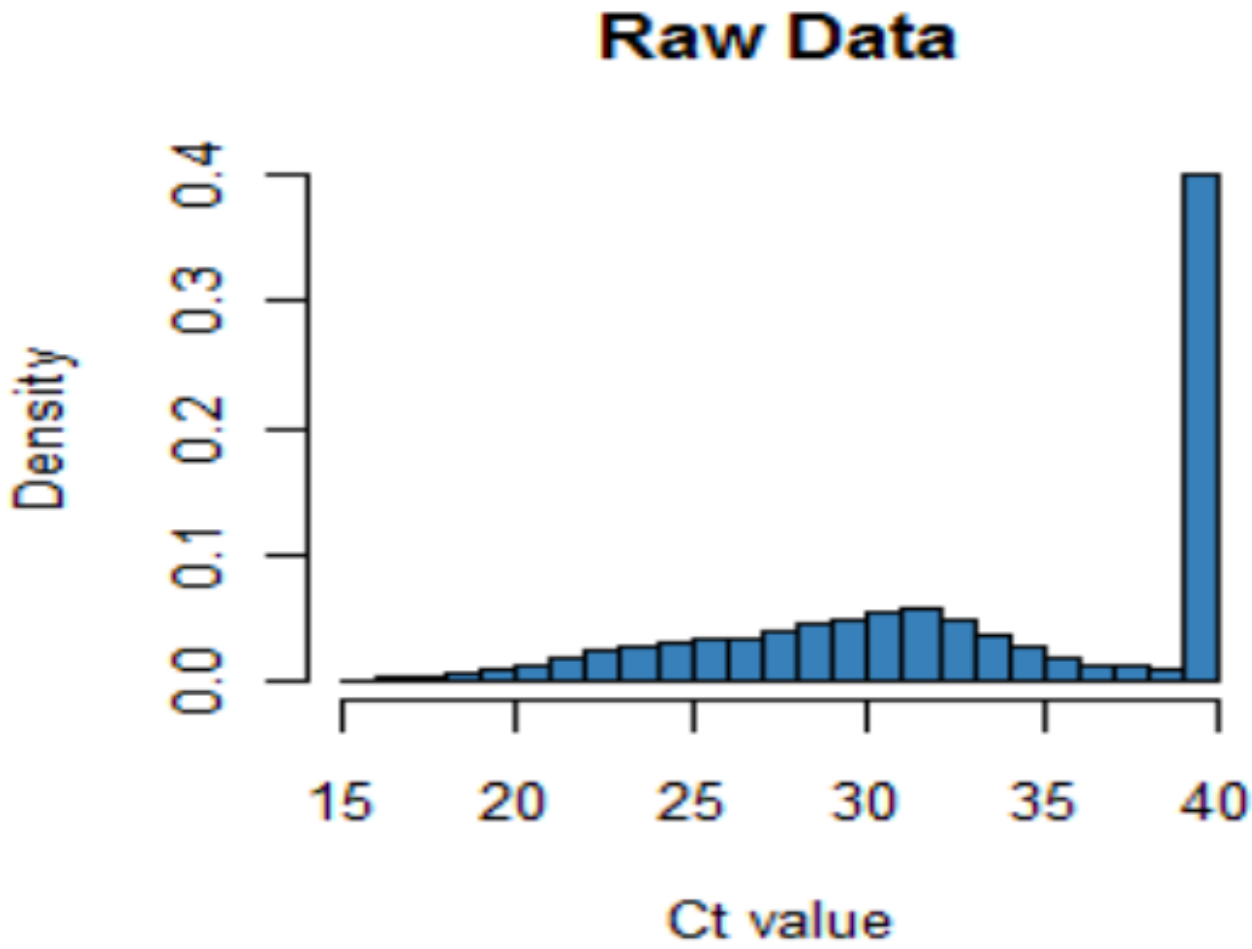
VARIABLES RELACIONADAS A EXPRESIÓN DE GENES

- Variables relacionadas a experimentos d expresión de genes.

1. Valor de Ct o Cq.
2. Delta Ct.
3. Delta Delta Ct.
4. 2^ddCT (Livak).
5. Fold change.
6. Log(2^ddCT).
7. Gene expression Ratio (Pfaffl).

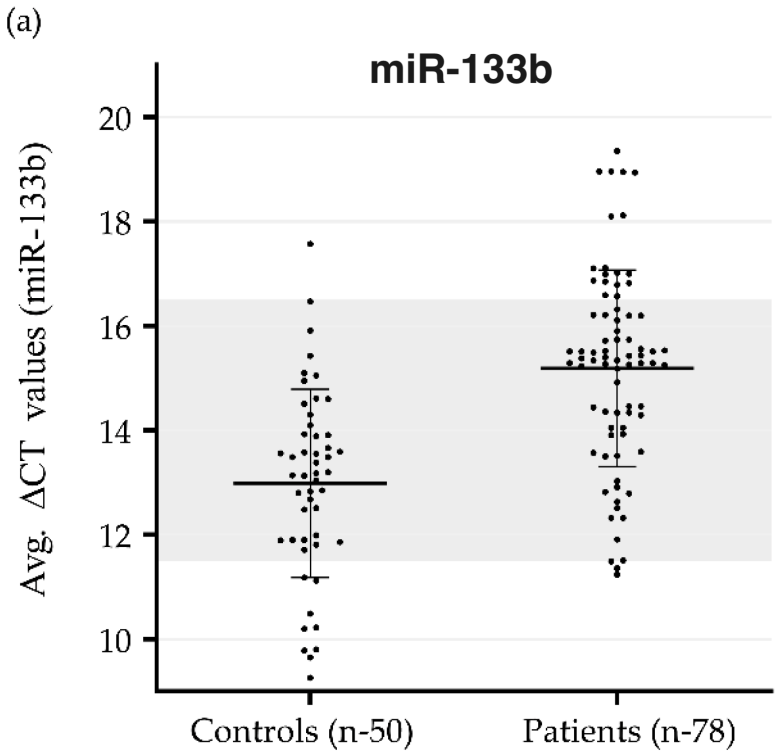
¿CÓMO DISTRIBUYE LA VARIABLE CT?

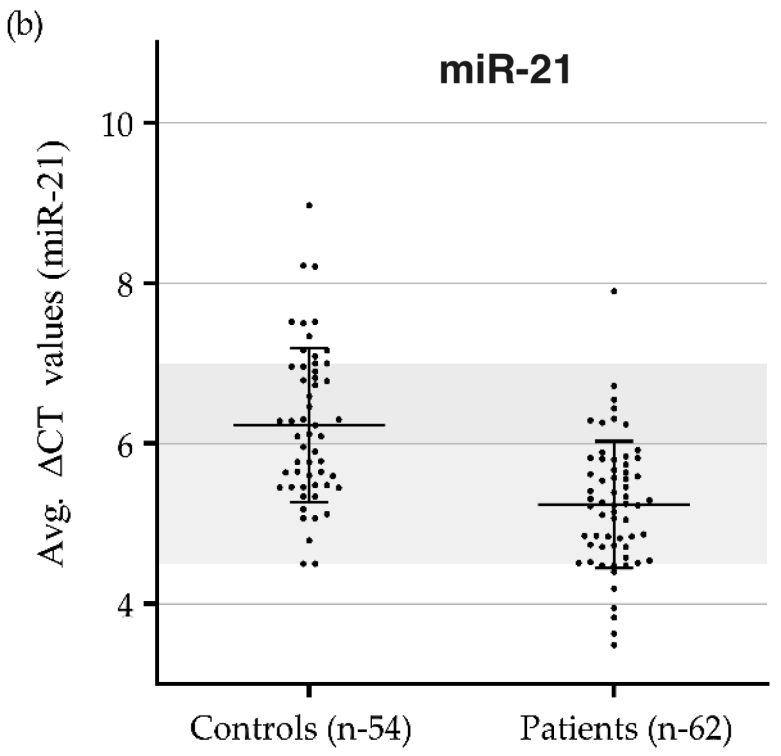
- Estudio de caso: 768 miRNAs colectados desde 19 pacientes normales y 19 enfermos de cancer a la próstata.



¿CÓMO DISTRIBUYE LA VARIABLE DELTA CT?

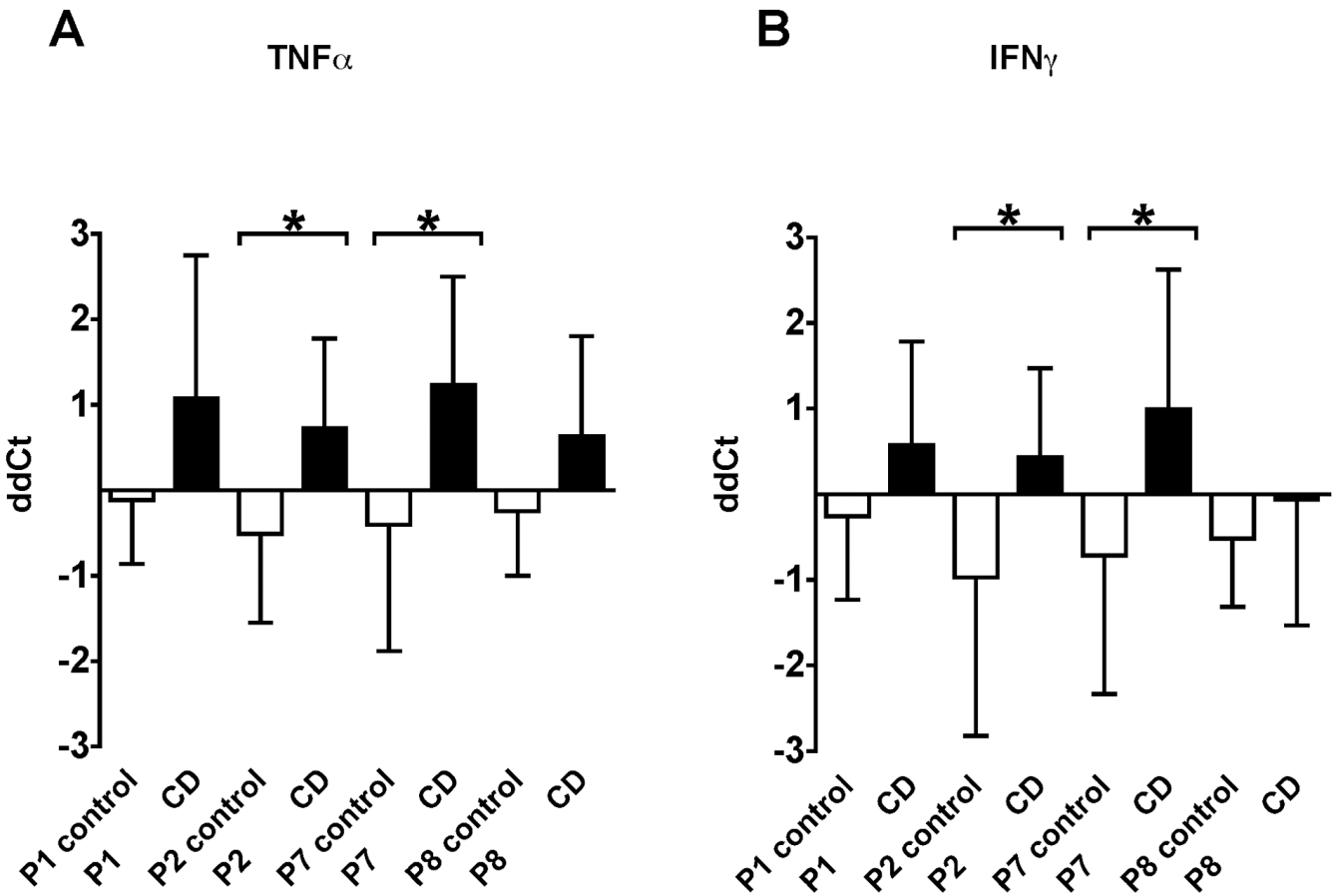
- Estudio de caso: Predicción de enfermedad coronaria con biomarcadores de miRNA.





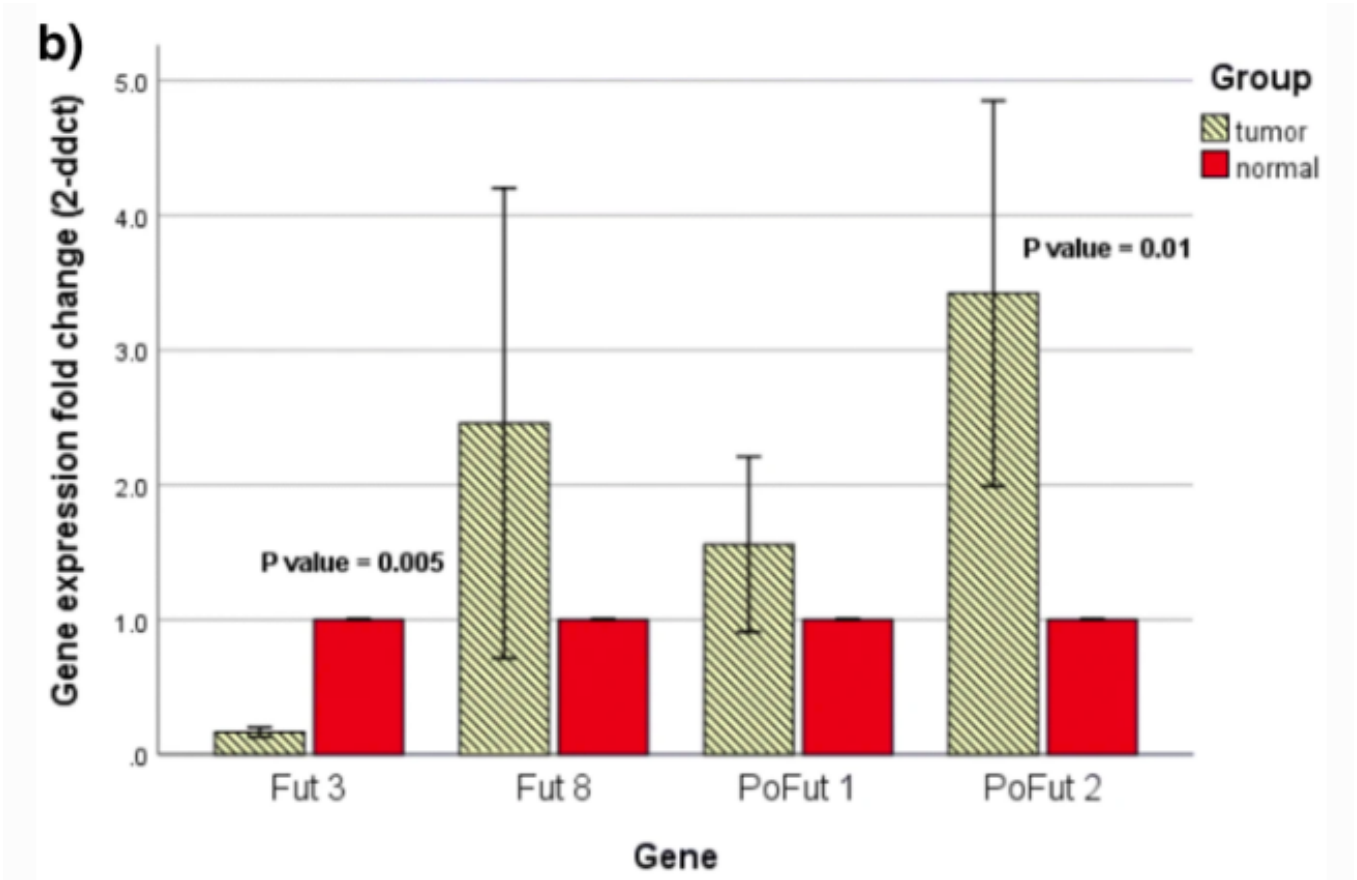
¿CÓMO DISTRIBUYE LA VARIABLE ddCT?

- Estudio de caso: Modulación de la respuesta inmune en pacientes pediátricos.



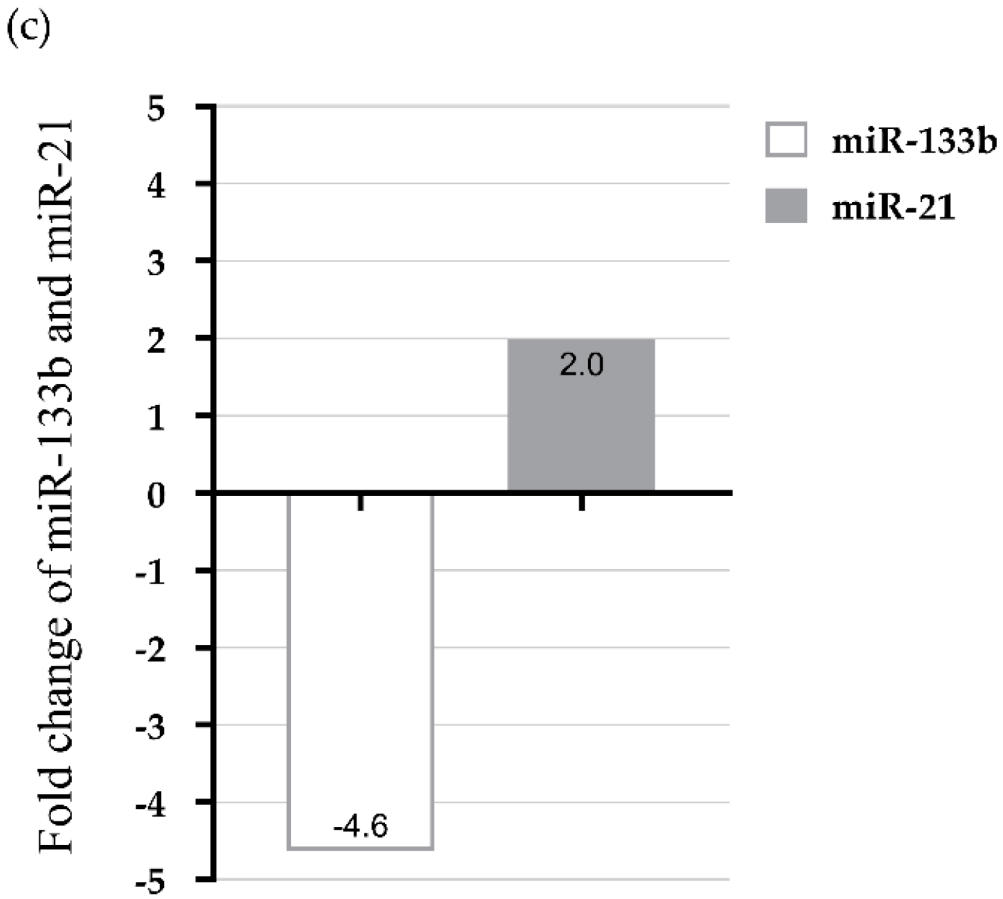
¿CÓMO DISTRIBUYE LA VARIABLE 2ˆddCT?

- Estudio de caso: regulación de Fucociltransferasa en línea celular de carcer.



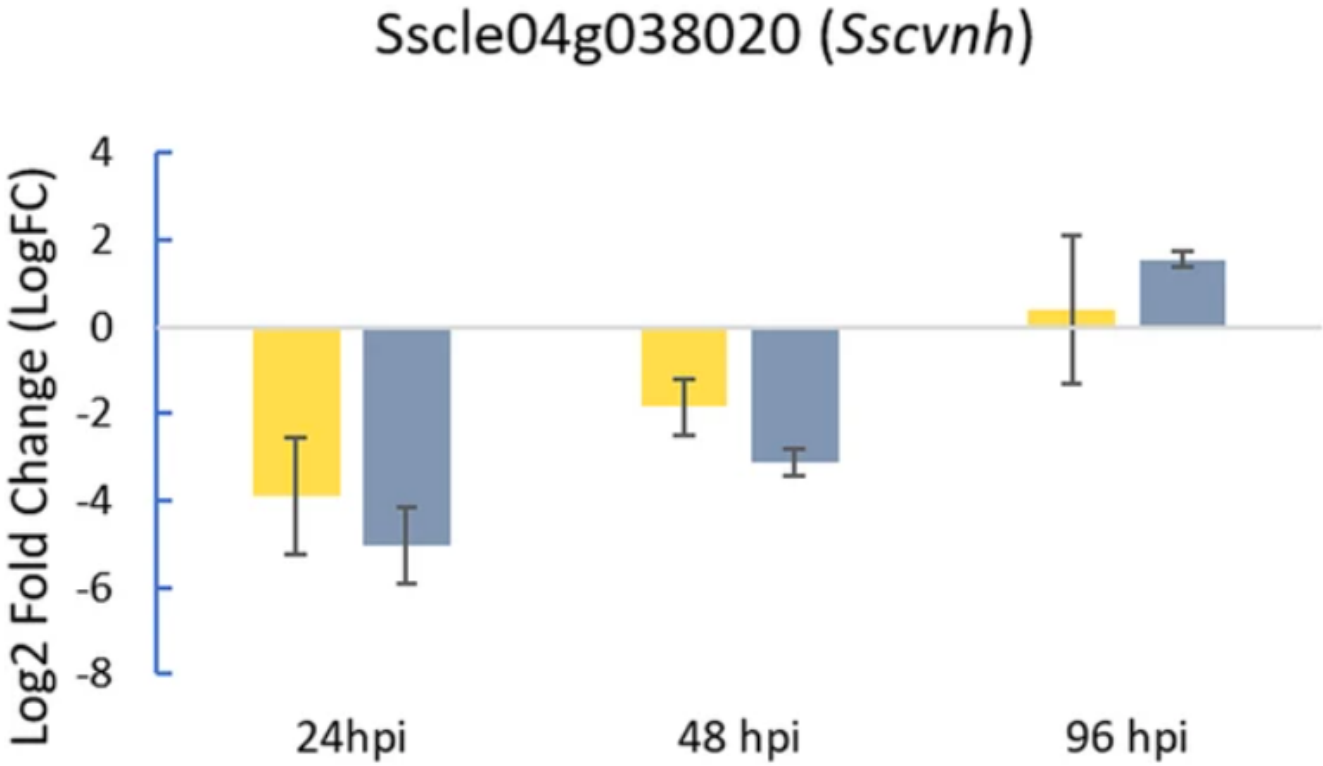
¿CÓMO DISTRIBUYE LA VARIABLE FOLD CHANGE?

- Estudio de caso: Predicción de enfermedad coronaria con biomarcadores de miRNA.



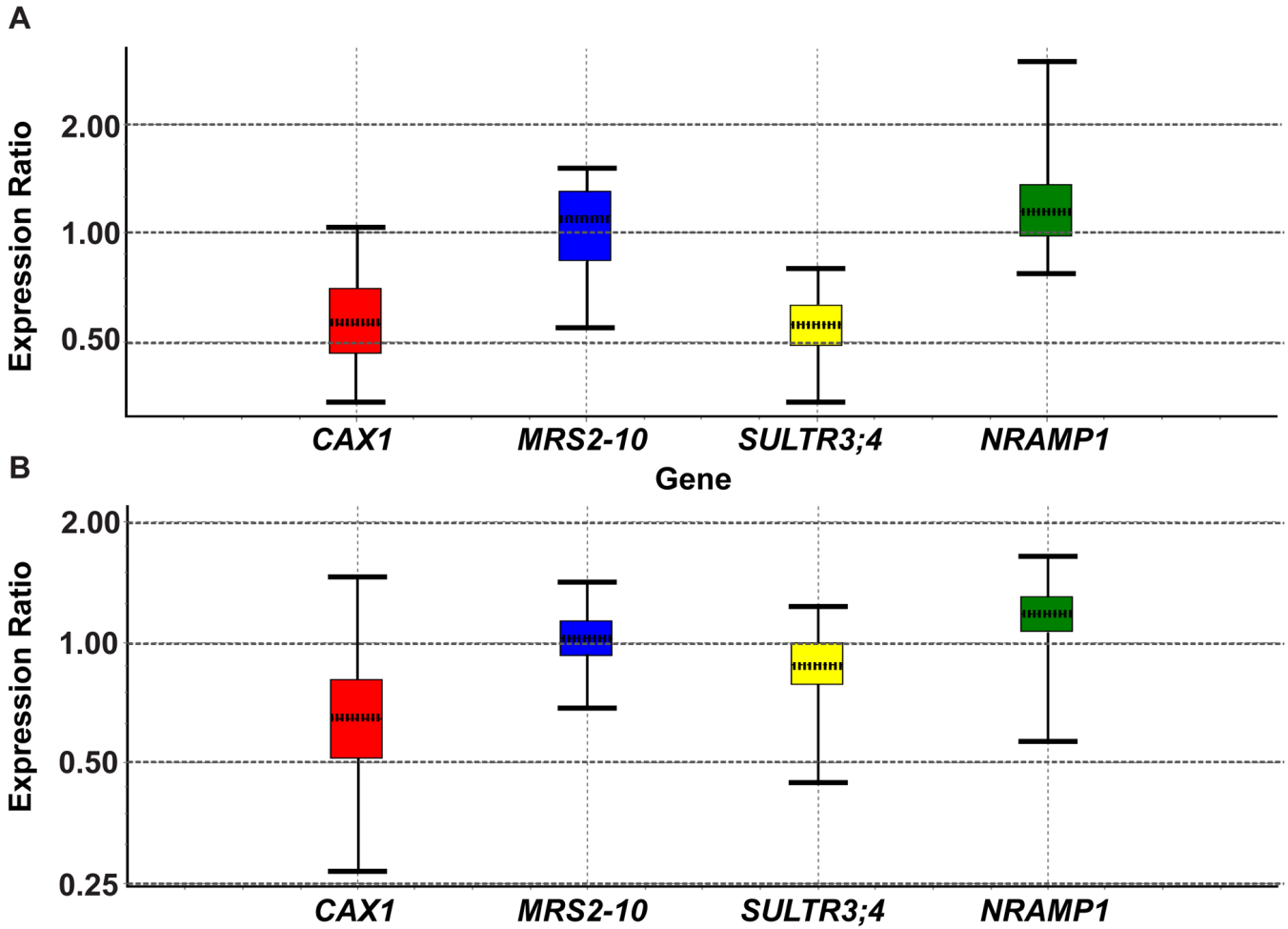
¿CÓMO DISTRIBUYE LA VARIABLE LOG(2^ddCT)?

- Estudio de caso: Respuesta a la infección en soya.



¿CÓMO DISTRIBUYE LA VARIABLE GENE EXPRESION RATIO?

- Estudio de caso: Respuesta de Arabidopsis al sulfato de magnesio.



IDENTIFICA CORRECTAMENTE TU VARIABLE

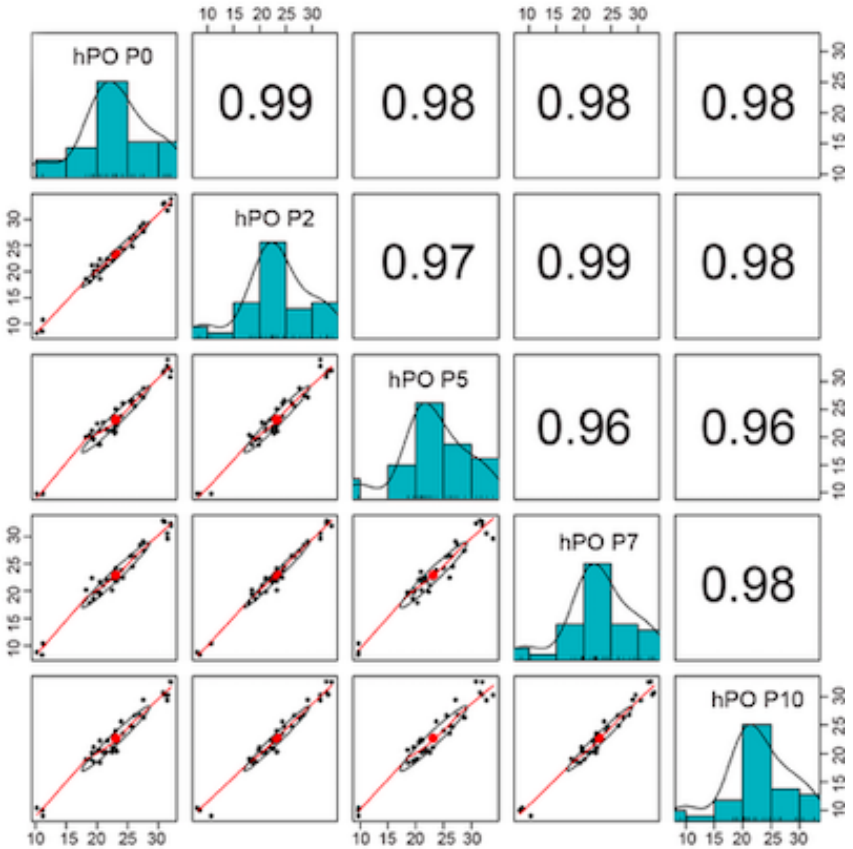
- Es importante identificar correctamente la variable de expresión de genes para evitar errores en la interpretación.
- Mucha precaución al interpretar resultados de 2^ddCT o Expresión Ratio, pequeñas diferencias hacia abajo de 1 representan grandes cambios en la expresión relativa de los genes.

TG Ct	HKG Ct	ΔCt	Calibrator	ΔΔCt	2^-(ΔΔCt) Livak	FC	Log (2^-(ΔΔCt))	GE Ratio pfaffl
26	17	9	15	-6	64,0	64	1,806	64,3
27	17	10	15	-5	32,0	32	1,505	32,7
28	17	11	15	-4	16,0	16	1,204	16,6
29	17	12	15	-3	8,0	8	0,903	8,3
29	17	12	15	-3	8,0	8	0,903	8,4
30	17	13	15	-2	4,0	4	0,602	4,5
31	17	14	15	-1	2,0	2	0,301	2,6
32	17	15	15	0	1,0	1	0,000	1,2
33	17	16	15	1	0,50	-2	-0,301	0,51
34	17	17	15	2	0,25	-4	-0,602	0,24
35	17	18	15	3	0,125	-8	-0,903	0,115
36	17	19	15	4	0,063	-16	-1,204	0,060
37	17	20	15	5	0,031	-32	-1,505	0,033
38	17	21	15	6	0,016	-64	-1,806	0,017

CORRELACIÓN CT ENTRE REPLICAS BIOLÓGICAS

¿Existe correlación entre réplicas biológicas?

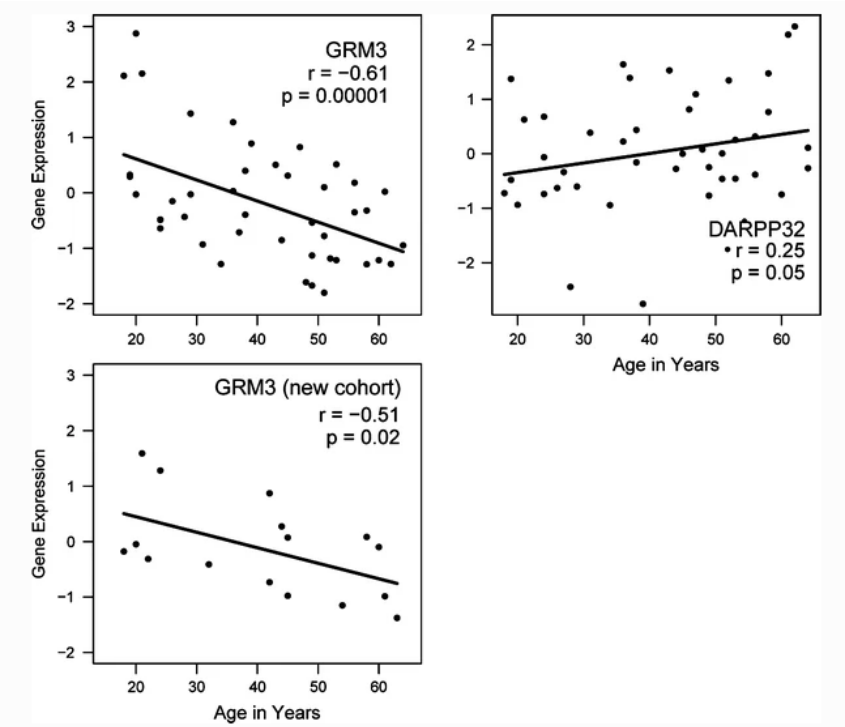
- Estudio de caso: Respuesta correlacionada entre pares de muestras (CT HK gene).



CORRELACIÓN EXPRESIÓN DE GENES Y FENOTIPO

¿Existe correlación entre expresión de mis genes y otras variables?

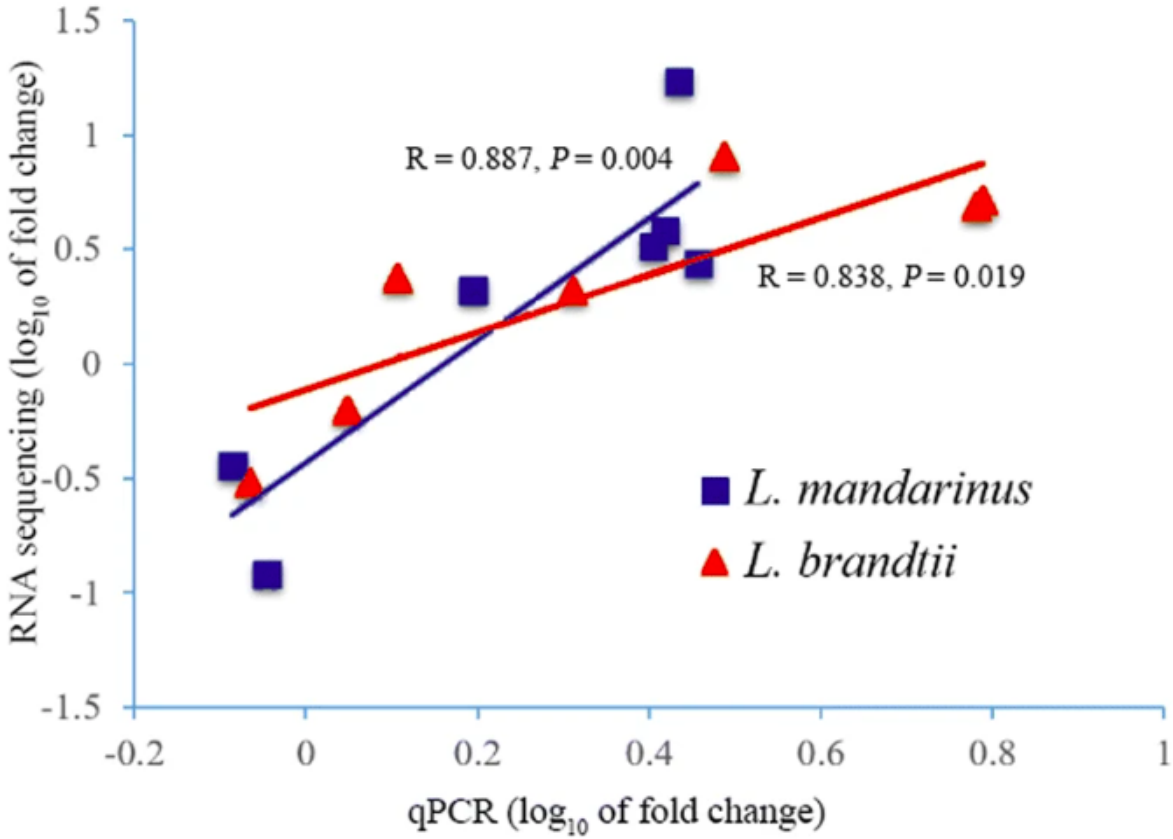
- Estudio de caso: Respuesta correlacionada entre expresión de genes relacionados con esquizofrenia y edad.



OTRAS CORRELACIONES: qPCR y RNAseq

¿Existe correlación entre expresión de genes por qPCR y RNAseq?

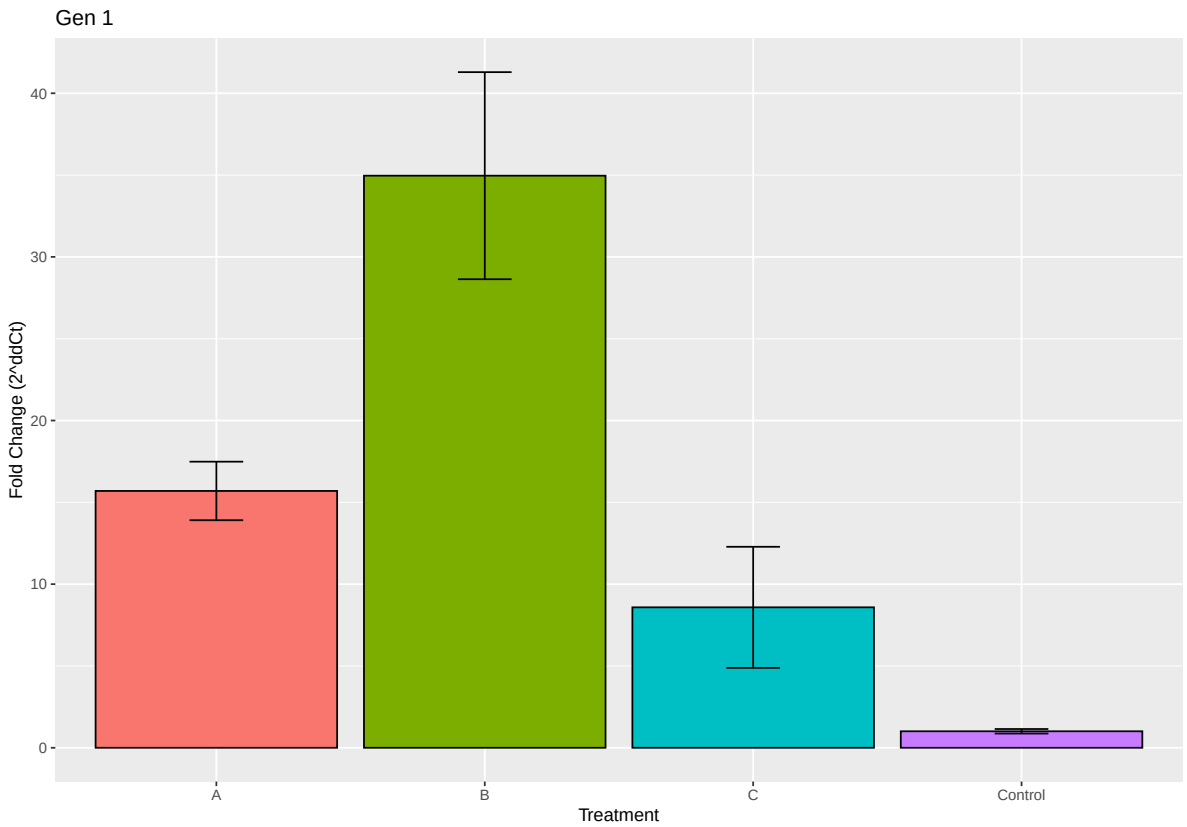
- Estudio de caso: Respuesta correlacionada entre RNAseq y qPCR.



COMUNICACION EFECTIVA DE TUS RESULTADOS

Principales problemas de las barras con error

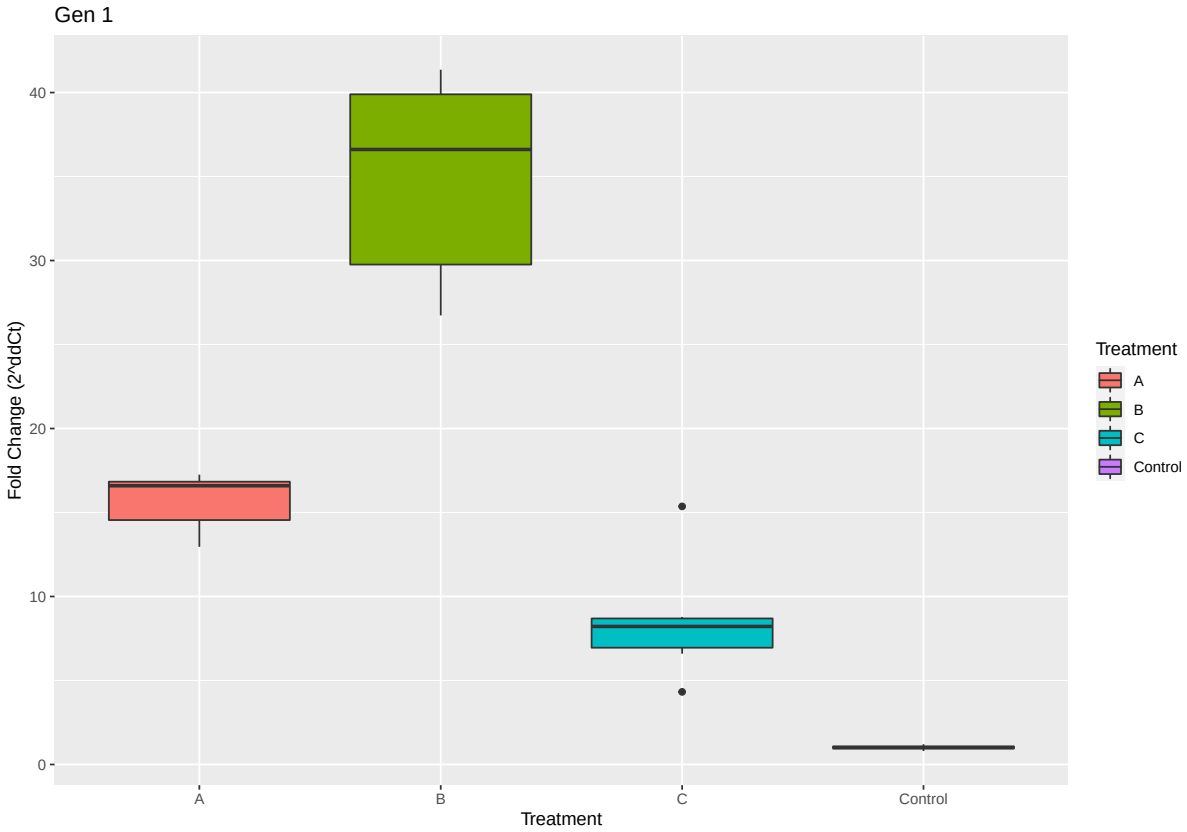
1. La escala del eje y.
2. Se enmascara la variabilidad.
3. Por lo tanto, no usar y desconfiar de ellas



COMUNICACION EFECTIVA: BOXPLOT

Pros y contras del boxplot

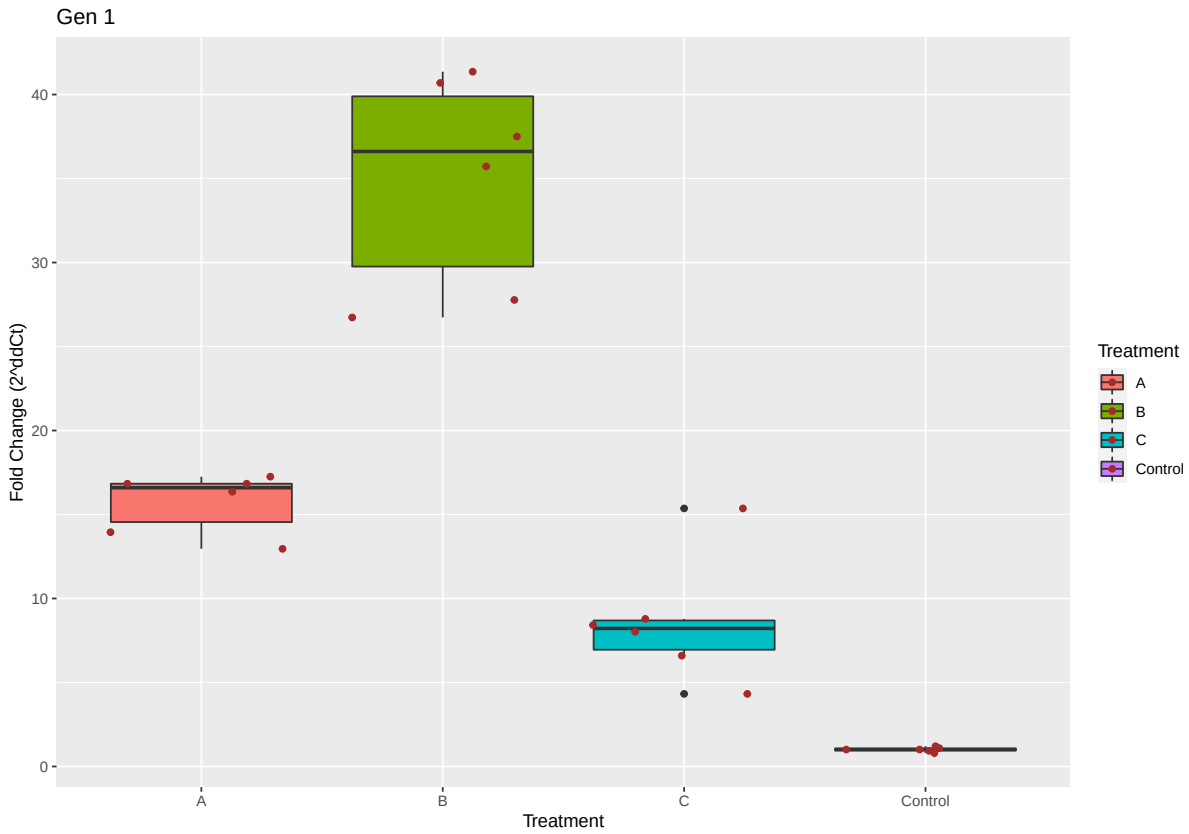
1. Muestra la escala real de y. 2. Permite detectar valores atípicos.
3. No muestra el número de observaciones.



COMUNICACION EFECTIVA: JITTER

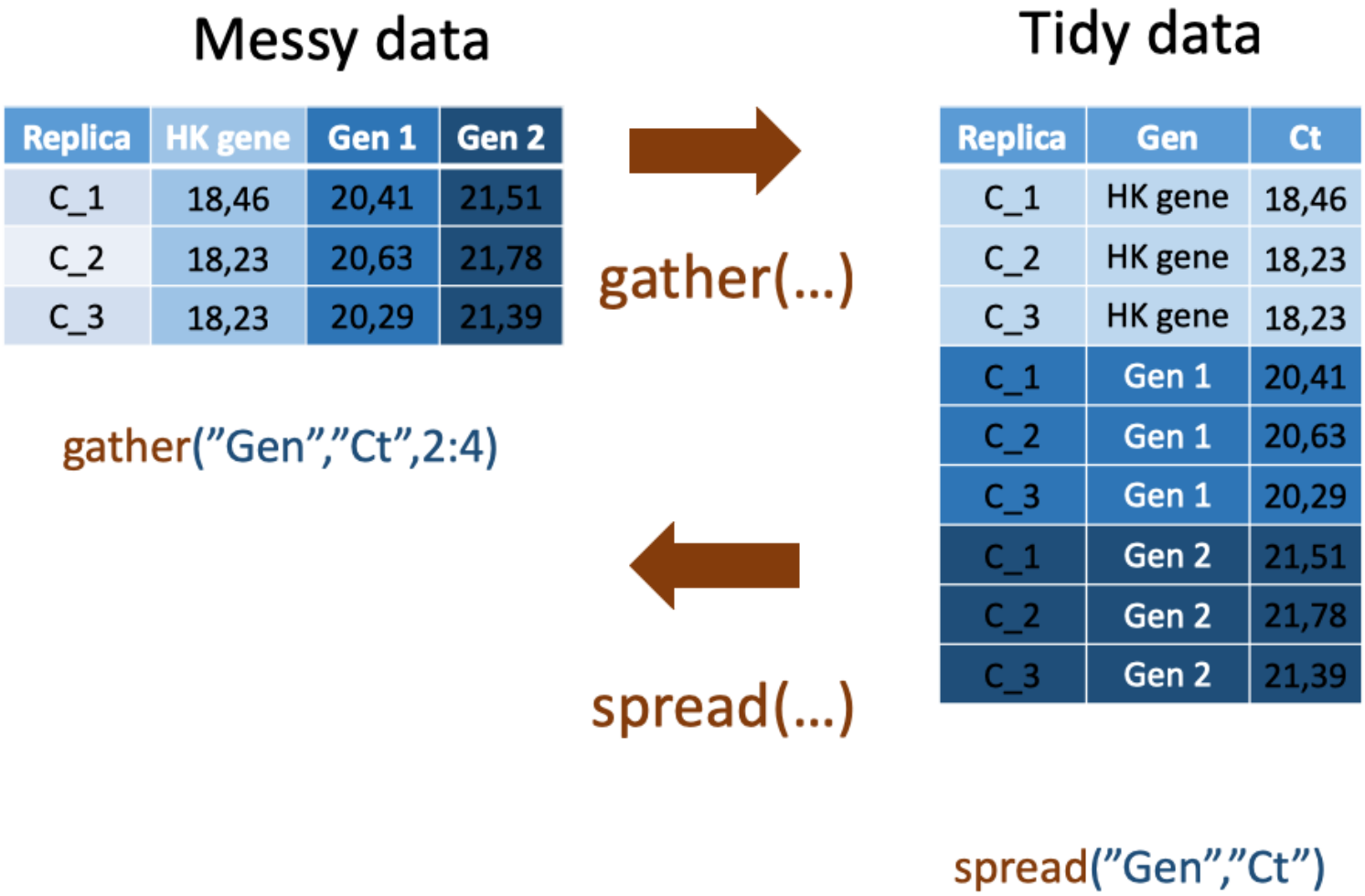
Pros de boxplot + jitter

1. Muestra la escala real de y. 2. Permite detectar valores atípicos.
3. Muestra el número de observaciones.



PAQUETE TIDYR: FUNCIONES AVANZADAS

gather(): Colapsa múltiples columnas para crear tidy data.
spread(): Separa una columna en múltiples columnas.



PRÁCTICA ANÁLISIS DE DATOS

Guía de trabajo en Rstudio.cloud.

RESUMEN DE LA CLASE

- Identificamos y clasificamos variables.
- Observamos la distribución de variables relacionadas a expresión de genes.
- Identificamos la importancia de realizar un análisis exploratorio de datos en expresión de genes.
- Reconocemos preguntas importantes de un EDA: Variación y correlación.
- Realizamos gráficas con ggplot2.