PEC 1 - Ana Alicia Martín García

1. **Selección de un dataset de metabolómica obtenido de un repositorio de github:** https://github.com/nutrimetabolomics/metaboData/

El conjunto de datos utilizado se ha obtenido a partir de un experimento de fosfoproteómica En el experimento se han analizado (3 + 3) modelos PDX de dos subtipos diferentes utilizando muestras enriquecidas con fosfopéptidos. Se ha realizado un análisis LC-MS de 2 duplicados técnicos en cada muestra. El conjunto de resultados consistió en abundancias normalizadas de señales de MS para aproximadamente 1400 fosfopéptidos Objetivo del análisis: **buscar fosfopéptidos que permitan diferenciar los dos grupos tumorales* Los datos se han proporcionado en un archivo de Excel: TIO2+PTYR-human-MSS+MSIvsPD.XLSX

```
##              M1_1_MSS   M1_2_MSS   M5_1_MSS   M5_2_MSS    T49_1_MSS  T49_2_MSS
## O00560       24.29438  44475.964      0.000   6269.141    1135.8169   21933.90
## O00560.1      0.00000  43138.904   2102.056  50355.051     248.9275    3239.16
## O00560.2   3412.60332 172143.040  77323.019 307637.429   98442.2773  192982.37
## O15264   220431.17880 145656.887 104287.815  75887.365  773377.4981  481165.54
## O15264.1  18254.77813   8529.755  35955.901  44102.316   57145.1682   34638.01
## O15551   644513.31840 261938.025 187023.484 124867.715 4487443.6920 2572575.27
##             M42_1_PD   M42_2_PD     M43_1_PD    M43_2_PD    M64_1_PD
## O00560         0.000       0.00     772.9056    2136.746    1820.724
## O00560.1    1315.904       0.00       0.0000       0.000       0.000
## O00560.2   24851.344   16547.95    5565.2821       0.000    3264.563
## O15264   1027196.292 1163747.38 4080239.1820 4885818.113 3093786.793
## O15264.1   21231.256   49499.70  666107.0448  379313.615  255792.117
## O15551    535809.187  434645.89   91361.8781   65997.913  243250.439
##              M64_2_PD
## O00560      1727.9098
## O00560.1     892.3565
## O00560.2    5901.9577
## O15264   2759104.5440
## O15264.1  579765.0018
## O15551    206632.6444
```

```
## # A tibble: 12 x 4
##    Sample...1 Sample...2 Individual Phenotype
##    <chr>      <chr>           <dbl> <chr>
##  1 M1_1       M1                  1 MSS
##  2 M1_2       M1                  1 MSS
##  3 M5_1       M5                  2 MSS
##  4 M5_2       M5                  2 MSS
##  5 T49_1      T49                 3 MSS
##  6 T49_2      T49                 3 MSS
##  7 M42_1      M42                 4 PD
##  8 M42_2      M42                 4 PD
##  9 M43_1      M43                 5 PD
## 10 M43_2      M43                 5 PD
## 11 M64_1      M64                 6 PD
## 12 M64_2      M64                 6 PD
```

2. **Creación de un contenedor del tipo SummarizedExperiment que contenga los datos y los metadatos (información acerca del dataset, las filas y las columnas). La clase SummarizedExperiment es una extensión de ExpressionSet y muchas aplicaciones o bases de datos (como metabolomicsWorkbench) lo utilizan en vez de usar expressionSet.**

```
## class: SummarizedExperiment
## dim: 1438 12
```

```
## metadata(0):
## assays(1): counts
## rownames(1438): O00560 O00560.1 ... Q13283.1 Q9NYF8.12
## rowData names(1): ProteinID
## colnames(12): M1_1_MSS M1_2_MSS ... M64_1_PD M64_2_PD
## colData names(1): SampleID
```

3. **Llevad a cabo una exploración del dataset que os proporcione una visión general del mismo en la línea de lo que hemos visto en las actividades.**

El objetivo de este estudio es encontrar aquellos fosfopéptidos con una expressión diferencial entre los dos grupos tumorales de ratones, los grupos se definen como:
- Grupo MSS: Muestras M1, M5 y T49 - Grupo PD: Muestras M42, M43 y M64 Con dos réplicas por muestra.

Figura 1. El gráfico representa la abundancia de fosfoproteinas segun la muestra. Diferenciado por colores podemos observar los dos grupos tumorales.

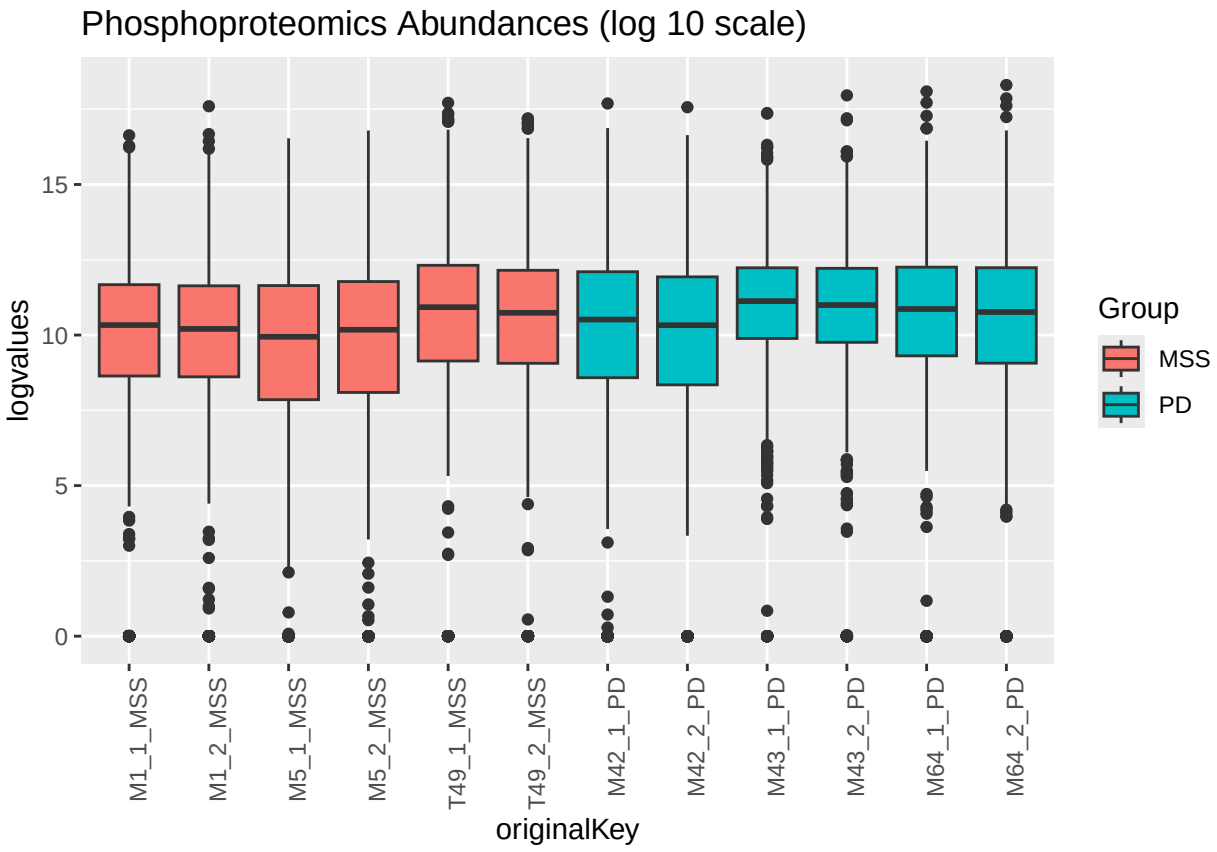


Figura 2. Análisis de Componentes Principales (PCA) sobre los datos de fosfoproteínas. Las muestras que pertenecen a un mismo grupo o condición tienden a agruparse juntas porque comparten características similares en los datos proteómicos. En la figura, observamos como las muestras del grupo con el mismo tipo de tumor (MSS) estan más cercanas entre si, y que respecto al otro grupo (PD) la agrupación no se aprecia más que entre las réplicas.

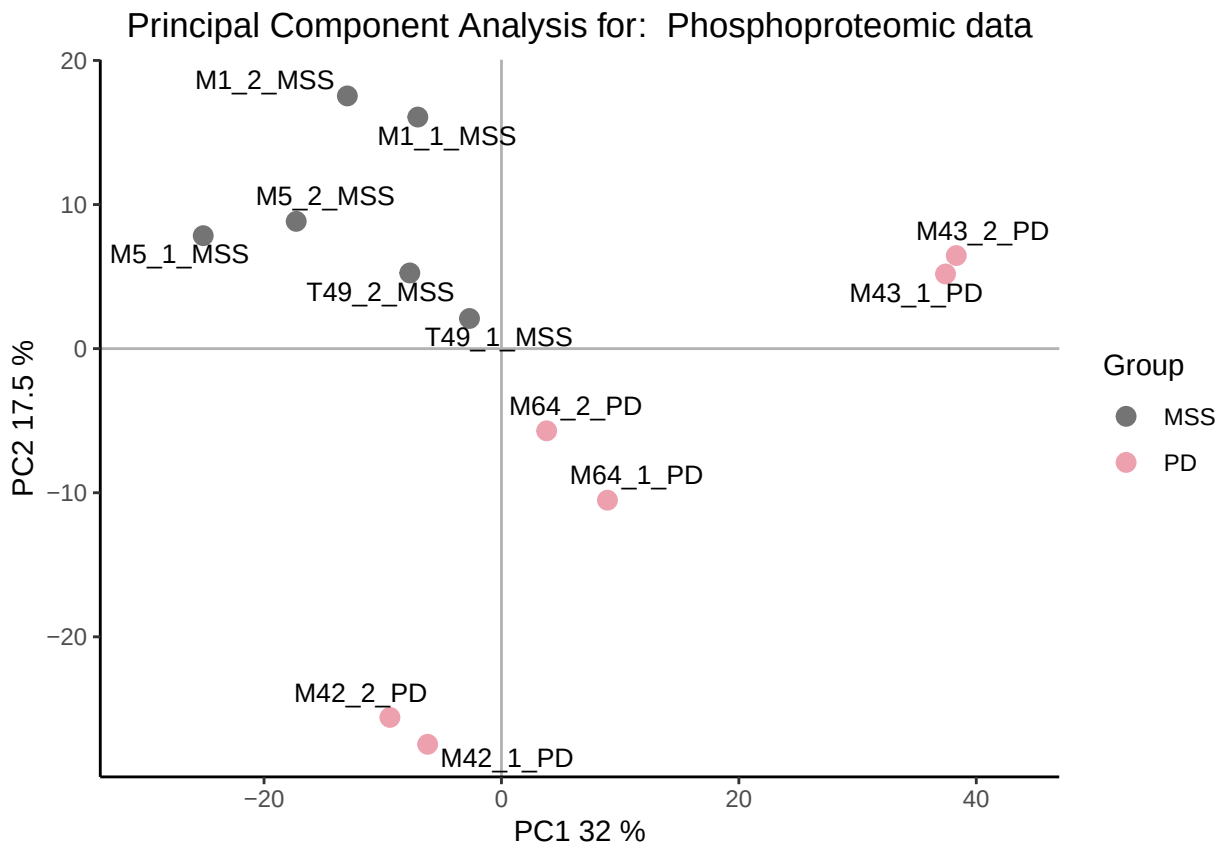


Figura 3. Distribución de los valores de abundancia de las proteínas detectadas.En el gráfico observamos una distribución simetrica en las muestras lo que podria sugerir cierta homogeneidad en los niveles de expresión de las proteínas en las diferentes muestras.

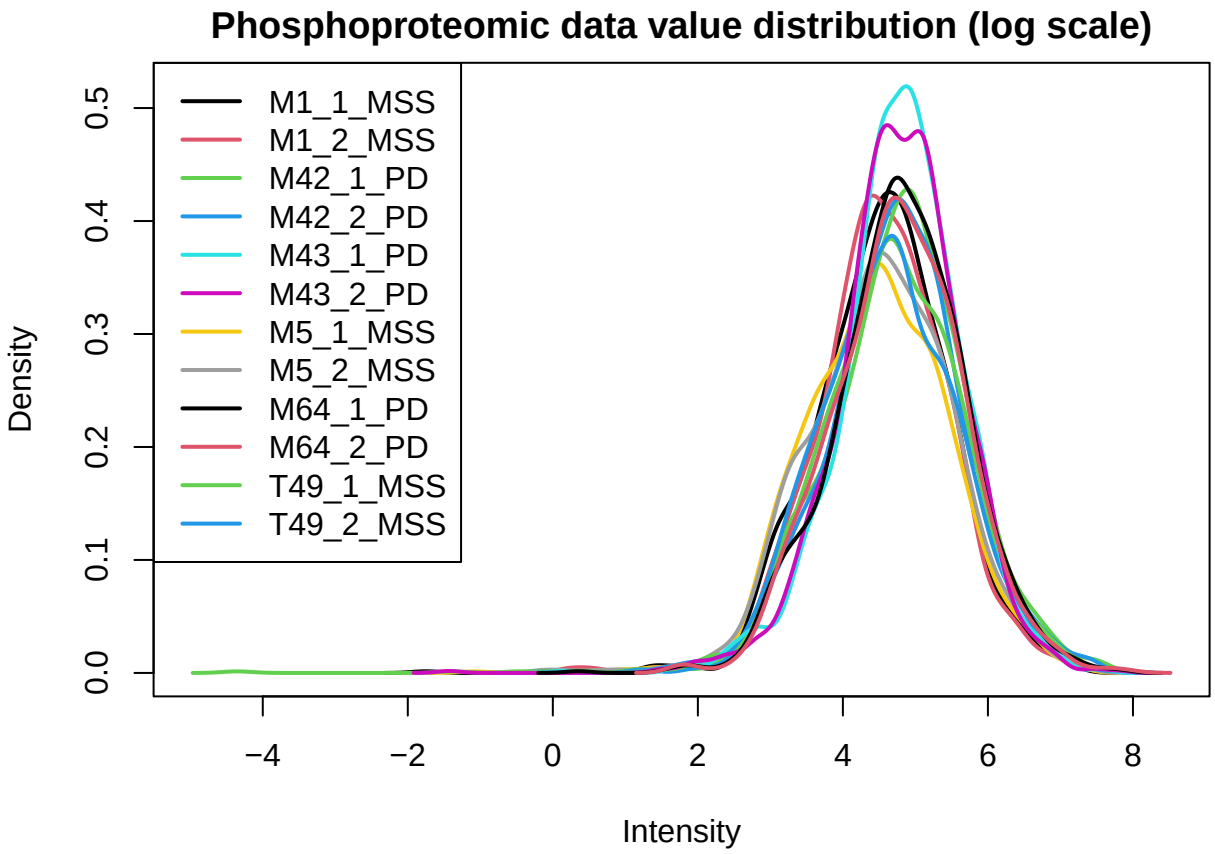
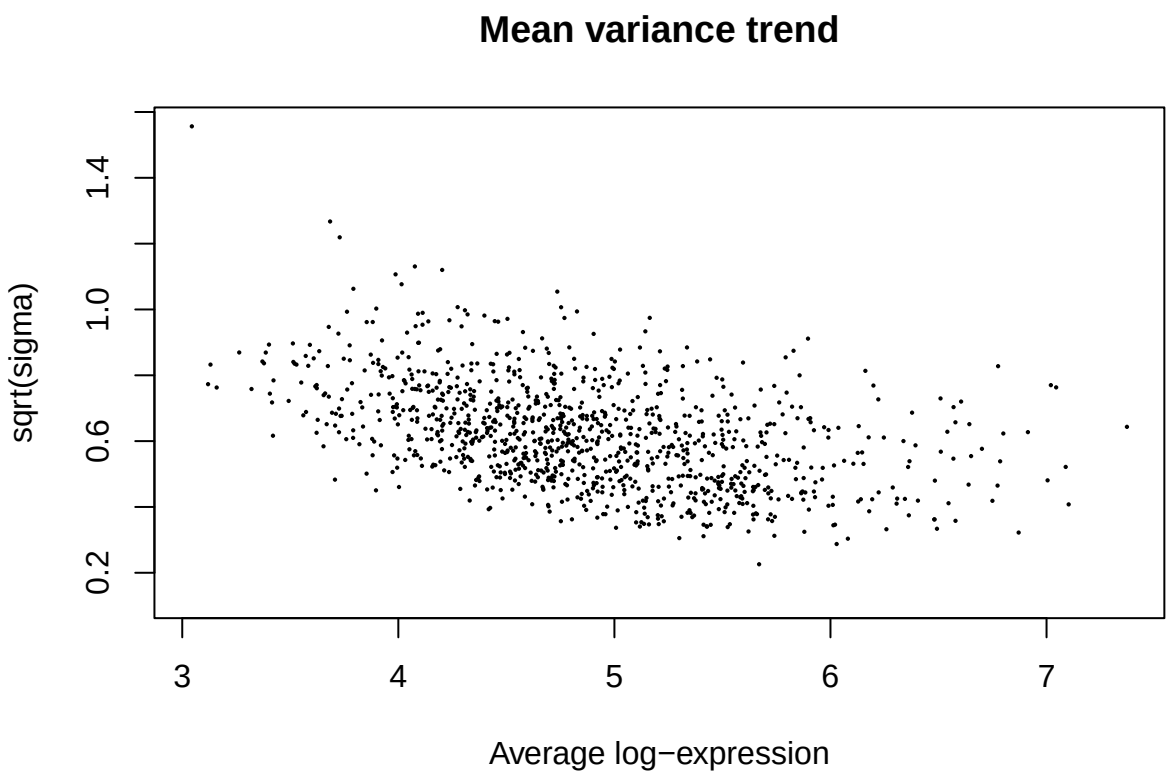


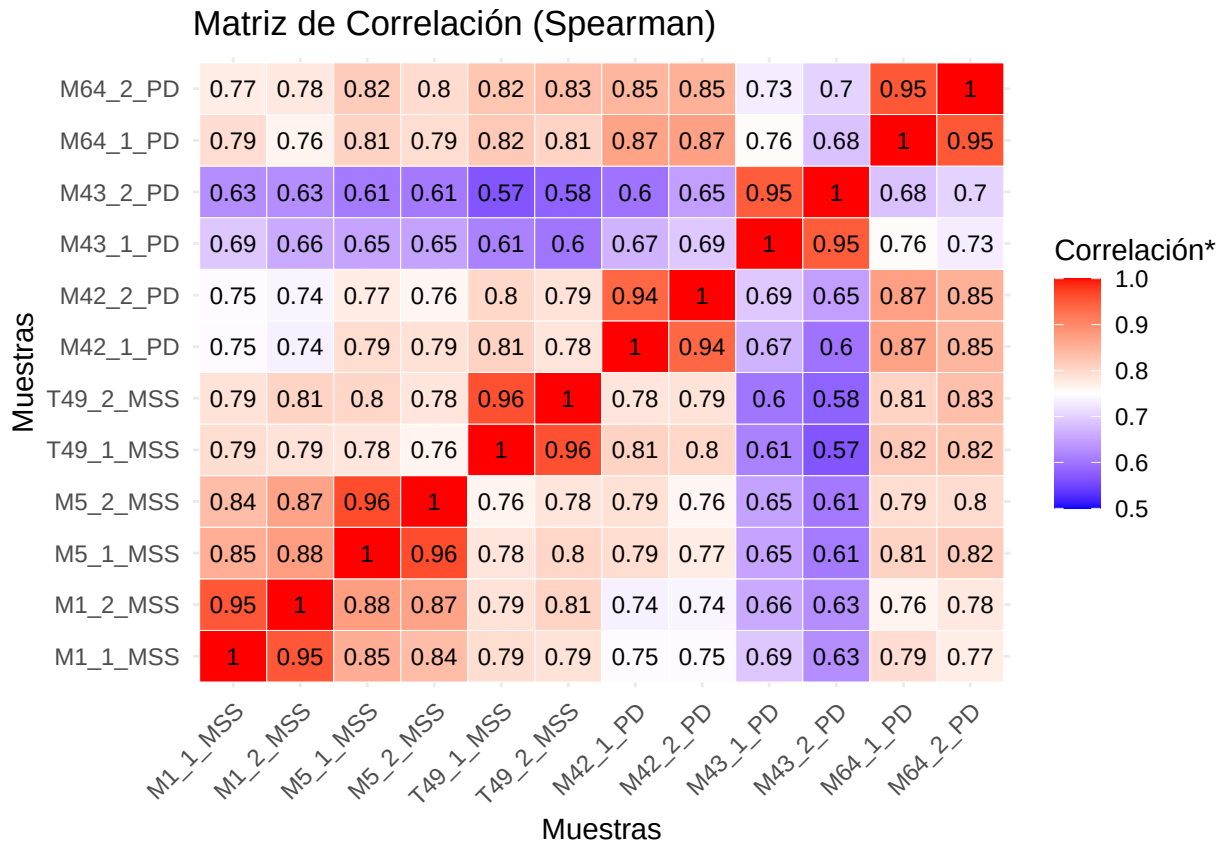
Figura 4. Gráfico de la tendencia de la media y la varianza. Este grafico indica que las proteínas con baja media de expresión (a la izquierda del gráfico) presentan una mayor varianza. Esto suele ser común en datos de expresión, donde las proteínas de baja abundancia suelen mostrar más variabilidad debido a la naturaleza ruidosa de la medición en estos rangos.



Finalmente para evaluar si hay diferencias entre los grupos resulta interesante realizar test estadísticos, en este caso primero realizaremos un test shapiro para evaluar si los valores en nuestras muestras siguen una distribución normal:

```
##     M1_1_MSS     M1_2_MSS     M5_1_MSS     M5_2_MSS    T49_1_MSS    T49_2_MSS
## 1.168445e-60 7.661533e-63 8.179926e-61 4.797945e-61 9.934279e-62 3.148262e-61
##     M42_1_PD     M42_2_PD     M43_1_PD     M43_2_PD     M64_1_PD     M64_2_PD
## 8.761241e-62 7.761307e-62 6.753852e-62 4.033348e-63 7.743173e-63 2.446375e-63
```

Con los p-valores obtenidos sumado a la observación en la distribución de los datos de las figuras anteriores, concluimos que nuestros datos no siguen una distribución normal y es por ello que realizamos un análisis de correlaciones entre las muestras utilizando el método de Spearman:



*Los valores en la correlación de Spearman van del -1 al 1 siendo aquellos más cercanos al 1 los más positivamente relacionados, al tratarse de datos tan correlacionados positivamente los colores se han ajustado a un rango menor del 0.5 al 1 para poder visualizar mejor las diferencias.

En la matriz podemos observar que los datos con una correlación positiva ligeramente mas débil son aquellos correspondientes a la muestra M43 con el resto de muestras.

4. **Elaborad un informe que describa el proceso que habéis realizado, incluyendo la descarga de los datos, la creación del contenedor, la exploración de los datos y la reposición de los datos en github. El nombre del repositorio tiene que ser el siguiente: APELLIDO1-Apellido2-Nombre-PEC1. Por ejemplo, en mi caso el repositorio se llamaría: "Sanchez-Pla-Alex-PEC1"**

5. **Cread un repositorio de github2 que contenga o el informe, o el objeto contenedor con los datos y los metadatos en formato binario (.Rda), o el código R para la exploración de los datos o los datos en formato texto y o los metadatos acerca del dataset en un archivo markdown.**

El repositorio ha sido creado en: https://github.com/anaaliciaUOC/MARTIN-GARCIA-ANAALICIA-PEC1.git