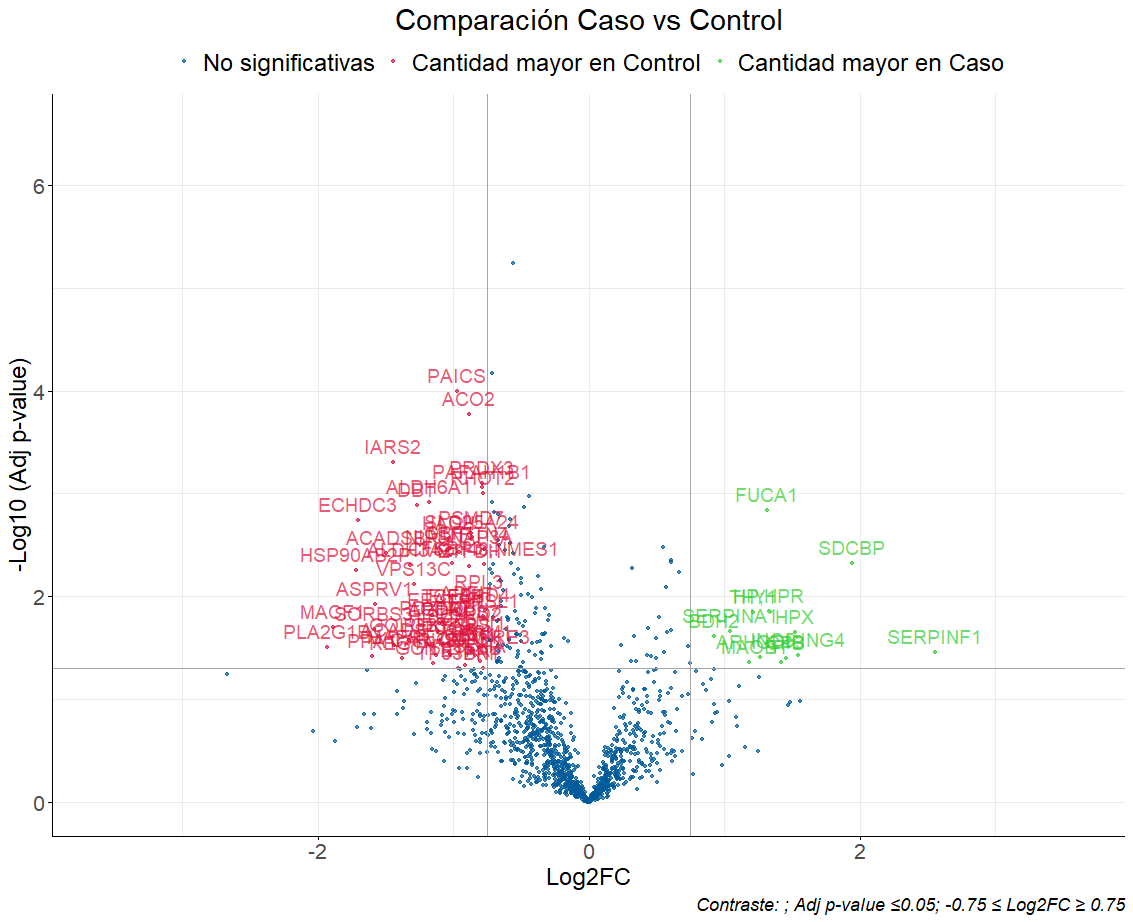
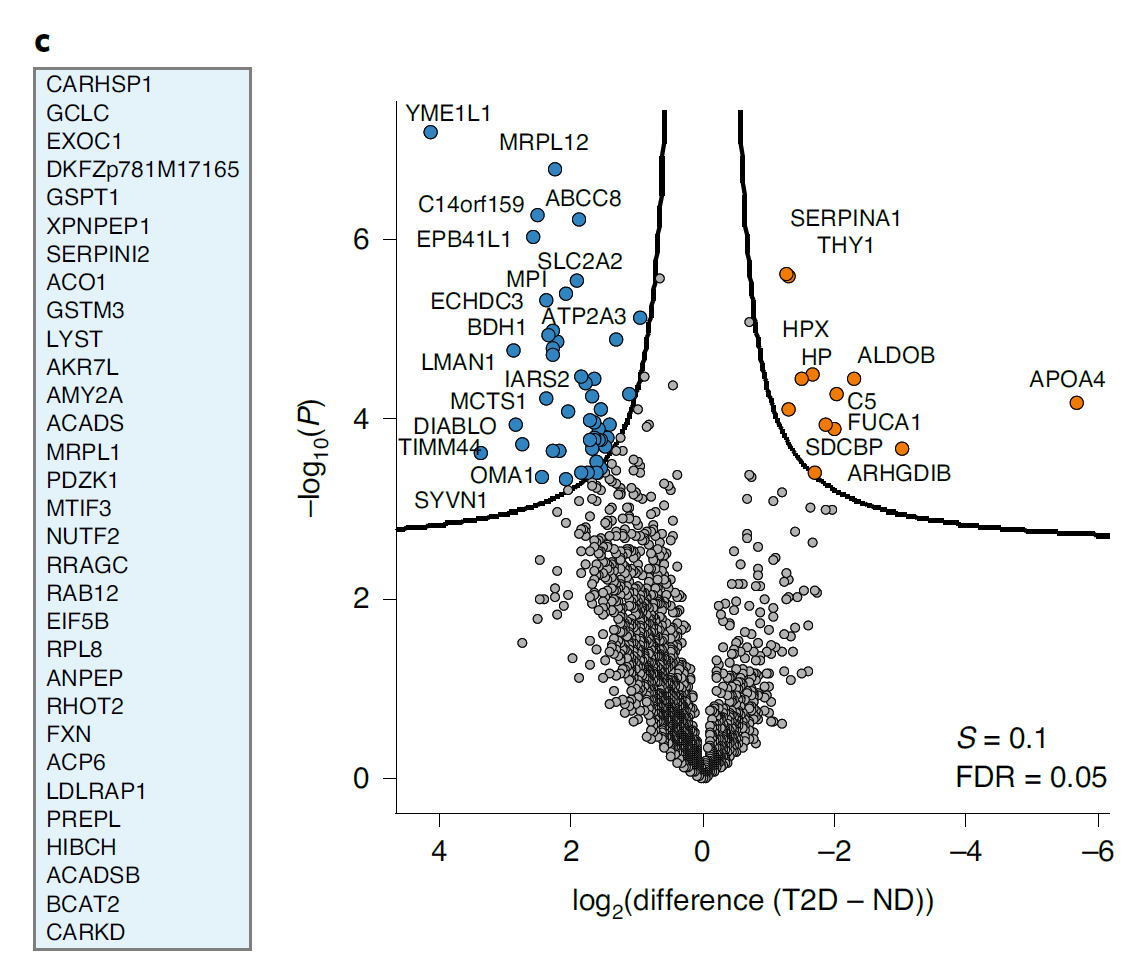
# 944\_401\_Scopus

Primera proba:

* Datos crudo no PRIDE, carpeta SEARCH que é saída de MaxQuant.
* Seleccionamos ProteinGroup.txt, onde hai datos para 40 mostras, as cales 20 pertencen a un grupo (T2DM) e outras 20 a outro (control), e dentro de cada grupo hay 5 replicados biolóxicos e 4 replicados técnicos.
* Analizamos con R seguindo os pasos que se fixeron en Perseus como se di no artículo:
  + Seleccionamos LFQ intensities
  + Eliminamos as proteínas identificadas coa base de datos de proteínas reversas (reverse), only identified by site e posibles contaminantes.
  + Para cada proteína e cada mostra, facemos un promedio dos seus 4 replicados técnicos sempre e cando haxa mínimo datos en 2 dos replicados. Se so hai datos nunha das réplicas, esa proteína considérase como missing value.
* Imputamos dunha distribución normal con width 0.3 e shift 1.8.
* Facemos test t entre T2DM e control, corrixindo por BH e obtemos a diferenza na media de expresión.
* Ao obter o volcano plot:
  + Se o facemos cos valores p axustados e co logFC, non hai proteínas relevantes.
  + Se o facemos con valores p sen axustar e coa diferenza de media de expresión, obtense unha figura parecida á do articulo e con certas coincidencias en proteínas significativas, ainda que con magnitude de diferenza de medias menor. Non entendo o eixo X indicado no volcano plot da Figura 4D: no pe de figura falan de log(FC), pero no eixo din log(difference(T2DM-Control)). Eu co logFC non obteño eses valores nin de broma. Ten pinta de que o que hai representado é a diferenza de medias, pero é imposible facer o log2 destes valores porque os hai negativos, e si se ponen en positivo a forma do volcano non sería esa nin de broma (probado en R).





SDCBP, HPX, SERPINA1, FUCA1, THY1, HP, HPX comunes na parte dereita; IARS2, ECDHDC3, ACADSB (tabla) na izquierda

Segundo intento

* Igual procesado, pero considerando as réplicas como mostras individuais, non vai ser que fose así a análise. Os resultados do volcano plot son moi diferentes, así que descartamos que usaran as réplicas técnicas como mostras individuais.

Análise final:

* Igual procesado, pero na parte de missing data soamente nos quedamos coas proteínas con datos completos. Aquí pode haber diferenzas porque eu considerei que polo menos tiña que haber datos en 2 das 4 réplicas técnicas para promediar.
* Facendo o test e obtendo o volcano con valores p e diferenza de medias siguen coincidindo algunhas proteínas coas do volcano da Figura 4, pero moitas non aparecen alí (polo menos as de datos completos deberían coincidir 100%)
* En fin, quedamosnos con proteínas completas (993 vs 1311 con 70% mínimo de valores en cada grupo), e eses datos son os usados na metanálise.

