ANEXO GSE164416

INFORMACIÓN SUPLEMENTARIA SOBRE ESTE PROYECTO

En el proyecto GSE164416[1] los autores describen como los perfiles multiómicos de células de islotes pancreáticos de donantes humanos revelan trayectorias heterogéneas hacia el desarrollo de la DM2. Uno de los objetivos del estudio era entender los mecanismos moleculares por los cuales las células de islotes, principalmente compuestos por células beta, se volvían deficientes en la secreción de insulina y se deterioraban a medida que progresaba la enfermedad, desde una regulación normal de la glucemia hasta una diabetes declarada con hiperglucemia. Para ello recogieron muestras de tejido pancreático tras cirugía de pacientes pancreatectomizados metabólicamente fenotipados. Se obtuvieron los perfiles de expresión génica de las células beta aisladas de islotes de cada paciente mediante secuenciación del transcriptoma, realizado con tecnología Illumina HiSeq 2500®. Los resultados obtenidos los filtraron y alinearon contra el genoma de referencia hasta obtener la matriz de conteos de genes totales y realizaron los análisis de calidad correspondientes para descartar aquellas librerías que no pasaron un umbral mínimo de calidad. Esta matriz depurada la subieron al repositorio GEO para compartir con la comunidad científica. Del total de participantes se quedaron con 133, los cuales se estratificaron en cuatro grupos en función de su estado diabético; 18 sujetos eran no diabéticos (ND), 41 presentaban alteraciones de la glucemia o prediabetes (IFG o IGT), 35 presentaban diabetes tipo 3c (por pancreatopatías) (T3cD) y 39 presentaban diabetes mellitus tipo 2 (DM2). Los grupos se asignaron en base a los umbrales definidos en las guías de la Asociación Americana de Diabetes (ADA por sus siglas en inglés)[2]. La cohorte de participantes estaba compuesta por un 51’9% hombres y un 48’1% de mujeres, la edad media era de 65’4±11’5 años. Para el análisis de expresión de los genes, primero filtraron para eliminar aquellos genes con niveles de expresión bajos y tras esto realizaron análisis de expresión diferencial entre las categorías ND versus IFG+IGT, ND vs DM2 y ND vs T3cD (comparaciones por pares), usando las funciones del paquete DESeq2[3] implementado en Bioconductor[4] para R software[5 6] para normalizar los datos de expresión y Limma, que emplea un modelo lineal, que al no ser fácilmente aplicable a datos discretos como los generados en un RNA-Seq incluye paso previo de transformación Voom[7], también del paquete Bioconductor[4]. Los genes con un *fold change* (FC) > 1.5 y un FDR (*False Discovery Rate*) <=0.05 fueron considerados diferencialmente expresados (DE) entre los grupos. Posteriormente realizaron el análisis de enriquecimiento funcional de genes en el par DM2 vs ND, mediante GSEA (*weighted gene set enrichment analysis*) en la lista de genes ordenada por orden decreciente. Los términos GO (*gene ontology*) y KEGG (*Kyoto encyclopedia of genes and genomes*) se realizaron empleando la librería ClusterProfiler[8] implementada en R software[5,6].

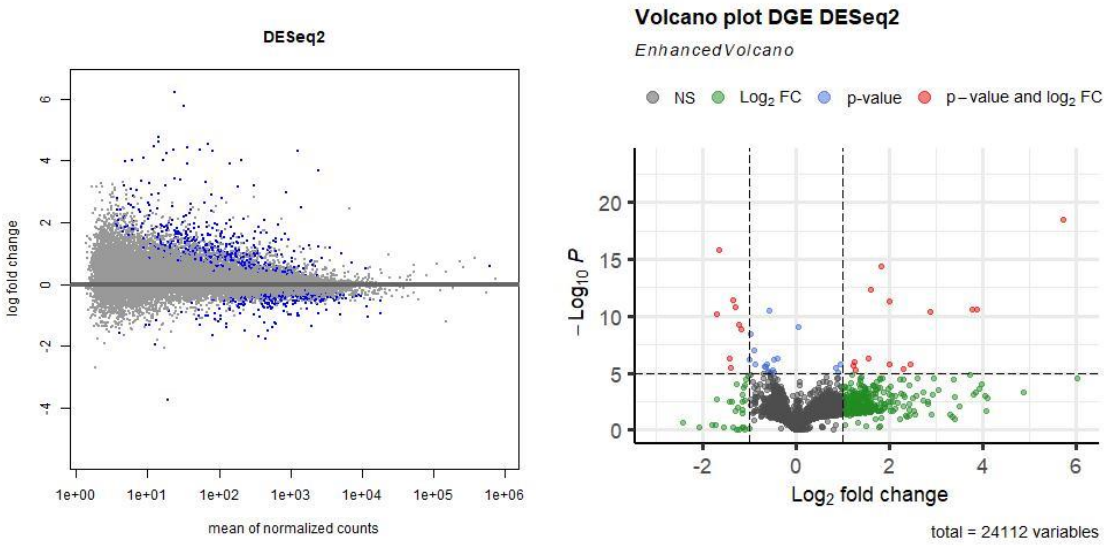
Resultados: Tras los análisis de expresión diferencial de genes observaron una exacerbada desregulación génica en deterioro del control glucémico, donde observaron aproximadamente 650 genes DE entre las categorías DM2 vs ND. Comprobaron que 7 genes conocidos, implicados en presentar un elevado riesgo, para DM2 se encontraban desregulados, siendo los genes *SGSM2* y *BCL2* sobreexpresados y *RASGRP1, G6PC2, SLC2A2 ZMAT4* y *PLUT* subexpresados en los pacientes con DM2. Mientras que el resto de los genes no habían sido previamente reportados en islotes pancreáticos de sujetos con DM2. En los análisis de enriquecimiento observaron vías sobreexpresadas relacionadas con interacción de la matriz extracelular, respuesta inmune y rutas de señalización. Y vías subexpresadas relacionadas con procesamiento del ARN, traducción de proteínas y fosforilación oxidativa mitocondrial.

Análisis de expresión diferencial del proyecto GSE164416 realizado en este TFM

Para el análisis DGE realizado en este trabajo, se empleó el subconjunto de las 57 muestras de interés seleccionadas de la matriz de conteos del proyecto GSE164416 original. Este análisis se llevó a cabo con el fin de comprobar que se obtenían resultados similares a los del trabajo original. Para ello, también se filtraron los datos de expresión por aquellos genes que presentaban bajos niveles de expresión, se seleccionaron solo aquellos genes que presentaban más de 5 conteos en al menos 5 muestras. Los datos de conteos de la matriz filtrada se normalizaron mediante tamaño de librería y fueron transformados por estabilización de las varianzas usando las herramientas del paquete DESeq2[3] implementado en Bioconductor[4]. El análisis de expresión diferencial se llevó a cabo entre las categorías DM2 vs ND, también usando las funciones del paquete DESeq2. Este paquete proporciona las herramientas necesarias para analizar expresión diferencial mediante el uso de modelos de regresión binomial negativos[3]. Para llevar a cabo la normalización, este método divide los conteos de cada gen en una muestra por el número total de lecturas en dicha muestra y para estimar la dispersión toma el valor máximo entre las estimaciones de dispersión individuales y la tendencia media de la dispersión[9]. La prueba estadística que emplea DESeq2 es la paramétrica de Wald. Los genes con un FC > 1 y un FDR ≤ 0,05 fueron considerados estadísticamente significativos (Ver Datos disponibles). Posteriormente se realizó el análisis de enriquecimiento funcional ORA (*over-representation analysis*) de la lista de genes resultado del análisis de expresión diferencial entre DM2 vs ND ordenados en orden decreciente. Los términos GO y KEGG se realizaron empleando la librería ClusterProfiler[8] implementada en R software[5,6].

Resultados: Tras los análisis de expresión diferencial de genes mediante esta metodología se observaron 2358 genes DE entre DM2 y ND (Tabla ResultadosDGE, https://github.com/AnaLagoSampedro/TFM), con 1426 genes sobreexpresados y 932 genes subexpresados que sobrepasaban el umbral p-valor ajustado ≤0.05 (Figura 1A) y siendo 594 los genes que presentaban un FC ≥1 (563 UP y 31 DOWN) (Figura 1B

Volcano-plot). Posiblemente, las diferencias en cuanto al número de genes observados, es porque en este caso se ha sido menos estricto a la hora de filtrar y de obtener los genes estadísticamente significativos. En todo caso, se observaron los mismos genes subexpresados en el análisis original (*RASGRP1, G6PC2, SLC2A2 ZMAT4 y PLUT*) en este análisis también. Así mismo ocurrió con el gen *BCL1* sobreexpresado en pacientes DM2, pero no siendo significativo el *SGSM2* en este caso. En los análisis de enriquecimiento se observaron también las vías relacionadas con interacción de la matriz extracelular y respuesta inmune sobreexpresadas (Figura 2). En el enriquecimiento con KEGG se observan rutas sobreexpresadas relacionadas con señalización celular (vía *PIK3-Akt-mTOR*, interacción citoquinas a su receptor, respuesta a daño oxidativo, rutas de respuesta inflamatoria) reportadas por estar implicadas en DM2 (Figuras 2 A)GO, B)KEGG y C)Reactome).

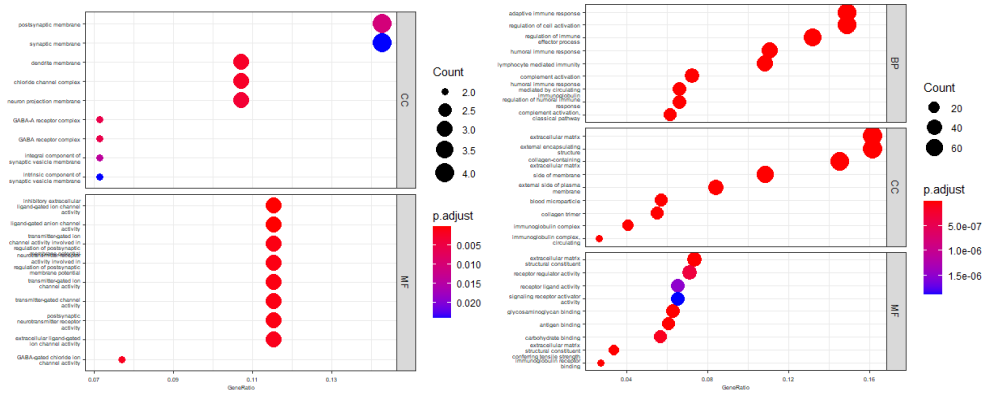


Figuras 1A DESeq2 y 1B Volcano plot.

A) GO

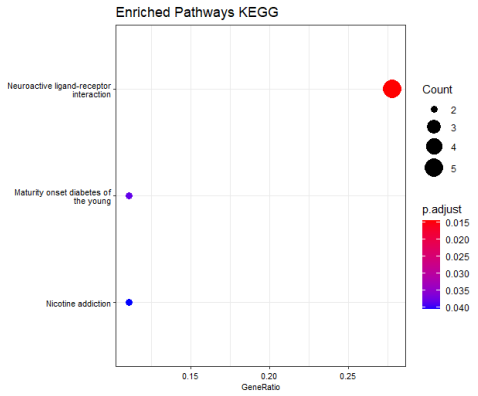
DOWN

UP

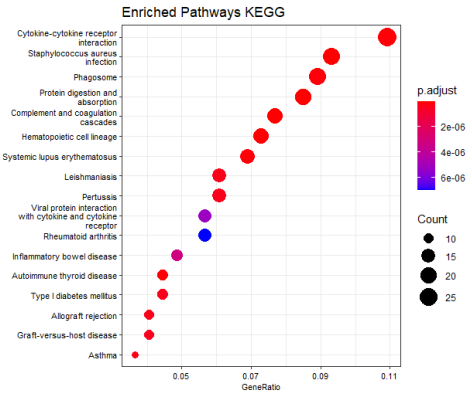


B)KEGG

DOWN

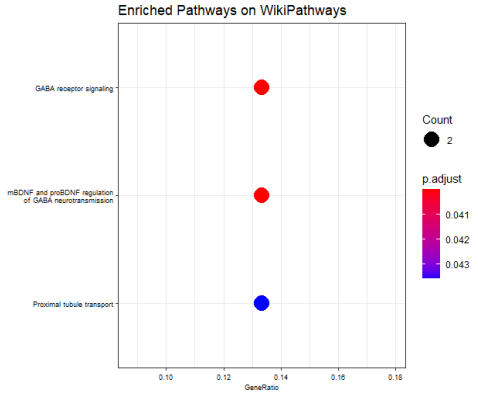


UP

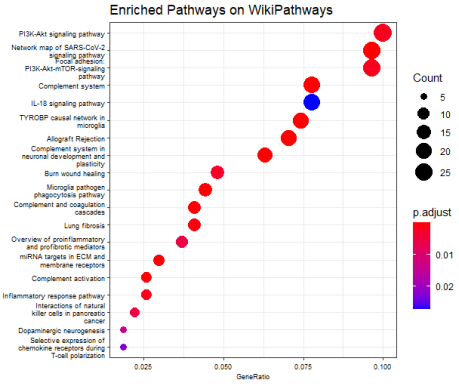


C)REACTOME

DOWN



UP



Figuras 2A GO, 2B KEGG y 2C REACTOME.

REFERENCIAS

1. Wigger, L. *et al.* Multi-omics profiling of living human pancreatic islet 1 donors reveals heterogeneous beta cell trajectories 2 toward type 2 diabetes 3 4. doi:10.1101/2020.12.05.412338.
2. Association, A. D. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* **28**, s37–s42 (2005).
3. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
4. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. (2015) doi:10.1038/NMETH.3252.
5. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
6. R: A Language and Environment for Statistical Computing.
7. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, (2004).
8. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. doi:10.1089/omi.2011.0118.
9. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092 (2019).