PREPROCESAMIENTO DE LAS BASES DE DATOS

# 1.Esquema general

Contamos con las siguientes bases de datos:

## ***ENVEJECIMIENTO***

## RATÓN

* The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature 583, 590–595 (2020). <https://doi.org/10.1038/s41586-020-2496-1> GEO Accession viewer (nih.gov)

## HUMANO

* A Single-Cell Transcriptomic Atlas of Human Skin Aging. Zhiran Zou, Xiao Long, Qian Zhao, Yandong Zheng, Moshi Song, Shuai Ma, Yaobin Jing, Si Wang, Yifang He, Concepcion Rodriguez Esteban, Nanze Yu, Jiuzuo Huang, Piu Chan, Ting Chen, Juan Carlos Izpisua Belmonte, Weiqi Zhang, Jing Qu, Guang-Hui Liu. Developmental Cell, Volume 56, Issue 3, 2021. <https://doi.org/10.1016/j.devcel.2020.11.002>. Los datos está naquí: <https://ngdc.cncb.ac.cn/gsa-human/browse/HRA000395>
* Solé-Boldo, L., Raddatz, G., Schütz, S. et al. Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. Commun Biol 3, 188 (2020). <https://doi.org/10.1038/s42003-020-0922-4>. Los datos están: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130973>.

## ***REPROGRAMACIÓN CELULAR***

## RATÓN

* OSKM y QUÍMICA (MSX1): Antoine Roux, Chunlian Zhang, Jonathan Paw, José Zavala-Solorio, Twaritha Vijay, Ganesh Kolumam, Cynthia Kenyon. Partial reprogramming restores youthful gene expression through transient suppression of cell identity. BioRxiv 2021. <https://doi.org/10.1101/2021.05.21.444556>. (<https://www.biorxiv.org/content/10.1101/2021.05.21.444556v2> )
* OSKM parcial: Guang-Hui Liu, Juan Carlos Izpisua Belmonte. In vivo partial cellular reprogramming enhances liver plasticity and regeneration, Cell Reports,Volume 39, Issue 4, 2022, <https://doi.org/10.1016/j.celrep.2022.110730>.
* QUÍMICA parcial: Acquisition of a Unique Mesenchymal Precursor-like Blastema State Underlies Successful Adult Mammalian Digit Tip Regeneration. VOLUME 52, ISSUE 4, P509-524.E9, FEBRUARY 24, 2020 <https://doi.org/10.1016/j.devcel.2019.12.004>

## HUMANO

* QUÍMICA (2022): Guan, J., Wang, G., Wang, J. et al. Chemical reprogramming of human somatic cells to pluripotent stem cells. Nature 605, 325–331 (2022). <https://doi.org/10.1038/s41586-022-04593-5>. -> También parcial (estado II).
  + Probar si la máquina puede con la concatenación de scRNAseq+scATACseq (por separado y luego, si se puede concatenar —es decir, que se hayan hecho las dos técnicas en las mismas células—, junto)
* OSKM Pluripotente: Novak, G., Kyriakis, D., Grzyb, K. et al. Single-cell transcriptomics of human iPSC differentiation dynamics reveal a core molecular network of Parkinson’s disease. Commun Biol 5, 49 (2022). <https://doi.org/10.1038/s42003-021-02973-7>. La info está en: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183248>. (TRANSCRIPTÓMICA Y PROTEÓMICA) -> CARPETAS VACÍAS
  + <https://www.biorxiv.org/content/10.1101/2021.09.01.458444v1.full> POSIBLE IDEA

# 2.Envejecimiento

## 2.1 Atlas de ratón

**The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature 583, 590–595 (2020).**

Encontramos datos que forman un atlas de ratón (mus musculus) seq-RNA de 23 tejidos y órganos. Además de los datos en bruto, existe una herramienta interactiva para explorar el conjunto <https://tabula-muris-senis.ds.czbiohub.org/>.

Los datos en bruto están en [Processed files (to use with scanpy) (figshare.com)](https://figshare.com/articles/dataset/Processed_files_to_use_with_scanpy_/8273102/2).

Tal y como hemos dicho, son referentes a 23 tejidos y órganos. Por cada muestra encontramos dos archivos .h5ad: facs y droplet.

El archivo facs contiene los resultados extraídos con citometría de flujo en Standford, mientras que el droplet contiene datos referidos a los cuerpos lipídicos?.

Leemos el archivo .h5ad y extraemos los datos necesarios. Por un lado, sacamos los barcodes (células) y gene\_id (nombres o id de los genes). En cuanto a la etiqueta, las células son jóvenes (3 meses) o adultas (24 meses), por lo que creamos un vector columna con la codificación 0 si es joven y 1 si es adulta. Finalmente, sacamos la matriz de conteos. **FALTA LA MATRIZ.**

## 2.2 Atlas de piel de humanos

**A Single-Cell Transcriptomic Atlas of Human Skin Aging.**

Tenemos un conjunto de datos de células de la piel humana, del contorno de alrededor de los ojos (dado que funcionan como marcador de la edad biológica), de mujeres sanas y edades variadas. Hay 11 tipos celulares (BC (KRT14+), mitotic cell (MC, MKI67+), vellus hair follicle cell (VHF, SOX9+, KRT6B+, and SFRP1+), melanocyte (ME, TYR+ and DCT+), SC (KRT10+), GC (FLG+), endothelial cell (EC, CLDN5+), immune cell (IC, PTPRC+), fibroblast (FB, PDGFRA+), pericyte (PT, RGS5+ and PDGFRB+) and orbicularis oculi muscle (OOM, CKM+) (S1J, and S1K)).

Los factores de transcripción controlan el envejecimiento, concretamente la inactivación de HES1 en los fibroblastos y KLF6 en los queratinocitos lleva a la senescencia (inactivos = envejecidos).

Se analizaron 35,678 células de “jóvenes”, “medianas” y “adultas” (3 etiquetas). Hay varios conjuntos para cada edad:

* Young (Y): edades 18, 22 y 23, cada uno con 4 conjuntos.
* Middle (M): edades 44, 47 y 48, cada uno con 4 conjuntos a su vez.
* Old (O): edades 70, 73 y 76, con 4 conjuntos a su vez.

Los datos se extraen de <https://ngdc.cncb.ac.cn/gsa-human/browse/HRA000395>**. ERROR DESCARGANDO, NO TERMINA.**

## 2.3 Atlas de piel de humanos II

**Solé-Boldo, L., Raddatz, G., Schütz, S. et al. Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. Commun Biol 3, 188 (2020).**

Se analizan 15000 células protegidas del sol divididas en joven (25 and 27 y/o) y adultas (53–70 y/o) de un hombre caucásico. Hablan de diferentes células de la piel, con varios tipos celulares, pero en el paper se centran en los fibroblastos.

Los datos están en <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130973>. Tenemos datos asociados a la salida del Cell Ranger filtrada y en bruto, concretamente genes\_id, barcodes y matrix (archivos .tsv y .mtx comprimidos). Según la página de información de Cell Ranger, las salidas:

* Unfiltered gene-barcode matrices: Contains every barcode from fixed list of known-good barcode sequences. This includes background and non-cellular barcodes.
* Filtered gene-barcode matrices: Contains only detected cellular barcodes.
* FILTERED MATRIX: Raw sequencing data was processed with Cell Ranger, version 2.1.0, from 10X Genomics. 16,062 cells passed the quality control steps performed by Cell Ranger. To remove possible cell doublets, we filtered out cells with more than 7500 expressed genes, and to remove potential apoptotic cells we discarded cells with more than 5% mitochondrial reads. The application of these filters resulted in a final dataset of 15,457 single-cell transcriptomes. **INFORMACIÓN DEL PAPER**

Por lo tanto, podemos aplicarlo a la matriz, y usamos la filtrada, para tener los conteos de las células detectadas y no tener en cuenta aquellas que han sufrido apoptosis.

Concretamente, encontramos información de 5 samples, dos asociadas a un hombre joven (y1=25y, y2=27y) y tres de adultos (o1=53y, o2=70y, o3=69y).

Extraemos los barcodes y genes\_id, al igual que la matriz de conteos leyendo y preprocesando los archivos. Analizando los barcodes y junto a la información del paper, vemos que gracias a dichos barcodes podemos saber a qué muestra pertenece cada célula. Concretamente,

* subject: y1, 25y/o -> the cell barcode of this samples for the processed data files is -1
* subject: y2, 27 y/o -> the cell barcode of this samples for the processed data files is -2
* subject: o1, 53 y/o -> the cell barcode of this samples for the processed data files is -3
* subject: o2, 70 y/o -> the cell barcode of this samples for the processed data files is -4
* subject: o3 -> the cell barcode of this samples for the processed data files is -5

Usamos esta información para crear la etiqueta Y, asignando un 0 a las muestras de los dos hombre jóvenes, y 1 a las muestras de los tres hombres mayores.

# 2.Reprogramación

## 2.1 OSKM y MSX1 ratón: TFG

**Partial reprogramming restores youthful gene expression through transient suppression of cell identity.**

## 2.2 OSKM Parcial de ratón

**Guang-Hui Liu, Juan Carlos Izpisua Belmonte. In vivo partial cellular reprogramming enhances liver plasticity and regeneration, Cell.**

Se utilizan los factores de Yamanaka en hepatocitos (células somáticas) para reprogramarlos de forma parcial in vivo (4F). Son datos de ratón (Mus musculus).

Encontramos los datos en [GEO Accession viewer (nih.gov)](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144600).

Hay dos archivos .h5, los cuales contienen la información relativa a las células de control (no reprogramadas, “dox minus”), y de las células reprogramadas (“dox plus”). Preprocesamos ambos archivos y extraemos las matrices de conteo, barcodes y gene\_id asociadas a cada tipo de células (reprogramadas o no).

En cuanto a la etiqueta, la generamos en función de si las células están reprogramadas o no, codificándolo de forma que si no están reprogramadas la salida sea 0, mientras que si están reprogramadas la salida sea 1.

En One Drive tenemos dos matrices de conteos, dos barcodes, dos gene\_id y dos variables Y. Únicamente tenemos que concatenar estos archivos antes de procesarlos (por la RAM de mi ordenador no podía).

## 2.3 Química Parcial ratón

**Acquisition of a Unique Mesenchymal Precursor-like Blastema State Underlies Successful Adult Mammalian Digit Tip Regeneration.**

En este experimento se usan ratones C57BL/6 en su mayoría, pero también encontramos otros dos organismos Dmp1CreERT2 y R26-LSL-TdT. Se pretende demostrar que se puede regenerar el extremo distal de los dedos de ratón. Concretamente, se basa en reprogramar células mesenquimales hasta que expresen Pdgfra, estado en el que crea un blastema y así tienen capacidad de regeneración.

Encontramos datos de células uninjured, regenerating, non regenerative y otras embriónicas de las patas traseras, embrionicas de la parte distal de las patas traseras y postnatal de las patas traseras. Cada dataset presenta un tiempo específico de reprogramación (DPA), tejido y edad.

Este experimento se equipara al realizado en humanos (siguiente, 2.4) (mismas moléculas).

## 2.4 Química Parcial humano

**Chemical reprogramming of human somatic cells to pluripotent stem cells. Nature 605, 325–331 (2022).**

En este experimento se reprograman químicamente diferentes tipos de células somáticas para llevarlas a un estado de pluripotencia. Concretamente, en seq-RNA las muestras son:

* HEFs-0330: Muestra de fibroblastos.
* hADSCs-0618: Muestra de adipocitos
* hCiPSCs: células madre pluripotentes humanas inducidas químicamente a partir de células mesenquimales humanas adultas derivadas del tejido adiposo
* stage I, stage II, stage III, stage IV

Los datos se pueden descargar de [GEO Accession viewer (nih.gov)](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178325). Dentro del archivo principal encontramos la información asociada a cada tipo celular, tanto de las células somáticas como de las células reprogramadas.

Extraemos las matrices de conteos, barcodes y genes\_id para cada tipo celular (fibroblastos embrionarios, células madre mesenquimales adultas derivadas de tejido adiposo y fibroblastos dérmicos adultos).

Para dar la etiqueta nos vamos a basar en si las células están reprogramadas o no. Aquellas células en el primer estado (sin estar en cultivo todavía o llevando 1 día) se consideran no reprogramadas, y le damos una etiqueta de valor 0. Por otro lado, las células en el último estado de cultivo se consideran reprogramadas, y le damos una etiqueta 1.

En el One Drive, dentro de cada línea celular encontramos dos matrices y dos genes\_id asociadas a las células reprogramadas y sin reprogramar. La etiqueta Y se encuentra como una columna más dentro de las matrices X.

## 2.5 OKSM Pluripotente humano

**Single-cell transcriptomics of human iPSC differentiation dynamics reveal a core molecular network of Parkinson’s disease. Commun Biol 5, 49 (2022).**

En este experimento, diferentes délulas de la piel, concretamente fibroblastos que presentan mutaciones del gen PINK1 (se piensa que puede estar relacionado con el Parkinson) se han reprogramado mediante Sendai virus (para transferir los factores de Yamanaka), generando IPSC (pluripotentes) y secuenciadas, para luego diferenciarse a neuronas.

Encontramos los datos en [GEO Accession viewer (nih.gov)](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183248).

17608/6 cell line de control (sin mutación, creo que si), se van diferenciando en paralelo con la muestra de control. Podríamos analizar las de control los primeros días (no hay sin reprogramar ni día 0 sino que es del día 6 -> 17608/6 D06) y ya reprogramadas, iPSC (17608/6 iPSC)

**CARPETAS DE DATOS VACÍAS AL DESCARGARLAS**