RESUM TFM PRIMERES SETMANES

El workflow a seguir està clar:

1. Carreguem les dades
2. Quality Control (QC)
3. Normalitzem i escalem
4. PCA & Nearest Neighbor Graph
   1. Clustering
   2. Low Dimensional Embedding
5. Avançat
   1. Cluster Marker Analysis
   2. Cell Type Identification
   3. Sub-clustering
   4. Trajectories
   5. Cell Cycle Analysis
   6. Cluster DGE

He aconseguit carregar les dades correctament (sembla), així que per aquesta banda molt bé, i entenc tot el que es fa de crear l’objecte, què conté l’objecte creat, etc.

Arriba el control de qualitat, aquí em perdo una mica més.

**Terminologia bàsica de scRNA:**

* Què són UMI counts? 🡪 Unique Molecular Identifiers (UMI): Molecular barcodes ares hort sequences used to uniquely tag each molecule in a sample library. Can be applied to detect and quantify the unique transcripts.
* Features = genes
* Barcodes = Single-cell sequencing experiments uses hort DNA barcode ‘tags’ to dientify reads that originate from the same cell . Barcode are unique to each cell.
* Count Matrix/Feature-Barcode Matrix/Gene-Barcode Matrix = A matrix of counts representing the number of unique observations of each feature within each cell barcode.
* Doublets = when two cells are encapsulated into one reaction volume.
* nCount🡪 són els trànscripts o molècules. Is the total number of molècules detected within a cell.
* nFeature 🡪 Counts de gens. Number of genes detected in each cell.

**Understanding a Seurat object structure**

Informació sobre cada cèl·lula, nombre de “features”, informació sobre les cèl·lules, etc. Al objecte meta.data

Les reduccions de dimension al slot de “reduction”

L’slot commands conté totes les comandes que s’han corregut al objecte Seurat.

**Quality Control**

És important veure el nombre de “features” i d’UMIs per cada cèl·lula. Aquests valors ens donen informació de quan hi ha cèl·lules amb baixa qualitat, ja que aquestes tindran uns valors baixos de UMIs detectats.

Quan tenim valors molt alts, això pot ser degut a Doublets o que en realitat són dues cèl·lules que es seqüencien juntes (s’hauria de filtrar).

El percentatge de gens mitocondrials, és important mirar-ho perquè les cèl·lules mortes tindran un alt percentatge de gens mitocondrials, així doncs aquelles cèl·lules que tinguin alts % de gens mitocondrials, voldrà dir que estan mortes, llavors aquestes ens estaran contaminant les mostres.

Valors baixos de nFeature\_RNA per una cèl·lula ens indicaria que aquesta cèl·lula està morta o en procés de morir o és un espai buit. Alts valors d’nCount\_RNA o/i nFeature\_RNA ens indiquen que aquella cèl·lula realment és un doublet o multiplet.

Gràfica nFeature\_RNA Vs nCount\_RNA 🡪 Aquesta, quan tenim punts que van a baix a la dreta, ens indiquen que són cèl·lules amb alts valors de trànscrips/molècules i pocs gens. El que ens indicaria que a lo millor es detecten les mateixes molècules d’RNA que es van seqüenciant i per això tens uns alts valors de transcripts, però el nombre de counts és baix (pocs gens diferents). Quan tenim punts a dalt a l’esquerre de la gràfica, ens indicaria que tenim alts continguts de gens, però aquests no són suficientment profunds o llargs com per ser identificats com a molècules (?).

Filtering, quin filtratge seria adequat aplicar???

nFeature\_RNA > 200 & nFeature\_RNA < 2500 (??) / percent.mt < 5 (aquest està clar) 🡪 Amb aquest filtre passem de 4272 cèl·lules a només 18.

He provat amb nFeature\_RNA < 4000 així obtinc fins 60 cèl·lules. El problema que li veig a les dades de J0 és que hi ha molts de transcripts/molècules, per molts de gens, semblaria ser que a lo millor hi ha duplets ?? A més, es veu que hi ha bastant de RNA ribosomal, què fem amb això? Jo pel que tinc entès no va gaire bé treure-ho.

nFEature < 5000 🡪 223 cèl·lules

A més, què són aquests gens que tenen el nom TsB??? Són marcadors de cèl·lules?