In [1]:
```
# Description:
"""
This script processes a single-nucleus RNA-seq dataset from afca, removes cells
and filters cell types with fewer than 200 total cells or fewer than 100 cells p
it filters out lowly expressed genes, while retaining those expressed in at leas
Each processed cell type is saved as a separate .h5ad file.
"""

# Import the libraries
import os
import pandas as pd
import scanpy as sc
import anndata as ad
import numpy as np
import re
```

In [2]:
```
# Set the functions
def split_by_batch_prefix(name):
    """
    Splits the index into two parts:
    - Part before 'AFCA' or 'FCA'
    - The rest starting with 'AFCA' or 'FCA'
    """
    match = re.search(r'(AFCA|FCA)', name)
    i = match.start()
    return name[i:]
```

In [3]:
```
# load the dataset data and get the metadata
adata = ad.read_h5ad("/hpc/shared/onco_janssen/dhaynessimmons/projects/ageing_fl
```

In [4]:
```
# print out the basics
print([i for i in adata.obs.columns])
print("shape of full data: ", adata.shape)
#value couns for important columns
print("\nAge value counts: ", adata.obs["age"].value_counts())
print("\nSex value counts: ", adata.obs["sex"].value_counts())
print("\nDataset value counts: ", adata.obs["dataset"].value_counts())
```

```
['tissue', 'sex', 'age', 'sex_age', 'n_genes_by_counts', 'total_counts', 'total_c
ounts_mt', 'pct_counts_mt', 'log1p_n_genes_by_counts', 'log1p_total_counts', 'log
1p_total_counts_mt', 'dataset', 'fca_annotation', 'afca_annotation', 'afca_annota
tion_broad']
shape of full data:  (276273, 15992)

Age value counts:  age
5     96594
30    84496
70    49963
50    45220
Name: count, dtype: int64

Sex value counts:  sex
female    148049
male      123879
mix         4345
Name: count, dtype: int64

Dataset value counts:  dataset
AFCA    179679
FCA      96594
Name: count, dtype: int64
```

In [5]:
```
# Get the unique afca cell types that are part of the enterocyte lineage
entero_subtypes = []
for cell_type in sorted(adata.obs['afca_annotation'].unique().tolist()):
    if "enterocyte" in cell_type:
        entero_subtypes.append(cell_type)
print("\nEnterocyte subtypes: ", entero_subtypes)
```

```
Enterocyte subtypes:  ['adult differentiating enterocyte', 'adult midgut enterocy
te', 'enterocyte of anterior adult midgut epithelium', 'enterocyte of posterior a
dult midgut epithelium', 'enterocyte-like']
```

In [6]:
```
# get cells where sex is neither F nor M
mix_adata = adata[(adata.obs.sex != "female")&(adata.obs.sex != "male")]
print(mix_adata.obs.shape)
```

```
(4345, 15)
```

In [7]:
```
# Remove them from the dataset
mf_adata = adata[~(adata.obs.index.isin(mix_adata.obs.index))&(adata.obs['afca_a
# Only keep the cells that are in teh enterocyte lineage
mf_adata = mf_adata[mf_adata.obs['afca_annotation'].isin(entero_subtypes)]

# get the indiv from the row name
mf_adata.obs['indiv'] = mf_adata.obs.index.map(lambda x: split_by_batch_prefix(x
print(mf_adata.obs.head())
```

```
                                                tissue     sex age    sex_age  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1   body  female  30  female_30
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1   body  female  30  female_30
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1   body  female  30  female_30
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1   body  female  30  female_30
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1   body  female  30  female_30

                                            n_genes_by_counts  total_counts  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1                1288        3819.0
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1                1167        3274.0
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1                1506        5080.0
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1                1206        3539.0
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1                1301        3662.0

                                            total_counts_mt  pct_counts_mt  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1              0.0       0.000000
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1              3.0       0.091631
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1              3.0       0.059055
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1              0.0       0.000000
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1              1.0       0.027307

                                            log1p_n_genes_by_counts  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1                 7.161622
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1                 7.063048
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1                 7.317876
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1                 7.095893
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1                 7.171657

                                            log1p_total_counts  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1            8.248006
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1            8.094073
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1            8.533263
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1            8.171882
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1            8.206038

                                            log1p_total_counts_mt dataset  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1                0.000000    AFCA
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1                1.386294    AFCA
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1                1.386294    AFCA
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1                0.000000    AFCA
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1                0.693147    AFCA

                                           fca_annotation  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1            nan
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1            nan
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1            nan
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1            nan
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1            nan

                                                                                afca_a
nnotation  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1  enterocyte of posterior adult midgut e
pithelium
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1   enterocyte of anterior adult midgut e
pithelium
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1  enterocyte of posterior adult midgut e
pithelium
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1  enterocyte of posterior adult midgut e
pithelium
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1  enterocyte of posterior adult midgut e
```

```
pithelium

                                                afca_annotation_broad  \
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1       epithelial cell
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1       epithelial cell
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1       epithelial cell
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1       epithelial cell
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1       epithelial cell

                                                                 indiv
AAAGTGAGTACTCGAT-1_AFCA_female_body_30_S1  AFCA_female_body_30_S1
AACGAAATCGTTAGTG-1_AFCA_female_body_30_S1  AFCA_female_body_30_S1
AAGGTAAGTAGCACAG-1_AFCA_female_body_30_S1  AFCA_female_body_30_S1
AATCGACGTCTCACGG-1_AFCA_female_body_30_S1  AFCA_female_body_30_S1
AATGAAGGTTGGGTAG-1_AFCA_female_body_30_S1  AFCA_female_body_30_S1
/tmp/ipykernel_436793/3865041347.py:7: ImplicitModificationWarning: Trying to mod
ify attribute `.obs` of view, initializing view as actual.
  mf_adata.obs['indiv'] = mf_adata.obs.index.map(lambda x: split_by_batch_prefix
(x))
```

In [8]:
```python
# Get the observation dataframe as a pandas dataframe
mf_adata_obs = mf_adata.obs.copy()
print(type(mf_adata_obs))
cell_list = []

# Set the save path
save_path = "/hpc/shared/onco_janssen/dhaynessimmons/projects/ageing_flies/data/
os.makedirs(save_path, exist_ok=True)

# Gene list of interest
gene_list = [
    "Su(var)205", "Su(var)3-9", "G9a", "HP1b", "HP1c", "HP4",
    "HP5", "HP6", "ADD1", "Su(var)2-HP2", "Su(var)3-7", "Lam",
    "LamC", "LBR", "Kdm4A", "Kdm4B", "His2Av", "His3.3A", "His3.3B"
]
```

```
<class 'pandas.core.frame.DataFrame'>
```

In [9]:
```python
# Evaluate the QC of the new adata object
cell_cnt = mf_adata_obs.shape[0]
print("number of cells: ", cell_cnt)
print("Min number of genes expressed : ", mf_adata_obs.n_genes_by_counts.min())
```

```
number of cells:  1361
Min number of genes expressed :  263
```

In [10]:
```python
# Check that each age group has at least 100 cells
age_grouped = mf_adata_obs.groupby('age', observed=False).size()
min_value = age_grouped.min()
print("Minimum number of cells in an age group: ", min_value)
if min_value < 100:
    print("Not enough cells in an age group to proceed with analysis")
else:
    print("Sufficient cells in each age group to proceed with analysis")
    # Create a new adata object with the cell type data
    cell_list.append(cell_type)
del mf_adata_obs
```

```
Minimum number of cells in an age group:  161
Sufficient cells in each age group to proceed with analysis
```

In [11]:
```
# Create a new adata object
cell_group_adata = mf_adata
print("\nshape of cell type data: ", cell_group_adata.shape)

# ----------- Custom gene filtering starts here ------------ #

# Compute how many cells express each gene
gene_expression_counts = np.array((cell_group_adata.X > 0).sum(axis=0)).flatten(

# Get gene names
gene_names = pd.Index(cell_group_adata.var_names)

# Genes expressed in >= 3 cells
genes_expressed_enough = gene_expression_counts >= 3

# Create a boolean mask to keep genes that are either:
# - expressed in enough cells
# - or present in the gene_list
gene_list_set = set(gene_list)
genes_in_list = gene_names.isin(gene_list_set)

# Combine masks
genes_to_keep = genes_expressed_enough | genes_in_list

# Filter genes
cell_group_adata = cell_group_adata[:, genes_to_keep].copy()

print("\nshape of cell type data after filtering: ", cell_group_adata.shape)

# ----------- End of custom filtering ------------ #

# Save the new adata object
cell_group_adata.write_h5ad(f"{save_path}entero_expr_set.h5ad")
print("\n\tSaved the new adata object to: ", f"{save_path}entero_expr_set.h5ad\n
```

```
shape of cell type data:  (1361, 15992)

shape of cell type data after filtering:  (1361, 10451)

        Saved the new adata object to:  /hpc/shared/onco_janssen/dhaynessimmons/p
rojects/ageing_flies/data/entero_expr_set.h5ad
```