In [ ]:
```
# 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 matplotlib.pyplot as plt
import pandas as pd
import scanpy as sc
import anndata as ad
```

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

In [ ]:
```
# Set the functions
def split_by_batch_prefix(name):
    """
    Splits the string 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 [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 [ ]:
```
# Get the unique afca cell types
for cell_type in sorted(adata.obs['afca_annotation'].unique().tolist()):
    print(cell_type)
```

```
16-cell germline cyst in germarium region 2a and 2b
CNS surface associated glial cell
adult Malpighian tubule principal cell
adult Malpighian tubule principal cell of initial segment
adult Malpighian tubule principal cell of lower segment
adult Malpighian tubule principal cell of lower ureter
adult Malpighian tubule stellate cell of main segment
adult alary muscle
adult differentiating enterocyte
adult fat body_body
adult glial cell
adult heart ventral longitudinal muscle
adult hindgut
adult midgut enterocyte
adult midgut-hindgut hybrid zone
adult oenocyte
adult peripheral nervous system
adult reticular neuropil associated glial cell_body
adult salivary gland
adult tracheal cell
adult ventral nervous system
anterior ejaculatory duct
antimicrobial peptide-producing cell
cardia (1)
cardia (2)
cardiomyocyte, working adult heart muscle (non-ostia)
cell body glial cell
copper cell
crop
cyst cell
ejaculatory bulb
ejaculatory bulb epithelium
enteroblast
enterocyte of anterior adult midgut epithelium
enterocyte of posterior adult midgut epithelium
enterocyte-like
enteroendocrine cell
eo support cell
epidermal cell that specialized in antimicrobial response
epithelial cell_body
escort cell
female reproductive system
follicle cell
follicle stem cell and prefollicle cell
germline cell
gustatory receptor neuron
hemocyte_body
indirect flight muscle
intestinal stem cell
leg muscle motor neuron
male accessory gland main cell
male accessory gland secondary cell
mechanosensory neuron of haltere
midgut large flat cell
multidendritic neuron
muscle cell
oviduct
perineurial glial sheath
pheromone-sensing neuron
polar follicle cell
```

```
posterior midgut*
prefollicle cell/stalk follicle cell
principal cell*
scolopidial neuron
seminal vesicle & testis epithelia
spermatid
spermatocyte
stalk follicle cell
subperineurial glial cell_body
unannotated
visceral muscle of the crop
visceral muscle of the midgut
young germ cell
```

The Kernel crashed while executing code in the current cell or a previous cell.

Please review the code in the cell(s) to identify a possible cause of the failur e.

Click here for more info.

View Jupyter log for further details.

In [ ]:
```
# get cells where sex is neither F nor M
mix_adata = adata[(adata.obs.sex != "female")&(adata.obs.sex != "male")].copy()
print(mix_adata.obs.shape)
# Remove them from the dataset
mf_adata = adata[~(adata.obs.index.isin(mix_adata.obs.index))&(adata.obs['afca_a
# get the batch from the row name
mf_adata.obs['batch'] = mf_adata.obs.index.map(lambda x: split_by_batch_prefix(x
print(mf_adata.obs.shape)
```

```
(4345, 15)
(265979, 15)
```

In [ ]:
```
# 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/data/ageing_flies/afca_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"
]

# Loop through the afca cell types and create a new adata object for each cell t
for cell_type in mf_adata_obs['afca_annotation'].unique():
    cell_type = cell_type.replace("/", "-")
    print("\n\nWorking on cell type: ", cell_type, "\n")

    # Get the indices of the cells that match the current cell type
    indices = mf_adata_obs[mf_adata_obs['afca_annotation'] == cell_type].index
    # Create new pandas dataframe with the indices
    cell_type_df = mf_adata_obs.loc[indices]
```

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

    if cell_cnt < 200:
        print("Not enough cells to proceed with analysis")
        continue
    else:
        print("Sufficient cells to proceed with analysis")

        # Check that each age group has at least 100 cells
        age_grouped = cell_type_df.groupby('age', observed=False).size()
        del cell_type_df
        min_value = age_grouped.min()
        del age_grouped
        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")
            continue
        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)

    # Create a new adata object for this cell type
    cell_type_adata = mf_adata[indices].copy()
    print("\nshape of cell type data: ", cell_type_adata.shape)

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

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

    # Get gene names
    gene_names = pd.Index(cell_type_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_type_adata = cell_type_adata[:, genes_to_keep].copy()

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

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

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

```
# Print list of successfully processed cell types
print(cell_list)
del mf_adata_obs
```

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


Working on cell type:  follicle cell

number of cells:  25251
Min number of genes expressed :  226
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  5826
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult fat body_body

number of cells:  31311
Min number of genes expressed :  224
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  2703
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult hindgut

number of cells:  683
Min number of genes expressed :  286
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  146
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult oenocyte

number of cells:  5925
Min number of genes expressed :  233
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  1081
Sufficient cells in each age group to proceed with analysis


Working on cell type:  hemocyte_body

number of cells:  2962
Min number of genes expressed :  213
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  560
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult tracheal cell

number of cells:  6939
Min number of genes expressed :  207
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  1154
Sufficient cells in each age group to proceed with analysis


Working on cell type:  crop

number of cells:  6226
```

```
Min number of genes expressed :  261
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  1051
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult ventral nervous system

number of cells:  22172
Min number of genes expressed :  271
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  3454
Sufficient cells in each age group to proceed with analysis


Working on cell type:  enteroendocrine cell

number of cells:  529
Min number of genes expressed :  268
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  75
Not enough cells in an age group to proceed with analysis


Working on cell type:  germline cell

number of cells:  3824
Min number of genes expressed :  208
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  435
Sufficient cells in each age group to proceed with analysis


Working on cell type:  enterocyte of posterior adult midgut epithelium

number of cells:  751
Min number of genes expressed :  263
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  40
Not enough cells in an age group to proceed with analysis


Working on cell type:  cardia (1)

number of cells:  1062
Min number of genes expressed :  298
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  42
Not enough cells in an age group to proceed with analysis


Working on cell type:  antimicrobial peptide-producing cell

number of cells:  104
Min number of genes expressed :  390
Not enough cells to proceed with analysis


Working on cell type:  female reproductive system
```

```
number of cells:  1219
Min number of genes expressed :  287
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  256
Sufficient cells in each age group to proceed with analysis


Working on cell type:  scolopidial neuron

number of cells:  1295
Min number of genes expressed :  278
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  211
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult salivary gland

number of cells:  583
Min number of genes expressed :  280
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  52
Not enough cells in an age group to proceed with analysis


Working on cell type:  intestinal stem cell

number of cells:  2132
Min number of genes expressed :  282
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  254
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult peripheral nervous system

number of cells:  5889
Min number of genes expressed :  255
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  737
Sufficient cells in each age group to proceed with analysis


Working on cell type:  mechanosensory neuron of haltere

number of cells:  635
Min number of genes expressed :  259
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  136
Sufficient cells in each age group to proceed with analysis


Working on cell type:  epithelial cell_body

number of cells:  42978
Min number of genes expressed :  211
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  5034
Sufficient cells in each age group to proceed with analysis
```

```
Working on cell type:  escort cell

number of cells:  756
Min number of genes expressed :  304
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  99
Not enough cells in an age group to proceed with analysis


Working on cell type:  enterocyte of anterior adult midgut epithelium

number of cells:  182
Min number of genes expressed :  281
Not enough cells to proceed with analysis


Working on cell type:  midgut large flat cell

number of cells:  73
Min number of genes expressed :  525
Not enough cells to proceed with analysis


Working on cell type:  eo support cell

number of cells:  2449
Min number of genes expressed :  268
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  407
Sufficient cells in each age group to proceed with analysis


Working on cell type:  pheromone-sensing neuron

number of cells:  672
Min number of genes expressed :  303
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  101
Sufficient cells in each age group to proceed with analysis


Working on cell type:  polar follicle cell

number of cells:  304
Min number of genes expressed :  318
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  48
Not enough cells in an age group to proceed with analysis


Working on cell type:  enteroblast

number of cells:  2404
Min number of genes expressed :  270
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  247
Sufficient cells in each age group to proceed with analysis
```

```
Working on cell type:  oviduct

number of cells:  1517
Min number of genes expressed :  256
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  257
Sufficient cells in each age group to proceed with analysis


Working on cell type:  muscle cell

number of cells:  53895
Min number of genes expressed :  201
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  5675
Sufficient cells in each age group to proceed with analysis


Working on cell type:  copper cell

number of cells:  149
Min number of genes expressed :  265
Not enough cells to proceed with analysis


Working on cell type:  enterocyte-like

number of cells:  259
Min number of genes expressed :  301
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  38
Not enough cells in an age group to proceed with analysis


Working on cell type:  follicle stem cell and prefollicle cell

number of cells:  430
Min number of genes expressed :  278
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  5
Not enough cells in an age group to proceed with analysis


Working on cell type:  leg muscle motor neuron

number of cells:  355
Min number of genes expressed :  384
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  64
Not enough cells in an age group to proceed with analysis


Working on cell type:  adult midgut-hindgut hybrid zone

number of cells:  576
Min number of genes expressed :  279
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  51
Not enough cells in an age group to proceed with analysis
```

```
Working on cell type:  adult midgut enterocyte

number of cells:  97
Min number of genes expressed :  367
Not enough cells to proceed with analysis


Working on cell type:  gustatory receptor neuron

number of cells:  695
Min number of genes expressed :  298
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  100
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult Malpighian tubule principal cell of lower segment

number of cells:  50
Min number of genes expressed :  317
Not enough cells to proceed with analysis


Working on cell type:  cardia (2)

number of cells:  60
Min number of genes expressed :  332
Not enough cells to proceed with analysis


Working on cell type:  adult Malpighian tubule principal cell

number of cells:  47
Min number of genes expressed :  238
Not enough cells to proceed with analysis


Working on cell type:  young germ cell

number of cells:  224
Min number of genes expressed :  216
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  24
Not enough cells in an age group to proceed with analysis


Working on cell type:  16-cell germline cyst in germarium region 2a and 2b

number of cells:  532
Min number of genes expressed :  366
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  82
Not enough cells in an age group to proceed with analysis


Working on cell type:  adult Malpighian tubule principal cell of lower ureter

number of cells:  35
Min number of genes expressed :  582
```

```
Not enough cells to proceed with analysis


Working on cell type:  stalk follicle cell

number of cells:  293
Min number of genes expressed :  276
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  33
Not enough cells in an age group to proceed with analysis


Working on cell type:  multidendritic neuron

number of cells:  210
Min number of genes expressed :  263
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  27
Not enough cells in an age group to proceed with analysis


Working on cell type:  posterior midgut*

number of cells:  114
Min number of genes expressed :  343
Not enough cells to proceed with analysis


Working on cell type:  prefollicle cell-stalk follicle cell

number of cells:  0
Min number of genes expressed :  nan
Not enough cells to proceed with analysis


Working on cell type:  epidermal cell that specialized in antimicrobial response

number of cells:  91
Min number of genes expressed :  268
Not enough cells to proceed with analysis


Working on cell type:  indirect flight muscle

number of cells:  17625
Min number of genes expressed :  211
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  773
Sufficient cells in each age group to proceed with analysis


Working on cell type:  cell body glial cell

number of cells:  2306
Min number of genes expressed :  223
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  357
Sufficient cells in each age group to proceed with analysis
```

```
Working on cell type:  visceral muscle of the midgut

number of cells:  1253
Min number of genes expressed :  275
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  227
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult alary muscle

number of cells:  2706
Min number of genes expressed :  226
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  362
Sufficient cells in each age group to proceed with analysis


Working on cell type:  visceral muscle of the crop

number of cells:  760
Min number of genes expressed :  274
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  67
Not enough cells in an age group to proceed with analysis


Working on cell type:  adult glial cell

number of cells:  2153
Min number of genes expressed :  252
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  282
Sufficient cells in each age group to proceed with analysis


Working on cell type:  cardiomyocyte, working adult heart muscle (non-ostia)

number of cells:  34
Min number of genes expressed :  465
Not enough cells to proceed with analysis


Working on cell type:  principal cell*

number of cells:  121
Min number of genes expressed :  223
Not enough cells to proceed with analysis


Working on cell type:  adult heart ventral longitudinal muscle

number of cells:  112
Min number of genes expressed :  280
Not enough cells to proceed with analysis


Working on cell type:  adult differentiating enterocyte

number of cells:  72
```

```
Min number of genes expressed :  460
Not enough cells to proceed with analysis


Working on cell type:  perineurial glial sheath

number of cells:  2181
Min number of genes expressed :  260
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  334
Sufficient cells in each age group to proceed with analysis


Working on cell type:  subperineurial glial cell_body

number of cells:  1431
Min number of genes expressed :  280
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  279
Sufficient cells in each age group to proceed with analysis


Working on cell type:  adult Malpighian tubule principal cell of initial segment

number of cells:  70
Min number of genes expressed :  295
Not enough cells to proceed with analysis


Working on cell type:  CNS surface associated glial cell

number of cells:  208
Min number of genes expressed :  271
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  31
Not enough cells in an age group to proceed with analysis


Working on cell type:  adult reticular neuropil associated glial cell_body

number of cells:  521
Min number of genes expressed :  255
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  29
Not enough cells in an age group to proceed with analysis


Working on cell type:  adult Malpighian tubule stellate cell of main segment

number of cells:  10
Min number of genes expressed :  675
Not enough cells to proceed with analysis


Working on cell type:  seminal vesicle & testis epithelia

number of cells:  635
Min number of genes expressed :  253
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  121
```

```
Sufficient cells in each age group to proceed with analysis


Working on cell type:  ejaculatory bulb

number of cells:  1401
Min number of genes expressed :  285
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  125
Sufficient cells in each age group to proceed with analysis


Working on cell type:  ejaculatory bulb epithelium

number of cells:  549
Min number of genes expressed :  334
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  56
Not enough cells in an age group to proceed with analysis


Working on cell type:  anterior ejaculatory duct

number of cells:  189
Min number of genes expressed :  280
Not enough cells to proceed with analysis


Working on cell type:  male accessory gland main cell

number of cells:  2015
Min number of genes expressed :  265
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  225
Sufficient cells in each age group to proceed with analysis


Working on cell type:  spermatid

number of cells:  131
Min number of genes expressed :  203
Not enough cells to proceed with analysis


Working on cell type:  cyst cell

number of cells:  201
Min number of genes expressed :  207
Sufficient cells to proceed with analysis
Minimum number of cells in an age group:  26
Not enough cells in an age group to proceed with analysis


Working on cell type:  male accessory gland secondary cell

number of cells:  20
Min number of genes expressed :  336
Not enough cells to proceed with analysis
```

```
Working on cell type:  spermatocyte

number of cells:  90
Min number of genes expressed :  224
Not enough cells to proceed with analysis
['follicle cell', 'adult fat body_body', 'adult hindgut', 'adult oenocyte', 'hemo
cyte_body', 'adult tracheal cell', 'crop', 'adult ventral nervous system', 'germl
ine cell', 'female reproductive system', 'scolopidial neuron', 'intestinal stem c
ell', 'adult peripheral nervous system', 'mechanosensory neuron of haltere', 'epi
thelial cell_body', 'eo support cell', 'pheromone-sensing neuron', 'enteroblast',
'oviduct', 'muscle cell', 'gustatory receptor neuron', 'indirect flight muscle',
'cell body glial cell', 'visceral muscle of the midgut', 'adult alary muscle', 'a
dult glial cell', 'perineurial glial sheath', 'subperineurial glial cell_body',
'seminal vesicle & testis epithelia', 'ejaculatory bulb', 'male accessory gland m
ain cell']
```