AFCA analysis

1. Overview of the Analysis Pipeline

a) Data Source

The primary dataset used for this analysis was the AnnData object adata_body_S_v1.0.h5ad, which is the publicly released single-nucleus RNA-seq dataset from the AFCA study (Lu et al., 2023). This dataset contains transcriptomic profiles from nuclei isolated from adult *Drosophila melanogaster* body tissues across multiple ages. The data was read into R using the zellkonverter (readH5AD) and Seurat (as.Seurat) packages.

Each cell in this dataset is annotated with metadata, including:

- age (5, 30, or 70 days)
- sex
- afca_annotation (cell type labels, e.g. “enterocyte of posterior adult midgut epithelium,” “intestinal stem cell,” etc.)

b) Genes of interest

The genes that are of particular interest to the analysis are the following: "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"

c) Pre-processing data

Initial preprocessing followed standard steps from the AFCA paper (lu et al., 2023), including quality control measures to filter out doublets, remove ambient RNA contamination, exclude cells with high mitochondrial content, and retain genes expressed in a minimum number of cells. Subsequently, additional preprocessing steps were introduced (afca_analysis_280425.ipynb): a unique sample identifier was generated for each individual fly to serve as a latent variable in the downstream MAST differential gene expression analysis; cell types with fewer than 100 cells per unique age group were excluded; and genes were filtered to include only those expressed in at least three cells per cell type. Genes of interest were retained irrespective of expression thresholds. The filtered, pre-processed datasets were then saved as cell-type-specific .h5ad files for efficient downstream analysis. This resulted in 31 cell-type specific datasets for downstream analysis, only 2 of which were pre-selected as potential cell types of interest.

For the Enterocyte-specific analysis, an additional preprocessing step was implemented. All cells annotated with “enterocyte” in the name were grouped into a single category, referred to as "Enterocyte". This grouping was performed to increase cell counts for this cell type, which had low cell numbers in the original dataset. The filtered, pre-processed dataset, was then saved as a .h5ad file for downstream analysis.

Adult alary muscle	**Adult fat body (body)**	Adult glial cell	Adult hindgut
Adult oenocyte	Adult peripheral nervous system	Adult tracheal cell	Adult ventral nervous system
Cell body glial cell	Crop	Ejaculatory bulb	Enteroblast
EO support cell	Epithelial cell (body)	Female reproductive system	Follicle cell
Germline cell	Gustatory receptor neuron	Hemocyte (body)	Indirect flight muscle
Intestinal stem cell	Male accessory gland main cell	Mechanosensory neuron of haltere	Muscle cell
Oviduct	Perineurial glial sheath	Pheromone-sensing neuron	Scolopidial neuron
Seminal vesicle & testis epithelia	Subperineurial glial cell (body)	Visceral muscle of the midgut	

d) Differential Gene Expression Analysis (DGEA)
Differential expression analysis and the following steps were executed using the R script (run_dgea_single_hpc.r), where each cell type was processed independently. First, metadata fields including age, indiv, sex_age, and afca_annotation are converted to factors for appropriate handling. DGEA is performed using Seurat's FindMarkers() function with the MAST test. The analysis compares 5-day-old samples (the "young group") vs. each of the other age groups found in that cell type subset. When running the MAST-based DGE analysis, the unique sample identifier "indiv" is included as a latent variable (covariate) to account for potential confounding effects.

e) Combining and Correcting p-values
After performing differential gene expression analyses for each cell type and age comparison (e.g., 5 vs. 30 days, 5 vs. 70 days, etc.), the resulting marker tables were appended into a single cumulative results file. Seurat's FindMarkers function was used with the MAST test, including individual sample ID (indiv) as a latent variable to account for biological variation between individual flies. Following the analysis, a Benjamini-Hochberg (BH) correction was manually applied to the raw p-values (p_val) to generate a new adjusted p-value column (p_val_adj) to control for the false discovery rate (FDR) rather than the family-wise error rate, which offers a better balance between identifying true positive results and limiting false positives. The differential expression results from all age comparisons for a given cell type were then combined and saved into a single output file, named markers_<cell_type>.csv, containing information on the gene tested, the associated p-values, BH-adjusted p-values, log fold changes, and the relevant cell type and comparison labels.

f) Generating the plots
Differential gene expression results were visualized using volcano plots generated in R (afca_analysis_viz_310325.Rmd & afca_analysis_entero_vis_170525.Rmd for the enterocyte category). For each plot, the log2 fold change (log2FC) relative to 5-day-old flies was plotted against the negative log10 of the adjusted p-value. Positive log2FC values indicate higher expression in 5-day-old flies compared to older groups, while negative values indicate lower expression in 5-day-old flies.
The plots were partitioned based on statistical significance (adjusted p-value ≤ 0.1) and effect size (|log2FC| ≥ 1). Genes of particular interest were labelled on the plots. Volcano plots were created using ggplot2, and separate plots were generated for each cell type and comparison. Final figures were saved individually in .png format for reporting.

2. Structure of the final Results

The final output csv files each include (among other columns):

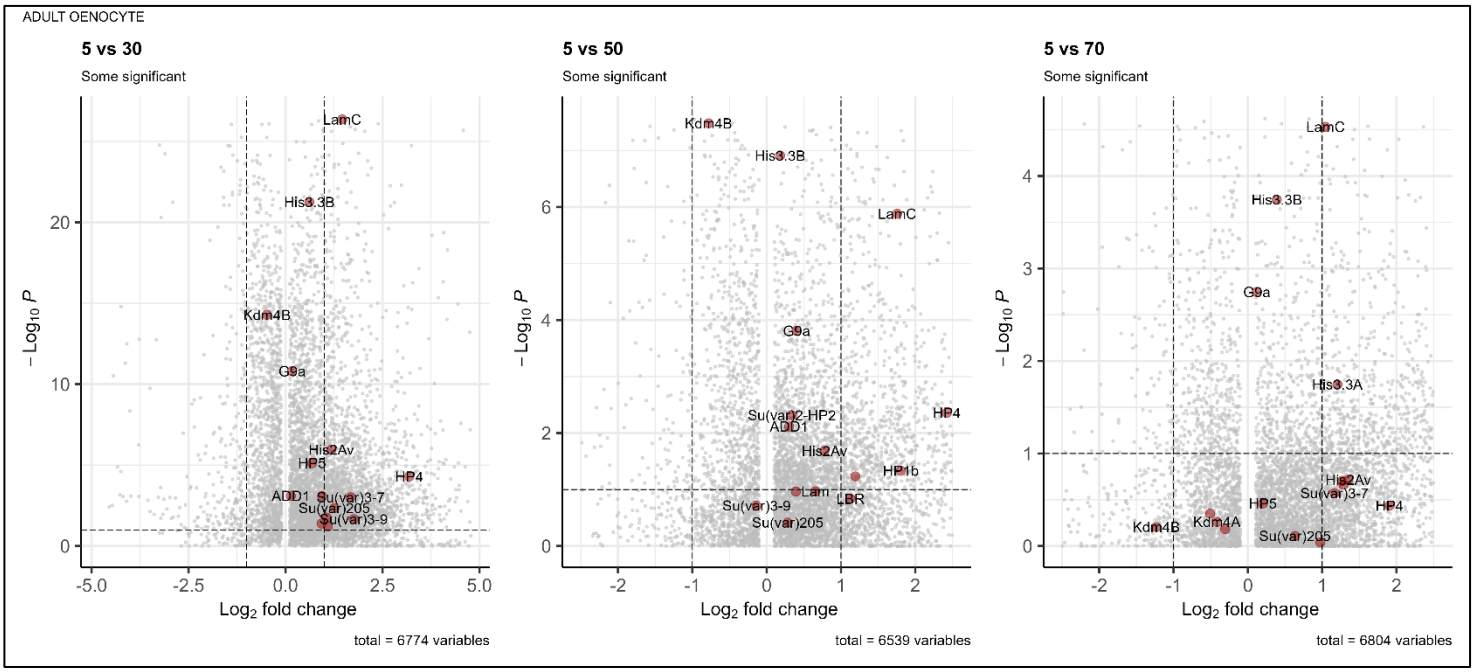
- **avg_log2FC**: The log-fold-change estimate (log2 scale) for 5-day-old vs. the other age. Positive values suggest higher expression in the 5-day-old group; negative values suggest lower expression in 5-day-old.
- ***cell_type*** (*not in Enterocyte analysis*): Which cell type the comparison is associated with
- **comparison**: The specific comparison made (e.g., 5 vs 70).
- **gene**: The gene name.
- **p_val_adj:** The Benjamini-Hochberg (BH) corrected p-value, controlling the false discovery rate (FDR) separately within each cell type.
- **Additional Columns**: Other statistical outputs from the MAST test such as the percentage of cells expressing the gene in each group, pct.1 (in 5 day) and pct.2 (in the other).

3. Key Findings

The differential gene expression analysis identified about 8 cell types out of the 30 where a more than 2 of the genes of interest showed significant (BH-adjusted p-value ≤ 0.1) and meaningful (|log2FC| ≥ 1) changes between young and old age.

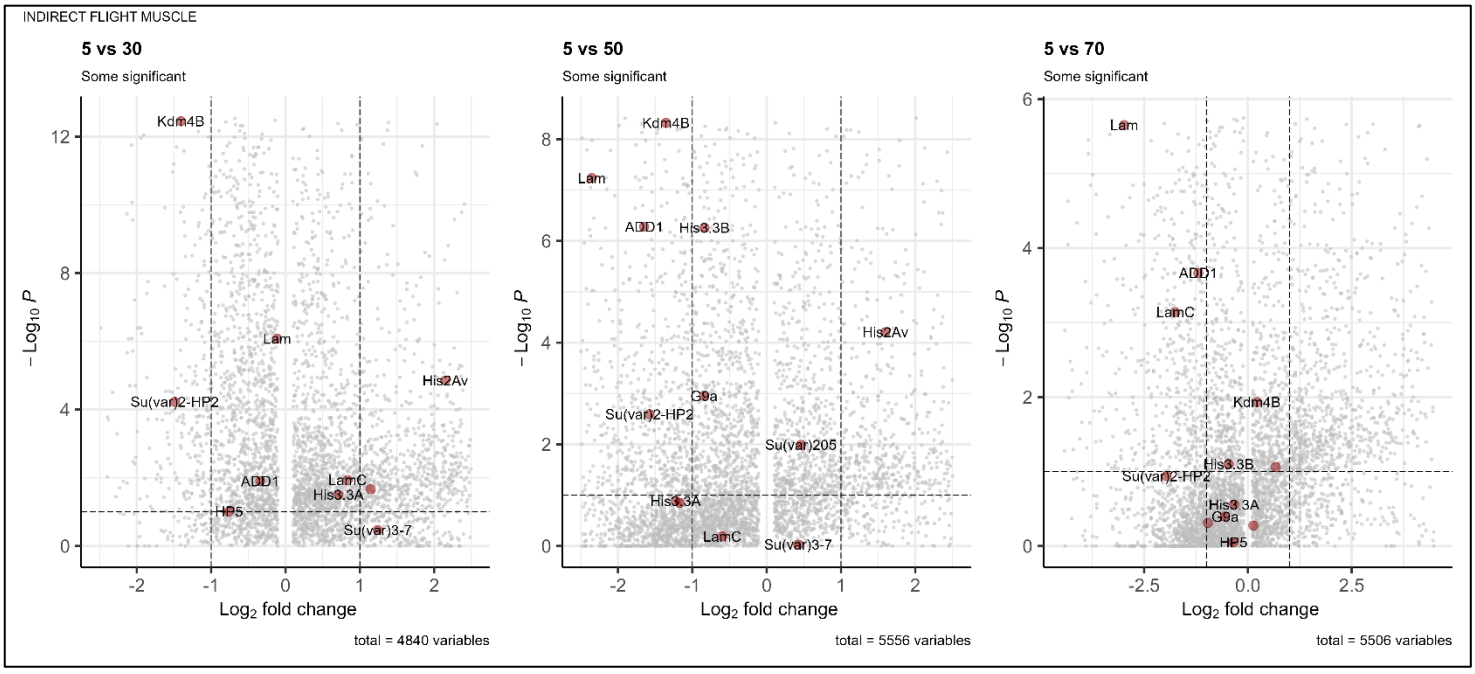
a) Adult Oenocyte:

The adult oenocyte exhibited the largest number of differentially expressed genes (9 genes). **Over-expression was primarily observed in 5-day-old flies compared 30 day-old flies** in a subset of the genes of interest (LamC, His2Av, HP4, Su(var)3-7, Su(var)205, Su(var)3-9, HP1b & LBR). This over-expression is similarly shown in 5 day-old flies compared to 50 day-old flies for subset of those genes (LamC, HP4, Su(var)3-7 & HP1b). The over-expression of the genes of interest in 5 day-old flies compared to 70 day-old flies is however only shown for LamC but also for His3.3A which was not demonstrated in the previous comparisons. This suggests that these genes of interest could be downregulated with age in adult oenocytes.

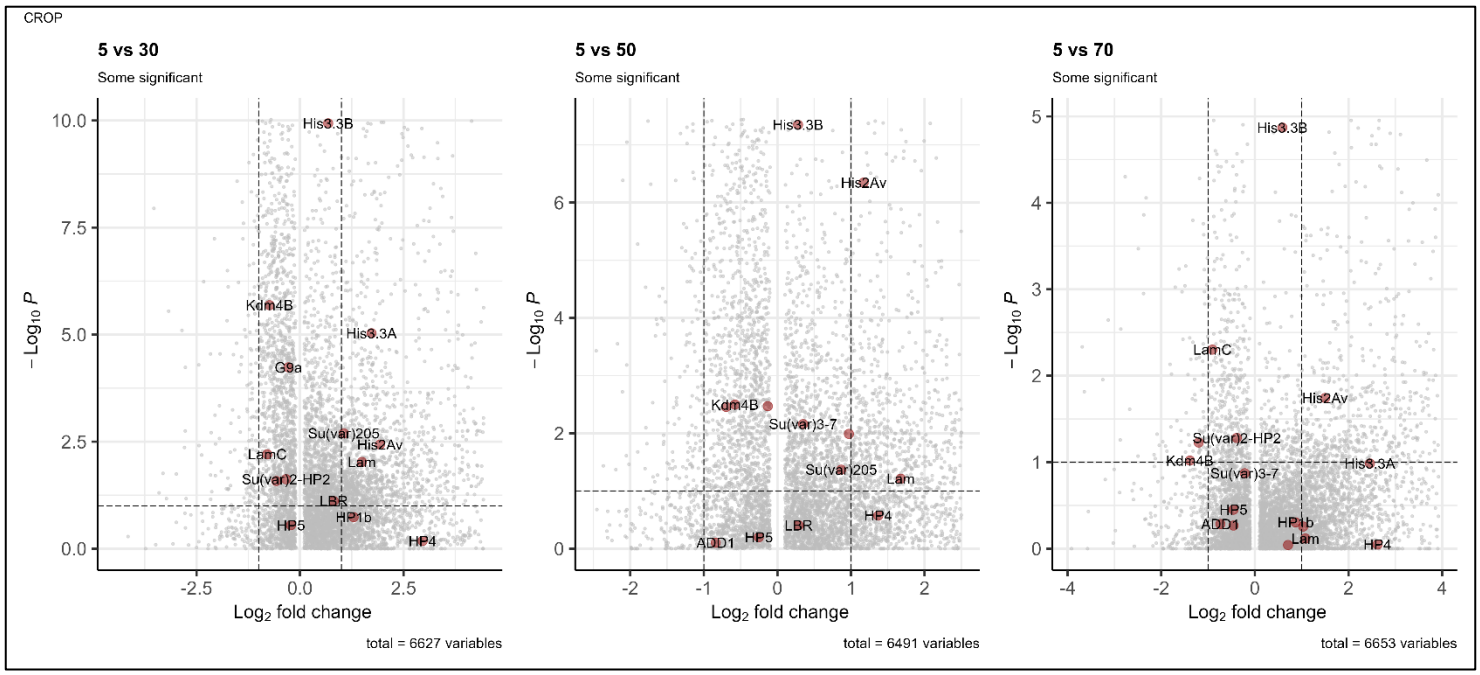


b) Indirect Flight Muscle:

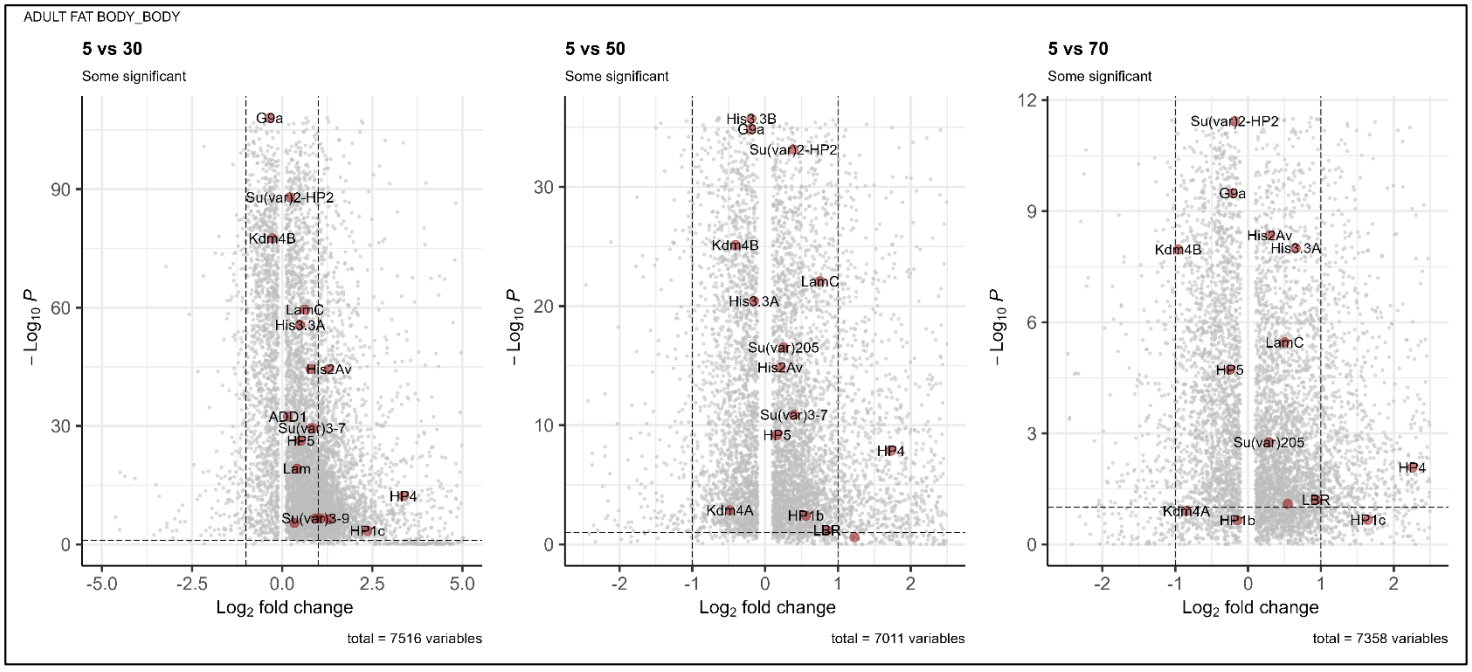
The indirect flight muscle exhibited differential expression of 7 genes of interest. **Over-expression in 5-day-old flies compared to 30-day-old flies was observed for *His2Av* and *Su(var)205*, while under-expression was observed for *Kdm4B* and *Su(var)2-HP2***. In comparisons between 5-day-old and 50-day-old flies, over-expression of *His2Av* persisted, whereas *Kdm4B*, *Lam*, *ADD1*, and *Su(var)2-HP2* were under-expressed, indicating a shift toward reduced expression of chromatin-associated genes in the 5 day old group. In the comparison between 5-day-old and 70-day-old flies, further under-expression in the younger group was observed for *Lam*, *ADD1*, and *LamC*, with no genes demonstrating over-expression. This suggesting that the selected genes in indirect flight muscles could be progressively upregulated with age.



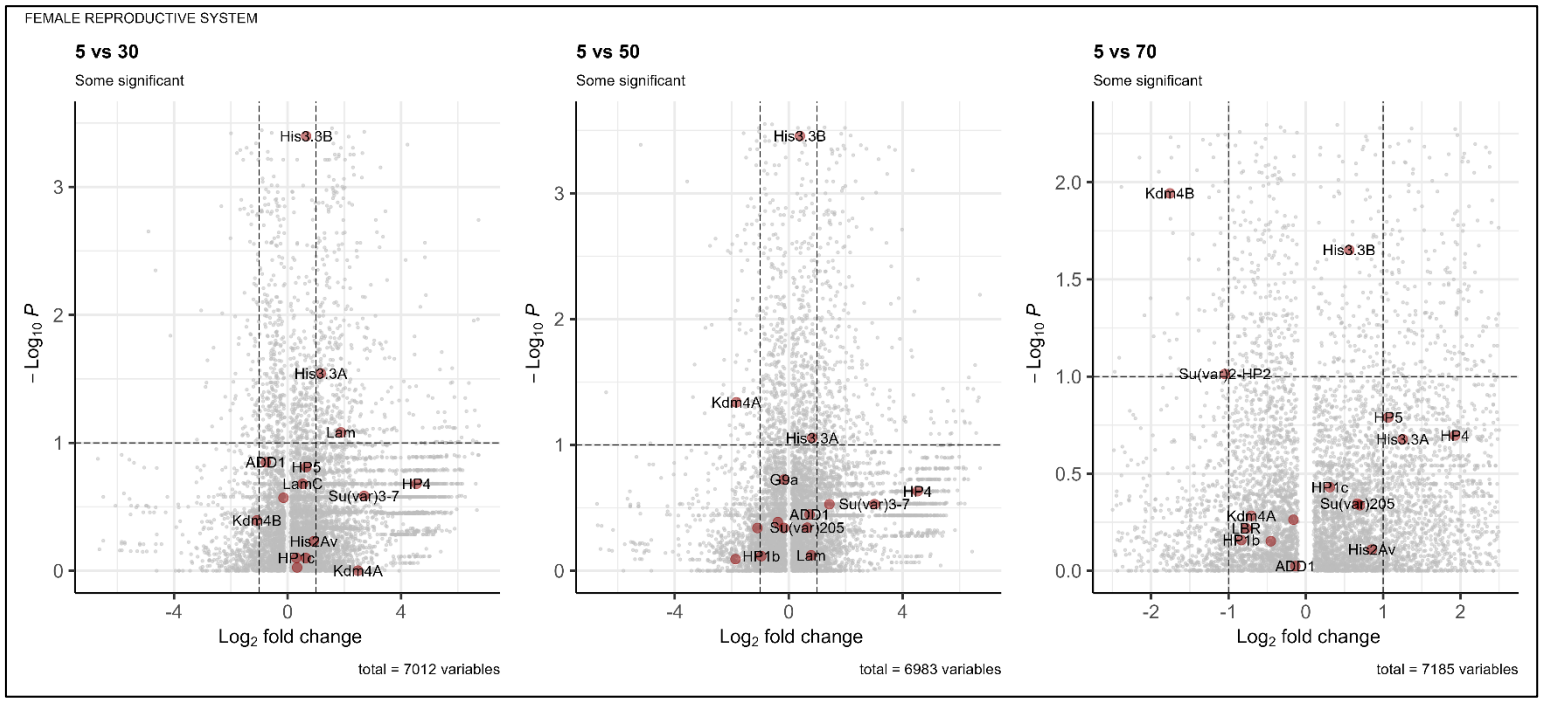
c) <u>Crop</u>:
The crop cells exhibited differential expression of 6 genes of interest. **Over-expression was primarily observed in 5-day-old flies compared to 30-day-old flies for several genes** (*His3.3A, Su(var)205, His2Av,* and *Lam*). In the comparison between 5-day-old and 50-day-old flies, over-expression of *His2Av* and *Lam* was again observed in the younger group**. Between 5-day-old and 70-day-old flies, over-expression of *His2Av* persisted, although under-expression emerged for *Kdm4A* and *Kdm4B***, which were not previously observed in earlier comparisons



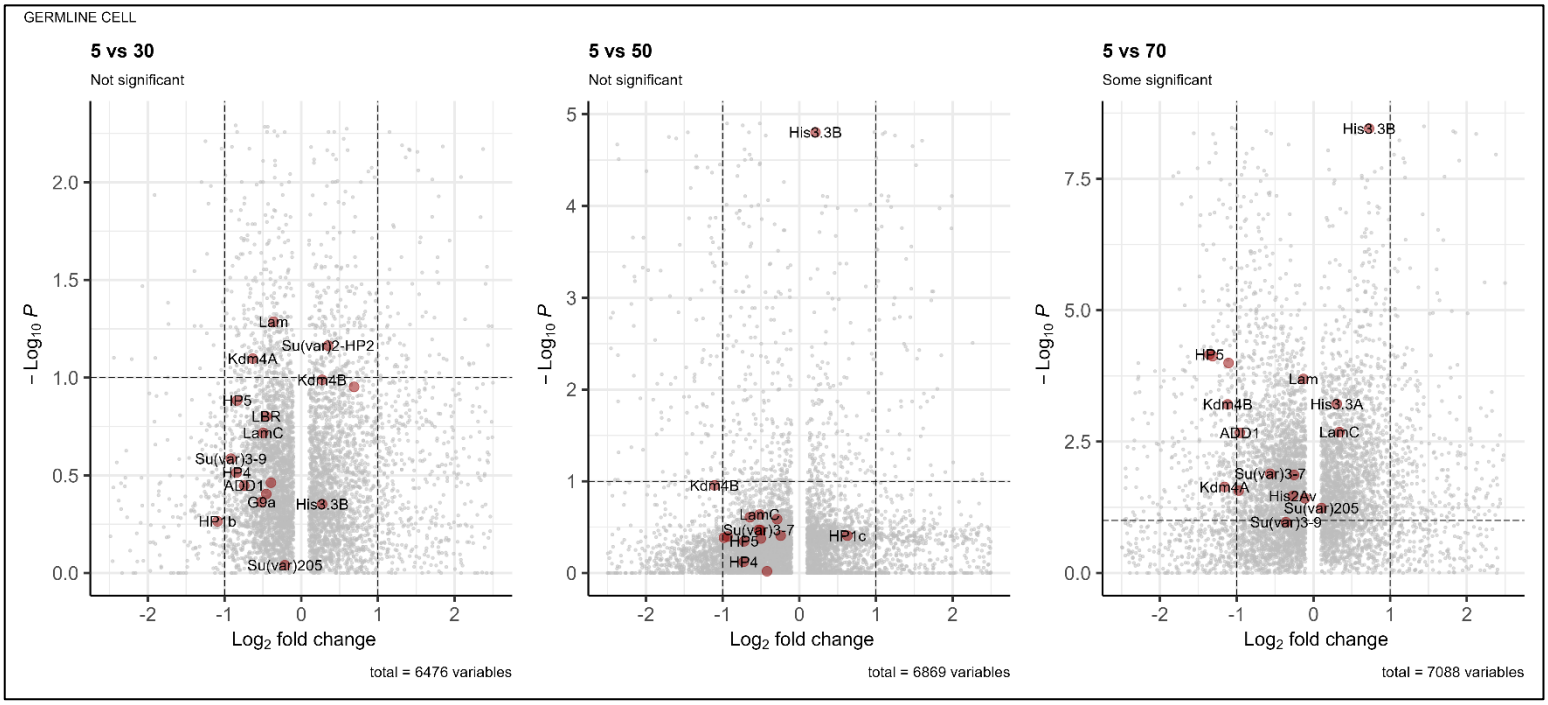
d) Adult Fat Body (body):
The adult fat body exhibited differential expression of 5 genes of interest**. Over-expression was primarily observed in 5-day-old flies compared to 30-day-old flies for *His2Av, HP4, HP1b, LBR,* and *HP1c***. In the comparison between 5-day-old and 50-day-old flies, over-expression of *HP4* was only observed. Between 5-day-old and 70-day-old flies, *HP4* remained over-expressed, suggesting that *HP4* expression is consistently reduced with age.



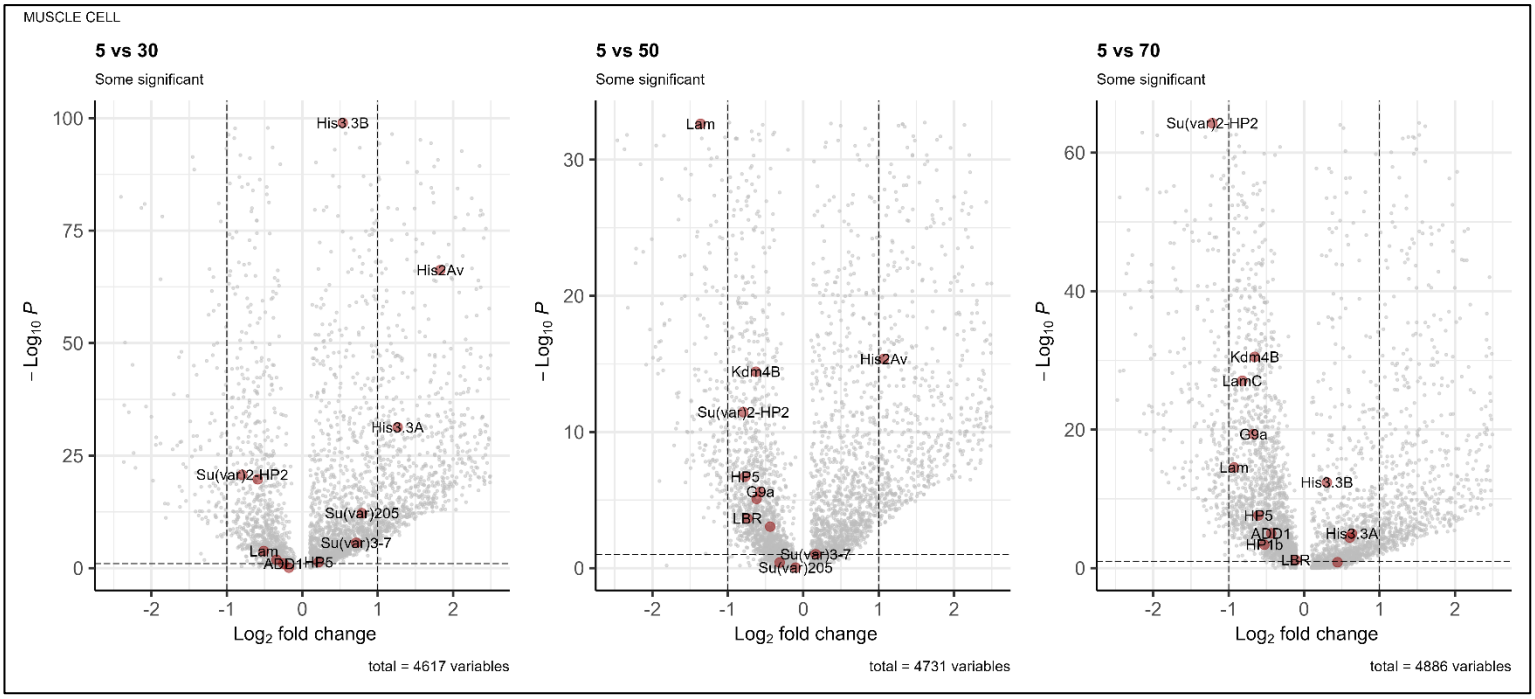
e) Female Reproductive System:
The female reproductive system exhibited differential expression of 5 genes of interest**. Over-expression was primarily observed in 5-day-old flies compared to 30-day-old flies for *His3.3A* and *Lam***. In the comparison between 5-day-old and 50-day-old flies, under-expression emerged for *Kdm4A*, indicating a shift toward increased expression of the factor at mid-life. Between 5-day-old and 70-day-old flies, under-expression was observed for *Kdm4B* and additionally *Su(var)2-HP2*.



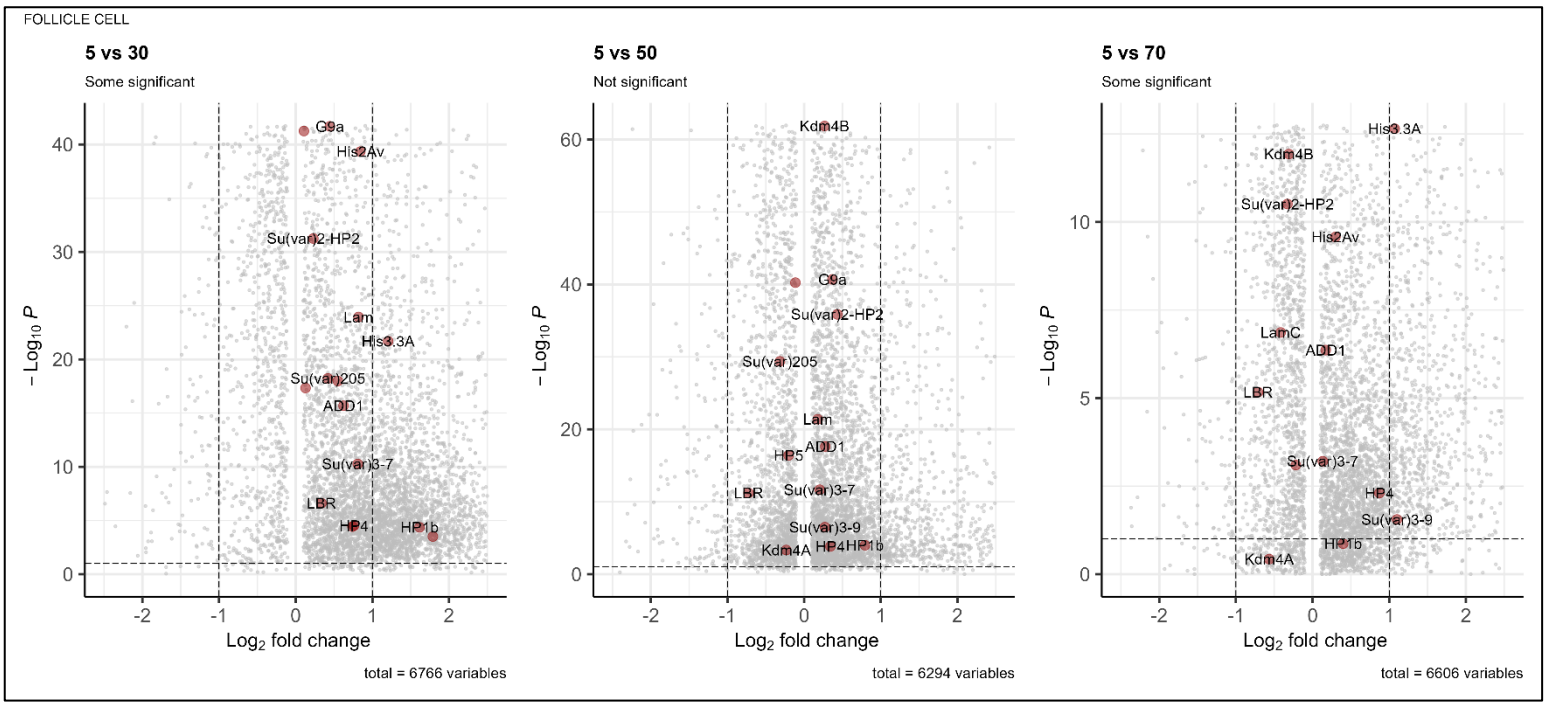
f) Germline Cell:
The germline cells exhibited differential expression of 5 genes of interest. **All observed changes were between 5-day-old and 70-day-old flies and involved under-expression in the 5-day-old group for *HP5, Su(var)2-HP2, HP4, Kdm4B,* and *Kdm4A***. This pattern indicates that these genes are expressed at higher levels in aged germline cells compared to young flies.



g) Muscle Cell:
The muscle cells exhibited differential expression of 4 genes of interest. **Over-expression in 5-day-old flies compared to 30-day-old flies was observed for *His2Av* and *His3.3A***. Between 5-day-old and 50-day-old flies, *His2Av* continued to be over-expressed, whereas *Lam* showed under-expression in the younger group. In the comparison between 5-day-old and 70-day-old flies, only under-expression of *Su(var)2-HP2* was observed in the young age group.



h) Follicle Cell:
The follicle cells exhibited differential expression of 4 genes of interest. **Over-expression in 5-day-old flies compared to 30-day-old flies was observed for *His3.3A, HP1b,* and *Kdm4A***. No differences in the gene set were observed between the 5 day-old and 50 day-old flies. Between 5-day-old and 70-day-old flies, over-expression of *His3.3A* was yet again observed along with that of *Su(var)3-9* in young flies even at advanced comparisons.



i) <u>Enterocyte Group</u>:
The Enterocyte cell group exhibited differential expression of several genes of interest. Specifically, Lam, HP4, Kdm4A and HP1c were over-expressed in 5-day-old flies compared to 30-day-old flies. However, no genes were significantly and meaningfully differentially expressed when comparing the 5 day-old age group to the 50 day-old age category and only His3.3A showed overexpression when compared to 70 day-old group.