**Submitted by: Aleena Khalid**

**Report: Analysis of SNP Density and Gene Expression in *Drosophila melanogaster***

**Overview**

This project focused on exploring the relationship between SNP density within exonic regions and gene expression levels in *Drosophila melanogaster*. For simplicity and efficiency, the analysis was limited to chromosome 2L. The ultimate objective was to develop a Python-based pipeline that could process genomic data (GFF, VCF, and expression files), calculate SNP density for each gene, and correlate it with average gene expression levels.

**Initial Plan and Challenges**

Initially, I aimed to perform a genome-wide analysis covering all chromosomes. However, this approach quickly became impractical. The GFF file was exceptionally large and couldn't be processed with standard tools or editors. Attempting to parse the entire file led to crashes or unresponsive scripts, primarily due to hardware limitations.

To address this, I narrowed the scope to chromosome 2L, which allowed for faster and more reliable processing. Chromosome 2L was selected because it contains a substantial number of genes and SNPs—making it both representative and manageable.

**Why Chromosome 2L?**

The decision to focus solely on 2L was driven by performance constraints. Working with the entire genome would have required more computational power than was available. Fortunately, 2L alone includes over 2,300 genes with complete information, providing a sufficient dataset for meaningful statistical analysis.

**File Differences and Gene Mapping Issues**

One major issue encountered was the mismatch between gene identifiers across the input files. The GFF file listed transcript IDs (FBtr) in the Parent= attribute, while the expression data used gene IDs (FBgn). Because of this inconsistency, early attempts to connect SNP data with expression values failed, producing empty results.

To resolve this, I implemented the following:

* Parsed both mRNA and exon features from the GFF file.
* Created a mapping between FBtr (transcript) IDs and FBgn (gene) IDs using parent-child relationships.
* Translated exon-level FBtr identifiers to gene-level FBgn identifiers using this mapping.

After these changes, the script successfully linked SNPs, exons, and expression data.

**Expression File Challenges**

The gene expression file (expression.tsv) posed additional problems. It included several time-point columns, but the script initially expected a different structure and failed to detect valid expression data, leading to errors like "No expression columns found!"

To fix this:

* I modified the script to automatically detect and select all numeric columns (excluding the first column, which contained gene IDs).
* Calculated the average expression per gene across all time points.

This update made the script more flexible and compatible with the dataset.

**Script Workflow**

The Python script executes the following steps:

1. Parses exon coordinates from the GFF file (limited to chromosome 2L).
2. Maps transcript IDs to gene IDs using extracted mRNA features.
3. Reads SNP positions from the chromosome 2L section of the VCF file.
4. Maps SNPs to exons and counts how many fall within each gene.
5. Computes exon lengths and calculates SNP density per gene.
6. Reads expression values and computes the average expression for each gene.
7. Combines SNP density and expression data into a single DataFrame.
8. Exports the result as a CSV file.
9. Performs Pearson correlation analysis and generates a scatter plot.

**Results Summary**

* **Total genes analyzed**: 2,340
* **Correlation coefficient (Pearson r)**: 0.0279
* **P-value**: ~0.178
* **Conclusion**: The correlation between SNP density and gene expression was weak and statistically insignificant.

**Scatter Plot Interpretation**

The scatter plot revealed that most genes clustered in the lower-left region—showing low SNP density and low expression. A few genes exhibited high expression with very low SNP density, which might indicate purifying selection. Overall, the plot lacked a clear trend, reflecting the low correlation value observed in the analysis.