**Comparative Analysis of SNP Distribution and Gene Expression in**

***Drosophila melanogaster***

**1. Introduction**

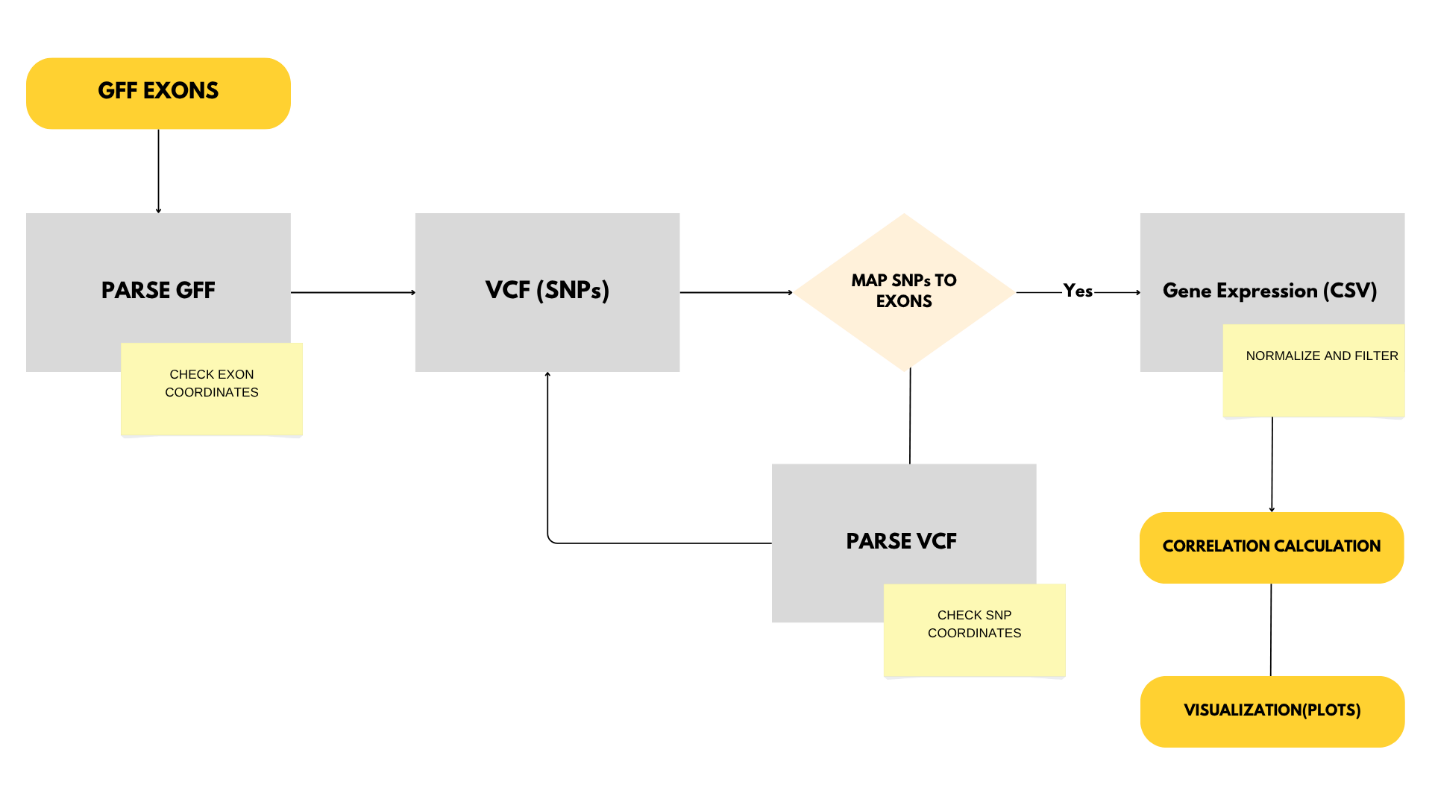
Single Nucleotide Polymorphisms (SNPs) are the most abundant genetic variations across individuals. They can be found in both coding and non-coding regions of the genome. Investigating the relationship between SNP density within exons and gene expression levels provides insights into molecular evolution and gene regulation in eukaryotes. In this study, we analyzed SNP density and gene expression in *Drosophila melanogaster*.

**2. Objectives**

* Parse genome annotation and extract exon features.
* Map SNPs to gene exonic regions.
* Calculate SNP density per gene.
* Integrate gene expression data and normalize IDs.
* Compute Pearson correlation between SNP density and expression.
* Visualize the correlation using a scatter plot.

**3. Methodology**

**Step-wise Workflow:**

****

1. **GFF Parsing:**  
   Extracted exon coordinates and gene IDs from the *D. melanogaster* GFF annotation file.
2. **VCF Parsing:**  
   Parsed SNP positions from the genome-wide VCF file.
3. **SNP-Exon Mapping:**  
   Identified SNPs falling within exonic regions and counted them per gene.
4. **Gene Expression Integration:**  
   Loaded normalized pseudotime expression data and computed average expression per gene.
5. **Gene ID Normalization & Mapping:**  
   Normalized RefSeq/transcript IDs and mapped them to FlyBase gene IDs using a BioMart file.
6. **Merging and Correlation:**  
   Merged datasets on mapped gene IDs and calculated Pearson correlation (log2-transformed).
7. **Visualization:**  
   Generated a scatter plot of log2-transformed SNP density vs. expression.

**Tools Used:**  
Python (Pandas, NumPy, Matplotlib, Seaborn), Data from Ensembl Metazoa and GEO (GSE263568)

**4. Results**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Exons parsed | 190,710 |
| SNPs parsed | 4,438,427 |
| Genes with mapped SNPs | 35,677 |
| Gene expression entries | 7,367 |
| Merged gene count | 19,212 |
| Pearson correlation (log2) | **0.0216** |

**Interpretation:**

* The Pearson correlation (r = **0.0216**) between log2-transformed SNP density and gene expression indicates a **very weak positive linear relationship**.
* This suggests that SNP accumulation in exonic regions of *Drosophila melanogaster* does **not significantly impact average gene expression levels**.
* Possible explanations include:
  + **Purifying selection** might remove harmful exonic SNPs before they affect expression.
  + **Regulatory regions** (e.g., promoters, enhancers) may play a more critical role than coding regions in controlling expression.
  + The effect of SNPs might be **non-linear or context-dependent**, requiring deeper functional annotation.

**5. Visualization**

**Scatter Plot:**

**Log2 SNP Density vs. Log2 Gene Expression**

The scatter plot reveals a cloud-like distribution of data points with no strong trend, supporting the weak correlation observed.

**6. Conclusion**

This analysis presents a genome-wide approach to evaluate the relationship between sequence variation and gene expression. In *Drosophila melanogaster*, SNP density in exonic regions shows **minimal correlation with gene expression**, hinting that other genomic elements (e.g., UTRs, promoters) may have more functional relevance. The methods demonstrated here can be applied to other species or genomic regions to further explore the regulatory architecture of genes.