

# Multiple Logistic Regression Model Pipeline for MRE Prediction Task for Dosage Compensation in Drosophila

**Romer Miranda and Prottoy Roy**

PHP1560 Final Project



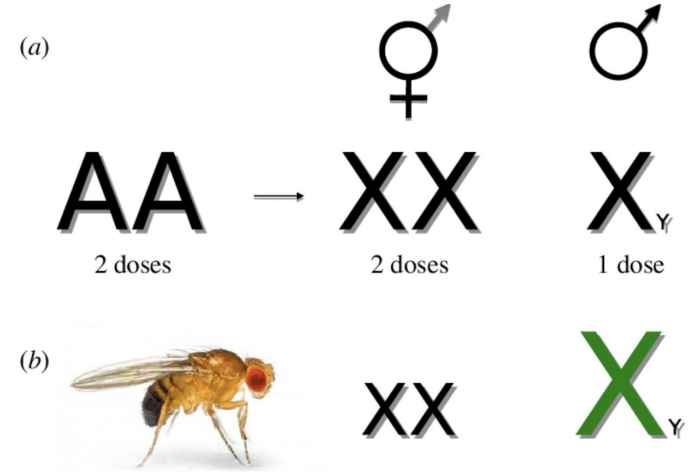
School of  
Public Health

# Outline

- **Motivation and Background**
- Data and Metrics
- Program Design
- Results and Interpretation
- Reflection and Conclusion

# Dosage Compensation in Drosophila

- **Dosage Compensation (DC)** equalizes gene expression in sex chromosomes
  - In Drosophila, DC upregulates single X chromosome in males (XY) to match females (XX)
- A transcription factor CLAMP binds to **MRE (Male-Specific Lethal) motif** to start DC machinery
  - However, MRE motif is not a strict sequence, making it difficult to locate.
- Certain **epigenetic markers** are markers of MREs
  - **Histone modifications** control chromatin environment → impacts where MRE can exist
  - **ChIP-seq** data reveals amount of histone modifications across all chromosomes



Reference: Philosophical Transactions B

### Research Question:

**Are epigenetic features, specifically histone modifications, accurate predictors of the presence of the MRE motif in *Drosophila* chromosomes?**

# Outline

- Motivation and Background
- **Data and Metrics**
- Program Design
- Results and Interpretation
- Reflection and Conclusion

# Data

- Epigenetic factors collected from ChIP-Seq data
  - 8 Histone modifications:  
  
h3k27ac, h3k27me3, h3k36me3, h3k4me1, h3k4me2, h3k4me3, h3k9me3, h4k16ac
- Data organized by **1 kilobase pair bins** each with values for epigenetic factors and labels
  - Label 0: No MRE
  - Label 1: MRE
- Labeled data is heavily **imbalanced** (80% 0s and 20% 1s)

# Model and Metrics

- Multiple Logistic Regression model used for **binary classification** of the MRE labels
- **Recall**: proportion of positive samples that were correctly classified
- **Precision**: proportion of positive-predicted samples that were actually positive samples
- Main metrics used to evaluate model performance:
  - **AUROC**: how well the model separates the two classes (Recall / False Positive Rate)
  - **AUPRC**: how well the model balances precision and recall with changing the decision threshold (Precision / Recall)

# Outline

- Motivation and Background
- Data and Metrics
- **Program Design**
- Results and Interpretation
- Reflection and Conclusion



# Workflow and Organization

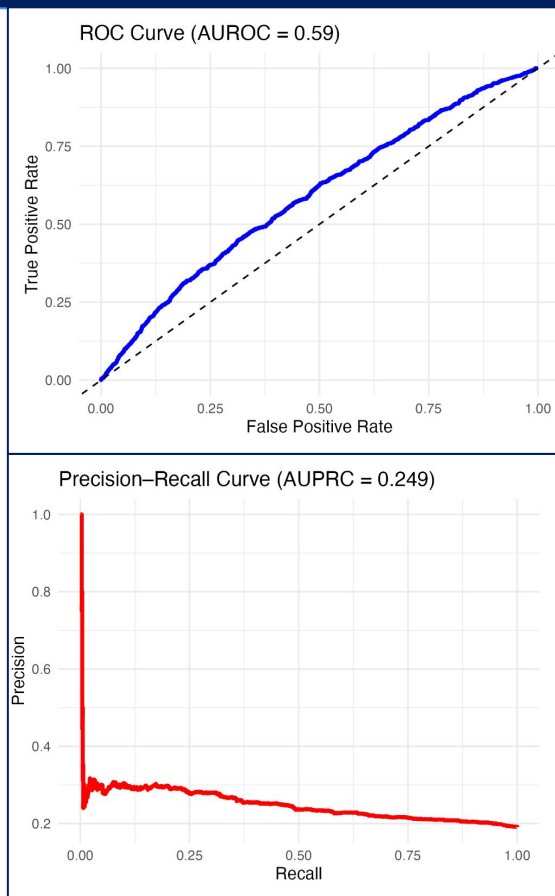
- preprocess.R
  - preprocess\_data () - loads the dataset and cleans it
  - filter\_chromosome\_data () - keeps only the rows of the selected chromosome
- split\_data.R
- train\_model.R
  - integrates downsampling of the unbalanced dataset
- evaluate\_model.R
- plotting.R
  - plot\_curves () - ROC and Precision-Recall curves
  - plot\_metric\_cross\_model () - barplots for a given metric for the cross evaluated model
- cross\_chromosome.R

# Outline

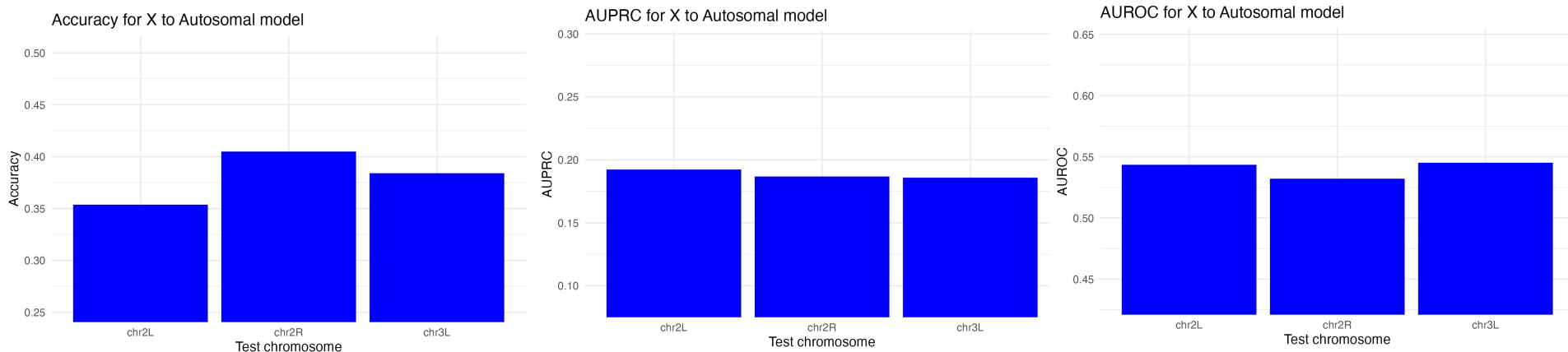
- Motivation and Background
- Data and Metrics
- Program Design
- **Results and Interpretation**
- Reflection and Conclusion

# Results

- AUROC: 0.59
  - Baseline: 0.5 - **So our model performs better than random guessing**
- AUPRC: 0.249
  - Not very good at balancing recall and precision



# Cross Chromosome Experiments



- Ran our X-trained model on 3 autosomal chromosomes to test whether X-linked MRE epigenetic patterns also appear on autosomes
- Results:
  - Accuracy is lower than 50% (baseline) for all the autosomal chromosomes
  - AUPRC is also low
  - AUROC near 0.5 means model cannot differentiate between MRE/non-MRE bins in autosomes

# Outline

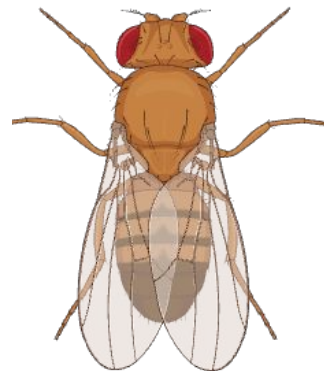
- Motivation and Background
- Data and Metrics
- Program Design
- Results and Interpretation
- **Reflection and Conclusion**

# Conclusion

- Results may indicate the histone modifications alone may not be the best predictors
  - **Future steps:** incorporate different data modalities (3D chromatin organization) or more complex models (neural networks)
  - More experiments: perturbations and examining coefficients
- Pipeline is **fully flexible** for any binary classification task on any ChIP-seq dataset!

## References:

1. Duan, J., & Larschan, E. N. (2019). Dosage compensation: How to be compensated... or not? Current Biology, 29(23), R1229–R1231.  
<https://doi.org/10.1016/j.cub.2019.09.065>
2. <https://medium.com/the-researchers-guide/modelling-binary-logistic-regression-using-tidymodels-library-in-r-part-1-c1bdce0ac055>



**Thank You!**

**Questions?**

