Summer Training in University of Toronto

Sponsored By KAUST Academy

Mohammed Khaled Shafeey



Table of Contents

Introduction 3

Program Agenda 3

Background 5

Project Aim 6

Methods 6

- Data 6
- Theoretical Framework 6
- Statistical Approaches 7
- Validation and Model Comparison 7
- Functional Analysis 8

Implementation 8

- Data Preparation 8
- Dimensionality Reduction and Modeling 8
- Functional Annotation 9
- Visualization and Communication 9

Results 9

- Sex-Specific Selection 9
- Comparison of Statistical Approaches 9
- Candidate Genes Under Selection 10
- Model Performance 10

Poster 10

Introduction

This summer I had the privilege of participating in the Summer Undergraduate Data Science (SUDS) Program, an intensive research training initiative sponsored by KAUST Academy. The program is designed to provide undergraduates with advanced training in bioinformatics and data science applications to biology, combining coursework, workshops, and hands-on research at the University of Toronto.

The selection process for the program was highly competitive: over 3,000 applicants completed a series of online and in-person courses in bioinformatics, statistical modeling, and research methods. Through this multi-stage training and evaluation, only six students were selected to continue their research placements at the University of Toronto. I was fortunate to be among these students.

The program lasted 11 weeks, during which I worked in the Sztepanacz Lab in the Department of Ecology and Evolutionary Biology. I collaborated closely with a fellow student, Mikaeel Abbas, under the supervision of Prof. Jacqueline Sztepanacz. Our project, titled Estimating Sex-Specific Selection on Gene Expression: A Comparison of Statistical Approaches, focused on applying cutting-edge computational methods to evolutionary biology, specifically understanding how natural selection acts on patterns of gene expression.

Program Agenda

June		
June 18 **3:30 - 4:30 pm ET**	**Professional Development Session**	***Workshopping your abstract*** Dr. Patricia Luedecke, Victoria College Writing Centre
June 20 **3 - 4 pm ET** **4-6 pm ET**	**Data Science @ Work** **followed by Data Science Social**	***Zia Babar*** Head of Data Science, Payfare
June 27 **3:00 - 4:00 pm ET**	**Professional Development Session**	***Your academic future*** Rory McKeown, Admissions and Recruitment Officer, Department of Computer Science, University of Toronto

July		
July 9		Deadline for SUDS Scholar to submit their SUDS project research abstract to SUDS.dsi@utoronto.ca
July 11 **3 - 4 pm ET**	**Data Science @ Work**	***Sepehr T. Sisakht,*** Chief Executive Officer and Co-Founder, Shyftbase Inc.
July 16 **3 -5 pm ET** Virtual	**Professional Development Session**	***Research Poster Design Workshop*** Dr. Sarah Treit
July 18 **2 - 4 pm ET**	**Data Science @ Work**	***In Collaboration with the School of Cities:*** **Lianne Harrison, MSc in Planning student, University of Toronto** **Tom Weatherburn, Project Manager, City of Toronto Michael Liu, Senior Advisor, Transit-Oriented Communities** **Ananmay Sharan, Web Developer and Data Analyst, Canadian Urban Institute** **Jue Wang, Assistant Professor, Geography, Geomatics and Environment, University of Toronto** **Tom Weatherburn, Project Manager, City of Toronto**
July 25 **3 - 4 pm ET**	**Professional Development Session**	***What's Next? Your Career Questions*** Saddiya Rose, MEd. Career Educator, UofT Career Centre
August		

Aug 8 **3 - 4 pm ET**	**Professional Development Sessions**	***Optimizing Resumes for Data Sciences*** Saira Banu Student Development Coordinator, Arts & Science Co-op University of Toronto Scarborough
August 15 **10 am - 2 pm ET**	**SUDS SHOWCASE DAY!**	

Background

A central goal of **evolutionary biology** is to understand how natural selection shapes traits within populations. Traits that improve survival or reproduction tend to increase in frequency over generations, while less advantageous traits decline. Traditionally, traits studied under selection include morphological features (e.g., body size, wing length) or behavioral traits (e.g., mating displays).

In recent years, however, researchers have become increasingly interested in **gene expression**—the process by which information from a gene is used to produce RNA and proteins. Unlike static traits, gene expression is **highly dynamic** and varies across individuals, tissues, developmental stages, and sexes. Because it directly influences cellular and physiological function, gene expression is a crucial link between the genome and organismal phenotype.

Importantly, gene expression also plays a role in **sexual dimorphism**—the differences between males and females of the same species. Genes expressed only in males (male-limited) or females (female-limited) can drive sex-specific adaptation, but may also create conflict if the same genes are expressed in both sexes but favored differently by selection.

Studying how selection acts on gene expression is therefore essential for answering broader evolutionary questions:

- Do males and females experience different strengths of selection on expression?
- Which genes or pathways are under the strongest selection, and do they relate to reproduction or survival?
- How can we accurately detect selection on thousands of traits simultaneously, given the complexity of genomic data?

To address these questions, researchers often use the **Lande–Arnold framework**, which relates traits to fitness using regression models. However, when thousands of genes are measured at once, traditional methods break down due to **high dimensionality** (too many predictors relative to sample size) and **correlation among genes**. This challenge has motivated the use of advanced statistical approaches—such as **dimensionality reduction, regularized regression, and network clustering**—to better estimate selection in genomic data.

Project Aim

The goal of this project was to investigate how natural selection acts on gene expression, a complex and highly variable trait that plays a central role in sexual dimorphism and evolutionary processes. Specifically, we aimed to:

- Estimate the strength and direction of selection on gene expression traits in Drosophila serrata, using previously published transcriptomic data.
- Compare multiple statistical approaches—including dimensionality reduction (PCA, fitness-associated PCA), regularized regression (ridge, lasso, elastic net), and clustering methods (WGCNA, stochastic block models)—to evaluate their ability to accurately capture selection gradients in high-dimensional datasets.
- Identify biologically meaningful genes and patterns, such as male- or female-limited expression, that may contribute to differences in selection pressures between the sexes.

By combining evolutionary biology theory with modern computational approaches, the project sought to determine which statistical methods best predict relative fitness from gene expression, and to uncover insights into the genetic basis of sexual dimorphism.

Methods

Data

We used previously published transcriptomic data from *Drosophila serrata*, comprising expression levels for nearly **6,000 genes** across **40 inbred lines**. Each line had associated **fitness measurements**, allowing us to link gene expression to reproductive success. To capture sex-specific patterns, genes were grouped into four datasets:

1. Male co-expressed genes
2. Female co-expressed genes
3. Male-limited genes
4. Female-limited genes

This subdivision allowed us to directly compare patterns of selection acting across sexes and gene categories.

Theoretical Framework

Our analysis was grounded in the **Lande–Arnold (1983) framework**, which models the relationship between phenotype and fitness as:

$$W = \alpha + \beta 1\, z1 + \beta 2 z2 + \cdots + \epsilon$$

where w is relative fitness, zi are standardized trait values (gene expression levels), and βi are the selection gradients representing the strength and direction of selection.

A major challenge of applying this framework to gene expression is **high dimensionality and multicollinearity** among thousands of traits. To address this, we implemented dimensionality reduction, penalized regression, and clustering approaches.

Statistical Approaches

1. **Dimensionality Reduction**
    - **Traditional PCA**: reduced expression data into orthogonal principal components, which were then regressed on fitness.
    - **Fitness-Associated PCA (Fit-PCA)**: optimized for components that best explained fitness variation, ensuring biologically relevant axes were emphasized.
2. **Regularized Regression**
    - **Ridge Regression ($\alpha = 0$)**: minimized squared error while penalizing large coefficients, controlling for multicollinearity without eliminating variables.
    - **Elastic Net Regression ($\alpha = 0.5$)**: combined ridge and lasso penalties, balancing shrinkage and sparsity in coefficient estimates.
    - **Lasso Regression ($\alpha = 1$)**: applied strong penalization that shrinks weaker coefficients to zero, effectively performing variable selection.
3. **Clustering Approaches**
    - **Weighted Gene Co-expression Network Analysis (WGCNA)**: identified modules of co-expressed genes, summarized by eigengenes, and related these module-level traits to fitness.
    - **Stochastic Block Model (SBM)**: a degree-corrected, probabilistic clustering method that partitions genes into modules based on co-expression structure, allowing us to test how selection acts on expression clusters.

Validation and Model Comparison

- For each method, **selection gradients (β)** were estimated and compared across the four datasets.
- To assess stability, we performed **100 bootstrap replicates**, reporting the mean and **95% confidence intervals**.
- To evaluate predictive performance, **15% of lines were held out** as test data. Predicted fitness was compared to observed fitness, with model accuracy quantified as **R^2 (variance explained)**.
- Directional similarity between methods was measured by calculating the **angles between β-vectors**, allowing us to assess concordance in estimated selection patterns.

Functional Analysis

For clustering methods (especially WGCNA), we performed **Gene Ontology (GO) enrichment analysis** to annotate co-expressed hub genes and identify functional themes. This provided biological interpretation of modules under the strongest estimated selection.

Implementation

The implementation of the project involved a combination of **data preparation, dimensionality reduction, model fitting, and biological interpretation**. Much of the work was carried out in **R**, with additional tools and databases used for functional annotation.

Data Preparation

We began by cleaning and standardizing the gene expression dataset from *Drosophila serrata*. This involved:

- Removing missing and extreme values.
- Standardizing expression values across lines to ensure comparability.
- Dividing the dataset into four biologically motivated subsets: male co-expressed, female co-expressed, male-limited, and female-limited genes.

To establish a baseline, we ran **univariate regressions** of each gene on relative fitness. This step allowed us to identify general trends, confirm data quality, and highlight potential candidate genes before applying more complex multivariate methods.

Dimensionality Reduction and Modeling

Because the dataset contained thousands of genes but only 40 samples, direct regression was not possible. To address this, we tested multiple data reduction strategies:

1. **Principal Component Analysis (PCA):** Reduced the dataset into a smaller number of orthogonal axes explaining most variance. PCs were then regressed on fitness.
2. **Weighted Gene Co-expression Network Analysis (WGCNA):** Grouped correlated genes into modules, which were summarized by eigengenes or hub genes. These module-level predictors were then regressed on fitness, preserving biological meaning in clusters.
3. **Regularized Regression:** Applied elastic net regression on principal components to balance variable selection (lasso) and shrinkage (ridge), improving robustness to multicollinearity.
4. **Weighted Nested Degree-Corrected Stochastic Block Model (wn-DC-SBM):** Implemented later in the project, this method treated genes as nodes in a network and identified nested, hierarchical clusters. The degree correction accounted for differences in connectivity, allowing us to reveal hidden gene-expression structures not captured by PCA or WGCNA.

Each approach produced reduced trait sets that were then used to estimate **selection gradients** following the Lande–Arnold framework. Bootstrapping (100 replicates) was used to assess stability and provide confidence intervals.

Functional Annotation

To interpret biological meaning, we automated the process of linking genes to known functions. We used **BLAT** to align expressed sequence tags (ESTs) to reference genes and retained the highest-quality matches. We then collected **Gene Ontology (GO) terms** for each gene, classifying their roles into biological processes, cellular components, and molecular functions. This provided a functional layer of interpretation for genes or modules under strong estimated selection.

Visualization and Communication

Throughout the implementation, we generated figures to visualize:

- Model performance (R^2 between predicted and observed fitness).
- Pairwise similarity in selection estimates across methods.
- Functional clustering of hub genes explaining variation in fitness.

In parallel, we prepared written summaries of key genes, built plots to accompany analyses, and began developing our **scientific poster**. The poster was designed to communicate both technical and biological insights in an accessible way to diverse audiences.

Results

Our analyses revealed consistent biological and statistical patterns across the different approaches, while also highlighting important contrasts between methods.

Sex-Specific Selection

- Across all approaches, **male gene expression** showed stronger directional selection than female expression.
- **Male-limited genes** were consistently identified as being under the strongest selection, suggesting they may play a central role in driving sexual dimorphism.

Comparison of Statistical Approaches

- **PCA and Ridge Regression** produced highly similar estimates of selection gradients, indicating both methods capture the main axes of selection in a consistent way.
- **Elastic Net and Lasso** provided comparable results but introduced additional sparsity, reducing the number of predictors considered important.
- **WGCNA** revealed functional clusters of co-expressed genes, with male co-expressed modules explaining up to **30% of the variation in fitness**. This suggests that module-level summaries can retain biological signal while reducing complexity.
- **wn-DC-SBM** identified nested, hierarchical clusters of genes not captured by PCA or WGCNA, offering a complementary perspective on network structure and potential hidden patterns of selection.
- Overall, **fitness-associated PCA and ridge regression** had the highest predictive accuracy in test data, explaining the largest proportion of variance in relative fitness.

Candidate Genes Under Selection

- Across multiple approaches, the gene **hp:CR32205** (a hairpin RNA involved in regulating male fertility genes) was consistently identified as being under strong positive directional selection for higher expression.
- Functional annotation of hub genes from WGCNA modules showed enrichment in categories related to **neuronal processes, apoptosis/proteolysis, and muscle structure**, indicating these biological processes may be important targets of selection in males.

Model Performance

- Predictive performance varied across methods:
    - **Fitness-associated PCA and Ridge Regression** achieved the best cross-validation results (highest R^2 in test data).
    - **Clustering approaches (WGCNA, wn-DC-SBM)** provided more biologically interpretable results but showed lower predictive accuracy compared to regression-based methods.
- Pairwise comparisons of β-vectors revealed high concordance between PCA and regression methods, while clustering-based approaches showed greater divergence in estimated selection directions.

Poster

