BRAUDE
College of Engineering, karmiel

Software Engineering Department

Braude College

Capstone Project Phase B

Gender-Based Analysis of Biological Pathways in Gene Expression

Supervisor: Mr. Gur Arye Yehuda

Mor Shmuel - Mor.Shmuel@e.braude.ac.il

Maayan Avittan - Maayan.Avittan@e.braude.ac.il

Table of Content

Git repository link:

https://github.com/avtn96/Final-Project.git

Abstract

Bioinformatics is an interdisciplinary field that integrates biology, computer science, and statistics to analyze large-scale biological data, such as genomics. Genes, as sequences of DNA, encode instructions for cellular functions, and differential gene expression refers to the varying levels at which these genes are expressed under different conditions or across groups, including between males and females. Sex-based differences at the gene level, influenced by sex chromosomes and hormones, shape key biological pathways—complex sequences of molecular interactions leading to specific cellular outcomes.

This research investigates how gender differences influence gene expression across multiple tissues using data from the Genotype-Tissue Expression (GTEx) project. We used Python-based pipelines, statistical analyses using t-tests, and machine learning models to identify genes with significant sex-biased expression across tissues, including adipose-visceral, muscle-skeletal, nerve-tibial, and liver. These genes were subsequently mapped to biological pathways using enrichment analysis with the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, providing insights into systemic and tissue-specific molecular mechanisms.

Key results revealed the recurring enrichment of immune-related pathways, such as Herpes Simplex Virus 1 Infection, and signal transduction pathways, including Wnt Signaling and Inositol Phosphate Metabolism, across multiple tissues. Additionally, tissue-specific pathways, such as Ubiquitin-Mediated Proteolysis in the liver and Insulin Signaling in skeletal muscle, underscored distinct metabolic and regulatory processes shaped by gender. A Random Forest machine learning model achieved high classification accuracy, particularly in adipose-visceral and skeletal muscle tissues, highlighting the predictive power of sex-biased gene expression patterns. Genes ranked highly in the model's feature importance aligned with findings from the enrichment analysis, further validating the biological significance of these pathways.

This study highlights how gender shapes immune regulation, metabolic processes, and cancer susceptibility at the molecular level. Integrating statistical and machine learning

approaches offers a robust framework for understanding sex-biased biological pathways. It lays the groundwork for precision medicine interventions that account for gender-specific genetic differences.

Keywords: Human Genetics, Genomic Data Analysis, Bioinformatics, Genetic Mapping, Statistical Analysis, Genome-wide Association Studies(GWAS), Gender Classification, Random Forest.

Introduction

Gender differences significantly influence biological processes, impacting disease susceptibility, treatment responses, and physiological functions. These differences arise from variations in gene expression between males and females, driven by sex chromosomes and hormones. For instance, autoimmune diseases are more prevalent in females, whereas certain non-reproductive cancers are more common in males. Understanding these differences is vital for advancing precision medicine and developing gender-specific therapeutic strategies.

This research project aims to explore the differences in gene expressions across multiple human tissues, focusing on identifying biological pathways that exhibit gender-specific characteristics. By leveraging the Genotype-Tissue Expression (GTEx) project—a comprehensive dataset providing RNA sequencing data from various human tissues—the study investigates how genetic expression varies between males and females. This data forms the foundation for analyzing tissue-specific gene expression differences and uncovering the underlying biological mechanisms.

To accomplish this, the study employs statistical methods, particularly the t-test, to identify genes with significant differential expression between genders. Using a Python-based pipeline, the project processes gene expression data from selected tissues, integrates phenotype information, and performs statistical analysis to highlight genes with sex-biased expression. The identified genes are then analyzed through a Random

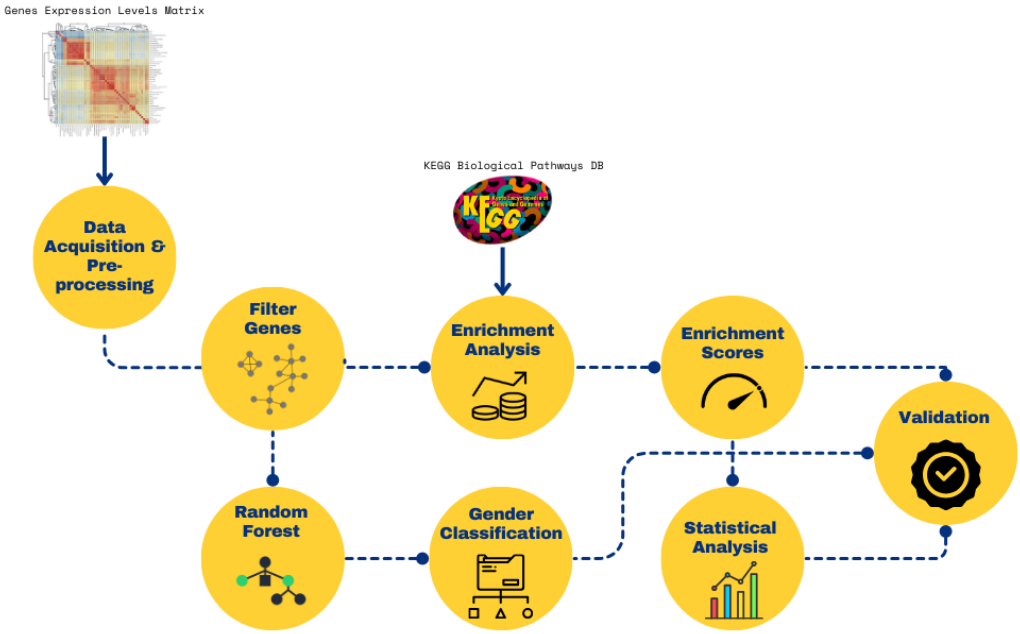
Forest model, a machine-learning algorithm used to predict sex based on gene expression patterns. This allows for the evaluation of gene importance in sex differentiation, providing insights into key genes that contribute to biological differences between genders.

Further, the identified sex-biased genes are inputted into an enrichment analysis tool, utilizing the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. KEGG is a curated resource that links genes, proteins, and small molecules to biological pathways and cellular mechanisms. By mapping the identified genes to KEGG pathways, the study highlights the molecular mechanisms and functional systems associated with gender-specific traits and conditions. Results from the enrichment analysis were visualized and analyzed to identify pathways that occur across multiple tissues, revealing consistent biological differences across different organ systems. To validate these findings, additional online research was conducted to confirm the identified pathways and their biological relevance.

The target audience for this framework includes researchers, medical professionals, and bioinformatics practitioners. By offering a detailed analysis of gender-biased gene expression and associated biological pathways, this study contributes to the broader understanding of gender-specific traits and conditions. Ultimately, it seeks to enhance the field of precision medicine by identifying potential biomarkers and therapeutic targets tailored to gender-specific needs.

Research Process

The research process for Phase B builds upon Phase A, focusing on the analysis of large-scale genomic datasets, extracting tissue-specific data, and conducting statistical analyses to identify gender-biased gene expression patterns. Key steps included preprocessing large datasets, performing statistical tests, mapping enriched pathways, and integrating machine learning for further analysis. This iterative process was designed to adapt to challenges such as memory limitations, data heterogeneity, and pathway interpretation.



Data Handling

To address the challenge of handling large datasets, the data was divided into manageable chunks. Using Python and Pandas, data files were read and processed iteratively to ensure efficient memory utilization. This approach allowed processing even on systems with limited resources.

Obstacles in Data Handling:

1. **Large File Sizes:** Large-scale genomic datasets often lead to memory issues. To mitigate this, chunk processing was implemented to load and process data in smaller segments.
2. **Data Integration:** Combining processed chunks into a unified dataset requires careful management to maintain consistency and accuracy.

Tissue-Specific Filtering

Metadata was utilized to filter data for specific tissues of interest, such as Adipose, Muscle, Nerve, and Liver. This step ensured that subsequent analyses focused only on relevant samples. Adipose tissue was used as a test case, with subject IDs extracted and matched to the corresponding gene expression data.
We acquired tissue-specific data for four tissues:

1. **Adipose - Visceral (Omentum):** Visceral adipose tissue plays a critical role in energy metabolism and has been linked to sex differences in metabolic conditions such as obesity and diabetes. Investigating this tissue provides insight into how sex affects fat storage and hormonal regulation.
2. **Muscle - Skeletal:** Skeletal muscle is essential for locomotion, energy expenditure, and metabolic health. Differences in muscle mass and composition between males and females make this tissue an important focus for understanding sex-specific adaptations.
3. **Nerve - Tibial:** The tibial nerve is a major component of the peripheral nervous system, involved in motor and sensory functions. Exploring this tissue helps to uncover potential sex-based differences in neurological health and disease susceptibility.
4. **Liver:** The liver is central to metabolism, detoxification, and hormone regulation. Gender differences in liver function have implications for diseases such as non-alcoholic fatty liver disease and liver cancer, making it a vital tissue for this study.

Statistical Analysis

For each tissue, we conducted a statistical analysis to identify genes with significant differential expression between males and females. This phase involved the following steps:

T-Test

The t-test is a robust statistical method used to determine whether there is a significant difference between the means of two groups. In this project, we employed the independent t-test to compare gene expression levels between male and female samples.

- **Null Hypothesis:** The expression levels of a given gene are the same in males and females.
- **Alternative Hypothesis:** The expression levels of a given gene differ between males and females.
- **Calculation:** For each gene, the t-statistic was calculated to measure the difference between group means relative to the variation within the groups.
- **P-Value Threshold:** A p-value threshold (<0.02) was applied to identify genes with statistically significant differences. Genes meeting this threshold were flagged for further analysis.
- **Batch Processing:** To optimize performance, we iteratively ran the t-tests for thousands of genes, leveraging Python libraries like SciPy and tqdm to handle large-scale computations efficiently.

Implementation

- **Metadata Integration:** Merged gene expression data with phenotype metadata, particularly the gender column, to group samples accurately.
- **Gene-Wise Analysis:** Ran t-tests for each gene in the dataset, generating a comprehensive list of significant genes for each tissue.
- **Result Storage:** Compiled the results into tissue-specific CSV files for downstream analysis and enrichment.

Enrichment Analysis

Biological pathways represent interconnected networks of molecular interactions that drive cellular processes and maintain organismal homeostasis. These pathways are critical for understanding how genes, proteins, and metabolites work together to regulate biological functions such as metabolism, immune responses, cell signaling, and energy production. Disruptions or alterations in these pathways often underlie diseases, making them valuable for identifying potential targets for therapeutic interventions. By examining pathways, we gain insight into how groups of genes interact and contribute to broader biological processes, helping to contextualize the molecular changes observed in our study.

The significant genes identified through the t-tests were inputted into **Enrichr**, a widely used tool for enrichment analysis. Enrichr maps input gene sets to databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) to identify enriched biological pathways. It ranks pathways based on statistical significance, providing metrics such as p-values and overlap counts to quantify how strongly the input genes are associated with each pathway. This allowed us to focus on pathways most relevant to our dataset, offering a clearer understanding of the biological functions and processes represented by the genes of interest.

The statistical foundation of Enrichr relies on the **hypergeometric distribution**, which calculates the likelihood of observing a certain overlap between the input gene set and a predefined pathway gene set. This approach considers the size of the input gene set, the size of the pathway gene set, and the total number of genes in the database. By comparing the observed overlap to what would be expected by random chance, Enrichr calculates a p-value to indicate the statistical significance of the enrichment. Lower p-values signify stronger evidence that the observed overlap is not due to chance, highlighting pathways that are most likely biologically relevant to the input gene set.

KEGG Pathway Mapping

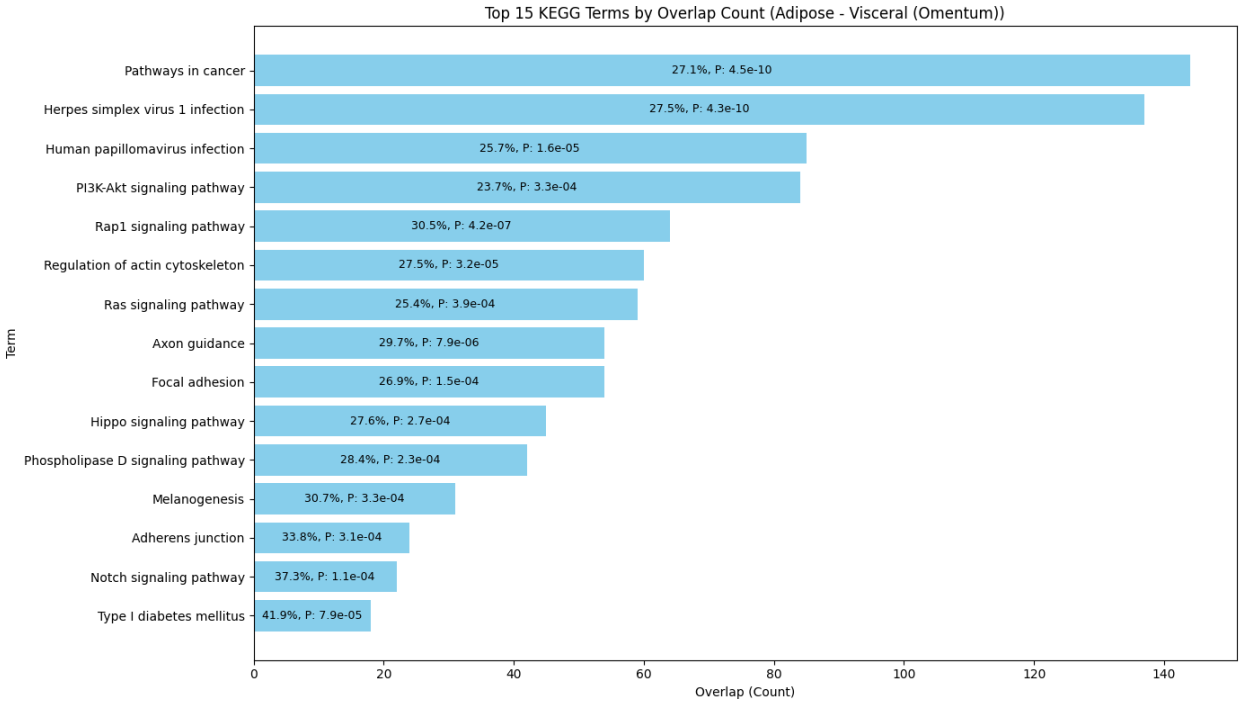
Using Enrichr, we employed KEGG pathway mapping to link significant genes to known biological pathways. This analysis provided insights into the broader biological functions and interactions of the identified genes. Pathways such as those involved in metabolism, immune regulation, and cancer emerged as particularly relevant to our study. Given our focus on gender-specific differences, we prioritized pathways that aligned with known or hypothesized sex-biased processes, as well as those that were easier to interpret and commonly studied in the literature.

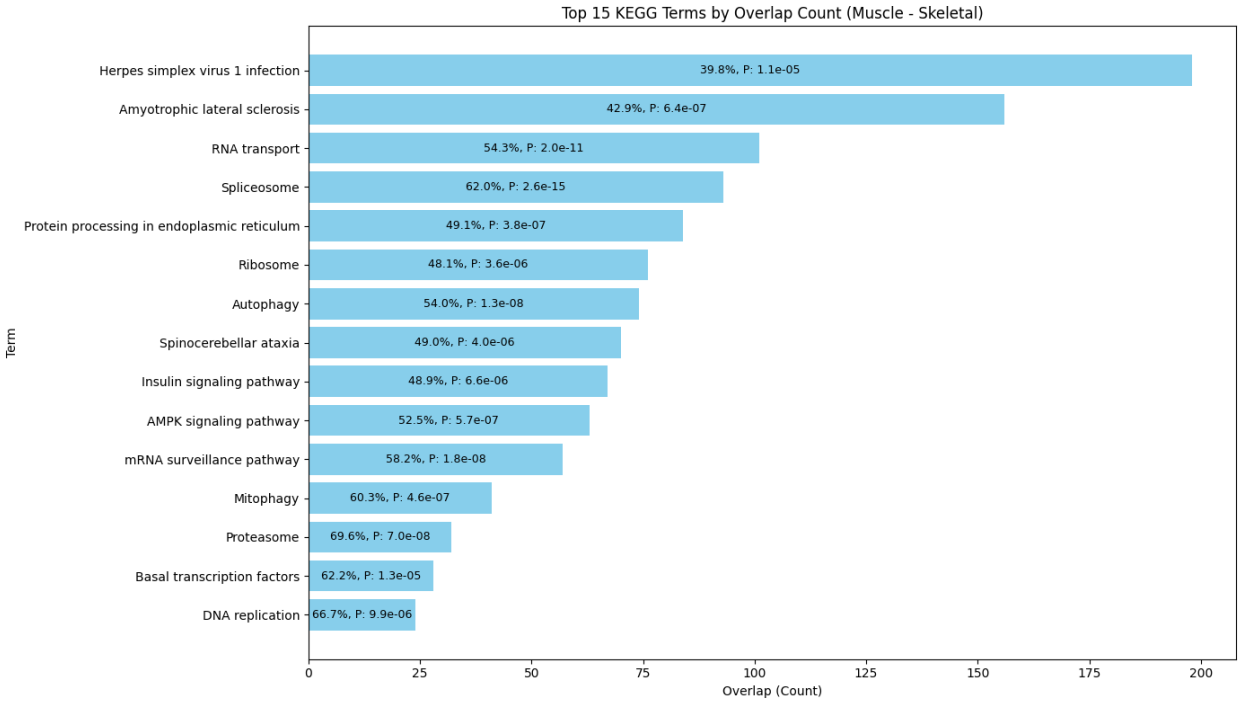
Pathways Selection and Validation

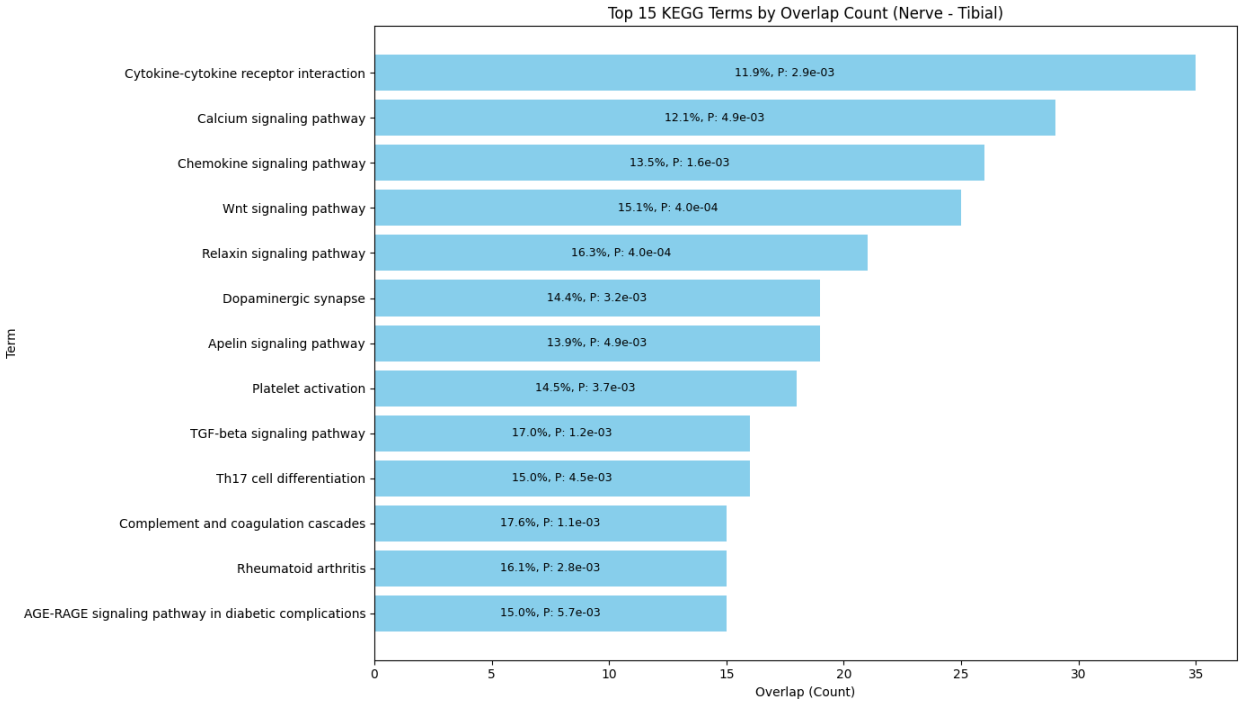
To ensure the validity of our findings, we performed an extensive literature review to contextualize the identified pathways and their biological significance. For pathways that appeared across multiple tissues, we performed cross-referencing to ensure consistency with existing knowledge. Recurring pathways, such as **Herpes Simplex Virus 1 Infection** and **Wnt Signaling**, were given special attention, as their presence across tissues provided stronger evidence for their systemic relevance to gender differences.

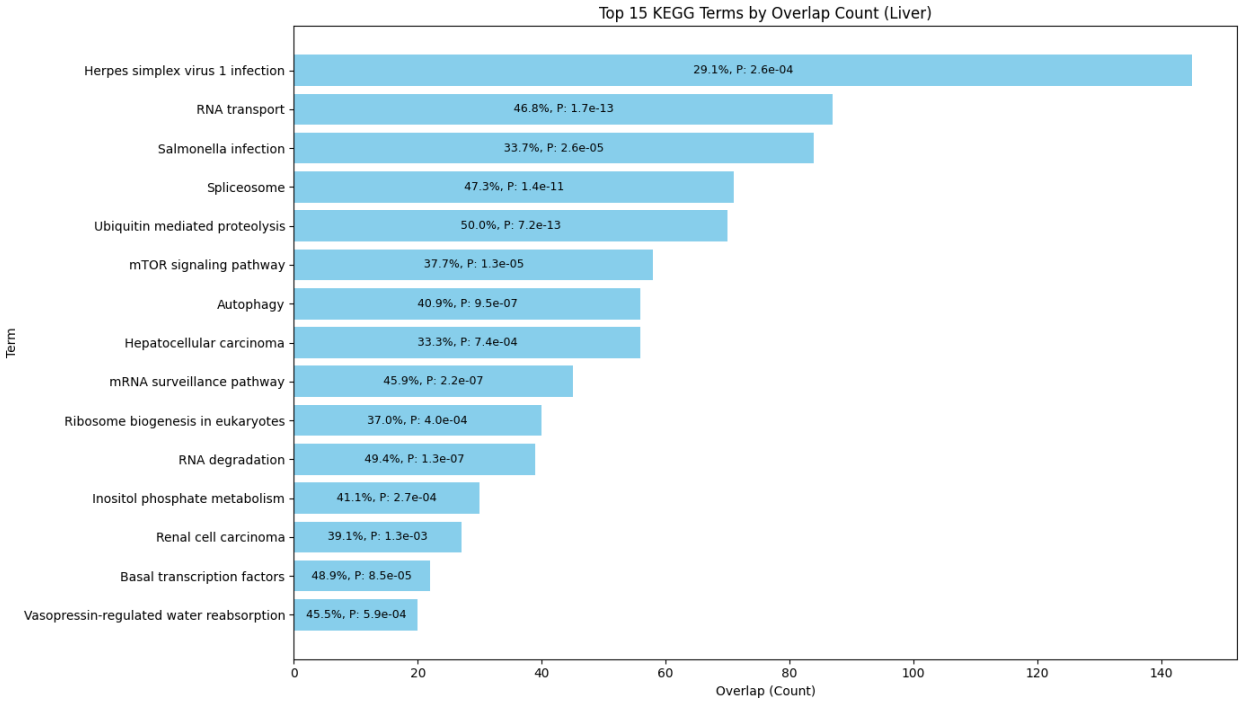
Visualization

To effectively communicate our findings, we created visualizations of the KEGG pathways, generating horizontal bar charts that displayed the top 15 pathways for each tissue based on p-value significance. These charts included annotations for overlap counts and p-values, highlighting the strength and biological relevance of each pathway. The visualizations provided an accessible way to present the results, helping to convey the connections between the significant genes and their broader biological contexts.

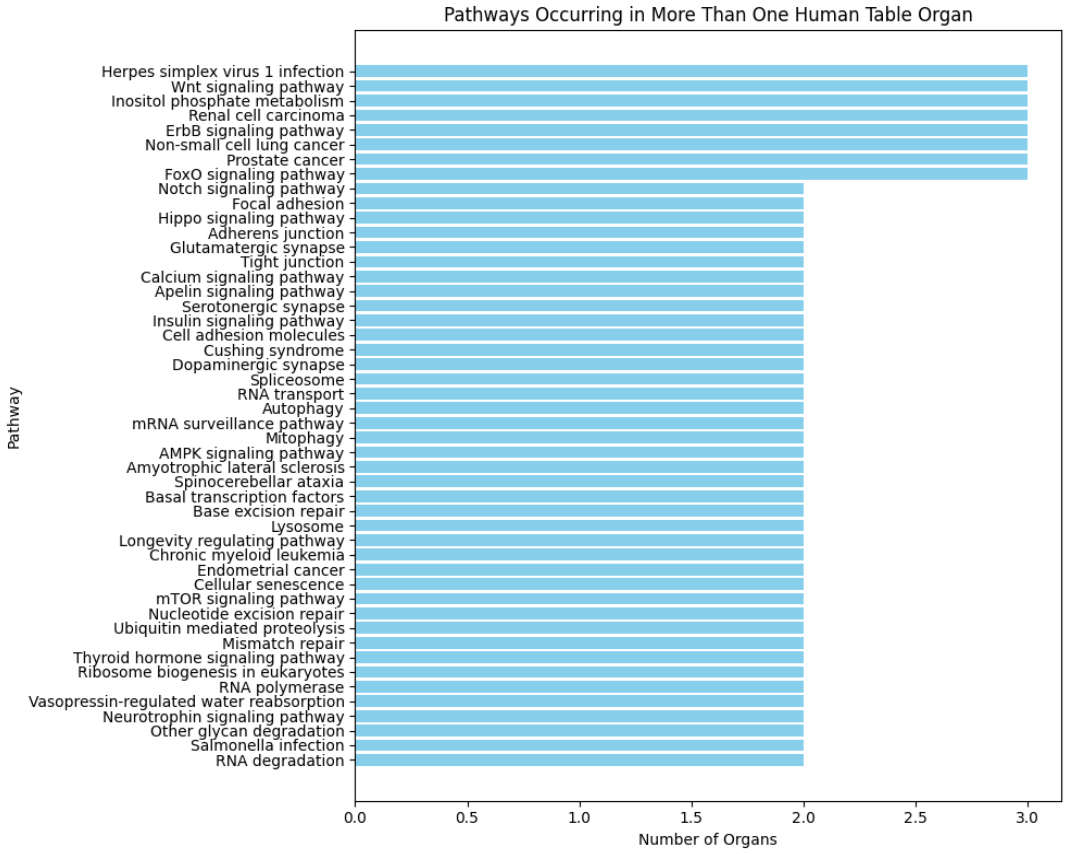

Top 15 KEGG Terms by Overlap Count (Adipose - Visceral (Omentum))
Term
Pathways in cancer
Herpes simplex virus 1 infection
Human papillomavirus infection
PI3K-Akt signaling pathway
Rap1 signaling pathway
Regulation of actin cytoskeleton
Ras signaling pathway
Axon guidance
Focal adhesion
Hippo signaling pathway
Phospholipase D signaling pathway
Melanogenesis
Adherens junction
Notch signaling pathway
Type I diabetes mellitus
27.1%, P: 4.5e-10
27.5%, P: 4.3e-10
25.7%, P: 1.6e-05
23.7%, P: 3.3e-04
30.5%, P: 4.2e-07
27.5%, P: 3.2e-05
25.4%, P: 3.9e-04
29.7%, P: 7.9e-06
26.9%, P: 1.5e-04
27.6%, P: 2.7e-04
28.4%, P: 2.3e-04
30.7%, P: 3.3e-04
33.8%, P: 3.1e-04
37.3%, P: 1.1e-04
41.9%, P: 7.9e-05
0
20
40
60
80
100
120
140
Overlap (Count)


Top 15 KEGG Terms by Overlap Count (Muscle - Skeletal)
Term
Herpes simplex virus 1 infection
Amyotrophic lateral sclerosis
RNA transport
Spliceosome
Protein processing in endoplasmic reticulum
Ribosome
Autophagy
Spinocerebellar ataxia
Insulin signaling pathway
AMPK signaling pathway
mRNA surveillance pathway
Mitophagy
Proteasome
Basal transcription factors
DNA replication
39.8%, P: 1.1e-05
42.9%, P: 6.4e-07
54.3%, P: 2.0e-11
62.0%, P: 2.6e-15
49.1%, P: 3.8e-07
48.1%, P: 3.6e-06
54.0%, P: 1.3e-08
49.0%, P: 4.0e-06
48.9%, P: 6.6e-06
52.5%, P: 5.7e-07
58.2%, P: 1.8e-08
60.3%, P: 4.6e-07
69.6%, P: 7.0e-08
62.2%, P: 1.3e-05
66.7%, P: 9.9e-06
0
25
50
75
100
125
150
175
200
Overlap (Count)


Top 15 KEGG Terms by Overlap Count (Nerve - Tibial)
Term
Cytokine-cytokine receptor interaction
Calcium signaling pathway
Chemokine signaling pathway
Wnt signaling pathway
Relaxin signaling pathway
Dopaminergic synapse
Apelin signaling pathway
Platelet activation
TGF-beta signaling pathway
Th17 cell differentiation
Complement and coagulation cascades
Rheumatoid arthritis
AGE-RAGE signaling pathway in diabetic complications
11.9%, P: 2.9e-03
12.1%, P: 4.9e-03
13.5%, P: 1.6e-03
15.1%, P: 4.0e-04
16.3%, P: 4.0e-04
14.4%, P: 3.2e-03
13.9%, P: 4.9e-03
14.5%, P: 3.7e-03
17.0%, P: 1.2e-03
15.0%, P: 4.5e-03
17.6%, P: 1.1e-03
16.1%, P: 2.8e-03
15.0%, P: 5.7e-03
0
5
10
15
20
25
30
35
Overlap (Count)


Top 15 KEGG Terms by Overlap Count (Liver)
Term
Herpes simplex virus 1 infection
RNA transport
Salmonella infection
Spliceosome
Ubiquitin mediated proteolysis
mTOR signaling pathway
Autophagy
Hepatocellular carcinoma
mRNA surveillance pathway
Ribosome biogenesis in eukaryotes
RNA degradation
Inositol phosphate metabolism
Renal cell carcinoma
Basal transcription factors
Vasopressin-regulated water reabsorption
29.1%, P: 2.6e-04
46.8%, P: 1.7e-13
33.7%, P: 2.6e-05
47.3%, P: 1.4e-11
50.0%, P: 7.2e-13
37.7%, P: 1.3e-05
40.9%, P: 9.5e-07
33.3%, P: 7.4e-04
45.9%, P: 2.2e-07
37.0%, P: 4.0e-04
49.4%, P: 1.3e-07
41.1%, P: 2.7e-04
39.1%, P: 1.3e-03
48.9%, P: 8.5e-05
45.5%, P: 5.9e-04
0
20
40
60
80
100
120
140
Overlap (Count)

To explore pathways recurring across multiple tissues, we conducted Pathway Organ Mapping, created a mapping between pathways and the tissues in which they appeared using KEGG pathway data from each tissue, and did an Identification of Common Pathways. We identified pathways present in more than one tissue to assess their systemic biological significance.



This visualization highlights pathways that are common across multiple tissues, providing strong evidence of systemic biological differences by gender. Examples include pathways related to signaling and metabolism, which appear in multiple tissues and are critical for understanding gender-specific health differences.

Machine Learning Integration

The machine learning integration phase of our research aimed to complement the statistical analysis by developing a predictive model to classify biological samples as male or female based on gene expression data. This approach provided a robust way to validate the biological relevance of the genes identified in earlier steps and to further uncover key contributors to gender-specific differences.

Preparing the Data

Gene expression data from each tissue served as the input for machine learning. The associated gender labels (`1` for male and `2` for female) were used as the target variable.

The data was standardized to ensure that all gene expression values had a mean of 0 and a standard deviation of 1. This step improves the performance and convergence of the machine learning model.

The dataset was split into training (70%) and testing (30%) subsets using stratified sampling to maintain balanced gender representation in both subsets.

Random Forest

Random Forest is an ensemble machine-learning algorithm that combines the predictions of multiple decision trees. Each decision tree is trained on a random subset of the data, and its predictions are aggregated (majority voting for classification) to produce the final result.

How It Works:

- Random subsets of the training data are created using a process called bootstrapping (sampling with replacement).
- For each subset, a decision tree is trained, but only a random subset of features is considered at each split. This randomness helps prevent overfitting.
- During prediction, each tree outputs a class label, and the Random Forest aggregates these predictions (e.g., by majority vote).

Random Forest handles high-dimensional datasets like gene expression data efficiently.

It is resistant to overfitting due to its ensemble nature, where multiple decision trees are trained on random subsets of the data.
The algorithm provides feature importance scores, allowing us to identify the genes that contribute most to distinguishing between males and females.

Model Training and Evaluation

The Random Forest model was trained on the standardized gene expression data from the training set. The number of decision trees was set to 100, and a fixed random state was used for reproducibility.

The trained model was evaluated on the test set using the following metrics:

- Accuracy: The percentage of correctly classified samples.
- Precision: The proportion of true positive predictions among all positive predictions.
- Recall: The proportion of true positives among all actual positive samples.
- F1-Score: The harmonic mean of precision and recall.
- Confusion Matrix: A visualization showing the number of correct and incorrect predictions for both male and female samples.

Cross-Validation

Cross-validation is a statistical method used in machine learning to assess the performance of a model by testing it on multiple subsets of the data. It ensures that the model is robust and performs well on unseen data, thereby reducing the risk of overfitting.
To ensure the robustness of the model, we performed 5-fold cross-validation. This approach provided a more reliable measure of the model's performance by testing it on different subsets of the data.

Tissue-Specific Results

The machine learning analysis was conducted separately for each tissue, giving the following results:

Classification Report: Detailed metrics (accuracy, precision, recall, F1-score) for each tissue provide insights into the model's performance.

Confusion Matrix: Displays the counts of true positives, false positives, true negatives, and false negatives for each tissue.

Cross-Validation Scores: Assesses the generalizability of the model across different data splits.

Feature Importance Visualization: Highlighted the top genes contributing to the model's decision-making for each tissue.

Conclusion

Phase B focused on refining the research process to handle large datasets effectively and perform tissue-specific analyses. By combining robust preprocessing, efficient data handling, targeted statistical tests, and integration of machine learning differences the phase aimed to uncover significant gender-based differences in gene expression across multiple tissues. These findings provide a strong foundation for downstream enrichment analysis and machine learning models in subsequent phases.

Challenges and Solutions

1. **Managing Large Data from GTEx**

   The GTEx dataset(V10) is extensive, containing a vast amount of gene expression data from various tissues. Initially, we encountered difficulties loading the entire dataset into Google Colab due to its size, which exceeded the platform's memory limitations.

   To address this issue, we split the dataset into smaller, manageable chunks. This approach allowed us to load and process the data incrementally. We then filtered these chunks to extract only the specific tissues of interest, significantly reducing the memory load and enabling efficient data analysis.

2. **Interpreting Pathways from the KEGG Database**

   During the enrichment analysis phase, we insert the list of differentially expressed genes into Enrichr and use the KEGG database to identify relevant pathways. However, many of the pathways retrieved were unfamiliar to us and lacked sufficient explanatory information, making them challenging to interpret, especially given our limited background in biology.

   To overcome this challenge, we focused on pathways that were more commonly studied and widely understood. We prioritized pathways that were familiar to us or had clear, detailed descriptions in the available literature and online resources. This selective approach ensured that our analysis remained focused and interpretable, allowing us to draw meaningful conclusions.

3. **Optimizing the T-Test Implementation**

   Running t-tests on thousands of genes for multiple tissues proved computationally intensive and time-consuming. The large number of statistical comparisons also increased the risk of false positives due to multiple tests.

   To optimize the t-test implementation, we utilized efficient libraries such as SciPy, which provided faster computation. Additionally, we leveraged Python's 'tqdm' library to monitor the progress of the tests and identify bottlenecks. Post-analysis, we applied a p-value threshold to reduce false positives and focused on the most statistically significant results for further exploration.

Results

The analysis of sex-biased gene expression revealed key biological pathways enriched across multiple human tissues, offering insights into the molecular mechanisms underlying gender differences. By integrating T-test results with machine learning approaches such as Random Forest and Random Forest modeling, we identified pathways where gender-related gene expression differences were particularly pronounced. These findings reinforce the systemic influence of gender on immune response, metabolic regulation, and cancer susceptibility.
Notably, several pathways, including Herpes Simplex Virus 1 (HSV-1) Infection, Wnt Signaling, Inositol Phosphate Metabolism, and Ubiquitin-Mediated Proteolysis, were highlighted in both statistical and machine learning analyses, reinforcing their biological relevance in sex-biased gene expression. A more in-depth exploration of these pathways and their biological implications will be presented in the subsequent sections.

A recurring theme was the prominence of immune-related pathways, particularly Herpes Simplex Virus 1 (HSV-1) Infection, which was significantly enriched in liver, visceral adipose (omentum), and skeletal muscle tissues. This pathway involves the host immune response to the HSV-1 virus, including processes such as viral entry, replication, and immune evasion. Studies suggest that males often experience worse outcomes in HSV-1 infections due to lower levels of interferon-gamma, a critical immune mediator influenced by testosterone, whereas females may benefit from estrogen-enhanced immune responses Han et al. (2001) [6]. The fact that HSV-1 Infection was highlighted in both T-tests and Random Forest analysis across multiple tissues suggests systemic sex-based differences in viral immunity, which could have implications for susceptibility and disease severity.

Beyond immune function, signal transduction pathways such as Wnt Signaling and Inositol Phosphate Metabolism emerged prominently across tissues. Wnt Signaling, which plays a crucial role in cell differentiation, tissue repair, and immune modulation, was detected in both statistical and machine learning approaches, underscoring its sex-

biased regulatory role Kim (2015) [7]. Hormonal influences, particularly estrogen, are known to modulate Wnt activity, enhancing its role in tissue-specific repair processes. Similarly, Inositol Phosphate Metabolism, a pathway that regulates cellular signaling and energy homeostasis, was repeatedly enriched across tissues. This pathway governs calcium release and secondary messenger production, processes that are critical for hormonal signaling and metabolic balance Subramanian et al. (2005) [10]. Its enrichment suggests distinct roles for male and female hormones in modulating metabolic control and cellular communication. The significance of these pathways in both T-tests and machine learning indicates that sex differences in tissue repair and metabolic regulation are strongly conserved across tissues, suggesting a systemic influence of sex hormones on gene expression dynamics.

Cancer-related pathways, such as Renal Cell Carcinoma, Non-Small Cell Lung Cancer, and Prostate Cancer, were also recurrently enriched across tissues, highlighting the profound impact of gender on cancer biology. Prostate cancer is male-specific due to the presence of the prostate gland, but the hormonal environment in males, particularly testosterone, can exacerbate cancer growth Lee (2023) [4]. Conversely, females typically have lower incidences of renal cell carcinoma and non-small cell lung cancer, potentially due to the protective effects of estrogen on DNA repair mechanisms Dohmen et al. (2003) [5]. The consistent enrichment of these pathways in both the statistical and Random Forest results suggests that hormonal and genetic factors contribute to sex-biased cancer risk and progression.

The Random Forest Model, developed to predict gender based on gene expression profiles, provided an additional layer of validation and biological insight. The model demonstrated exceptional accuracy in tissues such as adipose-visceral and muscle-skeletal (100% accuracy), suggesting clear and distinct sex-based differences in gene expression in these tissues. The confusion matrix in these cases confirmed that every male and female sample was correctly classified, highlighting the robustness of the gene expression data Breiman (2001) [3]. Conversely, in tissues such as nerve-tibial and liver, the model achieved lower accuracies (73% and 90%, respectively), as reflected in the confusion matrix and cross-validation scores. This lower accuracy may indicate

overlapping gene expression patterns or less pronounced sex-specific differences in these tissues.

The confusion matrix provides a summary of the model's predictions, showing the true positives (correctly predicted males and females) and false positives/negatives (misclassifications). In high-performing tissues like muscle-skeletal, there were no misclassifications, indicating that gene expression patterns were highly distinct between males and females. However, in tissues like nerve-tibial, some male samples were misclassified as female and vice versa, suggesting a more complex biological landscape where sex-biased gene expression is less pronounced or confounded by shared regulatory mechanisms Breiman (2001) [3].

Cross-validation scores further validated the model's performance by assessing its stability and generalizability across multiple iterations. The consistently high cross-validation scores in tissues like adipose-visceral and muscle-skeletal underscore the reliability of the model in identifying sex-biased gene expression Breiman (2001) [3]. In contrast, slightly lower cross-validation scores in nerve-tibial and liver tissues reflect the inherent challenges in these datasets, potentially due to external factors such as environmental influences or overlapping biological processes Breiman (2001) [3].

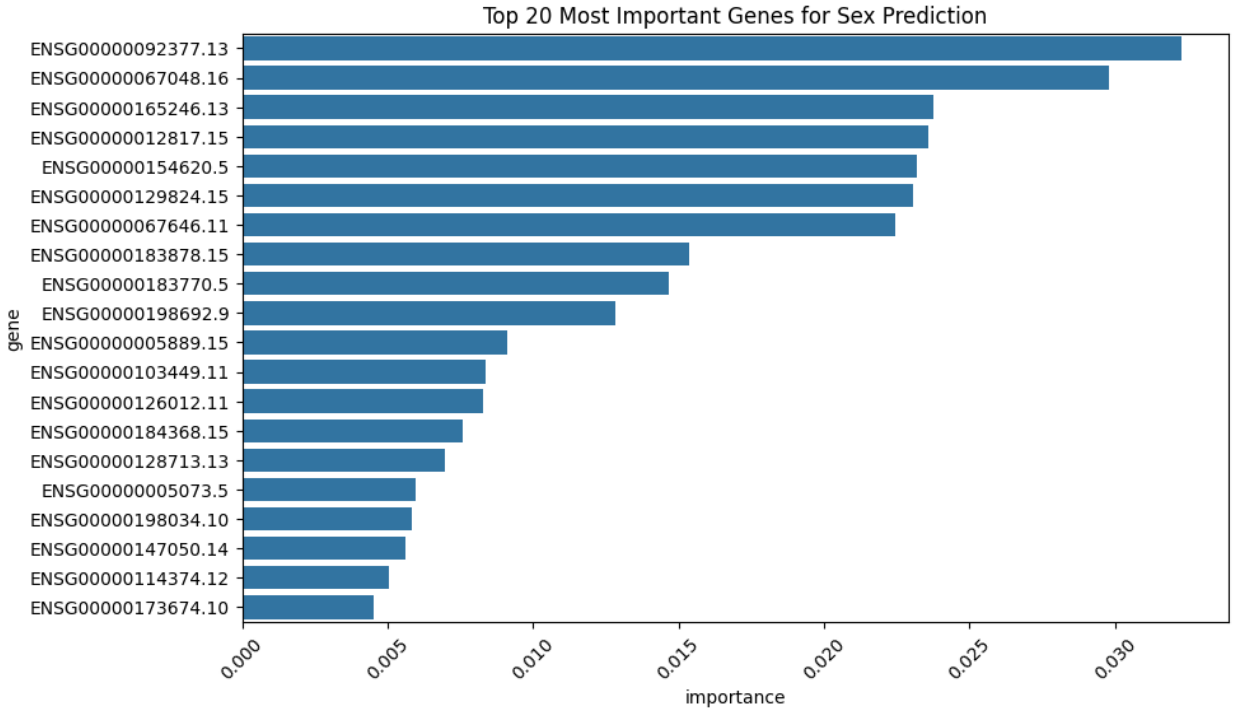
Importantly, genes identified as significant in the statistical analysis, such as those related to immune regulation and metabolic pathways, were among the most influential features in the model's predictions. This alignment between statistical findings and machine learning results strengthens the study's conclusions, highlighting the systemic and tissue-specific nature of sex-biased gene expression.

Tissue-specific findings further elucidate the influence of gender on biological pathways. In liver tissue, pathways such as Ubiquitin-Mediated Proteolysis (p-value: 7.2e-13) and Autophagy (p-value: 9.5e-07) were strongly enriched in both T-tests and Random Forest analysis, emphasizing their critical roles in cellular maintenance and metabolic regulation Chen et al. (2013) [2], Lista et al. (2011) [8]. Ubiquitin-mediated proteolysis ensures protein quality control, while autophagy supports the recycling of cellular components during metabolic stress. Hormonal differences likely modulate these

processes, explaining disparities in liver diseases such as hepatocellular carcinoma, which is more prevalent in males Dohmen et al. (2003) [5]. Similarly, in skeletal muscle, the enrichment of the Insulin Signaling Pathway (p-value: 6.6e-06) highlights sex differences in glucose metabolism, with females generally exhibiting greater insulin sensitivity Peer et al. (2021) [9]. These differences were reflected in the Random Forest model, where muscle-specific genes involved in these pathways strongly contributed to the model's predictions.

The recurrence of pathways such as Herpes Simplex Virus 1 Infection, Wnt Signaling, and Inositol Phosphate Metabolism across multiple tissues reflects systemic biological themes influenced by gender. These shared pathways suggest that sex-biased regulation of immune responses, signaling networks, and metabolic processes is coordinated across tissues rather than isolated to specific organs Subramanian et al. (2005) [10]. Hormonal and genetic differences, such as the presence of estrogen or testosterone, likely drive these systemic effects, shaping overarching biological principles and impacting health outcomes in both males and females.

In conclusion, the integration of enrichment analysis and machine learning provides compelling evidence of systemic and tissue-specific sex-biased biological processes. The results offer a deeper understanding of how gender shapes immune regulation, metabolic processes, and disease susceptibility, paving the way for more targeted and effective therapeutic strategies tailored to male and female physiology.

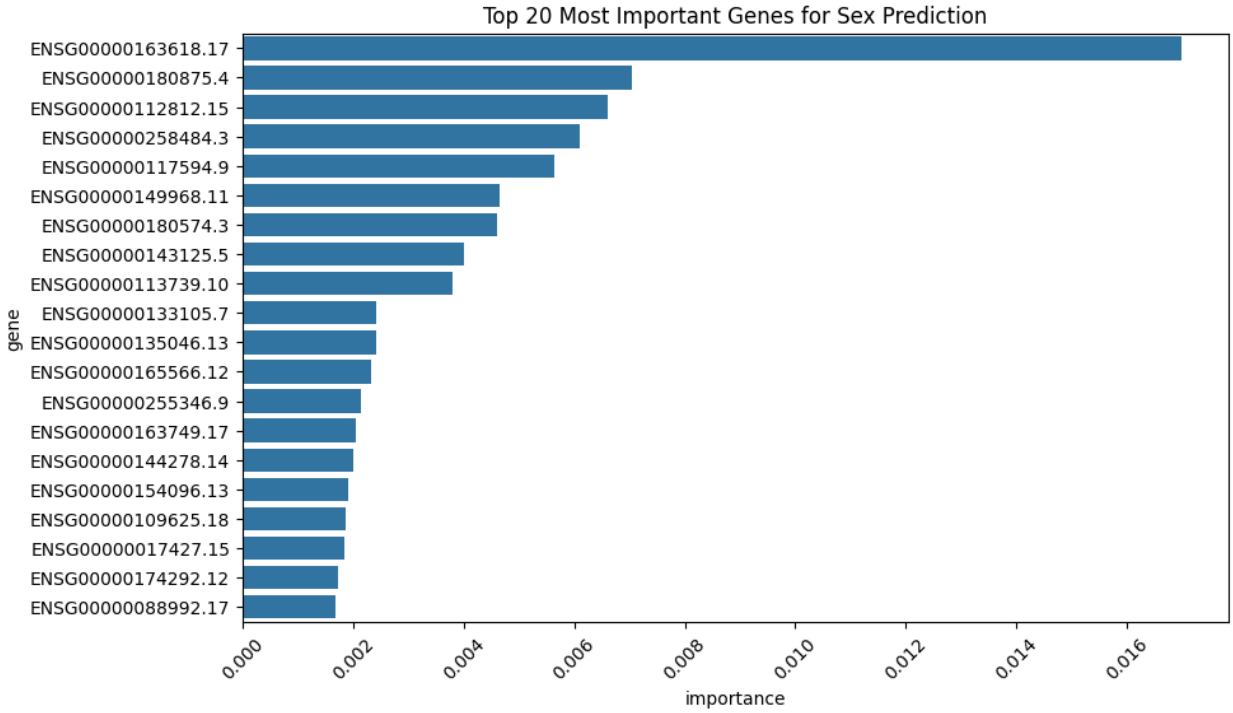



```
Results for Adipose - Visceral (Omentum):
Classification Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00       112
           2       1.00      1.00      1.00        51

    accuracy                           1.00       163
   macro avg       1.00      1.00      1.00       163
weighted avg       1.00      1.00      1.00       163

Confusion Matrix:
[[112   0]
 [  0  51]]
Cross-Validation Scores:
[1. 1. 1. 1. 1.]
```

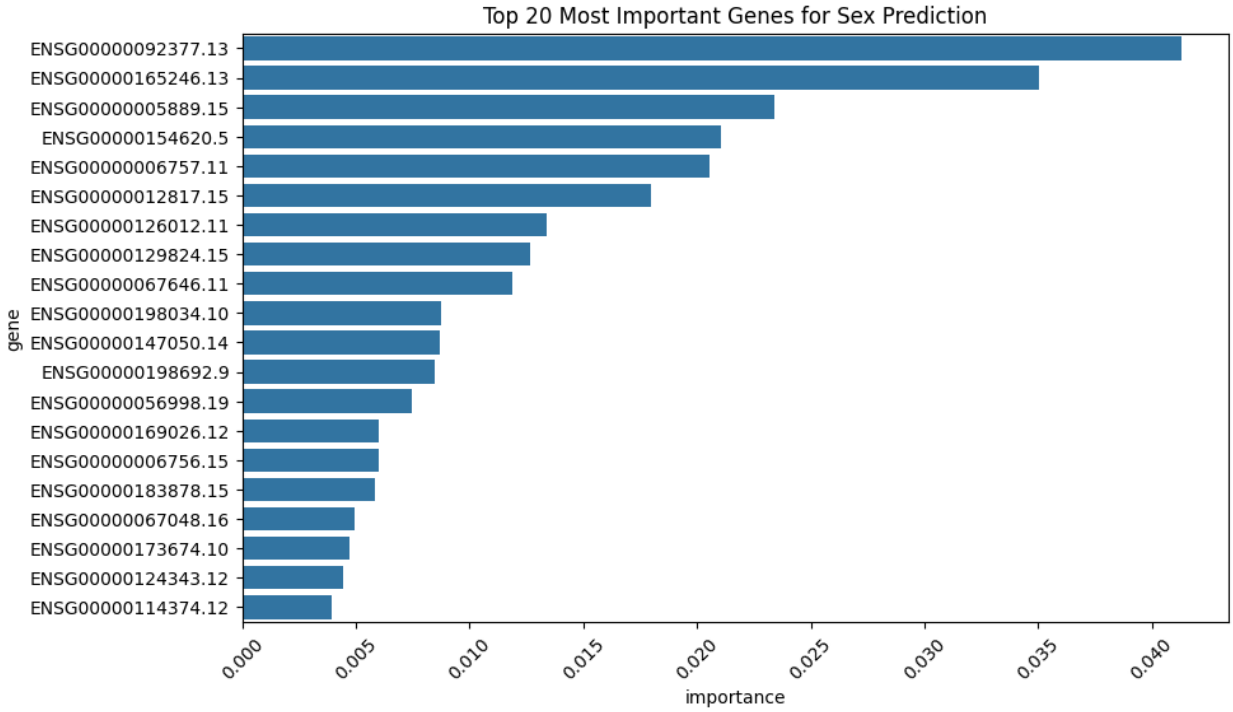



```
Results for Nerve - Tibial:
Classification Report:
              precision    recall  f1-score   support

           1       0.71      1.00      0.83       126
           2       1.00      0.15      0.26        60

    accuracy                           0.73       186
   macro avg       0.86      0.57      0.55       186
weighted avg       0.80      0.73      0.65       186

Confusion Matrix:
[[126   0]
 [ 51   9]]
Cross-Validation Scores:
[0.69354839 0.68548387 0.71774194 0.68548387 0.7398374 ]
```

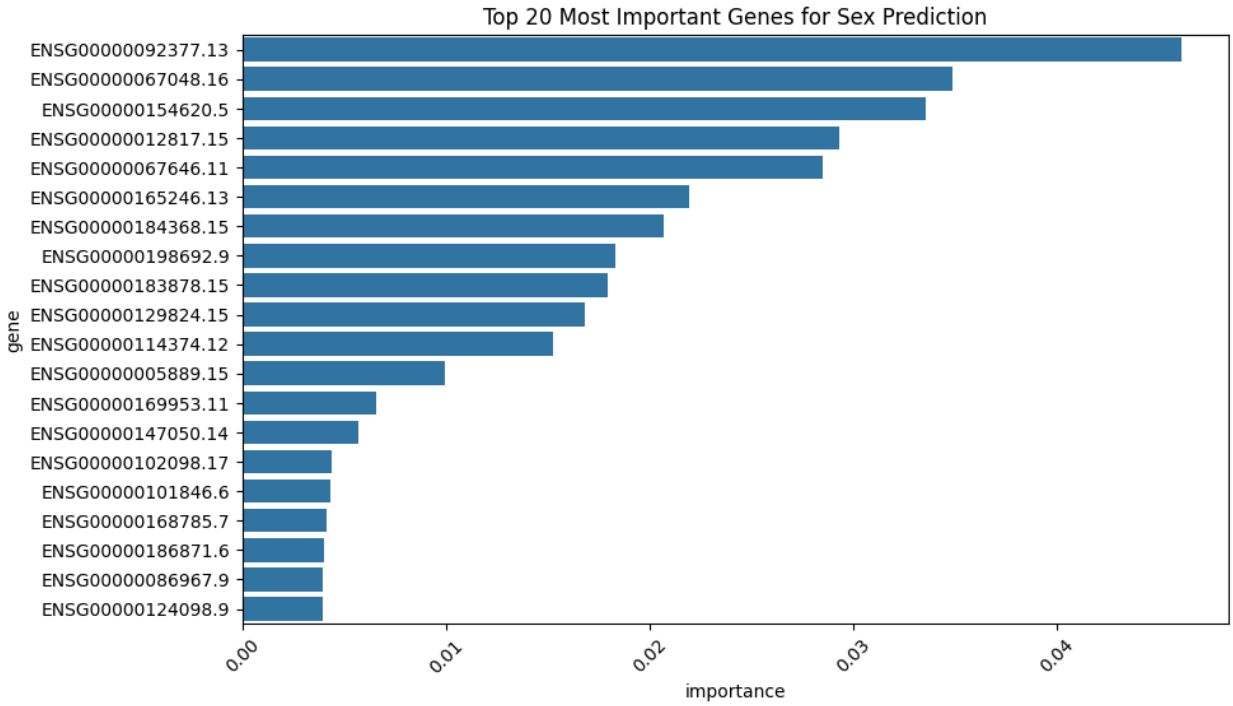



```
Results for Liver:
Classification Report:
              precision    recall  f1-score   support

           1       0.87      1.00      0.93        48
           2       1.00      0.65      0.79        20

    accuracy                           0.90        68
   macro avg       0.94      0.82      0.86        68
weighted avg       0.91      0.90      0.89        68

Confusion Matrix:
[[48  0]
 [ 7 13]]
Cross-Validation Scores:
[0.93478261 1.         1.         0.88888889 0.95555556]
```




```
Results for Muscle - Skeletal:
Classification Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00       163
           2       1.00      1.00      1.00        78

    accuracy                           1.00       241
   macro avg       1.00      1.00      1.00       241
weighted avg       1.00      1.00      1.00       241

Confusion Matrix:
[[163   0]
 [  0  78]]
Cross-Validation Scores:
[1. 1. 1. 1. 1.]
```

Conclusions

This study underscores the profound impact of gender on biological pathways, revealing both systemic and tissue-specific mechanisms that shape health, disease susceptibility, and physiological differences between males and females. By analyzing sex-biased gene expression across tissues and integrating the results with KEGG pathway enrichment analysis, we identified significant insights into how gender-specific factors influence fundamental biological processes.

The findings demonstrate a clear link between gender and susceptibility to infections and immune regulation. For instance, the enrichment of the Herpes Simplex Virus 1 (HSV-1) Infection pathway across liver, visceral adipose, and skeletal muscle tissues underscores the sex-based differences in immune responses. Males tend to experience worse outcomes and more frequent reactivation of HSV-1 due to less robust immune regulation, often influenced by testosterone's suppressive effects on key immune mediators like interferon-gamma. In contrast, females exhibit stronger immune responses, potentially due to estrogen's enhancing effects on immune cell activity. However, this heightened immune function in females comes with an increased predisposition to autoimmune diseases, as reflected by the enrichment of pathways such as Complement and Coagulation Cascades, which are frequently implicated in conditions like rheumatoid arthritis.

Metabolic regulation emerged as a key theme, with significant enrichment of pathways like Insulin Signaling and AMPK Signaling in skeletal muscle and adipose tissues. These pathways highlight the gender differences in energy regulation and glucose metabolism, with females often exhibiting greater insulin sensitivity, which provides a protective effect against Type 2 diabetes and related metabolic disorders. However, the Type I Diabetes Mellitus pathway, enriched in adipose tissue, points to a complex interplay between immune function and metabolic regulation, potentially explaining the higher prevalence of autoimmune diabetes in females. Hormonal fluctuations during life stages such as puberty, pregnancy, and menopause further modulate these pathways, influencing long-term metabolic outcomes in females.

Cancer-related pathways, including Renal Cell Carcinoma, Non-Small Cell Lung Cancer, and Prostate Cancer, were consistently enriched across tissues, reflecting the strong influence of gender on cancer biology. Males exhibit a higher prevalence of aggressive cancers, which may be attributed to higher oxidative stress levels, differences in cell proliferation rates, and hormonal effects. Testosterone, for example, has been shown to promote tumor growth in certain cancers, whereas estrogen provides a protective effect in some tissues but may contribute to hormone-sensitive cancers, such as breast and endometrial cancers in females. The findings also highlight how sex hormones act as both protectors and contributors depending on the cancer type and tissue context.

Pathways such as Ubiquitin-Mediated Proteolysis and Autophagy, enriched in liver and skeletal muscle, shed light on gender differences in cellular stress responses and maintenance. Males often exhibit higher oxidative stress and inflammation, leading to greater reliance on proteolysis for cellular repair. In contrast, females tend to maintain more robust autophagic processes, which enhance cellular repair and provide protection against age-related degenerative diseases. These differences are likely modulated by hormonal regulation, contributing to disparities in conditions such as neurodegenerative diseases and sarcopenia between males and females.

The integration of a machine learning model further strengthened these findings. The Random Forest model, trained on gene expression data, achieved 100% accuracy in predicting gender for tissues such as adipose-visceral and muscle-skeletal, indicating that these tissues exhibit clear, measurable differences between males and females. In contrast, the model showed lower accuracy for nerve-tibial (73%) and liver (90%), suggesting that gene expression differences in these tissues may be less pronounced or influenced by overlapping biological factors. This variability highlights how sex-biased processes may be more distinct in certain tissues, such as those involved in metabolism and immune regulation, while other tissues may reflect shared or less divergent biological mechanisms.

The recurrence of pathways such as Wnt Signaling, Inositol Phosphate Metabolism, and ErbB Signaling across multiple tissues suggests systemic coordination of sex-biased biological processes. These pathways regulate critical functions like cell communication,

differentiation, and survival, and their repeated enrichment implies that hormonal and genetic factors orchestrate widespread physiological responses. For instance, the Wnt signaling pathway's role in tissue repair and immune modulation highlights its relevance to both systemic and localized processes influenced by gender.

In summary, this study reveals a complex interplay of immune, metabolic, and regulatory pathways, demonstrating how gender-specific differences extend across tissues. The results from the enrichment analysis and the Random Forest model provide compelling evidence of systemic and tissue-specific distinctions in sex-biased gene expression. Tissues where the model achieved perfect accuracy likely reflect profound gender differences, whereas tissues with lower accuracy may represent shared or overlapping biological processes. These findings emphasize the importance of incorporating gender-based analyses into research and developing personalized therapeutic strategies that account for these differences, ultimately advancing the field of precision medicine.

Project Reflection

This project provided a unique opportunity to explore gender-based differences in gene expression and their impact on biological pathways. While the process was complex and often demanding, we successfully navigated challenges, made critical decisions, and met the goals we set out to achieve.

One of the key achievements of this project was meeting our primary objective: identifying and analyzing sex-biased biological pathways across multiple tissues. Through careful integration of bioinformatics tools, statistical methods, and enrichment analysis, we uncovered significant pathways and provided meaningful insights into how gender influences health and disease. This achievement reflects not only the technical rigor of our work but also our ability to synthesize data into actionable biological conclusions. Completing this ambitious task reaffirmed the value of interdisciplinary approaches and the potential of bioinformatics in addressing complex biological questions.

Handling challenges was a core component of our journey, as obstacles emerged at nearly every stage of the project. Data handling and processing were among the most significant challenges, particularly given the large size and complexity of the datasets. Dividing the data into manageable chunks allowed us to process iteratively, but this approach required careful decision-making to ensure consistency and accuracy. Another challenge was the quality of the data itself, with missing values, batch effects, and outliers posing threats to the reliability of our analyses. Addressing these issues required us to weigh different preprocessing techniques and select methods that balanced preserving biological variability while ensuring data integrity. These challenges taught us the importance of adaptability and critical thinking, as each decision had a direct impact on the validity of our results.

Evaluating our approach revealed areas of strength as well as potential improvements for future work. The pipeline we developed—incorporating data preprocessing, statistical analysis, and pathway enrichment—proved effective in meeting our research goals. Additionally, automating parts of the pipeline further could reduce manual

intervention, streamline the process, and minimize the risk of errors. These reflections highlight the iterative nature of research and the importance of continually refining methodologies to improve outcomes.

Meeting project metrics

From the outset, we established clear objectives, including processing large-scale datasets, identifying significant pathways, and generating interpretable visualizations. By systematically working through these goals, we ensured steady progress and maintained alignment with the project's broader aims. Metrics such as the number of tissues analyzed, pathways identified, and results validated provided concrete measures of our progress and success. The visualizations we created, summarizing key pathways and their significance, exemplify how our work not only met but exceeded expectations in terms of clarity and accessibility.

In conclusion, this project has been both challenging and rewarding, offering invaluable lessons in data analysis, decision-making, and research methodology. Our ability to overcome obstacles, evaluate our approach critically, and achieve our goals reflects the strength of our team and the processes we employed. Moving forward, the insights and skills gained from this project will serve as a strong foundation for tackling future interdisciplinary challenges, reinforcing our commitment to advancing knowledge in the field of gender-based biology.

References

1. Bryan, R. T., Evans, T., Dunn, J. A., Iqbal, G., Bathers, S., Collins, S. I., James, N. D., Catto, J. W. F., & Wallace, D. M. A. (2015). A comparative analysis of the influence of gender, pathway delays, and risk factor exposures on the long-term outcomes of bladder cancer. *European Urology Focus*, 1(1), 82–89. https://doi.org/10.1016/j.euf.2015.01.001

2. Chen, C., Hu, L.-X., Dong, T., Wang, G.-Q., Wang, L.-H., Zhou, X.-P., Jiang, Y., Murao, K., Lu, S.-Q., Chen, J.-W., & Zhang, G.-X. (2013). Apoptosis and autophagy contribute to gender difference in cardiac ischemia–reperfusion induced injury in rats. *Life Sciences*, 93(7), 265–270. https://doi.org/10.1016/j.lfs.2013.06.019

3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

4. Lee, K. (2023). Differences between the sexes in rheumatic disease. *Rheumatology.org*. https://rheumatology.org/patient-blog/differences-between-the-sexes-in-rheumatic-disease

5. Dohmen, K., Shigematsu, H., Irie, K., & Ishibashi, H. (2003). Longer survival in female than male with hepatocellular carcinoma. *Journal of Gastroenterology and Hepatology*, 18(3), 267–272. https://doi.org/10.1046/j.1440-1746.2003.02936.x

6. Han, X., Lundberg, P., Tanamachi, B., Openshaw, H., Longmate, J., & Cantin, E. (2001). Gender influences herpes simplex virus type 1 infection in normal and gamma interferon-mutant mice. *Journal of Virology*, 75(6), 3048–3052. https://doi.org/10.1128/JVI.75.6.3048-3052.2001

7. Kim, T. K. (2015). KoreaMed synapse. *Korean Journal of Anesthesiology*, 68(6), 540–546. https://doi.org/10.4097/kjae.2015.68.6.540

8. Lista, P., Straface, E., Brunelleschi, S., Franconi, F., & Malorni, W. (2011). On the role of autophagy in human diseases: A gender perspective. *Journal of Cellular and Molecular Medicine*, 15(7), 1443–1457. https://doi.org/10.1111/j.1582-4934.2011.01293.x

9. Peer, V., Schwartz, N., & Green, M. S. (2021). Sex differences in salmonellosis incidence rates—an eight-country national data-pooled analysis. *Journal of Clinical Medicine*, *10*(24). https://doi.org/10.3390/jcm10245767

10. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

11. Tower, J., Pomatto, L. C. D., & Davies, K. J. A. (2020). Sex differences in the response to oxidative and proteolytic stress. *Redox Biology*, *31*, 101488. https://doi.org/10.1016/j.redox.2020.101488

Tools

https://maayanlab.cloud/Enrichr/

https://chatgpt.com/

https://colab.google/

https://copilot.microsoft.com/

https://code.visualstudio.com/

User's Guide

Overview

This guide provides step-by-step instructions for running the provided code to analyse sex-biased gene expression across multiple tissues and visualize biological pathways using statistical and machine learning methods.

Operating Instructions

1. **Setting Up Your Environment**
    a. **Install Python:** Ensure Python 3.x is installed on your system. Download it from Python's website.
    b. **Required Libraries:** Install the following libraries using pip: pip install pandas scipy tqdm matplotlib seaborn scikit-learn
    c. **Colab Integration:** The code is designed for use in Google Colab, which provides a cloud environment for Python programming.
2. **Preparing Your Data**
    a. **File Structure:** Ensure your GTEx data files are organized as follows:

```
final_project_B/Data/
├── Adipose_Visceral
│   ├── mat.f.coding.Adipose - Visceral (Omentum).csv
│   └── pheno.f.Adipose - Visceral (Omentum).csv
├── Muscle_Skeletal
│   ├── mat.f.coding.Muscle - Skeletal.csv
│   └── pheno.f.Muscle - Skeletal.csv
├── Nerve_Tibial
│   ├── mat.f.coding.Nerve - Tibial.csv
│   └── pheno.f.Nerve - Tibial.csv
└── Liver
    ├── mat.f.coding.Liver.csv
```

└── pheno.f.Liver.csv

b. **Metadata:** The gene.f.csv file should contain the mapping of gene names to their descriptions.

3. Running the Code

a. **Mount Google Drive:** Ensure your data is in Google Drive and run the following:
from google.colab import drive
drive.mount('/content/drive')

b. **Gene Expression Analysis:**
   i. The script iterates through each tissue, performs t-tests to identify sex-biased genes, and saves the significant results in tissue-specific CSV files.
   ii. Outputs are stored in: /content/drive/My Drive/final_project_B/Data/diff_genes_<Tissue>.csv.

c. **KEGG Pathway Visualization:**
   i. Pathway visualization is generated for each tissue based on KEGG data files.
   ii. Charts highlight pathways with significant enrichment.

d. **Machine Learning:**
   i. A Random Forest Classifier predicts gender based on gene expression data.
   ii. Results include:
      1. Feature importance charts for top genes.
      2. Classification reports.
      3. Confusion matrices.
      4. Saved feature importance files for each tissue.

4. Visualizations

a. Pathway and gene-level insights are visualized as bar charts, saved as PNG files or shown interactively in Colab.

5. Output

a. **CSV Files:** Containing significant genes for each tissue.

b. **Charts:** Visual representation of KEGG pathways and top genes contributing to gender classification.

Maintenance Guide

Objective: Guide for maintaining, updating, and troubleshooting the codebase to ensure continued functionality.

System Requirements

- **Software:**
  - Python 3.x
  - Libraries: pandas, scipy, tqdm, matplotlib, seaborn, scikit-learn.
- **Hardware:**
  - 8 GB RAM (minimum recommended).
  - Google Colab or equivalent computing environment.

Maintenance Tasks

1. **Updating Libraries:** Periodically update Python libraries to their latest versions to ensure compatibility:
   pip install --upgrade pandas scipy tqdm matplotlib seaborn scikit-learn
2. **Modifying the Code:**
   a. **Adding Tissues**: To analyze new tissues, add their names and folder paths to organ_names and folder_names.
   b. **Changing Statistical Thresholds:** Modify the p-value threshold in the t-test logic for different sensitivity levels:
   diff_genes_df = res_df[res_df['p_value'] < 0.02].sort_values('p_value')
3. **Pathway Analysis Updates:**
   a. Adjust visualizations by editing the pathway filtering logic.
4. **Random Forest Model:**
   a. Experiment with model parameters like n_estimators or random_state in the RandomForestClassifier.

Troubleshooting

- **Memory Errors:** If Colab crashes, consider splitting the dataset

into smaller chunks or upgrading the environment to Colab Pro.

- **Missing Outputs:**
    - Check if file paths are correctly defined.
    - Ensure all required files exist in the specified directory.

Backup and Recovery

- **Backup:** Save processed data and generated CSV files to Google Drive or GIT.
- **Recovery:** Keep a backup of the code and data files. Restore from the backup in case of accidental deletions.