Leveraging Machine Learning Methods to Identify Target Genes of the Wnt Signalling Pathway

Cassandra Drazick, 1116756

BINF6999 – Bioinformatics Masters Project

Advisors: Dr. Andrew Hamilton-Wright, Dr. Terry Van Raay

Submitted Tuesday, August 13th, 2024

Table of Contents Page

Abstract 3

Introduction 3

Methods 10

Results 16

Discussion 22

References 27

Abstract

Understanding gene expression and signal transduction pathways is critical for unravelling cellular functions and disease mechanisms. This study addresses the challenges of identifying key target genes within signal pathways. I have focused on the Wnt signaling pathway, a deeply conserved and vital pathway involved in cell differentiation, proliferation and previously shown to be connected to a myriad of diseases, including many cancers. Previous methods for target gene identification have relied heavily on literature review and database usage to facilitate “next step” experimental design. These methods can be time consuming and may not capture a wide range of potential target genes. I have developed a bioinformatics pipeline that integrates RNA-seq data with machine learning techniques SVM, PCA, Random Forests and Neural Networks to enhance the identification of target genes across multiple datasets with varied experimental conditions. Results demonstrated that the integration of multiple machine learning models extracts a diverse idea of Wnt pathway target gene scope, potentially revealing genes that have not yet been identified for further study. Heatmaps generated to interpret gene patterning between machine learning models illustrated gene patterns across models, highlighting the ability of my pipeline to uncover both well-known and previously unidentified genes. This approach not only streamlines target gene identification but also broadens the scope of potential targets for experimental study, offering new insights into the Wnt signaling pathway and its associated diseases.

Introduction

The relationship between gene expression and signaling transduction pathway activity is bidirectional and crucial for maintaining cellular function. Signaling transduction pathways are intricate networks of molecular interactions that play crucial roles in cellular communication and function (Van De Stolpe et al., 2019). These pathways enable cells to respond to external signals, such as genetic, biochemical or environmental changes and translate them into explicit cellular responses. These pathways typically involve a series of sequential steps and consist of various components, with each factor possessing a specific role and function (Handy et al., 2016). These factors are gene expression products, such as transcription factors or enzymes, and the levels of these products affect the pathway’s ability to propagate and execute signals. In cases where a gene that is connected to a specific pathway’s activity is mutated or dysregulated, the impact on the signaling pathway can be profound and multifaceted. This can lead to disruption or alteration of pathway function, which is often linked to the formation of many diseases, including genetic disorders, cancers and developmental abnormalities (Van De Stolpe et al., 2019). Pathways are also often connected with others in complex systems, termed “cross talk”, and can be affected directly by dysregulation in other pathways in which they are connected to (Handy et al., 2016).

Understanding the intricacies of a particular pathway is very challenging, due to the dynamic interplay among its components. This can make it difficult for researchers to pinpoint which gene feature of a signal transduction pathway to focus their research on. Traditional methods for understanding key gene targets have previously relied heavily on the use of relevant literature and biomedical databases (Rosati et al., 2024). While these approaches provide valuable information, they can be time-consuming and may not always capture the dynamic changes occurring within pathways. This project aims to bridge this gap by utilizing RNA sequencing (RNA-seq) data and bioinformatics techniques to facilitate the identification of key target genes of a signal transduction pathway. RNA-seq is a high-throughput sequencing method used to quantify transcriptomes and is a technique that provides a broad understanding of gene expression under various experimental condition (Wang et al., 2010). Once RNA-seq has been performed, it generates extensive data that can reveal genes that are differentially expressed across conditions, providing insights into their roles within signaling pathways (Wang et al., 2010).

To enhance the identification and characterization of target genes, my project employs an originally designed pipeline that integrates RNA-seq data analysis with extensive feature extraction and validation, enabling a more precise and efficient discovery of key target genes. Several machine learning techniques are utilized within this pipeline: Support Vector Machine (SVM), Random Forests, Neural Networks and PCA dimensionality reduction. As a baseline figure, Student T-tests were also taken as a comparison for the output of machine learning models. These machine learning techniques utilize diverse methods for feature selection, increasing robustness of genes that are identified across models.

The Wnt Signaling Pathway

The Wnt signaling pathway is a signal transduction pathway that is highly conserved amongst species and regulates various aspects of cell behaviour, including differentiation, proliferation and migration (Liu et al., 2024). It is particularly important in developmental processes such as cell fate determination, tissue patterning and organogenesis (Liu et al., 2024). This pathway regulates the cytoplasmic levels of β-catenin, a protein crucial for coordinating gene transcription and cell-cell adhesion (Bell et al., 2024). It is typically described in two states: the OFF and the ON states. In the OFF state, a destruction complex targets β-catenin, leading to its degradation and preventing its accumulation in the cytoplasm. In the ON state, Wnt proteins bind to a receptor complex composed of Frizzled receptors, activating a cytoplasmic protein called Dishevelled which then transduces the signal into the cell, inhibiting the destruction complex (Liu et al., 2024). This inhibition causes β-catenin to amass within the cytoplasm, growing in levels until it is translocated into the nucleus where it interacts with TCF/LEF1 (T-cell factor/lymphoid enhancer-binding factor 1) (Liu et al., 2024). These are transcription factors that then activate target gene expression.

Due to its connection to various diseases, the Wnt pathway is well-studied. However, the complexity of its interactions and regulation continues to be an area of active research. Dysregulation in the Wnt pathway is connected to many diseases such as cancer, bone disorders, neurodegenerative diseases, developmental disorders, and possibly others that have yet to be uncovered (Zhan et al., 2016). The ability to accelerate the identification of target genes of this pathway would facilitate research that may lead to the development or discovery of novel treatments for various conditions caused by aberrant signaling (Sarker, 2021). This is the reason why I selected the Wnt pathway as the signal transduction pathway to use to construct and test my target gene identification pipeline. Datasets used to train the machine learning models are all datasets that deal with the induction of aberrant Wnt pathway signaling.

Datasets

Each dataset used in the pipeline construction and analysis features treated samples designed to induce aberrations in the Wnt pathway. These treatments encompass both molecular interventions and genetic modifications, such as knockouts and knockdowns. Incorporating a diverse range of experimental conditions was crucial for capturing the broad spectrum of pathway dysregulation and for evaluating the robustness of both model and overall pipeline performance. By including comparative opportunities between conditions that induce similar effects and those with opposing effects, this analysis gains deeper insights into genes that are regulated by the Wnt signaling pathway. In total, six datasets were selected to facilitate the identification of target genes and to validate the efficacy of the top genes pipeline.

CHIR99021 Dataset

CHIR99021 is a small molecule that has previously been found to be an effective and safe Wnt pathway activator (Wang et al., 2022). This small molecule acts by inhibiting GSK-3β, a vital component of the β-catenin destruction complex, allowing β-catenin to accumulate and translocate into the nucleus. Under normal physiological conditions, GSK-3β phosphorylates β-catenin, leading to its ubiquitination and subsequent degradation in the proteasome (Wang et al., 2022). The specificity of CHIR99021 for GSK-3β allows for targeted manipulation of the Wnt pathway without significantly affecting other signaling pathways (Wang et al., 2022). This makes it a valuable tool for studying Wnt signaling dynamics and its implications in development and disease.

WNT3A Dataset

WNT3A is a member of the Wnt family of signaling proteins, a secreted glycoprotein that functions as a key ligand and binds to Frizzled receptors, disrupting the destruction complex and the degradation of β-catenin, allowing levels to increase within the cytoplasm. This protein is often used in analyses to induce activation of the Wnt pathway to study effects of the pathway on various cellular processes, much like this study (Kaur et al., 2013). It is often used in stem cell, developmental and cancer research (He et al., 2015).

Nkd1and Axin2 Knockout Dataset

Nkd1 and Axin2 are regulators of the negative feedback loop that controls the overall function of the Wnt pathway (Bell et al., 2024). This negative feedback mechanism is a mode of self-regulation that uses the activation of the Wnt pathway transcription factors to target genes that code for Wnt deactivating proteins. Axin2 acts upon this mechanism as a β-catenin destabilizer, while Nkd1inhibits translocation of β-catenin into the nucleus. Knocking out both genes results in constitutive activation of the Wnt pathway and thus target gene up or downregulation (Bell et al., 2024).

C59 Dataset

C59 is a small molecule known for its role as an inhibitor of the Wnt signaling pathway. It specifically inhibits the enzyme Porcupine (PORCN), an enzyme essential for the secretion and processing of Wnt ligand activity (Motono et al., 2016). Through this action, ligands do not bind to the Frizzled receptor and thus the β-catenin destruction complex will remain to reduce β-catenin levels within the cytoplasm (Mukherjee et al., 2022). The C59 dataset serves as the only pathway inhibitor dataset within my study and is a critical component for comparing the effects of Wnt pathway inhibition against the stimulating activity of the other data set treatment conditions.

CEBPA and FOXA1 Datasets

CEBPA and FOXA1 are two transcription factors currently under investigation for their regulation by the Wnt pathway. It has been observed that their expression is suppressed by the β-catenin/TCF complex in liver cancer cells (Nakagawa et al., 2024). Emerging evidence suggests a compensatory mechanism at play: the suppression of CEBPA and FOXA1 by the Wnt pathway leads to an enhancement of other connected factors, and vice versa. This dataset includes gene expression data from experiments involving genetic knockdowns of CEBPA and FOXA1(Zhang et al., 2021). These knockdowns trigger the hyperactivation of the Wnt pathway, providing insights into the subsequent changes in gene expression levels of connected factors that occur with this activation (Nagakawa et al., 2024).

Feature Selection Using Machine Learning

Machine learning is a subsection of artificial intelligence that involves the development of algorithms and statistical models enabling computers to learn and make decisions from data without being explicitly programmed (Sarker, 2021) Large datasets allow machine learning models to identify patterns and make predictions while improving their performance over time through iterative training. Machine learning techniques can process complex, high-dimensional biological datasets, such as RNA-seq outputs to reveal significant relationships and features that may be challenging to detect using traditional analytic methods (Rosati et al., 2024). The machine learning methods selected for this project each offer diverse approaches to feature selection that could increase the validity of target genes identified across different models.

Support Vector Machines

Support vector machines are a class of supervised learning algorithms that are primarily used for classifications tasks. They aim to find the optimal decision boundary that best separates data points of different classes in a feature space. In feature selection, features (genes) are assigned weight coefficients that indicate their importance. Features with larger coefficients are considered more influential in the classification process (Alcaraz et al., 2022).

Random Forest

Random forest is an ensemble learning method used for classification tasks. It performs classification by building multiple decision trees and combining their outputs to improve accuracy and control overfitting of data. Decision trees are tree-like data structures that split data based on feature values to create “branches” and “leaves” of the trees. Important features are selected by averaging the decrease in impurity across all trees in the forest. Features that lead to a more significant reduction in impurity are marked as more important (Sarker, 2021).

Neural Network

Neural networks are a machine learning model that are structured to imitate the function of the human brain and are a powerful took to use for complex pattern recognition. Feature importance is inferred from network weights and activations. During model training, the network adjusts weights to minimize the loss function. Features that are more relevant for prediction tend to have larger weights and are deemed more important features (Huang et al., 2020).

Principal Component Analysis

PCA reduces the number of features while also preserving variance in the data. It identifies principal components, which are linear combinations of features) that capture the most variance and can highlight key features driving variance. By projecting data onto principal components, PCA aids in understanding underlying structure and selecting features that significantly contribute to the highest variance components (Schenone et al., 2017).

Methods

Data Acquisition

Datasets for analysis were selected through a comprehensive literature review, ensuring they met specific criteria. The review utilized NCBI’s search tools to identify studies involving alterations to Wnt pathway activity. The criteria for choosing datasets were as follows:

1. Cell Treatment: The study involved treatment of cells with either a Wnt pathway activator or inhibitor, achieved through molecular or genetic methods.
2. RNA Sequencing: The transcriptome analysis was performed using RNA sequencing.
3. Availability of Raw Data: Raw count data were accessible through databases such as ArrayExpress, NCBI’s Gene Expression Omnibus (GEO), or the Sequence Read Archive (SRA).

Data Preprocessing

CHIR99021 Treatment Dataset

The dataset, originally provided in a raw-counts matrix format, was uploaded into R Studio. This matrix comprised six samples: three treated with CHIR99021 and sequenced after twenty-four hours, and three control samples from an hESC cell line. Initial data cleaning involved removing any missing values or rows with only zero counts. The matrix contained 61,552 rows, each representing a unique gene identifier. Differential gene expression analysis was then conducted, followed by normalization of the data using counts per million (CPM) with the ‘EdgeR’ package. Subsequently, gene ontology analysis was performed to convert Ensembl gene IDs to human gene names, utilizing the gprofiler2 package in R. The final data frame was saved as a tab separated values (TSV) file for further analysis.

*nkd1;axin2* Double Knockout Dataset

The dataset was derived from raw FASTQ files obtained from RNA sequencing. Eight raw data files were retrieved, with four files representing treatment replicates from ***nkd1*** and ***axin2*** knockout samples and four files representing control replicates from the “Tubingen” line. A reference genome in FASTA format and an annotation file in GTF format were downloaded from Ensembl via FTP, specifically from assembly **GRCz11.112**, which is the latest version of the *Danio rerio* genome. These files were used to create an index file for alignment using the ‘STAR’ module. The alignment process resulted in eight BAM files. The BAM files were then imported into R, where the annotation file was used again to perform feature counting with the ‘**Rsubread’** package.

The resulting counts matrix included data from each of the eight samples and contained 26,985 rows representing unique gene identifiers. Subsequently, the data was cleaned by removing missing values and rows with only zero counts.

Differential expression analysis was performed using the ‘**DGElist’** package and counts per million normalization was applied. Gene ontology analysis was conducted using the ‘**biomaRt’** package, accessing both *Danio rerio* and *Homo sapien* marts. *Danio rerio* Ensembl gene codes were converted into *Homo sapiens* gene names, with duplicate genes aggregated by taking the mean. This final processing left 12,969 rows containing human gene IDs and gene expression data. The resulting data frame was saved as a TSV file for further analysis.

WNT3A and C59 Treatments Dataset

The dataset was derived from raw FASTQ files obtained from RNA sequencing. Fifteen raw data files were retrieved, categorized as follows: Three control replicates from embryonic stem cells, three early-stage treatment replicates with WNT3A, three late-stage treatment replicates with WNT3A, three early-stage treatment replicates with C59, three late-stage treatment replicates with C59.

*Figure 1. Workflow of bioinformatic pipeline to identify target genes of the Wnt signaling pathway. Outlining of each step from data acquisition, preprocessing, pipeline steps, and final data output of the pipeline utilized in this study.*

A reference genome and an annotation file were downloaded in FASTA and GTF formats from Ensembl via FTP. The assembly used for both files was GRCh38.112, the latest version of the *Homo sapien* genome. These files were used to create an index file with the ‘STAR’ module. The index file was utilized with ‘STAR’ for aligning the FASTQ files, resulting in BAM file outputs. The BAM files were processed using the ‘featureCounts’ tool within the ‘Rsubread’ package in R.

Data was then organized into separate data frames for each treatment method—WNT3A and C59—along with control samples. The resulting data frames contained: WNT3A Data Frame: Nine samples and 40,967 rows representing gene IDs and for the C59 Data Frame: Nine samples and 42,623 rows representing gene IDs.Data cleaning involved removing rows with all zero values. Gene ontology analysis was performed using the ‘gprofiler2’ package exchanging *Homo sapien* Ensembl gene codes into gene names. After final processing, the data frames included: 27,466 rows in the WNT3A data frame and 28,450 rows in the C59 data frame. The resulting data frames were saved as a TSV file for further analysis.

*CEBPA* and *FOXA1* Knockdown Dataset

RNA sequencing data were retrieved from fifteen raw files in the “.genes.results” format. These files included three control replicates from the HuH-7 cell line, two sets of samples with three replicates each from CEBPA knockdown, and two sets of samples with three replicates each from FOXA1 knockdown. The data were uploaded into R, and the ‘purrr’ package was used to combine these files into a single data frame. The combined data frame initially contained 60,617 rows, representing gene IDs, and encompassed all fifteen samples. To ensure data quality, rows containing only zeros were removed from the data frame. Differential gene expression analysis was subsequently conducted using the ‘edgeR’ package and counts per million (CPM) were calculated for normalization.

Following normalization, the dataset was split into two separate data frames. The first data frame contained the three HuH-7 control replicates alongside all CEBPA knockdown samples, while the second data frame included the HuH-7 control replicates and all FOXA1 knockdown samples. Gene ID row names were originally encoded with *Homo sapien* Ensembl gene codes and gene names but were modified to display only gene names for clarity. To address any potential issues with duplicate gene names, duplicates were aggregated by calculating the mean of their respective values. The final datasets consisted of 28,945 rows for the CEBPA knockdown data frame and 29,172 rows for the FOXA1 knockdown data frame.

Pipeline

Before analysis could begin, “conditions” files had to be generated for each dataset. These files were built as ‘.sh’ shell files containing labels for each sample, whether control or treatment. The first step of the pipeline uses supervised machine learning models SVM, Neural Network and Random Forest, principal component analysis (PCA) dimension reduction and T-tests to generate important features of the top 50 and top 2000 genes among each dataset. To begin the analysis, gene expression TSVs and conditions files were imported using the ‘pandas’ library. The gene expression datasets were read into a data frame with gene names as row indices and the conditions file was loaded into another data frame. Next, the data was organized to match sample labels. Separate lists were created for treated and control samples based on the conditions files, ensuring data was aligned with experimental groups.

Dimension reduction was conducted by PCA using the ‘sci-kit learn’ library’s ‘PCA’ function. The resulting loadings from PCA were extracted to compute absolute values, which were then summed to assess overall importance of each gene entry. The top genes, based on importance scores were identified and saved and the absolute loadings for each gene were plotted. In parallel, the three machine learning models were being employed to analyze the data. These models were validated by leave one out cross-validation. Performance metrics including accuracy, precision, recall and F1 scores were computed for each model for model evaluation and deposited into a comma separated values (CSV) file. To accompany the machine learning models, T-tests were performed as a statistical complement to the trained models. For T-tests, top genes were generated using Z-scores to assess the standard deviations from the means. To address the issue of multiple comparisons, p-values were adjusted using the Benjamini-Hochberg method. Important feature selection was performed for each model, dimension reduction and T-tests and placed into a separate CSV file showing the top genes discovered for each method.

For the next step of the pipeline, CSV files containing top genes lists from each dataset were utilized to find common occurrences of genes among models and datasets. Combined top gene lists were loaded into a ‘pandas’ data frame and transformed using the ‘melt’ function to reorganize the data and gather all gene entries into a single column. Any entries with missing values in the gene column were removed and occurrences of each gene across all models were then counted. A summary data frame was constructed that lists each gene alongside its frequency of appearance across top gene lists, genes that appear only once were filtered out to focus on those with more significant presence. For each model, the script created a dictionary to store the rank of each gene. It processed the data frame to capture and assign ranks to genes within each model’s top list. This ranking information was then used to compile a summary of where each gene appeared across the different models. A comprehensive rank summary was constructed by iterating through each gene in the filtered gene summary data frame. For each gene, the script compiled a list of models and ranks where the gene appeared. This gene summary was saved to a CSV file.

Results

After processing datasets through my pipeline, I generated comprehensive top genes lists for each dataset. These lists were categorized into two groups: top 50 and top 2000 features, and the results were systematically organized into CSV outputs. For each dataset, I assessed the top genes using several methodologies:

Machine Learning Models

I ranked genes based on their importance levels, which were determined by specific machine learning model used.

Principal Component Analysis (PCA):

I ranked genes according to their absolute loading values on the principal component (PC1), reflecting each gene’s contribution to the variance captured by PC1.

*A table of data with numbers

Description automatically generated with medium confidence*T-tests

*Figure 2. Top 20 genes across models from the CHIR99021 treated dataset. This list shows the output that was generated directly from the first step of my pipeline. Each of the six treated datasets (CHIR99021, CEBPA, C59, FOXA1, axin2;nkd1 and WNT3A) had the same output showing top genes generated. Methods for identification of top features are listed as column labels, and rank is shown as row labels. Rank is the level of importance of each gene among each method of discovery.*

I selected genes based on corrected p-values and ranked them according to their z-scores which represent deviance from the mean variance and representing their statistical significance.

Top Genes ListFor the CHIR99021 dataset, Figure 2 showcases the top 20 genes identified for each methodology employed. This figure illustrates the relative ranking of each gene, alongside their names, highlighting how different techniques prioritize genes differently.

Principal Component Analysis (PCA)

A graph of a number of blue lines

Description automatically generated with medium confidenceMy PCA results produced a bar plot representing the absolute loading values of the top genes (Figure 3). This plot provides a visual summary of how each gene influences the principal component. This absolute loading values indicate the extent of which each gene contributes to the variance captured by PC1, with higher values signifying greater influence.

***Figure 3. Absolute loading values of top identified genes using dimensionality reduction by PCA.*** *The bar plot displays the absolute loading values for the top genes identified through Principal Component Analysis (PCA). Each blue bar represents the magnitude of the absolute loading value for a specific gene on Principal Component 1 (PC1).*

T-test Analysis

A table of numbers and symbols

Description automatically generatedT-test result output provided detailed insights into the top genes across the datasets. For each gene in the top 50 and top 2000 lists, several important metrics were calculated. P-values and corrected P-values were obtained to verify statistical significance, and top genes were ranked according to these metrics. P-values were corrected by the Benjamini-Hochberg procedure was employed to correct the P-value. This step is crucial for minimizing the risk of false positive results. Additionally, t-statistic was calculated, representing the variance of the means between both sample conditions. The Z-score was included, which shows a standardized version of the t-statistic. Figure 4 illustrates a summary table of these metrics for the top genes identified in the CHIR99021 dataset.

*Figure 4. Summary of T-test metrics for Top Genes in the CHIR99021 treated dataset. This figure displays a table summarizing the t-test results, including top genes, p-value, corrected p-value, t-statistic, effect size and z-score. These metrics were calculated for each dataset funneled through the pipeline*

Model Evaluation Metrics

A table of numbers and letters

Description automatically generatedEvaluation metrics were retrieved for each model or statistical test across each dataset. These metrics are used to assess the overall performance quality of each model through each dataset. Figure 5 shows a full summary of model evaluation metric retrieved from the pipeline output. For each model, accuracy, precision, recall and the F1 score were retrieved. The values of each metric were shown to be seemingly identical, based on the number of significant figures used in the output for each model. Most models performed at a higher level, at a greater than 0.75 value for each metric. However, there were notable exceptions. The Neural Network model for the C59 dataset showed performance metrics at 0.667, noticeably lower than most others. The Neural Network model for the FOXA1 dataset performed at an even lower level, showing 0.11, repeating.

*Figure 5. Model evaluation metrics for each machine learning model across all datasets. These metrics allow you to evaluate the accuracy, precision, recall and F1 score for every model run during the pipeline. Values closer to 1 indicate better performance of the model.*

Heatmaps

Heatmaps, shown in figure 6, were generated using the complete top 50 and top 2000 genes lists produced by the top genes pipeline. These heatmaps were generated by the second portion of this pipeline, where occurrences of each gene across all methods and models was generated. This was done to visualize any patterns formed by gene identification by different models across the datasets containing samples with diverse treatment conditions. Genes identified in greater than two datasets were highlighted as datapoints on the heatmaps and are depicted in the lighter blue colour. The first column of the heatmap shows occurrence levels of genes plotted, ranging from 2-20 in the A screenshot of a graph

Description automatically generatedtop 2000 genes heatmap, and 2-13 in the top 50 genes heatmap. The red outlined samples represent the C59 dataset samples. This was done to identify those samples as a comparison between different directions of pathway dysregulation.

*Figure 6. Heatmaps showing gene trends across multiple models. A) Heatmap showing gene occurrence trends across the top 2000 genes list. Gene occurrence counts are shown in the first column and the inhibitory dataset C59 is outlined in red. B) Heatmap showing gene occurrence trends across the top 50 genes list. Gene occurrence counts are shown in the first column and the inhibitory data set C59 samples are outlined in red. Gene names are omitted from the y-axis on both heatmaps due to size constraints.*

Discussion

This study aimed to utilize atypical signaling of the Wnt pathway to construct a bioinformatics tool that would utilize machine learning models to identify key target genes across datasets with varying experimental conditions. Construction of this tool could streamline the process of gene identification and selection for further study, while simultaneously providing insights into the functional consequences of Wnt pathway dysregulation in diverse biological contexts (Rosati et al., 2024). The ability to facilitate the understanding of complex biological pathways such as Wnt is valuable to advancing research and therapeutic strategies. The resulting bioinformatics pipeline generated top genes list for each model for each dataset, a comprehensive list outlining genes that occur across datasets and identification methods. Identified genes generally aligned with literature, such as the CDH2 gene (cadherin 2) that appeared across 20 out of 30 top 2000 genes lists. This gene is known to be downregulated by Wnt and is connected to several malignancies and drug resistances in acute myeloid leukemia (Parker et al., 2023). Other highly identified genes, such as ACTB, or Actin Beta, which occurred in 10 out of 30 top 50 genes lists currently does not have any literature connecting it to the Wnt pathway, highlighting this pipeline’s ability to identify potential target genes that have not yet been studied.

Model evaluation metrics—accuracy, precision, recall, and F1 score—were found to be equal across all models. This consistency could be attributed to the balanced nature of the datasets or the binary classification methods employed by 'scikit-learn.' Given the small sample sizes (n=6-9) and the balanced classes between treated and control groups, these metrics may appear similar (Wei and Dunbrack, 2013). A more detailed analysis with increased significant figures might reveal minor differences between the metrics. Two Neural Network models performed poorly on the C59 and FOXA1 datasets. This performance can be attributed to several factors. RNA-seq data is characteristically high-dimensional due to the large number of genes represented in the transcriptome and gene expression data points. This high dimensionality can lead to overfitting and increased computational complexity (Wang et al., 2010). Additionally, Neural Networks lack direct importance scores for feature identification, which can complicate the selection process, particularly if the data is overfitted (Huang et al., 2020). To address these issues, future work should incorporate SHAP (Shapley Additive Explanations) values for feature identification. SHAP values provide a robust method for understanding feature importance by evaluating all possible feature combinations and their impact on model outputs (Wang et al., 2024). Additionally, performing PCA (Principal Component Analysis) for dimensionality reduction before feature selection, rather than in parallel with machine learning models, may improve handling of high-dimensional data and reduce the risk of overfitting, potentially enhancing Neural Network performance (Huang et al., 2020).

To further assess the consistency of gene identification across different models as well as gene patterns, heatmaps were generated for both 50 and 2000 top gene “total occurrences” lists. Datapoints on each heatmap are genes that are identified within top genes lists across models and treatment conditions. The clustering of apparent genes along these heatmaps can show whether certain models are identifying similar genes as compared with other models. What is immediately apparent by looking at these heatmaps is that PCA and Support Vector Machines methods are identifying genes that occur more frequently across datasets, especially within the same methods. This fact may emphasise PCA and SVMs effectiveness in detecting important features, and the consistency may suggest that these methods may be more reliable for gene identification in diverse conditions. Another observation of the heatmaps is that Random Forest and Neural Networks are identifying other less commonly occurring genes. The algorithms of these machine learning techniques may be identifying less common, yet potentially significant, genes that the other more classical models might overlook. This could be due to the way these algorithms are handling importance and feature selection. Random Forest’s ensemble approach and ability to handle high-dimensional data to identify important features using tree diversity can contribute to a broader exploration of feature space, leading to the identification of these less frequently identified possible target genes (Alkharabsheh et al., 2022). Neural networks utilize deep learning models that can capture complex and non-linear relationships within datasets (Alkharabsheh et al., 2022). This could allow the selection of genes that do not have straightforward associations but may be important for specific experimental conditions. The ability of Random Forest and Neural Networks to identify fewer common genes highlights their potential for discovering novel biomarkers or therapeutic targets that might not be immediately apparent from more commonly occurring genes and traditional feature identification methods. This aspect shows the benefit of utilizing multiple models and treatment conditions to analyze feature importance for signal transduction pathways. The approach taken by this pipeline construction to integrate different machine learning models, along with traditional methods such as PCA and T-tests ensures that both common and uncommon gene targets can be identified.

The main drawback of utilizing this type of data for model training is that the RNA-seq data often contains small sample sizes, as this limits the statistical power and the reliability of the target gene identification process. Small sample sizes can also lead to overfitting of the data reducing the reproducibility of the results (Pang and Jung, 2014). One way I’ve tried to overcome this issue is by employing Leave One Out Cross-Validation (LOOCV). LOOCV is a cross-validation technique that is tailored for training models on data with small sample sizes. This is due to its nature of using all available samples for training while reserving only one sample for testing. By increasing the number of cross-validation folds, LOOCV helps maximize the use of available data for trailing while providing hardy evaluation of the model’s performance on unseen data (Geroldinger et al., 2023). This approach aims to solve the problem of overfitting and improve the reliability of the results despite the limited sample size.

Other limitations of this study centre on model performance and evaluation. The performance of each machine learning model within my study may vary significantly, depending on parameters used and model training procedures. The way I have tried to mitigate this limitation is by the implementation of a model performance metric output. This allows the user to take both the gene list output and the evaluation metrics into consideration when deciding on a target gene of interest. However, the evaluation metrics used may not give an inclusive understanding of a model’s performance, especially using datasets with such balanced sample conditions. Other metrics that can be added to increase model evaluation include AUC-ROC, which measures the ability of the model to distinguish between classes across different thresholds, or the Matthews Correlation Coefficient (MCC) which uses a balanced measure to take true and false positives and negatives into account for binary classification tasks (Carrington et al., 2022).

To improve this pipeline for future use, I would integrate improved model evaluation metrics to increase understanding of model performance, integrate multi-omics approaches to deepen understanding of gene function, regulation and interaction and increase data sample sizes, perhaps incorporating metadata to provide context to filter or group samples based on specific characteristics. By refining model evaluation techniques and expanding the data framework, the pipeline will become more effective in identifying key target genes and understanding their roles in complex pathways such as the Wnt pathway. Ultimately, these advancements will support more accurate predictions, direct experimental design and contribute to the development of targeted therapeutic strategies.

References

1. Abusamra, H. (2013). A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma. *Procedia Computer Science*, *23*, 5–14. <https://doi.org/10.1016/j.procs.2013.10.003> (Abusamra., 2013)
2. Alcaraz, J., Labbé, M., & Landete, M. (2022). Support Vector Machine with feature selection: A multiobjective approach. *Expert Systems with Applications*, *204*, 117485. <https://doi.org/10.1016/j.eswa.2022.117485> (Alcaraz et al., 2022)
3. Alkharabsheh, K., Alawadi, S., Kebande, V. R., Crespo, Y., Fernández-Delgado, M., & Taboada, J. A. (2022). A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: A study of God class. *Information and Software Technology*, *143*, 106736. <https://doi.org/10.1016/j.infsof.2021.106736> (Alkharabsheh et al., 2022)
4. Bell, I., Khan, H., Stutt, N., Horn, M., Hydzik, T., Lum, W., Rea, V., Clapham, E., Hoeg, L., & Van Raay, T. J. (2024). Nkd1 functions downstream of Axin2 to attenuate Wnt signaling. *Molecular Biology of the Cell*, *35*(7), ar93. <https://doi.org/10.1091/mbc.E24-02-0059-T> (Bell et al., 2024)
5. Boonekamp, K. E., Heo, I., Artegiani, B., Asra, P., Van Son, G., De Ligt, J., & Clevers, H. (2021). Identification of novel human Wnt target genes using adult endodermal tissue-derived organoids. *Developmental Biology*, *474*, 37–47. <https://doi.org/10.1016/j.ydbio.2021.01.009> (Boonekamp et al., 2021)
6. Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., McInnes, M., & Holzinger, A. (2023). Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 329–341. <https://doi.org/10.1109/TPAMI.2022.3145392>
7. Geroldinger, A., Lusa, L., Nold, M., & Heinze, G. (2023). Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—A simulation study. *Diagnostic and Prognostic Research*, *7*(1), 9. <https://doi.org/10.1186/s41512-023-00146-0> (Geroldinger et al., 2023)
8. Handly, L. N., Yao, J., & Wollman, R. (2016). Signal Transduction at the Single-Cell Level: Approaches to Study the Dynamic Nature of Signaling Networks. *Journal of Molecular Biology*, *428*(19), 3669–3682. <https://doi.org/10.1016/j.jmb.2016.07.009> (Handy et al., 2016)
9. He, S., Lu, Y., Liu, X., Huang, X., Keller, E. T., Qian, C.-N., & Zhang, J. (2015). Wnt3a: Functions and implications in cancer. *Chinese Journal of Cancer*, *34*(3), 50. <https://doi.org/10.1186/s40880-015-0052-4> (He et al., 2015)
10. Huang, Y., Jin, W., Yu, Z., & Li, B. (2020). Supervised feature selection through Deep Neural Networks with pairwise connected structure. *Knowledge-Based Systems*, *204*, 106202. <https://doi.org/10.1016/j.knosys.2020.106202> (Huang et al., 2020)
11. Ito, K. K., Tsuruoka, Y., & Kitagawa, D. (2024). Predicting potential target genes in molecular biology experiments using machine learning and multifaceted data sources. *iScience*, *27*(3), 109309. <https://doi.org/10.1016/j.isci.2024.109309> (Ito et al., 2024)
12. Kaur, N., Chettiar, S., Rathod, S., Rath, P., Muzumdar, D., Shaikh, M. L., & Shiras, A. (2013). Wnt3a mediated activation of Wnt/β-catenin signaling promotes tumor progression in glioblastoma. *Molecular and Cellular Neuroscience*, *54*, 44–57. <https://doi.org/10.1016/j.mcn.2013.01.001> (Kaur et al., 2013)
13. Kleeman, S. O., Koelzer, V. H., Jones, H. J., Vazquez, E. G., Davis, H., East, J. E., Arnold, R., Koppens, M. A., Blake, A., Domingo, E., Cunningham, C., Beggs, A. D., Pestinger, V., Loughrey, M. B., Wang, L.-M., Lannagan, T. R., Woods, S. L., Worthley, D., Tomlinson, I., … Leedham, S. J. (2020). Exploiting differential Wnt target gene expression to generate a molecular biomarker for colorectal cancer stratification. *Gut*, *69*(6), 1092–1103. <https://doi.org/10.1136/gutjnl-2019-319126> (Kleeman et al., 2020)
14. Liu, J., Xiao, Q., Xiao, J., Niu, C., Li, Y., Zhang, X., Zhou, Z., Shu, G., & Yin, G. (2022). Wnt/β-catenin signalling: Function, biological mechanisms, and therapeutic opportunities. *Signal Transduction and Targeted Therapy*, *7*(1), 3. <https://doi.org/10.1038/s41392-021-00762-6> (Liu et al., 2022)
15. Marcilio, W. E., & Eler, D. M. (2020). From explanations to feature selection: Assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347. <https://doi.org/10.1109/SIBGRAPI51738.2020.00053> (Marcilio & Eler, 2020)
16. Motono, M., Ioroi, Y., Ogura, T., & Takahashi, J. (2016). WNT-C59, a Small-Molecule WNT Inhibitor, Efficiently Induces Anterior Cortex That Includes Cortical Motor Neurons From Human Pluripotent Stem Cells. *Stem Cells Translational Medicine*, *5*(4), 552–560. <https://doi.org/10.5966/sctm.2015-0261> (Motono et al., 2016)
17. Mukherjee, S., Luedeke, D. M., McCoy, L., Iwafuchi, M., & Zorn, A. M. (2022). SOX transcription factors direct TCF-independent WNT/β-catenin responsive transcription to govern cell fate in human pluripotent stem cells. *Cell Reports*, *40*(8), 111247. <https://doi.org/10.1016/j.celrep.2022.111247> (Mukherjee et al., 2022)
18. Nakagawa, S., Yamaguchi, K., Takane, K., Tabata, S., Ikenoue, T., & Furukawa, Y. (2024). Wnt/β-catenin signaling regulates amino acid metabolism through the suppression of CEBPA and FOXA1 in liver cancer cells. *Communications Biology*, *7*(1), 510. <https://doi.org/10.1038/s42003-024-06202-9> (Nakagawa et al., 2024)
19. Pagella, P., Söderholm, S., Nordin, A., Zambanini, G., Ghezzi, V., Jauregi-Miguel, A., & Cantù, C. (2023). The time-resolved genomic impact of Wnt/β-catenin signaling. *Cell Systems*, *14*(7), 563-581.e7. <https://doi.org/10.1016/j.cels.2023.06.004> (Pagella et al., 2023)
20. Pang, H., & Jung, S. (2013). Sample Size Considerations of Prediction‐Validation Methods in High‐Dimensional Data for Survival Outcomes. *Genetic Epidemiology*, *37*(3), 276–282. <https://doi.org/10.1002/gepi.21721> (Pang & Jung, 2013)
21. Parker, J., Hockney, S., Blaschuk, O. W., & Pal, D. (2023). Targeting N-cadherin (CDH2) and the malignant bone marrow microenvironment in acute leukaemia. *Expert Reviews in Molecular Medicine*, *25*, e16. <https://doi.org/10.1017/erm.2023.13> (Parker et al., 2023)
22. Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal*, *23*, 1154–1168. <https://doi.org/10.1016/j.csbj.2024.02.018> (Rosati et al., 2024)
23. Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. <https://doi.org/10.1007/s42979-021-00592-x> (Sarker, 2021)
24. Schenone, M., Dančík, V., Wagner, B. K., & Clemons, P. A. (2013). Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology*, *9*(4), 232–240. <https://doi.org/10.1038/nchembio.1199> (Schenone et al., 2013)
25. Taghizadeh, E., Heydarheydari, S., Saberi, A., JafarpoorNesheli, S., & Rezaeijo, S. M. (2022). Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinformatics*, *23*(1), 410. <https://doi.org/10.1186/s12859-022-04965-8> (Taghizadeh et al., 2022)
26. Van De Stolpe, A., Holtzer, L., Van Ooijen, H., Inda, M. A. D., & Verhaegh, W. (2019). Enabling precision medicine by unravelling disease pathophysiology: Quantifying signal transduction pathway activity across cell and tissue types. *Scientific Reports*, *9*(1), 1603. <https://doi.org/10.1038/s41598-018-38179-x> (Van De Stolpe et al., 2019)
27. Wang, B., Khan, S., Wang, P., Wang, X., Liu, Y., Chen, J., & Tu, X. (2022). A Highly Selective GSK-3β Inhibitor CHIR99021 Promotes Osteogenesis by Activating Canonical and Autophagy-Mediated Wnt Signaling. *Frontiers in Endocrinology*, *13*, 926622. <https://doi.org/10.3389/fendo.2022.926622> (Wang et al., 2022)
28. Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1), 44. <https://doi.org/10.1186/s40537-024-00905-w> (Wang et al., 2024)
29. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. <https://doi.org/10.1038/nrg2484> (Wang et al., 2009)
30. Wei, Q., & Dunbrack, R. L. (2013). The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, *8*(7), e67863. <https://doi.org/10.1371/journal.pone.0067863> (Wei & Dunbrack, 2013)
31. Zhan, T., Rindtorff, N., & Boutros, M. (2017). Wnt signaling in cancer. *Oncogene*, *36*(11), 1461–1473. <https://doi.org/10.1038/onc.2016.304> (Zhan et al., 2017)
32. Zhang, L.-M., Li, M., Tian, C.-C., Wang, T.-T., & Mi, S.-F. (2021). CCAAT enhancer binding protein α suppresses proliferation, metastasis, and epithelial-mesenchymal transition of ovarian cancer cells via suppressing the Wnt/β-catenin signaling. *Neoplasma*, *68*(03), 602–612. <https://doi.org/10.4149/neo_2021_210103N2> (Zhang et al., 2021)
33. Zhong, H., Duan, B.-H., Du, F.-M., Wang, W.-M., & Qiao, H. (2023). Identification of key genes, biological functions, and pathways of empagliflozin by network pharmacology and its significance in the treatment of type 2 diabetes mellitus. *Annals of Translational Medicine*, *11*(2), 123–123. <https://doi.org/10.21037/atm-22-6406> (Zhong et al., 2023)