BS831 Final Project

Introduction

Head and neck squamous cell carcinomas (HNSCCs) are characterized by heterogeneous tumors which arise from the squamous epithelium of the pharynx and oral cavity (Pulte et al, 2010). As of 2023, they are the 6th most common forms of cancer and are notoriously resistant to treatment beyond early stages (Choi et al, 2023). For this reason, investigations into the alterations of the genomic landscape at each stage of HNSCC progression have been particularly important in identifying molecular and biological pathways that could serve as both markers of the stage of cancer of particular cells as well as potential targets for treatments and drug development. The Cancer Genome Atlas Network's profiling of 279 HNSCCs and contribution in making the data freely available for individual analysis has allowed for significant progress towards better understanding the somatic genomic alterations and other bio pathogenic associations related to the presence and stage of HNSCC.

This dataset, consisting of RNA sequencing data displaying gene expression counts from sampled tumoral cells at different stages as well as feature and phenotypic metadata for both the laboratory samples and individual patients, allows for further analysis into the differential gene and pathway activity in HNSCC cases at each stage. For this project, differential expression analysis was run on the RNA sequencing expression set to investigate differences in gene transcription expression between grade 1 and grade 3 HNSCC cancer cells, and gene enrichment analysis was performed to further investigate which pathways were enriched as a result of the differences in gene expression. Hierarchical clustering was used to examine the similarities in up- and downregulated gene patterns between grade 1 and grade 3 samples. Finally, classification analysis was run using three methods, random forest, SVM, and elastic net, in order to train models to identify the grade of a sample when given the gene expression of a cell. Model performance between the three methods was compared. This experimental design allows for a two-fold approach to the analysis of the available data: 1) A supervised learning approach which investigates the differences in expression data between the two known labels (g1 vs g3), and 2) An unsupervised classification approach attempting to train a model that will be able to assign a label (either g1 or g3) to unknown cells when given their expression data. The aims of this project are to discover any significant gene and pathway enrichments, either positive or negative, that are different between the two grades of samples, and to identify the best model for predicting grade of a sample from the raw data.

Methods

Data Pre-processing and Normalization

Data was acquired and loaded into R as an ExpressionSet. The data contained 40 samples each for grades: AE, g1, g3, for a total of 120 samples. The data was filtered so only samples with phenotypes is grades g1 and g3 were kept, and any missing expression values were eliminated. As the distribution of the expression data was sharply skewed (Supplementary Figure 1a and 1b), further normalization steps were taken. The expression data was log2 transformed in order to stabilize variance and bring the data closer to a normal distribution in order to prepare it for further statistical analysis. Quantile normalization was then used to remove systematic biases such as differences in sequencing depth, and make the data comparable across the g1 and g3 levels. This was specifically chosen for the RNAseq data as it matches the distributions across expression values of the sample levels, but preserves the relative expression levels within each sample, allowing for later differential expression. A filter was also applied to the data in order to remove low variance genes in order to reduce noise and improve statistical power. Following pre-processing and normalization, the data displayed a more uniform expression pattern (Supplementary Figure 1c and 1d). Data handling and rearranging was done using the R packages Biobase_2.56.0, readr_2.1.4, dplyr_1.1.3, and reshape2_1.4.4. Graphical outputs and normalizations were run using ggplot2_3.5.1 and preprocessCore_1.58.0.

Differential Expression Analysis

Confounding variables were identified through the literature accompanying Cancer Genome Atlas Network's publication of the data. An attempt was made to assess the association of each metavariable to each other by calculating correlation coefficients in the case of continuous variables, or performing a t-test or ANOVA for the case of categorical variables with one or more levels respectively, but this kept crashing R no matter how many times it was run. For the sake of processing power and time, the variables "patient.number_pack_years_smoked" and "patient.age_at_initial_pathologic_diagnosis", were identified as the highest correlated and included as confounding factors. Both the variables "patient.number_pack_years_smoked" and "patient.age_at_initial_pathologic_diagnosis" as well as the metadata for "grade" were checked for completeness and any samples with NA values in these columns of the pData were eliminated. A dds object was then created with the filtered expression and phenotypic data, and the design included the columns listed above and grade. Reference level for grade was set at g1. Differential analysis was run using DESeq2_1.36.0 and significant genes defined as any adjusted p-value of under 0.05. And MA plot was created to visualize the mean-abundance relationship across all genes in the dataset and asses the quality of the data. A volcano plot was created to highlight the relationship between fold change and statistical significance for each gene, with genes that have a p-value <0.05 highlighted in green. Plotting was done using ggplot2. For DESeq only, raw data counts were used and DESeq package automatically applied its own normalization method by size factor and log transformation.

Hierarchical Clustering

The normalized counts of the top 50 significant genes were extracted from the DESeqDataSet and formatted for hierarchical clustering analysis and heatmap visualization. Two heatmaps were created, one displaying the Euclidean distance between each sample and grouped by grade g1 or g3, and another clustered by the Pearson correlation coefficient dendrogram. The R packages pheatmap_1.0.12 and reshape2_1.4.4 were used for this section.

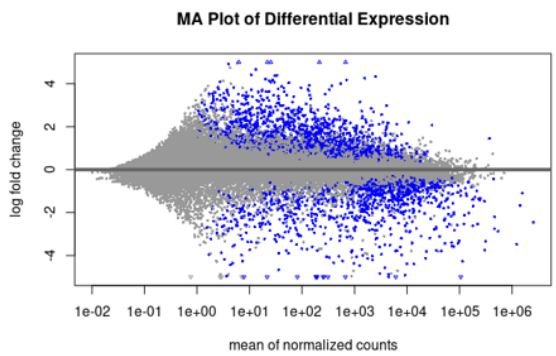
Gene Enrichment Analsys

This analysis used HALLMARK genesets for the species Homo sapiens to investigate differential pathway enrichment. HALLMARK gene sets were loaded in using R package msigdbr_7.5.1. Log fold change was used to rank expression data genes assuming the two conditions g1 and g3. The R package limma_3.52.4 was used to model a design matrix based on the two conditions and fit a model to the normalized expression data. HGNC symbols extracted from the fData of the ExpressionSet were used to match enriched genes to HALLMARK gene sets. Significant pathways were defined as those with p-value<0.05. Enriched pathways were plotted using ggplot2 and analysis was run using fgsea_1.22.0.

Classification

Expression data and grade labels were extracted from the ExpressionSet. The R package caret_6.0-92 was used to split the data into a training and testing set in a 70/30 ratio. The R package randomForest_4.7-1.1 was used to train a RandomForest model and obtain predictions from the test data. The R package glmnet_4.1-4 was used to train a Elastic Net model and obtain predictions from the test data. The R package e1071_1.7-11 was used to train a SVM model and obtain predictions from the test data. Confusion matrices were created for each model to display number of true positive, true negative, false positive, and false negative predictions regarding the class (g1 or g3) of the data given the expression counts for each gene of the sample. AUC (area under cruve) for ROC curves and F1 scores (combination scores for precision and recall) were calculated for each method for comparison, using R packages MLmetrics_1.13 and pROC_1.18.0. The 20 most important features identified in each of the models were also plotted for comparison using ggplot2. PCA plots of the data's predicted versus true labels were shown for each model.

A



B

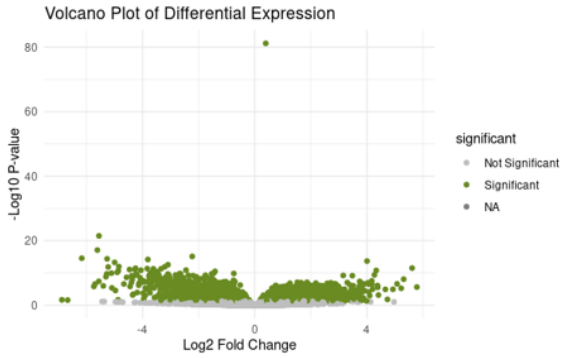


Figure 1: A) MA plot displaying distribution of log fold change in expression level versus mean of normalized counts for the data, with blue indicating a p-value <0.05. B) Volcano plot of differentially expressed genes according to DESe2 analysis, with green indicating a p-value<0.05.

Results

Analysis

Differential gene expression analysis with the confounding variables of number of years smoked and patient age at initial diagnosis showed 1762 genes (5.7% of the total 31,125) were upregulated with a positive log fold change and 1337 genes (4.3% of the total 31125) were downregulated with a negative log fold change (Figure 1b). The most significant gene was TTTY15, which was found to have a positive log fold change of 0.39 (p-value <0.05). The next top genes were VSIG8, ARG1, TMEM79, and SLURP1, all of which had negative log fold changes (p-value < 0.05). An MA plot was created to verify the quality of the data input, showing the log fold change in expression levels between the two grade levels for each gene and the average expression level of a gene across the grade levels. The MA plot displays a fairly symmetric distribution with a concentration of points at low expression levels, as is typical of RNAseq data (Figure 1a).

A hierarchical clustering analysis was performed with the gene expression data across both grade levels being clustered using the Pearson correlation coefficient (Figure 2a). The Pearson's correlation coefficient is used typically in RNAseq clustering analysis and is often used with a bottom-up approach that organizes the data into a dendrogram to visualize relationships between groups of genes across phenotypes. Figure 2a displays underexpression of several genes in g3 phenotypes with concurrent overexpression in g1 phenotypes. The phenotypes also tend to cluster together.

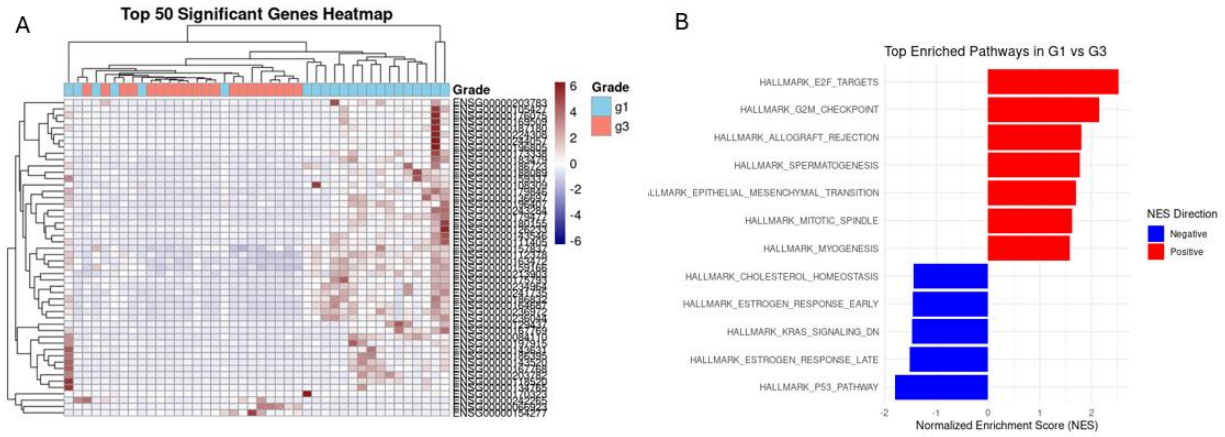


Figure 2: A) Hierarchical clustering heatmap of log fold expression change in the top 50 significant genes, labeled by phenotype. B) GSEA analysis results displaying top positively and negatively enriched pathways in G1 vs G3 phenotypes (p-value <0.05).

A gene enrichment analysis was run to identify the top positively or negatively enriched pathways in g3 phenotypes as opposed to g1 (Figure 2b). Pathways such as the G2M checkpoint, mitotic spindle, and epithelial to mesenchymal transition were found to be positively enriched, while pathways such as the p53 pathway were found to be negatively enriched (p-value < 0.05).

Classification

The data was plot 70/30 into training and testing sets with each point assigned at random. Three separate models, Random Forest, Elastic Net, and SVM, were fit to the data and evaluated for their capacity to correctly classify each sample to either the g1 or g3 phenotypes. ROC curves evaluating true positive rates (sensitivity) and false positive rates (specificity) were created for each sample (Figure 3a). Ideally, classifiers would have a curve that indicates high sensitivity and low specificity. The area under the curve score for each one as well as the model's precision, recall, and F1_Score (a combination of the two prior) is displayed in Table 1. The top features for each classifier were included in Supplementary Figure 2. Out of the classifiers tested, SVM captured the data best, with Random Forest a close second, and Elastic Net coming in last. The best accuracy was tied between SVM and Random Forest at 70.8%, which indicates an average quality that could be improved upon. Overall, none of the classifiers built today would be reliable enough to use in a medical setting.

Model	Accuracy	Precision	Recall	F1_Score	AUC
Random Forest	.708333	.857142	.5	.758620	.131944
Elastic Net	.666667	.750000	.5	.714286	.222222
SVM	.708333	.857142	.5	.758620	.138889

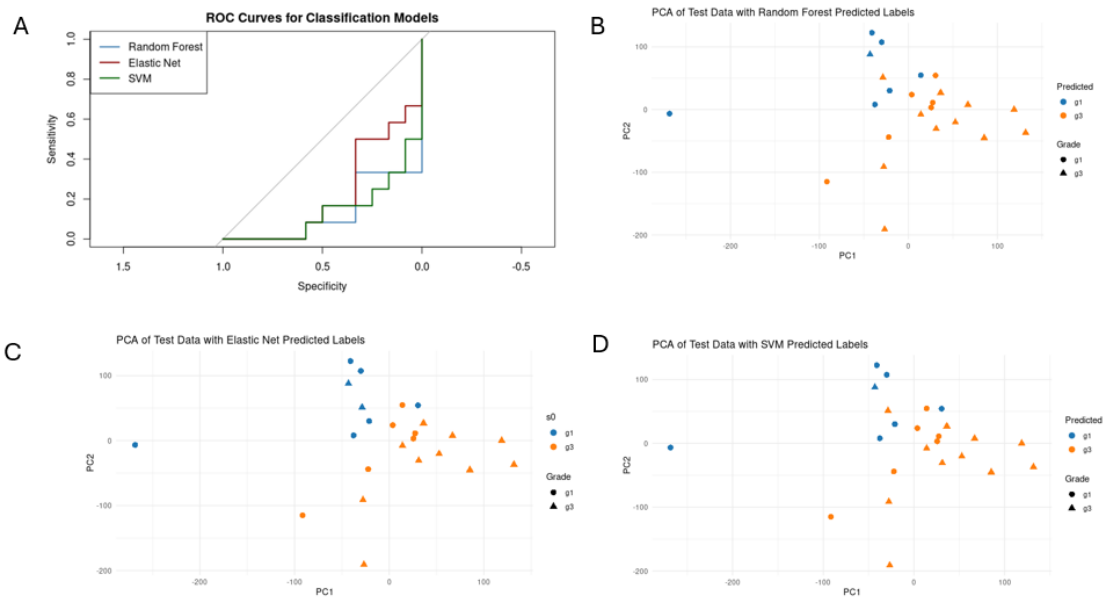


Figure 3: A) ROC curves for each model tested, B-D) PCA plot of predicted vs actual values for each model tested, color coded by predicted grade classification and shape coded by actual grade.

PCA plots of each model's predictions are displayed as a visual depiction of model accuracy and inter sample similarity (Figure 2b-d). Based on these plots, it appears that all models correctly classified most of the largely spatially separated points (indicating less inter-sample similarity and greater variance) and had a harder time distinguishing samples at the overlap (more inter-sample similarity).

Discussion

Analysis

The top individual differentially expressed genes identified included VSIG8, or V-Set and Immunoglobulin domain containing 8, which is a VISTA receptor. VISTA is a known immunotherapeutic target for cancer, autoimmune, and inflammatory diseases (Chen et al, 2022). The VSIG-8/VISTA coinhibitory pathway is thought to be a potential strategy for cancer treatment, so it is in accordance with known literature that our analysis indicated significantly lower expression levels of VSIG8 in g3 versus baselevel g1 phenotypes (log2FoldChange of -5.56). Likewise, ARG1, or the enzyme arginase 1, is known to be critical for cell proliferation, differentiation, and function (Wang et al, 2023). It can be released by tumor associated cells as well as expressed in BC cells, and its roles in cancer are still being explored, but the BC cell variant of ARG1 is a known tumor suppressor in breast cancer (Ming et al. 2020). ARG1 was found to display significantly lower expression levels in g3 versus baselevel g1 phenotypes (log2Fold change -5.61). Secreted Ly6/uPAR-related protein 1 (SLURP-1) is implicated in

control of cancer cell growth (Bychkov et al. 2021). It was found to be significantly less expressed in g3 versus baseline g1 phenotypes (log2Fold change -5.27). TTTY15, a male-specific long noncoding RNA, was found to inhibit lung cancer proliferation and metastasis via targeting T-box transcription factor 4 (Lai et al, 2019). Interestingly, it was found to be more expressed in g3 versus baselevel g1 phenotypes (log2Fold change 0.39) and was still significant even with a mixed gender dataset. TMEM79 is closely linked with immune checkpoints, drug sensitivity, and immunotherapy in hepatocellular carcinoma, and was found to also be less expressed in g3 versus baselevel g1 phenotypes (log2Fold change -2.23) (Wang et al. 2023). Altogether, this differentially expressed gene landscape seems heavily implicated in tumor regulation and growth and brings attention to some of the pathways that could be regulating HNSCC progression.

To further explore the downstream effects associated with the g3 phenotype, we performed hierarchical clustering of the gene expression data and ran a gene enrichment analysis. Using the Pearson correlation coefficient to cluster and visualize a heatmap of the gene expression across the phenotypes, it is evident that the g1 and g3 phenotypes tend to cluster together and that there is a decrease in expression of many of the top 50 significant genes in the g3 phenotype when compared to the g1 baseline (Figure 2a). The gene enrichment analysis conducted showed pathways such as the G2M checkpoint, mitotic spindle, and epithelial to mesenchymal transition to be positively enriched, while pathways such as the p53 pathway were found to be negatively enriched in the g3 phenotype as compared to the g1 baseline (Figure 2b). The epithelial to mesenchymal transition pathway in particular has been heavily implicated in HNSCC progression and metastasis (Gonzalez et al, 2021). P53 function is one of the most frequently altered pathways in HNSCC and has been considered a potential therapeutic target for treatment (de Bakker et al, 2021). Other pathways implicated in cell division and checkpoint were also differentially enriched in the g3 state as compared to the g1 baseline. It is of note that both g3 and g1 samples had present tumors and were not completely negative controls. As such, differences might relate to cancerous progression, but might be altogether more similar than when either phenotype is compared to a health control.

Overall, the results of the differential gene expression analysis, hierarchical clustering, and gene enrichment analysis conducted in this project aligned well with the findings of other studies and confirmed the potential for further causal studies and potential therapeutic targeting and drug development once causality is established. Future directions for this study might then be functional validation of the results in the lab to establish causality and explore knockdown/knockout or rescue techniques with any of the targets identified above. In terms of computational work, further exploration could be done in terms of gene regulatory networks and protein interactions networks to reveal regulators or pathways that may not have been identified here. Furthermore, alternative exon splicing might be considered, as well as experiments analysis single cell RNA sequencing at the individual cell level.

Classification

In order to identify the computational technique best suited to labeling unknown samples as g1 or g3 phenotype based on their expression data, we tested three separate unsupervised classification models, Random Forest, Elastic Net, and SVM, on the HNSCC dataset. Though Elastic Net pulled ahead based on the ROC area under curve metric, Random Forest and SVM placed better when evaluated for accuracy, precision, and recall. Recall, or the cost of False Negatives, was equivalent among them. Given the medical implications of false negatives in the diagnosis of cancer, that would have likely been the most important metric had there been a different. However ultimately, SVM did the best when the two evaluations are considered congruently. Even though it was the best of the three models tested, SVM still only displayed a 71% accuracy and an F1_score of .76 (Table 1). Accuracy between 70-90% and an F1_score between 0.5-0.8 are considered industry standard for a classifier, but would not be good enough to bring directly into clinical settings. As such, further work must be done to properly fit a model to this type of data for classification purposes. More models could be tested, such as KNN or logistic regression classifiers. More features could also be used besides just expression data – dimensions could be added for smoking status, age, et cetera.

One of the greatest limitations of this approach was lack of knowledge and experience in genomic analysis as well as classifier architecture. Though attempts have been made to verify the code is a typical RNA sequencing workflow and accurate in its analysis, there could be user error issues that I did simply did not realize, especially when constructing the heatmaps, which I was shaky on. Another limitation was lack of computing power, as observed by the attempt to identify confounding variables by calculating the association of each metavariable to each other to identify what was most closely correlated with the phenotype, which took several hours to run, freeze, and eventually crash R even when run on the SCC cluster. As such, confounding variables were instead chosen based on the literature associated with the dataset. Some of the limitations of the dataset itself include the inability to draw causal conclusions from association analysis and the very large number of NA variables in the metadata that made working with some of the pData very hard unless willing to sacrifice statistical power. With the latter I refer specifically to the NA values in the HPV status, which I would have liked to add to the confounding variables if possible as it is greatly discussed in the accompanying documentation. Future studies should take into account a higher dimensional analysis in both the exploratory and classification parts of this workflow for more accuracy in results and better fitted models (though it’s important to also avoid overfitting the training data). Altogether, one of the strongest limitations of this analysis workflow is the fact that most of this analysis has already been performed and is not new to the field, so it can only be used to confirm previous findings and emphasize points which have already been made. In that regard, all findings resulting from this project were in agreement with the surrounding literature and existing biological knowledge.

Bibliography

Alvares, C. T., Bettencourt, B. R., Silva, M. J., Costa, L. F., & Silva, R. M. (2018). hypeR: an R package for geneset enrichment workflows. Bioinformatics, 34(17), 3012-3014.

Berrar, D., & Flach, P. (2012). MLmetrics: Machine Learning Evaluation Metrics. R package version 1.1.1. Retrieved from https://CRAN.R-project.org/package=MLmetrics

Bolstad, B. M. (2020). preprocessCore: A collection of pre-processing functions. R package version 1.52.0. Retrieved from https://bioconductor.org/packages/release/bioc/html/preprocessCore.html

Bychkov ML, Shulepko MA, Shlepova OV, Kulbatskii DS, Chulina IA, Paramonov AS, Baidakova LK, Azev VN, Koshelev SG, Kirpichnikov MP, Shenkarev ZO, Lyukmanova EN. SLURP-1 Controls Growth and Migration of Lung Adenocarcinoma Cells, Forming a Complex With α7-nAChR and PDGFR/EGFR Heterodimer. Front Cell Dev Biol. 2021 Sep 14;9:739391. doi: 10.3389/fcell.2021.739391. PMID: 34595181; PMCID: PMC8476798.

Chen W, Qie C, Hu X, Wang L, Jiang J, Liu W, Liu J. A small molecule inhibitor of VSIG-8 prevents its binding to VISTA. Invest New Drugs. 2022 Aug;40(4):690-699. doi: 10.1007/s10637-022-01244-4. Epub 2022 Apr 11. PMID: 35404016.

Choi, JH., Lee, BS., Jang, J.Y. et al. Single-cell transcriptome profiling of the stepwise progression of head and neck cancer. Nat Commun 14, 1055 (2023). https://doi.org/10.1038/s41467-023-36691-x

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1-22.

González-González R, Ortiz-Sarabia G, Molina-Frechero N, Salas-Pacheco JM, Salas-Pacheco SM, Lavalle-Carrasco J, López-Verdín S, Tremillo-Maldonado O, Bologna-Molina R. Epithelial-Mesenchymal Transition Associated with Head and Neck Squamous Cell Carcinomas: A Review. Cancers (Basel). 2021 Jun 17;13(12):3027. doi: 10.3390/cancers13123027. PMID: 34204259; PMCID: PMC8234594.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., ... & Gatto, L. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods, 12(2), 115-121.

Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. Retrieved from https://CRAN.R-project.org/package=pheatmap

Kuhn, M. (2020). caret: Classification and Regression Training. R package version 6.0-86. Retrieved from https://CRAN.R-project.org/package=caret

Lai IL, Chang YS, Chan WL, Lee YT, Yen JC, Yang CA, Hung SY, Chang JG. Male-Specific Long Noncoding RNA TTTY15 Inhibits Non-Small Cell Lung Cancer Proliferation and Metastasis via TBX4. Int J Mol Sci. 2019 Jul 15;20(14):3473. doi: 10.3390/ijms20143473. PMID: 31311130; PMCID: PMC6678590.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics, 27(12), 1739-1740.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-8. Retrieved from https://CRAN.R-project.org/package=e1071

Ming Z, Zou Z, Cai K, Xu YI, Chen X, Yi W, Luo J, Luo Z. ARG1 functions as a tumor suppressor in breast cancer. Acta Biochim Biophys Sin (Shanghai). 2020 Dec 11;52(11):1257-1264. doi: 10.1093/abbs/gmaa116. PMID: 33128544.

Pulte, D. & Brenner, H. Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis. Oncologist 15, 994–1001 (2010).

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12(1), 77.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. BioRxiv, 060012.

The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517, 576–582 (2015). https://doi.org/10.1038/nature14129

Wang Y, Jin Q, Zhang S, Wang Y. Overexpression of TMEM79 combined with SMG5 is related to prognosis, tumor immune infiltration and drug sensitivity in hepatocellular carcinoma. Eur J Med Res. 2023 Nov 7;28(1):490. doi: 10.1186/s40001-023-01388-w. PMID: 37936239; PMCID: PMC10631028.

Wang, X., Xiang, H., Toyoshima, Y. et al. Arginase-1 inhibition reduces migration ability and metastatic colonization of colon cancer cells. Cancer Metab 11, 1 (2023). https://doi.org/10.1186/s40170-022-00301-z

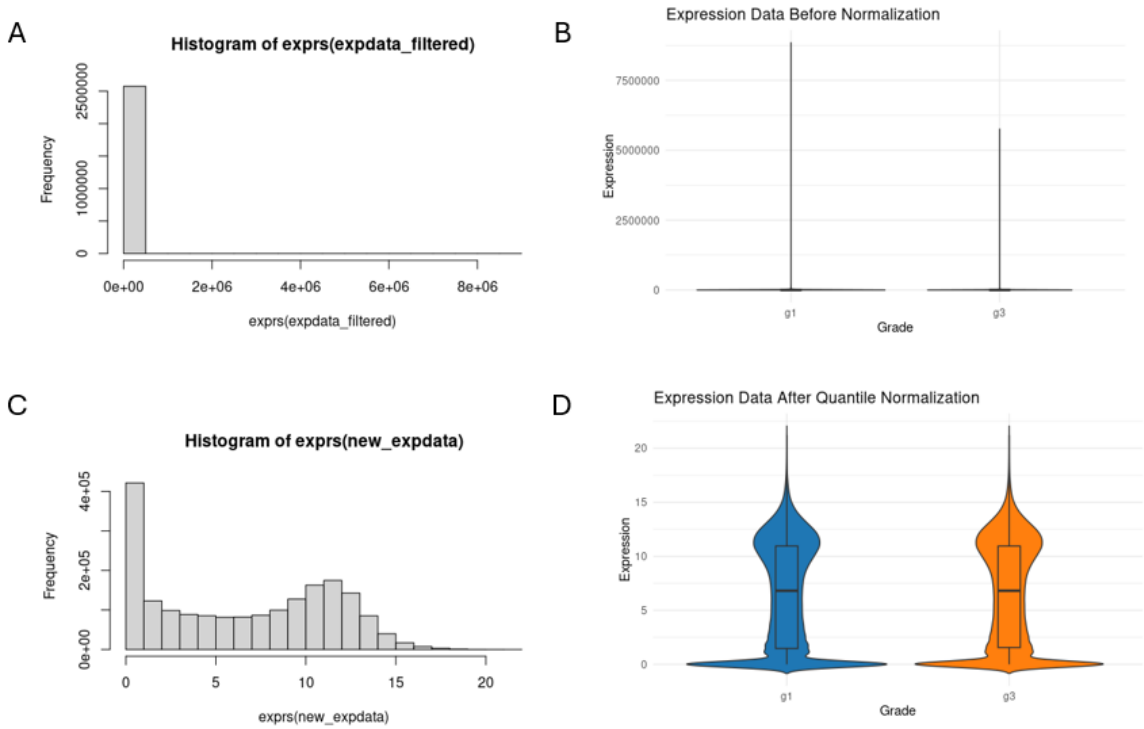
Wickham, H. (2007). Reshaping data with the reshape package. Journal of Statistical Software, 21(12), 1-20.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

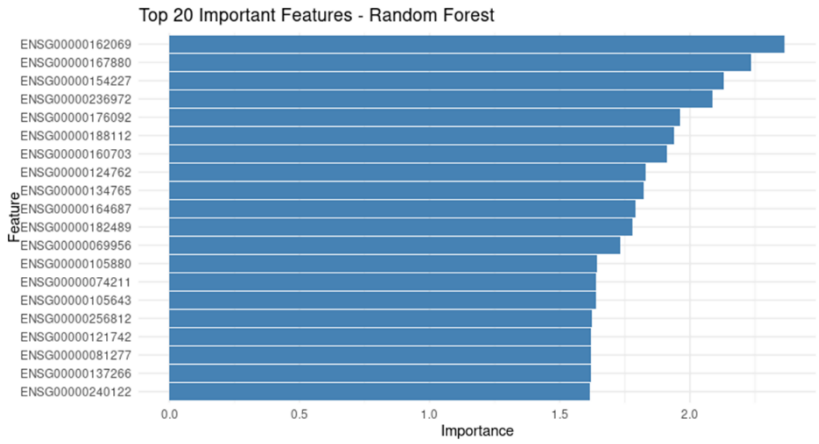
Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. Retrieved from https://CRAN.R-project.org/package=dplyr

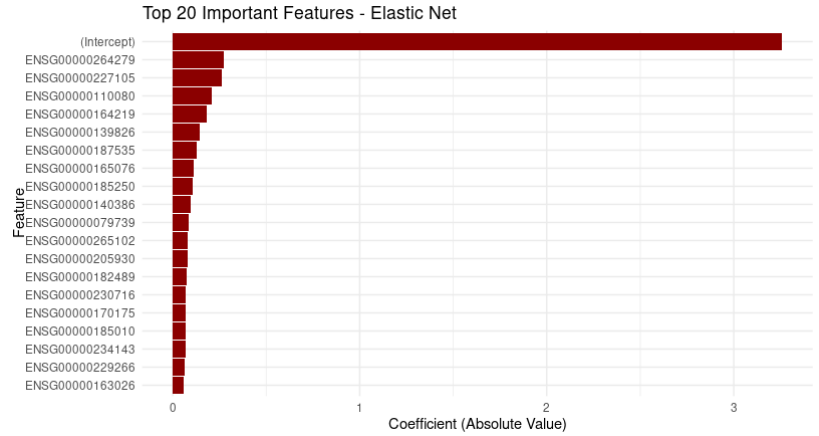
Wickham, H., Hester, J., & Francois, R. (2018). readr: Read Rectangular Text Data. R package version 1.3.1. Retrieved from https://CRAN.R-project.org/package=readr

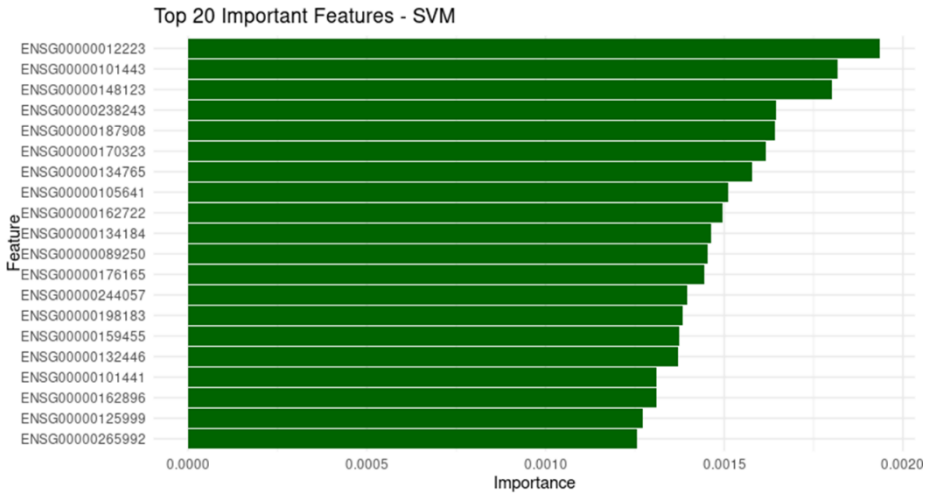
Supplementary Materials



Supplemental Figure 1: A) histogram of expression data pre-filtering and normalization, B) violin plot of expression data pre-filtering and normalization, C) histogram of expression data post filtering and normalization, D) violin plot of expression data post filtering and normalization.







Supplemental Figure 2: Top 20 most important features in classification for each model tested.