SCHOOL OF BIOLOGICAL SCIENCES

BS6202 Techniques in Biomedical Data Mining

Final Project Report

Understanding Biomarkers For Gastric Cancers due to overexpression of HOXD9 gene

Muhammad Irfan Bin Hajis

Li Ruisi

Hu Yuhan

Olivia Loh Jia Hui

Yang Yizhuo

27 October 2023

Introduction

Gastric Cancer (GC) is the fourth most diagnosed cancer whose etiology remains elusive (Machlowska et al., 2018). While there have been advancements in both diagnosis and multidisciplinary treatment approaches, the overall five-year survival rate for individuals diagnosed with GC typically falls within the range of 25% to 27%. Furthermore, the survival rate of patients with advanced stages of GC is even lower. Over the past few decades, research has demonstrated that the occurrence, recurrence, and metastasis of GC are the outcomes of complex interactions between host and environmental factors (Tan et al., 2014). These factors include phenotypic complexity, a multitude of influences, and a series of steps influenced by genetic heterogeneity and ethnic diversity. Hence, it is crucial to gain insights into the relationship between the biological mechanisms governing GC and the associated clinicopathological features. (Zheng et al., 2012) This understanding constitutes an essential step towards the development of targeted therapies and the enhancement of the quality of life for GC patients.

RNA-sequencing (RNA-seq) is a high-throughput sequencing technique used to quantify gene expression levels and identify differentially expressed genes (DEGs) between different treatment/s. RNA-seq has played a crucial role in understanding the biology of cancer, diagnosing cancer, identifying potential therapeutic targets, and developing personalised treatment strategies. RNA-seq experiments can aid to identify key genes and pathways involved in cancer development, identify potentially oncogenic fusion genes, discover novel biomarkers associated with specific cancer types used for early detection and improvement of treatment, and many more.

Previously, Li et al. performed a RNA-seq experiment and found that PAXIP1-AS1 was transcriptionally repressed by homeobox D9 (HOXD9) (Li et al, 2023). Consequently, PAXIP1-AS1 was significantly downregulated in GC tissues and cells, suppressing metastasis. Decreased expression of PAXIP1-AS1 was positively correlated with tumour progression (Li et al, 2023). PAXIP1-AS1 overexpression inhibited cell growth and metastasis both in vitro and in vivo, and significantly attenuated HOXD9-enhanced epithelial-to-mesenchymal transition (EMT), invasion and metastasis in GC cells (Li et al, 2023).

In our report, we attempt to understand the gene ontologies of DEGs and differentially alternative splicing events (ASEs) implicated in the overexpression of HOXD9 gene. We found that Li’s deposited RNA-seq data contained mid-level genomic contamination that consequently may obscure true DEGs. We also sought to determine the importance of each identified DEGS using The Cancer Genome Atlas (TCGA) within the context of GC. The source code and supplementary data are available https://github.com/micro-irfan/BS6202-Gastric-Cancer under the GPLv3 licence.

Methods

RNA-Seq Differentially Expressed Genes (DEGs) Analysis. The sequenced reads and Fragments Per Kilobase Million (FPKM) values were downloaded from Gene Expression Omnibus database (accession code GSE210016). Due to the limitations of FPKM for downstream DEGs Analysis, we performed our own quality check on the RNA-Seq data and alignment. Sequenced reads were downloaded using SRAtoolkit, quality tested using FastQC v0.11.0 and mapped to the GrCh38 human genome using the v2.7.10a STAR aligner (Dobin et al., 2012). Read alignment was performed using the default parameters. The genome index was constructed using the gene annotation provided with GrCh38 illumina iGenomes collection and sjdbOverhang value of 100. We used Qualimap and RSeQC to assess the quality of the RNA-seq alignments generated. Quantifications were performed using featureCounts v2.0.1 and differential gene expression was performed with DESeq2 v1.34.0 using the raw counts of the triplicates to compute within-group dispersion and contrasts to compare between overexpressed HOXD9-gene treated and untreated conditions. Significant DEGs were defined as having a False Discovery Rate (FDR) of < 0.05 and a log2 Fold Change of > 0.75.

RNA-Seq Alternative Splicing Events (ASE) Analysis. We used SplAdder 3.0.4 to identify differential ASE between two groups of samples (Kahles et al., 2016). Differentially spliced exons were defined using a splicing ratio |delta PSI (Percent-Spliced-In)| > 0.2 and FDR < 0.05 as thresholds.

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes Enrichment (KEGG) Analysis. To gain insights into the biological processes and functions associated with the DEGs and ASEs, we performed GO annotation using the enrichGO package in R. The enrichGO function identifies enriched GO terms, including Biological Processes (BP), Molecular Function (MF), and Cellular Components (CC) among the DEGs and ASEs. Significance was determined based on an adjusted p-value (FDR) < 0.05. We also conducted KEGG pathway enrichment analysis using the enrich KEGG function from the clusterProfiler package. Results were visualised using the ggplot2 package in R. We then ranked the DEGs according to their p-adj values and the DEGs with the top 700 smallest p-adj values were used for subsequent comparison with the TCGA dataset.

TCGA File Collection and Data Pre-processing. We downloaded a total of 448 transcriptome profiling sample files, 412 primary tumour samples and 36 normal tissue samples files, from the Genomic Data Commons (GDC) Data Portal for The Cancer Genome Atlas Stomach Adenocarcinoma (TCGA-STAD). In order to build our classification model, we used all 36 normal tissue samples and randomly selected 108 primary tumour samples, resulting in a total of 144 samples with each sample containing 19,962 gene expression data. From the 700 DEGs identified in the previous analysis, only 71 of them were identified in the TCGA dataset and these 71 genes and its 'fpkm_unstranded' expression values in the TCGA dataset were used to curate the machine learning dataset. We performed dimensionality reduction techniques, including Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE), to visualise the separation of the different sample type groups (Tumour vs Normal) using Scikit-learn library in Python3.

Machine Learning Models. Due to the imbalance in the dataset (3:1), we stratified our data when we split the dataset into training and testing dataset to train the classification models. We employed various Machine Learning Techniques, including logistic regression, support vector machine, decision tree, and random forests, to train a model to predict for malignant tumour in the context of GC. GridSearchCV was used to search for the optimal model hyperparameters and maximising cross-validation accuracy. In order to identify potential biomarkers, we extract the top 10 genes based on its coefficient generated from each tuned model. We discuss the potential of each identified gene to serve as a biomarker for GC based on evidence in the literature.

Results and Discussion

Raw Data Quality Assurance. To ensure the quality of the dataset, we performed a few quality checks. We observed the alignment and read assignments are consistent across the 6 samples, indicating a similar mapping rate and Reads Per Kilobase Million (RPKM) distribution (Figure 1c). We conducted a PCA on the RPKM and observed that there are two distinct groups (Figure 1a) with a variance of 34.5% (PC1) and 20.8% (PC2), although they are not clustered closely together (Figure 1b). We sought to explore possible reasons behind the poor clustering, relatively low uniquely mapping rate (65-70%), and high multimapping rate (10-15%) in STAR (Figure 1d). We inspected the level of rRNA contamination in each run using bbduk.sh in the bbmap toolbox (Bushnell et al., 2014) and identified that each run has a rRNA contamination percentage of between 17.10 - 24.09%, which is considered relatively higher than a properly depleted rRNA should have (~10%). On closer inspection of the alignment using the coverage plot on SplAdder and Integrative Genomics Viewer (IGV), we may have found that the sample contains genomic DNA (gDNA) contaminants. We observed a consistent low coverage (spiking at some regions) across all intronic regions in CRNDE gene (Figure 1f and 1g).

gDNA contamination is not reported in numerous RNA-seq studies. gDNA and/or rRNA contamination may unintentionally add to the gene count, as shown in situation 2 and 3 of Figure 1e, which generates more False Negatives (FN) during DEGs analysis (or False Positive for Novel mRNA Annotations) and may result in inaccurate quantitation of gene expression levels (Li et al., 2022). From a quick sequence alignment using BLAST, we found that rRNA and PAXIP1-AS has a local alignment with an e-value of 3e-89 and 85% percent identify between position 868 to 1162 on PAXIP1-AS gene (Figure 1h). This may have spiked the gene count slightly during gene/transcript quantification. Due to time constraint, we did not rerun the alignment and analysis using reads with rRNA contaminants removed to compare the difference it could make. Our results in the report are based on the original RNA-seq data.

Difference in Differentially Expressed Genes. FPKM values are not suitable for DEGs analysis using DeSeq2 due to the loss of the original count data, which is crucial for assessing statistical significance in differential expression analysis. The FPKM normalisation procedure can obscure the true differences between samples, making it difficult to accurately identify DEGs. We performed DEG analysis using DeSeq2 and the P-values are corrected using the multiple testing correction method with FDR < 0.05 to reduce erroneously reported false positive results. Li et al. defined significantly DEGs and transcript as having a false discovery rate of < 0.05 and a fold change of > 1.2 or < 0.83. Li et al. could have used a low Fold Change Threshold due to the observed contamination in the RNA-seq data.

We adjusted our abs(log2FC) threshold to > 0.27 to compare our results against Li. HOXD9 overexpression resulted in the downregulation of 18 (6 overlapping) lncRNAs and 576 (300 overlapping) protein-coding genes, whereas 2 (1 overlapping) lncRNAs and 370 (138 overlapping) protein-coding genes were upregulated. 6 lincRNAs were downregulated in both analysis including NEAT1, MALAT1, LINC00342, LINC01133 and LINC01605. PAXIP1-AS was downregulated in our analysis but disregarded due to FDR > 0.05 after checking for multiple testing corrections. We believe that PAXIP1-AS is a False Negative in our analysis due to gRNA or rRNA contamination since Li's paper was on downregulated PAXIP1-AS in GC.

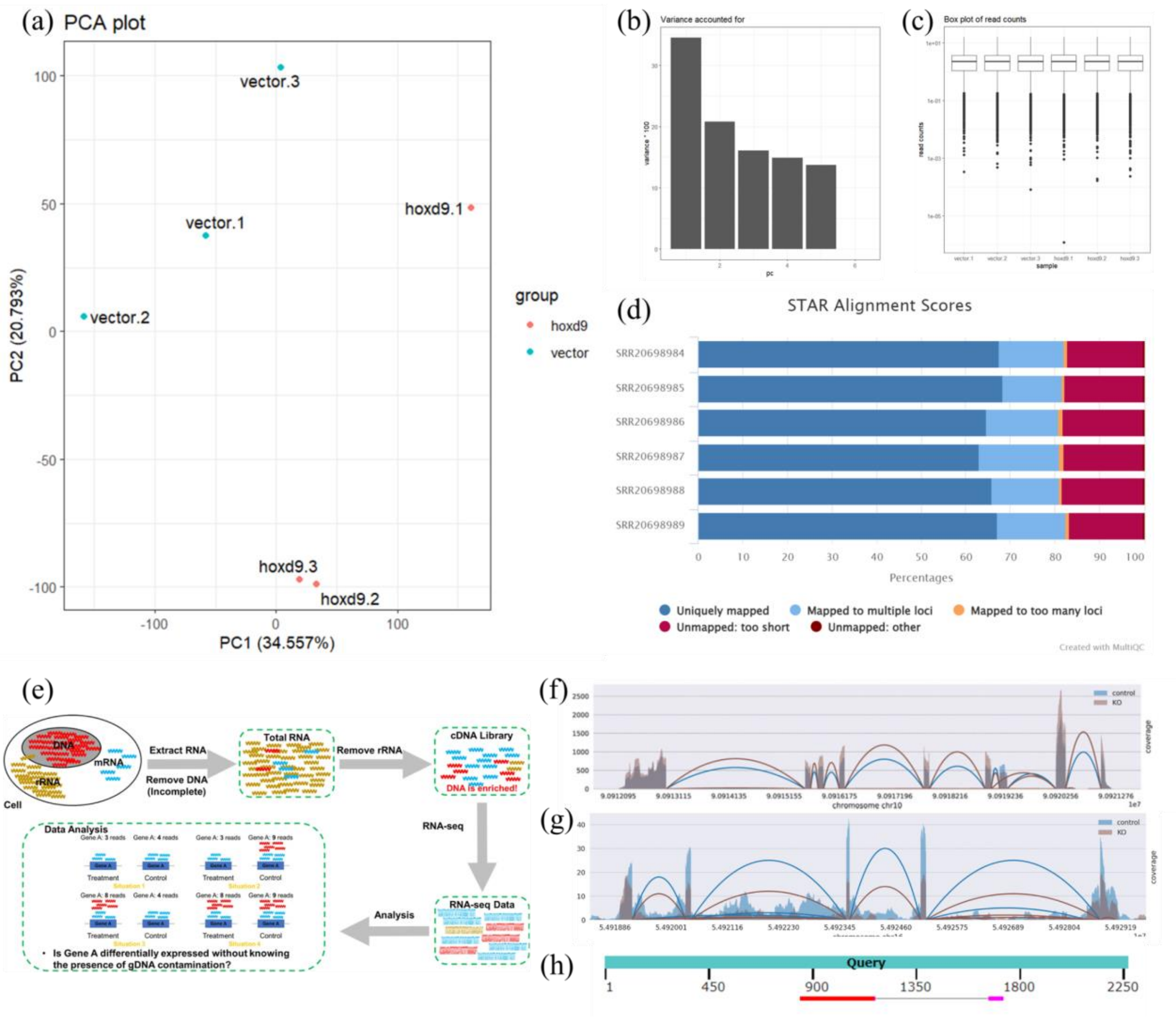

Figure 1. (a) PCA plot to separate sample group (hoxd9 and vector), (b) Variance for each PC, (c) Box plot of log2 RPKM, (d) Mapping rate of STAR Alignment, (e) When detecting DEGs between Treatment and Control groups, there are roughly four situations for one specific gene. Situations 1: Both the Treatment and Control are not contaminated by gDNA; Situation 2: Only Control is contaminated by gDNA contamination; Situation 3: Only Treatment is contaminated by gDNA; Situation 4: Both Treatment and Control are contaminated by gDNA. Different contaminating situations would result in different DEG detecting results for genes, e.g. gene A (Li et al., 2022), (f) An example of a clean Coverage plot with no alignment in the intronic region (taken from Variant Calling Methods and Protocols by Charlotte), (g) Plausible gDNA contaminants identified from low coverage across intronic regions for CRDNE gene, (h) BLAST local Alignment between PAXIP1-AS gene sequence and rRNA gene sequence

As Li only mentioned in their methods that they employed Ballgown for DEG analysis, we could only speculate the reasons behind the differences. Li has a lower DEGs for Protein Coding. This is expected as Ballgown is more sensitive to low gene level expression (Liu et al., 2022). As for LincRNAs, Li generated more LincRNAs than our analysis. Their pipeline involves an alignment step using TopHAT and a transcript assembler using either StringTie or CuffLinks, as advised on Ballgown's github page. Either method, gDNA contaminants may have contributed to the false detection of putative long non-coding RNAs (lncRNAs) and/or isoforms during transcript assembling (Iyer et al, 2015). This may have spiked the gene count during gene/transcript level quantification.

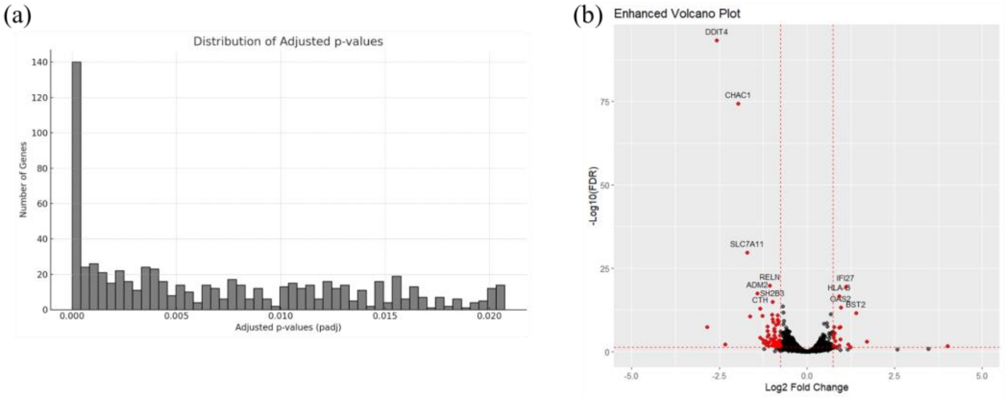

Figure 2. (a) Distribution of adjusted p-values for all the genes after multiple testing correction, (b) Volcano plot of differentially expressed genes

Differential Expression Analysis Results. In our analysis, DEGs have a FDR < 0.05 and abs(log2FC) > 0.75. The adj p-value is a measure of statistical significance that has been adjusted for multiple testing. A threshold of 0.05 is widely accepted in the scientific community as a balance between sensitivity and specificity. By maintaining this threshold, we ensure that the risk of false positives (Type I errors) remains controlled at 5%. The adjusted p-value distribution histogram for the DEG data is shown in Figure 2a. The histogram shows that 140 genes have an extremely low adjusted p-value.

Fold change represents the ratio of expression levels between conditions. The log2 transformation is commonly used to make the data symmetrical and easier to interpret. A threshold of abs(log2FC) > 0.75 ensures that we focus on genes with a biologically meaningful change. This threshold represents a fold change of approximately 1.68 or 0.595 in linear terms. By setting this threshold, we filter out genes with minor fluctuations in expression that might not be of biological relevance. This reduces the risk of false discoveries due to random noise or minor variations. The selected thresholds ensure that the gene selection method is robust, ie.

identified DEGs are both statistically significant and biologically relevant while reducing false discoveries. By setting stringent criteria, we reduce the likelihood of including genes that might show changes due to random noise or other non-relevant factors. The chosen thresholds align with commonly accepted standards in the field, ensuring that the results are comparable and interpretable by peers and reviewers.

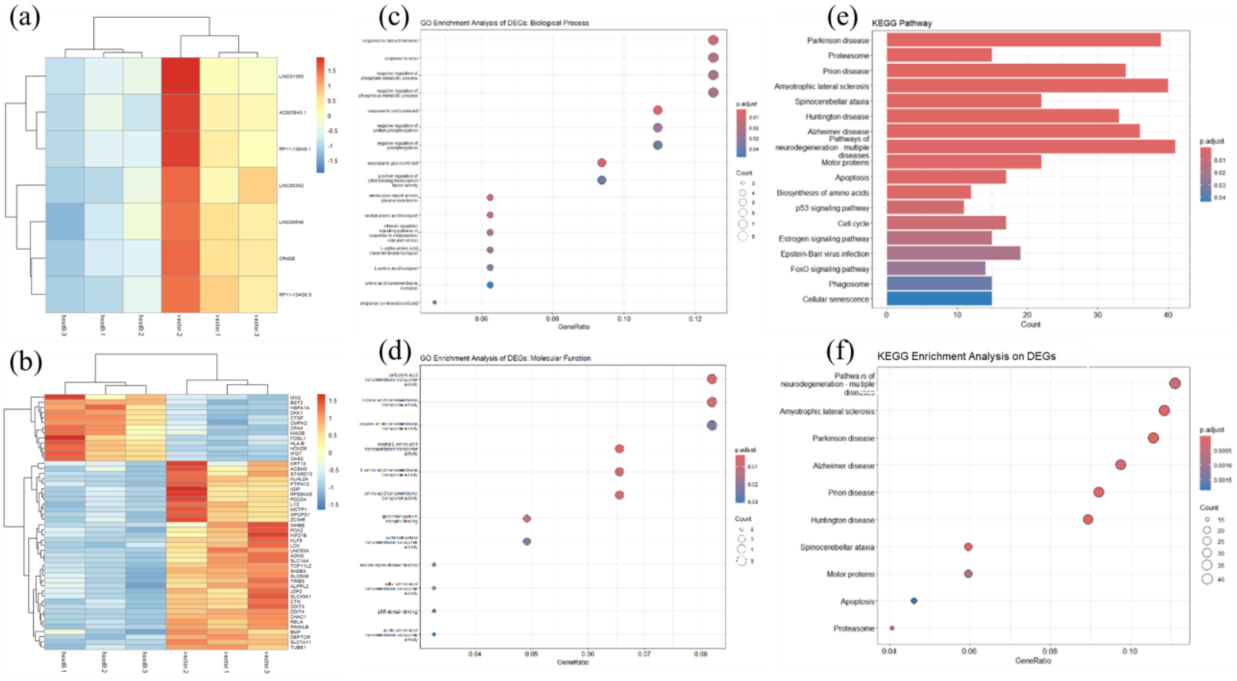

Figure 3. Total RNA was extracted from LV-HOXD9 and LV-Vector MKN-74 cells and subjected to RNA sequencing analysis using STAR Alignment and DeSeq2. The colour scale illustrates the relative expression of (a) lncRNAs and (b) top 50 protein coding genes with abs(log2FC) > 0.75 and FDR < 0.05. (c) Scatter Plot of GO Enrichment Analysis: Biological Process, (d) Scatter Plot of GO Enrichment Analysis: Molecular Function, (e) Histogram of KEGG pathway, (f) Scatterplot of KEGG pathway on DEGs

Due to the change in log2FC threshold, our number of DEGs reduced drastically. From Figure 2b, we observe that only a small number of genes are differentially expressed, mostly downregulated, due to the overexpression of the HOXD9 gene. As a quality measure, we observe that the HOXD9 gene is upregulated in the LV-HOXD9 samples as shown in Figure 3a despite the genome contamination in the RNA-seq data. HOXD9 overexpression resulted in the downregulation of 7 lncRNAs and 49 protein-coding genes, whereas 0 lncRNAs and 16 protein-coding genes were upregulated. Out of the 7 downregulated lincRNAs, AC083843.1, RP11-138A9.1, and RP11-134G8.5 have not been reported in association with GC to date among those genes. We found that AC083843.1 is located close to MIR30D and MIR30B in the genome and has the same transcriptional direction, suggesting that this lncRNA is a partial transcript of the pri-miR-30d~30b sequence. However, its specific role in GC is not clear from literature. We also found that lncRNA RP11 is highly expressed in colorectal cancer (CRC) tissues, and its expression increases with CRC stage in patients. RP11 regulates the migration, invasion and EMT of CRC cells positively in vitro and enhances liver metastasis in vivo.

GO and KEGG analysis. The x-axis refers to the GeneRatio. The larger the value, the higher the enrichment of genes enriched in this pathway. The y-axis represents the name of the enriched GO pathway. The size of the dot indicates the number of genes, and the larger the dot, the more genes are involved in the pathway; The colour represents the level of the P value, the larger -log10 (p-adj) the more significant the pathway. From the results of two ontologies, namely biological processes (BP) and molecular functions (MF) in GO enrichment results, we can find that their most enriching terms are "response to steroid hormone" and "carboxylic acid transmembrane transporter activity" (Figure 3c and 3d).

We used histograms and scatterplots to illustrate KEGG enrichment analysis on DEGs. The x-axis in the figure represents the number of DEGs under each pathway term, and the higher its value indicates the greater the degree of enrichment. The darker the colour of the bars and dots also represents the smaller the p -adj value, and the figure shows that the most enriched pathway is "Pathways of Neurodegeneration-Multiple Diseases" (Figure 3e and 3f).

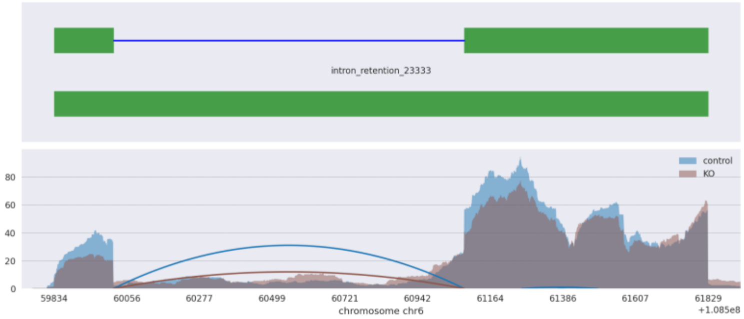

Figure 4. FOXO3 gene Retained Intron Events and its’ Coverage Plot

Differential Splicing Analysis. To discern any differences in the expression levels of AS, we compared the expression of AS between LV-HOXD9 and LV-Vector MKN-74 cells. There are five types of splicing pattern identified by SplAdder. However, only a total of 14 ASEs were identified. All ASEs are retained introns (RI). All ASEs were expressed at different levels in

GC and adjacent tissues, which were statistically significant (adjP < 0.05, abs(dPSI) ≥ 0.2). There were many AS events which were not regarded statistically significant when adjusted for multiple testing correction using the Benjamini-Hochberg (BH) procedure. This could be due to the high variability contained within each sample group itself. Two of the 15 ASEs was found to be downregulated (|log2FC| > 0.5 and FDR < 0.05) in the DEGs analysis. The set of ASEs could possibly be False Positives due to the gDNA contaminants detected earlier. There was no particular GO that was enriched based on the 14 ASEs.

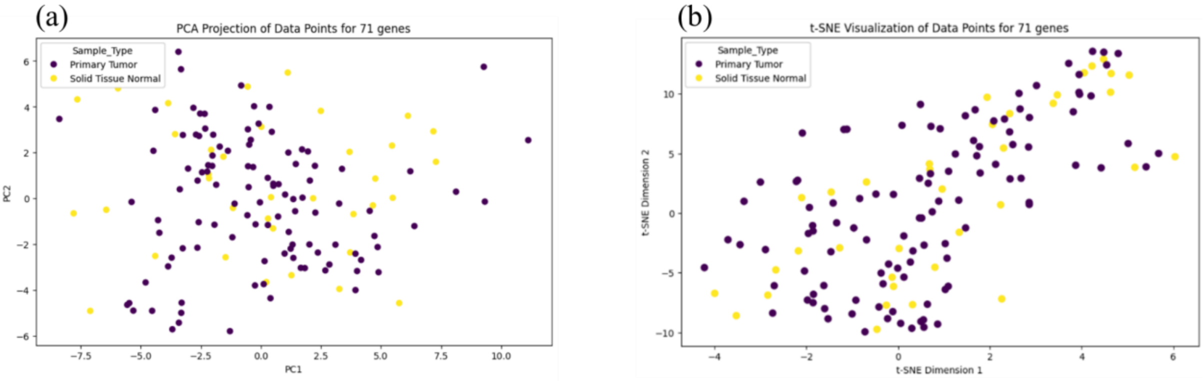

Figure 5. PCA and t-SNE Projection of Data Points for 71 genes. (a) Data visualisation in PC1 and PC2 coordinates. (b) Data visualisation in two t-SNE dimensions.

Relative Importance of DEGs as Biomarkers to Predict for Gastric Cancer. Due to the low number of DEGs in our previous analysis, we lowered our log2FC threshold and extracted 71 gene expression data from the TCGA dataset. This limited number of gene expression data can be attributed to inherent data variability, stringent DEG analysis criteria, biological diversity, differences in data sources, sample sizes, and preprocessing methods. We performed PCA and t-SNE to visualise the distribution of gene expression data in two-dimensional coordinates, to differentiate sample groups (Tumour vs Normal). Dimensionality reduction allows exploration of the underlying structure of the data and identification of potential clusters or relationships. Furthermore, we assessed the feasibility of distinguishing sample types in the reduced-dimensional coordinates, aiding in the preliminary evaluation of data separability for our classification model.

From Figure 5, The data points of the two samples are difficult to distinguish, with significant overlap in the transformed data shown in PCA (Figure 5a) and t-SNE visualisations (Figure 5b). In the PCA plot, normal sample data exhibited a more scattered distribution compared to tumour data. In the t-SNE plot, there is no evident difference in data distribution. The lack of distinct separation indicates some difficulties to classify normal and tumour samples because they cannot be easily distinguished from each other in a reduced-dimensional space.

For parameter tuning, we utilised GridSearchCV to search for optimal model parameter sets that maximise cross-validation accuracy. The best-tuned estimators for logistic regression and support vector machine classifier all demonstrate high test accuracies of 100% (Figure 6a and 6b). While random forest classifier models display relatively low test accuracy at 90% (Figure 6d). The decision tree classifiers exhibit the poorest generalisation performance on the test set (validation score: 0.95, test set accuracy: 0.86), which may be attributed to the overfitting tendency of the tree model. As we noticed from the decision process of the tree model, it only uses 4 genes to separate the two sample types (Figure 6c).

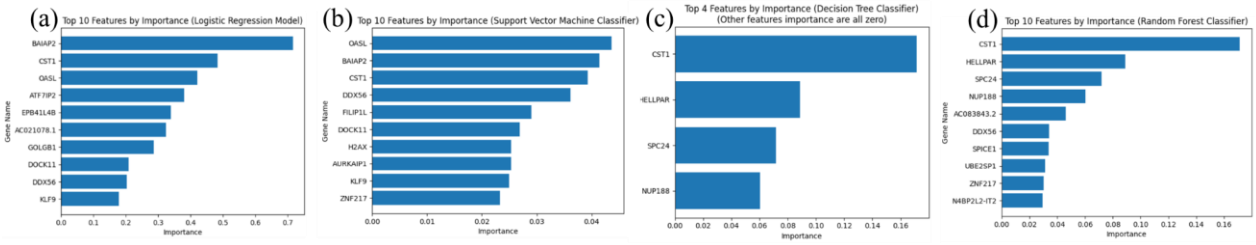

Figure 6. Horizontal bar plot for Top 10 genes and their weighted importance for each model. Top 10 Features by Importance (a) Logistic Regression (LR), (b) Support Vector Machine Classifier (SVC), (c) Decision Tree Classifier (DT), and (d) Random Forest Classifier (RF)

In Figure 6, we extracted and plotted the top 10 genes from each machine learning technique based on the weighted importance generated by the model. It is noteworthy that CST1 gene is listed with relatively high importance for all four models, which indicates the potential to serve as a GC biomarker. In our DEGs analysis, CST1 gene was similarly found to have been upregulated with a log2FC of 0.68. DDX gene, which appeared 3 out of 4 times in the top10 ranking for all our models, has a log2FC of -0.31 as well.

Based on the weighted importance obtained from the four models, we averaged the weighted importance value in the four classifier models to produce an average importance value and ranked each gene. CST1 has the highest average importance value and ranked the first, indicating it has the highest importance. CST1 was actively being studied and found to be related to GC. CST1 promoted GC cell migration and invasion through regulating transcription factor HOXC10 and activating the Wnt pathway (Kim et al., 2019 and Chen et al., 2021). CST1 encodes for cysteine protease inhibitors to halt the proteolytic activities of cysteine proteases and affect the modulation of immune responses, leading to the development of cancer (Choi et al., 2009). CST1 increases EMT by elevating fibronectin and vimentin expression and reducing E-cadherin and α-catenin expression, possessing similar EMT expression as in HOXD9. HOXC10 expression is upregulated in GC through DNA demethylation, and HOXC10 overexpression increases proliferation and migration of GC cells. CST1 was identified as one of the target genes regulated by HOXC10 in GC (Kim et al., 2019).

Conclusion

Using the RNA-seq dataset from Li et al., we discovered that the RNA-seq data was plausibly contaminated after our alignment and inspection. gDNA / rRNA contamination may lead to more False Negatives as observed in our comparison with Li's result. Due to time constraint, we were not able to re-align the RNA-seq data. We speculate that more DEGs would have a higher Log2FoldChange relative to our current analysis which would bring more meaningful results and biological discussion.

Additionally, we do not observe similar DEGs to Li et al. due to our stringent threshold (FDR < 0.05 and |log2FC| > 0.75) due to different methods employed. We performed DEGs analysis, GO and KEGG pathways analysis, and Differential Splicing Analysis using the RNA-seq data deposited in the GEO database (GSE210016). We found that the carboxylic acid transmembrane transporter activity biological process was enriched along with response to steroid hormones. There were also 14 alternative splicing events identified, which were mostly retained Introns. Subsequently, we selected genes from the TCGA-STAD dataset based on the DEG analysis to curate our machine learning dataset. TCGA-STAD contained only 71 identified in the DEGs dataset. Nevertheless, these 71 genes and their expression values were curated for machine learning model input. We employed four machine learning classifiers and found that CST1 gene has the highest weighted importance among these classifiers and has been studied extensively to be related with gastric cancer. Thus, we propose CST1 could be a potential biomarker for detecting gastric cancer malignant tumours.

Reference

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner*. United States.

Chen, S., Liu, Y., Zhang, K., & Chen, L. (2021). CST1 Promoted Gastric Cancer Migration and Invasion Through Activating Wnt Pathway. *Cancer management and research*, 13, 1901-1907. https://doi.org/10.2147/CMAR.S277770

Choi, E. H., Kim, J., Kim, J. H., Kim, S., Song, E. Y., Kim, J. W., Kim, S., Yeom, Y. I., Kim, Ik-Hwan., & Lee, H. Gu. (2009). Upregulation of the cysteine protease inhibitor, cystatin SN, contributes to cell proliferation and cathepsin inhibition in gastric cancer. *Clinica chimica acta*, 406(1), 45-51. https://doi.org/10.1016/j.cca.2009.05.008

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics,* 29 (1), 15-21. https://doi.org/10.1093/bioinformatics/bts635

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Presner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., & Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, 47(3), 199-208. https://doi.org/10.1038/ng.3192

Kahles, A., Ong, C. S., Zhong, Y., & Rätsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics*, 32(12), 1840–1847 (2016). https://doi.org/10.1093/bioinformatics/btw076

Kim, J., Bae, D., Kim, J. H., Song, K., Kim, Y. S., & Kim, S. (2019). HOXC10 overexpression promotes cell proliferation and migration in gastric cancer. *Oncology reports*, 42(1), 202-212. https://doi.org/10.3892/or.2019.7164

Li, J., Pei, M., Xiao, W., Liu, X., Hong, L., Yu, Z., Peng, Y., Zhang, J., Yang, P., Lin, J., Wu, X., Lin, Z., tang, W., zhi, F., Li, G., Xiang, L., Li, A., Liu, S., Chen, Y., & Wang, J. (2023). The HOXD9-mediated PAXIP1-AS1 regulates gastric cancer progression through PABPC1/PAK1 modulation. *Cell Death & Disease*, 14, 341. https://doi.org/10.1038/s41419-023-05862-5

Li, X., Zhang, P., Wang, H., & Ying, Y. (2022). Genes expressed at low levels raise false discovery rates in RNA samples contaminated with genomic DNA. *BMC Genomics*, 23(1), 554. https://doi.org/10.1186/s12864-022-08785-1

Liao, Y., Smyth, G. K., & Shi, W. (2014). Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liu, X., Zhao, J., Xue, L. et al. A comparison of transcriptome analysis methods with reference genome. BMC Genomics 23, 232 (2022). https://doi.org/10.1186/s12864-022-08465-0

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. https://doi.org/10.1186/s13059-014-0550-8

Machlowska, J., Maciejewski, R., & Sitarz, R. (2018). The pattern of signatures in gastric cancer prognosis. *International journal of molecular sciences*, 19(6), 1658. https://doi.org/10.3390/ijms19061658

Tan, I. B., Ng, I., Tai, WM., & Tan, P. (2012). Understanding the genetic basis of gastric cancer: recent advances. *Expert review of gastroenterology and hepatology*, 6(3), 335–341. https://doi.org/10.1586/egh.12.30

Wu, Y., Yang, X., Chen, Z., Tian, L., Jiang, G., Chen, F., Li, J., An, P., Lu, L., Luo, N., Du, J., Shan, H., Liu, H., & Wang, H. (2019). m6A-induced lncRNA RP11 triggers the dissemination of colorectal cancer cells via upregulation of Zeb1. *Molecular cancer*, 18(1), 87. https://doi.org/10.1186/s12943-019-1014-2

Zheng, L., Wu, C., Xi, P., Zhu, M., Zhang, L., Chen, S., Li, X., Gu, J., & Zheng, Y. (2014). The survival and the long-term trends of patients with gastric cancer in Shanghai, China. *BMC Cancer*, 14(1), 300. https://doi.org/10.1186/1471-2407-14-300

Contributions

Irfan contributed to the conception of the report scope. Irfan contributed to the raw data quality assurance, differences in DEGs results, and Differential Splicing Analysis subsections. Li Ruisi and Hu Yuhan contributed to the DEG analysis, and GO/KEGG analysis subsections. Olivia Loh and Yang Yizhuo contributed to the Relative Importance of DEGs as Biomarkers to Predict for Gastric Cancer subsection. All team members contributed to the writing of the report. Irfan and Olivia contributed to the final edits to the report.