Identification of Methylomic and Transcriptomic Biomarkers for Cancer Subtype Classification

Aashna Soni

Abstract

DNA methylation, or the addition of a methyl group to the fifth carbon of a cytosine or adenine nucleotide base, is considered to be the most important type of epigenetic modification in mammals due to its role in modulating the accessibility of chromatin to the cell's transcriptional machinery. Cancer cells display specific DNA methylation abnormalities, which can be summarized by the overall hypomethylation of the genome and hypermethylation of individual loci. Gene expression alterations are another hallmark of cancer: cancer cells are characterized by increased expression of oncogenes and decreased expression of tumor suppressor genes. These gene expression changes allow cancer cells to proliferate. Discovering novel methylomic and transcriptomic biomarkers has important clinical implications in cancer research, ranging from diagnosis, to treatment selection, to prediction of prognosis. Notably, identification of primary site in patients with cancers of unknown primary would allow these patients to receive more targeted, site-specific therapies that have much higher survival rates compared to conventional chemotherapy. In the present study, statistics and supervised learning techniques were used to identify gene expression and methylation biomarkers that are highly predictive of cancer primary site, or cancer subtype. These gene-level biomarkers were then investigated using pathway enrichment analysis, and the unsupervised learning technique of hierarchical clustering was used to identify groups of co-methylated and co-expressed genes that display different patterns between cancer subtypes. The results show that a very small number of DNA methylation and gene expression features are able to predict cancer subtype with greater than 90% accuracy. Additionally, developmental and nervous-system related pathways were found to be enriched using both the gene expression and methylation datasets. Furthermore, novel gene-level biomarkers such as cg25470758, CPNE2, TFAP2B, and CCDC18 were identified through this study, and hierarchical clustering revealed important pathway-level biomarkers such as osteoblast differentiation in thyroid cancer and axon injury response in brain cancer. The findings of this study demonstrate the utility of using methylation and gene expression features for cancer subtype determination. The novel gene-level and pathway-level biomarkers identified hold value for both the detection of primary site in patients with cancers of unknown primary, as well as for the development of personalized drugs that target specific transcriptomic and methylomic aberrations.

Introduction

Epigenetics is a field that focuses on understanding changes to the genetic material that do not involve direct changes to the nucleotide bases of DNA. Most epigenetic modifications fall into one of four categories—DNA methylation, histone modifications, noncoding RNA action, and mRNA methylation—out of which the first two have been most extensively studied. Acetylation and methylation of histones regulate transcription by altering the accessibility of chromatin to the cell's transcriptional machinery. Acetylation neutralizes histones' positive charge, making them adhere less strongly to the DNA, thus loosening the packaging of DNA and increasing transcription rates. Histone methylation can have either an activating or repressing effect on gene expression, depending on the site and degree of methylation (Jin et al., 2021). DNA methylation refers to the addition of a methyl group to the fifth carbon of a cytosine or adenine nucleotide base (Angermueller et al., 2017). When methyl groups are added to a region of DNA, they lead to the recruitment of chromatin-condensing proteins, which reduce the accessibility of chromatin to RNA polymerase and transcription factors, lowering the transcription levels of genes found within that region. Methyl groups are often added to CpG sites, which are regions where a guanine follows a cytosine in the DNA sequence (Angermueller et al., 2017). Stretches of DNA ranging from 300-3,000 base pairs in length where CpG sites occur very frequently are referred to as CpG islands (CGIs) (Janitz and Janitz, 2011). CGIs are located in or near approximately 40% of mammalian promoters, and play an important role in regulating gene expression by serving as binding sites for proteins involved in chromatin modification, known as CGI-reading proteins (Hughes et al., 2023).

Normal cells are characterized by distinct methylation signatures. Methylation marks differ between cells of different tissues, as methylation patterns determined during embryonic development help restrict a cell's differentiation potential into a specific lineage (Messerschmidt et al., 2014). Some common characteristics of normal cell methylomes can, however, be identified: CGIs remain hypomethylated and repetitive DNA sequences and gene bodies remain hypermethylated (Li et al., 2023). Modifications to these methylation patterns occur early on in the process of carcinogenesis and are relatively stable over time (Li et al., 2023).Tumor cells often display hypomethylated genomes and hypermethylation of individual loci, the latter often resulting in the inhibition of tumor suppressor genes (Li et al., 2023). Past studies have identified genes that are hypermethylated in different cancers—for instance, TFAP2A, a tumor suppressor gene that codes for a transcription factor, was shown to be commonly methylated in large B-cell lymphoma, renal cell carcinoma, and breast cancer (Dunwell et al., 2010). Such DNA methylation abnormalities can result from mutations to enzymes, like IDH1 and IDH2, which are involved in pathways that initiate and maintain DNA methylation patterns (Jin et al., 2021). The downstream products of IDH1 and IDH2 help activate histone demethylases that remove methyl groups from DNA. Gain-of-function IDH mutations result in the inhibition of histone demethylases, resulting in hypermethylation that can lead to cancer (Jin et al., 2021). Epigenetic drugs such as IDH inhibitors are a key focus of personalized therapy today as they target specific genetic abnormalities in epigenetic pathways (Jin et al., 2021).

Discovering novel epigenetic biomarkers holds significant clinical value. Notably, identification of primary site for patients with cancers of unknown primary is one focus of current cancer research. Cancers of unknown primary are metastatic cancers in which the primary site remains occult for unknown reasons (Moran et al., 2016). These cancers have very high mortality rates, and determination of primary site for patients with these cancers would allow caregivers to screen for site-specific mutations and administer site-specific therapies that have significantly better prognostic outcomes compared to trial-and-error chemotherapy (Moran et al., 2016).

In this age of multi-omics, characterized by the abundance of genomic, transcriptomic, methylomic, and proteomic data, statistics and machine-learning techniques have proven to be invaluable in biomarker discovery (Zhang et al., 2021). The primary challenge in processing methylation data, in particular, lies in the large number of features present. To reduce the number of input features for classification tasks, previous studies have used statistical tests like analysis of variance (ANOVA), univariate and multivariate regression analysis with regularization, random forests, and 2 sample t-tests (Adorján et al., 2002; Liu et al., 2019; Fan et al., 2019; Peng et al., 2020). Studies have also differed in the size of the features used: some have used individual CpG dinucleotides as features, while others have used promoters or entire genes as features (Adorján et al., 2002; Fan et al., 2019; Liu et al., 2019; Peng et al., 2020; Choi et al., 2023). Following data preprocessing, various machine learning and statistical techniques have been applied to methylation data, either to predict the presence or absence of cancer or to predict a patient's cancer subtype based on their epigenetic profile (Adorján et al., 2002; Liu et al., 2019; Fan et al., 2019; Peng et al., 2020). The primary algorithms that have been used for these classification tasks are support vector machines (SVMs), random forests, simple feed-forward neural networks, and logistic regression (Adorján et al., 2002; Liu et al., 2019; Fan et al., 2019; Peng et al., 2020). Additionally, some researchers have used unsupervised or semi-supervised learning techniques to gain insight from methylation data, thus being able to leverage the large quantities of unlabeled data available (Adorján et al., 2002, Choi et al., 2023). For instance, Adorján et al. applied hierarchical clustering to discover tumor subclasses based on differential methylation patterns, and Choi et al. used semi-supervised learning to create a cancer subtype classifier (Adorján et al., 2002; Choi et al., 2023).

In addition to machine learning, network analysis has been successfully applied to gain pathway-level insights from methylation data: Cui et al. used the methylation levels of methylation-related differentially-expressed genes (mrDEGs) to create a network that depicts clusters that constitute functional modules as well as identify crosstalk between these modules (Cui et al., 2019). Past studies have also integrated multiple data types to identify biomarkers: Fan et al. used mutation, DNA methylation, and gene expression data to identify genes that are mutated, differentially methylated, and differentially expressed between normal and cancerous samples (Fan et al., 2019), and Peng et al. used gene expression and methylation data of differentially-expressed genes to select genes that display strong correlations between gene expression and methylation patterns (Peng et al., 2020). Finally, studies in which pathway enrichment analyses were carried out following machine learning and network analysis were able to identify important cancer-related pathways that differentially methylated, differentially expressed, and mutated genes are implicated in, such as the Wnt pathway, NF-kB signaling, and axonal guidance signaling (Fan et al., 2019; Peng et al., 2020). These nervous system-related and developmental pathways were also found to be of importance in the present study.

While past studies have made significant strides in drawing insights from DNA methylation data, few biomarkers have been successfully implemented in the clinic (Koch et al., 2018). Reasons for this include a lack of thorough investigation into the genomic context and location of the methylation sites, which are needed to select optimal biomarkers (Koch et al., 2018). More can be done to annotate methylation sites as well as incorporate other data types, like mutation and gene expression data, to identify meaningful correlations and increase the strength of the biomarkers discovered. For example, understanding correlations between gene expression and mutation data could be highly informative: Li et al. suggest potential links between genome-wide hypomethylation in cancer cells, which reduces the stability of chromosomes, and the induction of mutations (Li et al., 2023). In the future, exploring the variety of functional regions in the genome for biomarkers could help discover novel insights into how hypermethylation or hypomethylation of different sites contributes to cancer.

Gene expression quantification is another key metric that displays altered patterns in cancerous tissue. In particular, tumor suppressor genes, or genes which code for proteins that negatively regulate the cell cycle, are largely underexpressed in cancer cells, and oncogenes, or mutant proto-oncogenes which code for proteins that overly stimulate the cell cycle, are largely overexpressed in cancer cells (Schriman et al., 2017). Traditional machine learning techniques, such as SVMs, random forests, and Naïve Bayes, have long been used for gene expression-based cancer detection or cancer subtype differentiation, performing at greater than 80% accuracy (Segal et al., 2003; Hijazi et al., 2013; Ram et al., 2017; Zhang et al., 2018; Yuan et al., 2020; Alharbi and Vakanski, 2023). A significant limitation of traditional machine-learning approaches is that they are much more reliant on proper feature engineering and processing of data compared to deep-learning approaches, which are better at identifying cancer-subtype-specific signatures from raw data (Alharbi and Vakanski, 2023). Deep learning techniques that have been successfully used to analyze gene expression data include multilayer perceptrons, convolutional neural networks, recurrent neural networks, and graph neural networks (Sathe et al., 2019; Koumakis et al., 2020; Zhu et al., 2020; Gunavathi et al., 2020).

This study aims to holistically investigate the biomarker discovery pipeline, from feature selection, classification, and clustering to identify potentially significant biomarkers, to pathway enrichment analysis to analyze these biomarkers in a biological context. The two data types used in this study are DNA methylation and gene expression data. For both data types, ANOVA and logistic regression with elastic net were used to identify the top gene features. From there, an XGBoost multiclass classifier was trained for cancer subtype prediction, and pathway enrichment analysis was conducted to understand the biological implications of the identified differentially-methylated and differentially-expressed genes. Finally, hierarchical clustering was used to identify groups of co-methylated and co-expressed genes that show different patterns of methylation or expression between the cancer subtypes. The findings of this study demonstrate that methylation and gene-expression biomarkers are highly predictive of cancer subtype. Novel gene-level and pathway-level biomarkers were identified, and gene expression biomarkers were found to show greater cancer subtype specificity compared to methylation biomarkers

Results

Novel DNA Methylation Gene-level and Pathway-level Biomarkers

To identify differentially-methylated genes that best distinguish the cancer subtypes under study, the ANOVA and elastic-net logistic regression feature selection methods were applied to the original dataset. Using ANOVA, 69,391 CpG sites with a p-value of 0.0 were identified. The top 1,400 CpG sites with a p-value of 0.0 were subjected to pairwise correlation analysis with a threshold of greater than 0.7 or less than -0.7 to reduce multicollinearity in the dataset. 939 highly-correlated features were identified and removed, leaving 461 features. These 461 features served as input features for a logistic regression with elastic net classifier. With an alpha value of 0.01 and an L1 ratio of 0.5, 160 features that appeared in all 3 bootstrapping runs of logistic regression were selected. On the test dataset, this multi-class logistic regression classifier performed at 100% accuracy. The confusion matrix for this model is shown in Table 1.

To create a cancer subtype prediction model, an XGBoost classifier with recursive feature elimination was trained on the methylation dataset with the 160 input features identified from logistic regression with elastic net. Recursive feature elimination identified the top 6 features. The XGBoost model trained on these features performed at 91% accuracy. The confusion matrix for this model is shown in Table 2, and the gene names for the 6 chosen features are shown in Table 3.

The 160 CpG sites identified through elastic-net logistic regression were annotated using Infinium Annotation (Zhou et al., 2017) to identify the genes they are located in. 131 unique genes were obtained from this annotation, and they were analyzed using Metascape, a pathway-enrichment software (Zhou et al., 2019), to understand their roles in biological pathways. The top pathways that these genes were found to be implicated in are shown in Table 9. Furthermore, a network was created using a subset of the genes inputted into the pathway enrichment software in order to identify functional gene modules. This network is provided in the supplementary files. The cluster found to be most highly enriched in the methylation network was endocrine factor-regulated processes (29 nodes). The most significant subclusters within this cluster are related to insulin response, thermogenesis, negative regulation of oxidative stress-induced cell death, protein kinase B signaling, and Fragile X syndrome.

Hierarchical clustering was applied to the methylation dataset with the 160 features identified from logistic regression with elastic net. The clustermap produced is shown in Figure 1. Six row clusters, which represent the 6 cancer subtypes, and 6 column clusters, which represent groups of co-methylated genes, were identified. Row clusters 1 through 6 represent the brain, thyroid gland, kidney, prostate gland, breast, and bronchus and lung primary sites, respectively. The column clusters were analyzed for pathway enrichment. The most heavily enriched pathway for each cluster was found, and the results are presented in Table 4. Cluster 5 did not show any enrichment.

Novel Gene Expression Gene-level and Pathway-level Biomarkers

First, ANOVA and elastic-net logistic regression were applied to the gene expression dataset. 6,981 genes with a p-value of 0.0 were identified. From there, the top 800 genes with a p-value of 0.0 were determined. Multicollinearity in the dataset was reduced by conducting pairwise correlation analysis with a threshold of greater than 0.7 or less than -0.7. 365 highly-correlated features were identified and removed, leaving 435 features. These 435 genes served as input features for a logistic regression with elastic net classifier. With an alpha value of 0.01 and an L1 ratio of 0.5, the top 160 features were obtained. The logistic regression model performed at 99.9% accuracy on the test dataset. The confusion matrix for this model is shown in Table 5.

An XGBoost model with recursive feature elimination was trained on the gene expression dataset with the 160 input features identified from elastic-net logistic regression. Recursive feature elimination identified the top 6 features. The XGBoost model trained on these features performed at an accuracy of 98.4%. The confusion matrix for this model is shown in Table 6, and the 6 features identified are shown in Table 7.

The 160 genes identified through elastic-net logistic regression were subjected to pathway-enrichment analysis via Metascape to understand their roles in biological pathways. Table 9 depicts the top 10 pathways these genes were found to be implicated in. From there, a network was created using a subset of the genes inputted into the pathway enrichment software in order to identify functional gene clusters. This network is provided in the supplementary files. The cluster found to be most heavily enriched was related to system development morphogenesis (42 nodes). The most significant subclusters within this cluster are related to RNA polymerase II transcription, response to testosterone, embryonic forelimb morphogenesis, limb morphogenesis, and appendage development.

Hierarchical clustering was applied to the gene expression dataset with the 160 features identified from logistic regression with elastic net. The resulting clustermap is shown in Figure 2. Six row clusters, which represent the 6 cancer subtypes, and 6 column clusters, which represent groups of co-expressed genes, were identified. Row clusters 1 through 6 represent the prostate gland, breast, bronchus and lung, brain, thyroid gland, and kidney primary sites, respectively. The column clusters were analyzed for pathway enrichment. The most heavily enriched pathways for all clusters are shown in Table 8.

Figures and Legends

Table 1. Confusion Matrix for DNA Methylation Elastic Net Classifier.

The accuracy of the DNA methylation logistic regression with elastic net regularization classifier, averaged across the 3 bootstrapping runs, was 100% on the test dataset.

	Brain	Breast	Bronchus and Lung	Kidney	Prostate Gland	Thyroid Gland
Brain	60	0	0	0	0	0
Breast	0	77	0	0	0	0
Bronchus and Lung	0	0	82	0	0	0
Kidney	0	0	0	66	0	0
Prostate Gland	0	0	0	0	50	0
Thyroid Gland	0	0	0	0	0	51

Table 2. Confusion Matrix for XGBoost Model Trained on Top 6 DNA Methylation Features.

The accuracy of the DNA methylation XGBoost classifier trained on the top 6 features was 91% on the test dataset.

	Brain	Breast	Bronchus and Lung	Kidney	Prostate Gland	Thyroid Gland
Brain	59	0	0	1	0	0
Breast	1	61	11	3	1	0
Bronchus and Lung	0	5	74	2	0	1
Kidney	1	1	2	60	2	0
Prostate Gland	0	0	0	0	50	0
Thyroid Gland	0	0	0	2	0	49

Table 3. Top 6 DNA Methylation Features.

Shown below are the 6 DNA methylation features identified from recursive feature elimination, the genes they are located within, and the pathway(s) the genes are involved in. Pathway information was retrieved from the GeneCards database.

CpG Site	Gene	GeneCards Pathway
cg06880930	CPNE2	Intracellular calcium signaling
cg10333400	AL023882.1	Uncharacterized
cg10970500	PLEKHG5	NFKB1 signaling
cg25470758	Near LOC105378621 and LINC01778	Uncharacterized lncRNAs
cg18768784	AC112504.2;GRK7	Deactivation of cone opsins in the retina
cg03579904	KLHDC4	Uncharacterized

Table 4. Top Enriched Pathways from DNA Methylation Clustermap.

Hierarchical clustering was used to divide the DNA methylation dataset with 160 features into groups of co-methylated genes. Six distinct gene clusters were identified. The enrichment results for the individual gene clusters are shown in the table.

Cluster	Pathway
1	Roof of mouth development
2	Neuronal system
3	Osteoblast differentiation
4	Rab regulation of trafficking
5	No enrichment
6	Positive regulation of cold-induced thermogenesis

Table 5. Confusion Matrix for Gene Expression Elastic Net Classifier.

The accuracy of the gene expression logistic regression with elastic net regularization classifier, averaged across the 3 bootstrapping runs, was 99.9% on the test dataset.

	Brain	Breast	Bronchus and Lung	Kidney	Prostate Gland	Thyroid Gland
Brain	59.67	0	0.33	0	0	0
Breast	0	77	0	0	0	0
Bronchus and Lung	0	0	82	0	0	0
Kidney	0	0	0	66	0	0
Prostate Gland	0	0	0	0	50	0
Thyroid Gland	0	0	0	0	0	51

Table 6. Confusion Matrix for XGBoost Model Trained on Top 6 Gene Expression Features.

The accuracy of the gene expression XGBoost model trained on the top 6 features was 98.4% on the test dataset.

	Brain	Breast	Bronchus and Lung	Kidney	Prostate Gland	Thyroid Gland
Brain	59	0	1	0	0	0
Breast	0	77	0	0	0	0
Bronchus and Lung	0	2	79	0	1	0
Kidney	1	0	1	65	0	0
Prostate Gland	0	1	0	0	49	0
Thyroid Gland	0	0	0	0	0	51

Table 7. Top 6 Gene Expression Features.

Shown below are the 6 gene features identified from recursive feature elimination and the pathway(s) they are involved in. Pathway information was retrieved from the GeneCards database.

Gene	**GeneCards Pathway**
CCDC198, protein-coding	Uncharacterized
SCGB2A2, protein-coding	Androgen receptor signaling
ROS1, protein-coding	Growth and differentiation
TG, protein-coding	Thyroid hormone production
MLC1, protein-coding	MAPK-ERK signaling, colorectal cancer metastasis
HOXA13, protein-coding	Regulation of morphogenesis and differentiation

Table 8. Top Enriched Pathways from Gene Expression Clustermap.

Hierarchical clustering was used to divide the gene expression dataset with 160 features into groups of co-expressed genes. Six distinct gene clusters were identified. The enrichment results for the individual gene clusters are shown in the table.

Cluster	**Pathway**
1	Mammary gland epithelium development
2	Embryonic morphogenesis
3	Thyroid hormone metabolic process
4	Nuclear receptors meta-pathway
5	Vascular transport
6	Response to axon injury

Table 9. Common Enriched Pathways Between DNA Methylation and Gene Expression Datasets.

Detailed below are the top 10 pathways found to be enriched in the DNA methylation and gene expression datasets, irrespective of cancer subtype. The number of nodes was used to rank the clusters by importance. In both workflows, enrichment is shown in developmental pathways, such as bone morphogenesis and epithelial cell development, and nervous-system related pathways, such as neuron chemotaxis locomotion and neuron apoptosis.

DNA Methylation		Gene Expression	
Pathway	**No. nodes**	**Pathway**	**No. nodes**
Endocrine factor regulated	29	System development morphogenesis	42
Alpha beta cell	21	Vascular transport metal	20
Regulation nervous system	12	Mammary gland epithelium	18
Bone morphogenesis positive	10	cAMP signaling pathway	10
Neuron chemotaxis locomotion	8	Epithelial cell morphogenesis	10
Sensory organ morphogenesis	8	Mucin type glycan	10
Negative regulation cell	6	Regulation myelination	10
Inositol phosphate metabolism	4	Thyroid hormone synthesis	9
Kidney development	4	Neuron apoptotic process	6

Heart development	2	Epithelial cell differentiation	5

Figure 1. DNA Methylation Clustermap.

The column clusters, which represent groups of genes that are co-methylated, are labeled 1-6. The row clusters are labeled with the cancer subtype overrepresented in the cluster.

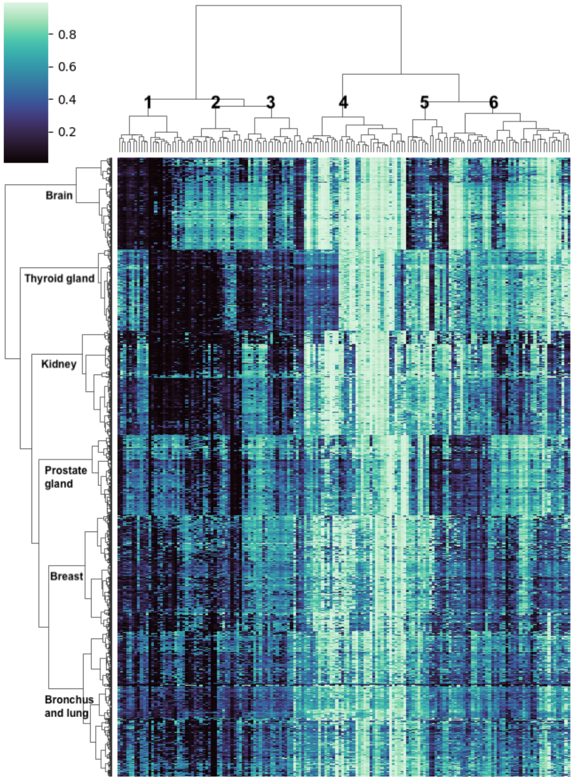


Figure 2. Gene Expression Clustermap.

The column clusters, which represent groups of genes that are co-expressed, are labeled 1-6. The row clusters are labeled with the cancer subtype overrepresented in the cluster.

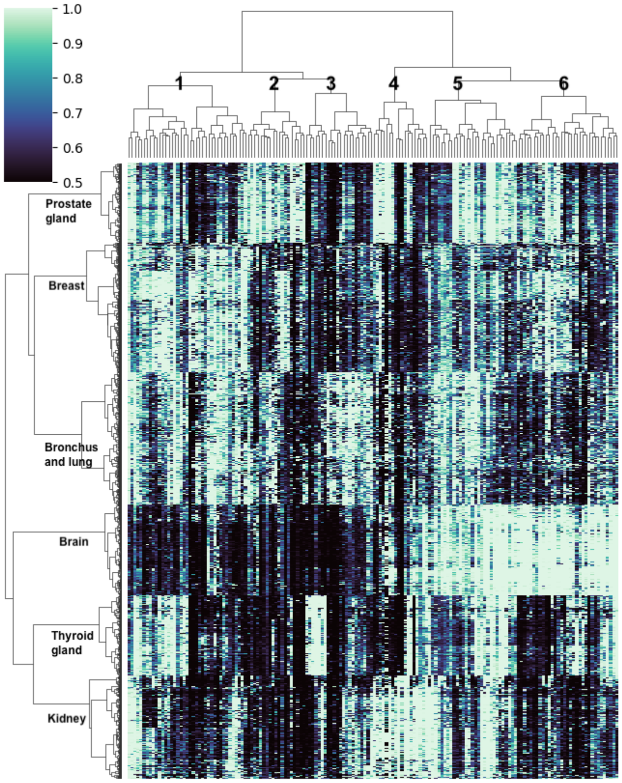


Figure 3. Recursive Feature Elimination Gene Expression Biomarkers.

Shown are the expression levels of the top 6 gene expression features across the 6 cancer subtypes. Error bars represent +/- 2 standard errors of the mean.

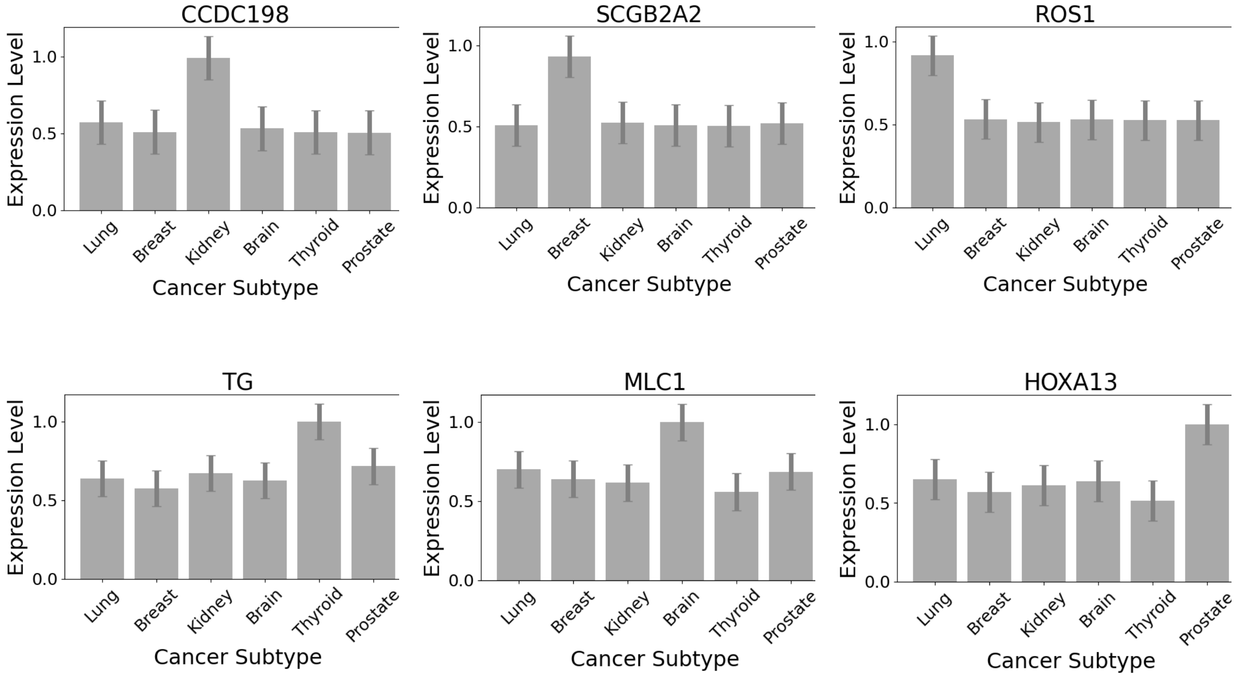
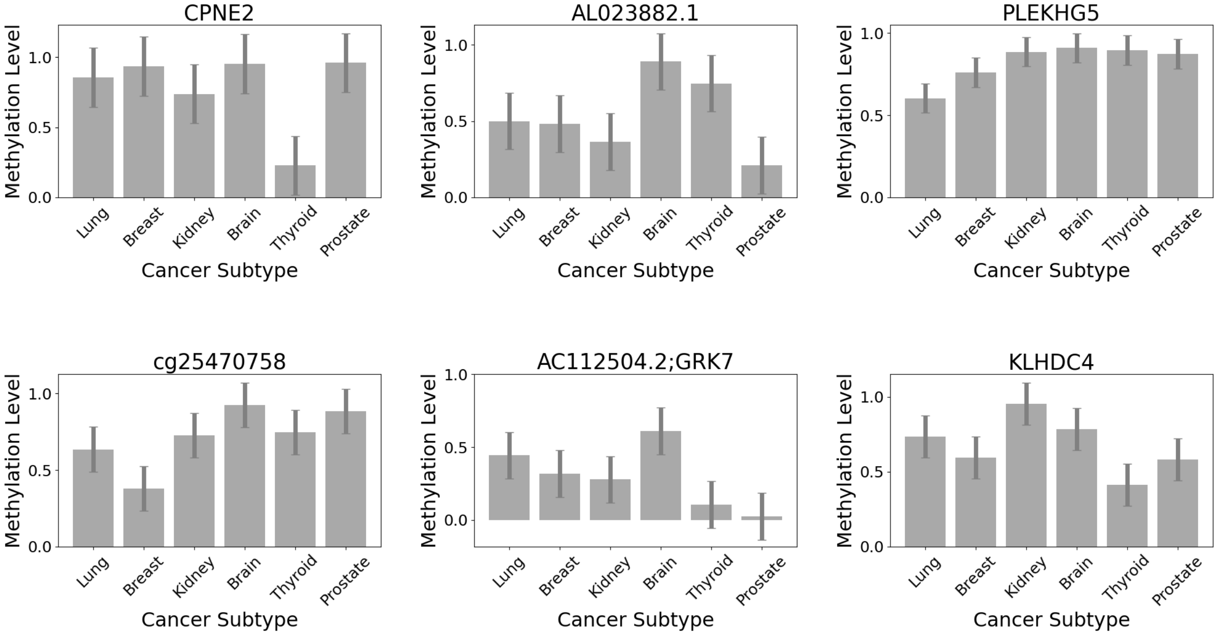


Figure 4. Recursive Feature Elimination DNA Methylation Biomarkers.

Shown are the methylation levels of the top 6 DNA methylation features across the 6 cancer subtypes. Error bars represent +/- 2 standard errors of the mean.



Discussion

To identify DNA methylation and gene expression biomarkers that best distinguish cancer subtypes, ANOVA and logistic regression with elastic net were first used as feature selection techniques. From there, XGBoost classification with recursive feature elimination was utilized for the task of primary site prediction. Finally, the most highly predictive genes were investigated using pathway enrichment analysis to understand their potential roles in carcinogenesis, and hierarchical clustering was used to identify both gene-level and pathway-level biomarkers that differentiate cancer subtypes.

Pan-Cancer Pathway Biomarkers

Pathway enrichment analysis on the DNA methylation and gene expression datasets yielded many similar classes of pathways to be enriched, as shown in Table 9, indicating these pathways to be significant across the cancer subtypes studied. Both workflows yielded enrichment in developmental pathways, such as kidney development and system development morphogenesis, and nervous-system related pathways, such as regulation of myelination and neuron apoptosis. These results are consistent with the literature—developmental pathways such as Wnt and Hedgehog signaling have been known to be re-activated in cancerous cells through either genetic or epigenetic alterations (Aiello and Stanger, 2016). The nervous system is also understood to be intricately linked to tumors, as nerves are a critical component of the microenvironment of tumor cells (Wang et al., 2021). Through the process of perineural invasion, tumor cells can migrate to other locations, resulting in aggressive metastases and poor prognostic outcomes (Wang et al., 2021).

The enrichment results from the DNA methylation and gene expression datasets were further investigated using interactive network visualization software to identify functional gene modules. The densest region of the methylation network was functionally annotated to be related to endocrine factor-regulated processes (29 nodes), and the densest region of the gene expression network was annotated to be related to system development morphogenesis (42 nodes). These results are supported by the literature, as endocrinal dysfunction and epithelial morphogenesis are broadly known to play a role in the onset of many cancers (Gray et al., 2011; Jiang et al., 2020). Interestingly, Fragile X syndrome, an intellectual disability characterized by silencing of the Fragile X Messenger Ribonucleoprotein gene (FMR1) (Hagerman et al., 2017), was identified to be an important pathway within the cluster of endocrine factor-regulated pathways in the present study, with a within-group enrichment of 14.008. Though Fragile X syndrome's link to cancer has been investigated in a population-based study, which demonstrated that individuals with this syndrome display a slightly decreased risk of cancer malignancy, this connection has been declared to need additional research (Sund and Pukkala, 2008). The results of the present study suggest that the pathway of genes that are implicated in Fragile X syndrome (AKT1, TSC2, AGO2, and DLGAP3) should be investigated further as potential cancer biomarkers.

High Predictive Power of 6 Genes and Novel Gene-level Biomarkers

Recursive feature elimination was used for the XGBoost models trained on the differentially-methylated and differentially-expressed genes to find the 6 genes most predictive of cancer subtype. Both models achieved greater than 90% accuracy, demonstrating that cancer subtype can be predicted with very high accuracy with a relatively small number of input features.

The differentially-expressed genes selected from recursive feature elimination were CCDC198, SCGB2A2, ROS1, TG, MLC1, and HOXA13. As shown in Figure 3, these genes have very high cancer subtype specificity, as each gene is indicative of a single cancer subtype. While SCGB2A2, ROS1, TG, HOXA13, and MLC1 are known biomarkers in breast, lung, thyroid, prostate, and brain cancers, respectively (Shi et al., 2004; Bubendorf et al., 2016; Prpić et al., 2018; Dong et al., 2017; Lattier et al., 2020), CCDC198, which was found to be overexpressed in the kidney, is a novel biomarker identified in this study. CCDC198 is a membrane-bound protein that is involved in kidney metabolic processes (Petersen et al., 2023). Petersen et al. suggested its potential role in cancer progression, though qualified that whether CCDC198 is implicated as a driver or passenger mutation, or whether it is involved in promoting or suppressing tumorigenesis, is an area that requires further investigation (Petersen et al., 2023). The present study suggests that CCDC198 is likely implicated as a driver gene in kidney cancer progression.

Using recursive feature elimination, the top 6 differentially-methylated features identified were CPNE2, AL023882.1, PLEKHG5, cg25470758, AC112504.2;GRK7, and KLHDC4. The relative methylation levels of these genes across the cancer subtypes are shown in Figure 4. CPNE2, a membrane protein which is expressed in almost all mammalian tissues (Tang et al., 2021), was found to be hypomethylated in the thyroid. The CPNE family of proteins is involved in multiple signaling pathways that direct the activities of numerous effector proteins (Tang et al., 2021). The present study demonstrates an important role for CPNE2 in cancer subtype differentiation among the 6 cancer subtypes under study. Additionally, AL023882.1, a lncRNA, is one of 11 lncRNAs that was identified to be a prognostic indicator in breast cancer by Yu et al. (Yu et al., 2023). The present study shows that AL023882.1 exhibits hypermethylation in the brain and thyroid gland, intermediate methylation in the lung, breast, and kidney, and hypomethylation in the prostate gland. However, overlap in error bars reveals that AL023882.1 does not show clear cancer subtype specificity, despite being a highly predictive feature in the XGBoost classifier. Finally, CpG site cg25470758, which was found to be hypomethylated in the breast compared to the other primary sites, was a novel predictive biomarker identified through this study. This CpG site was visualized using the UCSC Genome Browser (Kent et al., 2002), and was found to be located on chromosome 1 between the base pairs 30,807,296 and 30,807,297. cg25470758 is located 24,218 base pairs away from LOC105378621 and 16,392 base pairs away from LINC01778, the nearest genes. These genes are lncRNAs whose functions are yet to be determined, and it is possible that this CpG site is located within a regulatory region of these genes. The functions of these lncRNA genes should be explored further, as the present study identifies that they may play an important role in cancer subtype differentiation. This result demonstrates the utility of analyzing CpG sites throughout the genome, including those located far from known genes or in regulatory regions of genes, as their differential methylation may have yet unknown roles in carcinogenesis. In the future, increasing the range of CpG sites analyzed throughout the genome could yield additional novel biomarkers that are highly predictive of cancer subtype.

Comparing the biomarkers identified through the DNA methylation and gene expression workflows, the differentially-expressed genes were found to display stronger cancer subtype specificity compared to the differentially-methylated genes. In other words, while the gene expression features identified from recursive feature elimination were all clearly overexpressed in a given cancer subtype compared to the others, the differentially-methylated features did not point to exactly one cancer subtype in which the gene was significantly hypomethylated or hypermethylated. Rather, these features separated the cancer subtypes into 2 broad groups: those in which the gene was hypomethylated, and those in which the gene was hypermethylated. Thus, it is shown that gene expression data paints a clearer picture of cancer subtype differentiation compared to DNA methylation data.

Cancer Subtype-Specific Pathway Biomarkers

Hierarchical clustering was applied to the DNA methylation and gene expression datasets to identify groups of co-methylated and co-expressed genes that display different patterns between different cancer subtypes. For both the gene expression and DNA methylation datasets, the genes were grouped into 6 gene sets, or clusters, and pathway enrichment analysis was conducted on each gene set to understand which pathways are implicated in different cancer subtypes. The clustermaps produced from hierarchical clustering are shown in Figures 1 and 2. Many of the clustering results are highly consistent with expectation. For example, in the gene expression clustermap, cluster 1 genes, which are involved in mammary gland epithelium development, were overexpressed in the breast, and cluster 6 genes, which are involved in response to axon injury, were overexpressed in the brain. Some interesting results are detailed in the following paragraphs.

In the gene expression dataset, cluster 5 genes, which are enriched in vascular transport, were found to be overexpressed in the brain relative to the other primary sites. Two types of vascular transport are particularly important to brain function: transport across the blood-brain barrier and transport across the blood-cerebrospinal fluid barrier (Engelhardt and Sorokin, 2009). The literature support a role for the disruption to brain vasculature in the progression of glioblastoma, a highly aggressive and angiogenic brain cancer: disruption to the blood vessels leads to leakage of fluid and sub-optimal transport of nutrients to the brain (Guyon et al., 2021). Due to the relationship between angiogenesis, which is central to cancer progression, and vascular transport, overexpression of vascular transport genes may serve as a key pathway-level indicator of brain cancer. Additionally, cluster 4 genes, which are enriched in the nuclear receptors meta-pathway, were found to be generally overexpressed in the kidney. Nuclear receptors are intracellular steroid and thyroid hormone receptors which, upon ligand-binding, change shape and function as transcription factors (Data ref: https://www.wikipathways.org/instance/WP2882). Past research has suggested that the nuclear receptors meta-pathway is implicated in clear cell renal cell carcinoma (Ding et al., 2022). The present study builds upon the results obtained by Ding et al. by demonstrating, via computational analysis, that the genes involved in the nuclear receptors meta-pathway can serve as a pathway-level biomarker for kidney cancer.

Interestingly, in the methylation dataset, cluster 1 genes, which are enriched in roof of mouth development, were found to be hypomethylated in the brain relative to the other cancer subtypes. No literature support for such a link was found. The present study thus suggests that the roof of mouth development pathway may be directly implicated in brain cancer, or that the genes involved in roof of mouth development may have unrelated functions relevant to the brain and thus may be involved in brain cancer. Cluster 3 genes, which were enriched in osteoblast differentiation, were found to be hypomethylated in the thyroid. Past research has shown that the release of thyroid hormone stimulates the production of insulin-like growth factor 1 (IGF-1), which is positively involved in a pathway that promotes osteoblast proliferation (Deng et al., 2021). The role of thyroid hormone in regulating the development of bone tissue (Deng et al.,

2021) may be relevant to understanding thyroid cancer, as genes involved in osteoblast differentiation may also be related to the onset of thyroid cancer through this thyroid-mediated bone development pathway. Finally, cluster 4 genes, which were enriched in Rab regulation of trafficking, were found to be generally hypomethylated in the brain. The Rab family of proteins are small GTPases that play important regulatory roles in the vesicular transport of macromolecules in the cell (McCaffrey and Lindsay, 2013). It is possible that, in brain cancers, differential methylation to genes coding for Rab proteins could result in unconventional protein secretion (UPS), in which proteins that are not tagged with the necessary signal peptide are secreted from the cell without undergoing the traditional membrane trafficking pathway (Iglesia et al., 2022). Past research by Iglesia et al. shows that UPS may be implicated in gliomas, a highly aggressive brain cancer, by helping cancer cells develop resistance to therapies (Iglesia et al., 2022). The present study suggests that deregulation of Rab-mediated membrane trafficking through hypomethylation of genes that code for Rab proteins could be a strong indicator of brain cancer.

Cluster 5 in the DNA methylation clustermap was the only cluster that did not show any enrichment. This is likely due to the fact that cluster 5 contains only 15 genes, which is an insufficient number in order to gain statistically significant enrichment. Nevertheless, cluster 5 genes were shown to be significantly hypomethylated in the brain relative to the other cancer subtypes. Thus, further investigation into the functions of these genes and CpG sites (CKB, BCL11B, cg09112081, TMIE, AC109462.1;IRX6, cg00152577, GSE1, SHE;TDRD10, AL691442.1, WNT3A, THRB, ITPKA, and BICDL2) should be conducted in order to elucidate their roles in brain cancer.

Large Pool of Gene-Level Biomarkers

In addition to the 6 gene expression and DNA methylation features identified from recursive feature elimination, clustering analysis revealed many other features that are also highly indicative of cancer subtype. This is due to the fact that the top 6 genes identified through recursive feature elimination are co-methylated or co-expressed with other genes that are part of the same pathways. For example, for gene expression, clustering revealed TFAP2B and EPYC to be predictive of breast cancer, RSPH6A and IYD to be predictive of thyroid cancer, and SIM1, TRHDE, and SMIM24 to be predictive of kidney cancer. Similarly, clustering on the methylation dataset revealed TOX3, REC8, and SPSB4 to be predictive of brain cancer, LINCO1137 and TSC2 to be predictive of prostate cancer, ARHGEF10L to be predictive of breast cancer, ERICH1 to be predictive of kidney cancer, GILS2;PAM16 to be predictive of thyroid cancer, and AL008733.1;PRMD16 to be predictive of lung cancer. The predictive power of these genes can be shown through the comparison of their methylation or expression levels between the cancer subtypes, which are provided as bar plots in the supplementary files. Overall, these results demonstrate that, beyond the 6 most highly predictive genes, additional genes that are involved in the same pathways can also serve as highly predictive cancer subtype biomarkers.

Future Directions

In the future, the results of this study can be built upon through the analysis of additional cancer primary sites to identify both pan-cancer and additional cancer-subtype-specific biomarkers. Additionally, the incorporation of larger numbers of patients with metastatic cancers in the original dataset could help to identify methylation and gene expression patterns relevant to cancer cell invasion, which is particularly important to understand in the context of cancers of unknown primary. Finally, the DNA methylation and gene expression results of the present study can be analyzed in conjunction with DNA mutation data to understand correlations between mutations and both transcriptomic and methylomic abnormalities, which may help to elucidate the genetic causes of these aberrations that could serve as targets for personalized therapies.

Materials and Methods

Data Collection and Processing

The DNA methylation and gene expression data for this study were acquired from The Cancer Genome Atlas (TCGA) (Data ref: https://cancergenome.nih.gov/). The DNA methylation data was extracted using the Illumina 450k Methylation assay, and the gene expression quantification data was obtained using the STAR - Counts RNA-Seq workflow. Both data types were collected from the same 3,844 patients, whose cancers originated in 1 of 6 different primary sites: bronchus and lung, breast, brain, kidney, prostate gland, and thyroid gland. The samples differed in the type of tissue (primary tumor tissue, recurrent tissue, and metastatic tissue) used to collect either the methylation or gene expression data. In both the methylation and gene expression datasets, 3,801 samples were taken from primary tumors, 29 from recurrent tumors, and 14 from metastatic tissues. The cancer cases in the dataset were split approximately evenly between males and females (97 : 100 ratio, males to females). 10% of the total data was reserved for testing, and similar cancer subtype distributions were maintained in the training and test sets. The training set distribution of cancer subtypes was 0.212 : 0.199: 0.171 : 0.155 : 0.132 : 0.129, and the corresponding test set distribution was 0.215 : 0.204 : 0.172 : 0.149 : 0.132 : 0.129 (bronchus and lung : breast : kidney : brain : thyroid gland : prostate gland).

In the raw DNA methylation data, each training example had 485,577 features, which represent the methylation levels (beta values) of select CpG sites located throughout the genome. Beta values range from 0 (totally unmethylated) to 1 (totally methylated). Features for which more than 1% of the samples in the training data had missing values were removed, and any remaining missing values were imputed using median imputation. The columns removed from the training dataset were also removed from the test dataset, and any remaining missing values in the test dataset were imputed using the medians from the aggregated training and test datasets. The data imputation process reduced the number of CpG site features from 485,577 to 361,038.

In the raw gene expression data, each training example had 60,660 features, which represent the normalized gene expression levels (in transcripts per million) for select gene isoforms throughout the genome, each identified by a unique Ensemble ID. Genes whose sequences were completely encompassed by probes used to identify other gene sequences were removed, bringing the feature count down to 59,589. For all protein-coding genes, the different probes that were used to find the gene expression level were compared, and the probe which produced the highest mean expression level for that gene across the training examples was retained, while the other probes were dropped. This step eliminated 24 feature columns, which were also removed from the testing data. Because of the large range of the feature values in the gene expression dataset, logistic regression normalization was applied to the training and test datasets to bring the range of the gene expression features onto the same scale (between 0 and 1).

Feature Selection

In order to simplify the raw data to make training a classifier feasible as well as reduce the likelihood of overfitting, feature selection was implemented independently on the methylation and gene expression datasets. ANOVA (Kim, 2017), pairwise correlation analysis (Faizi and Alvi, 2023), and logistic regression with elastic net (Zou et al., 2005) were applied serially to the methylation training data to reduce the number of CpG sites from 361,038 to 160. The same procedure was applied to the gene expression training data, reducing the number of genes from 59,589 to 160. One-way ANOVA, a generalization of the two-sample t-test, is a statistical technique that determines whether there are significant differences between groups under study by comparing their means (Kim, 2017). For each feature (DNA methylation feature or gene expression feature), the training examples were resampled three times with replacement from the training data and grouped by cancer subtype. ANOVA was then used to rank the features based on their p-values. Ordering the p-values from least to greatest, the top 1,400 DNA methylation features and top 800 gene expression features were isolated. The multicollinearity of the resulting datasets was reduced using Pearson pairwise correlation with a threshold of greater than 0.7 or less than -0.7: in feature pairs in which the absolute value of the correlation score was greater than 0.7, 1 of the 2 features was removed. From there, 2 multiclass logistic regression models with elastic net regularization were trained on the resulting simplified data: one trained on the DNA methylation features, and one trained on the gene expression features Elastic net was utilized for regularization because it incorporates both ridge and LASSO regularization, and can thus reduce multicollinearity in the data as well as encourage sparsity (Narisetty, 2020).

Classification

Two XGBoost models (Lianglian et al., 2018) were trained for the purpose of predicting a patient's primary site: one was trained using the DNA methylation data, and the other was trained using the gene expression data. Recursive feature elimination was used to identify the minimum number of features needed to achieve greater than 90% accuracy. The performances of DNA-methylation-based and gene-expression-based classifiers were compared based on accuracy.

Pathway Enrichment and Clustering

Based on the features identified from ANOVA and logistic regression with elastic net, downstream analyses were carried out to understand the functional implications of differential methylation and differential expression to these features. The CpG sites were first annotated using Infinium Annotation (Zhou et al., 2017) to identify the genes they are located in. Using the 160 CpG sites, 131 unique genes were identified. Pathway enrichment analysis using Metascape (Zhou et al., 2019) was conducted on the differentially-methylated and differentially-expressed genes independently to identify pathways that are most highly implicated in carcinogenesis. The results were then compared to extract common classes of pathways found across both datasets. From there, networks of differentially-methylated and differentially-expressed genes were created using Metascape and analyzed using the Cytoscape network visualization tool (Shannon et al., 2003) to identify functional gene clusters as well as interesting subclusters within these clusters.

To identify pathway-level cancer-subtype biomarkers, as well as identify additional gene-level biomarkers, both the DNA methylation and gene expression datasets with the features identified from logistic regression with elastic net were used for clustering analysis. Hierarchical clustering (Robeva and Macauley, 2019) was applied on both the rows (training examples) and columns (features) to obtain a clustermap for each dataset. From there, each column cluster, which represents a group of co-methylated or co-expressed genes, was investigated using pathway enrichment analysis to determine the pathway it represents. To identify additional gene-level biomarkers beyond the 6 identified from recursive feature elimination, the methylation or expression levels of the genes in each cluster were visually compared between the cancer subtypes using bar plots. Genes that that showed high cancer subtype predictive power were identified and recorded as biomarkers.

Literature Cited

Adorján P, Distler J, Lipscher E, Model F, Müller J, Pelet C, Braun A, Florl AR, Gütig D, Grabs G, et al. (2002) Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res* 5

Aiello NM, Stanger BZ (2016) Echoes of the embryo: using the developmental biology toolkit to study cancer. *Dis Model Mech* 9: 105-14

Akman O, Comar T, Hrozencik D, Gonzales J (2019) Chapter 11—Data Clustering and Self-Organizing Maps in Biology in *MSE/Mathematics in Science and Engineering, Algebraic and Combinatorial Computationl Biology,* Robeva R, Macauley M (ed) pp 351-374. Academic Press

Alharbi F, Vakanski A (2023) Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. *Bioengineering (Basel)* 10

Angermueller C, Lee H J, Reik W, Stegle O (2017) DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* 67

Bubendorf L, Büttner R, Al-Dayel F, Dietel M, Elmberger G, Kerr K, López-Ríos F, Marchetti A, Öz B, Pauwels P, Penault-Llorca F, Rossi G, Ryška A, Thunnissen E (2016) Testing for ROS1 in non-small cell lung cancer: a review with recommendations. *Virchows Arch* 469: 489-503

Choi J M, Park C, Chae H (2023) meth-SemiCancer: a cancer subtype classification framework via semi-supervised learning utilizing DNA methylation profiles. *BMC Bioinformatics* 24

Cui Z, Zhou X, Zhang H (2019) DNA Methylation Module Network-Based Prognosis and Molecular Typing of Cancer. *Genes* 10

Deng T, Zhang W, Zhang Y, Zhang M, Huan Z, Yu C, Zhang X, Wang Y, Xu J (2021) Thyroid-stimulating hormone decreases the risk of osteoporosis by regulating osteoblast proliferation and differentiation. *BMC Endocr Disord* 21

Ding J, Zhao S, Chen X, Luo C, Peng J, Zhu J, Shen Y, Luo Z, Chen J (2022) Prognostic and Diagnostic Values of Semaphorin 5B and Its Correlation With Tumor-Infiltrating Immune Cells in Kidney Renal Clear-Cell Carcinoma. *Front Genet*. 13

Dong Y, Cai Y, Liu B, Jiao X, Li ZT, Guo DY, Li XW, Wang YJ, Yang DK (2017) HOXA13 is associated with unfavorable survival and acts as a novel oncogene in prostate carcinoma. *Future Oncol* 13: 1505-1516

Dunwell T, Hesson L, Rauch T A, Wang L, Clark R E, Dallol A, Gentle D, Catchpoole D, Maher E R, Pfeifer G P, et al. (2010) A Genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol Cancer* 44

Engelhardt B, Sorokin L (2009) The blood-brain and the blood-cerebrospinal fluid barriers: function and dysfunction. *Semin Immunopathol* 31: 497-511

Faizi N, Alvi Y (2023) Chapter 6—Correlation. In *Biostatistics Manual for Health Research*, Faizi N, Alvi Y (ed) pp 109-126. Academic Press

Fan S, Tang J, Li N, Zhao Y, Ai R, Zhang K, Wang M, Du W, Wang W (2019) Integrative analysis with expanded DNA methylation data reveals common key regulators and pathways in cancers. *NPJ Genom Med* 2

Gray RS, Cheung KJ, Ewald AJ (2010) Cellular mechanisms regulating epithelial morphogenesis and cancer invasion. *Curr Opin Cell Biol* 22: 640-50

Gunavathi C, Sivasubramanian K, Kerrthika P, Paramasivam C (2021) A Review on Convolutional Neural Network Based Deep Learning Methods in Gene Expression Data for Disease Diagnosis. *Mater. Today Proc.* 45: 2282-2285

Guyon J, Chapouly C, Andrique L, Bikfalvi A, Daubon T (2021) The Normal and Brain Tumor Vasculature: Morphological and Functional Characteristics and Therapeutic Targeting. *Front. Physiol.* 12

Hagerman RJ, Berry-Kravis E, Hazlett HC, Bailey Jr DB, Moine H, Kooy RF, Tassone F, Gantois I, Sonenberg N, Mandel JL, Hagerman PJ (2017) Fragile X syndrome. *Nat Rev Dis Primers* 3

Hijazi H, Chan C (2020) Using Class-Specific Feature Selection for Cancer Detection with Gene Expression Profile Data of Platelets. *Sensors* 20

Hughes AL, Szczurek AT, Kelly JR, Lastuvkova A, Tuberfield AH, Dimitrova E, Blackledge NP, Klose RJ (2023) A CpG island-encoded mechanism protects genes from premature transcription termination. *Nat Commun* 14

Iglesia RP, Prado MB, Alves RN, Escobar MIM, Fernandes CFL, Fortes ACDS, Souza MCDS, Boccacino JM, Cangiano G, Soares SR, de Araújo JPA, Tiek DM, Goenka A, Song X, Keady JR, Hu B, Cheng SY, Lopes MH (2022) Unconventional Protein Secretion in Brain Tumors Biology: Enlightening the Mechanisms for Tumor Survival and Progression. *Front Cell Dev Biol* 10

Janitz K, Janitz M (2011) Chapter 12—Assessing Epigenetic Information. In *Handbook of Epigenetics*, Tollefsbol T (ed) pp 173-181. Academic Press

Jiang SH, Zhang XX, Hu LP, Wang X, Li Q, Zhang XL, Li J, Gu JR, Zhang ZG (2020) Systemic Regulation of Cancer Development by Neuro-Endocrine-Immune Signaling Network at Multiple Levels. *Front Cell Dev Biol* 2020 8

Jiang X, Tan J, Li J, Kivimäe S, Yang X, Zhuang L, Lee PL, Chan MT, Stanton LW, Liu ET, Cheyette BN, Yu Q (2008) DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell* 6: 529-41

Jin N, George T L, Otterson G A, Verschraegen C, Wen H, Carbone D, Herman J, Bertino M E, He K (2021) Advances in epigenetic therapeutics with focus on solid tumors. *BMC* 83

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. *Genome Res* 12: 996-1006

Kim TK (2017) Understanding one-way ANOVA using conceptual figures. *Korean J Anesthesiol* 1: 22-26

Koch A, Joosten S C, Feng Z, de Ruijter T C, Draht M X, Melotte V, Smits K M, Veeck J, German J G, Neste L V et al. (2018) Analysis of DNA methylation in cancer: location revisited. *Nat Rev Clin Oncol* 15: 459-466

Koumakis L (2020) Deep Learning Models in Genomics; Are We There Tet? *Comput. Struct. Biotechnol. J.* 18: 1466-1473

Lattier JM, De A, Chen Z, Morales JE, Lang FF, Huse JT, McCarty JH (2020) Megalencephalic leukoencephalopathy with subcortical cysts 1 (MLC1) promotes glioblastoma cell invasion in the brain microenvironment. *Oncogene* 39: 7253-7264

Li Y, Fan Z, Meng Y, Liu S, Zhan H (2023) Blood-based DNA methylation signatures in cancer: A systematic review. *BBA* 1

Lianglian L, Ying Y, Zhihui F, Min L, Fang-Xiang W, Hong-Dong L, Yi P, Jianxin W (2018) An interpretable boosting model to predict side effects of analgesics for osteoarthritis. *BMC Syst Bio* 12

Liu B, Liu Y, Pan X, Li M, Yang S, Li SC (2019) DNA Methylation Markers for Pan-Cancer Prediction by Deep Learning. *Genes* 10

Martínez-Reyes I, Chandel NS (2021) Cancer metabolism: looking forward. *Nat Rev Cancer* 21: 669-680

McCaffrey MW, Lindsay AJ (2013) Rab Family. In *Encyclopedia of Biological Chemistry (Second Edition)*, Lennarz WJ, Lane MD (ed) pp 1-6. Academic Press

Messerschmidt DM, Knowles BB, Solter D (2014) DNA methylation dynamics during epigenetic reprogramming in the germling and preimplantation embryos. *Genes Dev* 28: 812-28

Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, Moutinho C, Heyn H, Diaz-Lagares A, de Moura MC, Stella GM, Comoglio PM, Ruiz-Miró M, Matias-Guiu X, Pazo-Cid R, Antón A, Lopez-Lopez R, Soler G, Longo F, Guerra I, Fernandez S, Assenov Y, Plass C, Morales R, Carles J, Bowtell D, Mileshkin L, Sia D, Tothill R, Tabernero J, Llovet JM, Esteller M (2016) Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *The Lancet* 17

Narisetty, NN (2020) Chapter 4—Bayesian model selection for high-dimensional data. In *Handbook of Statistics*, Rao ASRS, Rao CR (ed) pp 2017-248. Elsevier

Peng Y, Wu Q, Wang L, Wang H, Yin F (2020) A DNA methylation signature to improve survival prediction of gastric cancer. *Clin Epigenet* 15

Petersen J, Englmaier L, Artemov AV, Poverennaya I, Mahmoud R, Bouderlique T, Tesarova M, Deviatiiarov R, Szilvásay-Szabó A, Akkuratov EE (2023) A previously uncharacterized Factor Associated with Metabolism and Energy (FAME/C14orf105/CCDC198/1700011H14Rik) is related to evolutionary adaptation, energy balance, and kidney physiology. *Nat Commun.* 14

Prpić M, Franceschi M, Romić M, Jukić T, Kusić Z (2018) THYROGLOBULIN AS A TUMOR MARKER IN DIFFERENTIATED THYROID CANCER - CLINICAL CONSIDERATIONS. *Acta Clin Croat* 57: 518-527

Ram M, Najafi A, Shakeri MT (2017). Classification and Biomarker Genes Selection for

Cancer Gene Expression Data Using Random Forest. *Iran J. Pathol.* 12: 338-347

Sathe S, Aggarwal S, Tang J (2019) Gene Expression and Protein Function: A Survey of Deep Learning Methods. *SIGKDD Explor. Newsl.* 21: 23-38

Schirman D, Frumkin I, Pilpel Y (2017) Does cancer strive to minimize the cost of gene expression? *Oncotarget*

Segal NH, Pavlidis P, Noble WS, Antonescu CR, Viale A, Wesley UV, Busam K, Gallardo H, DeSantis D, Brennan MF, et al. (2003) Classification of Clear-Cell Sarcoma as a Subtype of Melanoma by Genomic Profiling. *JCO* 21: 1775-1781

Shannon P, Markiel A, Ozier O, Baliga NS, Wang, JT, Ramage D, Amin N, Schwikoswki B, Ideker T (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13: 2498-504

Shi CX, Long MA, Liu L, Graham FL, Gauldie J, Hitt MM (2004) The human *SCGB2A2* (Mammaglobin-1) promoter/enhancer in a helper-dependent adenovirus vector directs high levels of transgene expression in mammary carcinoma cells but not in normal nonmammary cells. *Molecular Therapy* 10: 758-767

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan-Golan Y, Kohn A, Rappaport N, Safran M, Lancet D (2016) The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* 54:1.30.1-1.30.33

Sund R, Pukkala E, Patja K (2008) Cancer incidence among persons with fragile X syndrome in Finland: a population-based study. *JIDR* 53: 85-90

Tang H, Pang P, Qin Z, Zhao Z, Wu Q, Song S, Li F (2021) The CPNE Family and Their Role in Cancers. *Front Genet* 12

Wang H, Zheng Q, Lu Z, Wang L, Ding L, Zia L, Zhang H, Wang M, Chen Y, Li G (2021) Role of the nervous system in cancers: a review. *Cell Death Discov* 7

Yu H, Liu Y, Zhang W, Peng Z, Yu X, Jin F (2023) A signature of cuproptosis-related lncRNAs predicts prognosis and provides basis for future anti-tumor drug development in breast cancer. *Transl Cancer Res* 12: 1392-1410

Zhang H, Kong Q, Wang J, Jiang Y, Hua H (2020) Complex roles of cAMP-PKA-CREB signaling in cancer. *Exp Hematol Oncol.* 9

Zhang X, Jonassen I, Goksøyr A. Chapter 4—Machine Learning Approaches for Biomarker Discovery using Gene Expression Data. In *Bioinformatics* [Internet], Helder IN (ed) Exon Publications

Zhang Y, Deng Q, Liang W, Zou X (2018) An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data. *BioMed Res. Int.*

Zhou W, Laird P W, Shen H (2017) Comprehensive characterization, annotation, and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 45: 22

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 10

Zhu W, Xie L, Han J, Guo X (2020) The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers* 12

Zou H, Hastie T (2005) Regularization and Variable Selection via the Elastic Net. *JSTOR* 67: 301-320

Data Availability

The datasets analyzed in the study are available from TCGA (https://cancergenome.nih.gov/).

The UCSC Genome Browser (http://genome.ucsc.edu) was used for visualization of genes.

WikiPathways (https://www.wikipathways.org/instance/WP2882) was used for pathway visualization of the nuclear receptors meta-pathway.

GeneCards (https://www.genecards.org/) was used to identify the pathways the top 6 gene expression and DNA methylation features are involved in.

Supplementary Files

Supplementary files for this study can be found at the following link:
https://drive.google.com/drive/folders/1f-IsPr08oB9ZdLYfBj4DzoJNEWoS9JCu?usp=drive_link