9-21-2017

Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers

Akram Mohammed
University of Nebraska-Lincoln, amohammed3@unl.edu

Greyson Biegert
University of Nebraska-Lincoln

Jiri Adamec
University of Nebraska-Lincoln, jadamec2@unl.edu

Tomáš Helikar
University of Nebraska-Lincoln, thelikar2@unl.edu

www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 49), pp: 85692-85715

Research Paper

Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers

Akram Mohammed[1], Greyson Biegert[1], Jiri Adamec[1] and Tomáš Helikar[1]

[1]Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

Correspondence to: Tomáš Helikar, **email:** thelikar2@unl.edu

Keywords: *cancer classification, biomarker identification, microarray gene expression, machine learning, cancer biomarker*

Received: *June 08, 2017* ***Accepted:*** *September 05, 2017* ***Published:*** *September 21, 2017*

ABSTRACT

Machine learning techniques for cancer prediction and biomarker discovery can hasten cancer detection and significantly improve prognosis. Recent "OMICS" studies which include a variety of cancer and normal tissue samples along with machine learning approaches have the potential to further accelerate such discovery. To demonstrate this potential, 2,175 gene expression samples from nine tissue types were obtained to identify gene sets whose expression is characteristic of each cancer class. Using random forests classification and ten-fold cross-validation, we developed nine single-tissue classifiers, two multi-tissue cancer-versus-normal classifiers, and one multi-tissue normal classifier. Given a sample of a specified tissue type, the single-tissue models classified samples as cancer or normal with a testing accuracy between 85.29% and 100%. Given a sample of non-specific tissue type, the multi-tissue bi-class model classified the sample as cancer versus normal with a testing accuracy of 97.89%. Given a sample of non-specific tissue type, the multi-tissue multi-class model classified the sample as cancer versus normal and as a specific tissue type with a testing accuracy of 97.43%. Given a normal sample of any of the nine tissue types, the multi-tissue normal model classified the sample as a particular tissue type with a testing accuracy of 97.35%. The machine learning classifiers developed in this study identify potential cancer biomarkers with sensitivity and specificity that exceed those of existing biomarkers and pointed to pathways that are critical to tissue-specific tumor development. This study demonstrates the feasibility of predicting the tissue origin of carcinoma in the context of multiple cancer classes.

INTRODUCTION

Cancer has been characterized as a heterogeneous disease that is categorized by many different types and subtypes. In the United States, cancer is the second leading cause of death. In 2016, over 1.6 million new cases of cancer were diagnosed and over 600,000 people died from this disease; the disease accounts for approximately 23% of all deaths in the US each year [1]. Successful treatment depends on the timely diagnosis, and the five-year survival rate significantly increases with early detection. Diagnosis typically begins with symptomology, is supported by imaging technology, and is confirmed histopathologically by biopsy. These methods, however, suffer from low sensitivity and high costs.

The identification of cancer-specific biomarkers is being evaluated as an alternative diagnostic and treatment option since it is minimally invasive and thus has the potential to lower the cost of diagnosis. Already, several biomarkers have been identified and used to some extent in diagnosis; however, they usually have low accuracy, selectivity, and specificity, and high false-positive, false-negative rates of diagnosis [2]. Therefore, improving the process and tools for the discovery of new biomarkers is essential for future improvement in cancer diagnostics and successful treatment.

While many strategies for discovering biomarkers exist, selecting useful biomarkers is a challenging task [3, 4]. Examples of these strategies include gene-expression profiling, mass-spectrometry-based

proteomic profiling, protein arrays and secreted protein approach [5]. Genomic and proteomic technologies have increased the number of potential biomarkers under investigation [6]. Furthermore, analysis of a single biomarker or a combination of only a few is increasingly being replaced by multiparametric analysis of genes, RNA, or proteins [7–10]. Specifically, high-throughput techniques such as microarrays and several machine-learning methods have been developed to study cancer classification and discovery of potential biomarkers [6, 11–20].

Many conventional biomarkers were established via discriminant analysis, through the comparison of cancerous tissues with normal tissues [21] or identifying nuanced differences among cancer subtypes [22, 23]. Progress in cancer biomarker identification has come through the application of machine learning to the analysis of high-throughput data from microarrays [24–29]. However, challenges remain in the application of machine learning to analyze biomarker data due to small sample sizes, the sheer size, and complexity of each dataset, as well as the diversity of experimental design [30].

The biomarker identification strategy outlined in this paper involves selecting genes whose differential expression in building cell structure, maintaining homeostasis, or the progression of cancer is a discriminating factor [31–34]. To achieve this, we developed single-tissue and multi-tissue machine learning cancer-versus-normal-classifiers using gene expression data that were used to identify tissue-specific cancer biomarkers. These biomarkers were obtained from machine learning models using gene expression data obtained and normalized from 2,175 samples and span nine tissue types. A feature selection method was identified to select informative genes (predicted biomarkers) from preprocessed data. Machine-learning models were identified through the analysis of gene expression samples from human cancer and non-cancerous tissue types that accurately distinguish malignant tissue from normal tissue and different malignant tissue types from each other. Using functional characterization and pathway analysis, the known tissue-specific cancer-related pathways were validated, and novel cancer-related pathways and functional groups for each of the tissue-specific predicted biomarkers were identified. The diagnostic capacity of the biomarkers predicted by the methods in this study (and later assessed by comparing their sensitivity and specificity to the sensitivity and specificity of known biomarkers for all tissue types) showed significant improvements over existing biomarkers. The development of our cancer prediction models and identification of the potential biomarkers may facilitate accurate, unbiased cancer diagnosis and effective treatment, ultimately improving cancer prognoses. Furthermore, the gene-expression signatures discovered by this classification approach may lead to new clinical reagents for successful tumor diagnosis.

RESULTS

Identification of the best feature selection algorithm

Of all the combinations of feature selection algorithms and feature thresholds tested (Step 4 in Figure 1), the Filtered Attribute Evaluator with Ranker method (FAER) used with a feature threshold of the top 1% genes performed the best (Supplementary Figure 1 shows the workflow for identification of the best feature selection algorithm and Supplementary File 1 provide the performance details of the feature selection algorithms; see Methods for the list of feature selection algorithms and feature thresholds). As such, FAER with a feature threshold of 1% was used for feature selection throughout this study.

Predictive power of the models

Single-tissue models

Given a sample of a specific tissue type, single-tissue models accurately classify the sample as cancer or normal. Each single-tissue model more accurately classified samples from the same tissue type (same-tissue) than it classified samples from other tissue types (across-tissues). The area under the ROC (receiver operating characteristics) curve for tissue-specific models ranged from 0.84 (Colon model) to 1 and is shown in Figure 2. Same-tissue testing accuracies ranged from 85.29% (Tongue Model) to 100% (Blood, Head and Neck and Lung Models). Across-tissues test accuracies ranged from 33.46% (Lung Model) to 88.68% (Gastric Model). (More details can be found in Supplementary Figure 2 and Supplementary Table 1).

Among the two classifiers, the random-forests classifier performed better than the Support Vector Machine classifier for each model except the Tongue model (the Random Forests classifier yielded an 85.29% same-tissue testing accuracy compared to 94.11% by the Support Vector Machine classifier). The Random Forests classifier outperformed the Support Vector Machine classifier in the across-tissues testing accuracies for each model. Supplementary Figure 3A and 3B show the same-tissue and across-tissues accuracies respectively. (To see differences between the performances of these classifiers, see Supplementary Table 2 and Supplementary Table 3). As a result, the models were constructed with Random Forests for the duration of this study.

A list of 244 genes (predicted biomarkers) was identified for each tissue type (See Supplementary File 2 for complete list of biomarkers for each tissue type and Table 2 for the number of characterized and uncharacterized genes for each tissue type) is given in Supplementary File 3 whereas, the list of uncharacterized genes is provided in Table 3.

Multi-tissue bi-class model

Given a sample of any of nine tissue types, the multi-tissue bi-class model accurately classifies the sample as cancer or normal. The area under the ROC curve for the multi-tissue bi-class model is 0.88 and is shown in Figure 3A. The multi-tissue bi-class model achieved training and testing accuracies of 97.33% and 97.89%, respectively. The model was more accurate in predicting a Cancer sample (Precision and Recall of 98.95% and 97.70%, respectively) than a Normal sample (Precision and Recall of 91.27% and 95.87%, respectively). (See Table 1 for Precision, Recall, and F1- score measures for both training and testing datasets.)

Multi-tissue multi-class model

Given a sample of any of the nine tissue types, the multi-tissue multi-class model accurately classifies the sample as cancer or normal, and as of a specific tissue type. The area under the ROC curve for the multi-tissue multi-class model is 0.97 and is shown in Figure 3B. The multi-tissue multi-class models achieved training and testing accuracies of 96.96% and 97.43%, respectively. The precision, recall, and F1- score for these models varied among classes (Figure 4). For the following classes, the model had 100% precision using the training dataset: blood-tumor, blood-normal, breast-tumor, gastric-tumor, gastric-normal, and head and neck-tumor. For the following classes, the model had 100% recall using the training dataset: blood-normal, gastric-normal, head-and neck-tumor, head-and-neck normal, lung-tumor, lung-normal, and tongue-normal. Out of all the classes, colon-normal (precision: 33.33%), prostate-normal (precision: 33.33%), and tongue-normal (precision: 28.57%) had the lowest precision using the training dataset (See Supplementary Table 4–6 for precision, recall, F1-score and confusion matrices).

Multi-tissue normal multi-class model

Given a normal sample of any of the nine tissue types, the multi-tissue normal multi-class model accurately

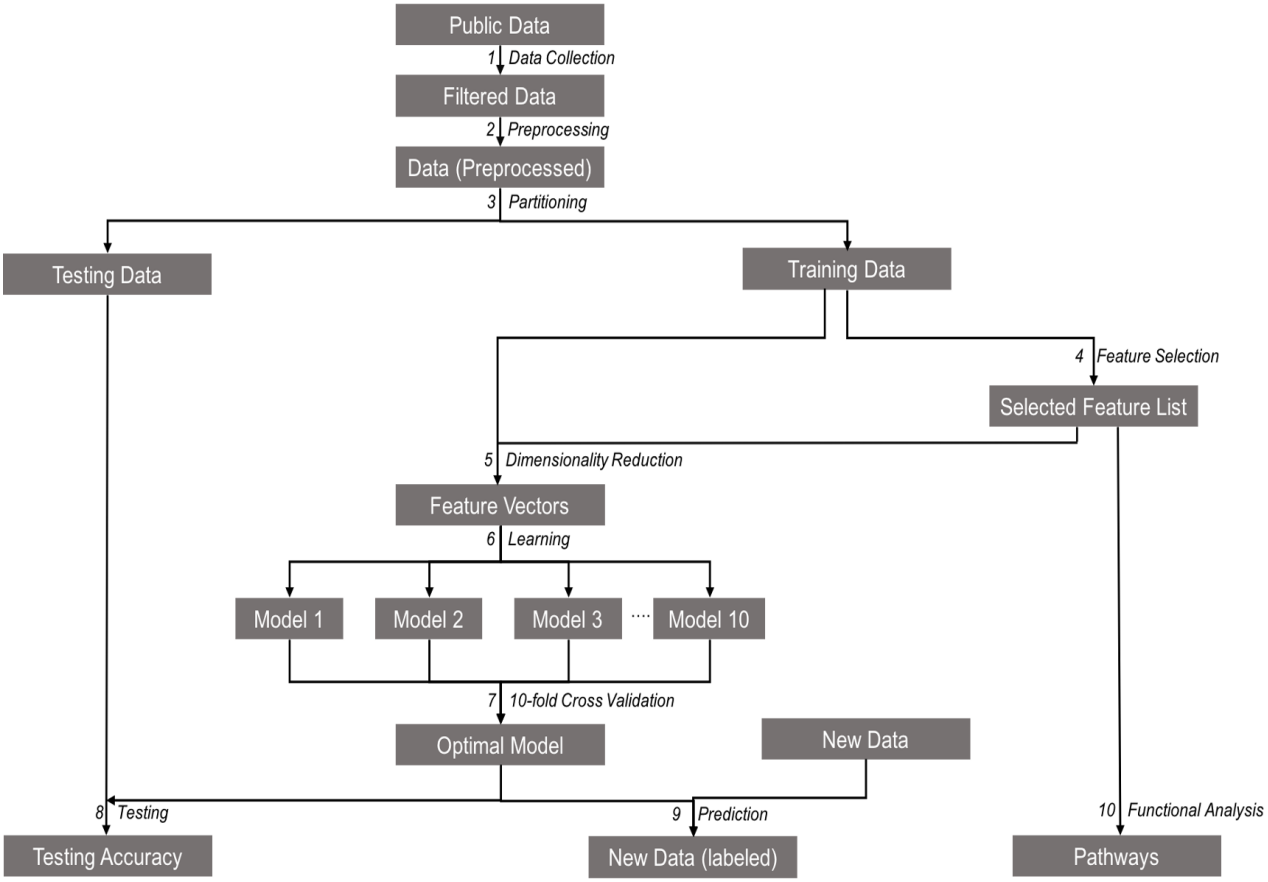

Figure 1: Schematic representation of the study workflow for each model. (1) Microarray gene expression data for each tissue type relevant to the model were collected from the NCBI Gene Expression Omnibus (GEO) repository. (2) The data were then normalized, and background correction was performed on these data. (3) The preprocessed data were then partitioned into training and testing sets. (4) Feature selection was conducted on the training dataset to extract the list of important genes. (5) The list of selected genes was then mapped to the training data to generate the feature vectors using a process called Dimensionality Reduction. (6) Feature vectors were trained to create multiple models. (7) Ten-fold cross-validation was used to identify the optimal model. (8) The model performance was assessed by the testing its accuracy using the testing dataset. (9) The model was used to predict the class labels for the samples in the unknown dataset. (10) The functional analysis was performed using the selected genes to retrieve the pathways and functional groups.

Table 1: Precision, recall and F1-Score for the multi-tissue bi-class model for training and testing data

	Training					Testing				
Class of Samples	# of Tumor Samples	# of Normal Samples	Precision (%)	Recall (%)	F1-Score	# of Tumor Samples	# of Normal Samples	Precision (%)	Recall (%)	F1-Score
Tumor	849	9	98.95	97.70	98.32	854	4	99.53	97.82	98.67
Normal	20	209	91.27	95.87	93.513	19	211	91.74	98.14	94.83

classifies the sample as of a particular tissue type. The area under the ROC curve for the multi-tissue normal multi-class model is 0.95 and is shown in Figure 3C. The multi-tissue normal multi-class models achieved training and testing accuracies of 97.88% and 97.35%, respectively. The models' precision and recall for each normal class using the testing dataset ranged from 87.5% to 100%, and from 95.45% to 100%, respectively (See Figure 5, Supplementary Tables 7–9 for details).

Functional analysis

Enrichment of cancer tissue-specific genes in metabolic and signaling pathways

A total of 104 KEGG (Kyoto Encyclopedia of Genes and Genomes, [35]) pathways were identified for the nine tissue types. The gastric tissue genes had the most pathways (38), whereas the blood and lung tissue genes had the fewest pathways (4). The colon-tissue genes had the second highest number (14) of KEGG pathways (Figures 6-8, Supplementary Figures 4–6, and Supplementary File 4).

Significant pathways for each tissue type are presented below.

Blood

Four pathways were identified using blood tissue genes. The only metabolic pathway identified was hsa00564: glycerophospholipid metabolism. The other three pathways are involved in intracellular signaling: hsa04015: Rap1 signaling pathway, hsa04064: NF-kappa B signaling pathway and hsa04080: neuroactive ligand-receptor interaction.

Breast

Six pathways were identified using the breast tissue genes. Some of these pathways are involved in inter- or intra-cellular structures: hsa04510: focal adhesion and hsa04810: regulation of the actin cytoskeleton. The rest were signaling motifs: hsa04670: leukocyte transendothelial migration, hsa03010: ribosome and hsa05131: shigellosis pathway.

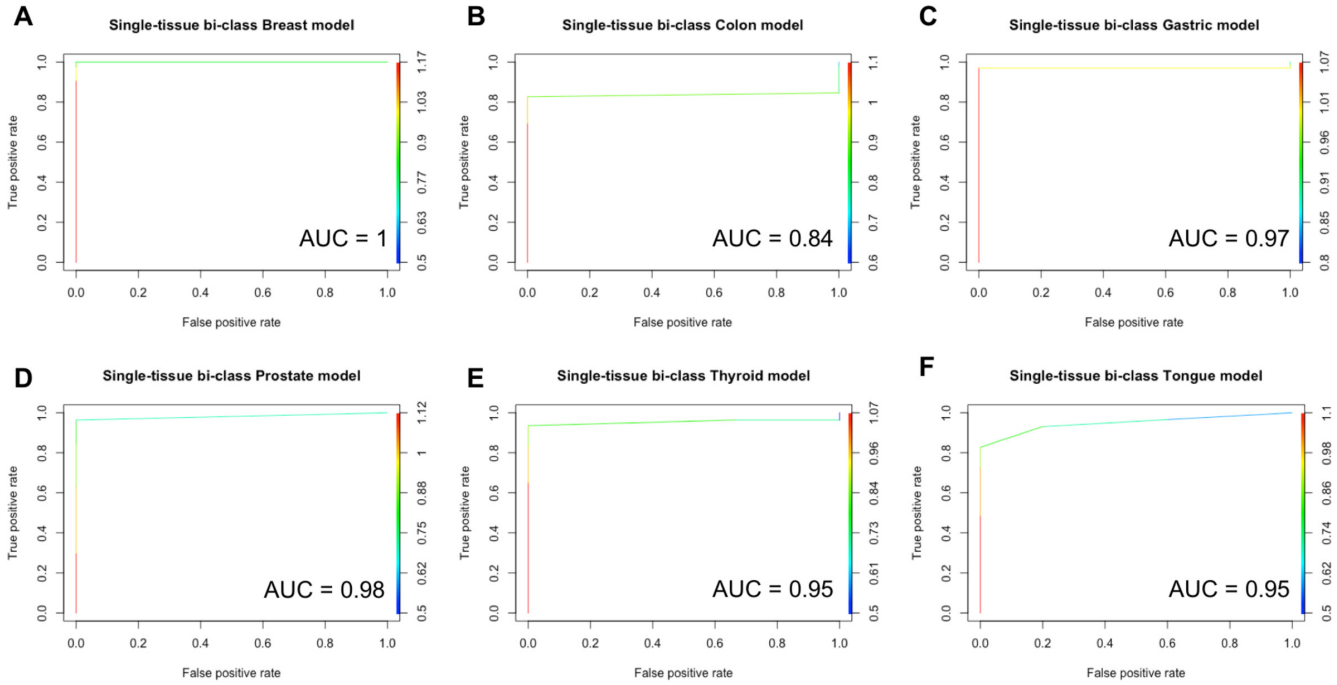

Figure 2: ROC for single-tissue specific models. The area under the ROC curves is shown for each model. (**A**) Breast, (**B**) Colon, (**C**) Gastric, (**D**) Prostate, (**E**) Thyroid, and (**F**) Tongue. The ROC for the Blood, Head & Neck and Lung models are not shown due to the due to their AUC = 1.

Table 2: Distribution of characterized and uncharacterized genes for each tissue type

Tissue	Predicted Biomarkers (Characterized Genes)	Predicted Biomarkers (Uncharacterized Genes)
Blood	170	74
Breast	239	5
Colon	240	4
Gastric	238	6
Head & Neck	243	1
Lung	224	20
Prostate	238	6
Thyroid	240	4
Tongue	237	7

Colon

Colon tissue genes were used to identify 14 pathways. One of this pathways was a signaling pathway: hsa04725: cholinergic synapse. Another pathway was involved in disease development, hsa05204: chemical carcinogenesis. Most of the remaining pathways were involved in diverse metabolic functions: hsa00830: retinol metabolism, hsa00982: drug metabolism–cytochrome P450, hsa00983: drug metabolism–other enzymes, hsa00053: ascorbate and aldarate metabolism, hsa00040: pentose and glucuronate interconversions, hsa00140: steroid hormone biosynthesis, hsa00860: porphyrin and chlorophyll metabolism, hsa00980: metabolism of xenobiotics by cytochrome P450 (For detailed results, refer to the Supplementary File 4).

Gastric

Gastric tissue genes were used to identify 38 pathways. Many of the pathways were involved in synaptic function: hsa04724: glutamatergic synapse, hsa04727: GABAergic synapse, hsa04725: cholinergic synapse, hsa04728: dopaminergic synapse, hsa04726:

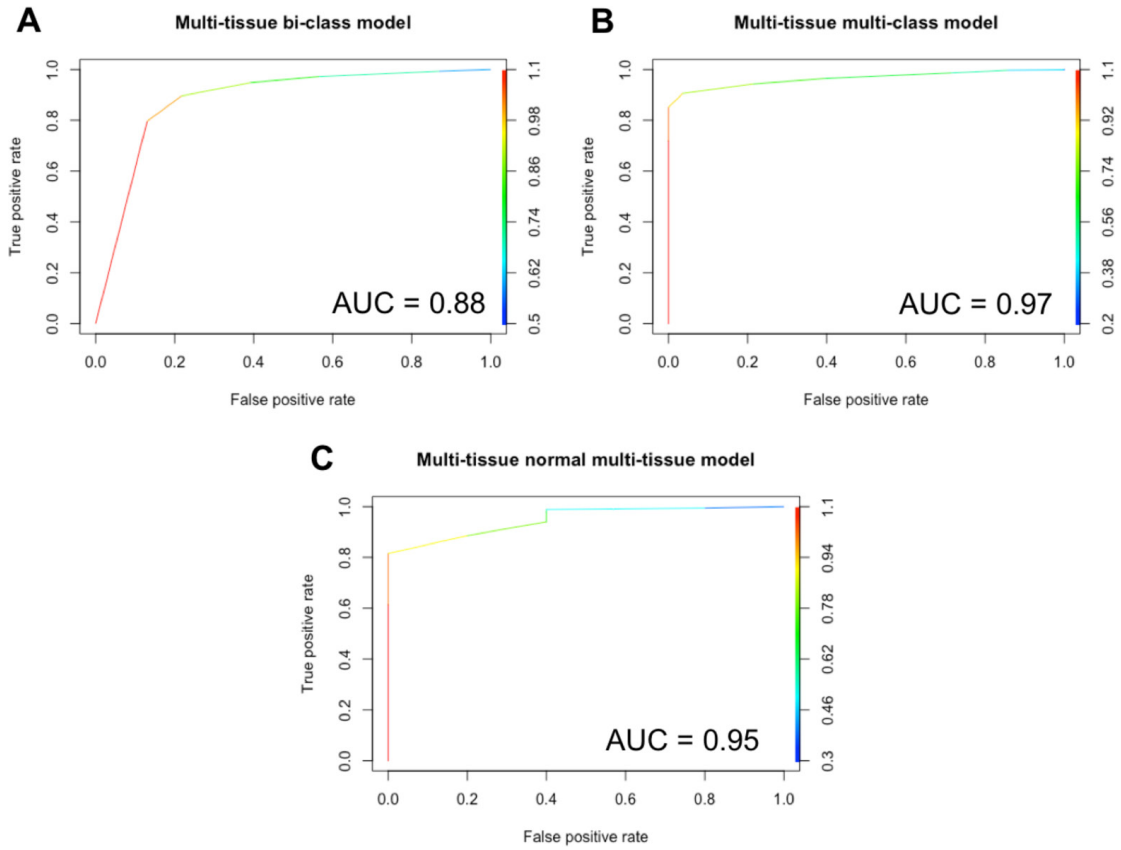

Figure 3: ROC for multi-tissue models. The area under the ROC curves is shown for each model. (**A**) multi-tissue bi-class model, (**B**) multi-tissue multi-class model, (**C**) multi-tissue normal multi-class model.

serotonergic synapse. Other pathways were involved in the different signaling aparati (hsa04062: chemokine signaling pathway, hsa04014: ras signaling pathway, hsa04070: phosphatidylinositol signaling system, hsa04151: PI3K-Akt signaling, hsa04071: sphingolipid signaling, hsa04744: phototransduction, and hsa04022: cGMP-PKG signaling pathway). The disease-related pathways included hsa05200: pathways in cancer, hsa05034: alcoholism, hsa05142: Chagas disease, hsa05146: amoebiasis, hsa04930: type II diabetes mellitus, hsa05213: endometrial cancer. The remaining pathways are involved in various forms of fatty acid chain metabolism: hsa00562: inositol phosphate metabolism, hsa00564: glycerophospholipid metabolism, hsa00592: alpha-Linolenic acid metabolism, hsa00565: ether lipid metabolism, hsa00563: glycosylphosphatidylinositol (GPI)-anchor biosynthesis, hsa00590: arachidonic acid metabolism (Supplementary File 4).

Head and neck

Using the Head and Neck tissue genes we identified ten pathways. Most of these pathways are specific to cellular signaling and regulation of signaling pathways: hsa04015: Rap1 signaling, hsa04610: complement and coagulation cascades, hsa04550: signaling pathways regulating pluripotency of stem cells, hsa04014: Ras signaling, hsa04151: PI3K-Akt signaling and has03018: RNA degradation. The disease-related pathways involve hsa05150: Staphylococcus aureus infection, hsa05218: melanoma, hsa05200: pathways in cancer, and hsa05217: Basal cell carcinoma.

Lung

Four pathways were identified using lung tissue genes. Three of these pathways were involved in signal transduction: hsa04080: neuroactive ligand-receptor

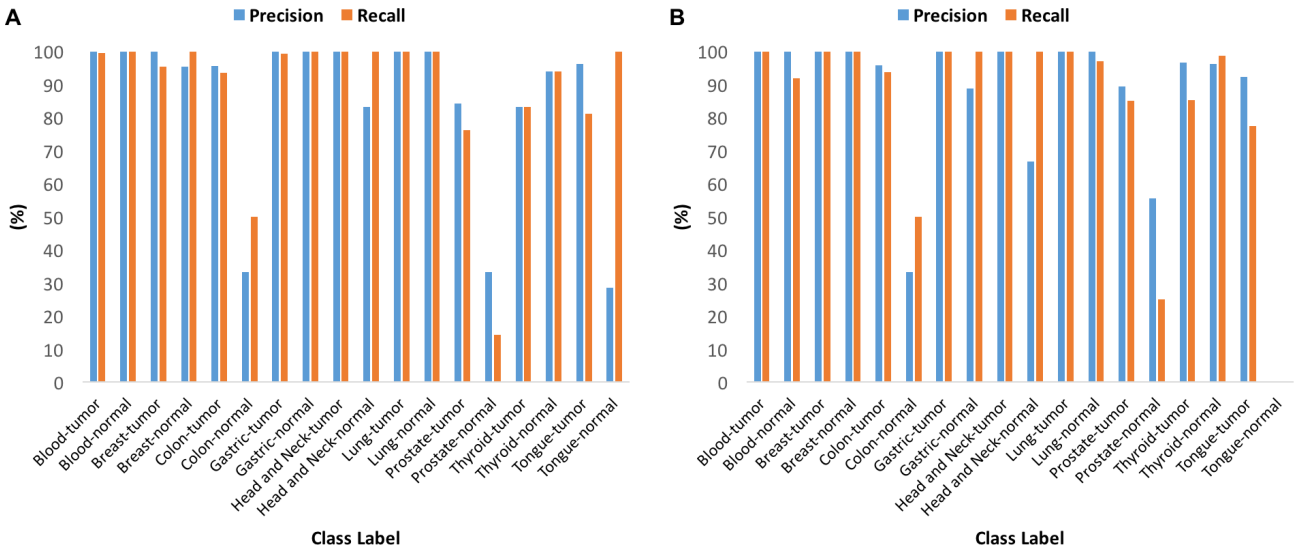

Figure 4: Performance of the multi-tissue multi-class models for each class. (**A**) Precision and recall using the training dataset. (**B**) Precision and recall using the testing dataset.

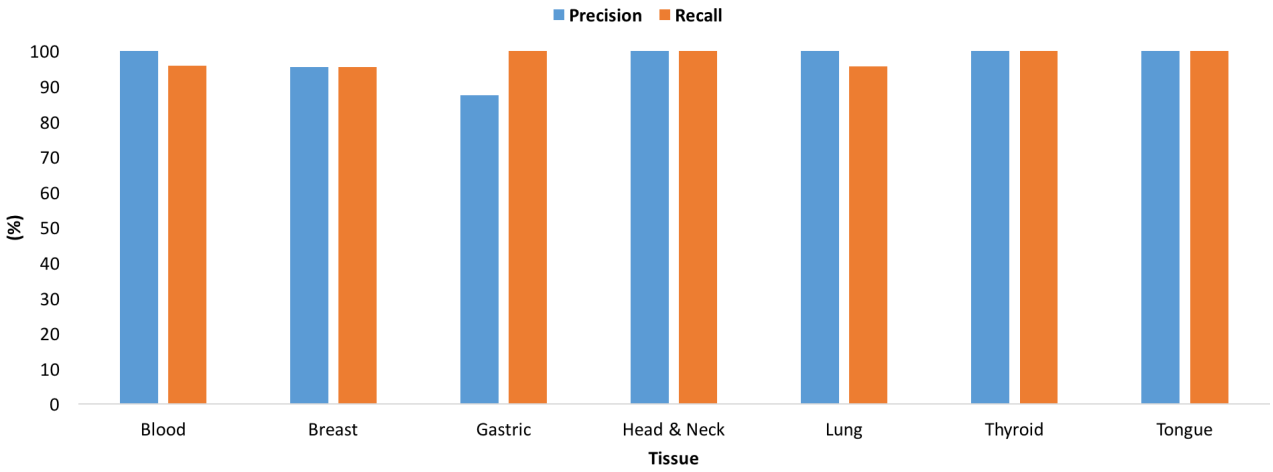

Figure 5: Performance of the multi-tissue normal multi-class model for each class. Precision and Recall values are shown for each of the nine tissue types using the testing dataset.

Table 3: List of uncharacterized genes predicted as biomarkers from different tissue types

Potential Biomarker (Uncharacterized genes)	Tissue
CTC-265F19.1	Blood
CTC-360P9.3	Blood
CTC-378H22.2	Blood
CTC-384G19.1	Blood
CTC-400I9.1	Blood
CTC-428G20.6	Blood
CTC-436K13.5	Blood
CTC-459F4.3	Blood
CTC-462L7.1	Blood
CTC-471C19.1	Blood
CTC-471F3.6	Blood
CTC-471J1.2	Blood
CTC-527H23.4	Blood
CTC-550B14.7	Blood
CTD-2002J20.1	Blood
CTD-2008P7.1	Blood
CTD-2012K14.6	Blood
CTD-2021H9.3	Blood
CTD-2033C11.1	Blood
CTD-2035E11.5	Blood
CTD-2036P10.3	Blood
CTD-2076M15.1	Blood
CTD-2083E4.4	Blood
CTD-2083E4.7	Blood
CTD-2118P12.1	Blood
CTD-2130O13.1	Blood
CTD-2196E14.6	Blood
CTD-2199O4.3	Blood
CTD-2199O4.7	Blood
CTD-2251F13.1	Blood
CTD-2256P15.2	Blood
CTD-2269F5.1	Blood
CTD-2281E23.2	Blood
CTD-2284J15.1	Blood
CTD-2286N8.2	Blood
CTD-2287O16.5	Blood
CTD-2293H3.1	Blood
CTD-2302E22.4	Blood
CTD-2310F14.1	Blood
CTD-2311B13.7	Blood
CTD-2313J17.5	Blood
CTD-2314B22.3	Blood
CTD-2325A15.5	Blood
CTD-2366F13.2	Blood
CTD-2373J6.1	Blood
CTD-2377D24.6	Blood
CTD-2520I13.1	Blood
CTD-2534I21.8	Blood
CTD-2537I9.16	Blood
CTD-2537I9.5	Blood
CTD-2540F13.2	Blood
CTD-2541J13.1	Blood
CTD-2541M15.1	Blood
CTD-2542L18.1	Blood
CTD-2547L24.4	Blood
CTD-2553C6.1	Blood
CTD-2554C21.3	Blood
CTD-2555O16.4	Blood
CTD-2561B21.11	Blood
CTD-2587H24.10	Blood
CTD-2587M23.1	Blood
CTD-2611O12.6	Blood
CTD-2616J11.10	Blood
CTD-2619J13.13	Blood

CTD-2619J13.17	Blood
CTD-2639E6.4	Blood
CTD-2647L4.1	Blood
CTD-3028N15.1	Blood
CTD-3046C4.1	Blood
LOC730139	Blood
LOC731424	Blood
LOC80154	Blood
LOC90834	Blood
LQFBS-1	Blood
AX746733	Breast
RP11-114H24.6	Breast
RP11-255C15.3	Breast
RP11-348B17.1	Breast
RP11-403P17.4	Breast
LA16C-381G6.1	Colon
LOC100652770	Colon
RP11-295M18.6	Colon
RP11-38P22.2	Colon
GS1-103B18.1	Gastric
GS1-111G14.1	Gastric
GS1-18A18.2	Gastric
GS1-124K5.9	Gastric
GS1-164F24.1	Gastric
GS1-304P7.2	Gastric
FLJ11292	Head And Neck
RP11-69I8.2	Lung
RP3-406C18.2	Lung
RP4-710M16.1	Lung
AC007967.3	Lung
LOC613037	Lung
LOC100127886	Lung
RP1-217P22.2	Lung
AC009947.3	Lung
RP11-770J1.4	Lung
RP11-209A2.1	Lung
RP5-1184F4.5	Lung
MGC13053	Lung
RP3-391O22.2	Lung
LOC649330	Lung
RP3-406P24.1	Lung
RP13-258O15.1	Lung
RP5-1118D24.2	Lung
GS1-124K5.9	Lung
RP1-190J20.2	Lung
RP1-192P9.1	Lung
AC004941.5	Prostate
LOC100506119	Prostate
RP1-101G11.2	Prostate
RP11-297L17.2	Prostate
AX746823	Prostate
RP11-96K19.4	Prostate
RP6-24A23.7	Thyroid
LOC100506558	Thyroid
LOC101930400	Thyroid
LOC102725271	Thyroid
CTA-384D8.35	Tongue
RP11-353N14.2	Tongue
CTC-444N24.11	Tongue
RP11-539I5.1	Tongue
LOC101928615	Tongue
GS1-111G14.1	Tongue
RP11-250B2.3	Tongue

interaction, hsa04024: cAMP signaling, and hsa04924: renin secretion. The fourth pathway is involved in hsa04260: cardiac muscle contraction.

Prostate

Prostate tissue genes were used to identify eight pathways. These include several metabolic pathways: hsa00480: glutathione metabolism, hsa00051: fructose and mannose metabolism, hsa00982: drug metabolism–cytochrome P450, hsa00030: pentose phosphate pathway and hsa00052: galactose metabolism. The other pathways are hsa04512: ECM-receptor interaction signaling pathway, hsa05200: pathways in cancer, and hsa04510: focal adhesion, a structural pathway.

Thyroid

Nine pathways were identified using thyroid tissue genes. The signaling pathways included hsa04512: ECM-receptor interaction and hsa04151: PI3K-Akt signaling. A few structural pathways were identified, including hsa04510: Focal adhesion, hsa05205: proteoglycans in cancer and hsa04360: axon guidance. The only metabolic pathway identified was hsa00350: tyrosine metabolism. The disease-related pathways include hsa05222: small cell lung cancer, hsa05200: pathways in cancer, and hsa05146: amoebiasis.

Tongue

Twelve pathways were identified using tongue-tissue genes. Many of the identified pathways were disease-related, including hsa05323: rheumatoid arthritis, hsa05146: amoebiasis, hsa05200: pathways in cancer, hsa05142: Chagas disease, hsa05132: Salmonella infection, hsa05222: small cell lung cancer, and hsa05140: leishmaniasis. The only structural pathway was hsa05205: proteoglycans in cancer. The following four signaling pathways were hsa04620: Toll-like receptor signaling, hsa04062: chemokine signaling, hsa04512: ECM-receptor interaction, and hsa04060: cytokine-cytokine receptor interaction.

Enrichment of cancer tissue-specific genes in various functional groups

Using tissue-specific genes, functional groups were identified related to protein kinase inhibitor activity (GO:0004860), negative regulation of JAK-STAT cascade (GO:0046426), myosin complex (GO:0016459), G-protein coupled receptor signaling pathway (GO:0007186), GTPase activity (GO:0003924), signal transducer activity (GO:0004871), flavone metabolic process (GO:0051552), tissue homeostasis (GO:0001894), amino acid transmembrane transporter activity (GO:0015171), regulation of MAPK cascade (GO:0043408), type I interferon signaling pathway (GO:0060337), and others. Figures 9–11 show the functional groups with the top five Gene Ontology (GO) groups with the total number of genes from each tissue-specific gene list. See Supplementary File 5 for full list of functional groups.

Predicted biomarkers perform better than existing biomarkers

A total of 244 potential biomarkers were identified for each tissue type distributed across the different cancer types (Supplementary File 2). The quality of these predictions was assessed by comparing the sensitivity and specificity of biomarkers to the sensitivity and specificity of existing biomarkers collected from the literature (Supplementary Tables 10–16). Biomarkers predicted by our machine learning models resulted in higher sensitivity and specificity for each tissue type than those of existing biomarkers (Figure 12).

DISCUSSION

In this study, machine learning models were developed to analyze a large-scale human gene-expression dataset to identify cancer biomarkers within nine tissue types. Given the presence of cancer, machine learning models were also equipped to distinguish between cancer types. A machine-learning method to select informative

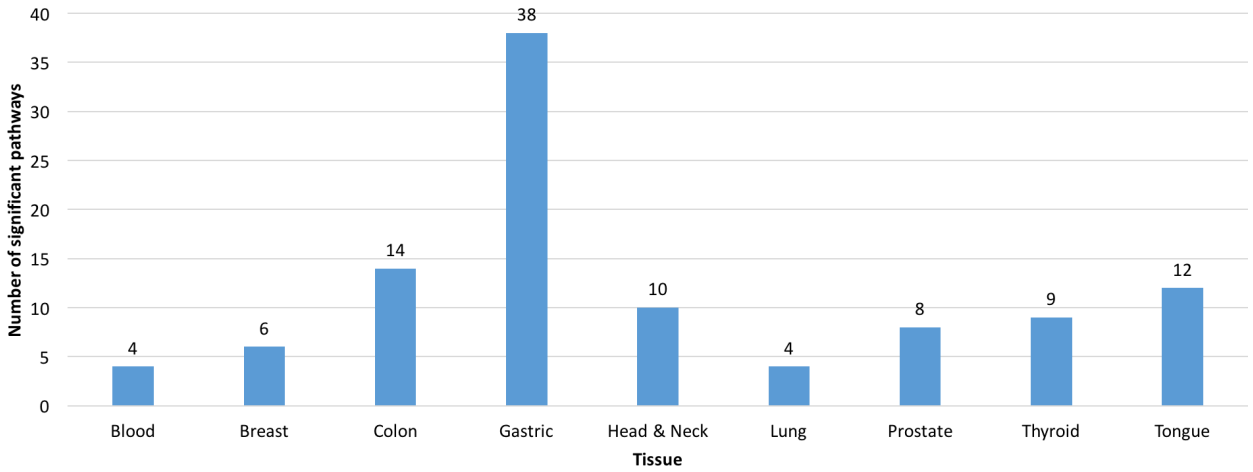

Figure 6: Number of significant pathways for the genes (predicted biomarkers) from each tissue type. A pathway was significant if its *p*-value was less than or equal to 0.05 and it had a minimum of three tissue-specific genes.

Figure 7: KEGG Pathway mapping for each tissue type using identified genes (potential biomarkers). A pathway was considered significant if its *p*-value was less than or equal to 0.05 and it had a minimum of three tissue-specific genes.

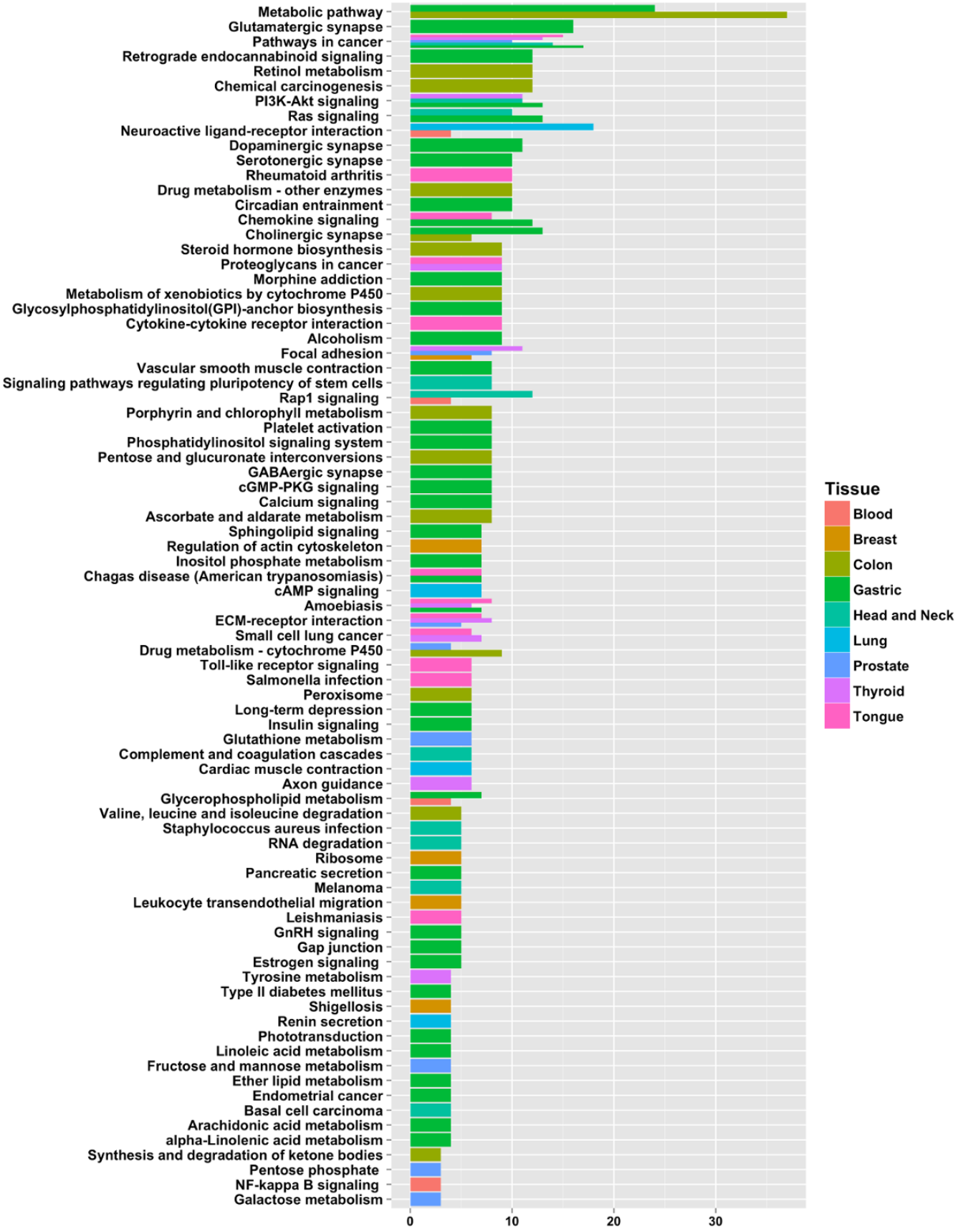

Figure 8: Number of selected genes (potential biomarkers) in pathways for each tissue type. A pathway was considered significant if its *p*-value was less than or equal to 0.05 and it had a minimum of three tissue-specific genes.

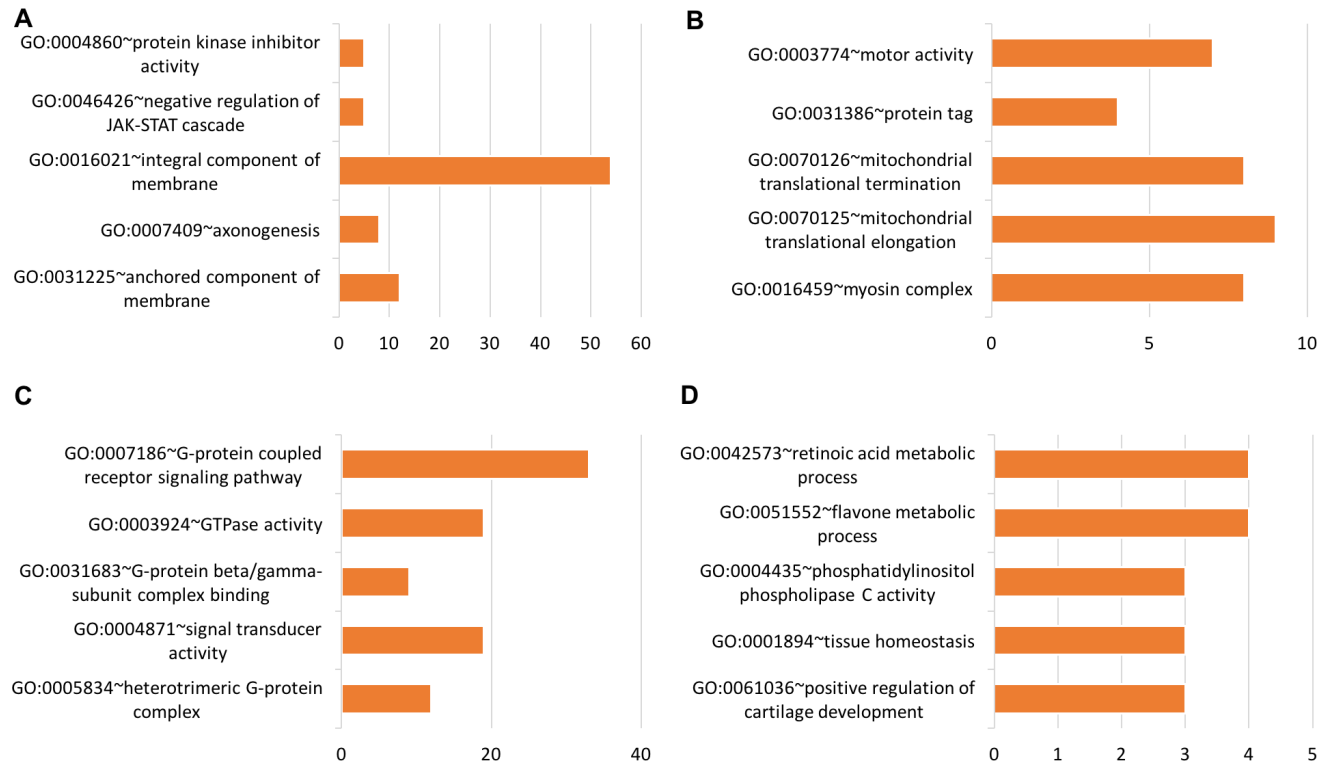

Figure 9: Top five significant Gene Ontology groups with the total number of predicted tissue genes. A functional group was considered significant if its *p*-value was less than or equal to 0.05 and if it had a minimum of three tissue-specific genes. (**A**) Blood, (**B**) Breast, (**C**) Colon, (**D**) Gastric

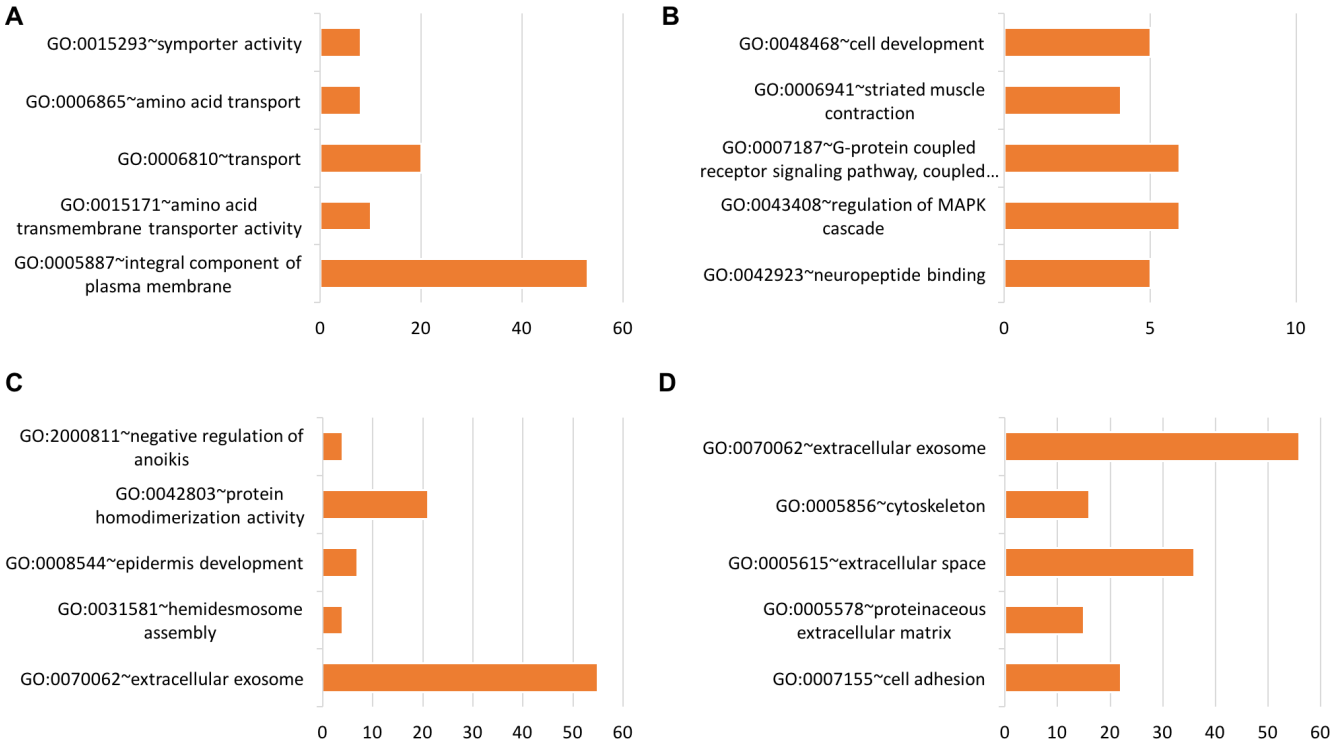

Figure 10: Top five significant Gene Ontology groups with the total number of predicted tissue genes. A functional group was considered significant if its *p*-value was less than or equal to 0.05 and it had a minimum of three tissue-specific genes. (**A**) Head & neck, (**B**) Lung, (**C**) Prostate, (**D**) Thyroid

genes (potential biomarkers) was identified for each tissue type. Four different classifiers were developed: (1) models which, given a sample of a specific tissue type, accurately classify the sample as cancer or normal ("single-tissue"), (2) a model which, given a sample of any of nine tissue types, accurately classifies the sample as cancer or normal ("multi-tissue bi-class"), (3) a model which, given a sample of any of the nine tissue types, accurately classifies the sample as cancer or normal, and as of a specific tissue type ("multi-tissue multi-class"), and (4) a model which, given a normal sample of any of the nine tissue types, accurately classifies the sample as of a particular tissue type (multi-tissue normal multi-class). (See Figure 13A and Supplementary Table 17 for distribution of samples among tissue types.) The classifiers, trained to incorporate a vast array of different tissue types, and the predicted biomarkers may facilitate accurate, unbiased cancer diagnosis and effective treatment, ultimately improving prognoses.

Machine learning methodology

The selection of relevant genes involved in different types of cancer remains a challenge [36, 37]. Moreover, for diagnostic purposes, it is important to find a small subset of genes that are sufficiently informative to distinguish between different cancer types. To extract useful gene information from cancer microarray data and reduce dimensionality, feature-selection algorithms were systematically investigated in this study. To this end, a feature selection method was identified (FAER with 1% feature threshold) from twelve feature selection algorithms to select informative genes (potential biomarkers) for each tissue type. As we showed, selecting relatively small subsets of genes significantly improved the performance of our classification models. The single-tissue models were tested using the testing data from all the nine tissues as part of the negative control. Each single-tissue model more accurately classified samples from the same tissue

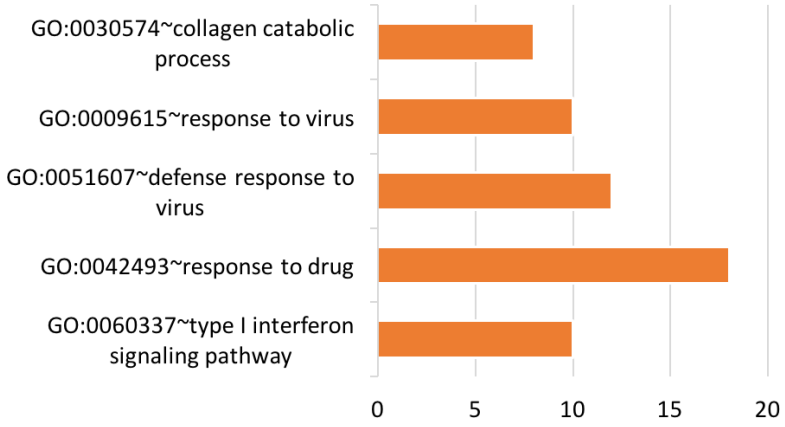

Figure 11: Top five significant Gene Ontology groups with the total number of predicted tongue tissue genes. A functional group was considered significant if its *p*-value was less than or equal to 0.05 and if it had a minimum of three tissue-specific genes.

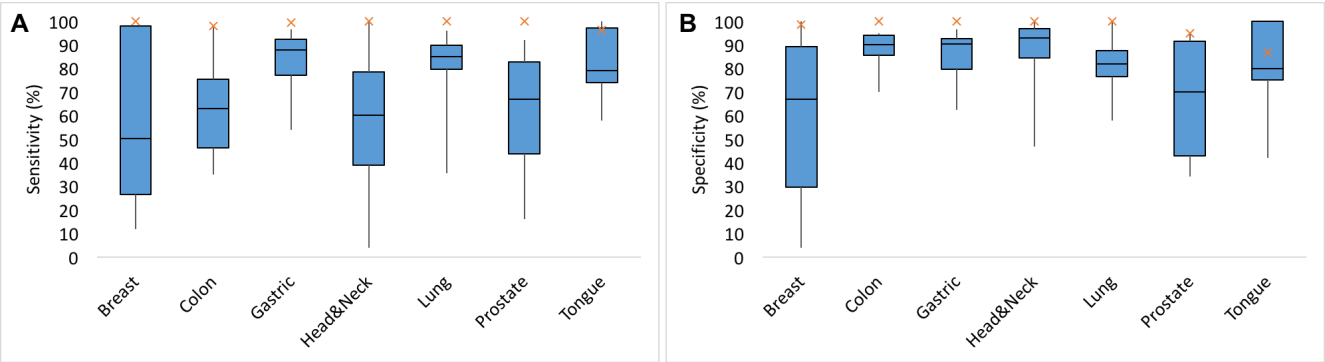

Figure 12: Performance of predicted biomarkers with known biomarkers for each tissue type. (**A**) Sensitivity of the existing biomarkers (breast: 50.1%, colon:63%, gastric:87.95%, head&neck:60%, lung:85%; prostate:67%, tongue:79%) is represented by box plot (blue) and sensitivity of our predicted biomarkers (breast:100%, colon:97.92%, gastric: 99.37 %, head & neck:100%, lung:100%; prostate:100%, tongue:96.3%) is represented by cross mark (orange). (**B**) specificity of the existing biomarkers (breast:66.89%, colon:90%, gastric: 90.3%, head& neck:92.9%, lung:82%; prostate:70%, tongue:80%) is represented by box plot (blue) and specificity of our predicted biomarkers (breast:98.46%, colon: 100%, gastric: 100%, head & neck:100%, lung:100%; prostate:95%, tongue:86.67%) is represented by cross (orange).

type (same-tissue) than it did samples from other tissue types (across-tissues). The multi-tissue bi-class and multi-class models were not only able to classify the sample as cancer or normal but also the tissue of origin. Moreover, this feature selection process also identified genes that are closely related to the pathways and functional groups of various cancers.

Metabolic pathways

The metabolism of a tumor depends on both the genotype and tissue of origin and has implications regarding the design of therapies targeting tumor metabolism [38]. Tissue-specific genes pointed to metabolic pathways that may be critical to tumor development in general and tissue-specific tumor development (see the list of pathways in Supplementary File 4). Metabolic rewiring is essential for the progression of many types of cancer [39, 40]. We discuss the metabolic pathways for selected tissues below.

Blood

The metabolic pathway identified using the blood tissue genes is glycerophospholipid metabolism, and there is an increase of acyl-glycerophospholipids in acute myeloid leukemia [41].

Colon

Most of the metabolic pathways identified using the colon tissue genes have known links to colon cancer. One such pathway is retinol metabolism; retinoids are known to play a role in the prevention and treatment of colorectal cancer [42–44]. Some of the colon-cancer genes identified also include steroid hormone biosynthesis, as the bacterial cells in the gut produce steroid hormones that can have implications for colon cancer [45]. Some colon cancer genes also identified include metabolism of xenobiotics; biotransformation of xenobiotics occurs in the human colon and rectum, and it is known to be associated with colorectal cancer [46–48]. Pentose and glucuronate interconversions were also identified using colon-tissue genes. The heightened metabolic demands of colon cancer cells are known to result in increased glucose uptake and glycolytic flux relative to normal tissues [49, 50]. One common feature of the altered metabolism in cancer is the increased glucose uptake and fermentation of glucose to lactate, a phenomenon known as the Warburg Effect [51, 52]. In tumor cells and other proliferating cells, the rate of glucose uptake dramatically increases, even in the presence of oxygen and fully functioning mitochondria.

Gastric

Many of the pathways identified using gastric tissue genes are involved in various forms of fatty acid chain metabolism: inositol phosphate metabolism, glycerophospholipid metabolism, ether lipid metabolism, glycosylphosphatidylinositol (GPI)-anchor biosynthesis, arachidonic acid metabolism and alpha-Linolenic acid metabolism. α-linolenic acid is known be the most effective in suppressing the growth of gastric cancer cells [53]. These results suggest that the metabolism of fatty acids may play a critical role in the tumorigenesis of gastric cancer. Levels of metabolism of fatty acids in cancer cells are known to vary across tissue types [54].

Prostate

Some of the metabolic pathways identified using the prostate tissue genes are glutathione metabolism and pentose phosphate metabolism. The glutathione S-transferases

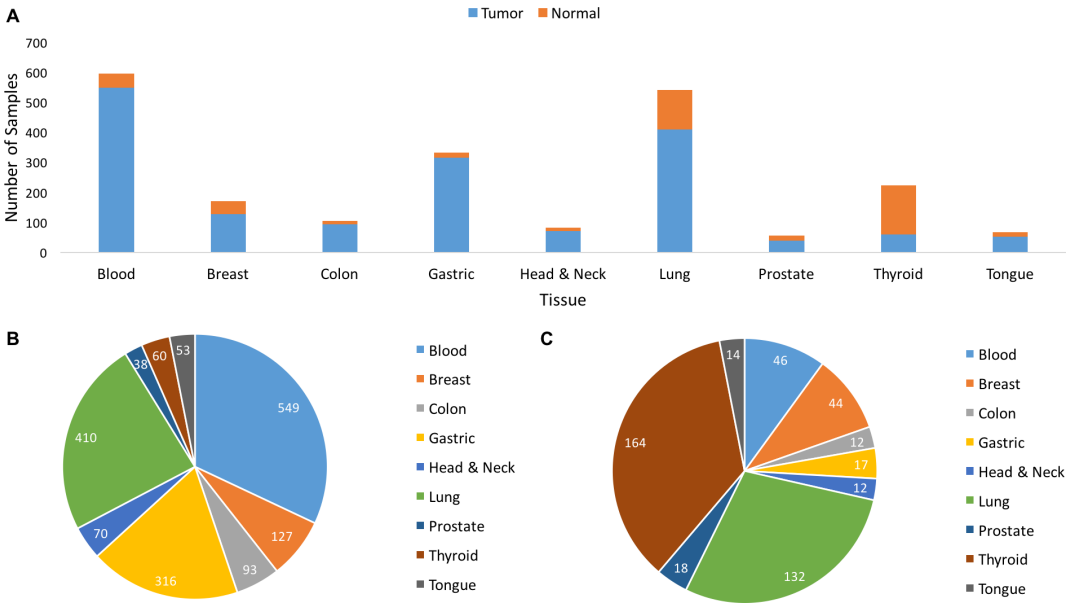

Figure 13: Sample Distribution of Tumor and Normal samples by tissue of origin. (**A**) Distribution of Tumor and Normal samples (2175) by tissue of origin. (**B**) Distribution of Tumor (1716) samples by tissue of origin. (**C**) Distribution of Normal (459) samples by tissue of origin.

(GSTs) enzymes are known to be involved in the metabolism of numerous potential prostate carcinogens [55, 56]. Cancer cells display an increased demand for glucose. Clinical data suggested that the glucose-6-phosphate dehydrogenase (G6PD), the rate-limiting enzyme in the pentose phosphate pathway, is upregulated in prostate cancer [57].

Thyroid

Using the thyroid tissue genes, we identified a tyrosine metabolism pathway where the thyroid gland uses tyrosine residues to generate T3 and T4, metabolic hormones known to be involved in thyroid cancer [58, 59].

Signaling pathways

Signaling pathways controlling cell growth, cell division, cell death, cell fate, and cell motility are almost invariably altered in cancer [60]. Many of the signaling pathways found in this study were identified using tissue-specific genes (discussed below).

Blood

The pathways identified using the blood tissue genes are involved in intracellular signaling (Rap1 signaling, NF-kappa B, and neuroactive ligand-receptor interaction signaling). Ras is known to induce myeloproliferative disorders and acute myeloid leukemia [61, 62]. Nuclear factor-kappaB is constitutively activated in human acute myeloid leukemia cells [63–68].

Gastric

Signaling pathways and interaction networks were altered in gastric cancer tissues [69, 70]. The pathways identified using the gastric tissue genes were involved in signaling aparati (chemokine signaling, Ras signaling, phosphatidylinositol signaling, PI3K-Akt signaling, sphingolipid signaling, phototransduction, and cGMP-PKG signaling). Phototransductive proteins are expressed to increase intracellular calcium in tumor cells for gastric cancer patients [71, 72].

Head and neck

Most of the pathways identified using head and neck genes are specific to cellular signaling and regulation of signaling pathways known to be involved in head and neck cancer, including complement and coagulation cascades, signaling pathways regulating pluripotency of stem cells, Ras signaling, PI3K-Akt signaling and RNA degradation. Rap1 signaling, Rap1, and Rap1GAP are known to play a role in the progression of squamous-cell carcinoma of the head and neck. Rap-1A pathway is also associated with survival, tumor progression, and metastasis of oral cavity squamous cell carcinoma patients [73, 74].

Infectious disease-related pathways

Many cancers have been attributed to infections [75–77]. Cancers caused by infections are thought to result from one or more of the following: immune suppression, chronic inflammation, and dysregulated inflammation [78–80]. Many of these infectious disease-related pathways were found using the tissue-specific genes identified in this study. For example, the Staphylococcus aureus (gram positive bacteria) pathway was found using head and neck genes. Staphylococcus aureus is known to be present in oral squamous-cell carcinoma tissue [81] and is also abundant in the blood of oral cancer patients [82]. The infectious disease-related pathways identified using the gastric tissue genes include Type II diabetes mellitus and Chagas disease. Type II diabetes mellitus is known to increase the risk of gastric cancer [83]. Chagas disease affects several gastrointestinal regions, but there is no apparent relationship with the growing incidence of cancer [84].

Gene ontology functional analysis

A Gene Ontology-based similarity assessment indicates that the selected genes for each tissue type are functionally diverse, further validating our gene selection method.

Blood

Many of the functional groups identified are known to be involved in cancer. For example, protein-kinase inhibitor activity (GO:0004860) and negative regulation of JAK/STAT cascade (GO:0046426) groups were identified using the blood tissue genes. Tyrosine kinase inhibitors are known to be useful in the treatment of acute myeloid leukemia [85]. The JAK/STAT signaling pathway is a known target for the treatment of leukemia [86].

Breast

One of the many functional groups found by the methods of this study was the breast-cancer gene list, which includes the myosin complex (GO:0016459). Myosin is known to promote breast cancer malignancy by enhancing tumor cell proliferation [87]. Mutant p53-associated motor protein myosin upregulation is known to promote breast cancer invasiveness and metastasis [88, 89]. Myosin light-chain kinase is known to play a role in the proliferation and migration of breast cancer cells [90].

Colon

A few of the many functional groups found using our colon cancer gene list include the G-protein coupled receptor signaling pathway (GO:0007186) and GTPase activity (GO:0003924). G-protein coupled receptor kinase-5 is known to regulate proliferation and chemokine gene expression in human colon cancer epithelial cells [91]. G-protein-coupled receptors for short-chain fatty acids are known to suppress colon cancer [92]. GTPase activation is known to be present in colon cancer [93].

Gastric

The flavone metabolic process (GO:0051552) was identified using gastric tissue genes. Flavone, derived from

plants, is known to induce apoptosis in human gastric-cancer cells [94].

Lung

Significant functional groups found using lung tissue genes include the G-protein coupled receptor signaling pathway (GO:0007187) and the regulation of MAPK cascade (GO:0043408). The G protein-coupled receptor is known to promote tumorigenesis and is highly expressed in lung cancer [95]. Overexpression of G protein-coupled receptors is known to correlate with poorer tumor differentiation and higher tumor proliferation in non-small-cell lung cancer [96]. Expression of Mitogen-Activated Protein Kinase is known to present in patients with small cell lung cancer [97–100].

Biomarkers

Biomarkers can be used in clinical settings for patient assessment, estimates of morbidity, screening for cancer, distinguishing benign tissue from malignant tissue, and determination of prognosis. The sensitivity and specificity of biomarkers identified in this study exceeded those of known biomarkers for all compared tissue types, suggesting that these predicted biomarkers are robust indicators of cancer. Further research may include the testing blood-based biomarkers from the list of biomarkers under consideration for this study. For example, blood-based biomarkers have been used for diagnosis, prognosis and treatment of colorectal cancer [7, 101], breast cancer [8, 102], prostate cancer ([103], ovarian cancer [104], and lung cancer [9].

Machine learning cancer prediction models were developed to identify potential biomarkers for unbiased cancer diagnosis and effective treatment, ultimately improving prognoses. Large publicly-available tissue-specific microarray gene expression data were used for cancer type prediction, as well as characterization of tissue-specific normal samples into their various tissues of origin. A logical next step in this work would be the application of machine learning to the generation of a working model of both homeostatic and cancer developmental processes for cancer biomarker detection and early diagnosis. Such work would require collection of numerous forms of data (such as methylation, metabolic and even miRNA data) from a diverse panel of patients including but not limited to, demographic information, normal tissue controls, tumor characteristics, different forms of cancer, subtypes of cancer, and perhaps even other inflammatory diseases such as rheumatoid arthritis, from patients at varying stages of disease progression and development.

MATERIALS AND METHODS

Data collection

Microarray gene expression data were collected from NCBI Gene Expression Omnibus (GEO) repository [105]. A total of 2,175 tissue samples, both normal and cancerous, were collected from nine distinct tissues: blood (595), breast (171), colon (105), gastric (333), head and neck (82), lung (542), prostate (56), thyroid (224), and tongue (67). The detailed sample distribution is shown in Figure 13, and Supplementary Table 16. The accession numbers for the data are as follows: blood data: GSE6891, GSE267, GSE43346, GSE63270; breast data: GSE5460, GSE2361, GSE20437, GSE43346; colon data: GSE64857, GSE4107, GSE2361, GSE43346; gastric data, GSE2361, GSE43346, GSE19826, GSE62254, GSE8167; head and neck data: GSE45153, GSE10300, GSE43346, GSE8987; lung data: GSE1133, GSE10072, GSE2361, GSE43346, GSE16538, GSE19804, GSE21369, GSE24206, GSE63074; prostate data: GSE46602, GSE6369, GSE1133, GSE2361, GSE43346; thyroid data: GSE33630, GSE5054, GSE58545, GSE2361, GSE43346, GSE60542, GSE3467, GSE3678, GSE35570; tongue data: GSE52915, GSE9844, GSE1133, GSE43346. Samples used in this study were collected directly from patients according to experimental design. The frequency of data derived from tissue samples was balanced across tissue classes and entered into a composite data set. The data were collected from the following three Affymetrix Human Genome: HG-U133_Plus_2, HG-U133A, and HG-U133A_2.

Normalization and background correction

Normalization and preprocessing are essential steps for the analyses of high-throughput data including microarrays. The Affy R module 1.54 [106] from Bioconductor package (https://bioconductor.org/packages/release/bioc/html/affy.html) was used to remove the technical variation from noisy data and background noise from signal intensities. The Quantile Normalization Method [107] was used to normalize the data, and the background correction was performed using the Robust Multi-Average (RMA) [108] parameter method. Quantile normalization method relies on the assumption that observed global changes across samples are due to unwanted technical variability. We used quantile normalization since it is a simple, fast, one-size-fits-all solution for transforming all the arrays to have a common distribution of intensities. The algorithm maps every value on any one chip to the corresponding quantile of the standard distribution. The intensities of all probes on each chip into one standard distribution shape, which is determined by pooling all the individual chip distributions. We used RMA because it has a smaller standard deviation at all levels of expression compared to dChip and MAS5.0 [108].

Probe to gene mapping

Using the information provided in Affymetrix annotation files (http://www.affymetrix.com/support/technical/annotationfilesmain.affx), probe names were replaced with their respective gene names. Since multiple

probes can also correspond to the same gene, the expression values for duplicate entries were averaged within samples. All preprocessed data were randomly divided into equal-sized subsets of training and testing datasets. Since the datasets are unbalanced across classes, class distributions are approximately preserved for each tissue using stratified partitioning for training and testing sets.

Identification of best feature selection algorithm

The key to construction of accurate and unbiased machine learning models from microarray gene expression data is identification of the features (genes) best able to predict tissue class and cancer status [109]. The test set must be kept separate from the model training set Support Vector Machine (SVM) [110], IBk K-nearest neighbor [111], and Naive Bayes [112] were used to identify the best feature selection algorithm. The following 12 feature selection algorithms were used to create the models: (Chi Squared_Ranker, ClassifierSubsetEvaluator_GeneticSearch,ConsistencySubsetEvaluator_BestFirst, ConsistencySubsetEvaluator_GeneticSearch,ConsistencySubsetEvaluator_LinearFWDSelection, FilteredAttributeEvaluator_Ranker, GainRatioAttributeEvaluator_Ranker, LatentSemanticAnalysis_Ranker, OneRAttributeEvaluator_Ranker, ReliefFAttributeEvaluator_Ranker, SymmetricalUncertAttributeEval_Ranker, WrapperSubsetEval_GeneticSearch) and 13 feature thresholds (Top 1%, 2%, 3%, 4%, 5%, 10%, 20%, 25%, 33%, 50%, 66%, 75%, 100%) is shown.

Machine learning classification model construction

Machine learning classification models can be categorized into the following four groups: (1) models which, given a sample of a specific tissue type, classify the sample as cancer or normal ("single-tissue"), (2) models which, given a sample of any of the nine tissue types, classify the sample as cancer or normal ("multi-tissue bi-class") (3) models which, given a sample of any of the nine tissue types, classifies the sample as cancer or normal and as of a specific tissue type ("multi-tissue multi-class") and (4) a model which, given a normal sample of any of the nine tissue types, classifies the sample as of a particular tissue type ("multi-tissue normal multi-class"). (See Figure 13A and Supplementary Table 17 for distribution of samples among tissue types). The overall workflow of the model construction is given in Figure 14. Models were constructed using Random Forests and Support Vector Machine. The configurable CancerDiscover software pipeline [113] was used to perform all the machine learning steps in this study.

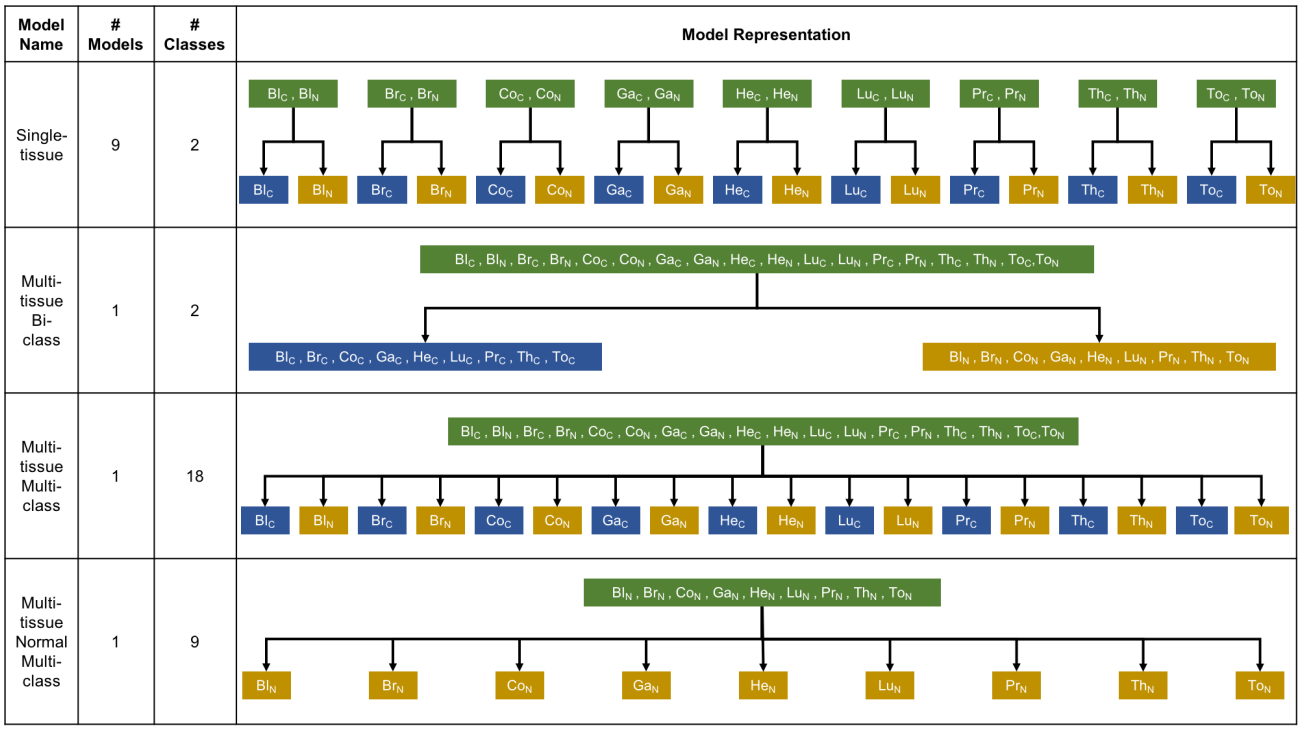

Figure 14: Types of machine learning classification model construction with the model name, the total number of models, the number of classes, and disease states of sample source for each model type. Green box: unlabeled data; blue box: cancer label; yellow box: normal label; Bl_C: blood-cancer, Bl_N: blood-normal, Br_C: breast-cancer, Br_N: breast-normal, Co_C: colon-cancer, Co_N: colon-normal, Ga_C: gastric-cancer, Ga_N: gastric-normal, He_C: head and neck-cancer, He_N: head and neck-normal, Lu_C: lung-cancer, Lu_N: lung-normal, Pr_C: prostate-cancer, Pr_N: prostate-normal, Th_C: thyroid-cancer, Th_N: thyroid-normal, To_C: tongue-cancer, To_N: tongue-normal.

Machine learning algorithms and framework

Support Vector Machines (SVMs) and Random Forests were used to construct the models for this study. These machine-learning methods were chosen because of their extensive and successful applications to datasets from genomic and proteomic domains [114, 115]. Some of the cancer classification tasks were binary (two classes), and the others were multiclass (more than two classes). Though SVMs are designed for binary classification, they can also be used for multiclass classification by a one-versus-rest approach [116]. The one-versus-rest approach for classification is known to be among the best-performing methods for multicategory classification for microarray gene expression [30]. Models were also constructed using Random Forests (RF), which can solve multicategory problems natively through direct application.

The Random Forests algorithm is well suited to the classification of genomic data because of the following advantages (i) it performs embedded feature selection (ii) it incorporates interactions between predictors: (iii) it allows the algorithm to accurately learn both simple and complex classification functions; (iv) it is applicable to both binary and multicategory classification tasks [117]. Feature selection and model construction was also accomplished using WEKA (Waikato Environment for Knowledge Analysis) [118] version 3.8.

Measures

Accuracy was defined as the overall ability of models to categorize testing sample data correctly. Reported measures included the numbers of *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FN). A true-positive count is the number of samples in a dataset which were correctly categorized *into* classes. A false-positive count is the number of samples in a dataset which were sorted into the wrong category. A true negative count represents the number of samples which were *not* classified into a class to which they do *not* belong, and false negatives are samples which are *not* classified into the class to which they do belong.

Accuracy, Sensitivity (or Recall), Specificity, Precision, and F1-score are derived from the measures mentioned above as follows: accuracy is the ratio of correctly predicted samples to the total number of samples. Sensitivity is the proportion of true positives that are predicted as positives. Specificity is the proportion of true negatives which are predicted as negatives, and Precision is the ratio of true positives to the total number of true negatives and true positives. Lastly, F1-score is defined as the harmonic mean of Precision and Recall and is calculated by first multiplying precision and recall values, then dividing the resulting value by the total of precision and recall, and finally, multiplying the result by two. The Accuracy, Sensitivity, Specificity, Precision, and F1-Score are given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall\ /\ Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Model selection and accuracy estimation

For model selection and accuracy estimation, we used 10-fold cross-validation [30, 115]. This technique separates data into ten parts and uses nine parts for the model generation while predictions are generated and evaluated by using the one part. This step is subsequently repeated ten times, so each part (internal test set) is tested against the other nine parts (internal train set). The average performance over the ten accuracies is accepted as an unbiased estimate of the model's performance.

Functional analysis

We used Database for Annotation, Visualization, and Integrated Discovery (DAVID) v6.8 [119] for functional analysis. For each of the nine tissue type provided to DAVID, the tissue-specific gene list consisting of top 244 (1%) of genes were used to classify samples of a particular tissue type as either cancerous or normal (See Supplementary File 2). Within DAVID, KEGG was chosen for pathway analysis. Of the pathways returned, only those with a *p*-value of less than or equal to 0.05 and with three or more of our genes were considered. Within DAVID, Functional Annotation analysis was used for sorting the genes according to functional groups. Of the functional groups returned, only those with a *p*-value of less than or equal to 0.05 and with three or more of the genes identified in this study were considered.

Abbreviations

FAER - Filtered Attribute Evaluator, Ranker; WEKA - Waikato Environment for Knowledge Analysis; DAVID - Database for Annotation, Visualization, and Integrated Discovery; KEGG - Kyoto Encyclopedia of Genes and Genomes; ML - Machine Learning; SVM - Support Vector Machine; RF - Random Forests; GEO - Gene Expression Omnibus; RMA – Robust Multi Average; GO – Gene Ontology; TP - True Positives; TN - True Negatives; FP - False Positives; FN - False Negatives; ROC- Receiver Operating Characteristics

Author contributions

TH and JA conceptualized the project. GB collected the data. AM and GB generated predicted biomarkers, machine learning models, performed functional annotation and pathway analysis. All authors wrote, reviewed, and revised the manuscript.

ACKNOWLEDGMENTS

We would like to thank Lara Appleby and David Pratt for the critical review of this manuscript. We would also like to acknowledge Holland Computing Center at the University of Nebraska-Lincoln for providing high-performance computing clusters for training and testing of machine learning models.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

Supported by NIH grant # 1R35GM119770-01 to TH.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J Clin. 2017; 67:7–30. https://doi.org/10.3322/caac.21387.
2. Wu L, Qu X. Cancer biomarker detection: recent achievements and challenges. Chem Soc Rev. 2015; 44:2963–97. https://doi.org/10.1039/c4cs00370e.
3. McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Hafen R, Ramey J, Rodland KD. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. Expert Opin Med Diagn. 2013; 7:37–51. https://doi.org/10.1517/17530059.2012.718329.
4. Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Informatics Assoc. 2013; 20:659–67. https://doi.org/10.1136/amiajnl-2012-001168.
5. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nat Clin Pract Oncol. 2008; 5:588–99. https://doi.org/10.1038/ncponc1187.
6. Zuo Y, Cui Y, Di Poto C, Varghese RS, Yu G, Li R, Ressom HW. INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery. Methods. 2016; 111:12–20. https://doi.org/10.1016/j.ymeth.2016.08.015.
7. Fung KYC, Tabor B, Buckley MJ, Priebe IK, Purins L, Pompeia C, Brierley GV, Lockett T, Gibbs P, Tie J, McMurrick P, Moore J, Ruszkiewicz A, et al. Blood-based protein biomarker panel for the detection of colorectal cancer. PLoS One. 2015; 10:e0120425. https://doi.org/10.1371/journal.pone.0120425.
8. Tang Q, Cheng J, Cao X, Surowy H, Burwinkel B. Blood-based DNA methylation as biomarker for breast cancer: a systematic review. Clin Epigenetics. 2016; 8:115. https://doi.org/10.1186/s13148-016-0282-6.
9. Birse CE, Lagier RJ, FitzHugh W, Pass HI, Rom WN, Edell ES, Bungum AO, Maldonado F, Jett JR, Mesri M, Sult E, Joseloff E, Li A, et al. Blood-based lung cancer biomarkers identified through proteomic discovery in cancer tissues, cell lines and conditioned medium. Clin Proteomics. 2015; 12:18. https://doi.org/10.1186/s12014-015-9090-9.
10. Yorker EE, Holdenrieder S, Gezer U. Blood-based biomarkers for diagnosis, prognosis and treatment of colorectal cancer. Clin Chim Acta. 2016; 455:26–32. https://doi.org/10.1016/j.cca.2016.01.016.
11. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl. 2009; 36:3240–7. https://doi.org/10.1016/j.eswa.2008.01.009.
12. Aliferis CF, Hardin D, Massion PP. Machine learning models for lung cancer classification using array comparative genomic hybridization. Proc AMIA Symp. 2002; 67–71. http://www.ncbi.nlm.nih.gov/pubmed/12463776.
13. Liu JJ, Cutler G, Li W, Pan Z, Peng S, Hoey T, Chen L, Ling XB. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics. 2005; 21:2691–7. https://doi.org/10.1093/bioinformatics/bti419.
14. Pirooznia M, Yang JY, Yang MQ, Deng Y, Guyon I, Weston J, Barnhill S, Vapnik V, Duan K, Rajapakse J, Wang H, Azuaje F, Liu H, et al. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008; 9:S13. https://doi.org/10.1186/1471-2164-9-S1-S13.
15. Mao Y, Zhou X, Pi D, Sun Y, Wong STC. Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. J Biomed Biotechnol. 2005; 2005:160–71. https://doi.org/10.1155/JBB.2005.160.
16. Peng Y. A novel ensemble machine learning for robust microarray data classification. Comput Biol Med. 2006; 36:553–73. https://doi.org/10.1016/j.compbiomed.2005.04.001.
17. Duan KB, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience. 2005; 4:228–33. https://doi.org/10.1109/TNB.2005.853657.
18. Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011; 12:507. https://doi.org/10.1186/1471-2164-12-507.

19. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, et al. ArrayExpress update-simplifying data submissions. Nucleic Acids Res. 2015; 43:D1113–6. https://doi.org/10.1093/nar/gku1057.

20. Gao J, Aksoy BBA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013; 6:pl1. https://doi.org/10.1126/scisignal.2004088.

21. Shi Y, Dai D, Liu C, Yan H. Sparse discriminant analysis for breast cancer biomarker identification and classification. Prog Nat Sci. 2009; 19:1635–41. https://doi.org/10.1016/j.pnsc.2009.04.013.

22. Bhowmick SS, Saha I, Maulik U, Bhattacharjee D. Biomarker identification using Next Generation Sequencing data of RNA. IEEE/ACM Trans Comput Biol Bioinform. 2016; 13:299–303.

23. Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, Sotiriou C. A three-gene model to robustly identify breast cancer molecular subtypes. J Natl Cancer Inst. 2012; 104:311–25. https://doi.org/10.1093/jnci/djr545.

24. Yousef M, Najami N, Abedallah L, Khalifa W. Computational Approaches for Biomarker Discovery. J Intell Learn Syst Appl. 2014; 6:153–61. https://doi.org/10.4236/jilsa.2014.64012.

25. Rykunov D, Beckmann ND, Li H, Uzilov A, Schadt EE, Reva B. A new molecular signature method for prediction of driver cancer pathways from transcriptional data. Nucleic Acids Res. 2016; 44:e110–e110. https://doi.org/10.1093/nar/gkw269.

26. Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. OMICS. 2013; 17:595–610. https://doi.org/10.1089/omi.2013.0017.

27. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2009; 26:392–8. https://doi.org/10.1093/bioinformatics/btp630.

28. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Gen. 2000; 24:227–35. https://doi.org/10.1038/73432.

29. Veer VL, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–6.

30. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005; 21:631–43. https://doi.org/10.1093/bioinformatics/bti033.

31. Solin LJ, Gray R, Goldstein LJ, Recht A, Baehner FL, Shak S, Badve S, Perez EA, Shulman LN, Martino S, Davidson NE, Sledge GW, Sparano JA. Prognostic value of biologic subtype and the 21-gene recurrence score relative to local recurrence after breast conservation treatment with radiation for early stage breast carcinoma: Results from the Eastern Cooperative Oncology Group E2197 study. Breast Cancer Res Treat. 2012; 134:683–92. https://doi.org/10.1007/s10549-012-2072-y.

32. Clark-Langone KM, Wu JY, Sangli C, Chen A, Snable JL, Nguyen A, Hackett JR, Baker J, Yothers G, Kim C, Cronin MT, Sorlie T, Perou C, et al. Biomarker discovery for colon cancer using a 761 gene RT-PCR assay. BMC Genomics. 2007; 8:279. https://doi.org/10.1186/1471-2164-8-279.

33. Cooperberg M, Simko J, Falzarano S, Maddala T, Chan J, Cowan J, Magi-Galluzzi C, Tsiatis A, Tenggara-Hunter I, Knezevic D, Baehner F, Kattan M, Shak S, et al. 2131 Development and Validation of the Biopsy-Based Genomic Prostate Score (Gps) As a Predictor of High Grade or Extracapsular Prostate Cancer To Improve Patient Selection for Active Surveillance. J Urol. 2013; 189:e873. https://doi.org/10.1016/j.juro.2013.02.2040.

34. Nguyen D V, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002; 18:39–50. https://doi.org/10.1093/bioinformatics/18.1.39.

35. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012; 40:D109–14. https://doi.org/10.1093/nar/gkr988.

36. Singh RK, Sivabalakrishnan M. Feature selection of gene expression data for cancer classification: A review. Procedia Comput Sci. 2015; 50:52–7. https://doi.org/10.1016/j.procs.2015.04.060.

37. Androulakis IP, Yang E, Almon RR. Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities. Annu Rev Biomed Eng. 2007; 9:205–28. https://doi.org/10.1146/annurev.bioeng.9.060906.151904.

38. Yuneva MO, Fan TWM, Allen TD, Higashi RM, Ferraris DV, Tsukamoto T, Matés JM, Alonso FJ, Wang C, Seo Y, Chen X, Bishop JM. The metabolic profile of tumors depends on both the responsible genetic lesion and tissue type. Cell Metab. 2012; 15:157–70. https://doi.org/10.1016/j.cmet.2011.12.015.

39. Liu J, Mi J, Zhou BP. Metabolic rewiring in cancer-associated fibroblasts provides a niche for oncogenesis and metastatic dissemination. Mol Cell Oncol. 2016; 3:e1056331. https://doi.org/10.1080/23723556.2015.1056331.

40. Coller HA. Is cancer a metabolic disease? Am J Pathol. 2014; 184:4–17. https://doi.org/10.1016/j.ajpath.2013.07.035.

41. Majeti R, Becker MW, Tian Q, Lee TLM, Yan X, Liu R, Chiang JH, Hood L, Clarke MF, Weissman IL. Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. Proc Natl Acad Sci U S A. 2009; 106:3396–401. https://doi.org/10.1073/pnas.0900089106.

42. Applegate CC, Lane MA. Role of retinoids in the prevention and treatment of colorectal cancer. World J Gastrointest Oncol. 2015; 7:184–203. https://doi.org/10.4251/wjgo.v7.i10.184.

43. Lidén M, Eriksson U. Understanding retinol metabolism: Structure and function of retinol dehydrogenases. J Biol Chem. 2006; 281:13001–4. https://doi.org/10.1074/jbc.R500027200.

44. Park EY, Wilder ET, Lane MA. Retinol inhibits the invasion of retinoic acid-resistant colon cancer cells *in vitro* and decreases matrix metalloproteinase mRNA, protein, and activity levels. Nutr Cancer. 2007; 57:66–77. https://doi.org/10.1080/01635580701268238.

45. Lin JH, Giovannucci E. Sex hormones and colorectal cancer: What have we learned so far? J Natl Cancer Inst. 2010; 102:1746–7. https://doi.org/10.1093/jnci/djq444.

46. Beyerle J, Frei E, Stiborova M, Habermann N, Ulrich CM. Biotransformation of xenobiotics in the human colon and rectum and its association with colorectal cancer. Drug Metab Rev. 2015; 2532:1–23. https://doi.org/10.3109/03602532.2014.996649.

47. Kaminsky LS, Zhang QY. The Small Intestine As a Xenobiotic-Metabolizing Organ. Drug Metab Dispos. 2003; 31:1520-5.

48. Bezirtzoglou EEV. Intestinal cytochromes P450 regulating the intestinal microbiota and its probiotic profile. Microb Ecol Health Dis. 2012; 23:182. https://doi.org/10.3402/mehd.v23i0.18370.

49. Carr RM, Qiao G, Qin J, Jayaraman S, Prabhakar BS, Maker AV. Targeting the metabolic pathway of human colon cancer overcomes resistance to TRAIL-induced apoptosis. Cell Death Discov. 2016; 2:16067. https://doi.org/10.1038/cddiscovery.2016.67.

50. Serra A, MacIà A, Romero MP, Reguant J, Ortega N, Motilva MJ. Metabolic pathways of the colonic metabolism of flavonoids (flavonols, flavones and flavanones) and phenolic acids. Food Chem. 2012; 130:383–93. https://doi.org/10.1016/j.foodchem.2011.07.055.

51. Liberti MV, Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? Trends Biochem Sci. 2016; 41:211–8. https://doi.org/10.1016/j.tibs.2015.12.001.

52. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009; 324:1029–33. https://doi.org/10.1126/science.1160809.

53. Dai J, Shen J, Pan W, Shen S, Das UN. Effects of polyunsaturated fatty acids on the growth of gastric cancer cells *in vitro*. Lipids Health Dis. 2013; 12:71. https://doi.org/10.1186/1476-511X-12-71.

54. Klil-Drori AJ, Ariel A. 15-Lipoxygenases in cancer: A double-edged sword? Prostaglandins Other Lipid Mediat. 2013; 106:16–22. https://doi.org/10.1016/j.prostaglandins.2013.07.006.

55. Rebbeck TR, Walker AH, Jaffe JM, White DL, Wein AJ, Malkowicz SB. Glutathione S -Transferase- • (GSTM1) and - * (GSTT1) Genotypes in the Etiology of Prostate Cancer 1. Cancer Epidemiol Biomarkers Prev. 1999; 8:283–7.

56. Gsur A, Haidinger G, Hinteregger S, Bernhofer G, Schatzl G, Madersbacher S, Marberger M, Vutuc C, Micksche M. Polymorphisms of glutathione-S-transferase genes (GSTP1, GSTM1 and GSTT1) and prostate-cancer risk. Int J Cancer. 2001; 95:152–5. https://doi.org/10.1002/1097-0215(20010520)95:3<152::AID-IJC1026>3.0.CO;2-S.

57. Tsouko E, Khan AS, White MA, Han JJ, Shi Y, Merchant FA, Sharpe MA, Xin L, Frigo DE. Regulation of the pentose phosphate pathway by an androgen receptor-mTOR-mediated mechanism and its role in prostate cancer cell growth. Oncogenesis. 2014; 3:e103. https://doi.org/10.1038/oncsis.2014.18.

58. Dong WK, Young SJ, Hye SJ, Hyo KC, Jung HS, Ki CP, Su HP, Jung HH, So YR, Gi RK, Lee SJ, Jo KW, Shong M. An orally administered multitarget tyrosine kinase inhibitor, SU11248, is a novel potent inhibitor of thyroid oncogenic RET/papillary thyroid cancer kinases. J Clin Endocrinol Metab. 2006; 91:4070–6. https://doi.org/10.1210/jc.2005-2845.

59. Cabanillas ME, Waguespack SG, Bronstein Y, Williams MD, Feng L, Hernandez M, Lopez A, Sherman SI, Busaidy NL. Treatment with Tyrosine Kinase Inhibitors for Patients with Differentiated Thyroid Cancer: the M. D. Anderson Experience. J Clin Endocrinol Metab. 2010; 95:2588–2595. https://doi.org/10.1210/jc.2009-1923.

60. Giancotti FG. Deregulation of cell signaling in cancer. FEBS Lett. 2014; 588:2558–70. https://doi.org/10.1016/j.febslet.2014.02.005.

61. Gyan E, Frew M, Bowen D, Beldjord C, Preudhomme C, Lacombe C, Mayeux P, Dreyfus F, Porteu F, Fontenay M. Mutation in RAP1 is a rare event in myelodysplastic syndromes. Leukemia. 2005; 19:1678–80.

62. Reuter CWM, Morgan MA, Bergmann L. Targeting the Ras signaling pathway: a rational, mechanism-based treatment for hematologic malignancies? Blood. 2000; 96:1655–69.

63. Zhou J, Ching YQ, Chng WJ. Aberrant nuclear factor-kappa B activity in acute myeloid leukemia: from molecular pathogenesis to therapeutic target. Oncotarget. 2015; 6:5490–500. https://doi.org/10.18632/oncotarget.3545.

64. Griessinger E, Frelin C, Cuburu N, Imbert V, Dageville C, Hummelsberger M, Sirvent N, Dreano M, Peyron JF. Preclinical targeting of NF-kappaB and FLT3 pathways in AML cells. Leukemia. 2008; 22:1466–9.

65. Cornelis M, Bosman J, Schepers H, Jaques J, Brouwers-vos AZ, Quax WJ, Schuringa JJ, Vellenga E. The TAK1-NF- k B axis as therapeutic target for AML. Blood. 2015; 124:3130–41.

66. Guzman ML. Nuclear factor-kappaB is constitutively activated in primitive human acute myelogenous leukemia cells. Blood. 2001; 98:2301–7. https://doi.org/10.1182/blood.V98.8.2301.

67. Francesconi M, Remondini D, Neretti N, Sedivy JM, Cooper LN, Verondini E, Milanesi L, Castellani G. Reconstructing networks of pathways via significance analysis of their intersections. BMC Bioinformatics. 2008; 9:S9. https://doi.org/10.1186/1471-2105-9-S4-S9.

68. Shi M, Wu M, Pan P, Zhao R. Network-based sub-network signatures unveil the potential for acute myeloid leukemia therapy. Mol Biosyst. 2014; 10:3290–7. https://doi.org/10.1039/c4mb00440j.

69. Fajardo AM, Piazza GA, Tinsley HN. The role of cyclic nucleotide signaling pathways in cancer: Targets for prevention and treatment. Cancers (Basel). 2014; 6:436–58. https://doi.org/10.3390/cancers6010436.

70. Wu M, Wu Y, Lan T, Jiang L, Qian H, Chen Y. Type II cGMP-dependent protein kinase inhibits EGF-induced JAK/STAT signaling in gastric cancer cells. Mol Med Rep. 2016; 14:1849–56. https://doi.org/10.3892/mmr.2016.5452.

71. Miyagawa Y, Ohguro H, Maruyama I, Takano Y, Yamazaki H, Ishikawa F, Metoki T, Mamiya K, Nakazawa M. Aberrantly Expressed Recoverin in Tumor Tissues from Gastric Cancer Patients. In: The Neural Basis of Early Vision. Tokyo: Springer Japan. 2003; pp 173–6.

72. Ohguro H, Odagiri H, Miyagawa Y, Ohguro I, Sasaki M, Nakazawa M. Clinicopathological features of gastric cancer cases and aberrantly expressed recoverin. Tohoku J Exp Med. 2004; 202:213–9. https://doi.org/10.1620/tjem.202.213.

73. Banerjee R, Russo N, Liu M, Van Tubergen E, D'Silva NJ. Rap1 and its regulatory proteins. Small GTPases. 2012; 3:192–7. https://doi.org/10.4161/sgtp.20413.

74. Chen CH, Chuang HC, Huang CC, Fang FM, Huang HY, Tsai HT, Su LJ, Shiu LY, Leu S, Chien CY. Overexpression of rap-1A indicates a poor prognosis for oral cavity squamous cell carcinoma and promotes tumor cell invasion via aurora-A modulation. Am J Pathol. 2013; 182:516–28. https://doi.org/10.1016/j.ajpath.2012.10.023.

75. Thun MJ, DeLancey JO, Center MM, Jemal A, Ward EM. The global burden of cancer: Priorities for prevention. Carcinogenesis. 2009; 31:100–10. https://doi.org/10.1093/carcin/bgp263.

76. Plummer M, de Martel C, Vignat J, Ferlay J, Bray F, Franceschi S. Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet Glob Health. 2016; 4:e609–16. https://doi.org/10.1016/S2214-109X(16)30143-7.

77. Parsonnet J. Bacterial infection as a cause of cancer. Environ Health Perspect. 1995; 103:263–8. https://doi.org/10.1289/ehp.95103s8263.

78. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell. 2010; 140:883–99. https://doi.org/10.1016/j.cell.2010.01.025.

79. Wang D, DuBois RN. Immunosuppression associated with chronic inflammation in the tumor microenvironment. Carcinogenesis. 2015; 36:1085–93. https://doi.org/10.1093/carcin/bgv123.

80. Goldszmid RS, Dzutsev A, Trinchieri G. Host Immune Response to Infection and Cancer: Unexpected Commonalities. Cell Host Microbe. 2014; 15:295–305. https://doi.org/10.1016/j.chom.2014.02.003.

81. Hooper SJ, Crean SJ, Lewis MAO, Spratt DA, Wade WG, Wilson MJ. Viable bacteria present within oral squamous cell carcinoma tissue. J Clin Microbiol. 2006; 44:1719–25. https://doi.org/10.1128/JCM.44.5.1719-1725.2006.

82. Panghal M, Kaushal V, Kadayan S, Yadav JP. Incidence and risk factors for infection in oral cancer patients undergoing different treatments protocols. BMC Oral Health. 2012; 12:22. https://doi.org/10.1186/1472-6831-12-22.

83. Shimoyama S. Diabetes mellitus carries a risk of gastric cancer: A metaanalysis. World J Gastroenterol. 2013; 19:6902–10. https://doi.org/10.3748/wjg.v19.i40.6902.

84. Matsuda NM, Miller SM, Evora PRB. The chronic gastrointestinal manifestations of Chagas disease. Clinics (Sao Paulo). 2009; 64:1219–24. https://doi.org/10.1590/S1807-59322009001200013.

85. Lainey E, Thépot S, Bouteloup C, Sébert M, Ads L, Tailler M, Gardin C, De Botton S, Baruchel A, Fenaux P, Kroemer G, Boehrer S. Tyrosine kinase inhibitors for the treatment of acute myeloid leukemia: Delineation of anti-leukemic mechanisms of action. Biochem Pharmacol. 2011; 82:1457–66. https://doi.org/10.1016/j.bcp.2011.05.011.

86. Chaudhari S, Desai JS, Adam A, Mishra P. Jak/Stat As a Novel Target for Treatment of Leukemia. Int J Pharm Pharmaceut Sci 2014;6: 1-7.

87. Ouderkirk-Pecone JL, Goreczny GJ, Chase SE, Tatum AH, Turner CE, Krendel M. Myosin 1e promotes breast cancer malignancy by enhancing tumor cell proliferation and stimulating tumor cell de-differentiation. Oncotarget. 2016;7:46419-46432. https://doi.org/10.18632/oncotarget.10139.

88. Arjonen A, Kaukonen R, Mattila E, Rouhi P, Hognas G, Sihto H, Miller BW, Morton JP, Bucher E, Taimen P, Virtakoivu R, Cao Y, Sansom OJ, et al. Mutant p53. associated myosin-X upregulation promotes breast cancer invasion and metastasis. J Clin Invest. 2014; 124:1069–82. https://doi.org/10.1172/JCI67280.

89. Cao R, Chen J, Zhang X, Zhai Y, Qing X, Xing W, Zhang L, Malik YS, Yu H, Zhu X. Elevated expression of myosin X in tumours contributes to breast cancer aggressiveness and metastasis. Br J Cancer. 2014; 111:539–50. https://doi.org/10.1038/bjc.2014.298.

90. Zhou X, Liu Y, You J, Zhang H, Zhang X, Ye L. Myosin light-chain kinase contributes to the proliferation and migration of breast cancer cells through cross-talk with activated ERK1/2. Cancer Lett. 2008; 270:312–27. https://doi.org/10.1016/j.canlet.2008.05.028.

91. Raghavendra PB, Parameswaran N. Novel signaling role of G protein-coupled receptor kinase-5 in human colon cancer cell line. FASEB J. 2013; 27:949.9–949.9.

92. Tang Y, Chen Y, Jiang H, Robbins GT, Nie D. G-protein-coupled receptor for short-chain fatty acids suppresses colon cancer. Int J Cancer. 2011. 128:847–56. https://doi.org/10.1002/ijc.25638.

93. Ogier-Denis E, Pattingre S, El Benna J, Codogno P. Erk1/2-dependent phosphorylation of G??-interacting protein stimulates its GTPase accelerating activity and autophagy in human colon cancer cells. J Biol Chem. 2000; 275:39090–5. https://doi.org/10.1074/jbc.M006198200.

94. Kim MJ, Kim DH, Na HK, Oh TY, Shin CY, Surh YJ. Eupatilin, a Pharmacologically Active Flavone Derived from Artemisia Plants, Induces Apoptosis in Human Gastric Cancer (AGS) Cells. J Environ Pathol Toxicol Oncol. 2005; 24:261–70. https://doi.org/10.1615/JEnvironPatholToxicolOncol.v24.i4.30.

95. Nishimura S, Uno M, Kaneta Y, Fukuchi K, Nishigohri H, Hasegawa J, Komori H, Takeda S, Enomoto K, Nara F, Agatsuma T. MRGD, a MAS-related g-protein coupled receptor, promotes tumorigenisis and is highly expressed in lung cancer. PLoS One. 2012; 7:e38618. https://doi.org/10.1371/journal.pone.0038618.

96. Nii K, Tokunaga Y, Liu D, Zhang X, Nakano J, Ishikawa S, Kakehi Y, Haba R, Yokomise H. Overexpression of G protein-coupled receptor 87 correlates with poorer tumor differentiation and higher tumor proliferation in non-small-cell lung cancer. Mol Clin Oncol. 2014; 2:539–44.

97. Vicent S, Garayoa M, López-Picazo JM, Lozano MD, Toledo G, Thunnissen FBJM, Manzano RG, Montuenga LM. Mitogen-activated protein kinase phosphatase-1 is overexpressed in non-small cell lung cancer and is an independent predictor of outcome in patients. Clin Cancer Res. 2004; 10:3639–49. https://doi.org/10.1158/1078-0432.CCR-03-0771.

98. Blackhall FH, Pintilie M, Michael M, Cell S, Cancer L, Leighl N, Feld R, Tsao M, Shepherd FA. Expression and Prognostic Significance of Kit, Protein Kinase B, and Mitogen-activated Protein Kinase in Patients with Small Cell Lung Cancer. Clin Cancer Res. 2003; 9:2241–7.

99. Papadimitrakopoulou V, Adjei AA. The Akt/mTOR and mitogen-activated protein kinase pathways in lung cancer therapy. J Thorac Oncol. 2006; 1:749–51. https://doi.org/10.1016/S1556-0864(15)30399-3.

100. Yan H, Zhu Y, Liu B, Wu H, Li Y, Wu X, Zhou Q, Xu K. Mitogen-activated protein kinase mediates the apoptosis of highly metastatic human non-small cell lung cancer cells induced by isothiocyanates. Br J Nutr. 2011; 106:1779–91. https://doi.org/10.1017/S0007114511002315.

101. Fung KYC, Purins L, Priebe IK, Pompeia C, Brierley GV, Tabor B, Lockett T, Gibbs P, Tie J, McMurrick P, Moore J, Ruszkiewicz A, Burgess A, et al. Analysis of 32 Blood-Based Protein Biomarkers for their Potential to Diagnose Colorectal Cancer. J Mol Biomark Diagn. 2014; S6:1–7.

102. Zawadzka AM, Schilling B, Cusack MP, Sahu AK, Drake P, Fisher SJ, Benz CC, Gibson BW. Phosphoprotein secretome of tumor cells as a source of candidates for breast cancer biomarkers in plasma. Mol Cell Proteomics. 2014; 13:1034–49. https://doi.org/10.1074/mcp.M113.035485.

103. Liong ML, Lim CR, Yang H, Chao S, Bong CW, Leong WS, Das PK, Loh CS, Lau BE, Yu CG, Ooi EJJ, Nam RK, Allen PD, et al. Blood-Based Biomarkers of Aggressive Prostate Cancer. PLoS One. 2012;7:e45802. https://doi.org/10.1371/journal.pone.0045802.

104. Visintin I, Feng Z, Longton G, Ward DC, Alvero AB, Lai Y, Tenthorey J, Leiser A, Flores-Saaib R, Yu H, Azori M, Rutherford T, Schwartz PE, et al. Diagnostic markers for early detection of ovarian cancer. Clin Cancer Res. 2008; 14:1065–72. https://doi.org/1078-0432.CCR-07-1569.

105. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30:207–10. https://doi.org/10.1093/nar/30.1.207.

106. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy - Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004; 20:307–15. https://doi.org/10.1093/bioinformatics/btg405.

107. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19:185.

108. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4:249–64. https://doi.org/10.1093/biostatistics/4.2.249.

109. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23:2507–17. https://doi.org/10.1093/bioinformatics/btm344.

110. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995; 20:273–97. https://doi.org/10.1023/A:1022627411411.

111. Cover TM, Hart PE. Nearest Neighbor Pattern Classification. IEEE Trans Inf Theory. 1967; 13:21–7. https://doi.org/10.1109/TIT.1967.1053964.

112. Rish I. An empirical study of the naive Bayes classifier. IJCAI 2001 Work Empir Methods Artif Intell. 2001; 335:41–2. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.330.2788.

113. Mohammed A, Biegert G, Adamec J, Helikar T. CancerDiscover: A configurable pipeline for cancer prediction and biomarker identification using machine learning framework. bioRxiv. 2017. http://www.biorxiv.org/content/early/2017/08/31/182998.article-info.

114. Statnikov A, Wang L, Aliferis C. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics. 2008; 9:319. https://doi.org/10.1186/1471-2105-9-319.

115. Mohammed A, Guda C. Application of a hierarchical enzyme classification method reveals the role of gut microbiome in human metabolism. BMC Genomics. 2015; 7:S16. https://doi.org/10.1186/1471-2164-16-S7-S16.

116. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995; 20:273–97. https://doi.org/10.1007/BF00994018.

117. Bishop CM. Pattern Recognition And Machine Learning. J Elect Imag. 2007; 738. https://doi.org/10.1117/1.2819119.

118. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. SIGKDD Explor Newsl. 2009; 11:10. https://doi.org/10.1145/1656274.1656278.

119. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane CH, Lempicki RA, Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4:P3. https://doi.org/10.1186/gb-2003-4-9-r60.