LITERATURE MINING SUSTAINS AND ENHANCES KNOWLEDGE DISCOVERY FROM OMIC STUDIES

by

Rick Matthew Jordan

B.S. Biology, University of Pittsburgh, 1996

M.S. Molecular Biology/Biotechnology, East Carolina University, 2001

M.S. Biomedical Informatics, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Rick Matthew Jordan

It was defended on

December 2, 2015

and approved by

Shyam Visweswaran, M.D., Ph.D., Associate Professor

Rebecca Jacobson, M.D., M.S., Professor

Songjian Lu, Ph.D., Assistant Professor

Dissertation Advisor: Vanathi Gopalakrishnan, Ph.D., Associate Professor

LITERATURE MINING SUSTAINS AND ENHANCES KNOWLEDGE DISCOVERY FROM OMIC STUDIES

Rick Matthew Jordan, M.S.

University of Pittsburgh, 2016

Genomic, proteomic and other experimentally generated data from studies of biological systems aiming to discover disease biomarkers are currently analyzed without sufficient supporting evidence from the literature due to complexities associated with automated processing. Extracting prior knowledge about markers associated with biological sample types and disease states from the literature is tedious, and little research has been performed to understand how to use this knowledge to inform the generation of classification models from 'omic' data. Using pathway analysis methods to better understand the underlying biology of complex diseases such as breast and lung cancers is state-of-the-art. However, the problem of how to combine literature-mining evidence with pathway analysis evidence is an open problem in biomedical informatics research.

This dissertation presents a novel semi-automated framework, named Knowledge Enhanced Data Analysis (KEDA), which incorporates the following components: 1) literature mining of text; 2) classification modeling; and 3) pathway analysis. This framework aids researchers in assigning literature-mining-based prior knowledge values to genes and proteins associated with disease biology. It incorporates prior knowledge into the modeling of experimental datasets, enriching the development process with current findings from the scientific community.

New knowledge is presented in the form of lists of known disease-specific biomarkers and their accompanying scores obtained through literature mining of millions of lung and breast cancer abstracts. These scores can subsequently be used as prior knowledge values in Bayesian modeling and pathway analysis. Ranked, newly discovered biomarker-disease-biofluid relationships which identify biomarker specificity across biofluids are presented. A novel method of identifying biomarker relationships is discussed that examines the attributes from the best-performing models. Pathway analysis results from the addition of prior information, ultimately lead to more robust evidence for pathway involvement in diseases of interest based on statistically significant standard measures of impact factor and p-values.

The outcome of implementing the KEDA framework is enhanced modeling and pathway analysis findings. Enhanced knowledge discovery analysis leads to new disease-specific entities and relationships that otherwise would not have been identified. Increased disease understanding, as well as identification of biomarkers for disease diagnosis, treatment, or therapy targets should ultimately lead to validation and clinical implementation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

ACKNOWLEDGEMENTS

I thank my advisor Vanathi Gopalakrishnan Ph.D. for helping and encouraging me down the path to completion of my dissertation. Her knowledge and experience were most valuable, and were of utmost importance in completion of my work.

Thank you to my committee members, Drs. Shyam Visweswaran, M.D., Ph.D. and Songjian Lu, Ph.D. for their encouragement, guidance, and most valuable suggestions. Thank you to Dr. Rebecca Jacobson, M.D., MS for serving on my committee, and for pushing me to completion.

Thank you to Toni Porterfield for the many times that she has helped make my life so much easier during my time at DBMI. And thank you Dr. James Lyons-Weiler for the opportunities, ideas, encouragement, and everything else that you have done for me.

Thank you to Jeya Balasubramanian, for his help with the Bayesian Rule Learner. I would also like to thank Haiwen Shi for providing me with the matching algorithm and Rahil Sethi for Perl scripts that I used.

Thank you to the people at Windber Research Institute, University of Pittsburgh Genomic and Proteomic Core Labs, Gaskin Forensics, and Molecular DX, for financial support, encouragement and in some cases, participation and helpfulness.

I thank my family and friends for understanding, encouragement, and helpfulness over the last decade.

To my parents thanks for your belief in me, and encouragement to achieve my goals. Thank you for instilling in me the values and qualities that have made me the person that I am today.

To my children, Jeb, Dane, and Jenna, I hope this can be an example to you that nothing worthwhile comes easy, and if you want something, to never give up until you achieve it.

And finally, to my wife Courtney, who has lived this experience along with me. Words cannot express my gratitude of everything you have done for me and our family. I’m certain I couldn’t have done this without you.

GLOSSARY

Area under the Receiver Operating Characteristic curve (AUC) – the Receiver Operating Characteristic (ROC) curve, is a plot of the True Positive Rate (TPR; Sensitivity or Recall) on the y-axis against the False Positive Rate (FPR; 1 – Specificity) on the x-axis. The area under the ROC curve (AUC) is a measure of a model's discriminative performance.

Bayesian network – is a directed acyclic graph where nodes represent variables and edges represent conditional dependencies. Each node has a probability attached to it which is conditional on the probability of the probability of its parent's nodes. Bayesian networks can be used to assign a probability to an event that has not happened yet based on prior knowledge of other events that have occurred.

binning – is a data pre-processing technique used to pool similar performing entities into a few groups. The pooled group can then be represented by one value instead of many. Binning can simplify large datasets into smaller groups for comparisons, however data loss usually results.

BioCreative corpora - is a large and structured set of texts used to perform assessments for evaluating text mining and information extraction activities using gene ontology annotation terms of human proteins.

breadth-first marker propagation – algorithm for traversing or searching tree structures. It starts at the tree root and explores the neighbor nodes first, before moving to the next level neighbors. Breadth-first searches require only one pass through the tree for model training, and thus are faster than other methods that do not.

conditional independencies - two events are conditionally independent if knowledge of one event occurring provides no information on the likelihood of the other occurring.

cross-validation - a modeling validation process where a dataset is subdivided into smaller training and test sets. Different combinations of samples are created by altering the groupings of samples. Training and testing of different combinations of samples are performed several times and the accuracies averaged. Cross-validation is performed to evaluate how a model will generalize to other datasets.

data mining - a bioinformatics technique for analyzing data and databases to discover significant patterns or relationships between biological entities such as genes or proteins

data parameters – a model consists of *variables* and *parameters* that attempt to explain relationships among variables. Variables are quantities that can be measured, and parameters are constants of essential properties (materials, equipment, or measures of central tendency, to name a few) of a given experiment.

data transformation – is where a mathematical function is applied to each data point in a dataset with a given probability distribution in order to convert the entire dataset into a different probability distribution.

decision rules – if/then expressions that represent a dependency between a condition and a decision. In modeling, these rules can be used to predict classification of subjects into groups.

enrichment analysis – a type of analysis where groups of genes or proteins are studied together to assign biological meaning to the group. The group is usually clustered together as a result of function, location, or some other area of interest. Analyzing as groups enables new biological patterns to emerge, or to determine whether a subset shows similar expression of a biological characteristic, or might belong to similar biological pathways.

entropy ranking methods – in literature mining, biomedical entities are ranked by frequency or relative entropy. Frequency ranking orders the entities with respect to the number of citations. The relative entropy ranking calculates the fraction between the documents containing the entity in the result set and the total number of documents containing the entity in the complete Medline document set.

false positive – in statistics, a false positive (type I error) occurs when the null hypothesis (H_0) is true, but is rejected. A false positive concludes that something exists when it truly does not (believe a falsehood).

feature selection – is the process of selecting a subset of the most relevant features for use in model construction. This is performed to eliminate redundant or irrelevant features found in datasets. Feature selection can be performed to reduce the amount of time computer algorithms spend analyzing irrelevant features.

Gene Ontology (GO) – is a bioinformatics naming convention that organizes gene and protein information. Gene Ontology provides structured terminology of gene and protein properties such as cellular components, molecular functions, and biological processes.

gene symbol disambiguation – in literature mining, resolving conflicts that arise from ambiguous gene names. The same gene may poses several aliases, while the same identifier may refer to two very different biological entities, such as ER referring to estrogen receptor, or emergency room.

greedy search – an algorithm that uses a heuristic that makes the locally optimal choice at each stage to search a space of classification models. In many cases, a greedy strategy does not produce an optimal solution, but may produce a locally optimal solution that approaches a global optimal solution in a reasonable amount of time. Finding a truly optimal solution may require many steps and a significant time investment.

hypergeometric distribution- is a probability distribution that describes the probability of k successes in n draws, *without* replacement, where each draw is either a success or a failure. The hypergeometric test uses the hypergeometric distribution to calculate the statistical significance of having drawn a specific number of k successes (out of n total draws). The test is used to identify which sub-populations are over- or under-represented in a sample.

local rule learning (local structure search) – used to learn Bayesian network structures. In local learning, one or more target variables of special interest are examined; the local structure of the target variable is of greater interest than the other variables. Each node in a tree model corresponds to an instance of a problem. At each node the local network structure is used to create a partial solution to the problem.

loop condition - conditional loops are a way for computer programs to repeat one or more steps depending on a condition. The 'while' loop and 'for' loop are the two most common types of conditional loops in most programming languages.

MEDLINE - (Medical Literature Analysis and Retrieval System Online) is a literature database of biomedical information maintained by the US National Library of Medicine. MEDLINE contains more than 18 million records from over 5,000 scientific publications from 1950 to the present. Each MEDLINE record is manually indexed with NLM's controlled vocabulary, known as Medical Subject Headings (MeSH).

NLPBA – acronym for Natural Language Processing in Biomedical Applications.

non-small cell lung cancer – a type of lung cancer consisting of adenocarcinomas, large cell carcinomas, and squamous cell carcinomas.

oncogene - a gene that when activated promotes tumorous cell growth.

ontology - a naming convention used to organize information. It can be used to define how to represent relationships among objects, concepts, and other entities belonging to a particular area of expertise. Gene ontology specifies processes, functions, and cellular locations of gene products.

over-fitting – occurs when a model with too many parameters produces a good fit with the sample data but a poor fit with new data.

parallel decision tree – in data mining, decision trees are tree-like models on which decisions are based. A decision tree is an undirected graph where edges exist between every two vertices. However, all attributes need to be sorted in order to choose the appropriate node at which to split the tree. Decision trees can require considerable amounts of time, memory, and computational resources when utilizing large data sets. Dividing the dataset into smaller pieces allows for parallel processing of trees resulting in increased speed, and fewer computational resource requirements.

PMID – acronym for PubMed Identification that is a unique identifier assigned to every article indexed in PubMed.

precision – in information retrieval, precision is the fraction of documents returned that are relevant. (precision = relevant / retrieved)

pre-processing parameters - pre-processing of data is an important step in the data mining process, to eliminate false, missing, or noisy data values. Data pre-processing includes filtering, normalization, transformation, and feature selection.

proto-oncogene – a normal gene that once mutated, has the potential to become an oncogene.

pruning step – in computer science, a technique that reduces the size of decision trees by removing sections that add little information to the final classifier. Pruning reduces the complexity of the classifier and improves the accuracy of the prediction by reducing the possibility of overfitting.

PubMed - a web-based literature retrieval service provided by the US National Library of Medicine. PubMed provides access to several biomedical literature databases, with MEDLINE being the largest.

recall – (a.k.a. sensitivity) in information retrieval, recall is the fraction of relevant documents that are retrieved. (recall = relevant documents retrieved / total number of relevant documents)

search heuristic - a heuristic is a technique for obtaining results of a search faster than existing methods, or for finding an immediate estimated solution when existing methods cannot produce a true solution. A heuristic, in general, sacrifices accuracy for speed.

sensitivity – (a.k.a. true positive rate) in experimental science, sensitivity of a test is the number of diseased people that are identified as such by the test, compared to the total number of diseased people tested. (sensitivity = diseased identified by test / total number of diseased tested).

small cell carcinoma - a type of lung cancer that is highly malignant, composed of small ovoid undifferentiated cells.

specificity – (a.k.a. true negative rate) in experimental science, specificity of a test is the number of healthy people identified as such by the test. (specificity = healthy identified by test / total number of healthy tested).

PREFACE

Many of the figures and text contained in this work can also be found in the Journal of Clinical Bioinformatics article titled ‘Semi-automated literature mining to identify putative biomarkers of disease from multiple biofluids’ (Jordan *et al.* 2014).

1.0 INTRODUCTION

In 2014, cancer surpassed heart disease as the leading cause of death worldwide, with 8.2 million deaths and 14.1 million new cancer cases reported (World Cancer Report 2014). Furthermore, worldwide cancer deaths are predicted to increase well into the future, with lung cancer currently being the leading cause of deaths in males, and breast cancer, the leading cause of deaths in females.

Lung cancer is the leading cause of cancer deaths worldwide, and the most common cancer in terms of incidence. In 2008, there were 1.61 million new cases, and 1.38 million deaths due to lung cancer, with the highest rates occurring in Europe and North America (Ferlay *et al.* 2010). The most common cause of lung cancer is long-term exposure to tobacco smoke (Merck Manual). Across the developed world, 91% of lung cancer deaths in men during the year 2000 were attributed to smoking and 71% for women (Peto *et al.* 2006). Lung cancer carries with it an unfortunate prognosis, due to the fact that it is usually not discovered until symptoms arise (~75%). Early stage (stages 0-3) diagnoses offer 80% - 10% five year survival rates, whereas late-stage survival averages <10% (Collins *et al.* 2007).

Breast cancer is the most common invasive cancer in women globally, comprising 22.9% of all cancers in women (World Cancer Report 2008). In 2008, breast cancer caused 458,503 deaths worldwide (13.7% of all cancer deaths in women) (World Cancer Report 2008). Unlike lung cancer, no specific activity can be attributed to causing breast cancer, with the majority of breast cancer cases, >80%, being classified as non-hereditary or sporadic. However, increases in incidence vary significantly around the world being lower in less-developed countries and greater in the well-developed countries; modern lifestyles have been implicated in causation (Laurance 2006). Breast cancer conveys much better prognoses compared to lung cancer, as it is more often discovered earlier, with early stage (stages 0-3) diagnoses providing 93% - 41% five year survival rates, while late stage survival averages 15% (Imaginis Corporation 2006).

While the number of worldwide cancer deaths have been increasing annually, cancer rates have decreased due to awareness and early detection methods. Early detection or screening is vital to surviving cancer. The most common lung cancer screening methods include low-dose spiral (helical) CT (Computed Tomography) chest scans, chest x-ray, and sputum cytology. Of these, helical CT appears to show the most promise as 20.3% fewer lung cancer deaths occur among those who were screened with low-dose helical CT compared with those who were screened with chest x-rays (http://www.cancer.gov/images/DSMB-NLST.pdf). This is due to helical CT using X-rays to obtain a multiple-image scan of the entire chest, while a standard chest X-ray produces a single image of the whole chest in which anatomic structures overlie one another (http://www.cancer.gov/news-events/press-releases/2011/NLSTprimaryNEJM). The most common breast cancer screening methods include self-exams, mammograms, clinical exams, and breast imaging via CT and MRI (Magnetic Resonance Imaging). In one study, breast cancer screening and management programs have been shown to improve survival rate ~18% (Kalager 2009). However, while current screening methods have led to better survival rates, millions of people still succumb to cancer each year; and room for improvement on current screening methods remains.

More recent disease screening methods have shown promise in the areas of molecular genomic/proteomic ('omic') testing for individual gene mutations, proteins and gene panels.

Biomarkers, which can be any measurable biological characteristic or substance that indicates a particular condition or process, now play a crucial role, with the emergence of personalized medicine. Individual biomarkers are currently tested for early detection of hereditary breast cancer (BRCA1/*breast cancer 1*, BRCA2/*breast cancer 2*, ESR1/*estrogen receptor*, PGR/*progesterone receptor*, HER2/*erb-b2 receptor tyrosine kinase 2*, and PARP/*poly (ADP-ribose) polymerase 1*); and heterogeneous nuclear ribonucleoprotein (hnRNP), a tumor-associated antigen found in what appears to be normal lung epithelium in lung cancer cases. Moreover, the relatively new field of Pharmacogenomics analyzes an individuals' genetic makeup to determine how they will metabolize or respond to certain drugs. This relatively new field combines pharmacology and genomics to develop effective, safe medications and doses that will be uniquely tailored to a given individual (https://ghr.nlm.nih.gov/primer/genomicresearch/pharmacogenomics). Examining the allelic makeup of individual genes can help physicians with disease management and treatment.

Multi-biomarker panels such as OncotypeDx (Lyman *et al.* 2007) for breast, colon, and prostate cancers; and PAM50 (Parker *et al.* 2009) for breast cancer, have become more common as increases in US Food and Drug Administration (FDA) approval have occurred. Since 2003 the FDA has cleared or approved ~1000 biomarker-based tests, including 139 tests that measure two or more biomarkers (http://www.amplion.com/biomarker-trends/biomarker-panels-the-good-the-bad-and-the-ugly/). However, with an average of less than 100 biomarker tests per year being approved for thousands of diseases, new methods of discovering biomarkers are desperately needed.

Biomarkers have been discovered in many different ways, ranging from clinical observation, to literature/data mining, disease modeling, gene clustering, and analysis of disease-related biological pathways. Literature mining provides a wealth of information as the amount of scientific knowledge continues to grow exponentially; however, the amount of information can also be a hindrance due to its enormity. Data mining identifies differential gene or protein expression in numerous disease states, but the findings are only representations of a moment in time when the sample was taken. Disease modeling provides precise measurements such as accuracy, sensitivity, and specificity, but a given model may not generalize to other/larger populations. Pathway analysis allows for visualization of biological processes and protein interactions, but is a snapshot of only one pathway in an intertwined biological network of pathways that interact and depend on one another. Hence, knowledge discovery (the process of discovering useful knowledge from a collection of data) frameworks which improve experimental findings by utilizing information obtained from several methods should be more desirable, reliable, and accurate. More accurate experimental findings should lead to more likely biomarker possibilities and better disease understanding at the molecular level, and ultimately in clinical improvements in prevention, maintenance, and treatment.

1.1 THE PROBLEM

New approaches are needed to improve methods which ultimately lead to biomarker discovery. One possibility is combining prior scientific knowledge from literature mining and experimental data to improve knowledge discovery, and subsequent disease modeling and pathway analysis. Prior knowledge can exist in many different forms such as facts, theories, beliefs, diagrams, charts, measurements, and calculated values. Several challenges are encountered when attempting to organize prior knowledge from literature mining: 1) obtaining and organizing information, 2) combining prior knowledge with experimental data, and 3) interpreting downstream results. Organizing and presenting others' findings into a single format is non-trivial. A carefully designed plan which can accommodate all types of reported findings is needed. Converting the prior knowledge in a way that will attribute a weighted significance to prior knowledge in relation to experimental data poses another issue. Lastly, many other issues arise in the interpretation of modeling and pathway analysis results.

1.1.1 Obtaining and organizing prior information from literature mining

In organizing prior knowledge several issues must be accounted for: 1) information source, 2) search space, 3) experimental design, and 4) scoring method. All of these issues will affect the outcome of the study, and must be clearly defined to allow accurate downstream interpretation.

Information source

Determining the type of text, as well as the source is an important first step in obtaining prior knowledge. Many types of information exist, such as meta-data, clinical reports, original research articles, review articles, opinion papers, books, and editorials. A researcher needs to determine which type of information can reveal the most relevant findings. Text availability is of concern because the full-text of all articles are not freely available; those that are not must be purchased from the publisher. Performing an exhaustive search of a substantial-sized corpus of full-text articles is currently cost-prohibitive ranging from $10-$40 per article, on average. Abstracts on the other hand, are easily-acquired and freely-available, and their size enables easy storage and processing.

Search space

Clearly defining a search space will allow for a truly representative and exhaustive search. Defining the search space provides the study specificity by limiting the information included to only that which is truly relevant. Without defining a search space, calculated values pertaining to system's performance measures such as accuracy, recall, and precision, cannot be correctly determined and could actually be incorrect. Many filters can be applied during this step to limit unwanted information from being obtained. Additional factors which may or may not be included in a search, not mentioned previously include: disease, disease stage, tissue type or biofluid, tumor type, gene or protein of interest, location in the body, treatments, age, sex, and diet.

Experimental design

Careful experimental design is crucial for any scientific study. All aspects of a study must be clearly defined such as: sample size, sample stratification, analysis method and relevant calculations, reporting of results, and others. Common issues known to occur in most experiments must be taken into account such as how exceptions, biases, confounding, and false positives will be handled. For example, in literature mining, false positives may be eliminated by using a set of articles not pertinent to the topic at hand (also known as the 'negative set'). An error in the experimental design phase may be fatal to the entire study.

Scoring method

Quantifying evidence from literature about disease-biomarker associations in the form of prior knowledge as numeric scores is challenging for several reasons. Defining what constitutes positive and negative findings is not a trivial process. Common practice in literature mining employs the use of a 'gold standard'. Gold standards may include previously established lists or dictionaries, actual expert assessments, or comparison of results to a previously annotated corpus. Obtaining a gold standard source can be quite difficult when studying less-researched diseases, or in new method assessments. Determining how mentions of one or more biological entities will be counted presents another issue. The question arises, should each mention be counted independently, or should all redundant mentions of an entity in one abstract be pooled together to provide a single count? Additional complications include score normalization, gene/protein ambiguity, the use of abbreviations versus full-length terms, as well as negation.

1.1.2 Combining prior knowledge with experimental data

Obtaining information and assigning a score to that information constitutes the first step in the process. Converting scores into meaningful values and incorporating those values into datasets follows. Transforming the values into prior probabilities poses a challenge as many possibilities exist such as using raw values, ranking the scores, weighting, or binning. Similarly, several options exist in incorporating knowledge into the dataset, such as including the values as a separate column in the data, adding or multiplying the experimental data values by the prior value, or any combination of the transformation and incorporation methods mentioned.

Additionally, the type of experimental data being used also comes with additional concerns. Array based data (microarray, protein array) is a common type of experimental data produced by researchers today. This allows for examination of thousands of biological entities (genes, proteins, miRNAs, and SNPs) simultaneously. However a statistical anomaly exists when comparing thousands of entities among tens or hundreds of samples in what is called the ‘multiple comparison problem’. This problem results in an increase in the false discovery rate when the set is considered as a whole (Benjamini & Hochberg 1995). Missing data measurements presents another problem, with possible solutions being to remove the entire entity from analysis, or use other measurements to impute a number to fill in the blank. None of the possible solutions are ideal. Lastly, many data analysis software packages were not created to handle millions of values simultaneously. This creates further issues with the best solution being feature selection. In feature selection, entities that do not appear to be relevant are removed from the analysis, thus decreasing noise and speeding up the analysis.

1.1.3 Interpreting downstream results

Machine learning algorithms have the ability to learn data patterns and use that information to make predictions. Modeling algorithms have been used to predict the stratification of patient samples by examining biological or environmental metrics. Standard models are usually accompanied by measures of accuracy, sensitivity, and specificity, in order to assess the models strengths and weaknesses in its classifications. While these classifiers have greatly enhanced the ability to make predictions, they are somewhat complicated; to many, they are a black box, which could lead to incorrect interpretation of important results. A concern exists in examining hundreds or thousands of entities; a model may appear to achieve desired levels of accuracy, sensitivity, and specificity, but may use so many modeling attributes that the findings may be a result of overtraining and the results may be biased. Another issue is that a model that correctly classifies samples using one dataset does not mean it will generalize well to other datasets, or the general population.

Pathway analysis provides a visual interpretation of known biological pathways and protein-protein interactions. However, many pathway analyses are misunderstood and enable some to draw false conclusions. A target gene/protein found to be upstream in a disease pathway or process may be thought to cause or influence other important downstream genes/proteins, but this may not be the case. Also a possibility exists that pathway information may not be sufficiently updated enough to reflect an entire process. Additionally, not every pathway may be represented, interactions between pathways may not be known, and only known pathways can be examined. Very few tool exist to predict new pathways.

Both, classification modeling and pathway analyses require follow-up and validation studies, requiring additional time and financial investment. For all of the reasons presented, one can see why biomarkers are lacking for many diseases. Different approaches and out-of-the-box thinking are needed to take biomarker discovery processes into the future. Defined methods or frameworks that can save time and obtain the most up-to-date and relevant information, which can then be utilized by well-informed researchers are highly desirable.

1.2 THE APPROACH

This dissertation provides a framework for combining prior knowledge with experimental datasets to ultimately aid in biomarker discovery. The overall goal is to speed up the time investment required to obtain prior knowledge, and then use this knowledge to uncover new relationships via classification models and identify pathway genes/proteins that otherwise would be missed by conventional analyses.

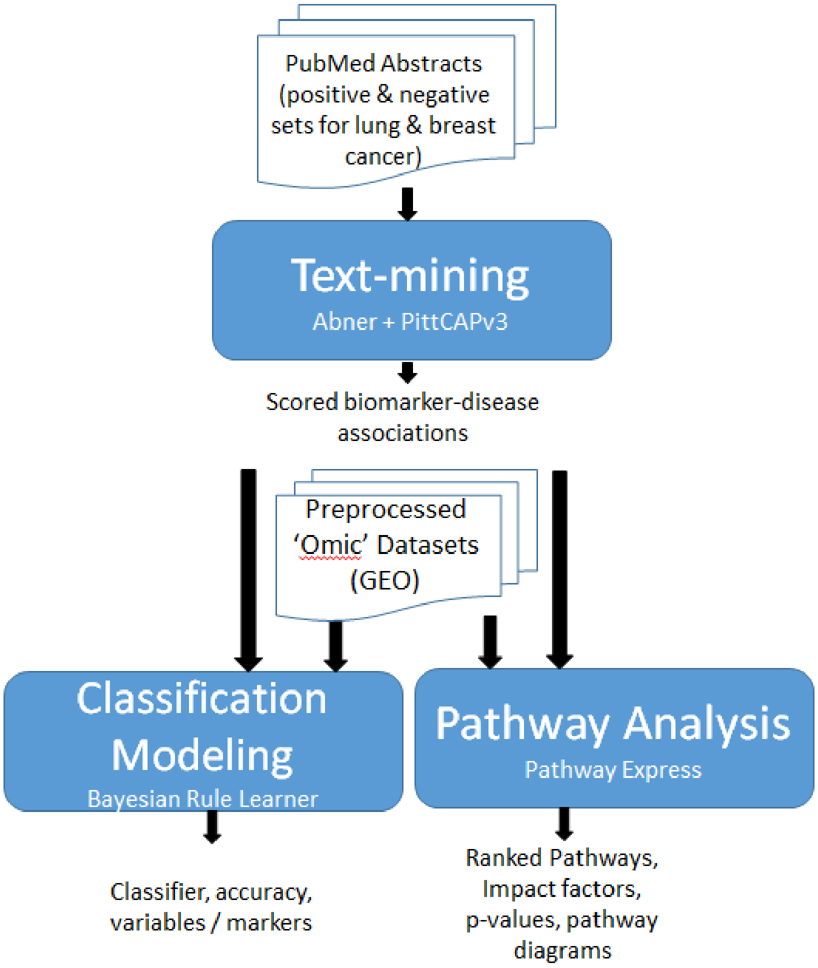

PubMed Abstracts (positive & negative sets for lung & breast cancer)
Text-mining
Abner + PittCAPv3
Scored biomarker-disease associations
Preprocessed 'Omic' Datasets (GEO)
Classification Modeling
Bayesian Rule Learner
Pathway Analysis
Pathway Express
Classifier, accuracy, variables / markers
Ranked Pathways, Impact factors, p-values, pathway diagrams

Figure 1. The KEDA framework components and information flow. The semi-automated text-mining component of KEDA takes as input positive and negative sets of abstracts returned from keyword searches of literature databases. The text-mining component outputs gene/protein names and counts obtained from all abstracts in each abstract set. The counts are transformed into prior probabilities; and preprocessed datasets are obtained. For classification modeling, the priors are incorporated into the datasets and used as input into a modeling algorithm, which outputs classifiers, model accuracy values, and the markers used to build the models. For pathway analysis, the datasets are analyzed first to obtain a single expression score per gene/protein. Each score is multiplied by the prior probability to obtain an updated expression value reflecting prior knowledge. These values accompany the gene/protein names in a list that is input into pathway analysis software which outputs pathway impact factors, p-values, and diagrams for each pathway.

The Knowledge Enhanced Data Analysis (KEDA) framework as shown in Figure 1 incorporates the following components: 1) text-mining of literature; 2) classification modeling; and 3) pathway analysis. The generalized flow of information is described in Figure 1. The current implementation of the KEDA framework is described as follows: in text-mining of the literature, research abstracts from PubMed (http://www.ncbi.nlm.nih.gov/pubmed) were examined for gene/protein mentions, scores were produced for each entity, and the scores transformed into prior probabilities.

Publicly-available datasets were normalized if needed, and features were selected using the J5 method (Patel *et al.* 2004) in caGEDA (Patel & Lyons-Weiler 2004). Remaining feature data was matched with prior probabilities, the datasets formatted if needed, and input into the Bayesian Rule Learner algorithm (Gopalakrishnan *et al.* 2010) for disease modeling exercises. In the modeling step, Bayesian networks (BNs) containing target nodes with zero parents are examined initially. Variables are added to the BN's as parent nodes and scored. The best-scoring BN's are retained on a beam for further analysis, stored by score. In this way, only the highest scoring BN structures that retain the ability to improve are explored. The beam is checked for models where the score can be improved by the addition of another parent variable. Greedy searches are performed by adding one more variable as an additional parent of the target, and scores are recalculated to see if the score of the model improved with the addition of a new parent variable. The process continues until Bayesian scores cannot improve further and the best scoring model is presented to the user in the form of a rule model (Gopalakrishnan *et al.* 2010). Model performance measures (accuracy, sensitivity, and specificity) were analyzed to determine the method's performance. Attributes of the best-performing models were compared to known disease biomarker lists to uncover novel relationships.

In the pathway analysis step, prior probabilities were multiplied to post-data analysis J5 results and input into Pathway Express. Pathway performance measures (number of input genes in pathway, impact factor, and p-value) were analyzed to determine the method's performance. Each pathway was visually examined to compare the number of genes and individual genes/proteins that were identified by the different methods examined.

1.2.1 Theses

The central thesis is that the KEDA Framework is sufficient for incorporating knowledge from literature mining into disease modeling from omic datasets and to enhance the results from pathway analyses.

Based on experiments performed on several array-based lung and breast cancer publicly available datasets of various experimental types, complexity, and sizes, the following specific claims are made:

Claim 1: The text-mining component in KEDA is a sufficient method of obtaining putative biomarkers, assigning a prior knowledge score per biomarker, and estimating biomarker specificity for biofluids.

Claim 2: **A)** Incorporation of prior information from literature mining does not degrade classifier modeling performance, on average.

B) Analyzing the attributes used to build the best-performing classification models leads to new biological relationships being uncovered.

C) Incorporation of prior information enhances pathway analysis results by identifying more input genes in disease-relevant pathways.

1.3 SIGNIFICANCE

From a bioinformatics perspective, this work is significant for a number of reasons:

A framework is developed and evaluated that utilizes a semi-automated text-mining method to produce a list of documented putative biomarker/biofluid relationships from millions of abstracts. If desired, researchers can apply the described methods to their own diseases of interest. The list of known disease-specific biomarkers, created as the gold standard, is novel, and was compiled from several databases for this work. Researchers now have a list to use for validation of their own work, when researching breast and/or lung cancer. Very few published works examine more than one or two biofluids at a time, whereas this work examines 14 biofluids simultaneously. Additionally, ranked, newly discovered biomarker-disease-biofluid relationships are presented; as well as biomarker specificity across biofluids. The genes/proteins presented are assumed to be disease specific (breast and lung cancer), and their accompanying z-scores are also novel. Researchers looking for values to use as informed priors now have a resource, and do not have to invest the time and effort to produce them, as it has already been done for them.

A new method of identifying possible biomarker relationships by examining the attributes from the best-performing models is described. Pathway analysis results were enhanced by the addition of prior information. Improved pathway analysis should ultimately lead to more robust disease-specific biomarkers, as well as improved disease-specific knowledge discovery. It is not apparent the extent to which others have investigated prior knowledge incorporation in pathway analysis. The pathway results presented here may be the first to show the improvement obtained by incorporation of prior knowledge.

1.4 DISSERTATION OVERVIEW

The rest of this work is organized as follows: Chapter 2 provides background information concerning lung and breast cancer, text-mining phases, modeling, pathway analysis, and current use of prior knowledge in molecular biology. Chapter 3 discusses the methodology used. Chapter 4 discusses the experimentation and evaluation methods, and examines the results. Chapter 5 is a discussion section focusing on conclusions and future work.

2.0 BACKGROUND

In this chapter, as this work centers on lung and breast cancer, background statistics are provided and a molecular perspective included. Additional background is given on the several phases of literature mining, as it is the process used to obtain prior knowledge. An introduction to Bayesian modeling is also presented, as this is the modeling method chosen for this work. Last, summary of the numerous ways that prior knowledge is currently being used in molecular biology is given.

2.1 STATISTICS & MOLECULAR ASPECTS OF LUNG AND BREAST CANCER

Lung Cancer

Lung cancer is characterized by uncontrolled cell growth in lung tissues. Most primary lung cancers are carcinomas, which are derived from epithelial cells. The two major forms of lung cancer are non-small cell lung cancer (NSCLC; 85% of total lung cancer cases) and small cell lung cancer (SCLC; 15% of total lung cancer cases). Non-small cell lung carcinomas (NSCLC) can be stratified into squamous cell lung carcinoma, adenocarcinoma, and large cell lung carcinoma subtypes. Squamous cell lung carcinomas account for 25% of lung cancers (Travis 2002), usually start near a central bronchus (Figure 2), and often grow more slowly than other cancer types (Vaporciyan *et al.* 2000). Adenocarcinoma accounts for 40% of non-small cell lung cancers (Travis 2002), and usually originate in peripheral lung tissue. Most cases of adenocarcinoma are associated with smoking, but adenocarcinoma is the most common form of lung cancer (Subramanian & Govindan 2007) among non-smokers as well (Horn *et al.* 2012).

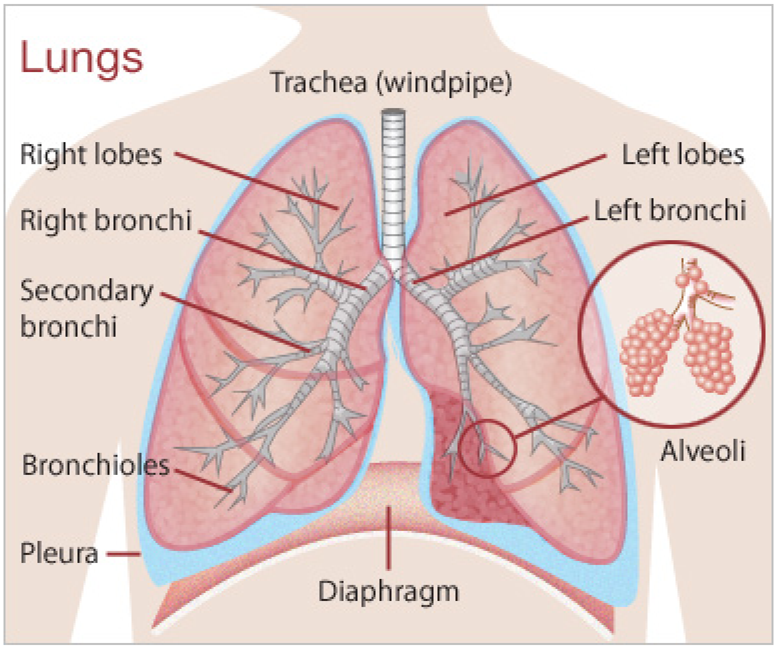


Figure 2. Diagram of the lung (www.abc.net.au).

Most small-cell lung carcinomas (SCLC) arise in the larger airways (primary and secondary bronchi; Figure 2) (Collins *et al.* 2007), grow quickly and spread early. 60–70% of SCLCs have metastatic disease at presentation, and are strongly associated with smoking (Horn *et al.* 2012).

Non-smokers account for 15% of lung cancer cases (Thun *et al.* 2006, 2008), which can be attributed to genetic factors (Gorlova *et al.* 2007; Hackshaw *et al.* 1997), radon gas (Catelinois *et al.* 2006), asbestos (O'Reilly *et al.* 2007), and air pollution (Kabir *et al.* 2007; Coyle *et al.* 2006; Chiu *et al.* 2006) including secondhand smoke (Carmona 2006; WHO, 2002). Second-hand smoking is a major cause of lung cancer in non-smokers. Studies have shown a significant increase in relative risk among those exposed to secondhand (passive) smoke (CDC

1986; Boffetta *et al.* 1998; Department of Health 1998; NHMRC 1994). Recent research of passive smoke suggests that it may be more dangerous than direct smoke inhalation (Schick & Glantz 2005). Radon is a colorless, odorless gas created from the breakdown of radium, which is the decay product of uranium. Radiation decay products can ionize DNA, causing mutations that can turn cancerous. Radon exposure is the second major cause of lung cancer, after smoking (Catelinois *et al.* 2006). Radon levels vary by location due to the composition of the soil and rocks. The Environmental Protection Agency (EPA) estimates that one in 15 homes in the U.S. has radon levels above the recommended guideline (EPA 2006). Asbestos, a silicate material, can cause a variety of lung diseases, including lung cancer. In the UK, asbestos accounts for 2–3% of male lung cancer deaths (Darnton *et al.* 2006). Asbestos can also cause mesothelioma. Outdoor air pollution has a small effect on increasing the risk of lung cancer. Fine particulates and sulfate aerosols, which may be released in traffic exhaust fumes, are associated with slightly increased risk (Alberg & Samet 2010; Chen *et al.* 2008). Outdoor air pollution is estimated to account for 1–2% of lung cancers (Alberg & Samet 2010). Factory and power plant emissions also pose potential risks (Kabir *et al.* 2007; Chiu *et al.* 2006).

The lung is a common place for metastasis from other parts of the body. Primary lung cancers most commonly metastasize to the adrenal glands, liver, brain, bone (Vaporciyan *et al.* 2000; Tan & Zander 2008), opposite lung, and kidneys (Greene 2002).

Non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC) prognoses are usually poor. The overall five-year survival for patients with SCLC is about 5% (Horn *et al.* 2012). Patients with more advanced SCLC have an average five-year survival rate of less than 1%. The median survival time for early-stage disease is 20 months, with a five-year survival rate of 20% (Merck Manual). According to the National Cancer Institute, the median age at diagnosis of lung cancer in the United States is 70 years (SEER 2010), and the median age at death is 72 years. Lung cancer, is the second most common cause of cancer-related death in women, causing ~12.0% of cancer deaths in women annually (Catelinois *et al.* 2006). The age group most likely to develop lung cancer is over-fifty with a history of smoking. The mortality rate in men has been declining for more than 20 years, while women's lung cancer mortality rates have been rising steadily over the last decades, but have recently begun to stabilize (Jemal *et al.* 2004). Eastern Europe has the highest lung cancer mortality among men, while northern Europe and the U.S. have the highest mortality among women. Lung cancer incidence is currently less common in developing countries (WHO 2004). With increased smoking in developing countries, the incidence is expected to increase in the next few years, notably in China (Zhang *et al.* 2011) and India (Behera & Balamugesh 2004).

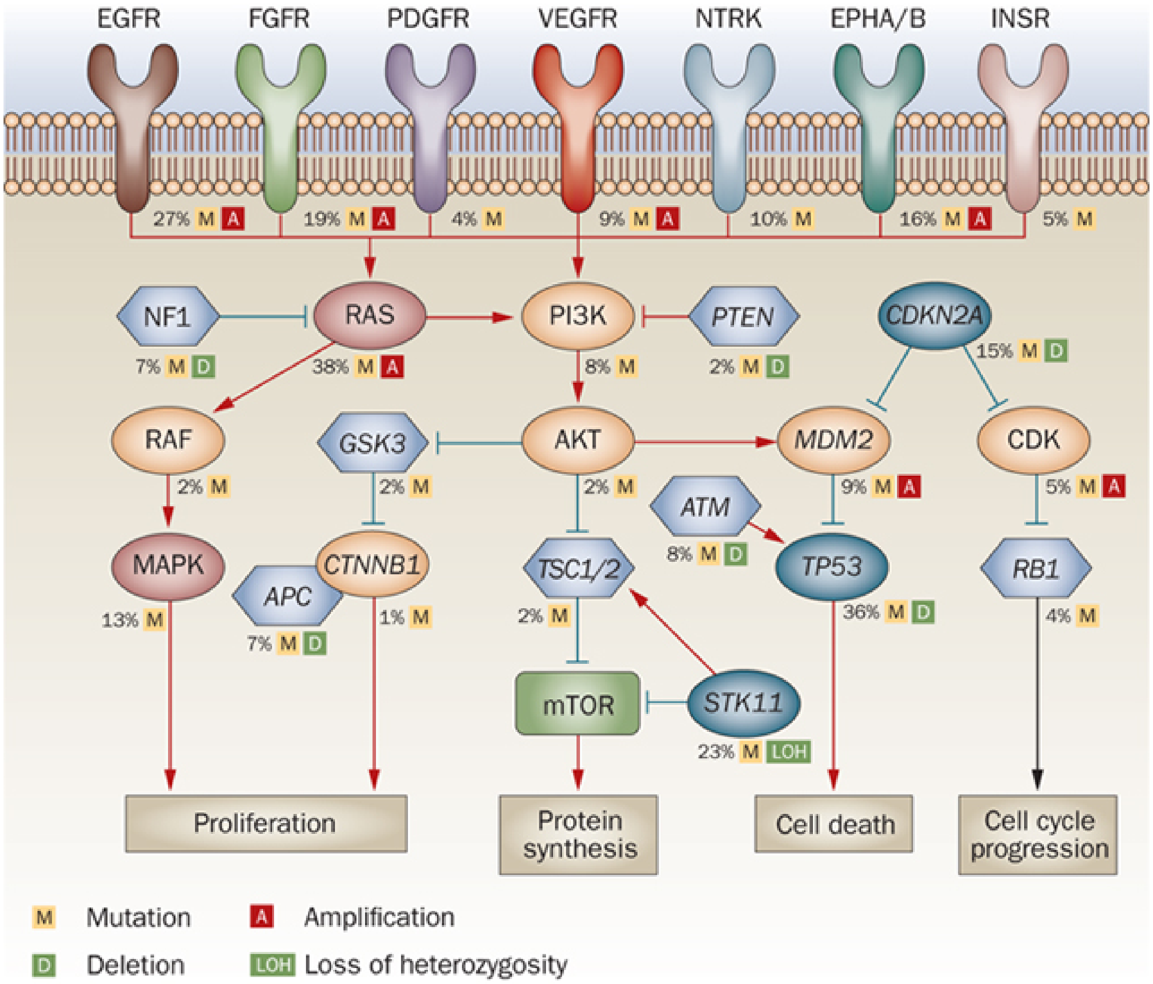


Figure 3. Mutated pathways in lung adenocarcinomas (Harris & McCormick 2010)

The main causes of any cancer include carcinogens (such as those in tobacco smoke), ionizing radiation, and viral infection. Exposures cause cumulative changes to the DNA in the epithelial lining of the lungs, and as more tissue becomes damaged, cancer develops (Vaporciyan *et al.* 2000). Also, nicotine seems to depress the immune response to malignant growths in exposed tissue (Sopori 2002). There is a genetic predisposition to lung cancer. In relatives of people with lung cancer, the risk is increased 2.4 times, and may be due to genetic polymorphisms (Kern & McLennan 2008). Lung cancer is initiated by activation of oncogenes (genes enabling susceptibility to cancer) or inactivation of tumor suppressor genes (Fong *et al.*

2003). Proto-oncogenes turn into oncogenes when exposed to particular carcinogens (Salgia & Skarin 1998). Mutations in the K-ras/*Kirsten rat sarcoma viral oncogene homolog* proto-oncogene are responsible for 10–30% of lung adenocarcinomas (Herbst *et al.* 2008; Aviel-Ronen *et al.* 2006). The EGFR/*epidermal growth factor receptor* regulates cell proliferation, apoptosis, angiogenesis, and tumor invasion. Mutations and amplification of EGFR are common in non-small-cell lung cancer (Figure 3) (Herbst *et al.* 2008). Chromosome damage can lead to loss of heterozygosity, and can cause inactivation of tumor suppressor genes. Damage to chromosomes 3p, 5q, 13q, and 17p are common in small-cell lung carcinoma (Qaiser 2012). The TP53/*tumor protein p53* tumor suppressor gene, located on chromosome 17p, is affected in 60-75% of cases (Devereux *et al*. 1996). Other genes that are often mutated or amplified are MET/*MET proto-oncogene, receptor tyrosine kinase*, NKX2-1/*NK2 homeobox 1*, STK11/*serine-threonine kinase 11*, PIK3CA/*phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit alpha*, and BRAF/*B-Raf proto-oncogene, serine/threonine kinase* (Herbst *et al.* 2008).

Breast Cancer

Breast cancer is a type of cancer that originates from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk (Sariego 2010). Cancers originating from ducts are known as ductal carcinomas; those originating from lobules are known as lobular carcinomas (Figure 4). While the majority of cases are women, men can also develop breast cancer.

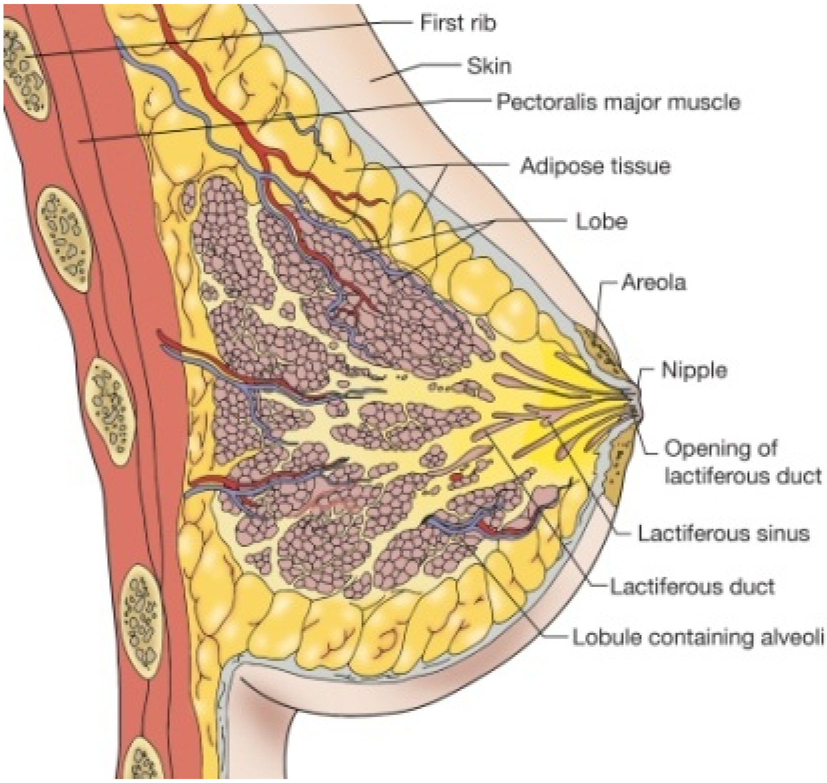


Figure 4. Diagram of the breast. (www.impressive-breast.com/blog/anatomy-female-breast/)

A woman's risk of breast cancer is increased if her mother, sister, or daughter had breast cancer, and the risk becomes significant if at least two close relatives had breast and/or ovarian cancer (Medew 2010). Family history accounts for ~10% of the cases, in general. In hereditary breast cancer syndrome, 10-20% of patients with breast cancer have a first- or second-degree relative with this disease. The most well-known of these, the BRCA1 and BRCA2 mutations, confer a 60-85% lifetime risk of breast cancer (http://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet). Approximately 2% of the female population carries the BRCA1 or BRCA2 gene mutation (Wooster & Weber 2003). The inherited mutation in BRCA1 or BRCA2 genes can interfere with or inhibit the repair of DNA cross links and DNA double strand breaks (Patel *et al.* 1998; Marietta *et al.* 2009; Theruvathu *et al*. 2005). Because of repair deficits, the risks from carcinogens and ionizing radiation can increase (Friedenson 2000; Friedenson 2012), allowing more mutations, which can lead to uncontrolled division, lack of attachment, and metastasis to distant organs (Dunning *et al.* 1999).

A woman who has had breast cancer in one breast is at an increased risk of getting a second breast cancer. Later age of first birth and not having children account for ~30% of U.S. breast cancer cases. Factors correlated with higher income contributed to 19% of cases (Madigan *et al.* 1995). Atypical hyperplasia and lobular carcinoma in situ (LCIS), which are found in benign breast conditions, are correlated with an increased breast cancer risk. Those with a normal body mass index (BMI) at age 20 who gained weight as they aged had nearly double the risk of developing breast cancer after menopause, compared to women who maintained their weight (NCI 2010). Hormone replacement therapy significantly increases the incidence of breast cancer (Sulik 2010). Additional risk factors include: being female (Giordano *et al.* 2004), choosing not to have children or breastfeed (Collaborative Group on Hormonal Factors in Breast Cancer 2002), increased hormone levels (Yager & Davidson 2006; Santoro *et al.* 2009), race, iodine deficiency (Venturi 2001; Aceves *et al.* 2005; Stoddard *et al.* 2008), high-fat diet (Chlebowski *et al. 2*006), alcohol intake (Boffetta *et al*. 2006), obesity, estrogen exposure (Cavalieri *et al.* 2006), radiation exposure (ACS 2005; Feig & Hendrick 1997; National Research Center for Women & Families 2009), shiftwork (WHO 2007), and other risk factors (Begg *et al. 2*008).

Breast cancer is around 100 times more common in women than in men, but males usually have poorer outcomes due to delays in diagnosis (World Cancer Report 2008; NCI 2011). Pre-menopausal women tend to have a worse prognosis than post-menopausal women (Peppercorn 2009). Unfortunately, sometimes breast cancer is not discovered until it has already metastasized. Common sites of breast cancer metastasis include bone, liver, lung and brain

(Lacroix 2006). The World Cancer Research Fund estimated that 38% of breast cancer cases in the US are preventable through reduced alcohol intake, increased physical activity, maintaining a healthy weight, and breastfeeding of children (Eliassen *et al.* 2010; American Institute for Cancer Research/ World Cancer Research Fund). Carcinogens take advantage of deficiencies in biological pathways that require normal BRCA1 and BRCA2 function. Avoiding these carcinogens reduce the risks for BRCA1/2 mutation carriers (Friedenson 2010).

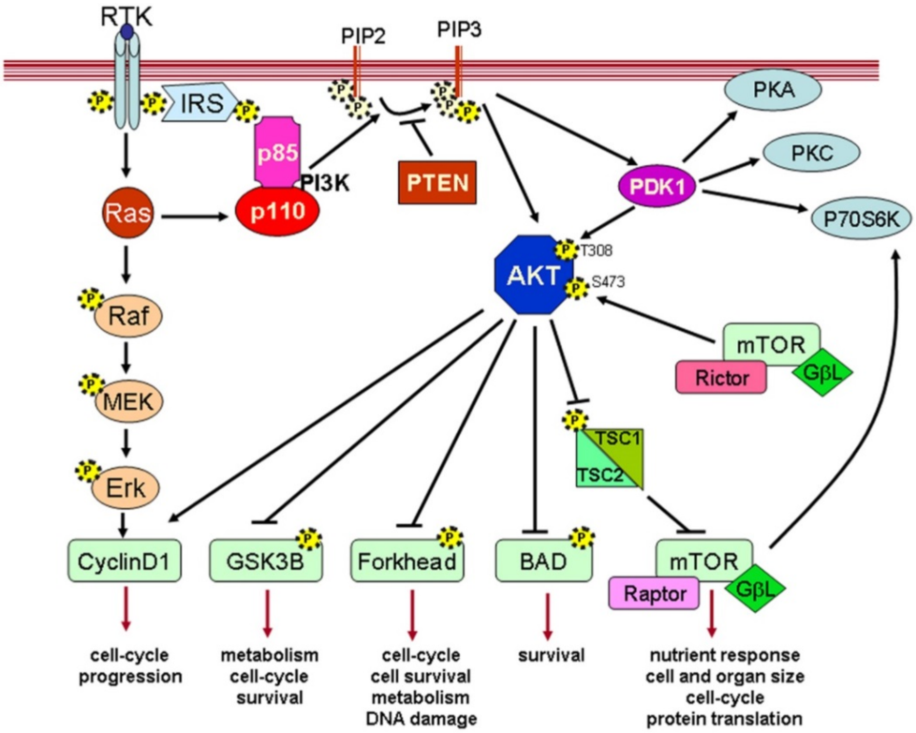


Figure 5. PI3K/AKT pathway diagram (journal.frontiersin.org).

Normal cells commit apoptosis when they are no longer needed. Until then, they are protected from cell suicide by several protein pathways. Two of the protective pathways are the PI3K/AKT pathway (Figure 5) and the RAS/MEK/ERK pathway (Figure 6). If the genes in these protective pathways are mutated, turning them permanently "on", the cell is incapable of committing suicide when it is no longer needed (SABCS 2009). Normally, the PTEN/*phosphatase and tensin homolog* protein turns off the PI3K/AKT pathway when the cell is ready for apoptosis. In some breast cancers, the gene encoding the PTEN protein is mutated, rendering the PI3K/AKT pathway being stuck in the "on" position, and the cancerous cell can no longer commit suicide (SABCS 2009).

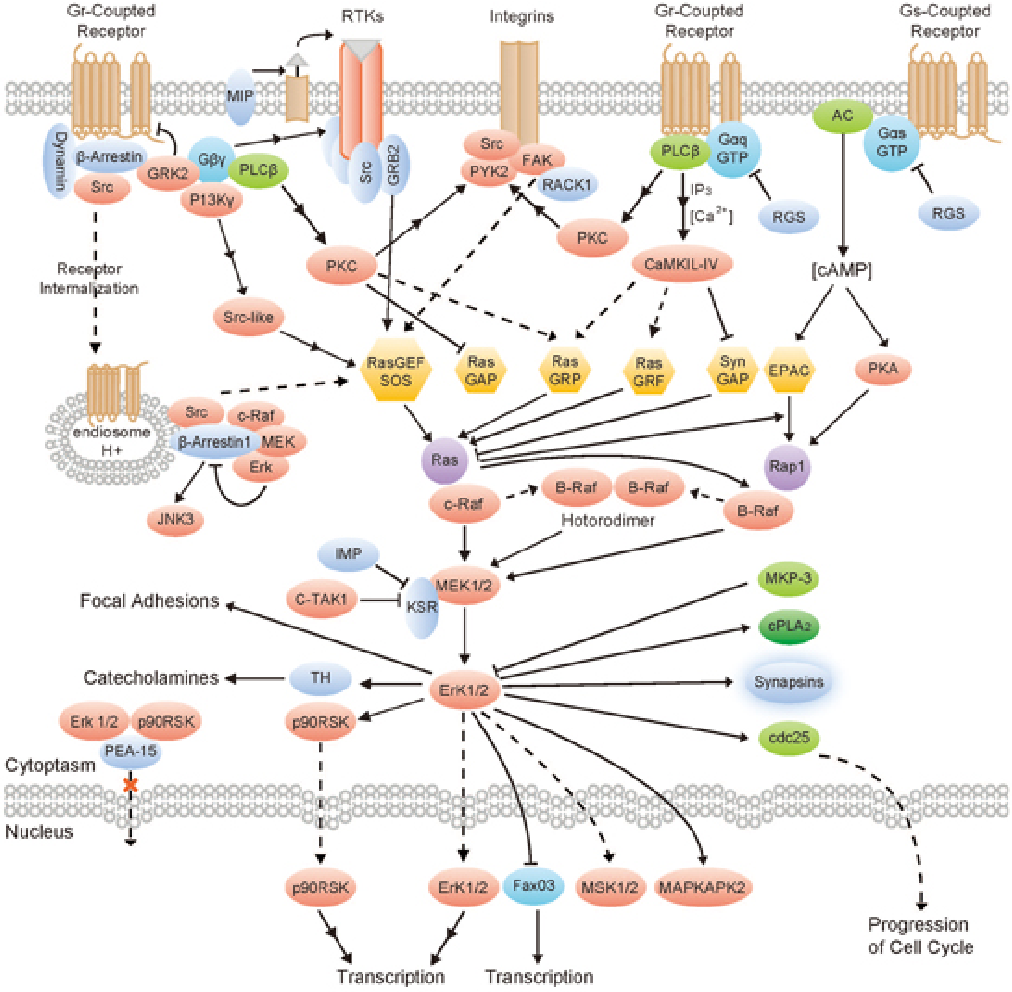


Figure 6. RAS/MEK/ERK pathway diagram (www.medchemexpress.com).

Breast cells have receptors on the surface, in the cytoplasm and on the nucleus. Hormones bind to these receptors causing cellular changes. Breast cancer cells may have estrogen receptor (ESR1), progesterone receptor (PGR), and HER2 receptors, or any combination of the three. Cells without these receptors are called triple negative, however they may possess other hormone receptors such as androgen and prolactin receptors.

2.2 LITERATURE MINING

Scientific information has become overwhelming in its extent and size, creating querying difficulties for scientists and physicians, as the literature mining process can be described as tedious at best. Many literature mining methods have been described (Adamic *et al.* 2002; Hirschman *et al.* 2002; Leonard *et al.* 2002; Novichkova *et al.* 2003; Srinivasan 2004; Wren *et al.* 2004; Cohen & Hersh 2005; Hristovski *et al.* 2005; Jensen *et al.* 2006; Xuan *et al.* 2007; and Krallinger *et al.* 2008, among others), and have created a solid foundation for future literature mining researchers.

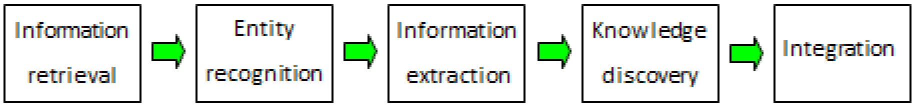


Figure 7. Literature mining process. *Information retrieval* - identify subset of articles from a much larger collection; *Entity recognition* - identifying biological entities (genes/proteins); *Information extraction* - identify relationship between a pair of biological entities; *Knowledge discovery* – aka ‘hypothesis generation’; drawing connections for novel relationships; *Integration* - integrate literature findings with other data types; potential for making biological discoveries.

Literature mining consists of several activities: information retrieval, entity recognition, information extraction, knowledge discovery, and integration. Delineation of the boundaries between the components is sometimes difficult, especially between information extraction and knowledge discovery (Figure 7).

INFORMATION RETRIEVAL

'Information retrieval' is the term given to the process of identifying relevant information. This information may be articles, abstracts, full text papers, or book chapters. Information retrieval is used to identify a subset of articles from a much larger collection. In text mining, information retrieval can be used to automatically extract features of interest from a set of documents. These features can in turn be used in combination with other algorithms to separate documents into relevant (positive) and non-relevant (negative) sets.

The two most common types of information retrieval techniques are, 'Boolean' and 'Vector Model'. The Boolean method retrieves all documents that contain user-defined keywords using the Boolean logic operators 'AND', 'OR', and 'NOT'. In the vector model, each term is assigned a value according to a frequency-based weighting scheme (Jensen *et al.* 2006).

Vector documents can be compared to a query that specifies the relative importance of each query term. Vectors can also be used as input for machine learning methods trained to discriminate between positive and negative documents by word content (Jensen *et al.* 2006). PubMed (http://www.pubmed.org) uses both the Boolean and vector models. Google Scholar (http://scholar.google.com) uses a ranking system for retrieval that ranks based on weighting of the full text, title, author, publication, and other citations in the literature (Beel & Gipp 2009).

Information retrieval has been heavily studied (Wilbur & Yang 1996; Stapley & Benoit 2000; Donaldson *et al.* 2003; and Kayaalp *et al.* [online]). MedMiner (Tanabe *et al.* 1999) (http://www.discover.nci.nih.gov/host/1999_medminer_abstract.jsp), XplorMed (Perez-Iratxeta *et al.* 2001) (http://bioinformatics.ca/links_directory/tool/10185/xplormed), Textpresso (Muller *et al.* 2004)

(http://www.textpresso.org), PubFinder (Goetz & von der Lieth 2005) (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160190/), and GeneInfoMiner (Xuan *et al.* 2005) are tools that have all been developed to aid in information retrieval from scientific literature.

NEGATIVE ABSTRACT SETS

False positive elimination from text mining findings can be aided by the use of negative abstract sets, which are abstracts that are not specifically about the relationship of interest. Descriptions of the implementation and use of negative sets of abstracts is sparse in the literature. This fact is somewhat puzzling due to the standard use of control sets in experimental design, in general. Nonetheless, the use of negative abstracts has been implemented in this work. This is a significant contribution of this work that not many others have explored in text mining. One of the benefits of using a negative set is the elimination of having to use other computational methods to minimize false discovery. A drawback is that bias that may exist in abstract selection.

A literature search identified only a few biomedical text mining papers that describe the use of negative sets of abstracts (Andrade & Valencia 1998; Adamic *et al.* 2002; Al-Mubaid & Singh 2005; Deyati *et al.* 2012; and Younesi *et al.* 2012).

Adamic (2002) described a statistical approach for finding novel gene-disease relationships. A frequency of occurrence count was discussed for relevant abstracts compared to a random set. Gene pairs and gene symbol disambiguation results were compared to a manually-edited breast cancer gene database.

Al-Mubaid and Singh (2005) covered a method for discovering protein-disease associations from MEDLINE abstracts. They employed a protein/disease dictionary and "positive" and "negative" abstract sets. The positive set was disease-relevant abstracts, determined by a PubMed keyword search; the negative set was a random set of abstracts that did not mention the disease. Their method identified disease-relevant proteins by comparing the frequency distributions of protein names in the positive set and the total set (union of the positive and negative sets), and selected proteins where the frequency distributions were statistically significantly different.

Andrade and Valencia (1998) annotated biological functions of protein sequences. In this article, the 'treatment of text with statistical methods' was discussed. The authors estimated word significance from a protein family set of abstracts by comparing each word's abundance and distribution to a background set of protein family abstracts.

Younesi *et al.* (Younesi *et al.* (2012); Deyati *et al.* 2012) divided the biomarker terminology into six concept classes (clinical management; diagnostics; prognosis; statistics; evidence; and antecedent). This extra level of stratification significantly reduced the number of retrieved relevant documents. Frequency and entropy ranking methods were used for acquired genelists with frequency ranking performing better than entropy ranking.

ENTITY RECOGNITION

Named entity recognition (NER), is the term given to the process of identifying biological entities (genes/proteins) mentioned in text. While the process may sound simple enough, a closer look reveals difficulties that exist. One of the main problems is a lack of standardization (Jensen *et al.* 2006). A complete dictionary of all biological entities does not exist. A given biological entity may have several names, abbreviations, or multi-word names. Similarly, the same word or phrase can refer to different entities (Cohen & Hersh 2005). Dealing with such ambiguity is not trivial. NER is possibly the most difficult task in biomedical text mining and is a prerequisite for both information extraction (IE) and information retrieval (IR) (Jensen *et al.* 2006).

NER systems are typically measured in terms of precision (P; correct predictions/total predictions) and recall (R; correct predictions/number of named entities in the text) (Cohen & Hersh 2005). Precision and recall often are combined into a F-score (F = 2PR / [P+R]) (Perez-Iratxeta *et al.* 2005), or can also be reported by balancing precision and recall levels (Cohen & Hersh 2005).

Rule-based methods (Fukuda *et al.* 1998; Narayanaswamy *et al.* 2003; and Tanabe & Wilbur 2002) and machine-learning algorithms using gene and protein tagged corpora (Tanabe & Wilbur 2002) (AbGene, P=85.7%, R66.7%); Collier *et al.* 2000; Zhou *et al.* 2004 (P=66.5%, R=66.6%); McDonald & Pereira 2005; and (Settles 2005; ABNER, http://www.cs.wisc.edu/~bsettles/abner, P=74.5%, R=77.8%) have been described. Dictionary (lexicon)-based methods (Donaldson *et al.* 2003; Chiang *et al.* 2004; Yu & Agichtein 2003; Cohen 2004; Liu & Friedman 2003; Yu *et al.* 2002; Schwartz & Hearst 2003; Chang *et al.* 2002) primarily used for synonym and abbreviation extraction have been extensively studied. Combinations of dictionaries with rule-based/statistical methods (Leonard *et al.* 2002; Mika &

Rost 2004; Finkel *et al.* 2005; and Chang *et al.* 2004 (GAPSCORE, http://bioinformatics.oxfordjournals.org/content/20/2/216.full.pdf+html, P=74%, R=81%) have been developed to reduce false positives. Other methods have been used to resolve ambiguity in biological names (Narayanaswamy *et al.* 2003; Eriksson *et al.* 2002; Hanisch *et al.* 2005) (IHOP; http://www.ihop-net.org/UniPub/iHOP/help.html).

Overall, the performance of state-of-the-art gene and protein NER systems achieve F-scores between 75 and 85 percent. While performance measures have not increased over the past few years, investigators are obtaining very consistent results using a variety of different approaches on different data sets (Cohen & Hersh 2005).

INFORMATION EXTRACTION

The goal of information extraction is to identify a relationship between a pair of biological entities. While the entity type is usually very specific (genes, proteins, and drugs) the relationship type can be general (biochemical association) or specific (regulatory relationship) (Cohen & Hersh 2005). Two methods of information extraction are common, co-occurrence or frequency-based scoring methods and NLP-based methods which combine analysis of semantics and syntax.

Manual template-based methods use patterns (usually in the form of regular expressions) generated by domain experts to extract concepts connected by a specific relation from the text (Yu *et al.* 2002). Automatic template methods utilize patterns from the text surrounding concept pairs that are known to have the relationship of interest (Yu & Agichtein 2003; Cohen 2004). Statistical methods identify relationships by looking for concepts that are found in combination with each other more often than predicted by chance (Lindsay & Gordon 1999).

Mining and mapping text from MEDLINE and UMLS Metathesaurus has been the focus of much of the work in this area (Hristovski *et al.* 2005; Srinivasan 2004; Stapley & Benoit 2000; Blaschke *et al.* 1999; Hristovski *et al.* 2001; Weeber *et al.* 2000; Weeber *et al.* 2003; Ding *et al.* 2002; Stephens *et al.* 2001). The UMLS Metathesaurus is the largest biomedical thesaurus, and provides biomedical knowledge consisting of concepts that are classified by semantic type and also employs hierarchical and non-hierarchical relationships among the concepts (Aronson 2001). MEDLINE is the National Library of Medicine (NLM) journal citation database. It was created in the 1960s, and now provides more than 22 million references to biomedical journal articles dating back to 1946. MEDLINE contains citations from more than 5,600 journals. The MEDLINE database can be accessed through PubMed as well as other services. What sets MEDLINE apart from the rest of PubMed is being able to use the NLM controlled vocabulary, Medical Subject Headings (MeSH), to index citations (https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html).

PubMed has been publicly available since 1996. It contains more than 25 million references including the MEDLINE database and additional citations: 1) in-process citations; 2) articles that are out-of-scope (general science and chemistry journals) from MEDLINE, 3) pre-print citations of MEDLINE indexed journals; 4) citations that precede the MEDLINE indexing of a journal; 5) citations that have not been updated with current MeSH headings and have not been converted to MEDLINE; 6) citations to added life sciences journals that submit full-text to PubMed Central (PMC); 7) citations to manuscripts published by NIH-funded researchers; 8) citations for the majority of books available on the NCBI Bookshelf (https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html).

Frequency-based methods usually produce better recall, but worse precision when compared to NLP-based methods. Frequency-based methods are unable to extract directional relationships, and also have difficulty distinguishing between direct and indirect relationships (Jensen *et al.* 2006).

NLP-based methods perform a substantial amount of sentence parsing (Stanford parser; http://nlp.stanford.edu/software/lex-parser.html) to break down the text into a structure where relationships can be easily extracted (Friedman *et al.* 2001).

Several different approaches have been described to identify interactions between genes and proteins based on frequently seen verbs in MEDLINE abstracts (Sekimizu *et al.* 1998); for automatic extraction focusing on protein-protein interactions (Blaschke *et al.* 1999); for combining a syntactic/semantic grammar in a single parsing process to extract a variety of gene pathway relationships (McDonald *et al.* 2004); to use dictionaries of proteins and interaction terms to identify protein-protein interactions within a sentence (Albert *et al.* 2003); to use NLP to extract causal relations between genes and diseases (Freudenberg & Propping 2002); for using a corpus (GENIA; semantically annotated; http://www.geniaproject.org/) for text-mining information extraction (Kim *et al.* 2003); and others (Andrade & Bork 2000; Hirschman *et al.* 2004; Yeh *et al.* 2004).

One common NLP-method uses a tree structure for each sentence to delineate noun phrases and represent interrelationships. A set of rules is then used to extract relationships based on the tree and semantic labels. One negative aspect of using NLP-based methods is that it is extremely difficult to extract relationships that span several sentences (Jensen *et al.* 2006).

KNOWLEDGE DISCOVERY

‘Knowledge discovery’ or ‘hypothesis generation’ is the next step in the literature mining process. Articles have been written about drawing implicit connections from separate literatures by extracting facts from different publications to infer new previously undiscovered relationships. Swanson and others (Zhu *et al.* 2006; Frijters *et al.* 2010; and Li & Liu 2012) have published numerous articles describing the implicit connections they uncovered. Several examples include: showing that fish oil can help patients with Raynaud syndrome (Swanson 1986), eleven neglected connections of migraine and magnesium (Swanson 1988), implicit connections between Somatomedin C and arginine (Swanson 1990), and connections linking estrogen to Alzheimer’s disease (Smalheiser & Swanson 1996). The software ARROWSMITH (http://arrowsmith.psych.uic.edu) which is a computer-assisted approach for formulating scientific hypotheses by identifying words shared between articles was also created (Smalheiser & Swanson 1998) to aid in this type of discovery process. It is quite probable that other novel relationships exist and are waiting to be discovered.

The main driver of development of hypotheses has been co-occurrence of terms from MEDLINE (Xuan *et al.* 2007; Hristovski *et al.* 2005; Srinivasan 2004; Stapley & Benoit 2000; Blaschke *et al.* 1999; Hristovski *et al.* 2001; Weeber *et al.* 2000; Weeber *et al.* 2003; Ding *et al.* 2002; Stephens *et al.* 2001.) Others have developed interesting methods for discovery as well. Jensen built a network of human genes (Jensen *et al.* 2006); Freudenberg described a similarity-based method for genome-wide prediction of disease-relevant human genes by clustering diseases based on their phenotypic similarity (Freudenberg & Propping 2002); Xuan developed MarkerInfoFinder to identify relationships between genetic markers and disease incorporating cytoband location, sequencing annotation, and diseases from OMIM (Xuan *et al.* 2007); and

Turner created POCUS to mine genomic sequence annotations to predict disease genes based on over-representation of annotation between loci for the same disease (Turner *et al.* 2003).

INTEGRATION

An integration framework combines data-mining approaches that integrate literature with other data types and has great potential for making biological discoveries (Jensen *et al.* 2006). Several methodologies are described here but overall, integration of text and data mining results has not been as extensively studied as the other literature mining components.

Perez-Iratxeta *et al.* (2002) described a method relating genes to inherited diseases using fuzzy relations in data mining, and established G2D as a tool for mining genes associated to disease (http://g2d2.ogic.ca/) (Perez-Iratxeta *et al.* 2005). Van Driel *et al.*'s (2003) method showed that given positional and expression/phenotypic data, it is possible to integrate data from several databases to produce an overview of interesting genes. Tiffin *et al.* (2005) integrated text- and data-mining using ontologies to successfully select disease gene candidates. Lustgarten *et al.* (2008) created the EPO-KB database to assist with identification and coordinate knowledge of validated biomarkers and their links to proteins, peptides, modifications, and disease. Further, they showed that 'using EPO-KB as a pre-processing method for biomarker selection found only in the biofluid of the proteomic dataset creates an increase in performance over no or random variable selection' (Lustgarten *et al.* 2009).

2.3 CLASSIFICATION MODELING

Classification is a supervised learning approach to biological data analysis which utilizes a training-set of samples to determine a specific set of measures or rules to use in placing new individuals in groups. Once the training of the model has been performed, the learned rules will be applied to a new set of samples, called the test set. Modeling performance measures can be calculated based on the correctness in placing the new samples into the correct groups. An algorithm which implements a mathematical function for classification is known as a classifier. Classification can also be referred to as modeling. Many modeling algorithms exist, with logistic regression (Cox 1958; Walker & Duncan 1967), Bayesian modeling (Bayes 1763; Pearl 1998), support vector machines (Vapnik & Chervonenkis 1964; Boser et al. 1992; Cortes & Vapnik 1995), decision trees (Quinlan 1983; 1987), and neural networks (McCulloch & Pitts 1943), being the most common. Classification is distinguished from clustering or unsupervised learning.

Clustering is an unsupervised learning approach to biological data analysis where the grouping of subsets of entities (genes or proteins) is accomplished by using a similarity measure. Once clustered, the members of the group will be more similar to each other than to entities in other groups, based on the similarity measure implemented. Cluster analysis encompasses many different clustering algorithms, with hierarchical (Sibson 1973; Defays 1977), k-means (Forgy 1965; Lloyd 1982), and density-based (Martin et al. 1996), being the most commonly used.

2.3.1 Bayesian analysis

Bayesian analysis is a statistical method which uses Bayes' theorem to assess the probability of an event occurring based on prior knowledge. Hypotheses are tested through probability distributions of scientific data. These distributions depend on unknown quantities called parameters. In Bayesian analysis, knowledge about model parameters is expressed by a probability distribution on the parameters, called the "prior distribution".

Prior probability is an assumption. Uncertainty may exist in using a prior, and may have an unknown effect on the results and subsequent conclusions. This uncertainty can be eliminated by using uniform or uninformed priors, or by not using prior information at all. However, including previous information in addition to experimental data may add another level for model building. Care must be taken, in creating priors, when attempting to attach a value to previously known information.

Uniform priors or 'un-informed' priors may be used when not much previous knowledge is available, or when a large amount of experimental data is available. No assumptions are made concerning the data, and a constant value is input for all entities, relegating the prior values to relatively irrelevant status. In this case, the greater importance is placed on the obtained experimental data, ignoring previous knowledge.

Informed prior probabilities can be used when previously known information is available, and when experimental data is limited. Additional effort is required to produce values, but the new information can be added to the experimental data, in hope of producing more informative results, with greater accuracy than when using uniform priors or no priors at all.

Prior information regarding model parameters is expressed as a 'likelihood', which is proportional to the distribution of the data given the model parameters. This information is combined with the prior to produce an updated probability distribution called the 'posterior distribution', on which all Bayesian inference is based (https://bayesian.org/Bayes-Explained). In Bayes' Theorem (Equation 1), the occurrence of an event given an observation (P(E|F)) is calculated by the probability of the occurrence of that observation given the event (P(F|E)), times the probability of the event (P(E)), divided by the probability of the observation (P(F)).

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

Equation 1. Bayes Theorem.

where P(E|F) is the conditional probability, the numerator P(F|E)P(E) is the joint probability, and the denominator P(F) is the marginal probability (Neapolitan 2004).

A Bayesian network (BN) is a directed acyclic graph consisting of a structure and probability parameters (Neapolitan 2004)**.** Bayesian modeling requires learning the structure and parameters of the model. Variables are represented as nodes, and relationships between the variables are represented as directed arcs. The BN consists of a child variable (target), and parent variables (of that target). Probability can be assigned to the child node based on the probability of the parent nodes. The probability distributions for all variables represent the joint probability distribution over all of the variables (Pearl 1998).

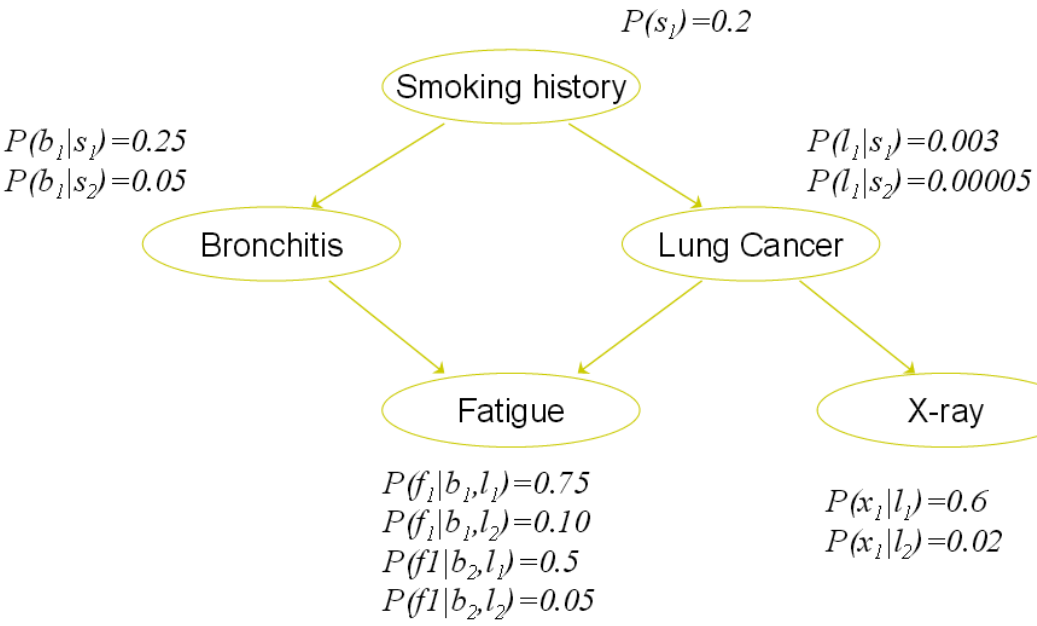


Figure 8. A hypothetical Bayesian Network example. (Neapolitan 2004)

For example, examine the BN (Figure 8) borrowed from Neapolitan 2004. Each node X_i has a conditional probability distribution P(X_i | Parents (X_i)) that calculates the effect of the parents on the node. Probabilities of each child node can be calculated based on the known probabilities of the parent node. The parameters are the probabilities in the conditional probability tables. Using a Bayesian network to calculate probabilities is known as Bayesian inference. Real-world or hypothetical conditions of any kind can be calculated using Bayesian inference, as long as the conditional probabilities are known.

Bayesian procedures can be utilized in classification modeling by enabling the calculation of group membership probabilities, which provide more information than just assigning a group-label to each new observation. There are several reasons that Bayesian methods are becoming more popular. Bayesian modeling is preferred because it is centered on probability theory; expert opinion and data are used to build models; model uncertainty is accounted for; and models can be updated when new knowledge is obtained.

On the other hand, prior probabilities are subjective and some statisticians see this as a drawback. However, powerful computational tools allow Bayesian methods to tackle complex statistical problems with relative ease (https://bayesian.org/Bayes-Explained).

Bayesian analysis utilizes prior knowledge to improve classification results. Several works described below utilize Bayesian methods and prior knowledge to improve performance measures compared to other methods.

Zhou & Zheng (2013) improved predictive performance and identified discrepancies between data, and achieved a prior known graph structure by examining network structures that represent biochemistry interactions. They proposed a Bayesian random graph-constrained model, rGrace that combines a priori network information with empirical evidence, to be used for pathway analysis.

Zhao *et al.* (2012) correctly identified the pathways reported to play essential roles in controlling bone mass by applying a Monte Carlo Markov Chain algorithm to a microarray data set, to improve understanding of the gene expression profile of osteoblasts at defined stages of differentiation. Their method used novel Bayesian models to integrate microarray data with KEGG pathway structures and gene-gene interactions from the literature.

Hill *et al.* (2012) achieved competitive variable selection performance using empirical Bayes with pathway-based priors. Prior biological knowledge was incorporated as weighted informative prior distributions over variable subsets using an empirical Bayes formula. The empirical Bayes method aided in variable selection and guarded against misspecification of priors.

Kim *et al.* (2012) proposed a Bayesian approach for identifying pathways related to different types of outcomes. They incorporated prior knowledge into a Bayesian hierarchical model and achieved more accurate coverage probability than likelihood-based approaches, especially when the sample size is small compared with the number of genes being studied. They suggested analyzing gene sets created based upon prior biological knowledge, as opposed to common statistical methods for microarray analysis that only consider one gene at a time, and may miss small gene-level changes.

Stingo *et al.* (2011) identified markers that would have been missed and improved the prediction accuracy of a Bayesian model by incorporating pathway and gene network information into analysis of DNA microarray data. The information was used for pathway summaries, specifying priors, and structuring the Markov chain Monte Carlo (MCMC) moves to fit the model. By integrating biological knowledge into the analysis they achieved a better understanding of underlying molecular processes.

Kim *et al.* (2011) inferred a signaling pathway related to lung cancer using Reverse Phase Protein Microarray (RPPM), which provided information about post-translational phosphorylation. The pathway was inferred by learning a Bayesian network and Protein-Protein Interaction (PPI) prior knowledge that was incorporated into a new scoring function based on the minimum description length (MDL) (Rissanen 1978). Their cluster-based Linear Programming Relaxation can search for optimal networks.

Parikh *et al.* (2010) discovered dependencies among genes while reducing the computational resources needed in processing high-throughput datasets by using a Bayesian framework to incorporate prior biological knowledge. The single-gene expansion algorithm ranked genes from a large gene-expression repository, as potential new members in search of new constituents of the known pathway. Inferring Bayesian networks from expression data is a powerful tool for learning complex genetic networks, since incorporation of prior knowledge can uncover dependencies among genes.

Jenkinson *et al.* (2010) produced statistical methods for estimating rate constants of a biochemical reaction system from time series data using perturbations. They introduced a Bayesian analysis approach for computing rate constants of a closed biochemical reaction system from experimental data and used a prior probability density function that integrated biophysical and thermodynamic knowledge.

Husmeier and Werhli (2007) improved reconstruction of gene regulatory networks from microarray data by integrating biological prior knowledge expressed as energy functions, from which a prior distribution over network structures were obtained as a Gibbs distribution.

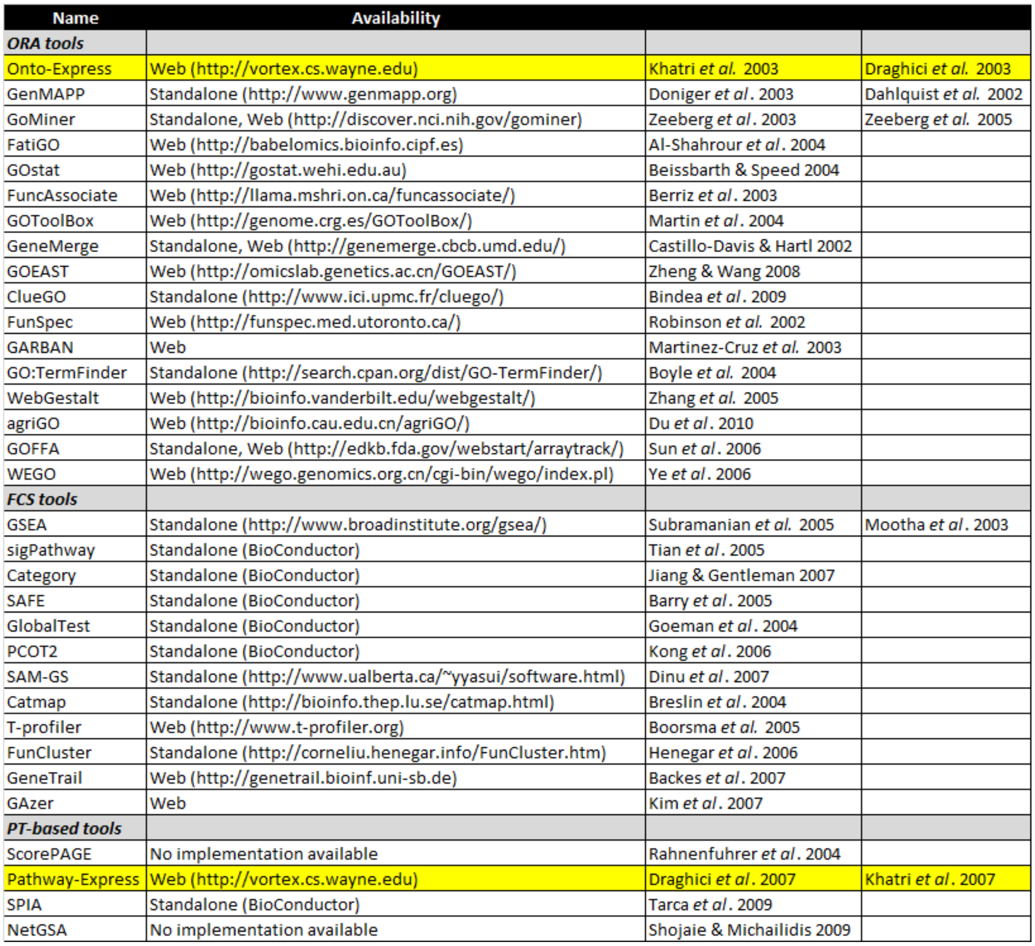
The hyperparameters of this distribution represent the weights associated with the prior knowledge relative to the data.

2.4 PATHWAY ANALYSIS

Pathway analysis has become an important step in the biological data analysis process. Pathways provide a visual representation of gene/protein interactions in physiological processes. High-throughput experimental profiling analyses produce lists of differentially expressed genes/proteins. Grouping long lists of genes/proteins into smaller sets of related genes/proteins implicated in similar pathways reduces the complexity of analysis from thousands of genes/proteins to hundreds of pathways (Khatri *et al.* 2012). Additionally, identifying pathways that differ between two conditions can have more illustrative power than a simple list of differentially expressed genes or proteins (Glazko & Emmert-Streib 2009). Knowledge-based pathway analysis identifies pathways that may be affected in a condition by associating information in a pathway database with gene expression patterns for the disease of interest. The result is differential expression of a set of genes or proteins rather than a list of individual genes (Khatri *et al.* 2012).

Pathway tools have been created to aid researchers in biological experimental data interpretation. By providing a visual representation, pathway tools allow researchers to determine upstream and downstream genes/proteins that affect or are affected by a gene/protein of interest, which ultimately may allow for discovery of targets for disease treatment.

Table 1. Pathway analysis tools. Names, access, and sources of common pathway analysis tools are provided. (Adapted from Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol* 2012, 8(2): e1002375).



Name	Availability		
ORA tools			
Onto-Express	Web (http://vortex.cs.wayne.edu)	Khatri *et al.* 2003	Draghici *et al.* 2003
GenMAPP	Standalone (http://www.genmapp.org)	Doniger *et al*. 2003	Dahlquist *et al.* 2002
GoMiner	Standalone, Web (http://discover.nci.nih.gov/gominer)	Zeeberg *et al*. 2003	Zeeberg *et al.* 2005
FatiGO	Web (http://babelomics.bioinfo.cipf.es)	Al-Shahrour *et al*. 2004	
GOstat	Web (http://gostat.wehi.edu.au)	Beissbarth & Speed 2004	
FuncAssociate	Web (http://llama.mshri.on.ca/funcassociate/)	Berriz *et al*. 2003	
GOToolBox	Web (http://genome.crg.es/GOToolBox/)	Martin *et al*. 2004	
GeneMerge	Standalone, Web (http://genemerge.cbcb.umd.edu/)	Castillo-Davis & Hartl 2002	
GOEAST	Web (http://omicslab.genetics.ac.cn/GOEAST/)	Zheng & Wang 2008	
ClueGO	Standalone (http://www.ici.upmc.fr/cluego/)	Bindea *et al*. 2009	
FunSpec	Web (http://funspec.med.utoronto.ca/)	Robinson *et al.* 2002	
GARBAN	Web	Martinez-Cruz *et al.* 2003	
GO:TermFinder	Standalone (http://search.cpan.org/dist/GO-TermFinder/)	Boyle *et al.* 2004	
WebGestalt	Web (http://bioinfo.vanderbilt.edu/webgestalt/)	Zhang *et al.* 2005	
agriGO	Web (http://bioinfo.cau.edu.cn/agriGO/)	Du *et al*. 2010	
GOFFA	Standalone, Web (http://edkb.fda.gov/webstart/arraytrack/)	Sun *et al*. 2006	
WEGO	Web (http://wego.genomics.org.cn/cgi-bin/wego/index.pl)	Ye *et al*. 2006	
FCS tools			
GSEA	Standalone (http://www.broadinstitute.org/gsea/)	Subramanian *et al.* 2005	Mootha *et al*. 2003
sigPathway	Standalone (BioConductor)	Tian *et al*. 2005	
Category	Standalone (BioConductor)	Jiang & Gentleman 2007	
SAFE	Standalone (BioConductor)	Barry *et al*. 2005	
GlobalTest	Standalone (BioConductor)	Goeman *et al*. 2004	
PCOT2	Standalone (BioConductor)	Kong *et al*. 2006	
SAM-GS	Standalone (http://www.ualberta.ca/~yyasui/software.html)	Dinu *et al*. 2007	
Catmap	Standalone (http://bioinfo.thep.lu.se/catmap.html)	Breslin *et al*. 2004	
T-profiler	Web (http://www.t-profiler.org)	Boorsma *et al.* 2005	
FunCluster	Standalone (http://corneliu.henegar.info/FunCluster.htm)	Henegar *et al*. 2006	
GeneTrail	Web (http://genetrail.bioinf.uni-sb.de)	Backes *et al*. 2007	
GAzer	Web	Kim *et al*. 2007	
PT-based tools			
ScorePAGE	No implementation available	Rahnenfuhrer *et al*. 2004	
Pathway-Express	Web (http://vortex.cs.wayne.edu)	Draghici *et al*. 2007	Khatri *et al*. 2007
SPIA	Standalone (BioConductor)	Tarca *et al*. 2009	
NetGSA	No implementation available	Shojaie & Michailidis 2009	

Several generations of pathway analysis approaches have been described in the literature. First-generation pathway approaches utilize Over-Representation Analyses (ORA). ORA methods evaluate the fraction of genes in a pathway that are found among the set of differently expressed genes (Table 1) (Khatri *et al.* 2012). ORA methods create an input list using a threshold or criteria (differentially expressed genes for a condition at a false discovery rate (FDR) of 5%). Then, for each pathway, input genes are counted. Every pathway is then tested for over or underrepresentation in the list of input genes. (Khatri *et al.* 2002; Draghici *et al.* 2003; Berriz *et al.* 2003; Beissbarth & Speed 2004; Boyle *et al.* 2004; Castillo-Davis & Hartl 2002; Martin *et al.* 2004; Doniger *et al.* 2003). The most commonly used tests are based on the hypergeometric, chi-square, or binomial distribution (Khatri *et al.* 2012). Comparisons of ORA tools can be found in Khatri & Draghici 2005; and Huang *et al.* 2009.

Second-Generation pathway approaches employ Functional Class Scoring (FCS). FCS is based on the premise that large changes in individual genes as well as smaller changes in functionally related gene sets (pathways) may have significant effects (Khatri *et al.* 2012). Most FCS methods use three steps (Ackermann & Strimmer 2009): Step 1) a gene-level statistic is calculated by computing differential expression of individual genes or proteins from experimental measurements. Gene-level statistics include: correlation of molecular measurements with phenotype (Pavlidis *et al.* 2004), ANOVA (Al-Shahrour *et al.* 2005), Q-statistic (Goeman *et al.* 2004), signal-to-noise ratio (Subramanian *et al.* 2005), *t*-test (Al-Shahrour *et al.* 2005; Tian *et al.* 2005), and Z-score (Kim & Volsky 2005).

Step 2) gene-level statistics of all pathway genes are combined into a single statistic. The statistic can represent interdependencies among genes (Kong *et al.* 2006; Lu *et al.* 2005; Xiong 2006; Hummel *et al.* 2008; Klebanov *et al.* 2007) or it can ignore them (Tian *et al.* 2005; Jiang & Gentleman 2007). The pathway-level statistic can depend on the number of differentially expressed genes, the size of the pathway, and the gene correlation within the pathway (Khatri *et al.* 2012).

Step 3) statistical significance of the pathway statistic is determined. Null hypothesis testing can be broken down into two categories: 1) competitive null hypothesis and 2) self-contained null hypothesis (Goeman & Buhlmann 2007; Ackermann & Strimmer 2009; Tian *et al.* 2005; Efron & Tibshirani 2007). A competitive null hypothesis permutes gene labels in the pathway, and compares the gene set in the pathway with another gene set not in the pathway. A self-contained null hypothesis permutes class labels (phenotypes) for each sample and compares the pathway gene set with itself, ignoring genes not in the pathway (Khatri *et al.* 2012).

Third Generation pathway approaches are Pathway Topology (PT)-Based. These approaches utilize protein-protein interaction databases in a given pathway, how the proteins interact, and where they interact within the cell. The databases include KEGG (www.genome.jp/kegg/pathway.html; Ogata *et al.* 1999; Kanehisa & Goto 2000), MetaCyc (Karp *et al.* 2002), Reactome (www.reactome.org/PathwayBrowser/; Joshi-Tope *et al.* 2003; Joshi-Tope *et al.* 2005), RegulonDB (Huerta *et al.* 1998), STKE (http://dictybase.org/STKE.htm), BioCarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways), and PantherDB (Thomas *et al.* 2003).

ORA and FCS methods only depend on the number of pathway genes or gene co-expression to identify significant pathways. They do not incorporate additional information. Therefore, as long as they contain the same set of genes, the two methods will produce the same results. Pathway topology (PT)-based methods utilize the additional information as well. PT-based methods are the same as FCS methods, but also incorporate the use of pathway topology to compute gene-level statistics (Khatri *et al.* 2012).

Pathway Express (vortex.cs.wayne.edu/Projects.html; Khatri *et al.* 2005; Khatri *et al.* 2007, Draghici *et al.* 2007) is a third generation pathway analysis approach that calculates an impact factor (***if***) in the analysis. The impact factor encapsulates the entire pathway by incorporating biological factors such as gene expression, types of interactions, and location of genes within the pathway (Draghici *et al.* 2007; Khatri *et al.* 2007). Impact factor analysis represents a pathway as a graph, with the nodes representing genes and edges signifying interactions between the nodes.

A perturbation factor (PF) for a gene is calculated as a sum of its differential expression and factors of all genes in the pathway (Equation 2). The impact factor is the sum of all perturbation factors for all genes in a pathway (Equation 3). Impact factor analysis was improved to address the effect of differential expression on the perturbation factor and the high false positive rate observed for small lists of input genes (Tarca *et al.* 2009).

Impact Factor Analysis

Impact factor analysis (Draghici *et al.* 2007; Khatri *et al.* 2007) computes a perturbation factor for each gene in each pathway as follows (Equation 2, Khatri *et al.* 2012):

$$PF(g_i) = \Delta F(g_i) + \sum_{j=1}^{n} \beta_{ji} \cdot \frac{PF(g_j)}{N_{ds}(g_j)}$$

Equation 2. Pertubation factor

$\Delta F(g_i)$, represents the normalized change in expression of gene g_i. The second term accounts for the topology of the pathway, where gene g_j is upstream of gene g_i. B_{ji} represents the interaction between g_i and g_j. If g_j activates g_i, $B_{ji} = 1$, and if it inhibits g_i, $B_{ji} = -1$. The PF of gene g_j is then normalized by the number of downstream genes it interacts with, $N_{ds}(g_i)$. The second term is repeated for every gene that is upstream of another gene. After computing PF for each gene, the impact factor (IF), is computed using Equation 3 (Khatri *et al.* 2012):

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g\in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Equation 3. Impact factor

In the first term, P_i is the probability of obtaining a statistic as extreme as the one observed for a true null hypothesis. In order for the IF to be large for severely impacted pathways (with small p-values), the first term uses $1/p_i$ rather than p_i. The log function converts the exponential scale of the p-values to a linear scale. The second term sums up all of the PFs for all genes in the pathway P_i, and is normalized by the number of differentially expressed genes in the pathway.

After computing the PFs for all genes in the pathway, Equation 3 is used to calculate the impact factor for each pathway. The impact factor of each pathway is then used to assess the impact of the gene expression data set on all of the pathways (with higher impact factors equating to the more significant the pathway) (Khatri *et al.* 2012).

2.5 PRIOR KNOWLEDGE USE IN BIOINFORMATICS

Many molecular biological experiments performed are exploratory in nature, examining thousands of genes or proteins at once with the goal of uncovering an individual or panel of biomarkers for a disease or condition. Bioinformatics techniques such as machine learning (Solomonoff 1957), pattern recognition (Carvalko & Preston 1972), image analysis (Exner & Hougardy 1988), or information retrieval are then performed to group genes/proteins or predict sample classifications. Lastly, incorporating experimental results with current scientific knowledge and pathway analysis allows for conclusions to be drawn and more-targeted experiments to follow.

Only recently has prior knowledge incorporation during analysis with other types of data become common. The use of prior biological knowledge can improve the classification results such as accuracy, reproducibility and interpretability. The addition of prior knowledge into tried-and-true molecular techniques has improved results, as well as enabled more creativity and produced some very innovative and intriguing concepts in many different areas of biological research. The following section describes the different types of prior knowledge and how they are used for different purposes in molecular biology analyses.

Bioinformatic Tools

Bioinformatic tools have been developed to aid researchers in processing and analysis of enormous amounts of information. These tools expedite the time investment, and streamline results based on relevancy.

Sun *et al.* (2015) developed the Drug-specific Signaling Pathway Network (DSPathNet) which combines prior drug knowledge and drug-induced gene expression via random walk algorithms. Drugs exert their effects through interconnected networks of multiple signaling pathways, but it is difficult to incorporate interwoven pathways into one network. DSPathNet can be used to construct drug-specific signal transduction networks and produce models for exploring signaling pathways, to assist in the understanding of drug action, disease pathogenesis, and identification of drug targets.

Johannes *et al.* (2011) introduced pathClass, which is a collection of different SVM-based classification methods to improve gene selection and classification performance. The methods contained in pathClass rely on gene expression data and also exploit gene network data.

Yang *et al.* (2007) introduced GS2PATH, a tool for gene-set enrichment from prior knowledge, such as gene ontology (GO) and pathway databases. GS2PATH can estimate gene set enrichment in GO terms from KEGG and BioCarta pathways, and allows users to compute and compare functional over-representations. Gene-set enrichment can be useful in metabolism, signal transduction, genetic and environmental information processing, cellular processing, and drug development.

Causal Pathways

Causal pathways can be used to map events or changes that can lead to disease. Four examples of causal pathways utilizing prior knowledge incorporation are provided: Causal networks constructed from individual relationships from scientific literature aid in gene-expression data interpretation.

Kramer *et al.* (2014) developed a method to predict downstream effects on biological functions and diseases. They presented tools for deducing and scoring regulator networks upstream of gene-expression data-based on a large-scale causal network derived from the Ingenuity Knowledge Base (IPA; http://www.ingenuity.com).

Catlett *et al.* (2013) described Reverse Causal Reasoning (RCR), a reverse engineering method to infer hypotheses from molecular profiling data. RCR aids in interpretation of gene expression profiling and provides an approach to the development of models of disease, drug action, and drug toxicity. Their methodology requires literature-curated cause-and-effect relationship prior knowledge that can link an upstream mechanism to downstream quantity. Whistle, can be used for the analysis of gene expression data using prior knowledge expressed in Biological Expression Language (BEL).

Silver *et al.* (2012) detected multivariate trait gene pathways, and used them to identify causal pathways that produce structural changes in the brains of Alzheimer's disease (AD) patients. The method known as pathways sparse reduced-rank regression (PsRRR) uses group lasso penalized regression to model the effects of genome-wide SNPs that are grouped into functional pathways using gene-gene interaction prior knowledge.

Martin *et al.* (2012) introduced Network Perturbation Amplitude (NPA) scoring. The NPA scoring method interprets high-throughput measurements and a priori literature-derived knowledge of cause and effect relationships in the form of network models to characterize the activity of biological processes at high-resolution. The relationships were used to create network models of biological processes, such as inflammation or cell cycle progression.

Clustering

Hierarchical clustering groups similar performing genes/proteins together based on function or expression. A similarity measure or metric is used to determine “closeness” of genes in relation to other genes. The addition of prior knowledge into clustering exercises has been shown to improve clustering results.

Milone *et al*. (2014) improved clustering of biological data by prior knowledge incorporation in a novel training algorithm that evaluated the biological connections of the data points while self-organizing maps (SOMs) clusters and biologically-inspired SOMs (bSOM) were being formed. Inclusion of biological information during training increased the biological value of the clusters, improved the results, and simplified further analysis.

Hwang *et al.* (2012) significantly improved the classification of disease phenotypes and disease pathway genes in experiments testing disease phenotype-gene associations in OMIM and KEGG. Phenotypes and genes were co-clustered to simultaneously detect associations between phenotype clusters and gene clusters. The algorithm created a phenotype-gene association matrix utilizing phenotype similarity and protein-protein interaction prior knowledge, disease classes and biological pathways.

Gene enrichment

In gene enrichment analysis, groups of genes are studied together to assign biological meaning to the group, as opposed to gene expression analysis where each gene is studied individually. The gene group is usually clustered together as a result of expression analysis, function, protein family, or some other area of interest using prior knowledge. Analyzing groups of genes enables new biological patterns to emerge, or to determine whether a subset shows similar expression of a biological characteristic, or might belong to similar biological pathways. Researchers now combine pathway, gene enrichment analysis and network-based approaches to identify relationships between different molecular mechanisms.

Huang *et al. (*2012) developed the Pathway and Gene Enrichment Database (PAGED), to enable disease-specific pathway, gene signature, microRNA target, and network queries by integrating gene-set prior knowledge from the genome, transcriptome, and proteome. PAGED explores relationships between gene-sets as gene-set association networks. This shows promise for developing tools which will perform even better than third-generation pathway analysis approaches, allowing for the discovery of molecular phenotypes for disease-associated pathway and gene enrichment analysis.

Gene-gene interaction

Gene-gene interaction (epistasis) occurs when the activity of one gene is dependent on the presence of other genes. Certain interactions among gene products or mutations within genes can result in downstream effects that can drastically alter biological processes.

Gomez-Vela and Diaz-Diaz (2014) developed GeneNetVal to assess the biological validity of gene networks using gene-gene interactions in KEGG metabolic pathways. Converting KEGG pathways into a gene association network with a distance measure of gene-gene interactions was proposed.

Ma *et al.* (2012) identified and validated an interaction affecting a complex trait in multi-ethnic populations, based on a knowledge-driven analysis of epistasis. Gene-gene interactions that affect lipid levels were tested, using prior knowledge of established GWAS hits, protein-protein interactions, and pathway information.

King *et al.* (2005) identified a "nexus" of genes that are attractive candidates for therapeutic targeting by using pathway techniques to study atherosclerosis as an integrated network of gene interactions. They describe their pathway development approach which is based on connectivity from language parsing of published literature, and ranking by differentially regulated genes in the network. The discussed a systems biology approach that accounts for gene interactions in atherosclerosis, incorporates non-transcriptionally regulated genes, and integrates prior knowledge. The results of their work show the advantages of a systems-based approach to analyzing complex diseases.

Genome-wide Association Studies

Genome-wide association studies (GWAS) can be used to examine genetic data as well as demographic data or other types of information or prior knowledge to determine if a variable is associated with a phenotype, condition, or disease. GWAS studies can be used to investigate an entire genome for single-nucleotide polymorphisms (SNPs) and variants associated with a disease.

Brenner *et al.* (2013) used a two-stage approach to investigate associations between variants in inflammatory pathway genes and lung cancer risk genes. Variants were identified using keyword and pathway searches of Gene Cards and Gene Ontology databases. Hierarchical modeling (HM) was used to incorporate variant prior information. A matrix of priors was constructed using: gene role in inflammation and immune pathways; physical properties such as location, conservation scores and amino acid coding; linkage disequilibrium (LD) with other variants; and heterogeneity.

Li *et al.* (2012) described a hybrid set based test (HYST) that combined the extended Simes' test and scaled chi-square test. The test combination of tests was used to produce a set of genome-wide association signals at multiple SNPs in order to determine the significance of association at gene/pathway levels. HYST can be used to examine SNP-sets based on prior biological knowledge, as well as evaluate statistical significance for protein-protein interactions to increase the power for detecting disease-susceptibility genes.

Liu *et al.* (2012) detected previously not-significant genes and determined novel drug targets and disease biomarkers using prior biological knowledge to restrict the set of candidate SNP pairs to be tested. They examined interactions among genomic loci (epistasis) as potential sources of missing heritability in genome-wide association studies, and presented four approaches to detect interactions involved in complex diseases: '(1) for each gene, a gene-specific set of SNPs produced a gene-based interaction model, (2) for each pathway, a pathway-specific gene-set of SNPs provided a pathway-based interaction model, (3) a disease-related gene-set of SNPs resulted in a network-based interaction model, and (4) a SNP function framework.'

Jia *et al.* (2011) tested GWAS association data integrated with human protein-protein interaction (PPI) network prior-knowledge using a dense module searching (DMS) method that identified gene-sets for complex diseases. Proteome studies were used to examine interactions between genes and the pathogenesis of complex diseases. Functional enrichment analysis showed that genes identified by DMS have higher association signal.

Chen *et al.* (2011) presented a GWAS framework that was more effective in identifying disease-associated genes than a single gene-based method. The Markov Random Field (MRF) model incorporated pathway topology for association analysis.

Being that GWAS usually focus on the analysis of single markers, which lack power to detect small effect sizes of most genetic variants, pathway-based approaches utilizing prior biological knowledge allow for more powerful analyses. Wang *et al.* (2010) reviewed the development of GWAS pathway-based approaches, and suggest that pathway-based approaches may also be useful for GWAS of sequencing data.

Metabolomics

Metabolomics studies specific cellular processes in cells, and provides a snapshot of cell physiology at the time the sample is taken. Van den Berg *et al.* (2009) explored relations between metabolome data and related metabolites, and an amino acid biosynthesis pathway.

They described consensus principal component analysis (CPCA) and canonical correlation analysis (CCA). CPCA searches for common metabolite concentrations. CCA identifies correlations between relevant metabolites and the rest of the metabolome. CCA and CPCA are complementary data analysis tools that can focus data analysis on metabolite groups.

Microarray / Gene Expression Analysis

In microarray/gene expression experiments, thousands of DNA/RNA/protein probes are affixed to a solid surface (slide or chip), as the sample is placed on the slide. The contents of the sample are allowed to interact and bind to the probes. The remaining unbound probe is washed away, and the amount of sample bound to the probes measured. This technology allows for tens of thousands of probes to be analyzed simultaneously however, issues arise in analyzing several thousand entities for a small number of samples. While incorporation of prior knowledge into data analysis has been deemed important, in practice, it has been extremely limited.

Yuryev (2015) advocated for causal reasoning methods to calculate cancer pathway activity signatures. Causal reasoning algorithms can transform microarray data into a small number of cancer hallmark pathways. They offer this as a solution for the 'curse of dimensionality’, which occurs when only a small number of samples are available for training sets, and a large number of genes are being measured, as happens often in the use of microarrays.

Chen *et al.* (2014) described a model that demonstrated better fitness than the state-of-the-art model, which relied on an initial random selection of genes, and showed the advantage of combining gene interactions from the literature with microarray analysis for generating gene regulatory networks. A genetic algorithm was used to optimize the strength of interactions using microarray data and an artificial neural network fitness function. Invasive ductile carcinoma

(IDCA) of the breast was used to query the literature and a microarray set containing gene expression changes in these cells over several time points was evaluated.

High-dimensional microarray datasets contain high levels of noise, causing problems for machine learning methods. Feature selection removes most of the irrelevant genes, and thus much of the noise. The most common feature extraction method is principal component analysis (PCA) (Hotelling 1933). Hira *et al.* (2014) proposed an a priori manifold learning method for finding a representative set of microarray data infused with KEGG pathway data. Manifold learning algorithms, such as Isomap (Tenenbaum *et al.* 2000) project data from a higher dimensional space to a lower dimension. The new manifold produced better classification results than either PCA or Isomap.

Chen and Wang (2009) showed that the prediction models constructed of gene-sets (prior knowledge integrated with gene expression values) outperformed prediction models of single-gene expression values, with improved prediction accuracy and interpretability. Gene id's were linked with annotation databases such as Gene Ontology (GO). 'Supergenes' for each category were constructed from outcome-related genes using a modified PCA method. These supergenes from each gene category represent the ability to predict survival outcome.

Kuffner *et al.* (2005) derived from the literature interpretations of expression measures with biological hypotheses. Gene clusters that exhibit significant gene expression as well as a coherent literature profile were identified, and were shown to be more sensitive and more specific than Gene Ontology categories of the same data. Their approach generalizes to real applications and does not rely on controlled vocabularies or pathway resources.

microRNAs

MicroRNAs (miRNA) are small RNAs that function in RNA silencing and gene expression regulation. MiRNAs can base-pair with complementary sequences in messenger RNAs (mRNAs), altering their function. Thousands of miRNAs are encoded in the human genome; and miRNAs are identified by the genes they affect. Qiu *et al.* (2011) developed the miR2Gene tool to examine gene patterns by analyzing prior knowledge of miRNA regulators. MiR2Gene is a useful tool that integrates miRNA knowledge for protein-coding gene analysis, and can be used for single, or multiple genes, as well as KEGG pathways. Sets of miRNAs were integrated with miR2Gene according to function, disease, and tissue specificity; and their enrichment evaluated.

Networks

Biological networks consist of many overlapping processes and pathways which make up the complicated systems involved in life. Recently, network-based approaches utilizing gene interaction information have emerged.

Barter *et al.* (2014) compared single-gene, gene-set, and network-based methods using gene expression microarray data from melanoma and ovarian cancer. Informative genes were identified using gene expression and network connectivity information combined with prior knowledge of protein-protein interactions; as well as informative sub-networks (small networks of interacting proteins from prior knowledge networks). The different methods tested were correctly classifying alternate subsets of patients in each cohort, in novel and patient-level analyses, leading to the conclusion that 'combination' classifiers that are capable of identifying which patients will be more accurately classified by one method or another are needed.

Hur *et al.* (2014) pushed for an integrative multiple analysis approach consisting of biochemical and pharmacological networks, and transcriptomic signatures for understanding drug safety and gene-drug interactions. Integrated pharmacology and biochemical networks could describe drug-induced rhabdomyolysis by incorporating prior knowledge with publicly available data. A list of rhabdomyolysis-inducing drugs (RIDs) was compiled. Proteins interacting with RID pharmacological targets were significantly enriched in cell cycle regulation, apoptosis, and ubiquitination functions. Transcriptomic analysis of RIDs revealed that multiple pathways are also perturbed by RIDs.

Jin & Zou (2013) identified new interactions among inflammatory factors and biological pathways by combining nonlinear ordinary differential equation (ODE)-based optimization with mutual information. They constructed an inflammatory regulatory network (IRN) during Influenza A virus (IAV) infection by integrating gene expression data with prior knowledge.

Ante *et al.* (2011) exhibited the role of spindle checkpoint-related pathways in breast cancer by performing validations of relevant pathways by creating a signaling network from TRANSPATH and a metabolic network from KEGG LIGAND, and incorporating Serial Analysis of Gene Expression (SAGE) expression data from breast cancer.

Pathway Intersections

In order to combine individual pathways into larger networks, similar entities need to exist in multiple pathways or in close proximity to several pathways in order to “link” them together. These linking entities may be genes, proteins, or similar processes, and are vital for network biology studies. The use of pathways and gene interaction networks has allowed for better understanding of the differences in gene expression profiles between samples from a systems biology perspective. The usefulness and accuracy of pathway analysis depend on understanding how genes interact with one another. That knowledge is continuously improving due to advances in next generation sequencing technologies and computational methods. While most approaches treat each genes or proteins as independent entities, pathways actually coordinate to perform essential functions in cells.

Liang *et al.* (2015) state that Sparse regression compares favorably to Weighted Correlation Network Analysis when gene association signals are weak. Sparse regression was used to find genes that are intermediary to and interact with at least two pathways. A gene is considered a shared neighbor of two pathways if it can be determined to interact with at least one gene in each pathway. Each pathway gene is modeled using a predictor gene-set, and a connection between the pathway gene and predictor gene occurs when the sparse regression coefficient is non-zero.

Francesconi *et al.* (2008) studied of networks of pathways. The networks were reconstructed based on significance of single pathways (nodes) and the intersection between them (edges). Groups of genes that interface between different pathways can be considered relevant even if the pathways they belong to are not significant alone.

Protein-Protein Interaction

Protein–protein interactions (PPIs) refer to known interactions between two or more proteins in biochemical events. Proteins must interact in a specific way in order for a process to be successful. If an important protein is missing or altered, the biological process will not be successfully completed, which could lead to disease or death. Protein interactions have been studied extensively, and databases of protein interactions exist. Being aware of these interactions has led to the creation of pathways and networks which improve the understanding of biological process and disease understanding, and can also lead to the discovery of possible drug targets.

Chronic obstructive pulmonary disease (COPD) is a highly-complex human disease with high mortality. Incorporation of network or pathway information into biomarker discovery might improve prediction performance. Hua and Zhou (2014) combined protein-protein interactions (PPI) information with a support vector machine (SVM) (Cortes & Vapnik 1995) (Ben-Hur *et al.* 2001) method to identify potential COPD-related genes that would enable determination of severe emphysema from mildly emphysematous lung tissue. When compared with another SVM method which did not use the prior PPI information, the prediction accuracy was significantly enhanced (AUC (Fawcett 2006) increased from 0.513 to 0.909). This shows that incorporating a prior knowledge network into gene selection can potentially significantly improves classification accuracy.

Zhao *et al.* (2014) showed that both the average accuracy (correctly predicted pathways / total number of pathways to which all the target genes were annotated) and the relative accuracy (percentage of the genes with all the annotated pathways being correctly predicted) for pathway predictions were increased with the number of the interacting neighbors. Protein-protein interactions and Gene Ontology (GO) databases were integrated for use as prior knowledge. KEGG pathways with interacting neighbors of target genes were chosen as candidate pathways. Pathways to which the target gene belonged were determined by testing whether genes in the candidate pathways were enriched in GO terms to which the target gene was annotated. Protein-protein interaction data obtained from the Human Protein Reference Database (HPRD; http://www.hprd.org/) and Biological General Repository for Interaction Datasets (BioGRID; http://thebiogrid.org/) was used to predict the pathway attributions of the target gene.

Kirouac *et al.* (2012) found that widely used pathway databases are highly inconsistent with respect to constituents and interactions. They assembled a network from multiple on-line resources of pathway and interactome databases (Cancer CellMap, GeneGo, KEGG, NCI-Pathway Interactome Database (NCI-PID), PANTHER, Reactome, I2D, and STRING) utilizing knowledge of proteins and protein interactions involved in inflammatory signaling networks. Wide inconsistencies among interaction databases, pathway annotations, and the numbers and identities of nodes associated with a given pathway pose major challenges in deriving causal insight from network graphs. As such, it is difficult to identify biologically meaningful pathways from interactome networks a priori; however by incorporating prior knowledge, it is possible to build out network complexity with increasing confidence.

SNPs/Variants

Different variations of a nucleotide at a given locus are called alleles. A single nucleotide polymorphism (SNP), is a variation at a given nucleotide at a given location in the genome, which occurs at least in 1% of the population. The possible alleles of a SNP are usually well-known. In contrast, a variant can be any variation (allele) at a given locus, but does not have to meet the qualification of being present in 1% of the population. SNPs have been extensively studied, to the point that the NCBI has created a database of SNPs called dbSNP. Over 10 million SNPs exist in a human genome, and many diseases such as sickle-cell anemia are known to be caused by SNPs.

Lin *et al.* (2014) proposed an efficient method that is less sensitive to neutral variants and direction effect of causal variants, and can zero in on a genomic region or a chromosome to a disease associated region. Genetic variants were scanned to identify the region most likely harboring a disease gene with rare or common causal variants. A score is given to each variant, and aggregate scores are used to identify regions with disease association. Using a Parkinson's case-control dataset, the proposed method has better power than three other tested methods, and also well-controlled type I error. The association of SNCA/*α-synuclein* gene with Parkinson's disease (p = 0.005) was also confirmed.

Li *et al.* (2011) indicated that the integration of network biology and genetic analysis provides bridges between genetic variants and candidate genes or pathways. Using a two-step approach, they detect differentially inherited SNP units from a SNP network. SNP-SNP interactions were identified using prior biological knowledge, such as chromosomal location or functional relationships of their genes. Disease-risk SNP units were ranked by their differentially inherited properties in IBD (Identity By Descent) profiles of affected and unaffected sib-pairs.

Namkung *et al.* (2011) showed that modeling local rather than global ancestry may be beneficial when controlling the population structure effect in rare variant association analysis. They evaluated different methods of rare variant analysis, including single-variant, gene-based, pathway-based analyses, and analyses that incorporate biological information. Using a Bayesian network and a collapsing receiver operating characteristic curve improved risk prediction for diseases caused by many rare variants.

Chen *et al.* (2011) summarized state-of-the-art approaches involved in integration of biological knowledge into rare variant association studies. The methods fell into three categories: (1) hypothesis testing of index scores by aggregating rare variants at the gene level, (2) variable selection techniques incorporating prior information, and (3) novel approaches that integrate prior information, such as pathway and single-nucleotide polymorphism (SNP) annotations. Similarities found between the methods were that gene-based analysis of rare variants was advantageous to single-SNP analysis and that the minor allele frequency threshold used to identify rare variants may influence the power of the test. A consistent increase in power was identified by considering only non-synonymous SNPs. It was demonstrated that integrating biological knowledge into statistical analyses enabled subtle improvements in the performance of statistical method applied to simulated data.

Other techniques

Pathway analysis incorporates prior biological knowledge to analyze genes/proteins in a biological context. However, hypotheses are often 1D in space. Yang et al (2014) developed direction pathway analysis (DPA), to test hypotheses in high-dimensional space to identify pathways that display distinct responses. DPA was used to study insulin action in adipocytes which regulates protein movement from the cell interior to the surface membrane. DPA determined that several insulin responsive pathways involved in plasma membrane trafficking are only partially dependent on the insulin-regulated kinase AKT. The findings were validated by targeted analysis of key proteins using immunoblotting and live cell microscopy.

Park et al. (2013) described functional knowledge transfer (FKT), and explain that state-of-art machine learning algorithms that utilize FKT improve accuracy in pathway membership prediction. FKT can help biologists integrate prior knowledge from diverse systems to direct targeted experiments. They showed that functional genomics can complement sequence similarity to improve gene annotation transfer between organisms. Their method transfers annotations when determined by genomic data and can be used with a prediction algorithm to combine transferred gene function knowledge with high-throughput data enabling function prediction.

Minn et al. (2012) identified components of the Raf kinase inhibitory protein (RKIP) signaling pathway, which can inhibit breast cancer metastasis, by utilizing statistical analysis of clinical data integrated with experimental validation. They showed how prior biological knowledge can be combined with genome-wide patient data to identify regulatory mechanisms that may control metastasis.

Cun and Frohlich (2012) showed that Reweighted Recursive Feature Elimination (RRFE) (Johannes *et al.* 2010) and average pathway expression led to clearly interpretable signatures; whereas on average, incorporation of pathway information or protein interaction data did not significantly improve their classification accuracies, but did affect the interpretability of gene signatures compared to other classical algorithms.

Morris et al. (2011) discuss a method that trains a protein pathway map, summarizing curated literature to context-specific biochemical data. They showed that fuzzy logic (cFL), can convert a prior knowledge network (from literature or interactome databases) into a model describing protein activation values across multiple pathways.

Zhu (2009) presented an algorithm designed to identify signaling pathways of low and concordant gene expression variation. The semi-supervised gene clustering algorithm extended and generalized the gene-shaving algorithm, so that prior knowledge of signaling pathways could be incorporated. Using pathway gene-sets as prior knowledge, the algorithm formed tight gene clusters with minimal variation across samples.

Pathway modeling may require the integration of multiple data types including prior knowledge. Guo et al. (2006) indicate information-based measures outperform graph structure-based measures for stratifying protein interactions. They assessed Gene Ontology (GO)-derived similarity measures for the characterization of direct and indirect interactions within human regulatory pathways. GO biological process and molecular function annotation measures can be used alone or together for the validation of protein interactions involved in pathways. Protein functional similarity within regulatory pathways decays rapidly as the path length increases.

In the section above, an extensive review of the current literature was performed. Many of the ways that prior biological knowledge is being implemented were described above. However, with all of this ongoing research on the different ways that prior knowledge is currently being utilized, no manuscripts have been identified that describe incorporating prior knowledge into datasets for use as input to enhance pathway analysis. The KEDA framework described herein provides insights into a set of methods developed and evaluated for this purpose, and complements extant methodologies.

3.0 IMPLEMENTATION OF KEDA FRAMEWORK

This section describes the KEDA Framework, and details the processes of literature-mining and the transformation of its results into priors leading to their use in modeling, and pathway analysis. The datasets used for modeling and pathway analysis were obtained from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/index.cgi) and are discussed. The Bayesian Rule-Learner algorithm (Gopalakrishnan *et al.* 2010) for modeling and the Pathway Express (vortex.cs.wayne.edu/Projects.html; Khatri *et al*. 2005; Khatri *et al.* 2007, Draghici *et al.* 2007) utility for pathway analysis are also examined.

3.1 KEDA FRAMEWORK OVERVIEW

The KEDA Framework (Figure 1) utilizes a semi-automated literature mining method to parse lung and breast cancer abstracts obtained from PubMed, to discover putative biomarkers in specific biofluids. Gene and protein mentions from millions of abstracts were tallied and transformed into prior probabilities. These ‘priors’ are incorporated with experimental data for use in the BRL to determine the best performing models. A comparison between the effects of prior information on model development from ‘omic’ datasets using informed prior, uniform prior and no prior results is performed. Additionally, pathway analysis is performed, and priors are incorporated into experimental data and a comparison made between results from prior information only, experimental data only, and prior information + experimental data combined. These subsequent results can be used to develop new methods of biomarker research/discovery.

3.2 DESCRIPTION OF KEDA COMPONENTS

The following sections describe the methodology behind the KEDA components: 1) literature mining, 2) classification modeling, and 3) pathway analysis.

3.2.1 Literature mining methodology

Computational methods for mining of biomedical literature can be useful in augmenting manual searches of the literature using keywords for disease-specific biomarker discovery from biofluids (Jordan *et al.* 2014). By counting the mentions of a gene/protein in disease-specific abstracts, a picture begins to emerge of what is already known in the scientific community about a given disease. Counts of gene/protein abstract mentions can be transformed into new knowledge, which can be used to further disease understanding. Verified findings from such exercises can contribute to the current body of knowledge, and possibly lead to new methods or areas of study for biomarker research and discovery.

In this work, *breast* and *lung cancer* searches were further stratified by biofluid mentions to increase the amount of relevant information. Added stratification enables us to not only determine genes and proteins involved breast and lung cancer, but also to discover within which biofluids the proteins may be found. This knowledge has potential clinical implications, by reducing the invasiveness of the method for obtaining a biofluid for testing. For example, there would be no reason to undergo the painful procedure for obtaining cerebrospinal fluid, if the same protein could be attained from blood or urine. This is very important knowledge, which has only recently been further pursued for lung and breast cancer (Veenstra *et al.* 2005; Zhou *et al.* 2005; Nicholas *et al.* 2006; Alterovitz *et al.* 2008; Xu & Veenstra 2008; Delaleu *et al*. 2008; Tyson & Ornstein 2008; Lee & Wong 2009; Gao *et al.* 2009; Sugimoto *et al.* 2010; Oumeraci *et* *al*. 2011; Nolen & Lokshin 2011; Lau *et al.* 2012; Aboud & Weiss 2013; Ramshankar & Krishnamurthy 2013; Tredwell *et al.* 2014; Jordan *et al*. 2014; Qin *et al.* 2015).

3.2.1.1 Information retrieval ***Defining the search space:*** It is important to examine all abstracts, both relevant (positive) and non-relevant (negative), within a given pre-defined search space, so that the results are exhaustive and so statistical significance measures can be accurately calculated. Figure 9 provides an example of defining a search space using 'urine', 'breast cancer' and/or 'lung cancer'.

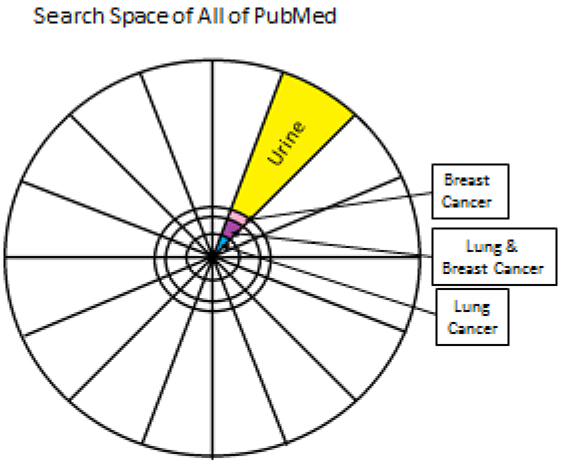


Figure 9. Defining the search space. The search space of all PubMed abstracts returned using the keyword 'urine'. Within those abstracts are those which also contain 'lung cancer'/'breast cancer', or both. Abstracts containing the terms 'urine' and 'lung cancer' and/or 'breast cancer' make up the positive set; the others make up the negative set.

Database searching

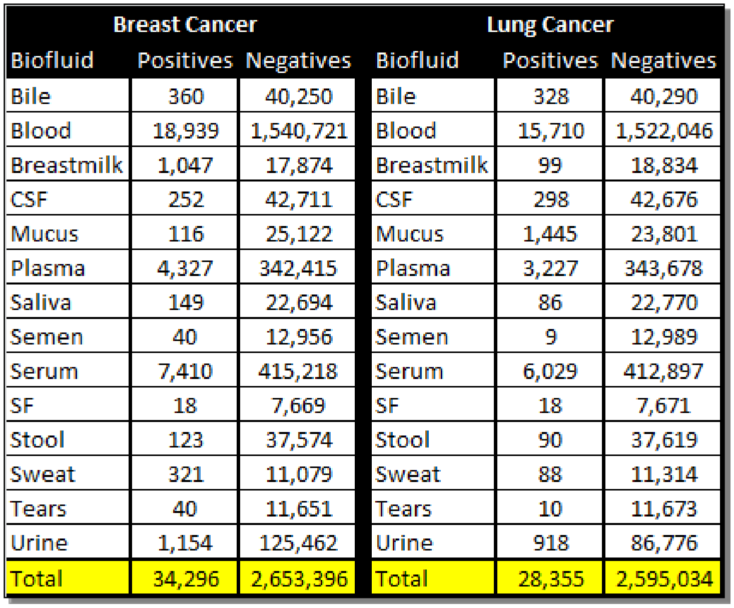
In literature mining, two primary search methods exist: subject heading and keyword. Due to strengths and weaknesses of both methods, some combination of the two are usually employed to achieve optimal results. In general, subject heading searches utilize a defined dictionary/thesaurus of controlled terms. MEDLINE terms are referred to as Medical Subject Headings (MeSH). Subject headings assemble possible synonyms and variations of a given term. For example, the term ‘cancer’ may also be described in articles as neoplasms, malignant, benign, tumor, or tumorous. MEDLINE (described earlier in the information extraction section), uses ‘neoplasm’ to return all variations of the term ‘cancer’. However, subject heading searches are more specific as irrelevant articles will not be returned (http://researchguides.uvm.edu/).

Keyword searches return all records containing the term or phrase. Keyword searching is useful in identifying citations missed by subject heading searches if the term is not found in the dictionary/thesaurus. Keyword searches usually return more information than subject heading searches, but the additional information may or may not be relevant.

PubMed utilizes a combination of the two methods. An example is provided. Entering the terms ‘***serum AND lung cancer***’ produces the following search scheme: *("serum"[MeSH Terms] OR "serum"[All Fields]) AND ("lung neoplasms"[MeSH Terms] OR ("lung"[All Fields] AND "neoplasms"[All Fields]) OR "lung neoplasms"[All Fields] OR ("lung"[All Fields] AND "cancer"[All Fields]) OR "lung cancer"[All Fields] AND (hasabstract[text] AND "humans"[MeSH Terms] AND English[lang]).*

In this work, the keyword ‘perspiration’ was used in addition to ‘sweat’; ‘stool’ with ‘feces’; ‘phlegm’ and ‘sputum’ in combination with ‘mucus’; and ‘lacrima’ with tears. PubMed (www.ncbi.nlm.nih.gov/pubmed) queries were performed with the following limits: Abstracts, English, and Human, to retrieve breast and lung cancer abstracts. Query results for diseases-biofluid combinations are found in Table 2. An abstract consists of a journal entry, title, authors, affiliations, text, copyright, and PMID. The sets of abstracts were obtained using criteria from the positive or negative queries (defined below).

Table 2. Size of the abstract sets returned from breast and lung cancer PubMed queries. CSF = cerebrospinal fluid; SF = synovial fluid.



Breast Cancer			Lung Cancer		
Biofluid	Positives	Negatives	Biofluid	Positives	Negatives
Bile	360	40,250	Bile	328	40,290
Blood	18,939	1,540,721	Blood	15,710	1,522,046
Breastmilk	1,047	17,874	Breastmilk	99	18,834
CSF	252	42,711	CSF	298	42,676
Mucus	116	25,122	Mucus	1,445	23,801
Plasma	4,327	342,415	Plasma	3,227	343,678
Saliva	149	22,694	Saliva	86	22,770
Semen	40	12,956	Semen	9	12,989
Serum	7,410	415,218	Serum	6,029	412,897
SF	18	7,669	SF	18	7,671
Stool	123	37,574	Stool	90	37,619
Sweat	321	11,079	Sweat	88	11,314
Tears	40	11,651	Tears	10	11,673
Urine	1,154	125,462	Urine	918	86,776
Total	34,296	2,653,396	Total	28,355	2,595,034

Positive Abstract Sets

A positive abstract set is defined as the set of abstracts obtained by using the keywords, ‘breast cancer AND (biofluid)’, for example breast cancer AND plasma; or ‘lung cancer AND (biofluid)’. From this point forward, all positive abstract sets will be referred to as “positive set(s)”. Positive set queries for breast cancer were performed on 4-29-2013 and for lung cancer on 5-2-2013. An assumption is made that a biomarker mention in these abstract sets is related to both disease and biofluid. PubMed queries output large text files, which were processed using the PittCAPv3 Python script.

Negative Abstract Sets

A negative abstract set is defined as a set of abstracts obtained by using the keywords ‘(biofluid) NOT breast cancer’ or ‘(biofluid) NOT lung cancer’. From this point forward, all negative abstract sets will be called “negative set(s)”. Negative set queries for breast cancer were performed on 4-29-2013 and for lung cancer on 5-2-2013.

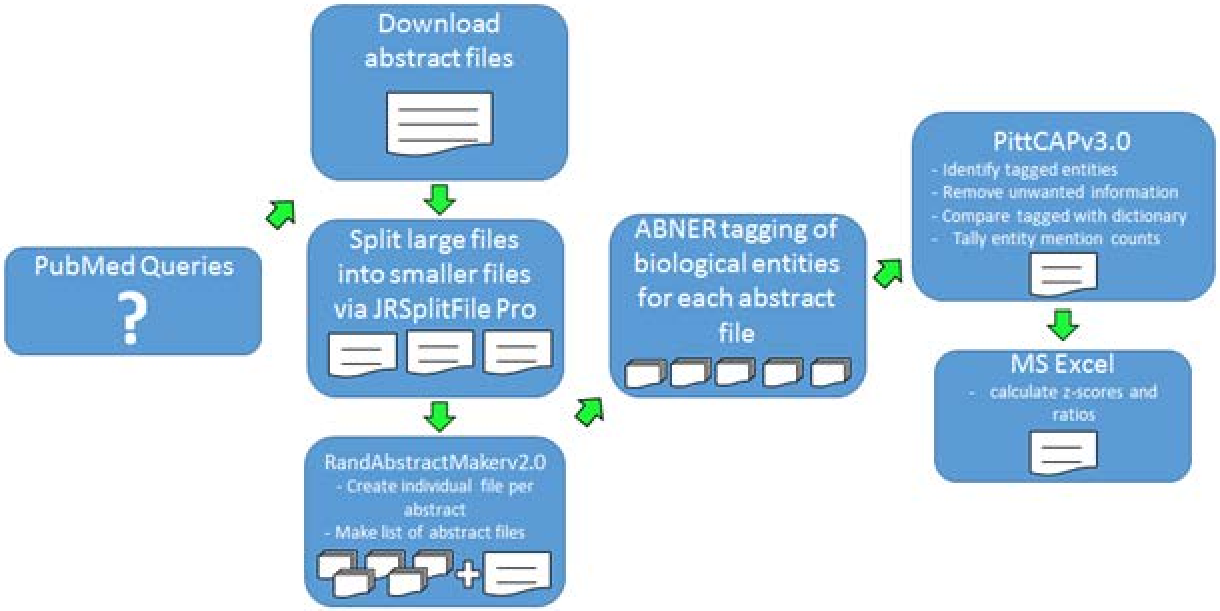

PubMed Queries
?
Download abstract files
Split large files into smaller files via JRSplitFile Pro
RandAbstractMakerv2.0
- Create individual file per abstract
- Make list of abstract files
+
ABNER tagging of biological entities for each abstract file
PittCAPv3.0
- Identify tagged entities
- Remove unwanted information
- Compare tagged with dictionary
- Tally entity mention counts
MS Excel
- calculate z-scores and ratios

Figure 10. Diagram of the KEDA text-mining process.

An overview of the KEDA text-mining process is shown in Figure 10. More than 5.3 million abstracts were obtained from PubMed and examined for biomarker-disease-biofluid associations (34,296 positive and 2,653,396 negative for breast cancer; 28,355 positive and 2,595,034 negative for lung cancer). Biological entity mentions in all positive abstracts were tagged and tallied, and compared to the same findings from negative abstracts. The counts were used to calculate ratios and z-scores for each entity.

Abstract output file

PubMed exports one large text file of all returned abstracts. Once the large files were downloaded from PubMed, JRSplitFile Pro (www.spadixbd.com/jsplit/index.htm) was used to split the large file into smaller 25 MB sized-files. A Python script entitled RandAbstractMaker2.0 was used to subdivide the smaller files even further, creating one individual file per positive abstract. At this point, a list of abstract files was created for input into PittCAPv3.0 (Appendix A).

3.2.1.2 Named entity recognition ABNERv1.5 "A Biomedical Named Entity Recognizer"; (Settles 2005; pages.cs.wisc.edu/~bsettles/abner/) was chosen to perform the entity recognition because of its batch processor which is extremely valuable when processing large numbers of files, and its proven performance pertaining to biological information. Individual abstract files were input to ABNER to tag mentions of proteins, DNA, RNA, cell lines, and cell types in the positive and negative sets. Version 1.5 trains on the NLPBA and BioCreative corpora. Documented ABNER performance measures range from 65.9-77.8 for protein recall and 68.1-74.5 for protein precision. The process described in this work only makes use of entities tagged as "Protein", "DNA", and "RNA".

3.2.1.3 Entity extraction The PittCAPv3 Python script was developed to reduce manual effort and eliminate errors involved in tallying the number of gene/protein mentions from the returned abstracts. The script takes as input a list of abstract file names and the dictionary filename, and performs the following functions: 1) identify tagged entities from the .sgml files output by ABNER and compare mentioned entities to the dictionary; 2) filters out unwanted characters, text, tags and duplicate biomarker mentions; 3) tallies the final count of all biological entity mentions; and 4) produce one final export list containing all confirmed biological entities and their individual counts. Additionally, relevant PMID's were retained for tracking and verification purposes.

3.2.1.4 Dictionary A dictionary file was utilized in order to identify molecular entities of interest, as well as merging the results obtained from different gene/protein aliases under one name. The Protein Nomenclature file was downloaded from the Human Protein Reference Database (HRPD) www.hprd.org/; Copyright © 2002-09, Johns Hopkins University and The Institute of Bioinformatics, for use as the dictionary file. This file contains 19,327 unique protein IDs. The format consists of the HPRD ID, gene symbol, RefSeq ID, and aliases (separated by semi-colons). The gene symbol was used for the consensus name for all accompanying aliases. Found entities were mapped to the dictionary via the PittCapv3 Python script (Appendix A).

3.2.1.5 Z-score calculation Counts were performed at abstract level, with a mention of a biomarker being given a count of 1, regardless of the number of mentions within the abstract. Al-Mubaid & Singh (2005) scoring method was adopted and modified for this work. Each z-score corresponds to a point in a normal distribution and can be associated by its deviation from the mean. The z-scores were computed as follows:

S_1 is the positive abstract set (i.e. disease/biofluid), $S_1 = \{A_1, A_2, \ldots, An\}$.

A is a given abstract,

S_p is the set of markers mentioned in the dictionary and found in the positive set S_1, $S_p = \{P_1, P_2, \ldots, P_m\}$.

S_2 is the negative abstract set.

For each marker P_i in S_p, compute the abstract frequency (af) of P_i in both sets S_1 and S_2 as:

af1(P_i) = number of S_1 abstracts in which P_i is mentioned,

af2(P_i) = number of S_2 abstracts in which P_i is mentioned,

aft(P_i) = af1(P_i) + af2(P_i).

For each marker in Sp compute expectation (ex) and evidence (ev) values:

$$\mathrm{ex}(P_i) = [\mathrm{aft}(P_i)/|S_1 + S_2|] * |S_1|, \quad \text{and} \quad \mathrm{ev}(P_i) = \mathrm{af1}(P_i)$$

Equation 4. Expectant calculation Equation 5. Evident calculation

ex calculates the expected number of mentions of P_i in the positive abstracts set S_1;

ev is a count of the S_1 positive set abstracts that P_i appears in.

The larger the difference in observed and expected abstract frequencies, ev(P_i) – ex(P_i), the more likely that the marker P_i and the disease are significantly associated.

The difference is normalized by:

$$\mathrm{f}(P_i) = (\mathrm{ev}(P_i) - \mathrm{ex}(P_i)) / \mathrm{aft}(P_i)$$

Equation 6. Normalization calculation

The z-score is calculated by:

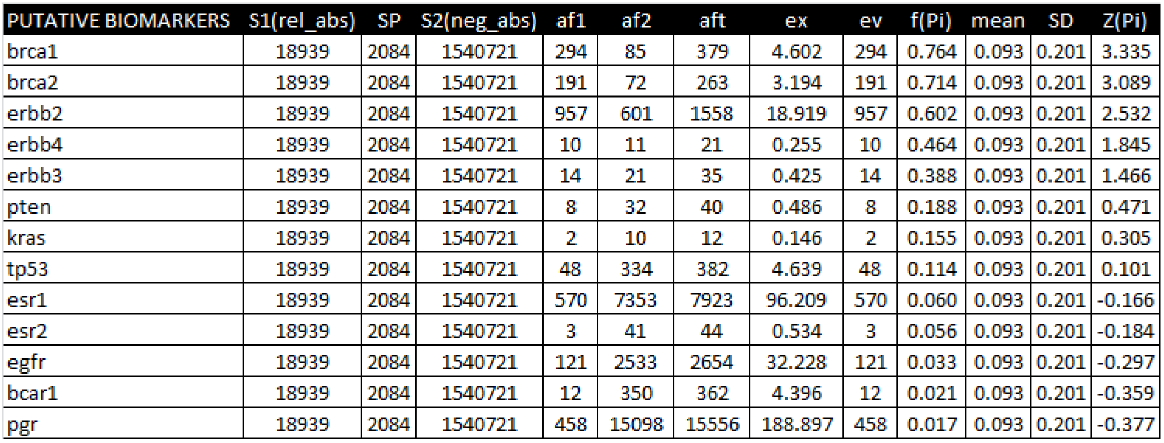
$$\mathrm{Z}(P_i) = [\mathrm{f}(P_i) - \mathrm{mean(f)}]/\mathrm{SD(f)}$$

Equation 7. *Z*-score calculation

where mean(f) = the mean of all f values of all proteins in S_p and SD(f) = the standard deviation of the f values.

A threshold value of 1.0 was established as a significance cut-off based on the results shown in Figure 18. The z-score values were ranked to determine the significance of the putative biomarkers, and to provide measures of disease specific relevance.

Table 3. Breast cancer-related genes from the text-mining final table. Examples measures from breast-cancer/blood final table show how the z-score calculation changes based on the number of positive and negative abstract counts. S1 = # of positive abstracts examined; SP = # of total positive markers; S2 = # negative abstracts examined; af1 = # of mentions in positive abstracts; af2 = # of mentions in negative abstracts; aft = # of mentions in all abstracts; ex = expected number of positive mentions; ev = actual number of positive mentions; f(Pi) = (ev-ex) / aft; mean = average f(Pi) of all biomarkers in table; SD = standard deviation of all biomarkers in table; Z(Pi) = calculated z-score.

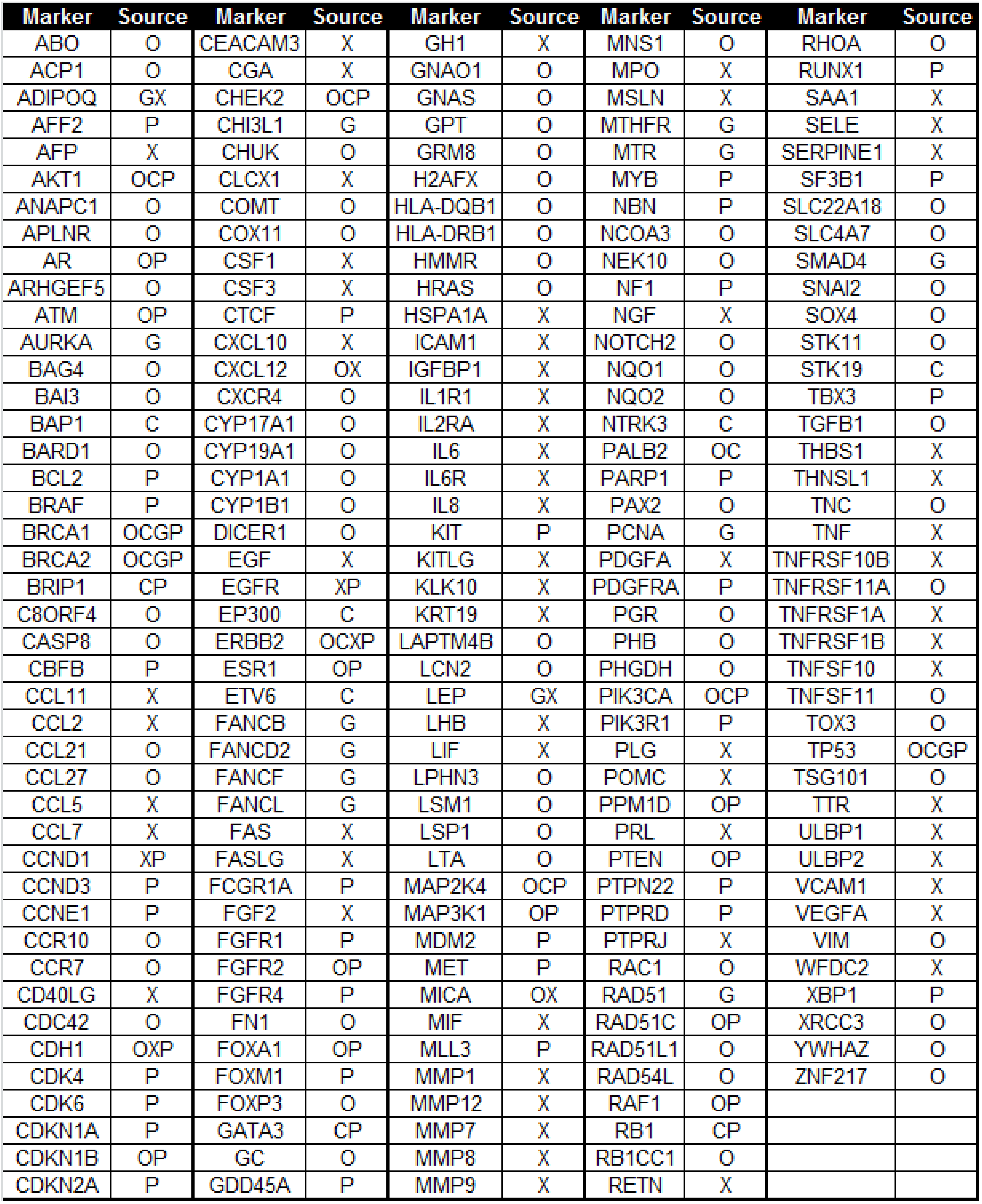


PUTATIVE BIOMARKERS	S1(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
brca1	18939	2084	1540721	294	85	379	4.602	294	0.764	0.093	0.201	3.335
brca2	18939	2084	1540721	191	72	263	3.194	191	0.714	0.093	0.201	3.089
erbb2	18939	2084	1540721	957	601	1558	18.919	957	0.602	0.093	0.201	2.532
erbb4	18939	2084	1540721	10	11	21	0.255	10	0.464	0.093	0.201	1.845
erbb3	18939	2084	1540721	14	21	35	0.425	14	0.388	0.093	0.201	1.466
pten	18939	2084	1540721	8	32	40	0.486	8	0.188	0.093	0.201	0.471
kras	18939	2084	1540721	2	10	12	0.146	2	0.155	0.093	0.201	0.305
tp53	18939	2084	1540721	48	334	382	4.639	48	0.114	0.093	0.201	0.101
esr1	18939	2084	1540721	570	7353	7923	96.209	570	0.060	0.093	0.201	-0.166
esr2	18939	2084	1540721	3	41	44	0.534	3	0.056	0.093	0.201	-0.184
egfr	18939	2084	1540721	121	2533	2654	32.228	121	0.033	0.093	0.201	-0.297
bcar1	18939	2084	1540721	12	350	362	4.396	12	0.021	0.093	0.201	-0.359
pgr	18939	2084	1540721	458	15098	15556	188.897	458	0.017	0.093	0.201	-0.377

The final table (Table 3) was created once the final tally lists were output from PittCAPv3.0. The information in this table was used to calculate z-scores and ratios, for use as prior knowledge, in subsequent KEDA processes (modeling and pathway analysis). Final tables for all biofluids examined can be found in Appendix C.

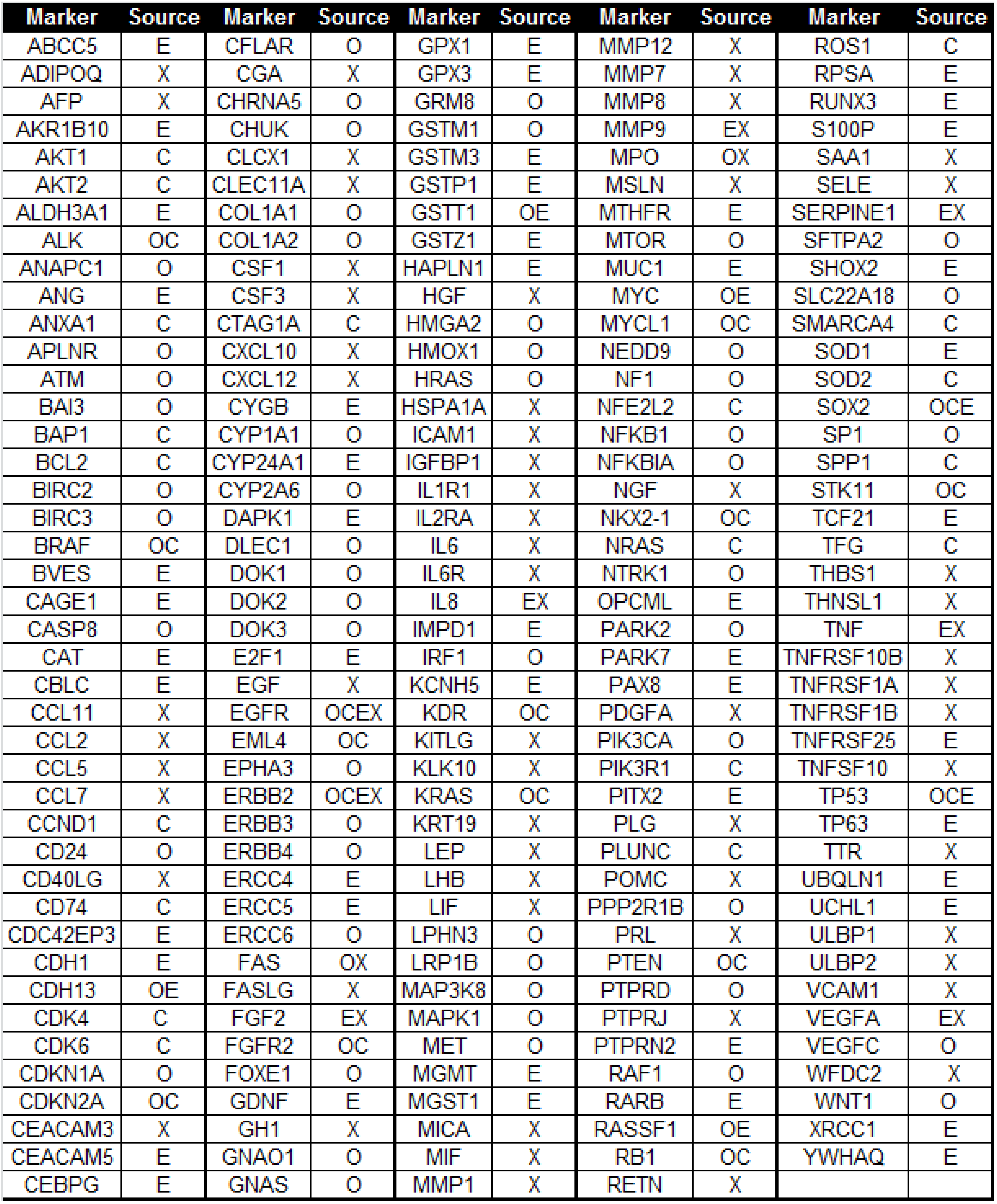
3.2.1.6 Verification of relationships Manual verification of relevant abstracts was performed to assess our method's performance, and to confirm true positive findings. Al-Mubaid & Singh 2005 removed from the abstract pool, 'verification documents' (specifically pertaining to a disease-protein relationship), and used these abstracts for verification. The verification described in this work, does not remove these abstracts, and verification instead, is performed by comparing found results to disease-specific known biomarker lists (Tables 4 & 5). The lists were created from the following sources: OMIM (O; www.ncbi.nlm.nih.gov/omim/; Wheeler *et al.* 2007), cancer gene annotation system for cancer genomics (CAGE(C); mgrc.kribb.re.kr/cage/pageHome.php?m=hm; Park *et al.* 2012) , NCBI's Genes & Disease ((G); www.ncbi.nlm.nih.gov/books/NBK22183/ ; NCBI 1998), NCI's Early Detection Research Network (EDRN (E); edrn.nci.nih.gov/; Wagner & Srivastava 2012), an expert provided list (X) of validated cancer markers, (Bigbee et al 2012), and a recently released breast cancer paper ((P) Cancer Genome Atlas Network 2012)). Markers found in one of these lists, in addition to the HPRD dictionary were considered verified. The breast cancer list was compiled using OMIM, CAGE, Genes & Disease, the expert provided list, and the paper. The lung cancer list was compiled using OMIM, CAGE, EDRN, and the expert provided list.

Table 4. Known Breast Cancer Biomarkers. O = OMIM; C = CAGE; G = NCBI's Genes & Disease; E = EDRN; X = expert provided list of validated cancer markers (Bigbee et al 2012); P = breast cancer paper (Cancer Genome Atlas Network 2012).



Marker	Source	Marker	Source	Marker	Source	Marker	Source	Marker	Source
ABO	O	CEACAM3	X	GH1	X	MNS1	O	RHOA	O
ACP1	O	CGA	X	GNAO1	O	MPO	X	RUNX1	P
ADIPOQ	GX	CHEK2	OCP	GNAS	O	MSLN	X	SAA1	X
AFF2	P	CHI3L1	G	GPT	O	MTHFR	G	SELE	X
AFP	X	CHUK	O	GRM8	O	MTR	G	SERPINE1	X
AKT1	OCP	CLCX1	X	H2AFX	O	MYB	P	SF3B1	P
ANAPC1	O	COMT	O	HLA-DQB1	O	NBN	P	SLC22A18	O
APLNR	O	COX11	O	HLA-DRB1	O	NCOA3	O	SLC4A7	O
AR	OP	CSF1	X	HMMR	O	NEK10	O	SMAD4	G
ARHGEF5	O	CSF3	X	HRAS	O	NF1	P	SNAI2	O
ATM	OP	CTCF	P	HSPA1A	X	NGF	X	SOX4	O
AURKA	G	CXCL10	X	ICAM1	X	NOTCH2	O	STK11	O
BAG4	O	CXCL12	OX	IGFBP1	X	NQO1	O	STK19	C
BAI3	O	CXCR4	O	IL1R1	X	NQO2	O	TBX3	P
BAP1	C	CYP17A1	O	IL2RA	X	NTRK3	C	TGFB1	O
BARD1	O	CYP19A1	O	IL6	X	PALB2	OC	THBS1	X
BCL2	P	CYP1A1	O	IL6R	X	PARP1	P	THNSL1	X
BRAF	P	CYP1B1	O	IL8	X	PAX2	O	TNC	O
BRCA1	OCGP	DICER1	O	KIT	P	PCNA	G	TNF	X
BRCA2	OCGP	EGF	X	KITLG	X	PDGFA	X	TNFRSF10B	X
BRIP1	CP	EGFR	XP	KLK10	X	PDGFRA	P	TNFRSF11A	O
C8ORF4	O	EP300	C	KRT19	X	PGR	O	TNFRSF1A	X
CASP8	O	ERBB2	OCXP	LAPTM4B	O	PHB	O	TNFRSF1B	X
CBFB	P	ESR1	OP	LCN2	O	PHGDH	O	TNFSF10	X
CCL11	X	ETV6	C	LEP	GX	PIK3CA	OCP	TNFSF11	O
CCL2	X	FANCB	G	LHB	X	PIK3R1	P	TOX3	O
CCL21	O	FANCD2	G	LIF	X	PLG	X	TP53	OCGP
CCL27	O	FANCF	G	LPHN3	O	POMC	X	TSG101	O
CCL5	X	FANCL	G	LSM1	O	PPM1D	OP	TTR	X
CCL7	X	FAS	X	LSP1	O	PRL	X	ULBP1	X
CCND1	XP	FASLG	X	LTA	O	PTEN	OP	ULBP2	X
CCND3	P	FCGR1A	P	MAP2K4	OCP	PTPN22	P	VCAM1	X
CCNE1	P	FGF2	X	MAP3K1	OP	PTPRD	P	VEGFA	X
CCR10	O	FGFR1	P	MDM2	P	PTPRJ	X	VIM	O
CCR7	O	FGFR2	OP	MET	P	RAC1	O	WFDC2	X
CD40LG	X	FGFR4	P	MICA	OX	RAD51	G	XBP1	P
CDC42	O	FN1	O	MIF	X	RAD51C	OP	XRCC3	O
CDH1	OXP	FOXA1	OP	MLL3	P	RAD51L1	O	YWHAZ	O
CDK4	P	FOXM1	P	MMP1	X	RAD54L	O	ZNF217	O
CDK6	P	FOXP3	O	MMP12	X	RAF1	OP		
CDKN1A	P	GATA3	CP	MMP7	X	RB1	CP		
CDKN1B	OP	GC	O	MMP8	X	RB1CC1	O		
CDKN2A	P	GDD45A	P	MMP9	X	RETN	X		

Table 5. Known Lung Cancer Biomarkers. O = OMIM; C = CAGE; G = NCBI's Genes & Disease; E = EDRN; X = expert provided list of validated cancer markers (Bigbee et al 2012); P = breast cancer paper (Cancer Genome Atlas Network 2012).



Marker	Source	Marker	Source	Marker	Source	Marker	Source	Marker	Source
ABCC5	E	CFLAR	O	GPX1	E	MMP12	X	ROS1	C
ADIPOQ	X	CGA	X	GPX3	E	MMP7	X	RPSA	E
AFP	X	CHRNA5	O	GRM8	O	MMP8	X	RUNX3	E
AKR1B10	E	CHUK	O	GSTM1	O	MMP9	EX	S100P	E
AKT1	C	CLCX1	X	GSTM3	E	MPO	OX	SAA1	X
AKT2	C	CLEC11A	X	GSTP1	E	MSLN	X	SELE	X
ALDH3A1	E	COL1A1	O	GSTT1	OE	MTHFR	E	SERPINE1	EX
ALK	OC	COL1A2	O	GSTZ1	E	MTOR	O	SFTPA2	O
ANAPC1	O	CSF1	X	HAPLN1	E	MUC1	E	SHOX2	E
ANG	E	CSF3	X	HGF	X	MYC	OE	SLC22A18	O
ANXA1	C	CTAG1A	C	HMGA2	O	MYCL1	OC	SMARCA4	C
APLNR	O	CXCL10	X	HMOX1	O	NEDD9	O	SOD1	E
ATM	O	CXCL12	X	HRAS	O	NF1	O	SOD2	C
BAI3	O	CYGB	E	HSPA1A	X	NFE2L2	C	SOX2	OCE
BAP1	C	CYP1A1	O	ICAM1	X	NFKB1	O	SP1	O
BCL2	C	CYP24A1	E	IGFBP1	X	NFKBIA	O	SPP1	C
BIRC2	O	CYP2A6	O	IL1R1	X	NGF	X	STK11	OC
BIRC3	O	DAPK1	E	IL2RA	X	NKX2-1	OC	TCF21	E
BRAF	OC	DLEC1	O	IL6	X	NRAS	C	TFG	C
BVES	E	DOK1	O	IL6R	X	NTRK1	O	THBS1	X
CAGE1	E	DOK2	O	IL8	EX	OPCML	E	THNSL1	X
CASP8	O	DOK3	O	IMPD1	E	PARK2	O	TNF	EX
CAT	E	E2F1	E	IRF1	O	PARK7	E	TNFRSF10B	X
CBLC	E	EGF	X	KCNH5	E	PAX8	E	TNFRSF1A	X
CCL11	X	EGFR	OCEX	KDR	OC	PDGFA	X	TNFRSF1B	X
CCL2	X	EML4	OC	KITLG	X	PIK3CA	O	TNFRSF25	E
CCL5	X	EPHA3	O	KLK10	X	PIK3R1	C	TNFSF10	X
CCL7	X	ERBB2	OCEX	KRAS	OC	PITX2	E	TP53	OCE
CCND1	C	ERBB3	O	KRT19	X	PLG	X	TP63	E
CD24	O	ERBB4	O	LEP	X	PLUNC	C	TTR	X
CD40LG	X	ERCC4	E	LHB	X	POMC	X	UBQLN1	E
CD74	C	ERCC5	E	LIF	X	PPP2R1B	O	UCHL1	E
CDC42EP3	E	ERCC6	O	LPHN3	O	PRL	X	ULBP1	X
CDH1	E	FAS	OX	LRP1B	O	PTEN	OC	ULBP2	X
CDH13	OE	FASLG	X	MAP3K8	O	PTPRD	O	VCAM1	X
CDK4	C	FGF2	EX	MAPK1	O	PTPRJ	X	VEGFA	EX
CDK6	C	FGFR2	OC	MET	O	PTPRN2	E	VEGFC	O
CDKN1A	O	FOXE1	O	MGMT	E	RAF1	O	WFDC2	X
CDKN2A	OC	GDNF	E	MGST1	E	RARB	E	WNT1	O
CEACAM3	X	GH1	X	MICA	X	RASSF1	OE	XRCC1	E
CEACAM5	E	GNAO1	O	MIF	X	RB1	OC	YWHAQ	E
CEBPG	E	GNAS	O	MMP1	X	RETN	X		

3.2.1.7 Error rate determination In order to assess the performance of our text-mining method, an error rate metric was sought. A true positive rate is currently unattainable as a comprehensive pool of breast or lung cancer biomarkers does not exist. To calculate the error rate of our method findings, the following equation was used:

$$\text{Error} = TP / (TP + FP)$$

Equation 8. Error rate calculation

where TP are true positives and FP are false positives.

The use of negative abstracts inherently eliminates some false positives. However, it was determined that manual examination of abstracts would be required to ensure that the results obtained were not false positives. Tracking the PubMed ID allowed for manual verification of relevant abstracts. In tracking the abstracts, three criteria were used to determine a pass/fail outcome. Abstracts were examined for mentions of biofluid, disease, and biomarker. All three criteria were required to be acceptable and counted as a TP. Synonyms or root words were also deemed acceptable. FN’s would include those genes/proteins that appear in the final list, but were missing one of the aforementioned criteria.

3.2.2 Classification modeling methodology

Prior information can be combined with experimental data and included in modeling exercises to add an additional level of confidence in modeling results. Gene-level experimental data was used due to informed prior ratios being developed at the gene-level. To ensure that findings were not experimental-type or platform-specific, several different breast and lung cancer datasets were examined. While discretization could yield a larger number of value ranges for a variable, thereby increasing the number of rules generated by BRL (Gopalakrishnan *et al.* 2010), it was not used in this work. Comparisons were made between datasets that either included informed priors, uniform priors, or no priors at all (data only). The following sections describe the datasets used, how the data was processed, and the implementation of the BRL modeling algorithm.

3.2.2.1 Experimental datasets In this section, a description of the 'omic' datasets used for model development and testing of the KEDA framework are presented. Publicly available breast and lung cancer experimental datasets were acquired via the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/index.cgi; detailed dataset descriptions can be found in Appendix B). GEO currently houses 3848 datasets (7-4-15). Several different types of data and platforms were examined to ensure the obtained results were not specific for a certain type of experiment or platform.

Datasets of interest were found using the following keyword search: human, gene expression whole blood, lung or breast cancer. Datasets containing the greatest number of samples were given highest priority, and subsequently downloaded and analyzed.

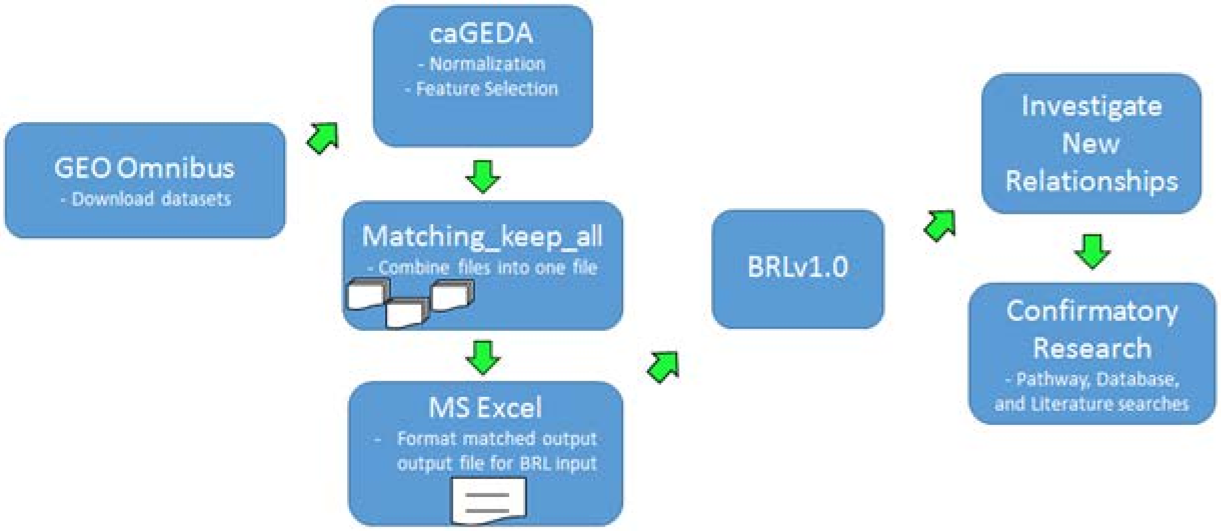

GEO Omnibus
- Download datasets
caGEDA
- Normalization
- Feature Selection
Matching_keep_all
- Combine files into one file
MS Excel
- Format matched output output file for BRL input
BRLv1.0
Investigate New Relationships
Confirmatory Research
- Pathway, Database, and Literature searches

Figure 11. Diagram of the KEDA classification modeling process.

Figure 11 provides a diagram of the KEDA modeling process. All subsequent sections will utilize the datasets and platforms mentioned earlier.

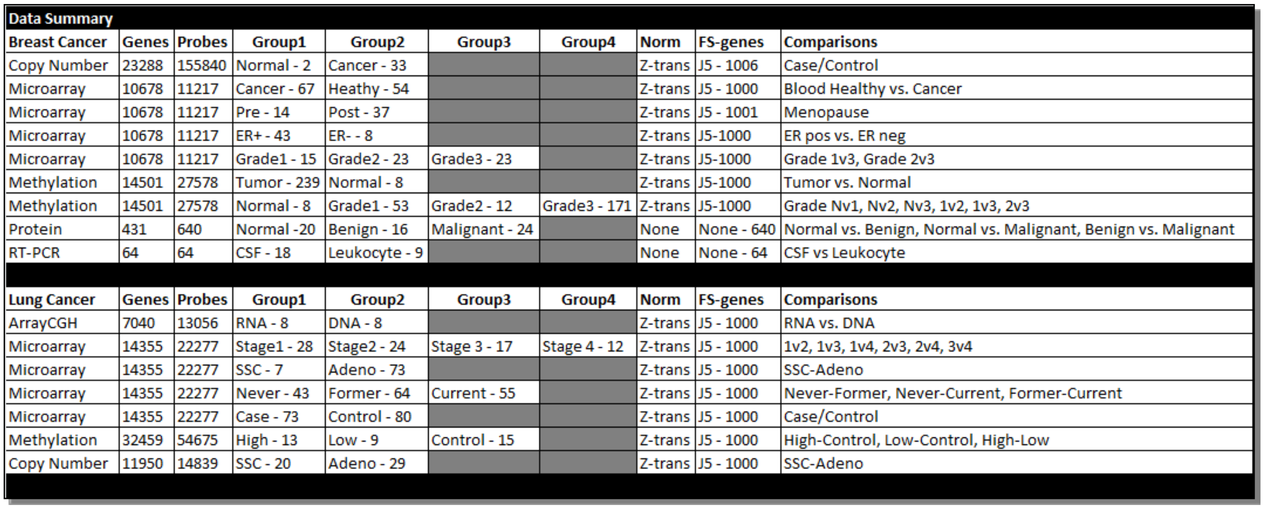
The following section provides a more in-depth description of the GEO datasets used in this work. The summaries accompanied the datasets and were provided by Gene Expression Omnibus.

Gene Expression

DNA which contains an organisms' blueprint in located in the cell's nucleus, which is encased in a membrane. DNA molecules are too large to leave the nucleus, so copies of sections (genes) of the DNA are 'transcribed'. These copies are messenger RNA (mRNA) transcripts. The mRNA molecules are small enough to leave the nucleus and enter the cytoplasm. In the cytoplasm, the mRNA will encounter ribosomes, where the mRNA sequence is 'translated' into proteins. Proteins are created in response to the cellular environment, and engage in controlling cellular behavior. Gene expression studies are important in determining which genes are active given certain environmental conditions.

Gene expression microarrays measure the specific amount of mRNA transcripts in a sample. Arrays vary in their technology, but generally speaking a microarray consists of thousands of gene-specific probes (DNA gene sequences or complimentary sequences) being hybridized to a surface (usually called a 'chip'). The mRNA in a cell is captured and labelled with an illuminescent that will emit light when activated by a laser. The mRNA that match the gene sequence will bind to the probes, with the unbound sequences being washed away. The chip is then exposed to the laser and the brightness of each spot is captured and calculated to produce an intensity value. The greater the intensity value, the more copies of the mRNA are believed to be found within the cells. In this way, researchers can determine which genes are 'active' or 'on' and which genes are 'not active' or 'off'. Conclusions can be made based on comparing intensity values from one cell type vs. another, or one environmental condition vs. another. In this work, one breast cancer and one lung cancer microarray dataset was examined, enabling several different comparisons to be made (see Table 6):

Table 6. Breast and lung cancer dataset summary. Gene-level data was used for comparisons. Several different comparisons could be made from one dataset. Norm = normalization method; FS = feature selection; CSF = cerebrospinal fluid; SSC = small cell carcinoma; Adeno = adenocarcinoma.



Data Summary

Breast Cancer	Genes	Probes	Group1	Group2	Group3	Group4	Norm	FS-genes	Comparisons
Copy Number	23288	155840	Normal - 2	Cancer - 33			Z-trans	J5 - 1006	Case/Control
Microarray	10678	11217	Cancer - 67	Heathy - 54			Z-trans	J5 - 1000	Blood Healthy vs. Cancer
Microarray	10678	11217	Pre - 14	Post - 37			Z-trans	J5 - 1001	Menopause
Microarray	10678	11217	ER+ - 43	ER- - 8			Z-trans	J5-1000	ER pos vs. ER neg
Microarray	10678	11217	Grade1 - 15	Grade2 - 23	Grade3 - 23		Z-trans	J5-1000	Grade 1v3, Grade 2v3
Methylation	14501	27578	Tumor - 239	Normal - 8			Z-trans	J5-1000	Tumor vs. Normal
Methylation	14501	27578	Normal - 8	Grade1 - 53	Grade2 - 12	Grade3 - 171	Z-trans	J5-1000	Grade Nv1, Nv2, Nv3, 1v2, 1v3, 2v3
Protein	431	640	Normal -20	Benign - 16	Malignant - 24		None	None - 640	Normal vs. Benign, Normal vs. Malignant, Benign vs. Malignant
RT-PCR	64	64	CSF - 18	Leukocyte - 9			None	None - 64	CSF vs Leukocyte

Lung Cancer	Genes	Probes	Group1	Group2	Group3	Group4	Norm	FS-genes	Comparisons
ArrayCGH	7040	13056	RNA - 8	DNA - 8			Z-trans	J5 - 1000	RNA vs. DNA
Microarray	14355	22277	Stage1 - 28	Stage2 - 24	Stage 3 - 17	Stage 4 - 12	Z-trans	J5 - 1000	1v2, 1v3, 1v4, 2v3, 2v4, 3v4
Microarray	14355	22277	SSC - 7	Adeno - 73			Z-trans	J5 - 1000	SSC-Adeno
Microarray	14355	22277	Never - 43	Former - 64	Current - 55		Z-trans	J5 - 1000	Never-Former, Never-Current, Former-Current
Microarray	14355	22277	Case - 73	Control - 80			Z-trans	J5 - 1000	Case/Control
Methylation	32459	54675	High - 13	Low - 9	Control - 15		Z-trans	J5 - 1000	High-Control, Low-Control, High-Low
Copy Number	11950	14839	SSC - 20	Adeno - 29			Z-trans	J5 - 1000	SSC-Adeno

Breast Cancer Microarray (Aarøe *et al.* 2010)

This dataset is titled: Gene expression profiling of peripheral blood cells for early detection of breast cancer (GSE16443). It utilized expression profiling by array. The platform is the ABI Human Genome Survey Microarray Version 2 (GPL2986). Multiple comparisons were performed with this dataset: Blood Healthy vs. Cancer, Menopause Status, ESR1 positive vs. ESR1 negative (see section 2.1), and Tumor Grade 1v3 and

2v3. Various sample sizes were used for the different comparisons with the largest sample size being 67 cancer bloods vs. 54 normal bloods. 11217 probes covering 10678 genes were analyzed.

Lung Cancer Microarray (Rotunno *et al.* 2011)

This dataset is titled: A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma (GSE20189). It utilized expression profiling by array. The platform is a commercial [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array (GPL571) with in situ oligonucleotides. Several comparisons were performed for this analysis: Stages 1-4; small cell carcinoma to adenocarcinoma; smoking histories (never, former, current); and case-control (cancer vs. normal) status. Varying sample size was available for each comparison with the largest sample size being 73 adenocarcinoma samples vs. 80 control samples. 22277 probes covering 14355 genes were analyzed.

Copy number

Each organism contains a specific number of chromosomes in each cell. For example, humans are known to have 23 pairs or 46 total chromosomes that make up their ‘genome’. Occasionally, or over time, the chromosome number can deviate from the norm due to deletions, insertions, inversions, and duplications which may result during cell replication. As a result, some cells may contain extra or limited numbers of certain chromosomes or chromosomal regions or genes. Copy number variation arrays measure chromosomal aberrations which may ultimately result in changes in the physical arrangement of genes on chromosomes (Feuk *et al.* 2006). Copy number is important because many medical complications can occur when chromosomal aberrations exist.

Polymerase chain reaction (PCR) is a molecular biology technique that is used to amplify regions of DNA exponentially, over and over again, ultimately resulting in millions of copies from minimal starting material. DNA usually exists in nature as a double-stranded molecule, with the two strands being held together by hydrogen bonds. Heating DNA to temperatures of 95°C ‘denatures’ the DNA by breaking the hydrogen bonds that hold the molecule together, which results in 2 molecules of single-stranded DNA. Once the DNA is denatured, the temperature is lowered to around 50-60°C, for the primers to ‘anneal’ or to bind to the single-stranded DNA molecules. Primers are short segments of DNA (~20-25 base pairs) that are complimentary to the DNA sequence that will bind to the DNA, and allow for an enzyme called DNA polymerase to bind. The temperature is raised to 75°C which is optimal for the enzyme to bind to the primer, and the ‘elongation step’ will occur where DNA polymerase will add nucleotides that are complimentary to the single-strand DNA molecule to create a complimentary DNA strand. In doing so one DNA molecule is replicated into two molecules. Thirty or so rounds of heating and cooling (thermal cycling) resulting in replication take place resulting in exponential numbers of the same DNA molecules.

Copy number variation assays utilize a real-time PCR system. Real-time PCR (qPCR) tracks the number of DNA molecules during the PCR process by labelling the DNA molecules with fluorescent dyes. In doing so it can determine if the number of molecules is more or less than expected. If the number is greater than expected, one can conclude that there were more than the normal number of copies (two copies of each gene total, one from father, and one from mother) of the DNA present initially. If the number is less than expected, one can conclude that there was less than the normal number of copies present initially. In this work, one breast cancer and one lung cancer copy number dataset was examined (see Table 6):

Breast Cancer Copy Number (Mathiesen et al. 2012)

This dataset is titled: High-resolution analysis of copy number changes in circulating and disseminated tumor cells in breast cancer patients (GSE27574). It utilized genome variation profiling by array. The platform is the Agilent-014693 Human Genome CGH Microarray 244A (GPL9128). Circulating tumor cells were compared to normal cells, across 155840 probes covering 23288 genes.

Lung Cancer Copy Number (Starczynowski *et al.* 2011)

This dataset is titled: DNA copy number and gene expression profiles of resected non-small cell lung cancer tumors (GSE31800). It utilized genome variation profiling by genome tiling array. The platform is custom-commercial Custom Rosetta-Affymetrix Human platform [rmhu01aa520485] (GPL14189) with spotted oligonucleotides. The comparison was small-cell carcinomas vs. adenocarcinomas. 14839 probes covering 11950 genes were analyzed.

RT-PCR

Reverse Transcription-Polymerase Chain Reaction (RT-PCR) is a molecular biology technique used to determine gene expression using RNA as a template and converting it into cDNA using an enzyme called reverse transcriptase. The cDNA is then amplified using the classical PCR technique.

A reaction mix containing nucleotides, primers, RNA, and enzyme is created. The reaction mix undergoes thermal cycling with the first cycle being the reverse transcription step which synthesizes single-strand cDNA. Inactivation of reverse transcriptase follows. Many cycles of denaturation, annealing and elongation occur which amplify the cDNA. Results are assessed by gel electrophoresis. In this work, one breast cancer RT-PCR dataset was examined (see Table 6):

Breast Cancer RT-PCR (Magbanua *et al.* 2013)

This dataset is titled: Molecular characterization of tumor cells from the cerebrospinal fluid and matched primary tumors from metastatic breast cancer patients with leptomeningeal disease (GSE46068). It utilized expression profiling by RT-PCR. The platform is the Custom Human TLDA 64-Circulating tumor cell associated gene panel (GPL17020). Tumor cells from CSF (n-18) vs. primary leukocytes (n=9) from metastatic breast cancer patients were compared. 64 genes were analyzed.

Methylation

Methylation occurs when a methyl group is added to a cytosine nucleotide of a DNA molecule by enzymes called methyltransferases. Methylation appears to occur most often in regions called cytosine-phosphate-guanine (CpG) islands, which can be found in gene regulatory regions. As such, this process plays a role in gene expression regulation by inhibiting the binding of proteins necessary for transcription. When methylation occurs in a promoter region, gene transcription is usually prohibited. Gene expression regulation is essential in different stages of cell progression. Certain genes need to be expressed early in development, but then are not needed again as aging occurs. Methylation aids in terminating the expression of such genes once they are no longer needed.

By being able to inhibit gene expression, methylation plays a key role in defending cells from detrimental conditions, interfering with viral-DNA expression for example. However, methylation can also be harmful if it occurs in the wrong area, such as silencing tumor suppressor genes. DNA methylation patterns can be inherited from mother cells to daughter cells. As such methylations can accumulate over time, thus methylation has been implicated as an indicator of the aging process.

Like other array processes, methylation arrays use probes designed specifically for predetermined loci, hybridized to a surface. These arrays are used to measure methylation intensity across the genome. The following is a description of the most common methylation assay technology from the Illumina Infinuim Methylation Assay website (http://www.illumina.com/technology/beadarray-technology/infinium-methylation-assay.html):

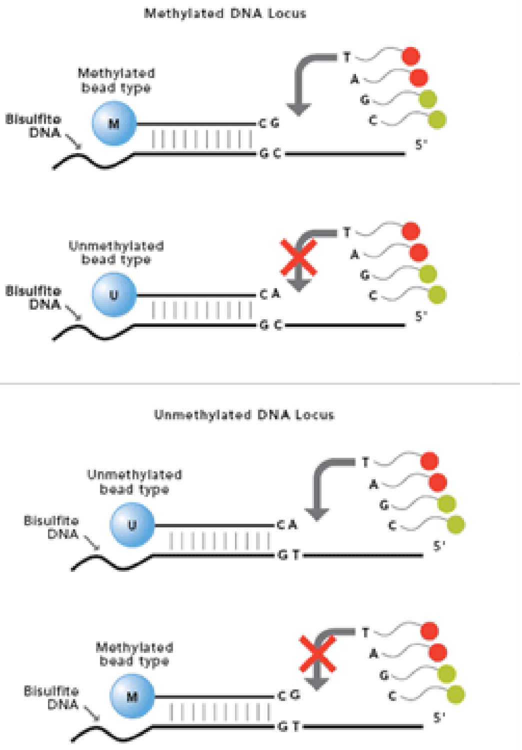


Figure 12. Infinium methylation assay bead technology. The Infinium Methylation Assay uses two different bead types to detect CpG methylation. The U bead type matches the unmethylated CpG site; the M bead type matches the methylated site. In the top figure, the unmethylated CpG target site matches with the U probe, enabling single-base extension and detection. It has a single-base mismatch to the M probe, which inhibits extension. If the CpG locus of interest is methylated (bottom figure), the reverse occurs. http://www.illumina.com/technology/beadarray-technology/infinium-methylation-assay.html

The assays work by "detecting cytosine methylation at CpG islands based on genotyping of bisulfite-converted genomic DNA. Following treatment with bisulfite, unmethylated cytosines are converted to uracil, while methylated cytosines remain unchanged. The loci are interrogated by using two site-specific probes (Figure 12), one for the methylated locus (M bead) and another for the unmethylated locus (U bead). Single-base extension of the probes incorporates a labeled nucleotide, which is tagged with a fluorescence reagent. The level of methylation for the interrogated locus can be determined by calculating the ratio of the fluorescent signals from the methylated vs unmethylated sites.” An aid to the above description is provided in Figure 12, as the wording can be complicated.

Methylation array technology has been utilized extensively in lung and breast cancer studies, with methylation signatures even being researched as possible biomarker panels. In this work, one breast cancer and one lung cancer methylation datasets was examined (see Table 6).

Breast Cancer Methylation (Dedeurwaerder *et al.* 2011)

This dataset is titled: Epigenetic portraits of human breast cancers (GSE20713). It utilized methylation profiling by array. The platform is the Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2) (GPL8490). Several comparisons were performed for this analysis: Tumor vs. Normal, as well as Tumor grade comparisons (Grade Nv1, Nv2, Nv3, 1v2, 1v3, and 2v3). Various sample sizes were used for the different comparisons with the largest sample size being 238 tumor samples vs. 8 normal. 27578 probes covering 14501 genes were analyzed.

Lung Cancer Methylation (Shames *et al.* 2006)

This dataset is titled: Genome-wide screen for hypermethylated genes in lung cancer (GSE5816). It utilized expression profiling by array. The platform is a commercial [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) with in situ oligonucleotides. Comparisons performed were based on treatment protocol: Control treatment group (DMSO qod for 6 days); Low dose treatment group (5-aza-2'-

deoxycytidine qod for 6 days); and High dose (1000nM) treatment group (5-aza-2'-deoxycytidine qod for 6 days). 54675 probes covering 32459 genes were analyzed.

ArrayCGH

Array comparative genomic hybridization (aCGH) is a molecular method for comparing copy number variations (CNVs; gain or loss of chromosomal regions) from two DNA samples such as a test sample and a reference. It allows for testing all chromosomes, locus-by-locus, for deletions and duplications, on a high-resolution scale (Pinkel & Albertson 2005) at a level of 5–10 kb (Ren et al. 2005).Two approaches exist for aCGH, whole genome and targeted areas. Whole-genome arrays are used more often in research applications, while targeted arrays are used to target specific areas of interest and can be used for clinical applications.

ACGH employs the same principles of competitive fluorescence in situ hybridization as traditional CGH. Equal quantities of DNA samples are labelled with different color fluorophores (usually red/Cyanine 5/Cy5 and green/Cyanine 3.Cy3) and are then used as probes that competitively hybridize to a microarray of thousands of spotted target sequences. After hybridization, the remaining unbound sample and fluorophore are washed away, and a digital imaging system is used to quantify the fluorescence intensities of each probe/target. The ratio of intensities is proportional to the ratio of the copy number in the sample and reference genomes. When the intensities are somewhat equal, the copy number of the samples compared is assumed to be equal. A yellow color on the array indicates no difference between the samples in that location (Strachan & Read 2010; Weiss et al 1999). A greater intensity of the sample color red indicates a loss of DNA, while a higher intensity of reference color green indicates a gain of

DNA at the given locus (Shinawi & Cheung 2008). In this work, one lung cancer methylation datasets was examined (see Table 6):

<u>Lung cancer ArrayCGH</u> (Medina *et al.* 2009)

This dataset is titled: Gene expression analysis & comparative genomic hybridization from lung cancer cell lines (GSE14079). It utilized expression and genome variation profiling by array. The platform is a non-commercial CNIO H. sapiens 13.6K Oncochip 1 (GPL1998) with spotted DNA/cDNA technology. DNA samples were compared to matched RNA samples, from Homo sapiens lung cancer cell lines, across 13056 probes covering 7040 genes.

<u>Protein arrays</u>

A protein microarray is a molecular biology technique developed to study protein interactions, and function. The array is usually a slide, nitrocellulose membrane, or microtiter plate. ’Capture proteins’ are attached to act as the probe molecules, and protein samples labeled with fluorescent dye, are added as the targets. When the fluorescent dye is hit by a laser, light of a specific wavelength is emitted, and the intensity is read by a scanner.

Protein arrays were developed because the quantity of mRNA in a cell doesn’t necessarily reflect the level of protein, and it is the proteins that are the functional entities of the cell. The surface coating anchors the capture proteins to the surface, as well as prevents protein denaturation, orients the protein for optimal binding, and inhibits non-specific binding to minimize background noise. The capture molecules used usually are antibodies, antigens, partial or full-length proteins. There are three types of protein microarrays that are commonly used:

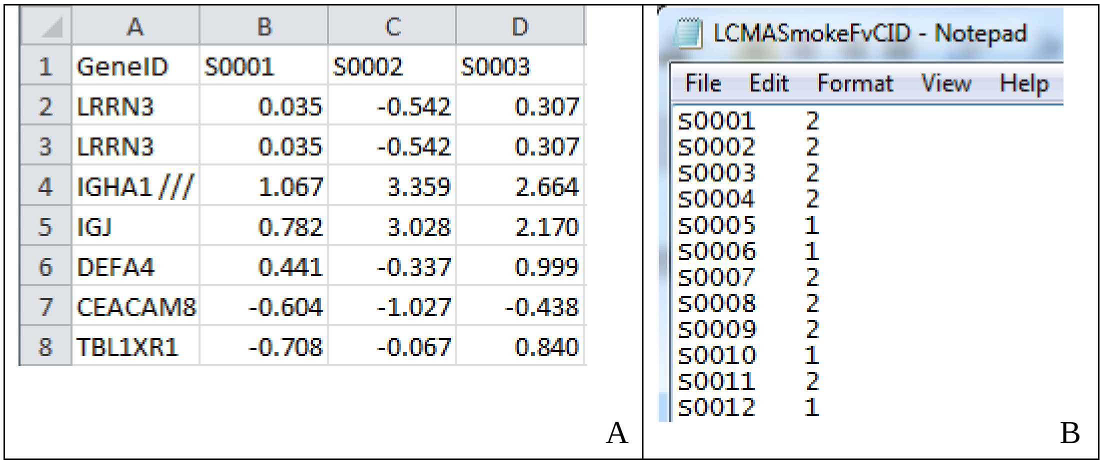
analytical (capture) arrays, functional (target) arrays, and reverse-phase arrays. In analytical arrays a library of antibodies is arrayed on the support surface as capture molecules. The array is probed with the sample containing proteins, and analysis provides information about the amount of protein as well as binding properties. In functional arrays purified proteins are used as probes to identify protein interactions, enzymatic activity and detect antibody specificity. Reverse phase arrays (RPAs) are used to study complex samples. Cell lysates is arrayed onto the surface and probed with antibodies to the target protein. Reference peptides are included on slides for protein quantification of the sample lysates. RPAs allow for the determination of the presence of altered proteins or other agents that may be results of a given disease. Post-translational modifications may also be detected using RPAs. In this work, one breast cancer protein array dataset was examined (see Table 6):

Breast Cancer Protein (no citation)

This dataset is titled: Evaluation of auto-antibody serum biomarkers for breast cancer screening (GSE34555). It utilized protein profiling by protein array. The platform is the Austrian Institution of Technology Protein Array 642 (GPL15009). Multiple comparisons were performed with this dataset: Normal vs. Benign, Normal vs. Malignant, and Benign vs. Malignant. Various sample sizes were used for the different comparisons with the largest sample size being 20 normal vs. 24 malignant. 640 probes covering 431 proteins were analyzed.

3.2.2.2 Pre-processing steps

Downloaded datasets were transformed/normalized for comparison purposes. CaGEDA (Patel & Lyons-Weiler 2004) is a resource developed to aid in data analysis. Downloaded datasets were formatted (ensuring datatype for each row and column was the required type) using Microsoft Excel 2010. The data was input as a text-file with the first column containing gene/protein IDs and subsequent columns containing data (Figure 13, A).



	A	B	C	D
1	GeneID	S0001	S0002	S0003
2	LRRN3	0.035	-0.542	0.307
3	LRRN3	0.035	-0.542	0.307
4	IGHA1 ///	1.067	3.359	2.664
5	IGJ	0.782	3.028	2.170
6	DEFA4	0.441	-0.337	0.999
7	CEACAM8	-0.604	-1.027	-0.438
8	TBL1XR1	-0.708	-0.067	0.840

```
S0001	2
S0002	2
S0003	2
S0004	2
S0005	1
S0006	1
S0007	2
S0008	2
S0009	2
S0010	1
S0011	2
S0012	1
```

Figure 13. Examples of portions of caGEDA file formats. A) Example of caGEDA Format 1 data spreadsheet. Duplicate values for GeneID's are averaged by caGEDA. B) Example of the caGEDA sample identification file. The sample name should be in column 1 and the class in column 2, identified as 1 (case) or 2 (control).

An additional file describing sample ID and group number (i.e. case = group 1; control = group 2) is also required (Figure 13, B). All datasets underwent z-transformation, and the J5 test was used for feature selection (to reduce the number of genes to be examined by BRL; see section 3.2.2.4).

The z-transformation is a function applied to every data point in a dataset that converts the values of a sample into z-scores using the formula

$$z_i = \frac{x_i - \bar{x}}{s}$$

Equation 9. Z-transformation formula

where z_i is the z-transformed sample observations, x_i is the original values of the sample, $\bar{x}$ is the sample mean, and s is the standard deviation of the sample. The z-transform of two datasets results in comparable distributions since both z-transformed distributions have a mean of 0.0 and a standard deviation of 1.0 (http://www.statistics4u.info/fundstat_eng/ee_ztransform.html).

The J5 test is a gene-specific ratio between the mean difference in expression intensity between two groups, A and B, to the average mean group difference of all m genes.

$$J5_i = \frac{\bar{A}_i - \bar{B}_i}{\frac{1}{m}\sum_{j=1}^{m}\left|\bar{A}_j - \bar{B}_j\right|}$$

Equation 10. The J5 formula

The J5 test is likely to be useful in pilot studies where, due to high variance, t-tests are likely to exhibit unacceptably low specificity (high false discovery rates) (Patel *et al.* 2004).

The z-transformed dataset containing new normalized values can be directly downloaded from the caGEDA website (http://bioinformatics.upmc.edu/GE2/GEDA.html; temporarily disabled) as a text file from the Analysis Results Page. A specific number of genes can be returned by setting a threshold which corresponds to the J5 Score. The dataset of retained genes can also be downloaded from the Results Page. To limit the number of genes that the modeling algorithm must examine, the feature selection cutoff was set to 1000 genes; the analyses were all performed at the gene-level, and thus experimental probe values pertaining to the same gene were averaged together to produce one value per gene per sample.

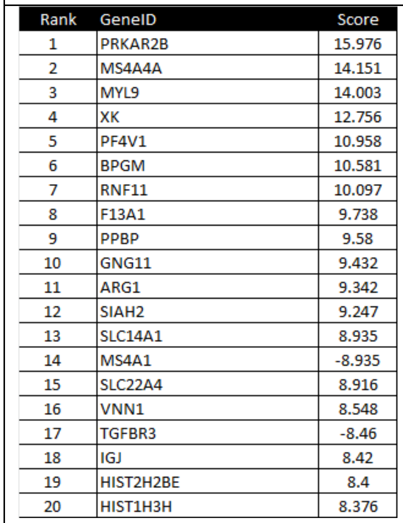
The caGEDA results page provides some valuable analysis metrics and plots. The between-mean array correlation, confounding index (after normalization), sample distribution box-whisker plots, global correlation graph of sample means, score histogram, and score frequency distribution (not necessary for this exercise), all help to ensure proper analysis procedure.

Matching script

Relevant information required for the modeling algorithm input file exists in different files. To combine all of the required information into one file, a Perl matching script named Matching_keep_all.pl was used. The script was provided by Haiwen Shi, Bioinformatics Core Labs, Genomic and Proteomic Core Labs, University of Pittsburgh. This Perl script takes two text files as input, and will match any common identifiers based on the user-defined columns of interest (Table 7). In this work, three lists (z-scores, feature selection list containing 1000 genes, and the normalized dataset), must be combined to obtain the necessary information, so the script was executed twice. The resulting file combines the necessary information from all three lists (Table 8).

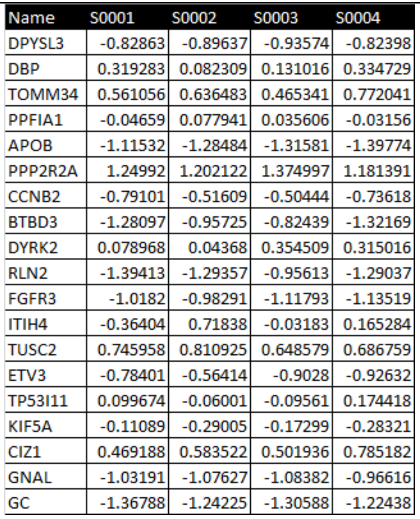
Table 7. Input files for matching script. The feature selection list and normalized dataset are output from caGEDA. The informed prior list comes from the text-mining exercise described earlier. Files were matched by using the GeneID/Name columns. Samples are depicted below.

Feature Selection List



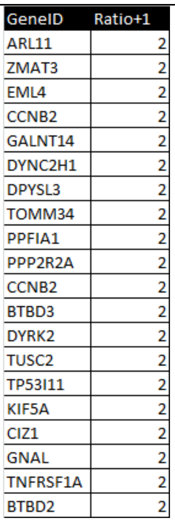
Rank	GeneID	Score
1	PRKAR2B	15.976
2	MS4A4A	14.151
3	MYL9	14.003
4	XK	12.756
5	PF4V1	10.958
6	BPGM	10.581
7	RNF11	10.097
8	F13A1	9.738
9	PPBP	9.58
10	GNG11	9.432
11	ARG1	9.342
12	SIAH2	9.247
13	SLC14A1	8.935
14	MS4A1	-8.935
15	SLC22A4	8.916
16	VNN1	8.548
17	TGFBR3	-8.46
18	IGJ	8.42
19	HIST2H2BE	8.4
20	HIST1H3H	8.376

Normalized Dataset



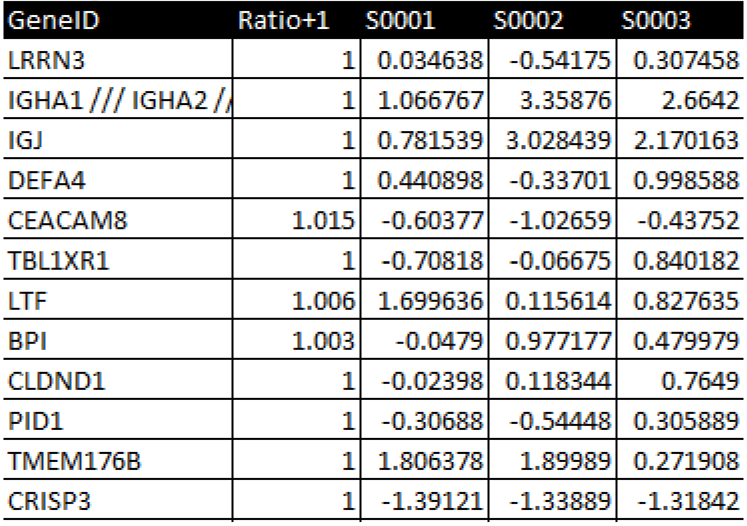
Name	S0001	S0002	S0003	S0004
DPYSL3	-0.82863	-0.89637	-0.93574	-0.82398
DBP	0.319283	0.082309	0.131016	0.334729
TOMM34	0.561056	0.636483	0.465341	0.772041
PPFIA1	-0.04659	0.077941	0.035606	-0.03156
APOB	-1.11532	-1.28484	-1.31581	-1.39774
PPP2R2A	1.24992	1.202122	1.374997	1.181391
CCNB2	-0.79101	-0.51609	-0.50444	-0.73618
BTBD3	-1.28097	-0.95725	-0.82439	-1.32169
DYRK2	0.078968	0.04368	0.354509	0.315016
RLN2	-1.39413	-1.29357	-0.95613	-1.29037
FGFR3	-1.0182	-0.98291	-1.11793	-1.13519
ITIH4	-0.36404	0.71838	-0.03183	0.165284
TUSC2	0.745958	0.810925	0.648579	0.686759
ETV3	-0.78401	-0.56414	-0.9028	-0.92632
TP53I11	0.099674	-0.06001	-0.09561	0.174418
KIF5A	-0.11089	-0.29005	-0.17299	-0.28321
CIZ1	0.469188	0.583522	0.501936	0.785182
GNAL	-1.03191	-1.07627	-1.08382	-0.96616
GC	-1.36788	-1.24225	-1.30588	-1.22438

Informed Prior List



GeneID	Ratio+1
ARL11	2
ZMAT3	2
EML4	2
CCNB2	2
GALNT14	2
DYNC2H1	2
DPYSL3	2
TOMM34	2
PPFIA1	2
PPP2R2A	2
CCNB2	2
BTBD3	2
DYRK2	2
TUSC2	2
TP53I11	2
KIF5A	2
CIZ1	2
GNAL	2
TNFRSF1A	2
BTBD2	2

Table 8. Matching script output file. Ratio+1 measure was used to eliminate any zeros from the dataset. A sample is depicted below.



GeneID	Ratio+1	S0001	S0002	S0003
LRRN3	1	0.034638	-0.54175	0.307458
IGHA1 /// IGHA2 //	1	1.066767	3.35876	2.6642
IGJ	1	0.781539	3.028439	2.170163
DEFA4	1	0.440898	-0.33701	0.998588
CEACAM8	1.015	-0.60377	-1.02659	-0.43752
TBL1XR1	1	-0.70818	-0.06675	0.840182
LTF	1.006	1.699636	0.115614	0.827635
BPI	1.003	-0.0479	0.977177	0.479979
CLDND1	1	-0.02398	0.118344	0.7649
PID1	1	-0.30688	-0.54448	0.305889
TMEM176B	1	1.806378	1.89989	0.271908
CRISP3	1	-1.39121	-1.33889	-1.31842

3.2.2.3 Transforming counts to prior probabilities Determining the most appropriate method to transform the literature mining counts into prior probabilities is essential for understanding the modeling results. Prior information values need to be within an acceptable range to appropriately be incorporated into the dataset without overwhelming the remaining data. In this work, three different transformation methods were tested: informed prior ratio (a ratio of number of positive biomarker mentions/number of negative biomarker mentions + 1 to eliminate zeros); uniform priors (value of 0.5 for all biomarkers); and no priors (data only).

3.2.2.4 Bayesian Rule Learner In the search for biomarkers, more accurate modeling should increase the chances of uncovering more-likely disease-specific markers. Using a Bayesian approach allows for prior information to be integrated into model learning. Additionally, rule learning is preferred because rules are easy to interpret and are easily applied. The Bayesian Rule Learner (BRL) algorithm is a probabilistic method for learning rules, and has been described in Gopalakrishnan *et al.* 2010. Models optimize a Bayesian score, which can be used to measure model uncertainty and rank and choose models; and ultimately translate the Bayesian network (BN) into a set of rules with scores.

The BRL utilizes the K2 Bayesian scoring measure and search heuristic (Cooper and Herskovits 1992). The K2 measure assumes discrete variables, independent cases, missing values, and a uniform prior probability distribution over all possible network structures. The K2 measure assumes every possible probability distribution over the values of a node given the state of its parents is equally likely. Under these assumptions, the Bayesian score is given by the following equation (Cooper and Herskovits 1992):

$$P(D|M) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i - 1)!}{\left(N_{ij} + r_i - 1\right)!}\prod_{k=1}^{r_i} N_{ijk}!$$

Equation 11. K2 Bayesian score

where *M* is the BN structure, *D* is the data used to learn *M*, *n* is the number of variables in *M*, q_i is the number of parent states of child variable *i*, r_i is the number of values or states of variable *i* and N_{ijk} is the number of instances in the training database *D* for which variable *i* has the value *k* and the parents of *i* have the value state denoted by index *j*. Also, N_{ij} is the sum over *k* of N_{ijk} (Cooper and Herskovits 1992).

The BRL process is described by Gopalakrishnan *et al.* 2010, with a summary given here: Bayesian networks containing a target node with zero parents are created and evaluated with Bayesian scores. Next, the list of variables in good scoring models which cannot be improved by adding a parent is initialized. A greedy search is implemented due to the difficulty of searching all possible BN structures. Models are searched by utilizing a beam to store the highest-scoring BNs. Beam size is user-defined, and the BNs are stored according to score. Because of beam size-restrictions, only the highest scoring BN structures which possess the ability to be improved upon by the addition of a parent variable are further examined. Additional searches are performed by adding one more variable as an additional parent of the target. For each target, its probability given each state of possible values of its parent variables is calculated. If the score of the model was improved with the addition of a new parent variable to the model structure, and the total number of parent variables in the model does not exceed the user-defined limit, additional searches are performed. If not, the model is placed on a priority queue containing final model structures ordered according to Bayesian scores.

$$P(D|M) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

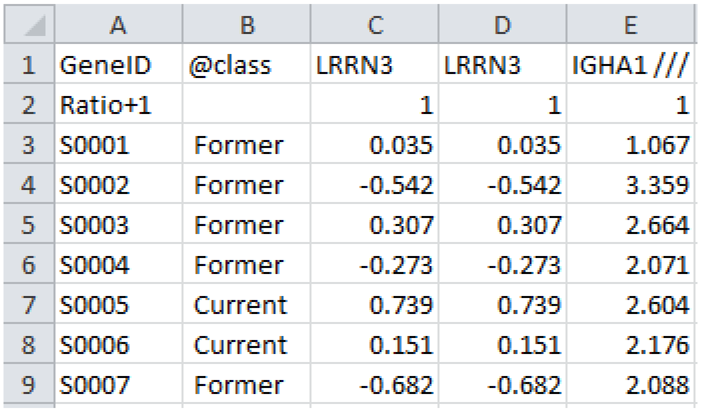
Equation 12. The BRL Score

The algorithm retains those sets of variables that cannot be improved upon further for reuse as parent variables. When no further improvements can be made to any model structure, the highest scoring models are returned to the user in the form of a rule models, which represent the probability that the model is valid given the data. The assumption is that if a predictor is found in a final rule, then it is unlikely to be a strong predictor in another rule.

The BRL was chosen as the modeling algorithm due to its many benefits also described in Gopalakrishnan *et al.* 2010: 1) evaluation of the entire rule set using a Bayesian score results in a whole model evaluation instead of a per rule evaluation; 2) creation of optimized probabilistic rules as opposed to the evaluation per rule; 3) incorporation of both structure and parameter priors; 4) prior knowledge with conditional independencies among variables can be applied, specifying the network structure; 5) returns parsimonious models with fewer variables or markers, without sacrificing classification performance; 6) fewer variables allows for less biological verification and validation; 7) more statistically significant results than other rule learning methods; 8) very efficient as it utilizes breadth-first marker propagation to only one pass through the training data once.; 9) ability to quantify uncertainty about the validity of a rule model using a Bayesian score; and 10) the use of prior probabilities into the rule-discovery process minimizes over-fitting.

3.2.2.5 Execution of the BRL algorithm BRL version 1 (2010-05-29) was used for the structure prior modeling exercises. The BRL is run as an executable jar file. The following arguments are given in the command line: –LP defines the learning parameters; –rgm 1 1 is the rule generation method, where the first 1 represents Bayesian local rule learning (local structure search), and the second 1 represents the decision tree parallel greedy search (PGS); –cv 10 represents 10-fold cross-validation; –d 0 0.5 specifies the discretization method, where 0 means no discretization, and the 0.5 is the default value for the structure prior parameter lambda; –beam 5000 specifies the size of the beam, or how many models can be retained at any one time; –PPP are pre-processing parameters; and –DP specifies data parameters used. So an example of the entire executable statement would be: java –jar BRLv1.jar –LP –rgm 1 1 –cv 10 –d 0 0.5 –beam 5000 –PPP –DP *filename*.

3.2.2.5.1 BRL input The BRL algorithm requires proper dataset formatting which was performed using Microsoft Excel 2010. The downloaded datasets must be transposed so that the rows represent samples and columns represent genes (Figure 14). The second column must be a sample class column. The second row must be the row of prior values if utilizing uniform or informed prior values. This row can be omitted if prior information is not desired. The file must be saved as a tab-delimited text file.



	A	B	C	D	E
1	GeneID	@class	LRRN3	LRRN3	IGHA1 ///
2	Ratio+1		1	1	1
3	S0001	Former	0.035	0.035	1.067
4	S0002	Former	-0.542	-0.542	3.359
5	S0003	Former	0.307	0.307	2.664
6	S0004	Former	-0.273	-0.273	2.071
7	S0005	Current	0.739	0.739	2.604
8	S0006	Current	0.151	0.151	2.176
9	S0007	Former	-0.682	-0.682	2.088

Figure 14. Example of a portion of a BRL input file. The @class column is the sample group identifier. In this example Former and Current represent smoking status, but any group identifier could be used.

3.2.2.5.2 BRL output BRL exports a number of informative files (cross-validation performance, cross-validation rules, and training data performance, prediction, and rules files). The performance and rules files were examined in this work. Performance files contain accuracy, sensitivity, specificity, balanced accuracy, and area under the curve measurements for the cross-validation and training data. The rules files contain the rules and attributes used to create the cross-validation and training models. An example of a rule is shown below. Rules were taken from a rules file obtained from a lung cancer microarray dataset of former vs. current smokers using informed priors: ((PCBP1 = 2.513..inf) (RTP4 = 1.045..inf) (TMEM161A = 1.054..1.364) (CBY1 = 0.661..inf) (SOD2 = 1.174..1.527) (AA654586 = 1.714..inf) (MFHAS1 = 0.627..inf) (WDR19 = 0.317..inf)) ==> (@class = Current). Samples whose data values fell within the described ranges for the 8 attributes (genes) in the rule were classified as Current smokers.

3.2.2.6 Confirmatory research The BRL modeling rules files contain the attributes (markers) used to create the best-scoring models. The attributes used to build the model were compared to the list of known disease biomarkers to look for markers common to both lists. When commonality was found, a potential relationship was assumed to exist (SOD2/CCL5). Further confirmatory research is then required to determine if the relationship is already known, novel, or a false-positive.

Pathway analysis was performed by examining KEGG, PID, and BioCarta pathway databases for SOD2/*superoxide dismutase 2* and CCL5/*C-C motif chemokine ligand 5* pathways. The protein interaction database String DB, string-db.org/, was searched for SOD2 and CCL5 protein interactions. Others examined include BioGRID, thebiogrid.org/, and IntAct, www.ebi.ac.uk/intact/main.xhtml. Entrez Gene, a genomic database www.ncbi.nlm.nih.gov/gene, was used to search the genes of interest. The information returned from this database includes gene summary, genomic context, genomic regions, transcripts, and products, pathways, ontology, and interactions. Lastly, a literature search of PubMed was performed. PubMed (www.ncbi.nlm.nih.gov/pubmed) contains more than 24 million citations for biomedical literature. PubMed can be searched using the marker IDs as keywords, and should return a somewhat comprehensive set of citations as results. Results can be further filtered by a number of parameters, one of which is species.

3.2.3 Pathway analysis methodology

Pathway analysis has become a standard step in biological data analysis. Some algorithms permit the user to input a set of relevant genes/proteins and their expression values, and output a diagram of the biological process where the input genes are highlighted. This allows the researcher to visualize what genes may be altered up- or down-stream for possible disease manipulation (prevention and/or treatment).

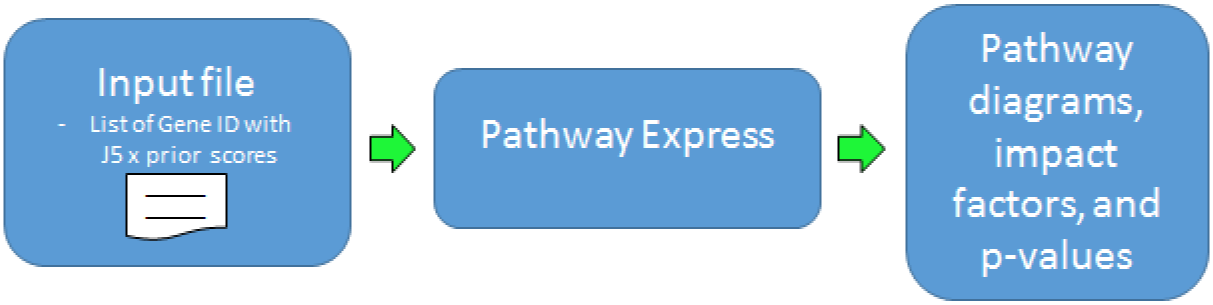

Input file
- List of Gene ID with J5 x prior scores
Pathway Express
Pathway diagrams, impact factors, and p-values

Figure 15. Diagram of the KEDA pathway analysis process.

Prior information can be incorporated with experimental data and analyzed for pathway analysis, increasing the likelihood that pathway findings and subsequent conclusions are more accurate. Figure 15 is a diagram of the pathway analysis process used in this work. The pathway analysis program utilized in this work was Pathway Express (PE).

3.2.3.1 Pathway Express Pathway Express (PE) is a freely available pathway analysis tool which is incorporated into the Onto-Tools suite. PE helps researchers find the most appropriate pathways for their genes of interest. PE takes as input a gene list with accompanying differentially expressed values, compares the list to existing pathways in the KEGG database, and outputs valuable pathway information such as impact factor, p-values, total number of input genes in the pathway, and pathway diagrams.

Khatri *et al*. 2005, briefly describe the impact factor calculation as follows: PE first calculates a perturbation factor (PF; Equation 2) for each input gene. The PF takes into account the normalized fold-change of the gene and the number and amount of perturbation of upstream and downstream genes. The PF reflects the relative importance of each differentially regulated gene. The impact factor (Equation 3) of the entire pathway includes a probabilistic term that takes into consideration the proportion of differentially regulated genes on the pathway and gene perturbation factors of all genes in the pathway. Pathways are ranked by impact factor before presentation to the user.

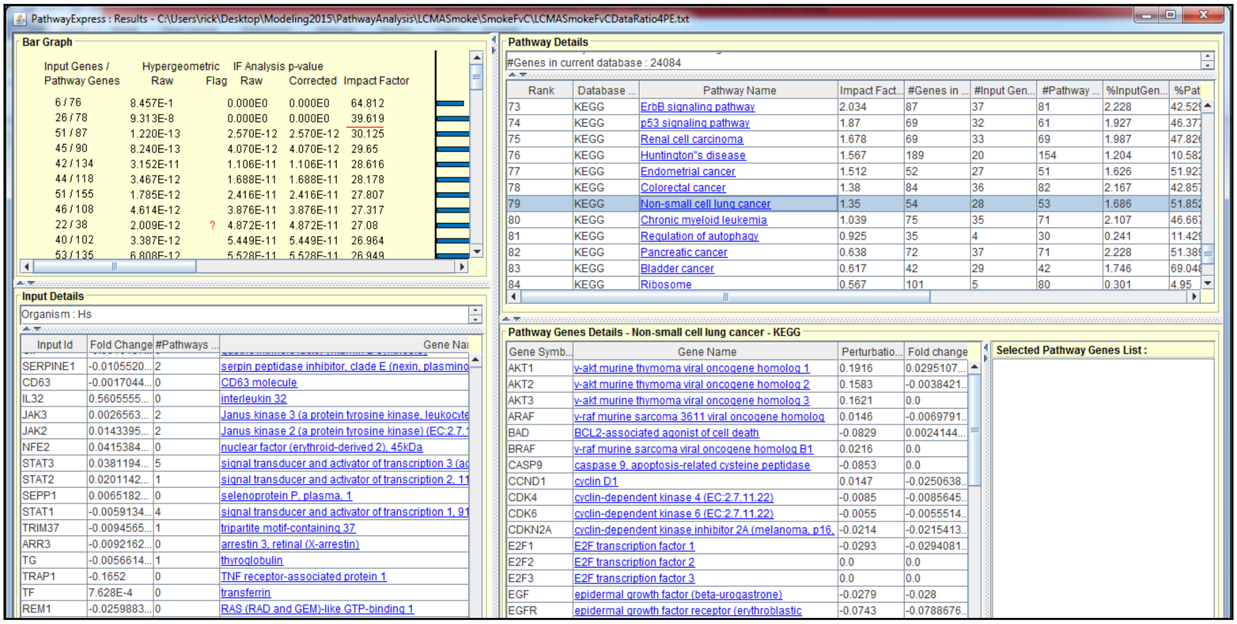
The PE p-value calculation is described in Khatri *et al.* 2007. PE performs a classical enrichment analysis based on a hypergeometric distribution in order to identify those pathways that contain a proportion of differentially expressed genes that is significantly different from what is expected by chance.

3.2.3.2 Execution of Pathway Express algorithm Pathway Express requires Java and is part of the Intelligent Systems and Bioinformatics Laboratory, found on the website http://vortex.cs.wayne.edu/projects.htm**.** Once the program is initiated, a menu will appear where the user chooses PE from Onto-Tools options menu. A subsequent PE input menu appears. From here the user enters the input file, reference file, reference array, organism, input type, and advanced options.

3.2.3.2.1 Pathway Express input Pathway Express allows the user to input a list of relevant genes/proteins for examination. The software assumes the entered values are fold change values, but any values may be used. In addition to genelists, a reference file is required, which is a list of all of the genes present on a given array. It uses this file for the perturbation and impact factor calculations described earlier.

Five different types of input genelists were studied to determine their effects on pathway results. Only a subset of the array-based experimental breast and lung cancer datasets mentioned earlier were examined. The first type of genelist contained differentially expressed genes and their accompanying J5 scores; which was considered data only. The second and third types of genelists contained genes with the z-scores and ratios obtained from the literature-mining exercises from whole blood, which were considered z-score only or ratio only. The fourth and fifth types of genelists used the product of the J5 score and z-score or ratios; considered data & z-score or data & ratio, respectively.

3.2.3.2.2 Pathway Express output Pathway Express outputs a menu of results as four windows (Figure 16). The upper left window orders affected pathways in decreasing order of their expected importance for the given condition (Khatri *et al*. 2005). The upper right window is the list of returned pathways, where the highlighted pathway is the active pathway. By clicking on the link, the actual pathway diagram is displayed, showing the input genes/proteins and their expression. The bottom right window displays the genes involved in the active pathway. The bottom left window displays the input genes/proteins, and the number of returned pathways that the genes/proteins are involved. Any particular window can be downloaded directly from the output menu.



PathwayExpress : Results - C:\Users\rick\Desktop\Modeling2015\PathwayAnalysis\LCMASmoke\SmokeFvC\LCMASmokeFvCDataRatio4PE.txt

Bar Graph

Input Genes / Pathway Genes	Hypergeometric Raw	Flag	IF Analysis p-value Raw	Corrected	Impact Factor
6 / 76	8.457E-1		0.000E0	0.000E0	64.812
26 / 78	9.313E-8		0.000E0	0.000E0	39.619
51 / 87	1.220E-13		2.570E-12	2.570E-12	30.125
45 / 90	8.240E-13		4.070E-12	4.070E-12	29.65
42 / 134	3.152E-11		1.106E-11	1.106E-11	28.616
44 / 118	3.467E-12		1.688E-11	1.688E-11	28.178
51 / 155	1.785E-12		2.416E-11	2.416E-11	27.807
46 / 108	4.614E-12		3.876E-11	3.876E-11	27.317
22 / 38	2.009E-12	?	4.872E-11	4.872E-11	27.08
40 / 102	3.387E-12		5.449E-11	5.449E-11	26.964
53 / 135	6.808E-12		5.528E-11	5.528E-11	26.949

Input Details

Organism : Hs

Input Id	Fold Change	#Pathways ...	Gene Nai
SERPINE1	-0.0105520...	2	serpin peptidase inhibitor, clade E (nexin, plasmino
CD63	-0.0017044...	0	CD63 molecule
IL32	0.5605555...	0	interleukin 32
JAK3	0.0026563...	2	Janus kinase 3 (a protein tyrosine kinase, leukocyte
JAK2	0.0143395...	2	Janus kinase 2 (a protein tyrosine kinase) (EC:2.7.1
NFE2	0.0415384...	0	nuclear factor (erythroid-derived 2), 45kDa
STAT3	0.0381194...	5	signal transducer and activator of transcription 3 (ac
STAT2	0.0201142...	1	signal transducer and activator of transcription 2, 11
SEPP1	0.0065182...	0	selenoprotein P, plasma, 1
STAT1	-0.0059134...	4	signal transducer and activator of transcription 1, 91
TRIM37	-0.0094565...	1	tripartite motif-containing 37
ARR3	-0.0092162...	0	arrestin 3, retinal (X-arrestin)
TG	-0.0056614...	1	thyroglobulin
TRAP1	-0.1652	0	TNF receptor-associated protein 1
TF	7.628E-4	0	transferrin
REM1	-0.0259883...	0	RAS (RAD and GEM)-like GTP-binding 1

Pathway Details

#Genes in current database : 24084

Rank	Database ...	Pathway Name	Impact Fact...	#Genes in ...	#Input Gen...	#Pathway ...	%InputGen...	%Pat
73	KEGG	ErbB signaling pathway	2.034	87	37	81	2.228	42.529
74	KEGG	p53 signaling pathway	1.87	69	32	61	1.927	46.377
75	KEGG	Renal cell carcinoma	1.678	69	33	69	1.987	47.826
76	KEGG	Huntington''s disease	1.567	189	20	154	1.204	10.582
77	KEGG	Endometrial cancer	1.512	52	27	51	1.626	51.923
78	KEGG	Colorectal cancer	1.38	84	36	82	2.167	42.857
79	KEGG	Non-small cell lung cancer	1.35	54	28	53	1.686	51.852
80	KEGG	Chronic myeloid leukemia	1.039	75	35	71	2.107	46.667
81	KEGG	Regulation of autophagy	0.925	35	4	30	0.241	11.429
82	KEGG	Pancreatic cancer	0.638	72	37	71	2.228	51.389
83	KEGG	Bladder cancer	0.617	42	29	42	1.746	69.048
84	KEGG	Ribosome	0.567	101	5	80	0.301	4.95

Pathway Genes Details - Non-small cell lung cancer - KEGG

Gene Symb...	Gene Name	Perturbatio...	Fold change
AKT1	v-akt murine thymoma viral oncogene homolog 1	0.1916	0.0295107...
AKT2	v-akt murine thymoma viral oncogene homolog 2	0.1583	-0.0038421...
AKT3	v-akt murine thymoma viral oncogene homolog 3	0.1621	0.0
ARAF	v-raf murine sarcoma 3611 viral oncogene homolog	0.0146	-0.0069791...
BAD	BCL2-associated agonist of cell death	-0.0829	0.0024144...
BRAF	v-raf murine sarcoma viral oncogene homolog B1	0.0216	0.0
CASP9	caspase 9, apoptosis-related cysteine peptidase	-0.0853	0.0
CCND1	cyclin D1	0.0147	-0.0250638...
CDK4	cyclin-dependent kinase 4 (EC:2.7.11.22)	-0.0085	-0.0085645...
CDK6	cyclin-dependent kinase 6 (EC:2.7.11.22)	-0.0055	-0.0055514...
CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16,	-0.0214	-0.0215413...
E2F1	E2F transcription factor 1	-0.0293	-0.0294081...
E2F2	E2F transcription factor 2	0.0	0.0
E2F3	E2F transcription factor 3	0.0	0.0
EGF	epidermal growth factor (beta-urogastrone)	-0.0279	-0.028
EGFR	epidermal growth factor receptor (erythroblastic	-0.0743	-0.0788676...

Selected Pathway Genes List :

Figure 16. Example of Pathway Express output menu. Upper left window orders affected pathways in decreasing order of their expected importance for the given condition. Upper right window is a list of relevant pathways returned. The highlighted pathway is the active pathway. Bottom left window is a list of the input genes/proteins and the number of returned pathways the gene/proteins are found. Bottom right window is a list of the genes found in active pathway.

Figure 17 is the KEGG pathway diagram for non-small cell lung cancer created using lung cancer microarray smoking former vs. current data + ratio dataset. The visual diagram allows researchers to view gene/protein interactions, as well as look for genes that may be possible targets for disease prevention and/or treatment. Genes/proteins from the user-provided genelist are highlighted to show their expression under given conditions.

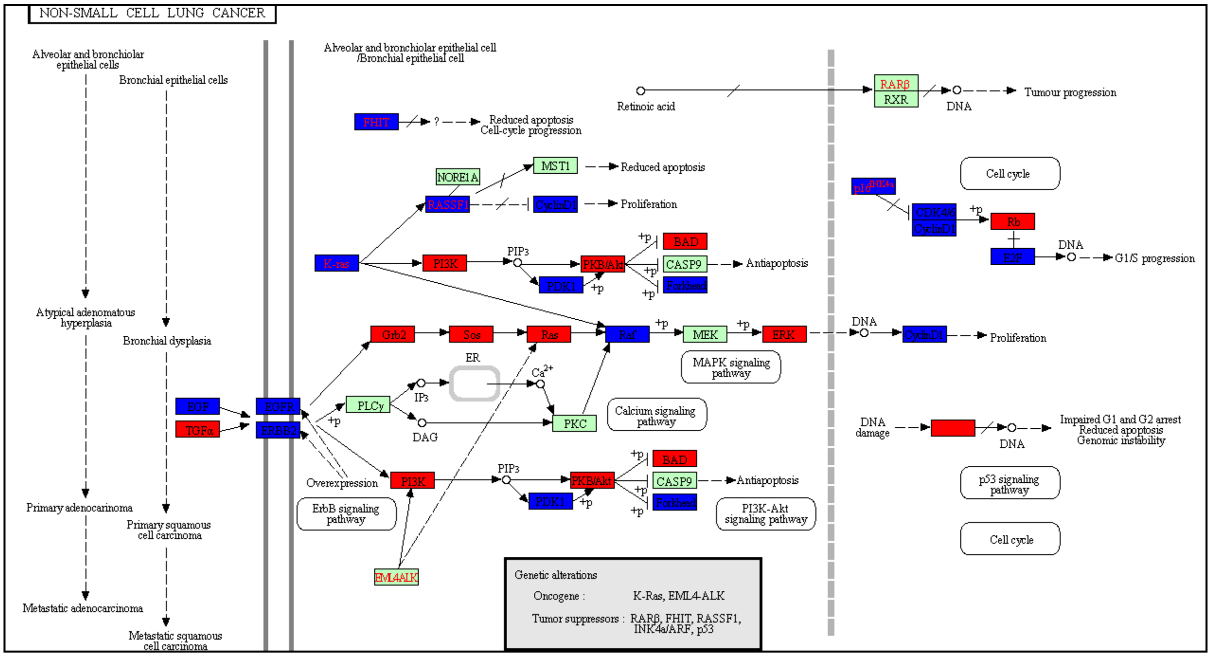


Figure 17. KEGG Non-small cell lung cancer diagram created using lung cancer microarray smoking former vs. current data + ratio dataset. Highlighted red = upregulated; blue = downregulated.

4.0 EVALUATION OF KEDA

The sections below present the experimentation undertaken to test the claims made above pertaining to literature-mining (4.1), modeling accuracy (4.2), and pathway analysis (4.3). The KEDA Framework was created to expedite the literature mining process to obtain values that can be used as prior knowledge in modeling and pathway analysis.

4.1 LITERATURE-MINING RESULTS

The goal of literature mining was to obtain prior knowledge values to aid in biomarker discovery via modeling and pathway exercises, while trying to acquire added new knowledge, in the form of disease and biofluid specific findings, in the process. Additionally, literature mining findings could also be utilized to identify potential biomarkers as a stand-alone process.

Biofluid-specific markers were identified from mining the literature, assigned relevance scores by frequency of occurrence, and validated using known biomarker lists and/or databases for lung and breast cancer. Biofluid specificity for each marker was calculated, and the performance of the semi-automated literature mining method assessed. In the following sections (4.1.1 - 4.1.9) the claim that text-mining is a sufficient method of obtaining potential biomarkers is tested (Claim 1).

4.1.1 Z- score threshold optimization

Gene/protein IDs in PubMed abstracts were identified and frequency of occurrence counts were converted to z-scores. Z-scores were selected as a measure because they provide more information that counts alone. Z-scores are a relative measure that provide a general idea about the number of standard deviations a data point varies from its mean; counts do not. A threshold was sought to determine the point at which z-scores would be considered significant. Because of the large number of markers found, only markers considered ‘significant’ were further pursued. Empirical findings were used to establish said threshold. Figure 18 is a plot of the number of known and new markers found by varying the z-score threshold in increments of 0.5. It can be seen that in both breast and lung cancer that a threshold value of 1.0 allows for the maximum number of new and known significant markers.

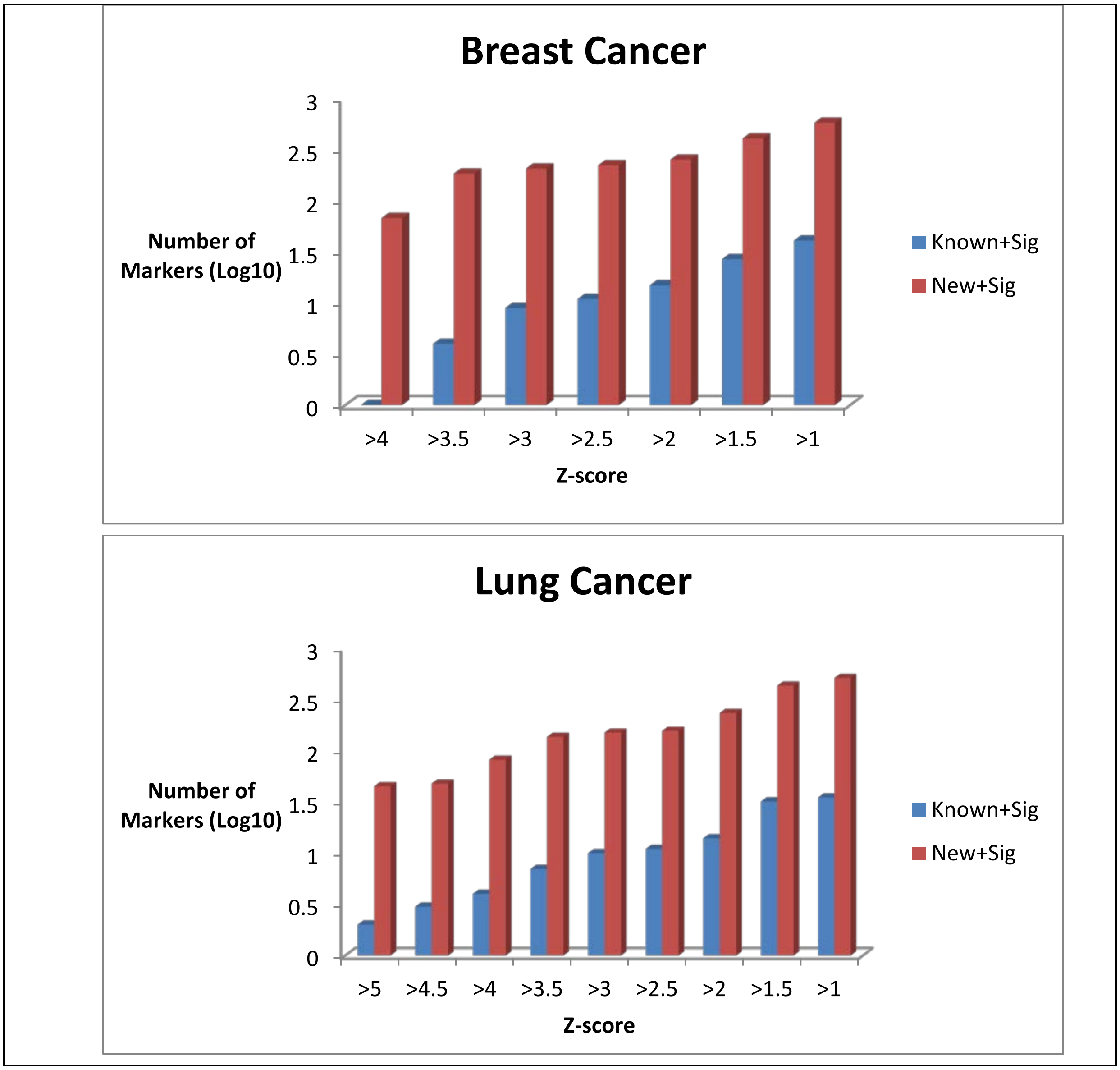


Figure 18. Number of markers identified across the range of possible Z-scores. Decreasing the Z-score threshold allows for more significant markers to be identified. Sig = significant.

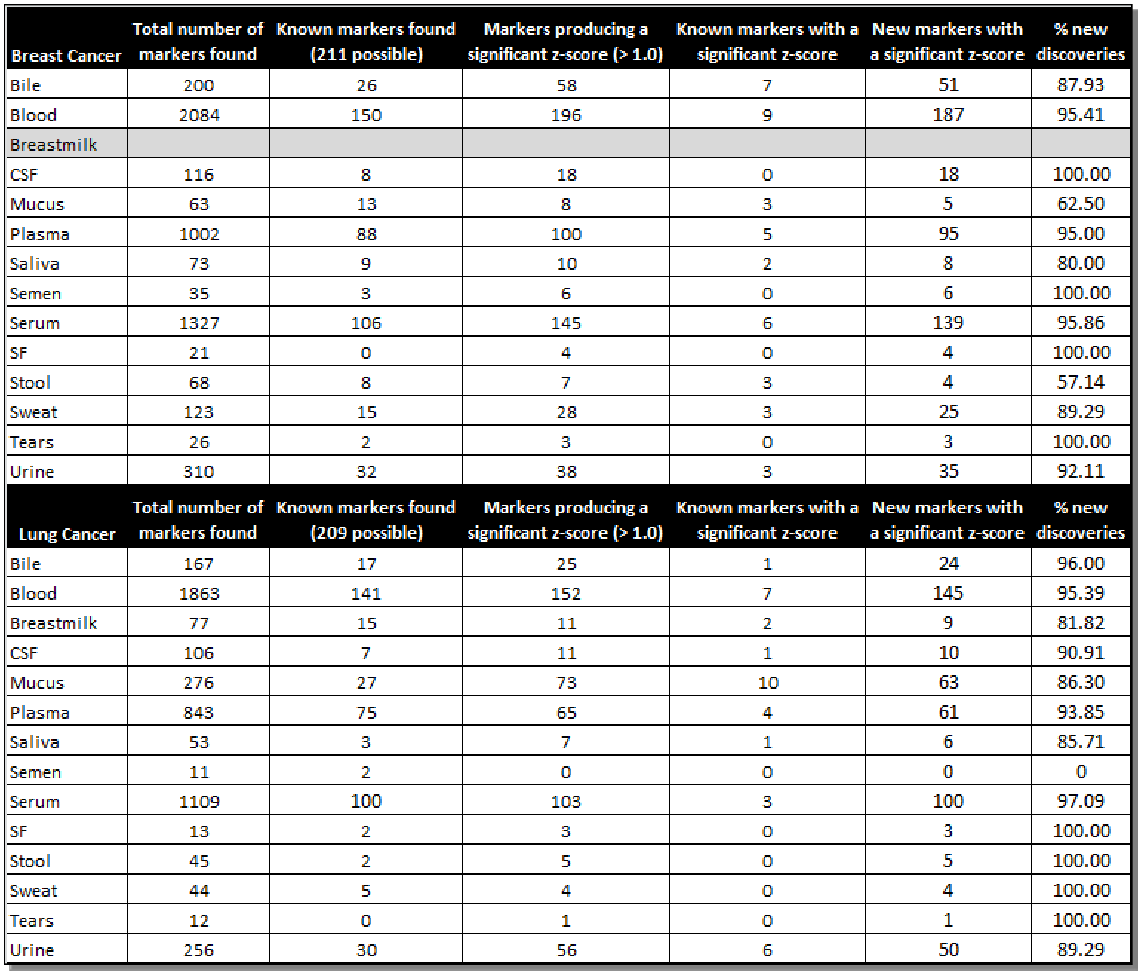
4.1.2 Known markers per biofluid

To estimate the performance of our semi-automated literature mining process, an error rate calculation was performed. By estimating the error rate in known biomarkers, it might be possible to infer the error rate to newly discovered unknown potential biomarkers. Known biomarker lists are combinations of several lists from well-known disease databases. The known breast cancer list contains 211 markers that mapped to our dictionary (Table 4; n=159), and the known lung cancer list has 209 markers that mapped to our dictionary (Table 5; n=145). Results presented in Table 9 were achieved by identifying putative biomarkers with a z-score exceeding the significance threshold (>1.0), and confirming the gene symbol's existence in a known disease biomarker list. Table 9 also provides a summary of each biofluid, markers with significant z-scores, the number of known markers found, and the calculated percent of new discoveries. Breastmilk was removed from breast cancer examination because the positive and negative search terms both contain the root word 'breast'.

4.1.3 Known markers found significant vs. non-significant

The next question to be asked was: Out of the total known markers that were identified by our methods, what percentage were being identified as significant by the proposed scoring method? By calculating this percentage using the counts provided in Table 9, it could be determined if the scoring threshold was too stringent or too lenient. For breast cancer, known/significant percentages ranged from 5% in plasma and serum to 37.5% in stool (for biofluids with known-significant markers; non-zero). In lung cancer the known/significant percentages ranged from 3% in serum to 37% in mucus. Based on these percentages, it was determined that the threshold was not too stringent because the it was not eliminating all of the found known markers, and it also was not too lenient in that it was reducing the number of markers for further study by more than half.

Table 9. Number of markers identified for disease-biofluid combinations. Known markers were confirmed by the presence of the gene symbol in our known biomarker lists (Tables 4 and 5). Significant markers have a z-score > 1.0.



Breast Cancer	Total number of markers found	Known markers found (211 possible)	Markers producing a significant z-score (> 1.0)	Known markers with a significant z-score	New markers with a significant z-score	% new discoveries
Bile	200	26	58	7	51	87.93
Blood	2084	150	196	9	187	95.41
Breastmilk						
CSF	116	8	18	0	18	100.00
Mucus	63	13	8	3	5	62.50
Plasma	1002	88	100	5	95	95.00
Saliva	73	9	10	2	8	80.00
Semen	35	3	6	0	6	100.00
Serum	1327	106	145	6	139	95.86
SF	21	0	4	0	4	100.00
Stool	68	8	7	3	4	57.14
Sweat	123	15	28	3	25	89.29
Tears	26	2	3	0	3	100.00
Urine	310	32	38	3	35	92.11
Lung Cancer	**Total number of markers found**	**Known markers found (209 possible)**	**Markers producing a significant z-score (> 1.0)**	**Known markers with a significant z-score**	**New markers with a significant z-score**	**% new discoveries**
Bile	167	17	25	1	24	96.00
Blood	1863	141	152	7	145	95.39
Breastmilk	77	15	11	2	9	81.82
CSF	106	7	11	1	10	90.91
Mucus	276	27	73	10	63	86.30
Plasma	843	75	65	4	61	93.85
Saliva	53	3	7	1	6	85.71
Semen	11	2	0	0	0	0
Serum	1109	100	103	3	100	97.09
SF	13	2	3	0	3	100.00
Stool	45	2	5	0	5	100.00
Sweat	44	5	4	0	4	100.00
Tears	12	0	1	0	1	100.00
Urine	256	30	56	6	50	89.29

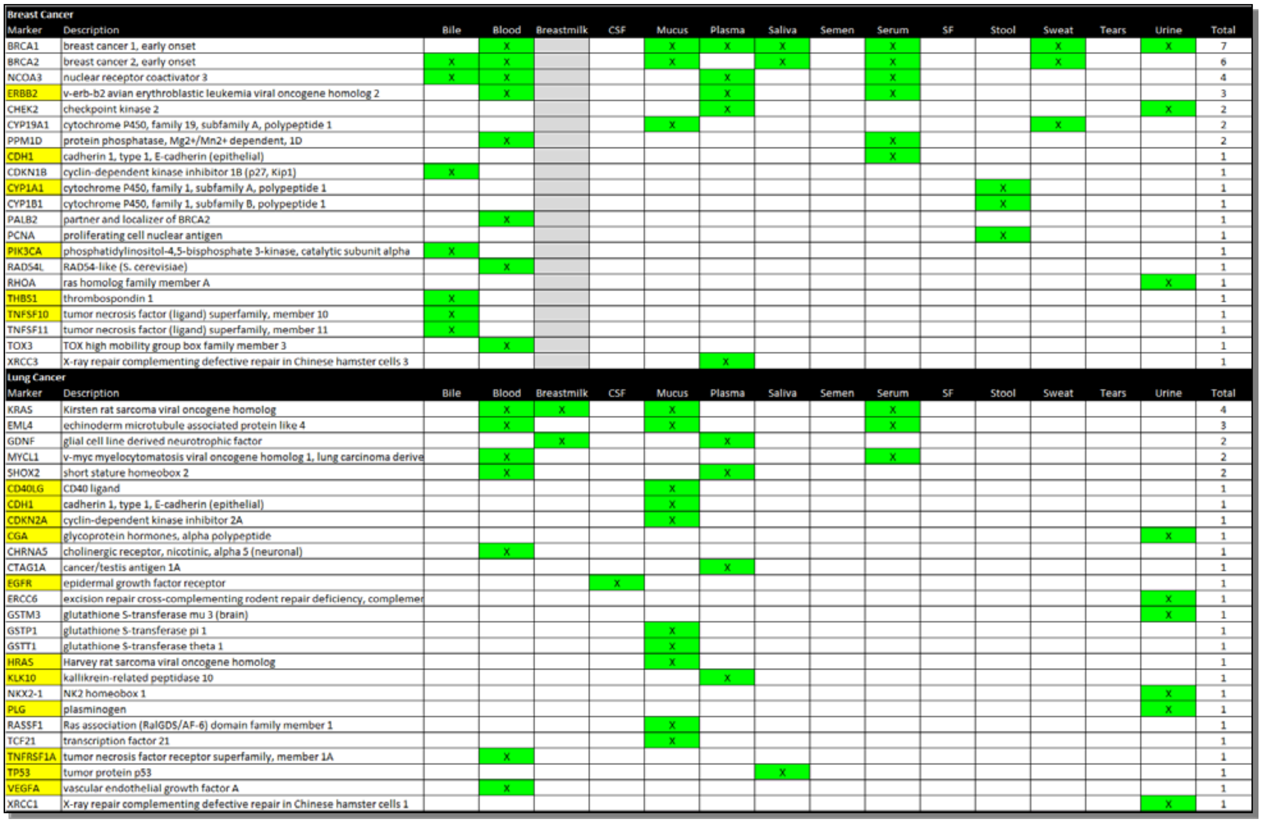
4.1.4 Newly discovered markers found significant vs. non-significant

The percentages of newly discovered markers (markers not found in known marker list) found to be significant vs. those that were identified but not found to be significant was calculated to determine if the error rate calculation for known markers could be extrapolated to apply to newly discovered unknown potential markers. For breast cancer, new/significant marker percentages ranged from 6.67% in stool to 29.3% in bile (for biofluids with known-significant markers; non-zero), and in lung cancer the new/significant percentages ranged from 7.9% in plasma to 27.2% in synovial fluid. The newly discovered/significant percentage ranges highly correlate with the known/significant percentage ranges. Based on this result, it was concluded that the error rate from known findings could be inferred to new findings.

4.1.5 Potential marker biofluid specificity

The search for additional information in the literature mining process led to breast and lung cancer findings being further subdivided into biofluids. Biomarker commonality and specificity was determined across biofluids. This is a significant finding as this information is novel as potential biomarker comparisons across more than a few biofluids are rarely seen in the scientific literature. This information could also prove very beneficial to breast and lung cancer researchers and clinicians in the future. Table 10 shows the known and significant biomarkers found within biofluid type for breast and lung cancer.

Table 10. Identification of the significant validated potential markers found to be in common to several biofluids or biofluid specific for breast and lung cancer. Yellow highlights are breast cancer markers found in the list of validated lung cancer biomarkers (Table 4), or lung cancer markers found in the list of validated breast cancer biomarkers (Table 5). It is doubtful that these markers are disease specific. Green highlights aid in identifying which markers were identified per biofluid. CDH1 is the only biomarker appearing in both breast and lung cancer lists. CSF = cerebrospinal fluid; SF = synovial fluid.



Breast Cancer																
Marker	Description	Bile	Blood	Breastmilk	CSF	Mucus	Plasma	Saliva	Semen	Serum	SF	Stool	Sweat	Tears	Urine	Total
BRCA1	breast cancer 1, early onset		X			X	X	X		X			X		X	7
BRCA2	breast cancer 2, early onset	X	X			X		X		X			X			6
NCOA3	nuclear receptor coactivator 3	X	X				X			X						4
ERBB2	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2		X				X			X						3
CHEK2	checkpoint kinase 2						X								X	2
CYP19A1	cytochrome P450, family 19, subfamily A, polypeptide 1					X							X			2
PPM1D	protein phosphatase, Mg2+/Mn2+ dependent, 1D		X							X						2
CDH1	cadherin 1, type 1, E-cadherin (epithelial)									X						1
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	X														1
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1											X				1
CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1											X				1
PALB2	partner and localizer of BRCA2		X													1
PCNA	proliferating cell nuclear antigen											X				1
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	X														1
RAD54L	RAD54-like (S. cerevisiae)		X													1
RHOA	ras homolog family member A														X	1
THBS1	thrombospondin 1	X														1
TNFSF10	tumor necrosis factor (ligand) superfamily, member 10	X														1
TNFSF11	tumor necrosis factor (ligand) superfamily, member 11	X														1
TOX3	TOX high mobility group box family member 3		X													1
XRCC3	X-ray repair complementing defective repair in Chinese hamster cells 3						X									1
Lung Cancer																
Marker	Description	Bile	Blood	Breastmilk	CSF	Mucus	Plasma	Saliva	Semen	Serum	SF	Stool	Sweat	Tears	Urine	Total
KRAS	Kirsten rat sarcoma viral oncogene homolog		X	X		X				X						4
EML4	echinoderm microtubule associated protein like 4		X			X				X						3
GDNF	glial cell line derived neurotrophic factor			X			X									2
MYCL1	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derive		X							X						2
SHOX2	short stature homeobox 2		X				X									2
CD40LG	CD40 ligand					X										1
CDH1	cadherin 1, type 1, E-cadherin (epithelial)					X										1
CDKN2A	cyclin-dependent kinase inhibitor 2A					X										1
CGA	glycoprotein hormones, alpha polypeptide														X	1
CHRNA5	cholinergic receptor, nicotinic, alpha 5 (neuronal)		X													1
CTAG1A	cancer/testis antigen 1A						X									1
EGFR	epidermal growth factor receptor				X											1
ERCC6	excision repair cross-complementing rodent repair deficiency, complemer														X	1
GSTM3	glutathione S-transferase mu 3 (brain)														X	1
GSTP1	glutathione S-transferase pi 1					X										1
GSTT1	glutathione S-transferase theta 1					X										1
HRAS	Harvey rat sarcoma viral oncogene homolog					X										1
KLK10	kallikrein-related peptidase 10						X									1
NKX2-1	NK2 homeobox 1														X	1
PLG	plasminogen														X	1
RASSF1	Ras association (RalGDS/AF-6) domain family member 1					X										1
TCF21	transcription factor 21					X										1
TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A		X													1
TP53	tumor protein p53							X								1
VEGFA	vascular endothelial growth factor A		X													1
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1														X	1

From Table 10, for breast cancer, nine biofluids produced known markers with significant scores. 21 known & significant putative markers were identified. 14 of these markers are only mentioned in combination with one biofluid, 3 with two biofluids, 1 with 3 biofluids (ERBB2; mentioned blood, plasma, and serum), 1 with 4 biofluids (NCOA3/*nuclear receptor coactivator 3*; mentioned in bile, blood, plasma, and serum), 1 with 6 biofluids (BRCA2; mentioned in bile, blood, mucus, saliva, serum, and sweat), and 1 with 7 biofluids (BRCA1; mentioned in blood, mucus, plasma, saliva, serum, sweat, and urine).

Also from Table 10, for lung cancer, eight biofluids produced known markers with significant scores. 26 known & significant putative markers were identified. 21 of these markers are only mentioned in combination with one biofluid, 3 with two biofluids, 1 with 3 biofluids (EML4/*echinoderm microtubule-associated protein-like 4*; mentioned in blood, mucus, and serum), and 1 with 4 biofluids (KRAS; mentioned in blood, breastmilk, mucus, and serum).

As we are interested in identifying disease-specific markers, it was important to ensure that the markers on the list were not common cancer markers, but breast or lung cancer specific. To do this, markers in the breast cancer list of Table 10 were compared to the list of validated lung cancer biomarkers (Table 5), and lung cancer markers found in the list of Table 10 were compared to the list of validated breast cancer biomarkers (Table 4). In doing so, six breast cancer markers (ERBB2, CDH1/*cadherin 1*, CYP1A1/*cytochrome P450 family 1 subfamily A member 1*, PIK3CA, THBS1/*thrombospondin 1*, and TNFSF10/*tumor necrosis factor superfamily member 10*) and eleven lung cancer markers (CD40LG/*CD40 ligand*, CDH1, CDKN2A/*cyclin-dependent kinase inhibitor 2A*, CGA/*glycoprotein hormones, alpha polypeptide*, EGFR, HRAS/*Harvey rat sarcoma viral oncogene homolog,* KLK10/*kallikrein-related peptidase 10*, PLG/*plasminogen*, TNFRSF1A/*tumor necrosis factor receptor superfamily member 1A*, TP53, and VEGFA/*vascular endothelial growth factor A*) were determined to not be either breast or lung cancer specific markers. However, even though they are not breast or lung cancer specific, most of them have been implicated in cancer biology in general, and should not be discarded from any cancer study.

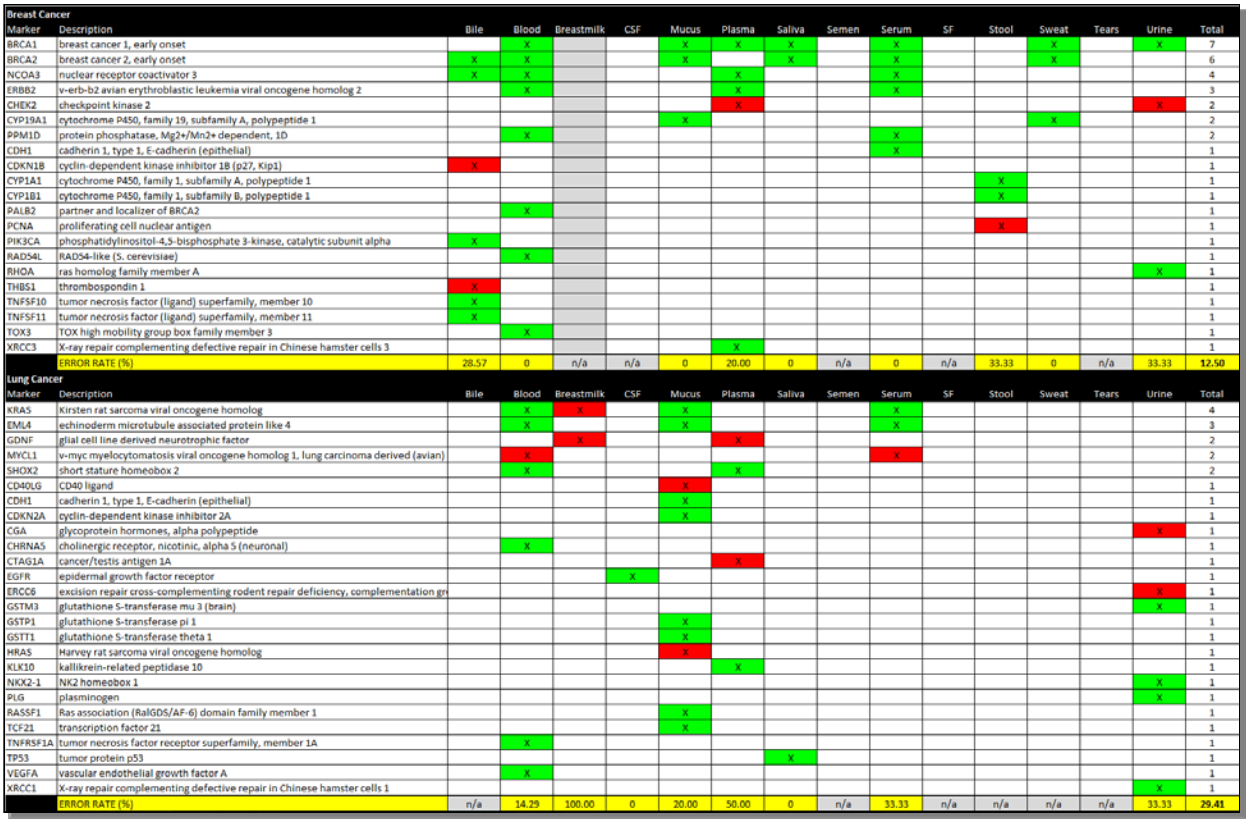
4.1.6 Manual verification of findings

Manual inspection of pertinent abstracts was performed to determine the reliability of the findings in Table 10. For each relevant finding in Table 10, the supporting PubMed abstracts were manually examined to verify that each abstract contained mention of the marker, biofluid, and disease. Only abstracts that mentioned all three entities were considered true positives in this study. The results can be seen in Table 11. In breast cancer, four known biomarkers (CHEK2/*checkpoint kinase 2* in both plasma and urine, CDKN1B/*cyclin-dependent kinase inhibitor 1B*, PCNA/*proliferating cell nuclear antigen*, and THBS1/*thrombospondin 1*) were identified as false positives (red); and in lung cancer, eight known biomarkers were identified as false positives (KRAS, GDNF/*glial cell line-derived neurotrophic factor* in both breastmilk and plasma, MYCL1/*v-myc avian myelocytomatosis viral oncogene lung carcinoma-derived homolog* in both blood and serum, CD40LG, CGA, CTAG1A/*cancer-testis antigen 1A*, ERCC6/*excision repair cross-complementation group 6,* and HRAS). KRAS is interesting in that it produced a false positive in association with breastmilk, but had verified positive findings in associations with blood, mucus, and serum.

4.1.7 Error rate estimation of new discoveries

The manual verification step enabled calculation of the error rates across the biofluid-disease combinations. Table 11 displays an average error rate for breast cancer of 12.5%, and an average error rate for lung cancer of 29.41%. Based on these error rates, it is estimated that 87.5% of the breast cancer new discoveries, and 70.59% of the lung cancer new discoveries from the proposed method could be trusted to be true positives.

Table 11. Manually verified biomarker table. Biomarker specific abstracts were manually examined for mentions of biofluid, disease, and biomarker. Omission of any of these terms resulted in a 'fail' or 'false positive' result. CSF = cerebrospinal fluid; SF = synovial fluid.



Breast Cancer

Marker	Description	Bile	Blood	Breastmilk	CSF	Mucus	Plasma	Saliva	Semen	Serum	SF	Stool	Sweat	Tears	Urine	Total
BRCA1	breast cancer 1, early onset		🟩 x			🟩 x	🟩 x	🟩 x		🟩 x			🟩 x		🟩 x	7
BRCA2	breast cancer 2, early onset	🟩 x	🟩 x			🟩 x		🟩 x		🟩 x			🟩 x			6
NCOA3	nuclear receptor coactivator 3	🟩 x	🟩 x				🟩 x			🟩 x						4
ERBB2	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2		🟩 x				🟩 x			🟩 x						3
CHEK2	checkpoint kinase 2						🟥 x								🟥 x	2
CYP19A1	cytochrome P450, family 19, subfamily A, polypeptide 1					🟩 x							🟩 x			2
PPM1D	protein phosphatase, Mg2+/Mn2+ dependent, 1D		🟩 x							🟩 x						2
CDH1	cadherin 1, type 1, E-cadherin (epithelial)									🟩 x						1
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	🟥 x														1
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1											🟩 x				1
CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1											🟩 x				1
PALB2	partner and localizer of BRCA2		🟩 x													1
PCNA	proliferating cell nuclear antigen											🟥 x				1
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	🟩 x														1
RAD54L	RAD54-like (S. cerevisiae)		🟩 x													1
RHOA	ras homolog family member A														🟩 x	1
THBS1	thrombospondin 1	🟥 x														1
TNFSF10	tumor necrosis factor (ligand) superfamily, member 10	🟩 x														1
TNFSF11	tumor necrosis factor (ligand) superfamily, member 11	🟩 x														1
TOX3	TOX high mobility group box family member 3		🟩 x													1
XRCC3	X-ray repair complementing defective repair in Chinese hamster cells 3						🟩 x									1
	ERROR RATE (%)	28.57	0	n/a	n/a	0	20.00	0	n/a	0	n/a	33.33	0	n/a	33.33	12.50

Lung Cancer

Marker	Description	Bile	Blood	Breastmilk	CSF	Mucus	Plasma	Saliva	Semen	Serum	SF	Stool	Sweat	Tears	Urine	Total
KRAS	Kirsten rat sarcoma viral oncogene homolog		🟩 x	🟥 x		🟩 x				🟩 x						4
EML4	echinoderm microtubule associated protein like 4		🟩 x			🟩 x				🟩 x						3
GDNF	glial cell line derived neurotrophic factor			🟥 x			🟥 x									2
MYCL1	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived (avian)		🟥 x							🟥 x						2
SHOX2	short stature homeobox 2		🟩 x				🟩 x									2
CD40LG	CD40 ligand					🟥 x										1
CDH1	cadherin 1, type 1, E-cadherin (epithelial)					🟩 x										1
CDKN2A	cyclin-dependent kinase inhibitor 2A					🟩 x										1
CGA	glycoprotein hormones, alpha polypeptide														🟥 x	1
CHRNA5	cholinergic receptor, nicotinic, alpha 5 (neuronal)		🟩 x													1
CTAG1A	cancer/testis antigen 1A						🟥 x									1
EGFR	epidermal growth factor receptor				🟩 x											1
ERCC6	excision repair cross-complementing rodent repair deficiency, complementation gr														🟥 x	1
GSTM3	glutathione S-transferase mu 3 (brain)														🟩 x	1
GSTP1	glutathione S-transferase pi 1					🟩 x										1
GSTT1	glutathione S-transferase theta 1					🟩 x										1
HRAS	Harvey rat sarcoma viral oncogene homolog					🟥 x										1
KLK10	kallikrein-related peptidase 10						🟩 x									1
NKX2-1	NK2 homeobox 1														🟩 x	1
PLG	plasminogen														🟩 x	1
RASSF1	Ras association (RalGDS/AF-6) domain family member 1					🟩 x										1
TCF21	transcription factor 21					🟩 x										1
TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A		🟩 x													1
TP53	tumor protein p53							🟩 x								1
VEGFA	vascular endothelial growth factor A		🟩 x													1
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1														🟩 x	1
	ERROR RATE (%)	n/a	14.29	100.00	0	20.00	50.00	0	n/a	33.33	n/a	n/a	n/a	n/a	33.33	29.41

The following factors support the idea that (Claim 1) literature mining is a sufficient method of obtaining potential biomarkers: 1) the search space was exhaustive, ensuring that all relevant abstracts were included in the analysis (methodology 3.2.1); 2) a gene/protein dictionary was implemented, to allow gene/protein aliases to also be included in the counts (methodology 3.2.4); 3) use of positive/negative abstract sets ensured that the findings were disease specific (methodology 3.2.4); 4) known biomarker lists were utilized as gold-standards, supporting the idea that the findings were true (methodology 3.2.6; experimentation 4.1.1 - 4.1.3); 5) manual verification allowed for further confirmation of true findings (methodology 3.2.6; experimentation 4.1.6); 6) an error rate was calculated to determine the number of newly discovered markers that can be trusted to be true positive findings (methodology 3.2.7; experimentation 4.1.7).

4.2 CLASSIFICATION MODELING RESULTS

The goal of modeling is to predict correct classification of samples into groups based on the examination of specific criteria. Several different evaluation measures were assessed in order to determine the performance of the models: accuracy, number of attributes, informativeness of the attributes, sensitivity, and specificity all contributed to the understanding and assessment of the results. The BRL determines the best performing models and presents only the best model(s) to the user. Several factors exist (dataset size, weighting of priors, and data type) which could bias results and need to be accounted for before proper assessment can be performed.

The following sections provide explanations of the comparisons and exercises performed. Sections 4.2.1 – 4.2.4 test the idea that incorporation of prior information did not on average, enhance or degrade the model performance (Claim 2, part A). Section 4.2.5 – 4.2.6 examine the claim that analyzing the attributes used to build the best-performing models may lead to possible new interactions (Claim 2, part B).

4.2.1 Experimental design using literature mining results

For the modeling exercises, the prior probabilities were incorporated into the normalized dataset. Figure 14 in section 3.2.2.5 provides an example of the BRL input file. The priors were incorporated as a column in-between the sample class column and the data columns. Each comparison required a separate dataset / prior combination. The customized BRL algorithm tested the prior column as its own variable, and then multiplied each data point with its applicable prior probability for all other variables (in this case gene/protein).

4.2.2 Dataset size effects on accuracy

An initial experiment was performed to determine if an optimal sized dataset exists and if so, would it influence the modeling results. The lung cancer microarray dataset case-control comparison (cancer vs. normal) was chosen for this exercise. The sample size consisted of 73 adenocarcinoma samples vs. 80 control samples. 22277 probes covering 14355 genes were analyzed. The dataset was Z-transformed, and feature selection was performed prior to analysis.

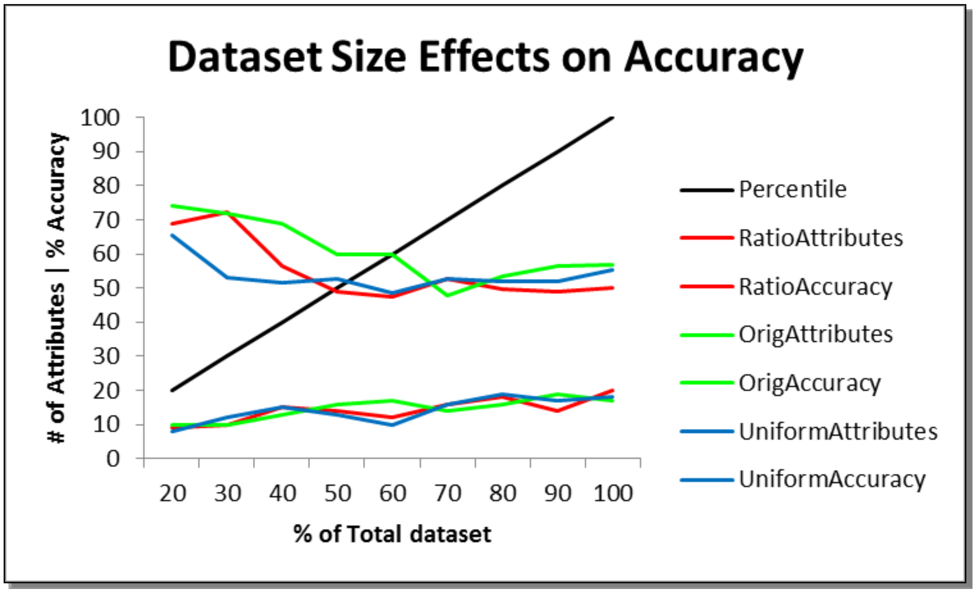


Figure 19. Overall modeling accuracy and the number of attributes found using different dataset sizes. Accuracy measurements can be found in the 40-80% range, while the number of attributes can be found in the 0-20 range. Results were obtained from BRL. Ratio = informed priors (red); Orig = no priors (green); Uniform = uniform priors (blue).

Smaller sized sub-datasets were manually created by randomly choosing samples to create datasets of pre-determined percentages of the original microarray dataset, ranging from

20% to 100%, in increments of 10%. For example, the 20% dataset would consist of 14 adenocarcinoma samples (out of 73 total) and 16 control samples (out of 80 total) randomly chosen. The BRL was executed with the previously discussed settings. Informed priors (Ratio), no priors (Orig), and uniform priors were tested and examined for classification accuracy. Figure 20 is a combined plot of the accuracy of the models tested (top group of curves), as well as the number of attributes tested (bottom group of curves), across the different sub-datasets. No consistent pattern of improvement in modeling accuracy or number of attributes is seen across the different dataset sizes. The greatest accuracy for the informed prior is obtained from the 30% dataset.

4.2.3 Weighting effects on accuracy

An experiment was performed to determine how weighting the informed prior ratio would influence the BRL modeling results. If the weights are too large, the prior values could overwhelm the data and models could be chosen based solely on the prior values, ignoring the data. Conversely, if the informed prior weights are too small, they might be overwhelmed by the data, and thus not contribute to the model either.

The same lung cancer microarray dataset case-control comparison (cancer vs. normal) was chosen for this exercise. However, the 30% dataset was used, being that the greatest informed prior accuracy was achieved with that dataset. The dataset was z-transformed, and feature selection was performed prior to analysis.

Informed prior ratios were weighted on an increasing scale beginning with no weight, and adding values of 2, 3, 4, 5, 10, 25, 50, and 100. Uniform priors were scaled beginning with 0.5, 1, 1.5, 2, 2.5, 5, 12.5, 25, and 50.

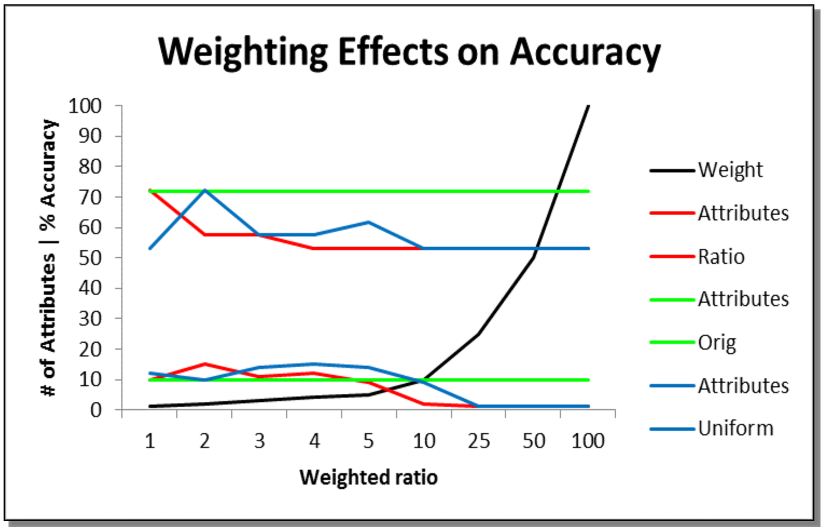


Figure 20. Overall modeling accuracy across various weights. Accuracy measurements are found in the 50-80% range, while the number of attributes can be found in the 0-20 range. Results were obtained from BRL. Ratio = informed priors (red); Orig = no priors (green); Uniform = uniform priors (blue).

The BRL was executed with the previously described settings. Figure 21, is a combined plot of the accuracy of the models tested (top group of curves), as well as the number of attributes tested (bottom group of curves), with different weights. As before, no consistent pattern of improvement in modeling accuracy or number of attributes is seen using different weights. The greatest accuracy for the informed prior is obtained using a weight of 1 (weighted ratio = 1 in Figure 21).

4.2.4 Accuracy across different types of datasets

The possibility exists that the results obtained up to this point may be data-type specific, and that the results could change drastically if a different data-type is examined. To confirm that previous findings were not data-type specific, an experiment was performed to determine modeling accuracy across different data-types or platforms. Eight comparisons and four different data-types were examined. In this experiment, the entire datasets were used. The dataset was z-transformed, and feature selection was performed prior to analysis.

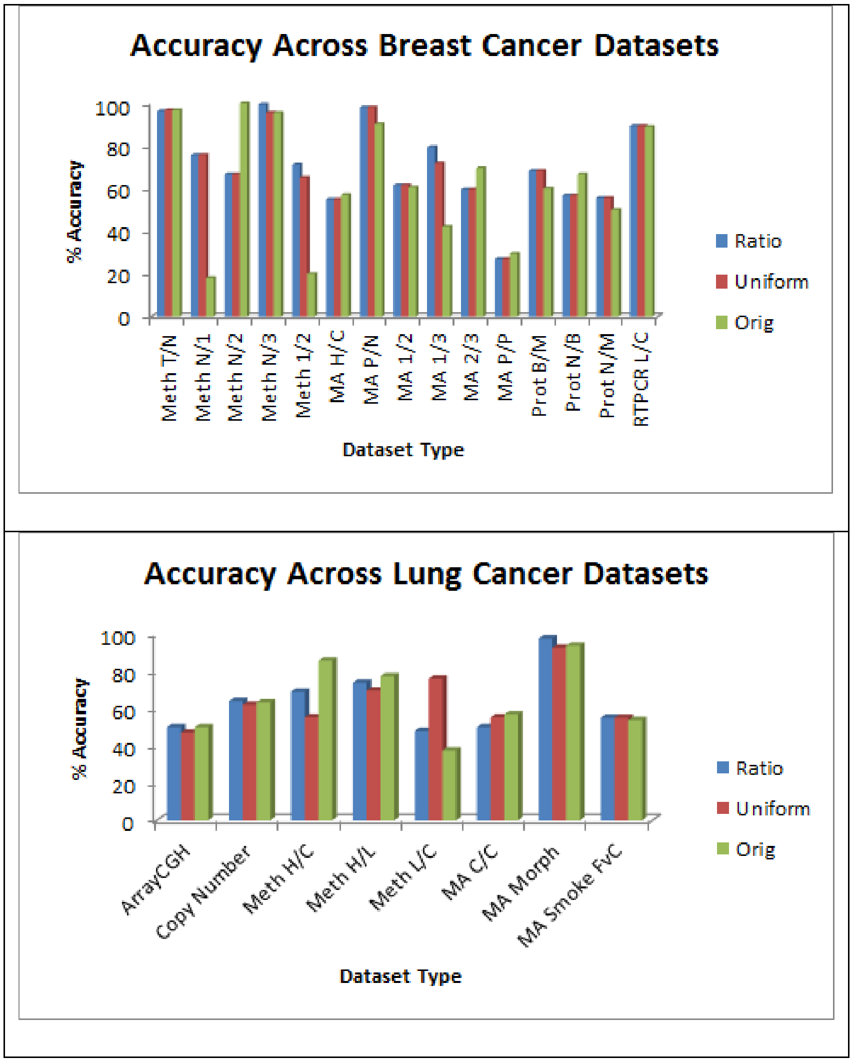


Figure 21. Overall modeling accuracy across different types of experimental data. Results were obtained from BRL. Results from breast cancer datasets found in the top plot, and lung cancer in the bottom. Breast cancer abbreviations: Meth = methylation; MA = microarray; Prot = protein; T/N = tumor vs. normal; N/1 = normal vs. grade 1; N/2 = normal vs. grade 2; N/3 = normal vs. grade 3; 1/2 = grade 1 vs. grade 2; H/C = healthy vs. cancer; P/N = ER+ vs. ER-; P/P = menopause pre vs. post; B/M = benign vs. malignant; N/B = normal vs. benign; N/M = normal vs. malignant; L/C = leukocyte vs. CSF. Lung cancer abbreviations: CGH = comparative gene hybridization; Meth = methylation; MA = microarray; H/C = high vs. control comparison; H/L = high vs. low comparison; L/C = low vs. control comparison; C/C = case vs. control comparison; Morph = morphology comparison (adenocarcinoma vs. scc; Smoke = smoking history; FvC = former vs. current; Ratio = informed priors; Orig = data only.

The BRL was executed with the previously described settings. Figure 21, is a plot of the accuracy of the models tested, across different data types and comparisons. Once again, no consistent pattern of improvement in modeling accuracy is seen across the different data types. The greatest accuracy for the informed prior from all of the dataset types tested is obtained from the microarray dataset comparing lung cancer morphology.

4.2.5 Statistical significance

While individual comparisons can pinpoint a few examples where informed priors increased the modeling accuracy (Figure 20), the accuracy of the models does not appear to improve with the addition of the prior information, on average. Paired student's t-tests were performed using Microsoft Excel 2010 to support the conclusion of no improvement and no decrease in modeling accuracy with the addition of prior information. Accuracy measures from the dataset type comparisons in Figure 21 were used to perform the significance test. None of the test comparisons (informed (I) vs. uniform (U), informed vs. data only (D), uniform vs. data only) showed statistical significant findings based on the t-tests (Breast I-U; p=0.36; Breast I-D; p=0.48; Breast U-D; p=0.51; Lung I-U; p=0.89; Lung I-D; p=0.65; Lung U-D; p=0.92).

4.2.6 Examining modeling attributes

While the lack of improved modeling accuracy was disappointing, scientific value remains. Attributes from the best performing models (Appendix C.27 and C.28) were examined and compared to known biomarker lists (Tables 4 and 5). When two attributes from the same disease model were also identified in the lists of known biomarkers, an assumption was made that a possible disease–specific relationship may exist. Analyzing the results from the lung cancer microarray dataset comparison between former vs. current smokers revealed a possible relationship between SOD2 and CCL5. This possible relationship was found in cross-validation models using both informed prior as well as a uniform prior. No similar finding was achieved from breast cancer models.

4.2.7 Confirmatory research

While the ideal scenario is that the possible relationship discovered is a novel finding, further research and experimentation is warranted. Pathway analysis, database searches, and literature searches were performed in search of further insight into the possible relationship finding. Several pathway databases were explored (KEGG, REACTOME, WikiPathways, BIOCYC, PID, and BioCarta) for pathways containing SOD2 and CCL5. No pathways were found to contain both SOD2 and CCL5 from the databases examined (although Di Renzo *et al.* 2014 refers to SOD in general as part of the Human Oxidatitive Stress Pathway, SOD2 in particular was not mentioned). Several protein interaction databases were searched (StringDB, BioGRID, IntAct) for known interactions and/or interacting entities of SOD2 and CCL5, and lists created for examination.

There does not appear to be a documented interaction between SOD2 and CCL5, nor does there seem to be any direct interacting entities in common obtained from the interaction databases. The genomic database Entrez Gene, www.ncbi.nlm.nih.gov/gene, was investigated for genomic information pertaining to SOD2 and CCL5. Gene ontologies were examined in search of common functions, processes, and/or components. No common function, process, or component ontologies between SOD2 and CCL5 was found.

A query was performed to determine if the SOD2-CCL5 relationship has been discussed previously in any scientific literature. PubMed was searched using the official symbols SOD2 and CCL5, as keywords, and implementing a filter of ‘Homo sapiens’. Seven scientific abstracts were returned as a result of said search.

Kitaya *et al*. (2007) discussed how genes are regulated by Interferon (IFN)-gamma in human uterine microvascular endothelial cells. IFN-gamma plays a critical role in murine uterine spiral artery remodeling for successful pregnancy, and a link to a human role was sought. Treatment with IFN-gamma induced a significant > or =2-fold change in 29 genes in pooled human uterine microvascular endothelial cells; of the 20 genes that were up-regulated, was the chemokine CCL5, and the enzyme SOD2. The results suggest that IFN-gamma regulates the gene expression involved in natural killer cell recruitment, embryo and trophoblast migration, endometrial decidualization, angiogenesis, angiostasis, and anti-viral infection in human uterine microvascular endothelial cells. However, no relationship between SOD2 and CCL5 was discussed; and it was mentioned that the SOD2 protein is not expressed in uterine microvascular endothelial cells *in vivo*.

Qui *et al.* (2009) described a relationship not between SOD2 but between SOD1/*copper-zinc superoxide dismutase* and CCR5/CCL5. They identified SOD1 as mediating CCR5/*C-C chemokine receptor 5* activation by CCL5 in macrophages. They discussed that CCL5/CCR5 are known to play a vital role in regulating leukocyte trafficking, engendering the adaptive immune response and contributing to the pathogenesis of a variety of diseases. While SOD1 was discussed in great detail, SOD2 was not directly mentioned in the entire manuscript.

Jin *et al.* (2010) discussed molecular signatures of maturing dendritic cells. Dendritic cells (DCs) are often produced by GM-CSF/*granulocyte-macrophage colony-stimulating factor* and IL-4/*interleukin 4* stimulation of monocytes. They analyzed the kinetics of DC maturation by LPS/*lipopolysaccharide* and IFN-gamma/*interferon gamma* induction in order to characterize the usefulness of mature DCs (mDCs) for immune therapy and to identify biomarkers for assessing the quality of mDCs. After 24 hours of LPS and IFN-gamma stimulation, Th1 attractant genes such as CCL5 were up-regulated during maturation. The expression of SOD2 was also up-regulated throughout maturation. They concluded that DCs, matured with LPS and IFN-gamma, were characterized by increased levels of Th1 attractants and may be particularly effective for adoptive immune cancer therapy. However, other than mentioning that SOD2 and CCL5 are both classic mature dendritic cell biomarkers, no direct or indirect relationship are discussed, with SOD2 being rarely mentioned at all.

Shah *et al.* (2011) conferred that oxidative stress and chemokines are important factors involved in the development of various clinical features found in patients with systemic lupus and arthritis, chronic inflammatory autoimmune disorders. The anti-oxidant activity of SOD was significantly reduced, and antioxidant molecules showed a negative association with CCL5 in both diseases. They concluded that excessive production of ROS/*reactive oxygen species* disturbs redox status and can modulate the expression of inflammatory chemokines leading to inflammatory processes, and affecting tissue damage in autoimmune diseases, as exemplified by their strong association with disease activity. A general term SOD was mentioned throughout the manuscript, but no delineation was defined between SOD 1 and SOD 2.

Kumar *et al.* (2012) discussed how reactive oxygen species mediate microRNA-302 regulation of cellular proliferation during transitions between normal cell growth phases' quiescence and proliferation. They discussed CCL5 as a target for miR-302, and describe the best possibility of the SOD2-CCL5 relationship via miR-302. MiR-302 levels are decreased significantly by overexpression of SOD2. Because SOD2 converts superoxide to hydrogen peroxide, overexpression of SOD2 is anticipated to increase hydrogen peroxide levels which may lead to ROS sensitivity of miR-302 regulation of ARID4a/*AT-rich interaction domain 4A* and CCL5 mRNA levels.

Di Renzo *et al*. (2014) discussed SOD2 and CCL5 in the context of the positive effect of red wine intake on oxidized-LDL and gene expression. SOD2 and CCL5 were two of six genes examined, but no direct link between the two genes was discussed. They found that when red wine is taken in, values of ox-LDL are lowered ($P < 0.05$) and expression of antioxidant genes is increased, while CCL5 expression is decreased ($P < 0.05$). While a negative correlation in gene expression was revealed between SOD2 and CCL5, no direct link was mentioned.

Kim *et al*. (2014) also examined miR-302's regulation of cell proliferation and cell-cycle progression in adipose tissue-derived mesenchymal stem cells using microarray technology and other assays. They found that miR-302 induces cell proliferation and inhibits oxidant-induced cell death through a reduction in CCL5 expression. However, SOD2 was only mentioned once in the manuscript, as one of the anti-oxidant molecules being tested. Transfection of miR-302 did not affect the expression of SOD molecules. While treatment of the stem cells with $CoCl_2$ increased the gene expression of SOD1 and SOD2, transfection with miR-302 inhibited the $CoCl_2$-induced increase in SOD1 and SOD2. However, no direct relationship was discussed between CCL5 and SOD2.

Results presented above support the idea that incorporation of prior information did not enhance or degrade the model performance, on average (Claim 2, part A; experimentation 4.2.1 – 4.2.4). A consistent pattern of increased modeling accuracy was NOT observed when dataset size, weighting, and different data types were examined (experimentation 4.2.1 - 4.2.3).

However, these results cannot be generalized because only one dataset was tested. Statistically significant differences in modeling accuracy were NOT achieved (experimentation 4.2.4). An example is shown where analyzing the attributes used to build the best-performing models did lead to a possible new interaction (Claim 2, part B; experimentation 4.2.5). A very likely relationship may exist between SOD2 and CCL5 (experimentation 4.2.5, Kumar *et al*. 2012). While the two entities have been examined together in several studies, with miR-302 acting as an intermediary, it is possible that SOD2 may indirectly regulate CCL5 (experimentation 4.2.6). Similarly, a direct relationship has been described between CCL5 and SOD1. One would assume that SOD1 and SOD2 exhibit very similar molecular traits and behaviors. Additional confirmatory research is needed to support the idea since it is hypothesized, but not clearly defined in the literature (experimentation 4.2.6). The ’not so promising results’ that were achieved from the limited experimentation led to a change in direction of study (pathway analysis) as described in section 4.3.

4.3 PATHWAY ANALYSIS RESULTS

The goal of pathway analysis is to present a visual diagram of a biological process or pathway to enable researchers to better understand the process, as well as the biological entities involved in the process, and identify targets for altering the process. Three different evaluation measures were assessed in order to determine the performance from the pathway analyses: number of input genes in each pathway, impact factor of each pathway, and individual pathway p-values. These three measures contributed to the understanding and assessment of the results. Pathway Express allows for expression values to accompany the gene/protein name, which can be incorporated as an added benefit of the pathway analysis.

The following sections describe the various comparisons and exercises performed, relating to pathway analysis. Sections 4.3.1 – 4.3.1.3 test the idea that incorporation of prior information enhances pathway analysis results by identifying more input genes in breast cancer-relevant pathways (Claim 2, part C). Sections 4.3.2 – 4.3.2.3 test the idea that incorporation of prior information enhances pathway analysis results by identifying more input genes in lung cancer-relevant pathways (also Claim 2, part C).

Pathway Express returned 93 KEGG pathways from our breast and lung cancer data input. While a comprehensive analysis is always optimal, the study was limited to 22 pathways directly pertinent to breast cancer and 23 pathways directly pertinent to lung cancer. These pathways were chosen due to established biology relationships (ERBB2 and breast cancer); as being described as significant by Guille *et al*. 2013; or as being returned identified by a disease search of 'breast cancer' or 'lung cancer' using the KEGG website.

4.3.1 Experimental design using literature mining results

Prior incorporation was performed differently for pathway analysis compared to the modeling exercises. Pathway Express takes as input a list of gene/protein ID's and one value. In order to achieve a representative value, the dataset for each comparison was analyzed first using caGEDA. The resulting differentially expressed value was in the form of a J5 score. The J5 score for each variable (gene/protein) was multiplied by the informed or uniform prior value (or by 1 where no prior was used), to achieve the value used for Pathway Express input.

4.3.2 Breast cancer datasets

Seven previously described (Table 6, Section 3.2.2.1) breast cancer datasets were analyzed: 1) copy number case vs. control (BCCNCvN); 2) microarray grade 1 vs. grade 2 (BCMA1v2); 3) microarray grade 1 vs. grade 3 (BCMA1v3); 4) microarray grade 2 vs. grade 3 (BCMA2v3); 5) microarray blood healthy vs. cancer (BCMABlood); 6) microarray ER status positive vs. negative (BCMAER); 7) microarray menopausal status pre vs. post (BCMAMeno).

Analysis results underwent post-analysis processing by one of five different methods: data-only (D; 835 most differentially expressed J5 scores); ratio (R; 835 genes from literature mining); z-scores (Z; 835 genes from literature mining); the product of data & ratio (DR; highest scoring 835 genes when J5 score multiplied by literature mining ratio); and the product of data & z-scores (DZ; highest scoring 835 genes when J5 score multiplied by literature mining z-score). In most cases the ratio (R) and z-score (Z) return the same genes, however, due to the mathematical sign of the value, (ratio always being a positive number, whereas the z-score may be positive or negative number), the gene expression values may be different. The same phenomenon applies to data & ratio (DR) and data & z-score (DZ), where the same genes are usually returned but the expression values and their signs may differ. Post-analysis processed values accompanied the gene ID's in the genelists from the breast cancer microarray grade 1 vs. 3 dataset and were input into Pathway Express.

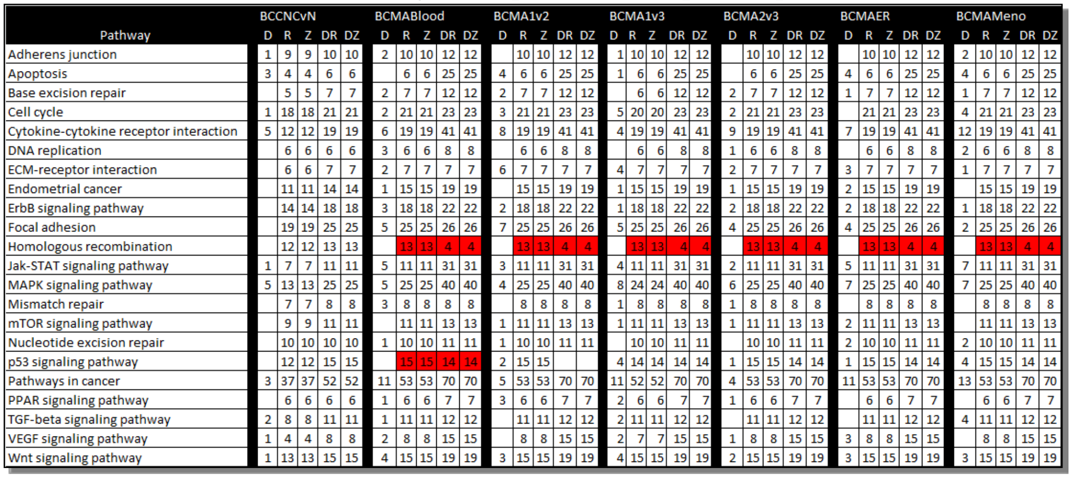
All 22 relevant breast cancer pathways were assessed for each post-processing method. The Apoptosis Pathway produced from breast cancer microarray grade 1 vs. 3 dataset will be used as an in-depth example. The Apoptosis Pathway was chosen because 1) a steady increase in the number of input genes in the pathway is shown, from D (n=1) to R/Z (n=6) to DR/DZ (n=25); 2) the impact factor (***if***) increased by around a factor of 10, from D (***if***=3.07), R (***if***=2.70), and Z (***if***=2.25) to DR (***if***=27.26) and DZ (***if***=27.64); and 3) the p-values changed from non-significant, D (p=0.98) and R/Z (p=0.17), to significant values of p=2.275 X 10^{-12} for DR/DZ.

'Data only' results are what would normally be obtained from a typical pathway analysis using only experimental data array findings. Combining prior knowledge with experimental data array findings adds an additional layer of confidence to the reported pathway analysis findings, which allows the researcher more avenues to pursue in drawing conclusions, as well as more areas to focus on for future work.

Assessment of pathway measures

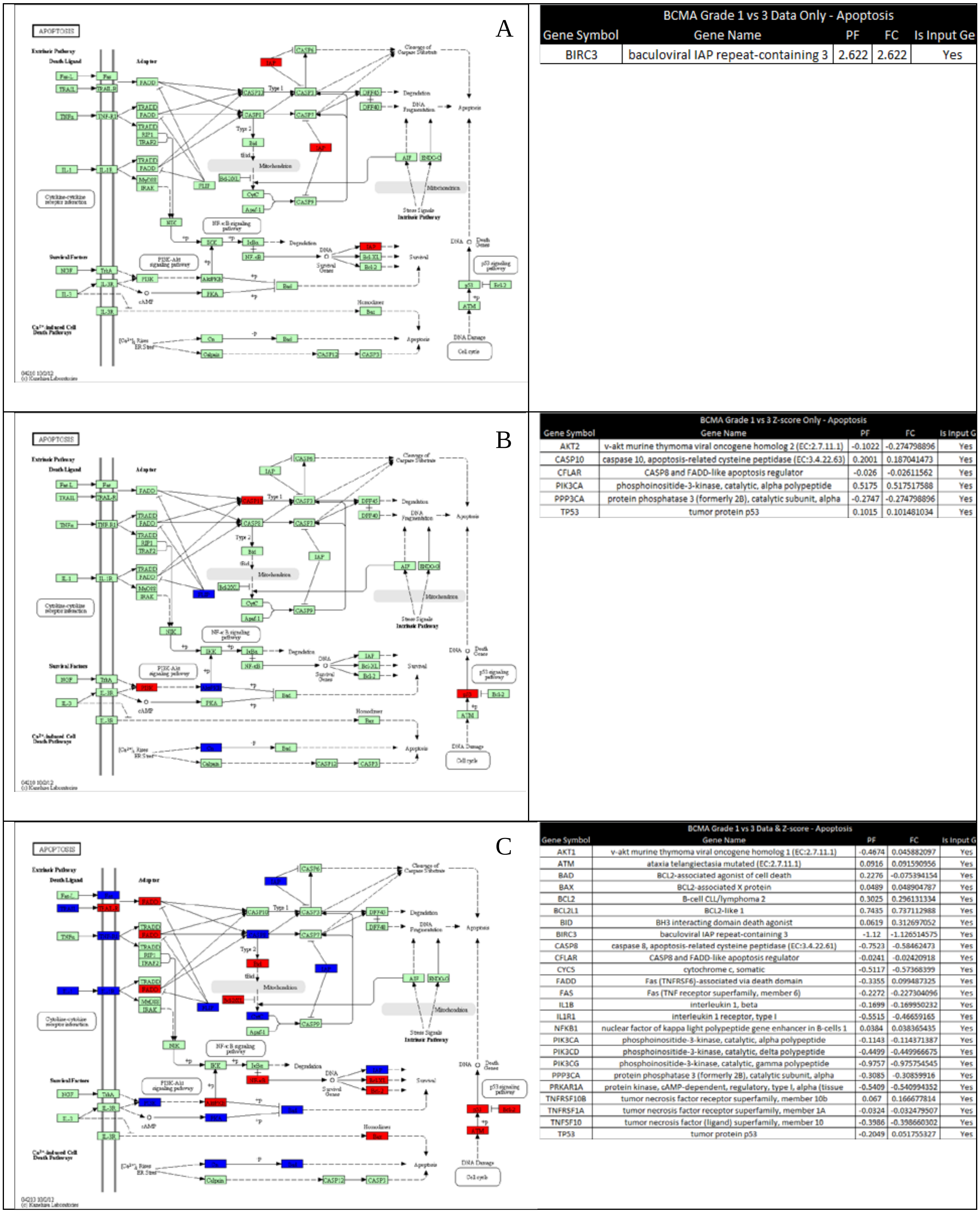
In assessing the pathway measures, the most important measure to be examined is the number of input genes in pathway. It is from this count that all other measures are calculated.

Table 12. Number of input genes in pathway from breast cancer datasets. D = data only; R = ratio only; Z = z-score only; DR = data & ratio; DZ = data & z-score. Red highlights cases where the number of input genes in the pathway decreased from left (R/Z only) to right (DR/DZ). BCCNCvN = breast cancer copy number case vs. control; BCMA1v2 = breast cancer microarray grade 1 vs. grade 2; BCMA1v3 = breast cancer microarray grade 1 vs. grade 3; BCMA2v3 = breast cancer microarray grade 2 vs. grade 3; BCMA Blood = breast cancer microarray blood healthy vs. cancer; BCMAER = breast cancer microarray ESR1 status positive vs. negative; BCMAMeno = breast cancer microarray menopausal status pre vs. post.



	BCCNCvN					BCMABlood					BCMA1v2					BCMA1v3					BCMA2v3					BCMAER					BCMAMeno				
Pathway	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ
Adherens junction	1	9	9	10	10	2	10	10	12	12		10	10	12	12	1	10	10	12	12		10	10	12	12		10	10	12	12	2	10	10	12	12
Apoptosis	3	4	4	6	6		6	6	25	25	4	6	6	25	25	1	6	6	25	25		6	6	25	25	4	6	6	25	25	4	6	6	25	25
Base excision repair		5	5	7	7	2	7	7	12	12	2	7	7	12	12		6	6	12	12	2	7	7	12	12	1	7	7	12	12	1	7	7	12	12
Cell cycle	1	18	18	21	21	2	21	21	23	23	3	21	21	23	23	5	20	20	23	23	2	21	21	23	23		21	21	23	23	4	21	21	23	23
Cytokine-cytokine receptor interaction	5	12	12	19	19	6	19	19	41	41	8	19	19	41	41	4	19	19	41	41	9	19	19	41	41	7	19	19	41	41	12	19	19	41	41
DNA replication		6	6	6	6	3	6	6	8	8		6	6	8	8		6	6	8	8	1	6	6	8	8		6	6	8	8	2	6	6	8	8
ECM-receptor interaction		6	6	7	7	2	7	7	7	7	6	7	7	7	7	4	7	7	7	7	2	7	7	7	7	3	7	7	7	7	1	7	7	7	7
Endometrial cancer		11	11	14	14	1	15	15	19	19		15	15	19	19	1	15	15	19	19	1	15	15	19	19	2	15	15	19	19		15	15	19	19
ErbB signaling pathway		14	14	18	18	3	18	18	22	22	2	18	18	22	22	1	18	18	22	22	2	18	18	22	22	2	18	18	22	22	1	18	18	22	22
Focal adhesion		19	19	25	25	5	25	25	26	26	7	25	25	26	26	5	25	25	26	26	4	25	25	26	26	4	25	25	26	26	2	25	25	26	26
Homologous recombination		12	12	13	13		13	13	4	4		13	13	4	4		13	13	4	4		13	13	4	4		13	13	4	4		13	13	4	4
Jak-STAT signaling pathway	1	7	7	11	11	5	11	11	31	31	3	11	11	31	31	4	11	11	31	31	2	11	11	31	31	5	11	11	31	31	7	11	11	31	31
MAPK signaling pathway	5	13	13	25	25	5	25	25	40	40	4	25	25	40	40	8	24	24	40	40	6	25	25	40	40	7	25	25	40	40	7	25	25	40	40
Mismatch repair		7	7	8	8	3	8	8	8	8		8	8	8	8	1	8	8	8	8	1	8	8	8	8		8	8	8	8		8	8	8	8
mTOR signaling pathway		9	9	11	11		11	11	13	13	1	11	11	13	13	1	11	11	13	13	1	11	11	13	13	2	11	11	13	13		11	11	13	13
Nucleotide excision repair		10	10	10	10	1	10	10	11	11	1	10	10	11	11		10	10	11	11		10	10	11	11	2	10	10	11	11	2	10	10	11	11
p53 signaling pathway		12	12	15	15		15	15	14	14	2	15	15			4	14	14	14	14	1	15	15	14	14	1	15	15	14	14	4	15	15	14	14
Pathways in cancer	3	37	37	52	52	11	53	53	70	70	5	53	53	70	70	11	52	52	70	70	4	53	53	70	70	11	53	53	70	70	13	53	53	70	70
PPAR signaling pathway		6	6	6	6	1	6	6	7	7	3	6	6	7	7	2	6	6	7	7	1	6	6	7	7		6	6	7	7		6	6	7	7
TGF-beta signaling pathway	2	8	8	11	11	1	11	11	12	12		11	11	12	12	2	11	11	12	12		11	11	12	12		11	11	12	12	4	11	11	12	12
VEGF signaling pathway	1	4	4	8	8	2	8	8	15	15		8	8	15	15	2	7	7	15	15	1	8	8	15	15	3	8	8	15	15		8	8	15	15
Wnt signaling pathway	1	13	13	15	15	4	15	15	19	19	3	15	15	19	19	4	15	15	19	19	2	15	15	19	19	3	15	15	19	19	3	15	15	19	19

Table 12 shows the number of input genes found in a given pathway using a given genelist. In general, a steady progression exists from left to right with D genelists providing the least amount of input genes in the returned pathways; the R/Z genelists producing more input genes in the returned pathways; and DR/DZ combined genelists producing the most genes in a pathway.



BCMA Grade 1 vs 3 Data Only - Apoptosis				
Gene Symbol	Gene Name	PF	FC	Is Input Ge
BIRC3	baculoviral IAP repeat-containing 3	2.622	2.622	Yes

BCMA Grade 1 vs 3 Z-score Only - Apoptosis				
Gene Symbol	Gene Name	PF	FC	Is Input G
AKT2	v-akt murine thymoma viral oncogene homolog 2 (EC:2.7.11.1)	-0.1022	-0.274798896	Yes
CASP10	caspase 10, apoptosis-related cysteine peptidase (EC:3.4.22.63)	0.2001	0.187041473	Yes
CFLAR	CASP8 and FADD-like apoptosis regulator	-0.026	-0.02611562	Yes
PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	0.5175	0.517517588	Yes
PPP3CA	protein phosphatase 3 (formerly 2B), catalytic subunit, alpha	-0.2747	-0.274798896	Yes
TP53	tumor protein p53	0.1015	0.101481034	Yes

BCMA Grade 1 vs 3 Data & Z-score - Apoptosis				
Gene Symbol	Gene Name	PF	FC	Is Input G
AKT1	v-akt murine thymoma viral oncogene homolog 1 (EC:2.7.11.1)	-0.4674	0.045882097	Yes
ATM	ataxia telangiectasia mutated (EC:2.7.11.1)	0.0916	0.091590956	Yes
BAD	BCL2-associated agonist of cell death	0.2276	-0.075394154	Yes
BAX	BCL2-associated X protein	0.0489	0.048904787	Yes
BCL2	B-cell CLL/lymphoma 2	0.3025	0.296131334	Yes
BCL2L1	BCL2-like 1	0.7435	0.737112988	Yes
BID	BH3 interacting domain death agonist	0.0619	0.312697052	Yes
BIRC3	baculoviral IAP repeat-containing 3	-1.12	-1.126514575	Yes
CASP8	caspase 8, apoptosis-related cysteine peptidase (EC:3.4.22.61)	-0.7523	-0.58462473	Yes
CFLAR	CASP8 and FADD-like apoptosis regulator	-0.0241	-0.02420918	Yes
CYCS	cytochrome c, somatic	-0.5117	-0.57368399	Yes
FADD	Fas (TNFRSF6)-associated via death domain	-0.3355	0.099487325	Yes
FAS	Fas (TNF receptor superfamily, member 6)	-0.2272	-0.227304096	Yes
IL1B	interleukin 1, beta	-0.1699	-0.169950232	Yes
IL1R1	interleukin 1 receptor, type I	-0.5515	-0.46659165	Yes
NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	0.0384	0.038365435	Yes
PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	-0.1143	-0.114371387	Yes
PIK3CD	phosphoinositide-3-kinase, catalytic, delta polypeptide	-0.4499	-0.449966675	Yes
PIK3CG	phosphoinositide-3-kinase, catalytic, gamma polypeptide	-0.9757	-0.975754545	Yes
PPP3CA	protein phosphatase 3 (formerly 2B), catalytic subunit, alpha	-0.3085	-0.30859916	Yes
PRKAR1A	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue	-0.5409	-0.540994352	Yes
TNFRSF10B	tumor necrosis factor receptor superfamily, member 10b	0.067	0.166677814	Yes
TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A	-0.0324	-0.032479507	Yes
TNFSF10	tumor necrosis factor (ligand) superfamily, member 10	-0.3986	-0.398660302	Yes
TP53	tumor protein p53	-0.2049	0.051755327	Yes

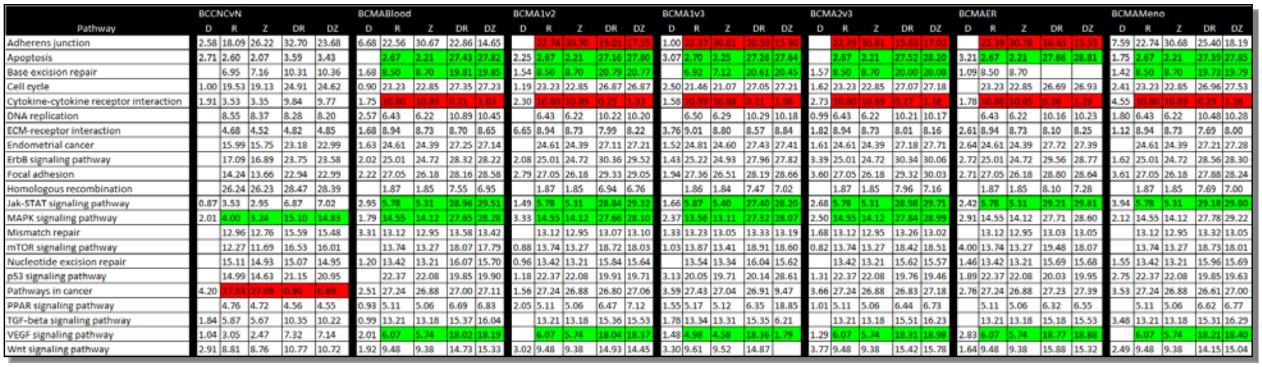
Figure 22. KEGG Apoptosis Pathway diagrams and input gene tables created using breast cancer microarray grade 1 vs. 3 dataset. A = data-only (D); B = z-score only (Z); C = data & z-score (DZ). Input genes highlighted red = upregulated and blue = downregulated. PF = perturbation factor; FC = fold-change.

The KEGG Apoptosis Pathway diagram, output from Pathway Express, contains 89 total genes (Figure 22, A). When the data only (D) genelist is analyzed, only one input gene is highlighted in the output Apoptosis pathway (BIRC3/*baculoviral IAP repeat containing 3*, in three different roles). When the literature mining z-score (Z) genelist is input (Figure 22, B), six input genes are highlighted in the output Apoptosis Pathway (AKT2/*v-akt murine thyoma viral oncogene homolog 2*, CASP10/*caspase 10*, CFLAR/*CASP8 and FADD-like apoptosis regulator*, PIK3CA*, PPP3CA/*protein phosphatase 3 catalytic subunit alpha*, and TP53*). Finally, when the combined DZ genelist is used as input (Figure 22, C), 25 input genes are highlighted in the output Apoptosis Pathway (AKT1*/*v-akt murine thyoma viral oncogene homolog 1*, ATM*/*ATM serine-threonine kinase*, BAD/*BCL2-associated agonist of cell death*, BAX/*BCL2-associated X protein*, BCL2*/*B-cell CLL-lymphoma 2*, BCL2L1/*BCL2-like 1*, BID/*BH3 interacting domain death agonist*, BIRC3, CASP8*/*caspase 8*, CFLAR, CYCS/*cytochrome c, somatic*, FADD/*Fas-associated via death domain*, FAS*/*Fas cell surface death receptor*, IL1B/*interleukin 1 beta*, IL1R1*/*interleukin 1 receptor type 1*, NFKB1/*nuclear factor of kappa light polypeptide gene enhancer in B-cells 1*, PIK3CA*, PIK3CD/*phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit delta*, PIK3CG/ *phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit gamma*, PPP3CA, PRKAR1A/*protein kinase c-AMP-dependent type 1 regulatory subunit alpha*, TNFRSF10B*/*tumor necrosis factor receptor superfamily member 10b*, TNFRSF1A*, TNFSF10*, TP53*).

* Indicates gene/protein found in list of known breast cancer biomarkers Table 4

The actual up-regulation or down-regulation of the genes is not of great importance for this exercise. The z-score measure and J5 scores produce both positive and negative values, whereas the ratio only produces positive values. As such, the sign of the input value will depend on the two values being multiplied together, and so careful interpretation will be required of the researcher. The major point of emphasis here is the enriched input dataset, which produces more-relevant results, in the output pathways, when data and prior knowledge are combined.

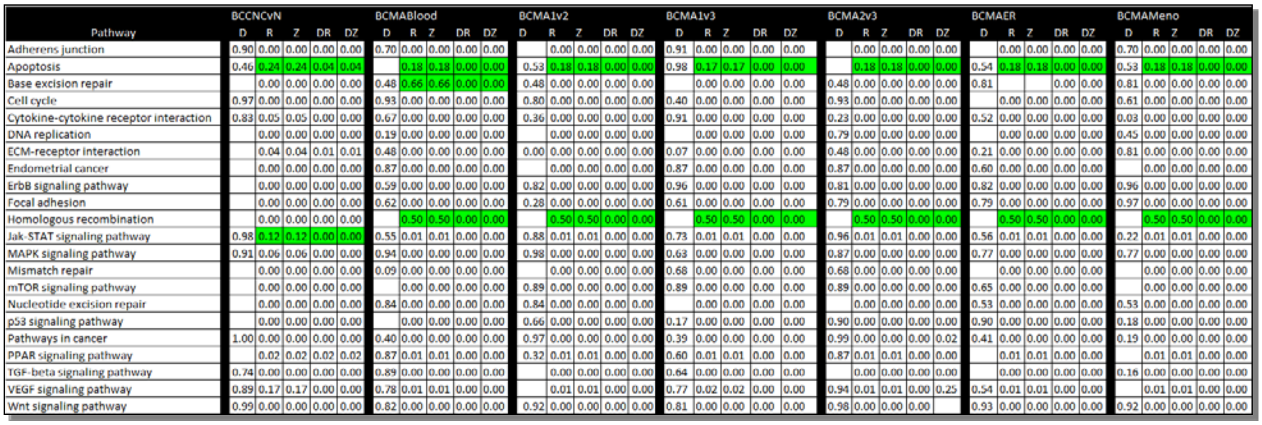
Table 13. Impact factors from breast cancer datasets. D = data only; R = ratio only; Z = z-score only; DR = data & ratio; DZ = data & z-score. Red highlights cases where the impact factor decreased from left (R/Z only) to right (DR/DZ). Green highlights cases where the impact factor increased by a factor of three or greater. BCCNCvN = breast cancer copy number case vs. control; BCMA1v2 = breast cancer microarray grade 1 vs. grade 2; BCMA1v3 = breast cancer microarray grade 1 vs. grade 3; BCMA2v3 = breast cancer microarray grade 2 vs. grade 3; BCMA Blood = breast cancer microarray blood healthy vs. cancer; BCMAER = breast cancer microarray ESR1 status positive vs. negative; BCMAMeno = breast cancer microarray menopausal status pre vs. post.



	BCCNCvN					BCMABlood					BCMA1v2					BCMA1v3					BCMA2v3					BCMAER					BCMAMeno				
Pathway	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ
Adherens junction	2.58	18.09	26.22	32.70	23.68	6.68	22.56	30.67	22.86	14.65		22.74	30.70	19.01	17.55	1.00	22.37	30.81	20.50	15.90		22.39	30.81	15.63	17.02		22.39	30.78	16.45	15.53	7.59	22.74	30.68	25.40	18.19
Apoptosis	2.71	2.60	2.07	3.59	3.43		2.67	2.21	27.43	27.82	2.25	2.67	2.21	27.16	27.80	3.07	2.70	2.25	27.26	27.64		2.67	2.21	27.52	28.20	3.21	2.67	2.21	27.86	28.81	1.75	2.67	2.21	27.39	27.85
Base excision repair		6.95	7.16	10.31	10.36	1.68	8.50	8.70	19.81	19.85	1.54	8.50	8.70	20.79	20.77		6.92	7.12	20.61	20.45	1.57	8.50	8.70	20.00	20.08	1.09	8.50	8.70			1.42	8.50	8.70	19.73	19.79
Cell cycle	1.00	19.53	19.13	24.91	24.62	0.90	23.23	22.85	27.35	27.23	1.19	23.23	22.85	26.87	26.87	2.50	21.46	21.07	27.05	27.21	1.62	23.23	22.85	27.07	27.18		23.23	22.85	26.69	26.93	2.41	23.23	22.85	26.96	27.53
Cytokine-cytokine receptor interaction	1.91	3.53	3.35	9.84	9.77	1.75	10.80	10.69	0.21	1.03	2.30	10.80	10.69	0.29	1.33	1.58	10.99	10.88	0.21	1.08	2.73	10.80	10.69	0.27	1.18	1.78	10.80	10.69	0.26	1.26	4.55	10.80	10.69	0.29	1.26
DNA replication		8.55	8.37	8.28	8.20	2.57	6.43	6.22	10.89	10.45		6.43	6.22	10.22	10.20		6.50	6.29	10.29	10.18	0.99	6.43	6.22	10.21	10.17		6.43	6.22	10.16	10.23	1.80	6.43	6.22	10.48	10.28
ECM-receptor interaction		4.68	4.52	4.82	4.85	1.68	8.94	8.73	8.70	8.65	6.65	8.94	8.73	7.99	8.22	3.76	9.01	8.80	8.57	8.84	1.82	8.94	8.73	8.01	8.16	2.61	8.94	8.73	8.10	8.25	1.12	8.94	8.73	7.69	8.00
Endometrial cancer		15.99	15.75	23.18	22.99	1.63	24.61	24.39	27.25	27.14		24.61	24.39	27.11	27.21	1.52	24.81	24.60	27.43	27.41	1.61	24.61	24.39	27.18	27.71	2.64	24.61	24.39	27.72	27.39		24.61	24.39	27.21	27.28
ErbB signaling pathway		17.09	16.89	23.75	23.58	2.02	25.01	24.72	28.32	28.22	2.08	25.01	24.72	30.36	29.52	1.43	25.22	24.93	27.96	27.82	3.39	25.01	24.72	30.34	30.06	2.72	25.01	24.72	29.56	28.77	1.62	25.01	24.72	28.56	28.30
Focal adhesion		14.24	13.66	22.94	22.99	2.22	27.05	26.18	28.16	28.58	2.79	27.05	26.18	29.33	29.05	1.94	27.36	26.51	28.19	28.66	3.60	27.05	26.18	29.32	30.03	2.71	27.05	26.18	28.80	28.64	3.61	27.05	26.18	27.88	28.24
Homologous recombination		26.24	26.23	28.47	28.39		1.87	1.85	7.55	6.95		1.87	1.85	6.94	6.76		1.86	1.84	7.47	7.02		1.87	1.85	7.96	7.16		1.87	1.85	8.10	7.28		1.87	1.85	7.69	7.00
Jak-STAT signaling pathway	0.87	3.53	2.95	6.87	7.02	2.95	5.78	5.31	28.96	29.51	1.49	5.78	5.31	28.84	29.32	1.66	5.87	5.40	27.40	28.20	2.68	5.78	5.31	28.98	29.71	2.42	5.78	5.31	29.21	29.61	3.94	5.78	5.31	29.18	29.80
MAPK signaling pathway	2.01	4.00	3.24	15.10	14.83	1.79	14.55	14.12	27.65	28.28	3.33	14.55	14.12	27.66	28.10	2.37	13.56	13.11	27.32	28.07	2.50	14.55	14.12	27.84	28.99	2.91	14.55	14.12	27.71	28.60	2.12	14.55	14.12	27.78	29.22
Mismatch repair		12.96	12.76	15.59	15.48	3.31	13.12	12.95	13.58	13.42		13.12	12.95	13.07	13.10	1.33	13.23	13.05	13.33	13.19	1.68	13.12	12.95	13.26	13.02		13.12	12.95	13.03	13.05		13.12	12.95	13.32	13.05
mTOR signaling pathway		12.27	11.69	16.53	16.01		13.74	13.27	18.07	17.79	0.88	13.74	13.27	18.72	18.03	1.03	13.87	13.41	18.91	18.60	0.82	13.74	13.27	18.42	18.51	4.00	13.74	13.27	19.48	18.07		13.74	13.27	18.73	18.01
Nucleotide excision repair		15.11	14.93	15.07	14.95	1.20	13.42	13.21	16.07	15.70	0.96	13.42	13.21	15.84	15.64		13.54	13.34	16.04	15.62		13.42	13.21	15.62	15.57	1.46	13.42	13.21	15.69	15.68	1.55	13.42	13.21	15.96	15.69
p53 signaling pathway		14.99	14.63	21.15	20.95		22.37	22.08	19.85	19.90	1.18	22.37	22.08	19.91	19.71	3.13	20.05	19.71	20.14	28.61	1.31	22.37	22.08	19.76	19.46	1.89	22.37	22.08	20.03	19.95	2.75	22.37	22.08	19.85	19.63
Pathways in cancer	4.20	27.53	27.09	0.90	0.69	2.51	27.24	26.88	27.00	27.11	1.56	27.24	26.88	26.80	27.06	3.59	27.43	27.04	26.91	9.47	3.66	27.24	26.88	26.83	27.18	2.76	27.24	26.88	27.23	27.39	3.53	27.24	26.88	26.61	27.00
PPAR signaling pathway		4.76	4.72	4.56	4.55	0.93	5.11	5.06	6.69	6.83	2.05	5.11	5.06	6.47	7.12	1.55	5.17	5.12	6.35	18.85	1.01	5.11	5.06	6.44	6.73		5.11	5.06	6.32	6.55		5.11	5.06	6.62	6.77
TGF-beta signaling pathway	1.84	5.87	5.67	10.35	10.22	0.99	13.21	13.18	15.37	16.04		13.21	13.18	15.36	15.53	1.78	13.34	13.31	15.35	6.21		13.21	13.18	15.51	16.23		13.21	13.18	15.18	15.53	3.48	13.21	13.18	15.31	16.29
VEGF signaling pathway	1.04	3.05	2.47	7.32	7.14	2.01	6.07	5.74	18.02	18.19		6.07	5.74	18.04	18.37	1.48	4.98	4.58	18.36	1.79	1.29	6.07	5.74	18.31	18.98	2.83	6.07	5.74	18.77	18.88		6.07	5.74	18.21	18.40
Wnt signaling pathway	2.91	8.81	8.76	10.77	10.72	1.92	9.48	9.38	14.73	15.33	3.02	9.48	9.38	14.93	14.45	3.30	9.61	9.52	14.87		3.77	9.48	9.38	15.42	15.78	1.64	9.48	9.38	15.88	15.32	2.49	9.48	9.38	14.15	15.04

The impact factor calculation relies on the number of differentially regulated genes in the pathway and perturbation factors of all genes in the pathway. Similarly to section 4.3.1.1 results, for most pathways examined, a steady progression exists from left to right with D genelists producing the smallest impact factor; the R/Z genelists producing larger impact factors than D; and DR/DZ combined genelists producing the largest impact factors, in general. Some exceptions do exist however, as can be seen in Table 13. Red highlights indicate where the impact factor was greater for the ratio/z-score only input genelists than it was for the combined DR/DZ input genelists.

Table 14. P-values from breast cancer datasets. D = data only; R = ratio only; Z = z-score only; DR = data & ratio; DZ = data & z-score. Green highlights cases where the p-values went from non-significant to significant in going from left (R/Z only) to right (DR/DZ). BCCNCvN = breast cancer copy number case vs. control; BCMA1v2 = breast cancer microarray grade 1 vs. grade 2; BCMA1v3 = breast cancer microarray grade 1 vs. grade 3; BCMA2v3 = breast cancer microarray grade 2 vs. grade 3; BCMA Blood = breast cancer microarray blood healthy vs. cancer; BCMAER = breast cancer microarray ESR1 status positive vs. negative; BCMAMeno = breast cancer microarray menopausal status pre vs. post.



	BCCNCvN					BCMABlood					BCMA1v2					BCMA1v3					BCMA2v3					BCMAER					BCMAMeno				
Pathway	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ
Adherens junction	0.90	0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.00
Apoptosis	0.46	0.24	0.24	0.04	0.04		0.18	0.18	0.00	0.00	0.53	0.18	0.18	0.00	0.00	0.98	0.17	0.17	0.00	0.00		0.18	0.18	0.00	0.00	0.54	0.18	0.18	0.00	0.00	0.53	0.18	0.18	0.00	0.00
Base excision repair		0.00	0.00	0.00	0.00	0.48	0.66	0.66	0.00	0.00	0.48	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.48	0.00	0.00	0.00	0.00	0.81			0.00	0.00	0.81	0.00	0.00	0.00	0.00
Cell cycle	0.97	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.00
Cytokine-cytokine receptor interaction	0.83	0.05	0.05	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
DNA replication		0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.79	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.45	0.00	0.00	0.00	0.00
ECM-receptor interaction		0.04	0.04	0.01	0.01	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.48	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.81	0.00	0.00	0.00	0.00
Endometrial cancer		0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.60	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00
ErbB signaling pathway		0.00	0.00	0.00	0.00	0.59	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.81	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00
Focal adhesion		0.00	0.00	0.00	0.00	0.62	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Homologous recombination		0.00	0.00	0.00	0.00		0.50	0.50	0.00	0.00		0.50	0.50	0.00	0.00		0.50	0.50	0.00	0.00		0.50	0.50	0.00	0.00		0.50	0.50	0.00	0.00		0.50	0.50	0.00	0.00
Jak-STAT signaling pathway	0.98	0.12	0.12	0.00	0.00	0.55	0.01	0.01	0.00	0.00	0.88	0.01	0.01	0.00	0.00	0.73	0.01	0.01	0.00	0.00	0.96	0.01	0.01	0.00	0.00	0.56	0.01	0.01	0.00	0.00	0.22	0.01	0.01	0.00	0.00
MAPK signaling pathway	0.91	0.06	0.06	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.63	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00
Mismatch repair		0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00
mTOR signaling pathway		0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.65	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00
Nucleotide excision repair		0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.53	0.00	0.00	0.00	0.00	0.53	0.00	0.00	0.00	0.00
p53 signaling pathway		0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.66	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00
Pathways in cancer	1.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.02	0.41	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00
PPAR signaling pathway		0.02	0.02	0.02	0.02	0.87	0.01	0.01	0.00	0.00	0.32	0.01	0.01	0.00	0.00	0.60	0.01	0.01	0.00	0.00	0.87	0.01	0.01	0.00	0.00		0.01	0.01	0.00	0.00		0.01	0.01	0.00	0.00
TGF-beta signaling pathway	0.74	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00
VEGF signaling pathway	0.89	0.17	0.17	0.00	0.00	0.78	0.01	0.01	0.00	0.00		0.01	0.01	0.00	0.00	0.77	0.02	0.02	0.00	0.00	0.94	0.01	0.01	0.00	0.25	0.54	0.01	0.01	0.00	0.00		0.01	0.01	0.00	0.00
Wnt signaling pathway	0.99	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.81	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00		0.93	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00

The p-value calculation identifies pathways that contain a proportion of differentially expressed genes that are significantly different from what is expected at random. A significance threshold of ≤ 0.05 is applied. Again, similarly to sections 4.3.1.1 and 4.3.1.2 results, for most pathways examined, a steady progression exists from left to right with D genelists producing the least significant pathway p-values; the R/Z genelists producing more significant pathway p-values than D; and DR/DZ combined genelists producing the most significant pathway p-values, in general. In Table 14, green highlights cases where pathway p-values for R/Z produced genelists could be considered non-significant, but pathway p-values for the DR/DZ produced genelists could be considered significant.

4.3.3 Lung cancer datasets

Five previously described lung cancer datasets were analyzed: 1) microarray case vs. control (LCMACvC); 2) microarray smoking never vs. former (LCMANvF); 3) microarray smoking former vs. current (LCMAFvC); 4) microarray smoking never vs. current (LCMANvC); 5) microarray morphology small cell carcinoma vs. adenocarcinoma (LCMAMorph).

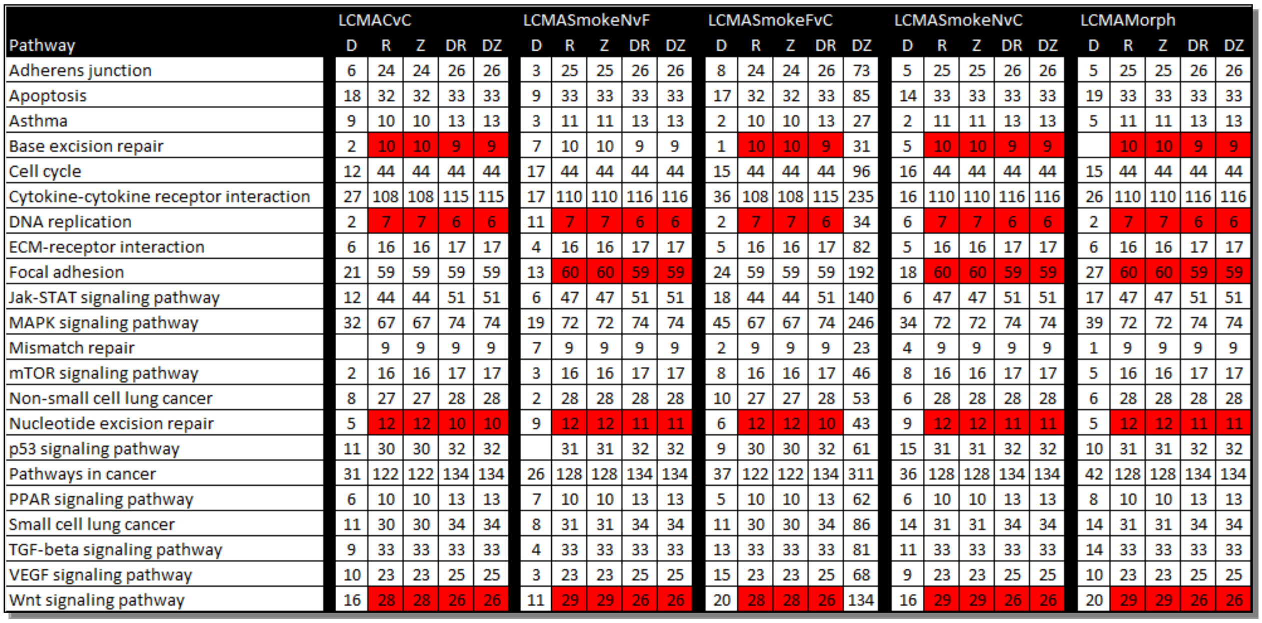
Analysis results underwent post-analysis processing by one of five different methods: data-only (D; 1660 most differentially expressed J5 scores); ratio (R; 1660 genes from literature mining); z-scores (Z; 1660 genes from literature mining); the product of data & ratio (DR; highest scoring 1660 genes when J5 score multiplied by literature mining ratio); and the product of data & z-scores (DZ; highest scoring 1660 genes when J5 score multiplied by literature mining z-score). Post-analysis processed values accompanied the gene ID's in the genelists from the lung cancer microarray case vs. control dataset and were input into Pathway Express.

All 23 relevant lung cancer pathways were assessed for each post-processing method. The PPAR/*peroxisome proliferator activated receptor alpha* Signaling Pathways produced from lung cancer microarray case vs. control datasets will be used as an in-depth example. The PPAR Signaling Pathway was chosen because 1) a steady increase in the number of input genes in the pathway is shown, from D (n=6) to R/Z (n=10) to DR/DZ (n=13); 2) the impact factor (***if***) doubled, from D (***if***=1.27), R (***if***=2.21), and Z (***if***=2.27) to DR (***if***=4.59) and DZ (***if***=4.62); and 3) the p-values changed from non-significant, D (p=0.741) and R/Z (p=0.187), to significant values of p=0.026 for DR/DZ.

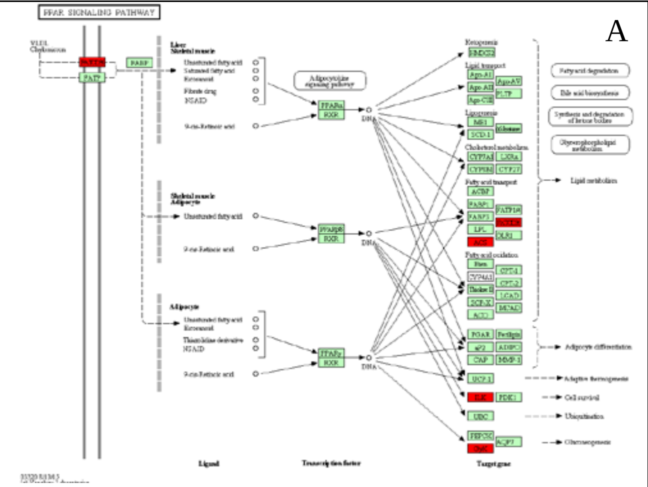
Assessment of pathway measures

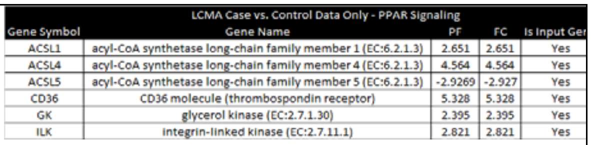
Table 15 shows the number of input genes found in a given pathway using a given genelist. In general, a steady progression exists from left to right with D genelists providing the least amount of input genes in the returned pathways; the R/Z genelists producing more input genes in the returned pathways; and DR/DZ combined genelists producing the most genes in a given pathway.

Table 15. Number of input genes in pathway from lung cancer datasets. D = data only; R = ratio only; Z = z-score only; DR = data & ratio; DZ = data & z-score. Red highlights cases where the number of input genes in the pathway decreased from left (R/Z only) to right (DR/DZ). LCMACvC = lung cancer microarray case vs. control; LCMANvF = lung cancer microarray smoking never vs. former; LCMAFvC = lung cancer microarray smoking former vs. current; LCMANvC = lung cancer microarray smoking never vs. current; LCMAMorph = lung cancer microarray morphology small cell carcinoma vs. adenocarcinoma.

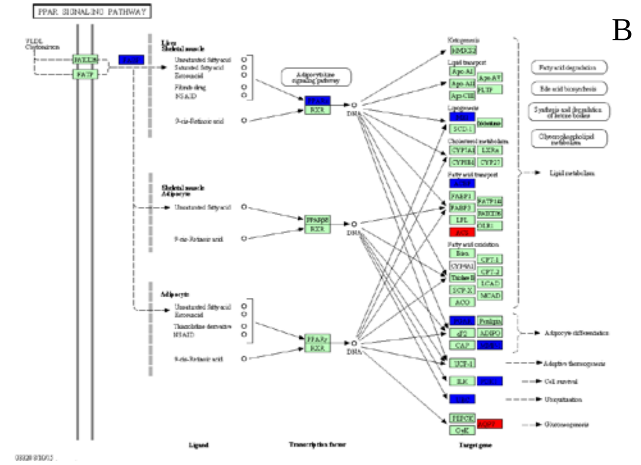


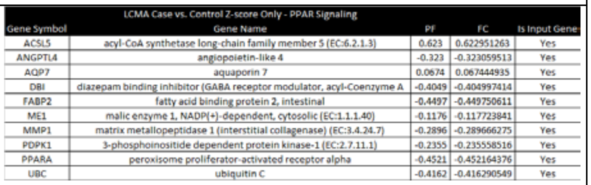
	LCMACvC					LCMASmokeNvF					LCMASmokeFvC					LCMASmokeNvC					LCMAMorph				
Pathway	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ
Adherens junction	6	24	24	26	26	3	25	25	26	26	8	24	24	26	73	5	25	25	26	26	5	25	25	26	26
Apoptosis	18	32	32	33	33	9	33	33	33	33	17	32	32	33	85	14	33	33	33	33	19	33	33	33	33
Asthma	9	10	10	13	13	3	11	11	13	13	2	10	10	13	27	2	11	11	13	13	5	11	11	13	13
Base excision repair	2	10	10	9	9	7	10	10	9	9	1	10	10	9	31	5	10	10	9	9		10	10	9	9
Cell cycle	12	44	44	44	44	17	44	44	44	44	15	44	44	44	96	16	44	44	44	44	15	44	44	44	44
Cytokine-cytokine receptor interaction	27	108	108	115	115	17	110	110	116	116	36	108	108	115	235	16	110	110	116	116	26	110	110	116	116
DNA replication	2	7	7	6	6	11	7	7	6	6	2	7	7	6	34	6	7	7	6	6	2	7	7	6	6
ECM-receptor interaction	6	16	16	17	17	4	16	16	17	17	5	16	16	17	82	5	16	16	17	17	6	16	16	17	17
Focal adhesion	21	59	59	59	59	13	60	60	59	59	24	59	59	59	192	18	60	60	59	59	27	60	60	59	59
Jak-STAT signaling pathway	12	44	44	51	51	6	47	47	51	51	18	44	44	51	140	6	47	47	51	51	17	47	47	51	51
MAPK signaling pathway	32	67	67	74	74	19	72	72	74	74	45	67	67	74	246	34	72	72	74	74	39	72	72	74	74
Mismatch repair		9	9	9	9	7	9	9	9	9	2	9	9	9	23	4	9	9	9	9	1	9	9	9	9
mTOR signaling pathway	2	16	16	17	17	3	16	16	17	17	8	16	16	17	46	8	16	16	17	17	5	16	16	17	17
Non-small cell lung cancer	8	27	27	28	28	2	28	28	28	28	10	27	27	28	53	6	28	28	28	28	6	28	28	28	28
Nucleotide excision repair	5	12	12	10	10	9	12	12	11	11	6	12	12	10	43	9	12	12	11	11	5	12	12	11	11
p53 signaling pathway	11	30	30	32	32		31	31	32	32	9	30	30	32	61	15	31	31	32	32	10	31	31	32	32
Pathways in cancer	31	122	122	134	134	26	128	128	134	134	37	122	122	134	311	36	128	128	134	134	42	128	128	134	134
PPAR signaling pathway	6	10	10	13	13	7	10	10	13	13	5	10	10	13	62	6	10	10	13	13	8	10	10	13	13
Small cell lung cancer	11	30	30	34	34	8	31	31	34	34	11	30	30	34	86	14	31	31	34	34	14	31	31	34	34
TGF-beta signaling pathway	9	33	33	33	33	4	33	33	33	33	13	33	33	33	81	11	33	33	33	33	14	33	33	33	33
VEGF signaling pathway	10	23	23	25	25	3	23	23	25	25	15	23	23	25	68	9	23	23	25	25	10	23	23	25	25
Wnt signaling pathway	16	28	28	26	26	11	29	29	26	26	20	28	28	26	134	16	29	29	26	26	20	29	29	26	26



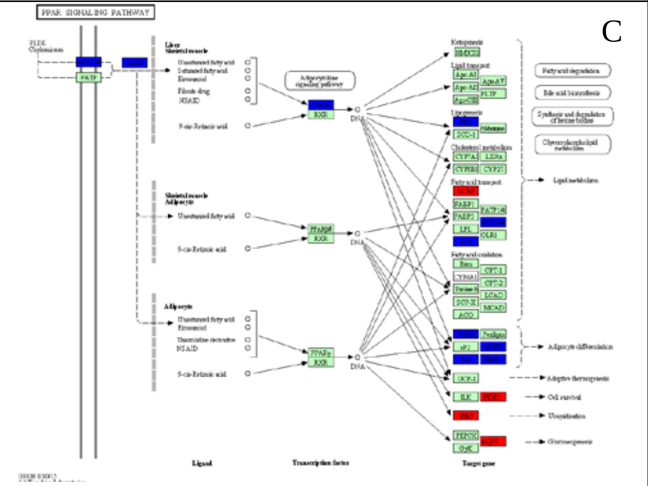


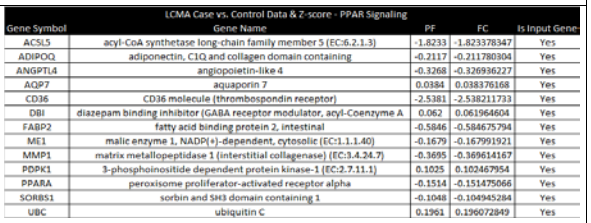
LCMA Case vs. Control Data Only - PPAR Signaling				
Gene Symbol	Gene Name	PF	FC	Is Input Ger
ACSL1	acyl-CoA synthetase long-chain family member 1 (EC:6.2.1.3)	2.651	2.651	Yes
ACSL4	acyl-CoA synthetase long-chain family member 4 (EC:6.2.1.3)	4.564	4.564	Yes
ACSL5	acyl-CoA synthetase long-chain family member 5 (EC:6.2.1.3)	-2.9269	-2.927	Yes
CD36	CD36 molecule (thrombospondin receptor)	5.328	5.328	Yes
GK	glycerol kinase (EC:2.7.1.30)	2.395	2.395	Yes
ILK	integrin-linked kinase (EC:2.7.11.1)	2.821	2.821	Yes





LCMA Case vs. Control Z-score Only - PPAR Signaling				
Gene Symbol	Gene Name	PF	FC	Is Input Gene-
ACSL5	acyl-CoA synthetase long-chain family member 5 (EC:6.2.1.3)	0.623	0.622951263	Yes
ANGPTL4	angiopoietin-like 4	-0.323	-0.323059513	Yes
AQP7	aquaporin 7	0.0674	0.067444935	Yes
DBI	diazepam binding inhibitor (GABA receptor modulator, acyl-Coenzyme A	-0.4049	-0.404997414	Yes
FABP2	fatty acid binding protein 2, intestinal	-0.4497	-0.449750611	Yes
ME1	malic enzyme 1, NADP(+)-dependent, cytosolic (EC:1.1.1.40)	-0.1176	-0.117723841	Yes
MMP1	matrix metallopeptidase 1 (interstitial collagenase) (EC:3.4.24.7)	-0.2896	-0.289666275	Yes
PDPK1	3-phosphoinositide dependent protein kinase-1 (EC:2.7.11.1)	-0.2355	-0.235558516	Yes
PPARA	peroxisome proliferator-activated receptor alpha	-0.4521	-0.452164376	Yes
UBC	ubiquitin C	-0.4162	-0.416290549	Yes





LCMA Case vs. Control Data & Z-score - PPAR Signaling				
Gene Symbol	Gene Name	PF	FC	Is Input Gene-
ACSL5	acyl-CoA synthetase long-chain family member 5 (EC:6.2.1.3)	-1.8233	-1.823378347	Yes
ADIPOQ	adiponectin, C1Q and collagen domain containing	-0.2117	-0.211780304	Yes
ANGPTL4	angiopoietin-like 4	-0.3268	-0.326936227	Yes
AQP7	aquaporin 7	0.0384	0.038376168	Yes
CD36	CD36 molecule (thrombospondin receptor)	-2.5381	-2.538211733	Yes
DBI	diazepam binding inhibitor (GABA receptor modulator, acyl-Coenzyme A	0.062	0.061964604	Yes
FABP2	fatty acid binding protein 2, intestinal	-0.5846	-0.584675794	Yes
ME1	malic enzyme 1, NADP(+)-dependent, cytosolic (EC:1.1.1.40)	-0.1679	-0.167991921	Yes
MMP1	matrix metallopeptidase 1 (interstitial collagenase) (EC:3.4.24.7)	-0.3695	-0.369614167	Yes
PDPK1	3-phosphoinositide dependent protein kinase-1 (EC:2.7.11.1)	0.1025	0.102467954	Yes
PPARA	peroxisome proliferator-activated receptor alpha	-0.1514	-0.151475066	Yes
SORBS1	sorbin and SH3 domain containing 1	-0.1048	-0.104945284	Yes
UBC	ubiquitin C	0.1961	0.196072849	Yes

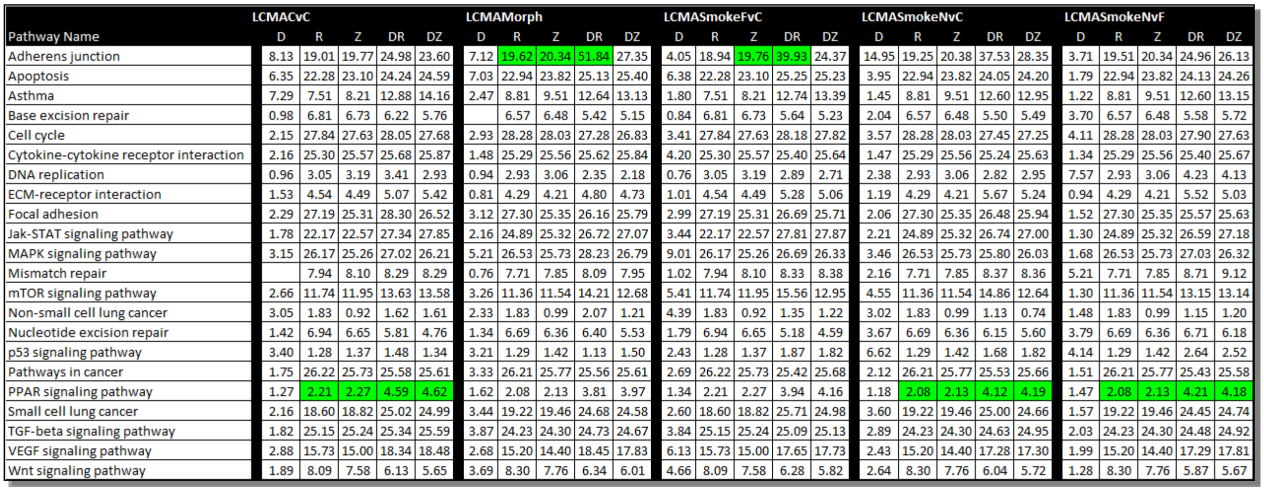
Figure 23. KEGG PPAR Signaling Pathway diagrams and input gene tables created using lung cancer microarray case vs. control. A = data-only (D); B = z-score only (Z); C = data & z-score (DZ). Input genes highlighted red = upregulated and blue = downregulated. PF = perturbation factor; FC = fold-change.

The KEGG PPAR Signaling Pathway diagram contains 70 total genes. As can be seen in Figure 23 when the data only (D) genelist is submitted (Figure 23, A), six input genes appear in the output PPAR Signaling Pathway (ACSL1/*acyl-CoA synthetase long-chain family member 1*, ACSL4/*acyl-CoA synthetase long-chain family member 4*, ACSL5/*acyl-CoA synthetase long-chain family member 5*, CD36, GK/*glycerol kinase*, ILK/*integrin-linked kinase*). When the literature mining z-score (Z) genelist is used as input (Figure 23, B), ten input genes appear in the output PPAR Signaling Pathway (ACSL5, ANGPTL4/*angiopoietin-like 4*, AQP7/*aquaporin 7*, DBI/*diazepam binding inhibitor (GABA receptor modulator, acyl-CoA binding protein*), FABP2/*fatty acid binding protein 2*, ME1/*malic enzyme 1, NADP(+)-dependent, cytosolic*, MMP1*/*matrix metallopeptidase 1*, PDPK1/*3-phosphoinositide-dependent protein kinase 1*, PPARA/*peroxisome proliferator-activated receptor alpha*, and UBC/*ubiquitin C*). Finally, when the combined DZ genelist is used as input (Figure 23, C), 13 input genes appear in the output PPAR Signaling Pathway (ACSL5, ADIPOQ*/*adiponectin, C1Q and collagen domain-containing*, ANGPTL4, AQP7, CD36, DBI, FABP2, ME1, MMP1*, PDPK1, PPARA, SORBS1/*sorbin and SH3 domain containing 1*, UBC).

* Indicates gene/protein found in list of known lung cancer biomarkers Table 5

Again, the up-regulation or down-regulation of the genes is not of great importance for this exercise. The major point of emphasis here is the enriched input dataset, which produces more-relevant results, in the output pathways, when data and prior knowledge are combined.

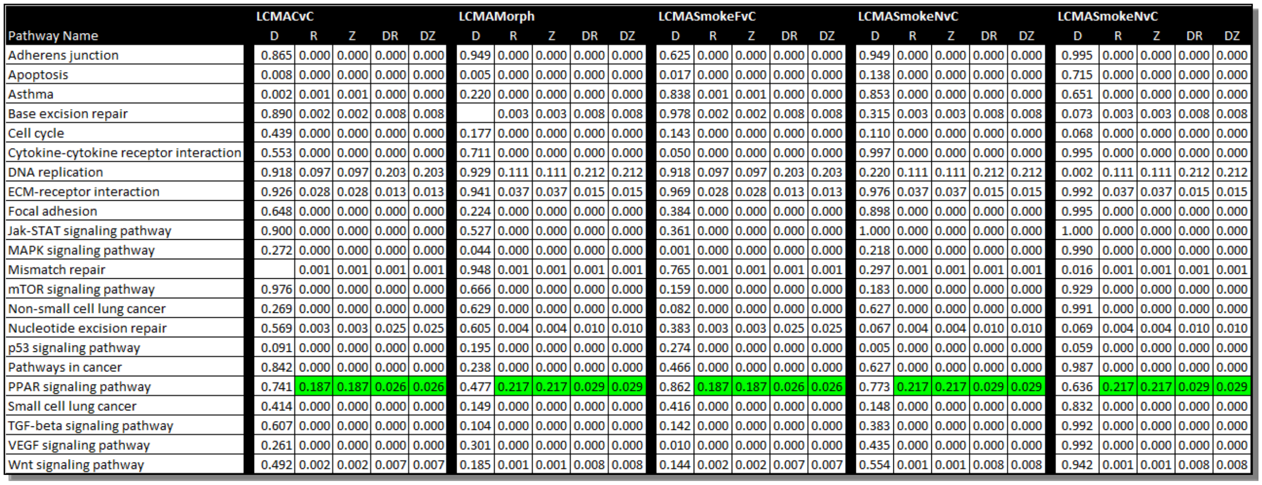
Table 16. Impact factors from lung cancer datasets. D = data only; R = ratio only; Z = z-score only; DR = data & ratio; DZ = data & z-score. Green highlights cases where the impact factor increased by a factor of two or greater. LCMACvC = lung cancer microarray case vs. control; LCMANvF = lung cancer microarray smoking never vs. former; LCMAFvC = lung cancer microarray smoking former vs. current; LCMANvC = lung cancer microarray smoking never vs. current; LCMAMorph = lung cancer microarray morphology small cell carcinoma vs. adenocarcinoma.



	LCMACvC					LCMAMorph					LCMASmokeFvC					LCMASmokeNvC					LCMASmokeNvF				
Pathway Name	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ
Adherens junction	8.13	19.01	19.77	24.98	23.60	7.12	19.62	20.34	51.84	27.35	4.05	18.94	19.76	39.93	24.37	14.95	19.25	20.38	37.53	28.35	3.71	19.51	20.34	24.96	26.13
Apoptosis	6.35	22.28	23.10	24.24	24.59	7.03	22.94	23.82	25.13	25.40	6.38	22.28	23.10	25.25	25.23	3.95	22.94	23.82	24.05	24.20	1.79	22.94	23.82	24.13	24.26
Asthma	7.29	7.51	8.21	12.88	14.16	2.47	8.81	9.51	12.64	13.13	1.80	7.51	8.21	12.74	13.39	1.45	8.81	9.51	12.60	12.95	1.22	8.81	9.51	12.60	13.15
Base excision repair	0.98	6.81	6.73	6.22	5.76		6.57	6.48	5.42	5.15	0.84	6.81	6.73	5.64	5.23	2.04	6.57	6.48	5.50	5.49	3.70	6.57	6.48	5.58	5.72
Cell cycle	2.15	27.84	27.63	28.05	27.68	2.93	28.28	28.03	27.28	26.83	3.41	27.84	27.63	28.18	27.82	3.57	28.28	28.03	27.45	27.25	4.11	28.28	28.03	27.90	27.63
Cytokine-cytokine receptor interaction	2.16	25.30	25.57	25.68	25.87	1.48	25.29	25.56	25.62	25.84	4.20	25.30	25.57	25.40	25.64	1.47	25.29	25.56	25.24	25.63	1.34	25.29	25.56	25.40	25.67
DNA replication	0.96	3.05	3.19	3.41	2.93	0.94	2.93	3.06	2.35	2.18	0.76	3.05	3.19	2.89	2.71	2.38	2.93	3.06	2.82	2.95	7.57	2.93	3.06	4.23	4.13
ECM-receptor interaction	1.53	4.54	4.49	5.07	5.42	0.81	4.29	4.21	4.80	4.73	1.01	4.54	4.49	5.28	5.06	1.19	4.29	4.21	5.67	5.24	0.94	4.29	4.21	5.52	5.03
Focal adhesion	2.29	27.19	25.31	28.30	26.52	3.12	27.30	25.35	26.16	25.79	2.99	27.19	25.31	26.69	25.71	2.06	27.30	25.35	26.48	25.94	1.52	27.30	25.35	25.57	25.63
Jak-STAT signaling pathway	1.78	22.17	22.57	27.34	27.85	2.16	24.89	25.32	26.72	27.07	3.44	22.17	22.57	27.81	27.87	2.21	24.89	25.32	26.74	27.00	1.30	24.89	25.32	26.59	27.18
MAPK signaling pathway	3.15	26.17	25.26	27.02	26.21	5.21	26.53	25.73	28.23	26.79	9.01	26.17	25.26	26.69	26.33	3.46	26.53	25.73	25.80	26.03	1.68	26.53	25.73	27.03	26.32
Mismatch repair		7.94	8.10	8.29	8.29	0.76	7.71	7.85	8.09	7.95	1.02	7.94	8.10	8.33	8.38	2.16	7.71	7.85	8.37	8.36	5.21	7.71	7.85	8.71	9.12
mTOR signaling pathway	2.66	11.74	11.95	13.63	13.58	3.26	11.36	11.54	14.21	12.68	5.41	11.74	11.95	15.56	12.95	4.55	11.36	11.54	14.86	12.64	1.30	11.36	11.54	13.15	13.14
Non-small cell lung cancer	3.05	1.83	0.92	1.62	1.61	2.33	1.83	0.99	2.07	1.21	4.39	1.83	0.92	1.35	1.22	3.02	1.83	0.99	1.13	0.74	1.48	1.83	0.99	1.15	1.20
Nucleotide excision repair	1.42	6.94	6.65	5.81	4.76	1.34	6.69	6.36	6.40	5.53	1.79	6.94	6.65	5.18	4.59	3.67	6.69	6.36	6.15	5.60	3.79	6.69	6.36	6.71	6.18
p53 signaling pathway	3.40	1.28	1.37	1.48	1.34	3.21	1.29	1.42	1.13	1.50	2.43	1.28	1.37	1.87	1.82	6.62	1.29	1.42	1.68	1.82	4.14	1.29	1.42	2.64	2.52
Pathways in cancer	1.75	26.22	25.73	25.58	25.61	3.33	26.21	25.77	25.56	25.61	2.69	26.22	25.73	25.42	25.68	2.12	26.21	25.77	25.53	25.66	1.51	26.21	25.77	25.43	25.58
PPAR signaling pathway	1.27	2.21	2.27	4.59	4.62	1.62	2.08	2.13	3.81	3.97	1.34	2.21	2.27	3.94	4.16	1.18	2.08	2.13	4.12	4.19	1.47	2.08	2.13	4.21	4.18
Small cell lung cancer	2.16	18.60	18.82	25.02	24.99	3.44	19.22	19.46	24.68	24.58	2.60	18.60	18.82	25.71	24.98	3.60	19.22	19.46	25.00	24.66	1.57	19.22	19.46	24.45	24.74
TGF-beta signaling pathway	1.82	25.15	25.24	25.34	25.59	3.87	24.23	24.30	24.73	24.67	3.84	25.15	25.24	25.09	25.13	2.89	24.23	24.30	24.63	24.95	2.03	24.23	24.30	24.48	24.92
VEGF signaling pathway	2.88	15.73	15.00	18.34	18.48	2.68	15.20	14.40	18.45	17.83	6.13	15.73	15.00	17.65	17.73	2.43	15.20	14.40	17.28	17.30	1.99	15.20	14.40	17.29	17.81
Wnt signaling pathway	1.89	8.09	7.58	6.13	5.65	3.69	8.30	7.76	6.34	6.01	4.66	8.09	7.58	6.28	5.82	2.64	8.30	7.76	6.04	5.72	1.28	8.30	7.76	5.87	5.67

Similar to section 4.3.2.1 results, for most pathways examined, a steady progression exists from left to right with D genelists producing the smallest impact factor; the R/Z genelists producing larger impact factors than D; and DR/DZ combined genelists producing the largest impact factors, in general. In Table 16, green highlights indicate where the impact factor was greater by at least a factor of two for the DR/DZ than it was for the R/Z input genelists.

Table 17. P-values from lung cancer datasets. D = data only; R = ratio only; Z = z-score only; DR = data & ratio; DZ = data & z-score. Green highlights cases where the p-values went from non-significant to significant in going from left (R/Z only) to right (DR/DZ). LCMACvC = lung cancer microarray case vs. control; LCMANvF = lung cancer microarray smoking never vs. former; LCMAFvC = lung cancer microarray smoking former vs. current; LCMANvC = lung cancer microarray smoking never vs. current; LCMAMorph = lung cancer microarray morphology small cell carcinoma vs. adenocarcinoma.



	LCMACvC					LCMAMorph					LCMASmokeFvC					LCMASmokeNvC					LCMASmokeNvC				
Pathway Name	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ	D	R	Z	DR	DZ
Adherens junction	0.865	0.000	0.000	0.000	0.000	0.949	0.000	0.000	0.000	0.000	0.625	0.000	0.000	0.000	0.000	0.949	0.000	0.000	0.000	0.000	0.995	0.000	0.000	0.000	0.000
Apoptosis	0.008	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.000	0.138	0.000	0.000	0.000	0.000	0.715	0.000	0.000	0.000	0.000
Asthma	0.002	0.001	0.001	0.000	0.000	0.220	0.000	0.000	0.000	0.000	0.838	0.001	0.001	0.000	0.000	0.853	0.000	0.000	0.000	0.000	0.651	0.000	0.000	0.000	0.000
Base excision repair	0.890	0.002	0.002	0.008	0.008		0.003	0.003	0.008	0.008	0.978	0.002	0.002	0.008	0.008	0.315	0.003	0.003	0.008	0.008	0.073	0.003	0.003	0.008	0.008
Cell cycle	0.439	0.000	0.000	0.000	0.000	0.177	0.000	0.000	0.000	0.000	0.143	0.000	0.000	0.000	0.000	0.110	0.000	0.000	0.000	0.000	0.068	0.000	0.000	0.000	0.000
Cytokine-cytokine receptor interaction	0.553	0.000	0.000	0.000	0.000	0.711	0.000	0.000	0.000	0.000	0.050	0.000	0.000	0.000	0.000	0.997	0.000	0.000	0.000	0.000	0.995	0.000	0.000	0.000	0.000
DNA replication	0.918	0.097	0.097	0.203	0.203	0.929	0.111	0.111	0.212	0.212	0.918	0.097	0.097	0.203	0.203	0.220	0.111	0.111	0.212	0.212	0.002	0.111	0.111	0.212	0.212
ECM-receptor interaction	0.926	0.028	0.028	0.013	0.013	0.941	0.037	0.037	0.015	0.015	0.969	0.028	0.028	0.013	0.013	0.976	0.037	0.037	0.015	0.015	0.992	0.037	0.037	0.015	0.015
Focal adhesion	0.648	0.000	0.000	0.000	0.000	0.224	0.000	0.000	0.000	0.000	0.384	0.000	0.000	0.000	0.000	0.898	0.000	0.000	0.000	0.000	0.995	0.000	0.000	0.000	0.000
Jak-STAT signaling pathway	0.900	0.000	0.000	0.000	0.000	0.527	0.000	0.000	0.000	0.000	0.361	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
MAPK signaling pathway	0.272	0.000	0.000	0.000	0.000	0.044	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.218	0.000	0.000	0.000	0.000	0.990	0.000	0.000	0.000	0.000
Mismatch repair		0.001	0.001	0.001	0.001	0.948	0.001	0.001	0.001	0.001	0.765	0.001	0.001	0.001	0.001	0.297	0.001	0.001	0.001	0.001	0.016	0.001	0.001	0.001	0.001
mTOR signaling pathway	0.976	0.000	0.000	0.000	0.000	0.666	0.000	0.000	0.000	0.000	0.159	0.000	0.000	0.000	0.000	0.183	0.000	0.000	0.000	0.000	0.929	0.000	0.000	0.000	0.000
Non-small cell lung cancer	0.269	0.000	0.000	0.000	0.000	0.629	0.000	0.000	0.000	0.000	0.082	0.000	0.000	0.000	0.000	0.627	0.000	0.000	0.000	0.000	0.991	0.000	0.000	0.000	0.000
Nucleotide excision repair	0.569	0.003	0.003	0.025	0.025	0.605	0.004	0.004	0.010	0.010	0.383	0.003	0.003	0.025	0.025	0.067	0.004	0.004	0.010	0.010	0.069	0.004	0.004	0.010	0.010
p53 signaling pathway	0.091	0.000	0.000	0.000	0.000	0.195	0.000	0.000	0.000	0.000	0.274	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.059	0.000	0.000	0.000	0.000
Pathways in cancer	0.842	0.000	0.000	0.000	0.000	0.238	0.000	0.000	0.000	0.000	0.466	0.000	0.000	0.000	0.000	0.627	0.000	0.000	0.000	0.000	0.987	0.000	0.000	0.000	0.000
PPAR signaling pathway	0.741	0.187	0.187	0.026	0.026	0.477	0.217	0.217	0.029	0.029	0.862	0.187	0.187	0.026	0.026	0.773	0.217	0.217	0.029	0.029	0.636	0.217	0.217	0.029	0.029
Small cell lung cancer	0.414	0.000	0.000	0.000	0.000	0.149	0.000	0.000	0.000	0.000	0.416	0.000	0.000	0.000	0.000	0.148	0.000	0.000	0.000	0.000	0.832	0.000	0.000	0.000	0.000
TGF-beta signaling pathway	0.607	0.000	0.000	0.000	0.000	0.104	0.000	0.000	0.000	0.000	0.142	0.000	0.000	0.000	0.000	0.383	0.000	0.000	0.000	0.000	0.992	0.000	0.000	0.000	0.000
VEGF signaling pathway	0.261	0.000	0.000	0.000	0.000	0.301	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.435	0.000	0.000	0.000	0.000	0.992	0.000	0.000	0.000	0.000
Wnt signaling pathway	0.492	0.002	0.002	0.007	0.007	0.185	0.001	0.001	0.008	0.008	0.144	0.002	0.002	0.007	0.007	0.554	0.001	0.001	0.008	0.008	0.942	0.001	0.001	0.008	0.008

Similarly to sections 4.3.2.1 and 4.3.2.2 results, for most pathways examined, a steady progression exists from left to right with D genelists producing the least significant pathway p-values; the R/Z genelists producing more significant pathway p-values than D; and DR/DZ combined genelists producing the most significant pathway p-values, in general. In Table 17, green highlights cases where pathway p-values for R/Z produced genelists could be considered non-significant, but pathway p-values for the DR/DZ produced genelists could be considered significant.

The idea that incorporation of prior information enhances pathway analysis results (Claim 2, part C), is not without merit. It is shown above that with few exceptions, consistent patterns emerge in quality metrics when analyzing the results. An increase in the number of input genes and impact factor, with a corresponding decrease in p-values, is relatively consistent when comparing pathway results from data-alone, ratio/z-scores, and data + (ratio/z-score), respectively.

Using the combined method of prior knowledge and experimental data array results returned more genes compared to either method individually. This may be beneficial to researchers as genes of interest may be missed using only experimental results or only literature mining results.

The following factors support the idea that (Claim 2, part C; experimentation 4.3.1 - 4.3.2) incorporation of prior information enhances pathway analysis by identifying more input genes in disease-relevant pathways: 1) in almost all of the breast and lung cancer datasets examined, the number of input genes in pathways increased with the incorporation of prior information (experimentation 4.3.1.1 & 4.3.2.1); 2) the impact factors increased with the incorporation of prior information (experimentation 4.3.1.2 & 4.3.2.2) ; and 3) the p-values decreased (became more significant) with the incorporation of prior information (experimentation 4.3.1.3 & 4.3.2.3).

4.4 SUMMARY

The first claim in this dissertation is to determine whether literature mining is a sufficient method of obtaining prior information for use in modeling and pathway analysis. The arguments that support this claim can be found in sections 4.1.2 – 4.1 5.

The second claim is to investigate whether incorporating prior knowledge into modeling and pathway analysis enhances results compared to using experimental data only. The arguments that support this claim can be found in the following sections:

a) Incorporation of prior information does not degrade modeling performance, on average. Section 4.2.1 – 4.2.4 strongly supported this claim.

b) Analyzing the attributes used to build the best-performing models leads to new biological relationships being uncovered. Section 4.2.5 strongly supports this claim.

c) Incorporation of prior information enhances pathway analysis results by identifying more input genes in disease-relevant pathways. Sections 4.3.1.1 and 4.3.2.1 strongly support this claim as there is robust evidence for pathway involvement in the disease of interest based on statistically significant standard measures such as impact factors and p-values.

5.0 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

The sections below discuss the conclusions, limitations, and future work of this dissertation. The benefits of using the KEDA framework are presented in section 5.1. Limitations and assumptions as they pertain to this dissertation are identified in section 5.2. Lastly, directions for future work in KEDA-related areas are addressed in section 5.3.

5.1 CONCLUSIONS

The age of personalized medicine is upon us. The field of pharmacogenomics is ready to explode by predicting via DNA analysis, individual patients that will benefit from given medications, and which will not. Moreover, biomarkers of all types are needed for disease prediction, diagnosis, and treatment. New methods for obtaining disease biomarkers are desperately needed as current approaches are very time-consuming, and are not producing enough reliable biomarkers at a sufficient rate.

Literature mining can aid in the search for new biomarkers as it has been shown to be an appropriate method of obtaining prior scientific knowledge. In this work, the KEDA framework, which utilizes a semi-automated literature mining method to examine PubMed abstracts and establish a list of putative biomarkers for breast and lung cancer, is described. Biological entity mentions in the abstracts were tallied and used to calculate scores for use as prior knowledge. Counts were used to determine biomarker specificity for a biofluid, rank the biomarkers by score, and to establish an error rate to predict the accuracy of novel future discoveries.

Disease modeling can also aid in the search for new biomarkers by emulating biological systems. The literature mining counts were converted to prior probabilities, and incorporated into publicly available lung and breast cancer datasets in modeling exercises. While neither improvement nor degradation of modeling performance measures was observed in models incorporating prior information, examining the attributes of the best-performing models uncovered a new potential interaction.

Pathway analysis can also aid in the search for new biomarkers by presenting researchers with maps of biological processes. Literature mining prior probabilities were combined with experimental data and incorporated into pathway analysis. The combination of prior information and data analysis results produced far superior increases in performance measures such as impact factor and p-value, when compared to data analysis results and/or prior information alone. This is a significant finding as the use of pathway analysis results as prior knowledge is well-documented, however documentation of the use of prior knowledge as input into pathway analysis had not been found.

This dissertation has provided a significant contribution to the scientific and bioinformatics community. The KEDA literature mining results presented in this work can be used by breast and lung cancer researchers as a valuable source of information. The KEDA computational tools can be extrapolated to obtain prior knowledge for any disease, provided the appropriate keyword are used in the initial search. Examining the attributes used to build the best-performing models provides a different approach to biomarker discovery, and incorporating prior knowledge with experimental results has been shown to greatly improve pathway analysis findings.

5.2 LIMITATIONS

The results presented in this dissertation show that the KEDA framework is a useful approach for biomarker discovery. That being said, the conclusions must be drawn carefully based on the following limitations of the study:

a) Aliases of biological entities continue to complicate literature-mining findings. For example, ceacam5 and ceacam8 were both identified with the CEA alias. Even the most up to date dictionary may still not contain every alias used by the scientific community for a given entity.

b) Negation was not addressed in this work. Some biomarker mentions in positive abstract sets may actually be in a negative context.

c) A biomarker associated with any other disease (even another type of cancer or a non-specific cancer marker) might negate a positive finding for breast or lung cancer due to the scoring method used.

d) Verification databases may be far from exhaustive. This could be a reason why the list of known/significant biomarkers is not as large as expected. Also, in limiting our search to 'breast or lung cancer specific' markers, many biomarkers common to several or all types of cancers may have accidentally been omitted from the study.

e) Due to lack of access to full text articles, only abstracts were examined. Access to full text of scientific articles remains a limiting factor for many researchers.

f) The BRL algorithm works optimally when discretization is employed. However, discretization was not implemented into the described modeling process as much information would be lost trying to convert experimental and prior knowledge values into discrete ones.

g) Only one algorithm was used for each KEDA process. Algorithms used were chosen for ease-of-use or familiarity. Other available algorithms may exist which may improve results.

h) For modeling, the initial analysis only examined results from one dataset. Therefore, the modeling results presented cannot be generalized further.

5.3 FUTURE WORK

The work described in this dissertation exposes several different avenues to biomarker discovery. The following suggestions could be pursued for future research:

5.3.1 Informatics

Informatic resources and tools to be utilized for future research are discussed in the following sections.

5.3.1.1 Resources *A database of disease/biofluid-associated ratios/z-scores*

For others to benefit from the work presented above, the information must be readily accessible. Online access to a database containing tables of putative disease biomarkers obtained using the KEDA framework would be extremely beneficial to other researchers. This database could serve as a valuable resource for scientists performing modeling or pathway analysis where prior information may be warranted; or as a starting point of genes/proteins that may be implicated in a given disease.

Verification databases

The databases of known cancer biomarkers used in this dissertation may be far from exhaustive, as well as too specific. Constant updating of the list of known biomarkers may improves KEDA results. These updated lists should be posted for others to use to avoid repetitive efforts.

5.3.1.2 Tools Literature mining, modeling, and pathway analysis tools and notions are presented in the following sections.

5.3.1.2.1 Literature mining *Further automation of the semi-automated process*

While the process performs well currently, improvements where manual intervention is currently required would speed up the process and reduce the chance for manual errors. Further automation could be added in searching for and downloading of relevant abstracts, parsing the abstract file to create an individual file per abstract, and calculating z-scores and ratios from final tallies. Ideally, one should be able to run PittCAP, enter the keywords from its execution window, and the algorithm should automatically perform all duties and return the list of putative disease markers with their z-scores and ratios.

Negation

Negation was not addressed in this work. An assumption was made that in the short amount of space allotted for an abstract text, writers would not write about negative findings. However that may not necessarily be the case and thus some of the biomarker mentions in positive abstract sets may actually be in a negative context. This is a well-known issue in the literature mining field, and is not easily fixed. Time and effort will need to be invested to tend to these issues. Adapting the literature mining component to account for negation will produce more accurate prior estimates.

Applying the KEDA framework to other datasets and diseases

The KEDA framework can be adapted to obtain and process abstracts for any disease. Many diseases exist where prior information from literature mining may be useful for disease modeling and pathway analysis.

Adapting the KEDA framework to use other resources

While one implementation of the KEDA framework was described above, KEDA can be can be adapted as many other implementation possibilities exist. Alternatives can be implemented for the abstract resource in place of PubMed; the tagger in place of Abner; the data source in place of Gene Expression Omnibus; the classification modeling algorithm in place of the Bayesian Rule Learner; and the pathway analysis algorithm in place of Pathway Express.

5.3.1.2.2 Modeling *Follow-up with BRL*

While a good deal of time and effort was invested in disease modeling using BRL, the results were less than expected. Classification accuracy was expected show at-least minimal improvement in accuracy and other quality measures. However, that was not observed in this work, or in several other in-house efforts which utilized structure priors. Further investigation is needed to determine why classification accuracy obtained from this work was not greater.

Determine how use of priors alters the search order

The fact that the SOD2-CCL5 relationship was observed in models utilizing informed priors and uniform priors but not found when no prior was used speaks to the fact that the use of priors changes the order of the search. This is an interesting occurrence that should be investigated further.

5.3.1.2.3 Pathway analysis *Pathway analysis from modeling results*

It would be interesting to perform pathway analyses based on modeling results. In this situation, the input file for Pathway Express could be limited to the attributes found from the best-performing models. Another interesting concept would be to search for modeling attributes common to more than one pathway.

5.3.2 Laboratory verification

Confirmatory laboratory experimentation and further research

Informatics can only take the biomarker discovery so far. Eventually, in order to verify potential findings, actual wet-lab experimentation will need to be performed. For example, experimentation seeking to identify a direct relationship between SOD2 and CCL5 is now warranted. Follow-up experimentation of pathway analysis findings may also prove beneficial.

SOD2/CCL5 follow-up research

It appears a possibility that a relationship may exist between SOD2 and CCL5 in lung cancer. While a documented direct relationship was not identified, several facts seem to support the theory due to SOD2 being an oxidative agent which aids in ridding the body of toxins, and CCL5 being a chemokine involved in the immunoregulatory and inflammatory processes.

Shah et al. (2011) studied the SOD/CCL5 relationship in the chronic inflammatory autoimmune disorders, systemic lupus and arthritis. In both diseases, SOD's anti-oxidant activity was significantly reduced, and antioxidant molecules showed a negative association with CCL5. It was concluded that excessive production of ROS disturbs redox status and can modulate the expression of inflammatory chemokines leading to inflammatory processes, and affecting tissue damage in autoimmune diseases.

Kumar *et al.* (2012) studied the microRNA miR-302 and how it regulates transition between cellular quiescence and proliferation. SOD2, is an antioxidant enzyme well-known to regulate cellular reactive oxygen species levels by converting superoxide into hydrogen peroxide. Overexpression of SOD2 is believed to increase H_2O_2 levels which may lead to ROS-sensitive regulation of CCL5 mRNA levels. SOD2 expression increases in quiescence, and it is suggested that SOD2-signaling activates CCL5 expression. Moreover, miR-302 levels decreased significantly by overexpression of SOD2, as well as from ionizing radiation, increasing the CCL5 mRNA levels.

Kim *et al.* (2014) also examined miR-302's regulation of cell proliferation and cell-cycle progression in adipose tissue-derived mesenchymal stem cells. They found that miR-302 induces cell proliferation and inhibits oxidant-induced cell death through a reduction in CCL5 expression.

DiRenzo *et al.* (2014) also discussed SOD2 and CCL5. They discussed that antioxidants reduced CCL5 mRNA expression, and mentioned the Human Oxidative Stress Pathway. The SOD2 gene related to antioxidant defense antioxidants effectively suppressed CCL5 mRNA expression. While more experimental evidence is required for confirmation, the possibility of a relationship in lung cancer persists.

APPENDIX A

PYTHON SCRIPTS FOR KEDA TEXT-MINING COMPONENT

PYTHON SCRIPT ‘RANDABSTRACTMAKERV2.0’ CODE

```
# Rand-Neg Abstract Maker2.0

import math
import string

#********************** Function definition *****************************
#************************** Main Program *********************************
def main():
    print "Rand-Neg Abstract Maker1.0"
    print "The DirList should contain a list of abstract files to be used"
    print

    NewAbFile = []

    # Read in abstract file
    AbFile = raw_input("Enter the abstract file: (.txt)")
    outfileB = open("List.txt", 'w')
    print AbFile
    print "Working..."
    infile = open(AbFile, 'r')
    things = 0
    SepAb = []
    for line in infile:
        line = string.split(line)
        NewAbFile.append(line)       # append lines to new abstract file
    infile.close()

    # Create new abstract files
    NewAbOut = []
    for line in NewAbFile:
        key = line[0:1]
        key = str(key)
        ID = line[1:2]
        ID = str(ID)
        NewAbOut.append(line)
        if key == "['PMID:']":
            name = (ID + ".txt")
            outfile = open(name, 'w')
            for line in NewAbOut:
                outfile.write(str(line))
                outfile.close
                NewAbOut = []
            outfileB.write(str(name)+'\n')

    outfileB.close
    print "All done."

            main()
```

PYTHON SCRIPT ‘PITTCAPV3.0’ CODE

```
#PittCAPv3
# CancerAbstractPicker

import math
import string

#********************** Function definition *****************************
#************************** Main Program ********************************
def main():
    print "CancerAbstractPicker"
    print "The DirList should contain a list of abstract files to be used"
    print

    DIRList = []

    # Read in list of abstracts
    AbDirFile = raw_input("Enter the list of abstracts: (.txt)")
    Dict = raw_input("Enter the dictionary filename: (.txt)")
    print AbDirFile
    infile = open(AbDirFile, 'r')
    things = 0
    for line in infile:
        line = string.split(line)
        DIRList.append(line)      # list each abstract file
        things = things + 1
```

```
infile.close()

# Confirm number of abstract files
print ("There is(are) ", things, " abstracts in the file(s).")

# Open individual abstract files
j = 0
outfile = open('PickerResults.txt', 'w')
keep = []
for report in DIRList:
   A = []
   report = str(report)
   report = report.rstrip("\n")
   report = report.strip("[]")
   report = report.strip("'")
   report = report.strip("''")
   A = open(report, 'r')
   print report

   # Parse the file
   flag = 0
   for line in A:
       line = str(line)
       line = string.split(line)
       ID = "PMID:"+ report[2:-11]
       for item in line:
         if item == '<PROTEIN>':
            flag = 1
         if item == '<CELL_LINE>':
            flag = 1
         if item == '<CELL_TYPE>':
            flag = 1
         if item == '<DNA>':
            flag = 1
         if item == '<RNA>':
            flag = 1
         if item == '</PROTEIN>':
            keep.append('\t')
            keep.append(ID)
            keep.append('\n')
            flag = 0
         if item == '</CELL_LINE>':
            keep.append('\t')
            keep.append(ID)
            keep.append('\n')
            flag = 0
```

```
                if item == '</CELL_TYPE>':
                    keep.append('\t')
                    keep.append(ID)
                    keep.append('\n')
                    flag = 0
                if item == '</DNA>':
                    keep.append('\t')
                    keep.append(ID)
                    keep.append('\n')
                    flag = 0
                if item == '</RNA>':
                    keep.append('\t')
                    keep.append(ID)
                    keep.append('\n')
                    flag = 0
                if flag == 1:
                    keep.append(str(item)+' ')
        j = j + 1
    for line in keep:
        outfile.write(str(line))
    outfile.close()
    print "Picker done."

# CleanerUpper

    print "CleanerUpper"
    print

    # Read in list of abstracts
    infile = open('PickerResults.txt', 'r')

    # Clean up tags
    firstlist = []
    comblist = []
    for line in infile:
        line = str(line)
        line = string.split(line)
        if line[0] == "<PROTEIN>":
            newline = str(line[1:])+'\n'
            firstlist.append(newline)
        if line[0] == "<DNA>":
            newline = str(line[1:])+'\n'
            firstlist.append(newline)
        if line[0] == "<RNA>":
            newline = str(line[1:])+'\n'
            firstlist.append(newline)
    if line[0] == "<CELL_LINE>":
        newline = str(line[1:])+'\n'
        firstlist.append(newline)
    if line[0] == "<CELL_TYPE>":
        newline = str(line[1:])+'\n'
        firstlist.append(newline)

# Remove extra characters
for line in firstlist:
    line = str(line)
    lineindex = line.index('\n')
    n = 0
    newstring = ""
    while n <= lineindex:
        if line[n]=="[":
            n=n+1
        elif line[n]=="""":
            n=n+1
        elif line[n]==",":
            n=n+1
        elif line[n]=="]":
            n=n+1
        elif line[n]=="'":
            n=n+1
        else:
            newstring = newstring + line[n]
            n = n+1
    comblist.append(newstring)

#Remove dups
biodict = []
for item in comblist:
    item = str(item)
    item = string.split(item)
    A = item
    A = str(A)
    A = A.lower()
    newA = ''
    if A in biodict:
        continue
    else:
        newA = str(A)
    biodict.append(newA)

#Write to outfile
biodict.sort()
```

```
    outfile = open('CleanerResults.txt', 'w')
    for line in biodict:
        line = str(line)
        lineindex = line.index(']')
        n = 0
        newstring = ""
        while n <= lineindex:
            if line[n]=="[":
                n=n+1
            elif line[n]=='"':
                n=n+1
            elif line[n]==",":
                n=n+1
            elif line[n]=="]":
                n=n+1
            elif line[n]=="'":
                n=n+1
            else:
                newstring = newstring + line[n]
                n = n+1
        outfile.write(str(newstring))
        outfile.write('\n')
    outfile.close()
    print "Cleaner done."

# Dictionary Checker

    print "Dictionary Checker"
    print "The dictionary list should have the gene/protein abbreviation in the first column, and aliases to the right - tab delimited"
    print "The input file should have one putative biomarker per line"
    print

    Dictionary = []
    Biomarkerlist = []

    # Read in dictionary
    dictfile = open(Dict, 'r')
    for line in dictfile:
        Dictionary.append(line)       # list each line
    dictfile.close()

    # Read in dictionary
    biofile = open('CleanerResults.txt', 'r')
    for line in biofile:
        Biomarkerlist.append(line)       # list each line
    biofile.close()

    # Match lists
    Keeplist = []
    for item in Biomarkerlist:
        item = str(item)
        item = item.split('\t')
        for entity in item:
            locate = entity.index('pmid')
            entity = entity[0:(locate-1)]
            for line in Dictionary:
                line = str(line)
                line = line.lower()
                line = line.split('\t')
                if entity in line:
                    Keeplist.append(line[0])

    # Tally up counts
    Finallist = []
    Keeplist.sort()
    oldcount = 1
    oldthing = ''
    for thing in Keeplist:
        a = Keeplist.count(thing)
        if thing == oldthing:
            continue
        else:
            oldthing = thing
            oldcount = a
            b = oldthing + '\t' + str(oldcount)
        Finallist.append(b)

    # Print output
    outfile = open('DictionaryCheckerResult.txt', 'w')
    outfile.write("PUTATIVE BIOMARKERS")
    outfile.write('\n')
    for item in Finallist:
        item = str(item)
        outfile.write(str(item))
        outfile.write('\n')
    outfile.close()
    print "All done."

main()
```

APPENDIX B

INPUT DATASET DESCRIPTIONS FROM GEO

Breast Cancer – Copy Number

Series GSE27574

Status Public on Nov 19, 2012

Title High-resolution analysis of copy number changes in circulating and disseminated tumor cells in breast cancer patients

Organism Homo sapiens

Experiment type Genome variation profiling by array

Summary The aim of this study was to establish a single-cell array comparative genomic hybridization (SCaCGH) method providing in-depth genomic analysis of circulating tumor cells (CTCs) and disseminated tumor cells (DTCs). The robustness and resolution limits of the method were estimated with different cell amounts of the breast cancer cell line SKBR3 using 44k and 244k arrays. Subsequent adjustments of the method were conducted analyzing the copy number profiles of 28 CTCs in combination with four hematopoietic cell (HC) controls from eight metastatic patients and of 24 DTCs, three probable HCs, and five HC controls from seven breast cancer patients and one healthy donor. The frequency of the major genomic gains and losses of the analyzed DTC revealed similarities to primary breast tumor samples with some evident differences. Three of the patients had available profiles for DTC and the corresponding primary tumor. In 2/3 of the examined DTCs, equivalent genomic changes and common aberration breakpoints were disclosed, both to each other and to the corresponding primary tumors. Interestingly, similar copy number changes were found in DTCs taken at time of diagnosis or in DTCs collected at 3-years relapse-free follow up. Residual immunomorphological characterized tumor cells showed balanced profiles with only minor aberrations. Three cells with unclear morphological identification showed either balanced profiles (n=2) or aberrations comparable to the primary tumor and DTC (n=1). SCaCGH may be a powerful tool for molecular characterization of immunostained and morphological identified CTCs and DTCs to explore the malignant potential, therapeutic targets and tumor heterogeneity of single tumor cells.

Overall design 24 DTCs, 3 probable HCs, and 5 HCs from 7 early-stage breast cancer patients, 28 CTCs and 4 HCs from 8 metastatic breast cancer patients, and 1 healthy donor were analyzed. Comparison with the primary tumor was done in 3 patients. The reference for the patients was DNA from multiple anonymous female donors. This submission does not include the SKBR3 data obtained from the 44k array.

Contributor(s) Baumbusch LO, Naume B, Speicher MR, Pantel K, Børresen-Dale A, Lingjærde OC, Mauermann O, Obenauf AC, Schneider IJ, Rye IH, Borgen E, Liestøl K, Riethdorf S, Geigl JB, Due EU, Fjelldal R, Mathiesen RR
Citation(s) Mathiesen RR, Fjelldal R, Liestøl K, Due EU et al. High-resolution analyses of copy number changes in disseminated tumor cells of patients with breast cancer. Int J Cancer 2012 Aug 15;131(4):E405-15. PMID: 21935921
Submission date Feb 28, 2011
Last update date Nov 14, 2014
Contact name Randi Mathiesen
E-mail randi.mathiesen@rr-research.no
Phone +4745290525
Organization name Institute for cancer research Oslo University Hospital Radiumhospitalet
Department Dept. of genetics
Street address Montebello
City Oslo
ZIP/Postal code 0310
Country Norway
Platforms (2)

GPL8841	Agilent-014950 Human Genome CGH Microarray 4x44K (Probe Name version)
GPL9128	Agilent-014693 Human Genome CGH Microarray 244A (Probe name version)

Samples (79)
Relations BioProject PRJNA181273

<u>Breast Cancer – Microarray</u>

Series GSE16443
Status Public on Jan 15, 2010
Title Gene expression profiling of peripheral blood cells for early detection of breast cancer
Organism Homo sapiens
Experiment type Expression profiling by array
Summary Purpose: Early detection of breast cancer is key to successful treatment and patient survival. We have previously reported the potential use of gene expression profiling of peripheral blood cells for early detection of breast cancer. The aim of the present study was to validate these findings using a larger sample size and a commercially available microarray platform.
Overall design Experimental Design: Blood samples were collected from 121 females referred for diagnostic mammography following an initial suspicious screening mammogram. Diagnostic work-up revealed that 67 of these women had breast cancer while 54 had no malignant disease. Additionally, 9 samples from 6 healthy female controls (three pregnant women, one breast-feeding woman and two healthy controls at different time points in their menstrual cycle) were included. Gene expression analyses were conducted using high-density oligonucleotide microarrays. Partial Least Square Regression was used for model building and predictors were identified using a Jackknifing approach. Prediction performance was determined by a 20-fold double cross validation approach
Contributor(s) Aarøe J, Lindahl T, Dumeaux V, Sæbø S, Hagen N, Tobin D, Skaane P, Lönneborg A, Sharma P, Børresen-Dale A

Citation(s) Aarøe J, Lindahl T, Dumeaux V, Saebø S et al. Gene expression profiling of peripheral blood cells for early detection of breast cancer. Breast Cancer Res 2010;12(1):R7. PMID: 20078854
Submission date Jun 04, 2009
Last update date Nov 12, 2012
Contact name Jørgen Mømb Aarøe
E-mail jorgen.aaroe@rr-research.no
Phone +4791774653
Fax +4722934440
Organization name The Norwegian Radium Hospital
Department Genetics
Lab Genetics
Street address Montebello
City Oslo
ZIP/Postal code N-0310
Country Norway

Platforms (1)
GPL2986 ABI Human Genome Survey Microarray Version 2
Samples (130)
Relations BioProject PRJNA116345

Breast Cancer – Methylation

Series GSE20713
Status Public on Oct 19, 2011
Title Epigenetic portraits of human breast cancers
Organism Homo sapiens
Experiment type Expression profiling by array; Methylation profiling by array
Summary This SuperSeries is composed of the SubSeries listed below.
Overall design Refer to individual Series
Citation(s) Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK et al. DNA methylation profiling reveals a predominant immune component in breast cancers. EMBO Mol Med 2011 Dec;3(12):726-41. PMID: 21910250
Submission date Mar 09, 2010
Last update date Jun 02, 2015
Contact name Benjamin Haibe-Kains
E-mail benjamin.haibe.kains@utoronto.ca
Phone +14165818626
Organization name Princess Margaret Cancer Centre
Department Princess Margaret Research
Lab Bioinformatics and Computational Genomics
Street address 610 University Avenue
City Toronto
State/province Ontario
ZIP/Postal code M5G 2M9
Country Canada

Platforms (2)
GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
GPL8490 Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2)
Samples (381)
Relations BioProject PRJNA124907

<u>Breast Cancer – RT-PCR</u>
Series GSE46068
Status Public on Jan 09, 2014
Title Molecular characterization of tumor cells from the cerebrospinal fluid and matched primary tumors from metastatic breast cancer patients with leptomeningeal disease
Organism Homo sapiens
Experiment type Genome variation profiling by genome tiling array; Expression profiling by RT-PCR
Summary We purified tumor cells in the CSF (“CSFTC”) from 15 metastatic breast cancer patients diagnosed with leptomeningeal disease using a two-step method involving immunomagnetic enrichment and fluorescence-activated cell sorting (IE/FACS). Magnetic beads coated with mAb to the epithelial cell adhesion marker (EPCAM) were used to enrich for tumor cells and were further purified by FACS analysis.
For DNA profiling, isolated CSFTC were subjected to molecular characterization through genome-wide copy number analyses. Genomic analyses were then compared with those performed on the corresponding archival primary tumors.
For RNA profiling, isolated CSFTC were then subjected to molecular characterization through gene expression profiling via QPCR analysis of 64 cancer-related genes
Overall design CGH: 17 CSFTC samples from 13 patients were successfully profiled, 1 patient had 5 time points, 6 of 13 patients had matched to copy number data archival tumors
RNA: 18 samples from 5 patients had successful gene expression data of the 64 genes measured in triplicates. For non-tumor controls, 9 of the samples had matching gene expression data from sorted leuckocytes (+CD45 cells) obtained from the same draw.
Contributor(s) Magbanua MJ, Scott JH, Hauranieh L, Melisko M, Sosa EV, Kablanian A, Roy R, Park JW
Citation(s) Magbanua MJ, Melisko M, Roy R, Sosa EV et al. Molecular profiling of tumor cells in cerebrospinal fluid and matched primary tumors from metastatic breast cancer patients with leptomeningeal carcinomatosis. Cancer Res 2013 Dec 1;73(23):7134-43. PMID: 24142343
Submission date Apr 15, 2013
Last update date Jan 09, 2014
Contact name Mark Magbanua
Organization name UCSF/Helen Diller Family Comprehensive Cancer Center
Department HemOnc
Lab Park
Street address 1450 3rd Street, PO Box 589001
City San Francisco
State/province CA
ZIP/Postal code 94158-9001
Country USA
Platforms (2)

GPL6359 UCSF Cancer Center HumArray3.2
GPL17020 Custom Human TLDA 64-Circulating tumor cell associated gene panel
Samples (57)
Relations BioProject PRJNA197197

Breast Cancer – Protein

Series GSE34555
Status Public on Dec 21, 2011
Title Evaluation of auto-antibody serum biomarkers for breast cancer screening
Organism Homo sapiens
Experiment type Protein profiling by protein array

Summary Using protein microarrays, derived from 642 His-tag proteins, we could distinguish sera from breast-nodule positive patients and healthy control individuals.

Overall design Each Protein microarray was divided in to 4 sub-arrays. Each protein was spotted in duplicates in each sub-array. For evaluation 24 malignant, 16 benign breast cancer serum samples and 20 healthy control serum samples were used.

Contributor(s) Weinhäusel A, Syed P
Citation missing Has this study been published? Please login to update or notify GEO.
Submission date Dec 19, 2011
Last update date Mar 23, 2012
Contact name Parvez Syed
E-mail parvez.syed@ait.ac.at
Organization name Austrian Institution of Technology
Street address Muthgasse 11
City Vienna
ZIP/Postal code 1180
Country Austria
Platforms (1)
GPL15009 Austrian Institution of Technology Protein Array 642
Samples (60)
Relations BioProject PRJNA151535

Lung Cancer – Microarray

Series GSE20189
Status Public on Sep 23, 2011
Title A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma
Organism Homo sapiens
Experiment type Expression profiling by array

Summary Affordable early screening in subjects with high risk of lung cancer has great potential to improve survival from this deadly disease. We measured gene expression from lung tissue and peripheral whole blood (PWB) from adenocarcinoma cases and controls to identify dysregulated lung cancer genes that could be tested in blood to improve identification of at-risk patients in the future. Genome-wide mRNA expression analysis was conducted in 153 subjects (73 adenocarcinoma cases, 80 controls) from the Environment and Genetics in Lung cancer Etiology (EAGLE) study using PWB and paired snap-frozen tumor and non-involved lung tissue samples. Analyses were conducted using unpaired t-tests, linear mixed effects and ANOVA models. The area under the receiver operating characteristic curve (AUC) was computed to assess the predictive accuracy of the identified biomarkers. We identified 50 dysregulated genes in stage I adenocarcinoma versus control PWB samples (False Discovery Rate ≤0.1, fold change ≥1.5 or ≤0.66). Among them, eight (TGFBR3, RUNX3, TRGC2, TRGV9, TARP, ACP1, VCAN, and TSTA3) differentiated paired tumor versus non-involved lung tissue samples in stage I cases, suggesting a similar pattern of lung cancer-related changes in PWB and lung tissue. These results were confirmed in two independent gene expression analyses in a blood-based case-control study (n=212) and a tumor-non tumor paired tissue study (n=54). The eight genes discriminated patients with lung cancer from healthy controls with high accuracy (AUC=0.81, 95% CI=0.74-0.87). Our finding suggests the use of gene expression from PWB for the identification of early detection markers of lung cancer in the future.

Overall design Samples from 164 subjects were initially included in the study. Two samples with poor quality profile based on quality assessment (described in Supplemental Material 2) were excluded before normalization. The remaining 162 samples were processed and normalized with the Robust Multichip Average (RMA) method. Nine additional subjects were excluded after data normalization because of reclassification to non-adenocarcinoma morphology during histologic review. The final analyses were based on 73 adenocarcinoma cases and 80 controls. All 22,277 probe sets based on RMA summary measures were used in the analyses.

Contributor(s) Rotunno M, Hu N, Su H, Wang C, Goldstein AM, Bergen AW, Consonni D, Pesatori AC, Bertazzi P, Wacholder S, Shih J, Caporaso NE, Taylor PR, Landi M

Citation(s) Rotunno M, Hu N, Su H, Wang C et al. A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. Cancer Prev Res (Phila) 2011 Oct;4(10):1599-608. PMID: 21742797

Submission date Feb 04, 2010

Last update date Mar 06, 2015

Contact name Melissa Rotunno

E-mail rotunnom@mail.nih.gov

Phone 301-402-1622

Fax 301-402-4489

Organization name NIH/NCI

Department DCEG

Lab GEB

Street address 6120 Executive Blvd

City Rockville

State/province MD

ZIP/Postal code 20892

Country USA

Platforms (1)

GPL571 [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array

Samples (162)

Relations BioProject PRJNA125685

Lung Cancer – ArrayCGH

Series GSE14079

Status Public on Mar 03, 2009
Title Gene expression analysis & Comparative genomic hybridization from Lung cancer Cell lines
Organism Homo sapiens
Experiment type Expression profiling by array; Genome variation profiling by array
Summary Gene expression and Comparative genomic hybridization (CGH) microarrays performed in a set of 8 Lung cancer Cell lines.
Overall design The search for oncogenes is becoming increasingly important in cancer genetics because they constitute suitable targets for therapeutic intervention. To identify novel oncogenes, activated by gene amplification, we performed high-resolution CGH (Comparative Genome Hybridization) analysis on cDNA microarrays and compared DNA copy number and mRNA expression levels in lung cancer cell lines. We have performed both microarrays (expression and CGH) in a set of 8 human lung cancer cell lines: Calu3, H23, H441, A427, H522, A549, H1299, and H2126.
Contributor(s) Medina PP, Castillo S, Sanchez-Cespedes M
Citation(s) Medina PP, Castillo SD, Blanco S, Sanz-Garcia M et al. The SRY-HMG box gene, SOX4, is a target of gene amplification at chromosome 6p in lung cancer. Hum Mol Genet 2009 Apr 1;18(7):1343-52. PMID: 19153074
Submission date Dec 21, 2008
Last update date Mar 20, 2012
Contact name Pedro P Medina-Vico
E-mail ppmedinavico@gmail.com
Organization name Yale University
Street address 266 Whitney Ave, 938 KBT
City New Haven
State/province CT
ZIP/Postal code 06511
Country USA
Platforms (1)
GPL1998 CNIO H. sapiens 13.6k Oncochip 1
Samples (16)
Relations BioProject PRJNA112505

Lung Cancer – Methylation

Series GSE5816
Status Public on Jan 03, 2007
Title A Genome-wide Screen for Hypermethylated Genes in Lung Cancer
Organism Homo sapiens
Experiment type Expression profiling by array
Summary Abstract

Background: Promoter hypermethylation coupled with loss of heterozygosity at the same locus results in loss of gene function in many tumor cells. The “rules” governing which genes are methylated during the pathogenesis of individual cancers, how specific methylation profiles are initially established, or what determines tumor-type specific methylation are unknown. However, DNA methylation markers that are highly specific and sensitive for common tumors would be useful for the early detection of cancer, and those required for the malignant phenotype identify pathways important as therapeutic targets.

Methods and Findings: In an effort to identify new cancer-specific methylation markers, we employed a high throughput global expression profiling approach in lung cancer cells. We identified 132 genes that have 5’ CpG islands, are induced from undetectable levels by 5-aza-2’-deoxycytidine (5-aza) in multiple non-small cell lung cancer cell lines, and are expressed in immortalized human bronchial epithelial cells. As expected, these genes were also expressed in normal lung, but often not in companion primary lung cancers. Methylation analysis of a subset (45/132) of these promoter regions in primary lung cancer (N=20) and adjacent non-malignant tissue showed that 31 genes had acquired methylation in the tumors, but did not show methylation in normal lung or lymphocytes. We studied the eight most frequently and specifically methylated genes from our lung cancer data set in breast cancer (N=37), colon cancer (N=24), and prostate cancer (N=24) along with counterpart non-malignant tissues. We found that seven loci were frequently methylated in both breast and lung cancers, with four showing extensive methylation in all four epithelial tumors.

Conclusions: By using a systematic biological screen we identified multiple genes that are methylated with high penetrance in primary lung, breast, colon, and prostate cancers. The cross-tumor methylation pattern we observed for these novel markers suggests that we have identified a partial promoter hypermethylation signature for these common malignancies. These data suggest that while tumors in different tissues vary substantially with respect to gene expression, there may be commonalities in their promoter methylation profiles that represent potential targets for early detection screening or therapeutic intervention.

Keywords Cell line comparison

Overall design Drug treatment: control, 100 nM, 1 uM

Cancer vs. Normal Comparison: NSCLC vs. Normal

Contributor(s) Shames DS, Girard L, Gao B, Sato M, Lewis CM, Shivapurkar N, Jiang A, Perou CM, Kim YH, Pollack JR, Fong KM, Lam CD, Wong M, Shyr Y, Nanda R, Olopade OL, Gerald W, Euhus DM, Shay JW, Gazdar AF, Minna JD

Citation(s) Shames DS, Girard L, Gao B, Sato M et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. PLoS Med 2006 Dec;3(12):e486. PMID: 17194187

Submission date Sep 12, 2006
Last update date Mar 20, 2015
Contact name David S Shames
E-mail shames.david@gene.com
Phone 650-225-7559
Organization name Genentech Inc.
Department Oncology Biomarker Development
Lab Shames
Street address 1 DNA Way
City South San Francisco
State/province CA
ZIP/Postal code 94080
Country USA

Platforms (1)

GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (42)
Relations BioProject PRJNA97201

Lung cancer – Copy Number

Series GSE31800
Status Public on Sep 12, 2011
Title DNA copy number and gene expression profiles of resected non-small cell lung cancer tumors
Organism Homo sapiens
Experiment type Genome variation profiling by genome tiling array; Expression profiling by array
Summary This SuperSeries is composed of the SubSeries listed below.
Overall design Refer to individual Series
Citation(s) Starczynowski DT, Lockwood WW, Deléhouzée S, Chari R et al. TRAF6 is an amplified oncogene bridging the RAS and NF-κB pathways in human lung cancer. J Clin Invest 2011 Oct;121(10):4095-105. PMID: 21911935
Submission date Aug 31, 2011
Last update date Jan 18, 2013
Contact name Raj Chari
E-mail rchari@bccrc.ca
Organization name BC Cancer Research Centre
Department Cancer Genetics and Developmental Biology
Lab Wan Lam Lab
Street address 675 West 10th Avenue
City Vancouver
State/province BC
ZIP/Postal code V5Z 1L3
Country Canada
Platforms (2)
GPL14189 Custom Rosetta-Affymetrix Human platform [rmhu01aa520485]
GPL14360 BCCRC whole genome tilling path array v2 (March 2006 build)
Samples (320)
This SuperSeries is composed of the following SubSeries:
GSE31798 DNA copy number profiles of NSCLC tumors
GSE31799 Gene expression profiles of NSCLC tumors
Relations BioProject PRJNA145473

Series GSE31798
Status Public on Sep 12, 2011
Title DNA copy number profiles of NSCLC tumors
Organism Homo sapiens
Experiment type Genome variation profiling by genome tiling array
Summary Whole genome tiling path array CGH was used to measure the copy number profiles of 271 NSCLC tumors
Overall design 271 microdissected NSCLC tumors
Contributor(s) Chari R, Lockwood WW, Lam WL

Citation(s) Starczynowski DT, Lockwood WW, Deléhouzée S, Chari R et al. TRAF6 is an amplified oncogene bridging the RAS and NF-κB pathways in human lung cancer. J Clin Invest 2011 Oct;121(10):4095-105. PMID: 21911935
Submission date Aug 31, 2011
Last update date Mar 23, 2012
Contact name Raj Chari
E-mail rchari@bccrc.ca
Organization name BC Cancer Research Centre
Department Cancer Genetics and Developmental Biology
Lab Wan Lam Lab
Street address 675 West 10th Avenue
City Vancouver
State/province BC
ZIP/Postal code V5Z 1L3
Country Canada
Platforms (1)
GPL14360 BCCRC whole genome tilling path array v2 (March 2006 build)
Samples (271)
This SubSeries is part of SuperSeries: GSE31800 DNA copy number and gene expression profiles of resected non-small cell lung cancer tumors
Relations BioProject PRJNA155045

Series GSE31799
Status Public on Sep 12, 2011
Title Gene expression profiles of NSCLC tumors
Organism Homo sapiens
Experiment type Expression profiling by array
Summary A custom microarray was used to measure the gene expression of NSCLC tumors. This represents a subset of samples which also have matched DNA copy number profiles from array CGH experiments
Overall design 49 microdissected NSCLC tumor samples
Contributor(s) Chari R, Lockwood WW, Lam WL
Citation(s) Starczynowski DT, Lockwood WW, Deléhouzée S, Chari R et al. TRAF6 is an amplified oncogene bridging the RAS and NF-κB pathways in human lung cancer. J Clin Invest 2011 Oct;121(10):4095-105. PMID: 21911935
Submission date Aug 31, 2011
Last update date Jan 18, 2013
Contact name Raj Chari
E-mail rchari@bccrc.ca
Organization name BC Cancer Research Centre
Department Cancer Genetics and Developmental Biology
Lab Wan Lam Lab
Street address 675 West 10th Avenue
City Vancouver
State/province BC
ZIP/Postal code V5Z 1L3

Country Canada

Platforms (1)

GPL14189 Custom Rosetta-Affymetrix Human platform [rmhu01aa520485]

Samples (49)

This SubSeries is part of SuperSeries: GSE31800 DNA copy number and gene expression profiles of resected non-small cell lung cancer tumors

Relations BioProject PRJNA155047

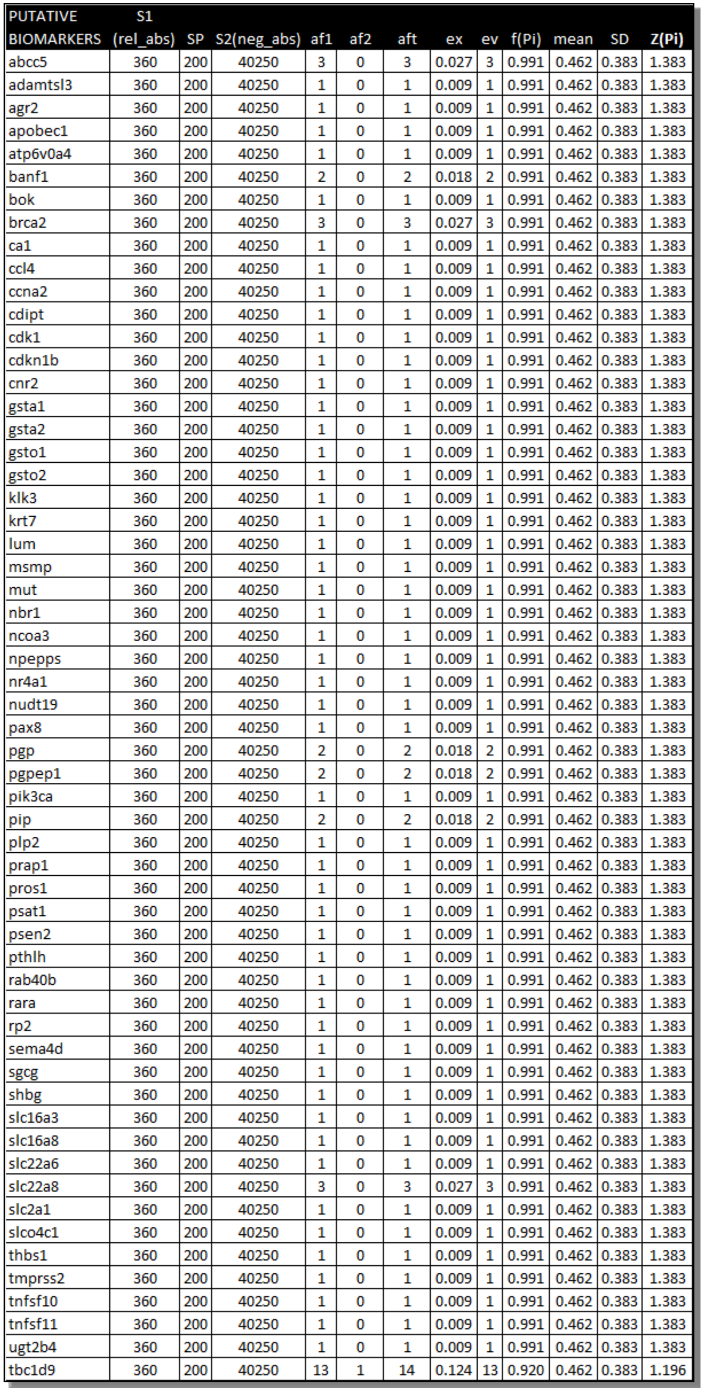
APPENDIX C

BIOMARKERS FOUND BY KEDA LITERATURE-MINING COMPONENT

Only significant biomarkers ($z \geq 1.0$) are presented in these tables

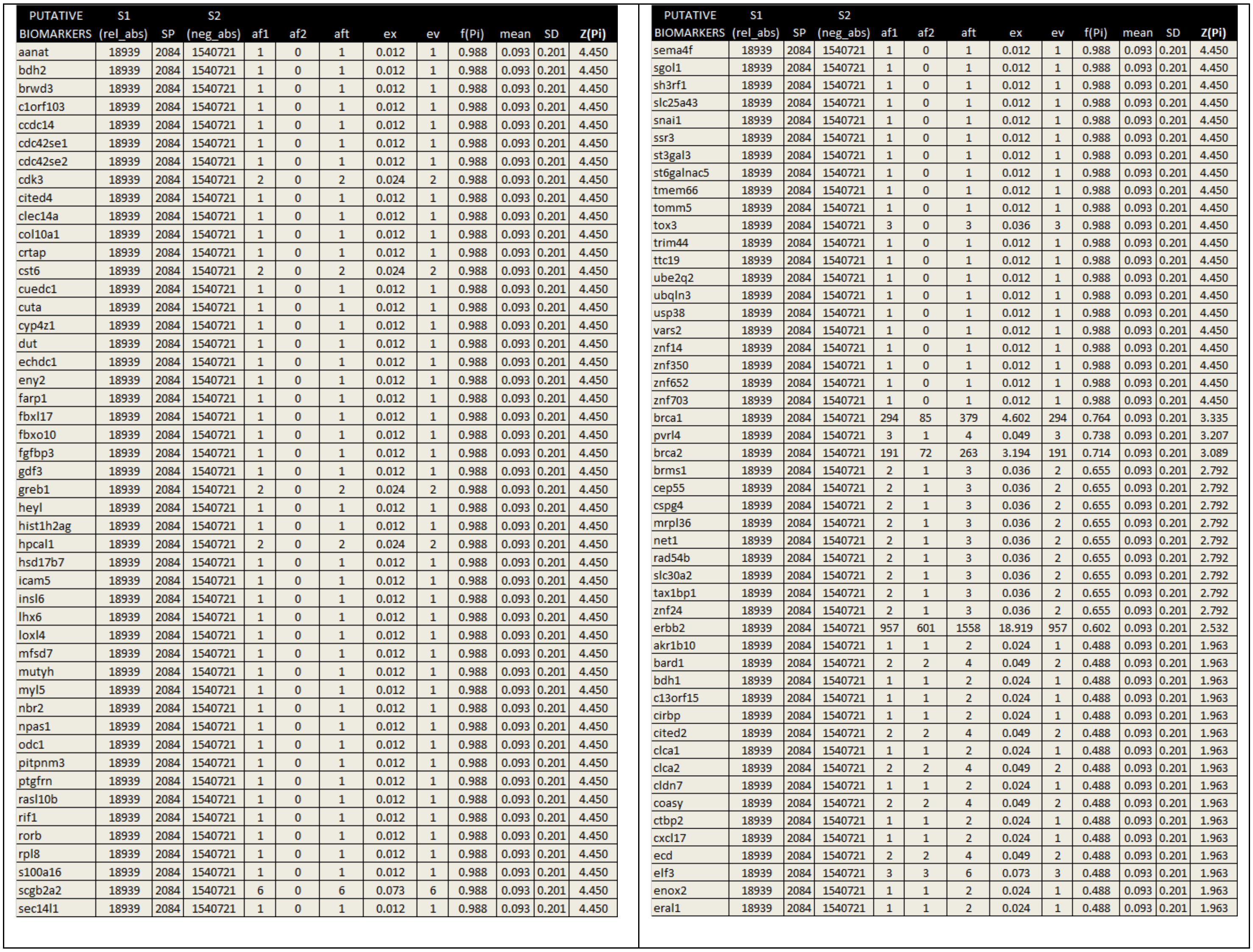
BREAST CANCER

Table C.1. Breast cancer/bile text-mining results and calculations.

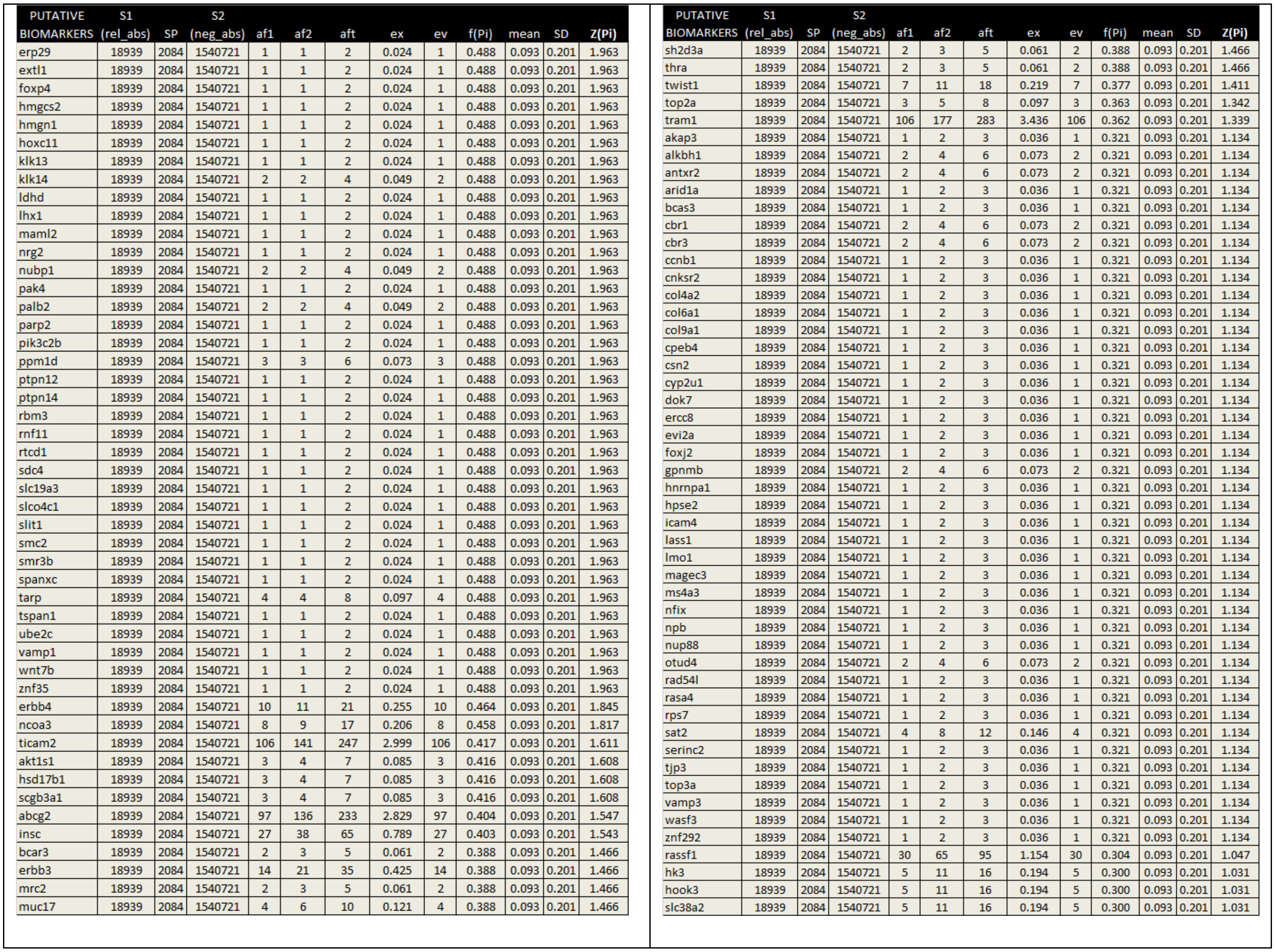


PUTATIVE BIOMARKERS	S1 (rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
abcc5	360	200	40250	3	0	3	0.027	3	0.991	0.462	0.383	1.383
adamtsl3	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
agr2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
apobec1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
atp6v0a4	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
banf1	360	200	40250	2	0	2	0.018	2	0.991	0.462	0.383	1.383
bok	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
brca2	360	200	40250	3	0	3	0.027	3	0.991	0.462	0.383	1.383
ca1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
ccl4	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
ccna2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
cdipt	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
cdk1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
cdkn1b	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
cnr2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
gsta1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
gsta2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
gsto1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
gsto2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
klk3	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
krt7	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
lum	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
msmp	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
mut	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
nbr1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
ncoa3	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
npepps	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
nr4a1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
nudt19	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
pax8	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
pgp	360	200	40250	2	0	2	0.018	2	0.991	0.462	0.383	1.383
pgpep1	360	200	40250	2	0	2	0.018	2	0.991	0.462	0.383	1.383
pik3ca	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
pip	360	200	40250	2	0	2	0.018	2	0.991	0.462	0.383	1.383
plp2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
prap1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
pros1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
psat1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
psen2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
pthlh	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
rab40b	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
rara	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
rp2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
sema4d	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
sgcg	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
shbg	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
slc16a3	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
slc16a8	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
slc22a6	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
slc22a8	360	200	40250	3	0	3	0.027	3	0.991	0.462	0.383	1.383
slc2a1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
slco4c1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
thbs1	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
tmprss2	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
tnfsf10	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
tnfsf11	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
ugt2b4	360	200	40250	1	0	1	0.009	1	0.991	0.462	0.383	1.383
tbc1d9	360	200	40250	13	1	14	0.124	13	0.920	0.462	0.383	1.196

Table C.2. Breast cancer/blood text-mining results and calculations.



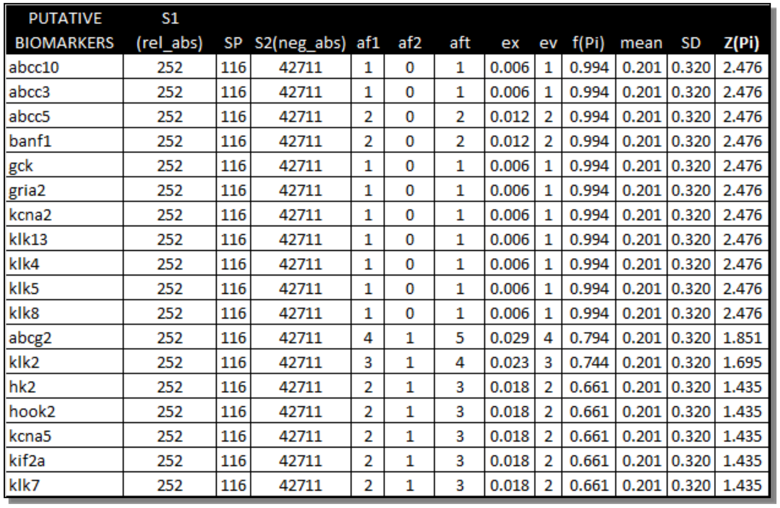
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
aanat	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
bdh2	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
brwd3	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
c1orf103	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
ccdc14	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
cdc42se1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
cdc42se2	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
cdk3	18939	2084	1540721	2	0	2	0.024	2	0.988	0.093	0.201	4.450
cited4	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
clec14a	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
col10a1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
crtap	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
cst6	18939	2084	1540721	2	0	2	0.024	2	0.988	0.093	0.201	4.450
cuedc1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
cuta	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
cyp4z1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
dut	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
echdc1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
eny2	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
farp1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
fbxl17	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
fbxo10	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
fgfbp3	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
gdf3	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
greb1	18939	2084	1540721	2	0	2	0.024	2	0.988	0.093	0.201	4.450
heyl	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
hist1h2ag	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
hpcal1	18939	2084	1540721	2	0	2	0.024	2	0.988	0.093	0.201	4.450
hsd17b7	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
icam5	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
insl6	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
lhx6	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
loxl4	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
mfsd7	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
mutyh	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
myl5	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
nbr2	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
npas1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
odc1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
pitpnm3	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
ptgfrn	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
rasl10b	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
rif1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
rorb	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
rpl8	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
s100a16	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
scgb2a2	18939	2084	1540721	6	0	6	0.073	6	0.988	0.093	0.201	4.450
sec14l1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
sema4f	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
sgol1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
sh3rf1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
slc25a43	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
snai1	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
ssr3	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
st3gal3	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
st6galnac5	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
tmem66	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
tomm5	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
tox3	18939	2084	1540721	3	0	3	0.036	3	0.988	0.093	0.201	4.450
trim44	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
ttc19	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
ube2q2	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
ubqln3	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
usp38	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
vars2	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
znf14	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
znf350	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
znf652	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
znf703	18939	2084	1540721	1	0	1	0.012	1	0.988	0.093	0.201	4.450
brca1	18939	2084	1540721	294	85	379	4.602	294	0.764	0.093	0.201	3.335
pvrl4	18939	2084	1540721	3	1	4	0.049	3	0.738	0.093	0.201	3.207
brca2	18939	2084	1540721	191	72	263	3.194	191	0.714	0.093	0.201	3.089
brms1	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
cep55	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
cspg4	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
mrpl36	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
net1	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
rad54b	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
slc30a2	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
tax1bp1	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
znf24	18939	2084	1540721	2	1	3	0.036	2	0.655	0.093	0.201	2.792
erbb2	18939	2084	1540721	957	601	1558	18.919	957	0.602	0.093	0.201	2.532
akr1b10	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
bard1	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
bdh1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
c13orf15	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
cirbp	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
cited2	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
clca1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
clca2	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
cldn7	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
coasy	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
ctbp2	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
cxcl17	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
ecd	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
elf3	18939	2084	1540721	3	3	6	0.073	3	0.488	0.093	0.201	1.963
enox2	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
eral1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963



PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
erp29	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
extl1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
foxp4	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
hmgcs2	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
hmgn1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
hoxc11	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
klk13	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
klk14	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
ldhd	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
lhx1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
maml2	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
nrg2	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
nubp1	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
pak4	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
palb2	18939	2084	1540721	2	2	4	0.049	2	0.488	0.093	0.201	1.963
parp2	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
pik3c2b	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
ppm1d	18939	2084	1540721	3	3	6	0.073	3	0.488	0.093	0.201	1.963
ptpn12	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
ptpn14	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
rbm3	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
rnf11	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
rtcd1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
sdc4	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
slc19a3	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
slco4c1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
slit1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
smc2	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
smr3b	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
spanxc	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
tarp	18939	2084	1540721	4	4	8	0.097	4	0.488	0.093	0.201	1.963
tspan1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
ube2c	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
vamp1	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
wnt7b	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
znf35	18939	2084	1540721	1	1	2	0.024	1	0.488	0.093	0.201	1.963
erbb4	18939	2084	1540721	10	11	21	0.255	10	0.464	0.093	0.201	1.845
ncoa3	18939	2084	1540721	8	9	17	0.206	8	0.458	0.093	0.201	1.817
ticam2	18939	2084	1540721	106	141	247	2.999	106	0.417	0.093	0.201	1.611
akt1s1	18939	2084	1540721	3	4	7	0.085	3	0.416	0.093	0.201	1.608
hsd17b1	18939	2084	1540721	3	4	7	0.085	3	0.416	0.093	0.201	1.608
scgb3a1	18939	2084	1540721	3	4	7	0.085	3	0.416	0.093	0.201	1.608
abcg2	18939	2084	1540721	97	136	233	2.829	97	0.404	0.093	0.201	1.547
insc	18939	2084	1540721	27	38	65	0.789	27	0.403	0.093	0.201	1.543
bcar3	18939	2084	1540721	2	3	5	0.061	2	0.388	0.093	0.201	1.466
erbb3	18939	2084	1540721	14	21	35	0.425	14	0.388	0.093	0.201	1.466
mrc2	18939	2084	1540721	2	3	5	0.061	2	0.388	0.093	0.201	1.466
muc17	18939	2084	1540721	4	6	10	0.121	4	0.388	0.093	0.201	1.466

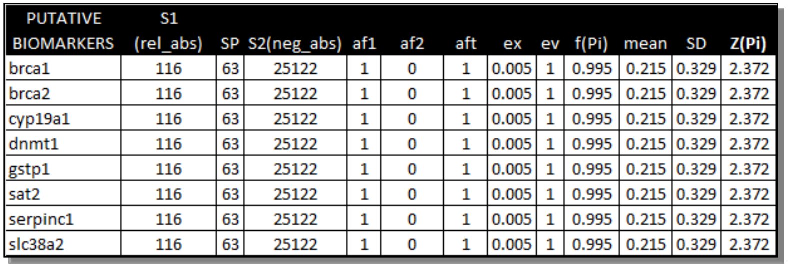
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
sh2d3a	18939	2084	1540721	2	3	5	0.061	2	0.388	0.093	0.201	1.466
thra	18939	2084	1540721	2	3	5	0.061	2	0.388	0.093	0.201	1.466
twist1	18939	2084	1540721	7	11	18	0.219	7	0.377	0.093	0.201	1.411
top2a	18939	2084	1540721	3	5	8	0.097	3	0.363	0.093	0.201	1.342
tram1	18939	2084	1540721	106	177	283	3.436	106	0.362	0.093	0.201	1.339
akap3	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
alkbh1	18939	2084	1540721	2	4	6	0.073	2	0.321	0.093	0.201	1.134
antxr2	18939	2084	1540721	2	4	6	0.073	2	0.321	0.093	0.201	1.134
arid1a	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
bcas3	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
cbr1	18939	2084	1540721	2	4	6	0.073	2	0.321	0.093	0.201	1.134
cbr3	18939	2084	1540721	2	4	6	0.073	2	0.321	0.093	0.201	1.134
ccnb1	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
cnksr2	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
col4a2	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
col6a1	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
col9a1	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
cpeb4	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
csn2	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
cyp2u1	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
dok7	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
ercc8	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
evi2a	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
foxj2	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
gpnmb	18939	2084	1540721	2	4	6	0.073	2	0.321	0.093	0.201	1.134
hnrnpa1	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
hpse2	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
icam4	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
lass1	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
lmo1	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
magec3	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
ms4a3	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
nfix	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
npb	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
nup88	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
otud4	18939	2084	1540721	2	4	6	0.073	2	0.321	0.093	0.201	1.134
rad54l	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
rasa4	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
rps7	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
sat2	18939	2084	1540721	4	8	12	0.146	4	0.321	0.093	0.201	1.134
serinc2	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
tjp3	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
top3a	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
vamp3	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
wasf3	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
znf292	18939	2084	1540721	1	2	3	0.036	1	0.321	0.093	0.201	1.134
rassf1	18939	2084	1540721	30	65	95	1.154	30	0.304	0.093	0.201	1.047
hk3	18939	2084	1540721	5	11	16	0.194	5	0.300	0.093	0.201	1.031
hook3	18939	2084	1540721	5	11	16	0.194	5	0.300	0.093	0.201	1.031
slc38a2	18939	2084	1540721	5	11	16	0.194	5	0.300	0.093	0.201	1.031

Table C.3. Breast cancer/CSF text-mining results and calculations.



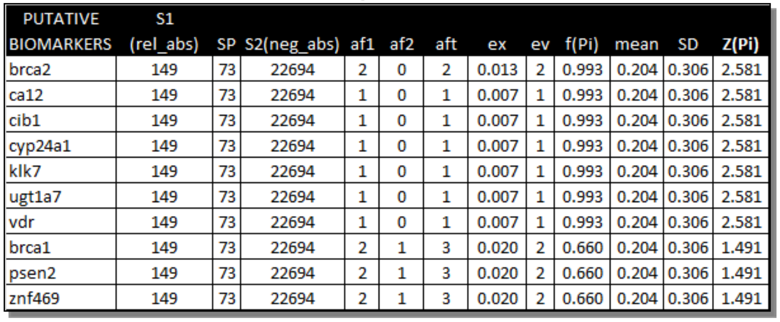
PUTATIVE BIOMARKERS	S1 (rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
abcc10	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
abcc3	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
abcc5	252	116	42711	2	0	2	0.012	2	0.994	0.201	0.320	2.476
banf1	252	116	42711	2	0	2	0.012	2	0.994	0.201	0.320	2.476
gck	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
gria2	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
kcna2	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
klk13	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
klk4	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
klk5	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
klk8	252	116	42711	1	0	1	0.006	1	0.994	0.201	0.320	2.476
abcg2	252	116	42711	4	1	5	0.029	4	0.794	0.201	0.320	1.851
klk2	252	116	42711	3	1	4	0.023	3	0.744	0.201	0.320	1.695
hk2	252	116	42711	2	1	3	0.018	2	0.661	0.201	0.320	1.435
hook2	252	116	42711	2	1	3	0.018	2	0.661	0.201	0.320	1.435
kcna5	252	116	42711	2	1	3	0.018	2	0.661	0.201	0.320	1.435
kif2a	252	116	42711	2	1	3	0.018	2	0.661	0.201	0.320	1.435
klk7	252	116	42711	2	1	3	0.018	2	0.661	0.201	0.320	1.435

Table C.4. Breast cancer/mucus text-mining results and calculations.



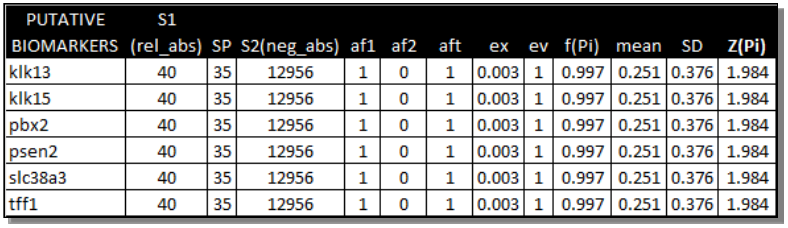
PUTATIVE BIOMARKERS	S1 (rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
brca1	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372
brca2	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372
cyp19a1	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372
dnmt1	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372
gstp1	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372
sat2	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372
serpinc1	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372
slc38a2	116	63	25122	1	0	1	0.005	1	0.995	0.215	0.329	2.372

Table C.5. Breast cancer/saliva text-mining results and calculations.



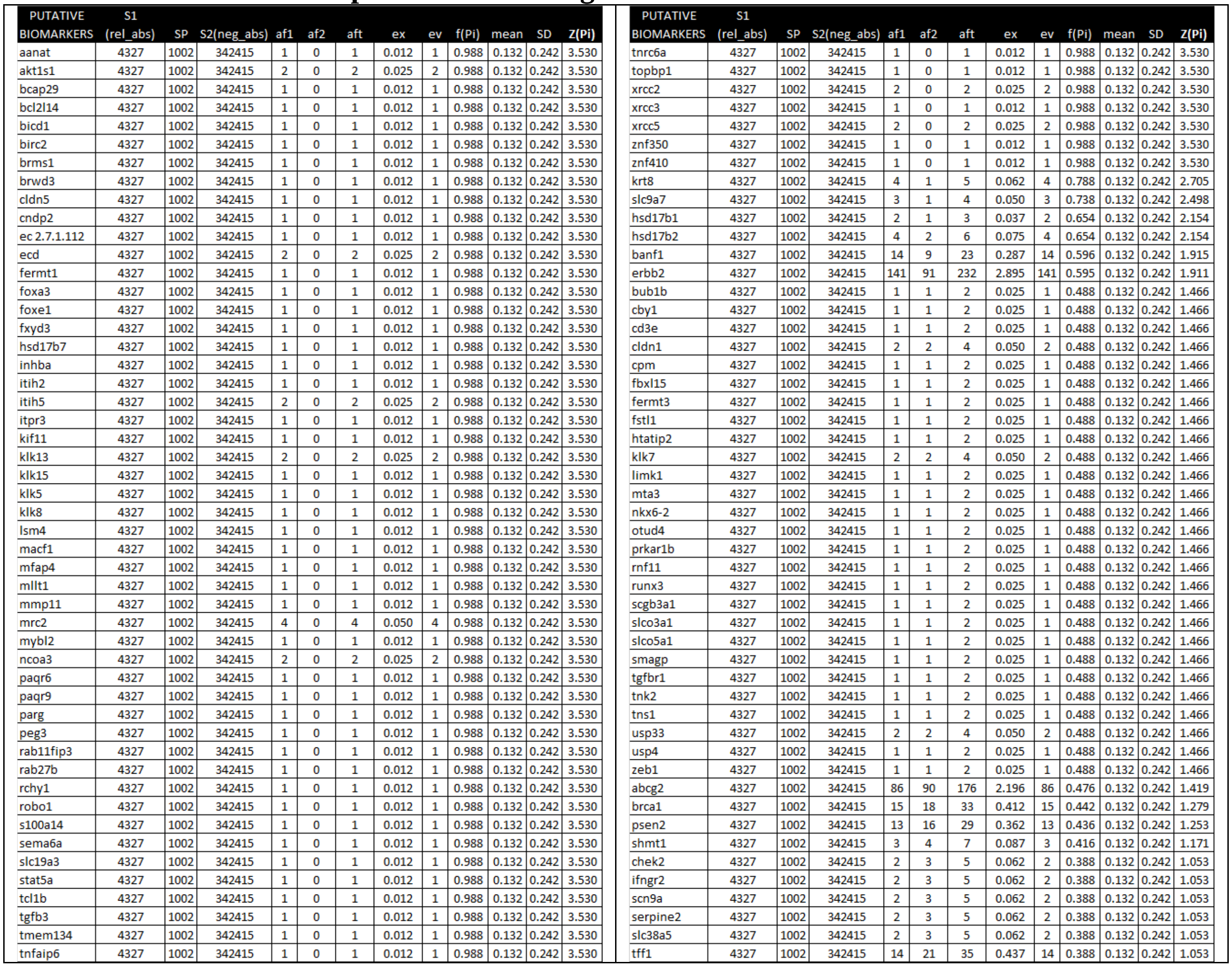
PUTATIVE BIOMARKERS	S1 (rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
brca2	149	73	22694	2	0	2	0.013	2	0.993	0.204	0.306	2.581
ca12	149	73	22694	1	0	1	0.007	1	0.993	0.204	0.306	2.581
cib1	149	73	22694	1	0	1	0.007	1	0.993	0.204	0.306	2.581
cyp24a1	149	73	22694	1	0	1	0.007	1	0.993	0.204	0.306	2.581
klk7	149	73	22694	1	0	1	0.007	1	0.993	0.204	0.306	2.581
ugt1a7	149	73	22694	1	0	1	0.007	1	0.993	0.204	0.306	2.581
vdr	149	73	22694	1	0	1	0.007	1	0.993	0.204	0.306	2.581
brca1	149	73	22694	2	1	3	0.020	2	0.660	0.204	0.306	1.491
psen2	149	73	22694	2	1	3	0.020	2	0.660	0.204	0.306	1.491
znf469	149	73	22694	2	1	3	0.020	2	0.660	0.204	0.306	1.491

Table C.6. Breast cancer/semen text-mining results and calculations.



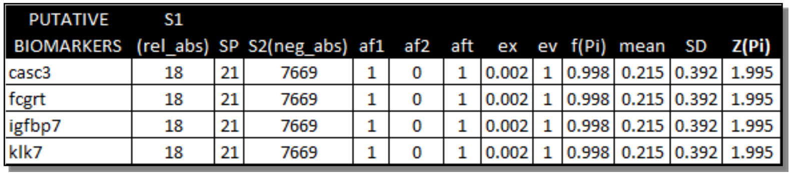
PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
klk13	40	35	12956	1	0	1	0.003	1	0.997	0.251	0.376	1.984
klk15	40	35	12956	1	0	1	0.003	1	0.997	0.251	0.376	1.984
pbx2	40	35	12956	1	0	1	0.003	1	0.997	0.251	0.376	1.984
psen2	40	35	12956	1	0	1	0.003	1	0.997	0.251	0.376	1.984
slc38a3	40	35	12956	1	0	1	0.003	1	0.997	0.251	0.376	1.984
tff1	40	35	12956	1	0	1	0.003	1	0.997	0.251	0.376	1.984

Table C.7. Breast cancer/plasma text-mining results and calculations.



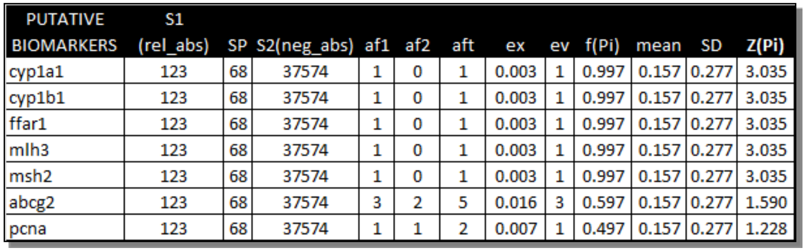
PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
aanat	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
akt1s1	4327	1002	342415	2	0	2	0.025	2	0.988	0.132	0.242	3.530
bcap29	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
bcl2l14	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
bicd1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
birc2	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
brms1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
brwd3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
cldn5	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
cndp2	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
ec 2.7.1.112	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
ecd	4327	1002	342415	2	0	2	0.025	2	0.988	0.132	0.242	3.530
fermt1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
foxa3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
foxe1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
fxyd3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
hsd17b7	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
inhba	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
itih2	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
itih5	4327	1002	342415	2	0	2	0.025	2	0.988	0.132	0.242	3.530
itpr3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
kif11	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
klk13	4327	1002	342415	2	0	2	0.025	2	0.988	0.132	0.242	3.530
klk15	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
klk5	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
klk8	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
lsm4	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
macf1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
mfap4	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
mllt1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
mmp11	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
mrc2	4327	1002	342415	4	0	4	0.050	4	0.988	0.132	0.242	3.530
mybl2	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
ncoa3	4327	1002	342415	2	0	2	0.025	2	0.988	0.132	0.242	3.530
paqr6	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
paqr9	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
parg	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
peg3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
rab11fip3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
rab27b	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
rchy1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
robo1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
s100a14	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
sema6a	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
slc19a3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
stat5a	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
tcl1b	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
tgfb3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
tmem134	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
tnfaip6	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
tnrc6a	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
topbp1	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
xrcc2	4327	1002	342415	2	0	2	0.025	2	0.988	0.132	0.242	3.530
xrcc3	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
xrcc5	4327	1002	342415	2	0	2	0.025	2	0.988	0.132	0.242	3.530
znf350	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
znf410	4327	1002	342415	1	0	1	0.012	1	0.988	0.132	0.242	3.530
krt8	4327	1002	342415	4	1	5	0.062	4	0.788	0.132	0.242	2.705
slc9a7	4327	1002	342415	3	1	4	0.050	3	0.738	0.132	0.242	2.498
hsd17b1	4327	1002	342415	2	1	3	0.037	2	0.654	0.132	0.242	2.154
hsd17b2	4327	1002	342415	4	2	6	0.075	4	0.654	0.132	0.242	2.154
banf1	4327	1002	342415	14	9	23	0.287	14	0.596	0.132	0.242	1.915
erbb2	4327	1002	342415	141	91	232	2.895	141	0.595	0.132	0.242	1.911
bub1b	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
cby1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
cd3e	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
cldn1	4327	1002	342415	2	2	4	0.050	2	0.488	0.132	0.242	1.466
cpm	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
fbxl15	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
fermt3	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
fstl1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
htatip2	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
klk7	4327	1002	342415	2	2	4	0.050	2	0.488	0.132	0.242	1.466
limk1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
mta3	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
nkx6-2	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
otud4	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
prkar1b	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
rnf11	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
runx3	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
scgb3a1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
slco3a1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
slco5a1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
smagp	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
tgfbr1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
tnk2	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
tns1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
usp33	4327	1002	342415	2	2	4	0.050	2	0.488	0.132	0.242	1.466
usp4	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
zeb1	4327	1002	342415	1	1	2	0.025	1	0.488	0.132	0.242	1.466
abcg2	4327	1002	342415	86	90	176	2.196	86	0.476	0.132	0.242	1.419
brca1	4327	1002	342415	15	18	33	0.412	15	0.442	0.132	0.242	1.279
psen2	4327	1002	342415	13	16	29	0.362	13	0.436	0.132	0.242	1.253
shmt1	4327	1002	342415	3	4	7	0.087	3	0.416	0.132	0.242	1.171
chek2	4327	1002	342415	2	3	5	0.062	2	0.388	0.132	0.242	1.053
ifngr2	4327	1002	342415	2	3	5	0.062	2	0.388	0.132	0.242	1.053
scn9a	4327	1002	342415	2	3	5	0.062	2	0.388	0.132	0.242	1.053
serpine2	4327	1002	342415	2	3	5	0.062	2	0.388	0.132	0.242	1.053
slc38a5	4327	1002	342415	2	3	5	0.062	2	0.388	0.132	0.242	1.053
tff1	4327	1002	342415	14	21	35	0.437	14	0.388	0.132	0.242	1.053

Table C.8. Breast cancer/SF text-mining results and calculations.



PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
casc3	18	21	7669	1	0	1	0.002	1	0.998	0.215	0.392	1.995
fcgrt	18	21	7669	1	0	1	0.002	1	0.998	0.215	0.392	1.995
igfbp7	18	21	7669	1	0	1	0.002	1	0.998	0.215	0.392	1.995
klk7	18	21	7669	1	0	1	0.002	1	0.998	0.215	0.392	1.995

Table C.9. Breast cancer/stool text-mining results and calculations.

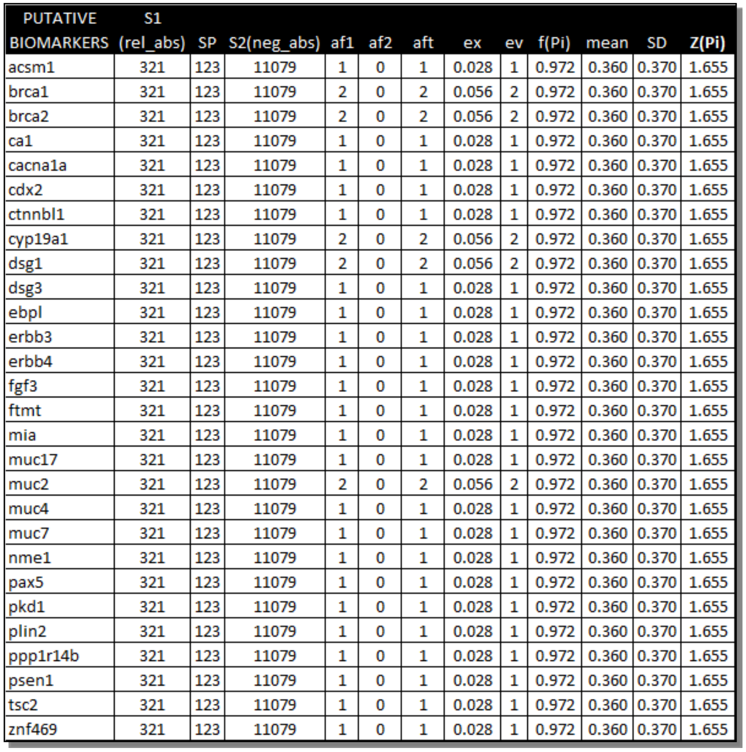


PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
cyp1a1	123	68	37574	1	0	1	0.003	1	0.997	0.157	0.277	3.035
cyp1b1	123	68	37574	1	0	1	0.003	1	0.997	0.157	0.277	3.035
ffar1	123	68	37574	1	0	1	0.003	1	0.997	0.157	0.277	3.035
mlh3	123	68	37574	1	0	1	0.003	1	0.997	0.157	0.277	3.035
msh2	123	68	37574	1	0	1	0.003	1	0.997	0.157	0.277	3.035
abcg2	123	68	37574	3	2	5	0.016	3	0.597	0.157	0.277	1.590
pcna	123	68	37574	1	1	2	0.007	1	0.497	0.157	0.277	1.228

Table C.10. Breast cancer/serum text-mining results and calculations.

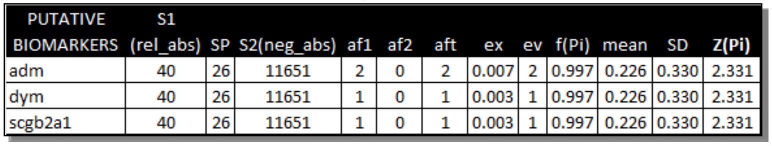
PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
arhgap1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
atg10	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
atg12	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
bdh1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
bdp1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
brms1	7410	1327	415218	3	0	3	0.053	3	0.982	0.131	0.230	3.698
brwd3	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
bst2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
ccdc14	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
cdk19	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
ceacam19	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
cirbp	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
clip1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
crtap	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
cspg5	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
cst6	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
dnm1l	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
ece2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
echdc1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
edn2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
fis1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
foxa3	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
gpaa1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
gper	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
grb14	7410	1327	415218	2	0	2	0.035	2	0.982	0.131	0.230	3.698
grb7	7410	1327	415218	2	0	2	0.035	2	0.982	0.131	0.230	3.698
greb1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
hecw1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
hipk2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
hmgcl	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
hmgcs2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
hpse2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
hrg	7410	1327	415218	2	0	2	0.035	2	0.982	0.131	0.230	3.698
jup	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
kcnj3	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
kcnj6	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
lmo7	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
meis2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
mfn1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
ndufaf4	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
nek3	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
nfix	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
nrg2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
ovca2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
pbov1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
pcyt2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
pde3b	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
ppm1d	7410	1327	415218	2	0	2	0.035	2	0.982	0.131	0.230	3.698
ppm1f	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
pvrl4	7410	1327	415218	3	0	3	0.053	3	0.982	0.131	0.230	3.698
rpa2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
s1pr2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
sepw1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
sgol1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
slc16a6	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
slc2a13	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
slc30a2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
st8sia1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
tmem66	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
vrk2	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
wif1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
wwp1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
zar1	7410	1327	415218	1	0	1	0.018	1	0.982	0.131	0.230	3.698
akr1b10	7410	1327	415218	2	1	3	0.053	2	0.649	0.131	0.230	2.250
atf5	7410	1327	415218	2	1	3	0.053	2	0.649	0.131	0.230	2.250
esrra	7410	1327	415218	2	1	3	0.053	2	0.649	0.131	0.230	2.250
htatip2	7410	1327	415218	2	1	3	0.053	2	0.649	0.131	0.230	2.250
spata1	7410	1327	415218	6	3	9	0.158	6	0.649	0.131	0.230	2.250
tpd52l1	7410	1327	415218	2	1	3	0.053	2	0.649	0.131	0.230	2.250
klk14	7410	1327	415218	3	2	5	0.088	3	0.582	0.131	0.230	1.960
pgrmc1	7410	1327	415218	3	2	5	0.088	3	0.582	0.131	0.230	1.960
postn	7410	1327	415218	3	2	5	0.088	3	0.582	0.131	0.230	1.960
erbb2	7410	1327	415218	375	254	629	11.028	375	0.579	0.131	0.230	1.944
brca1	7410	1327	415218	38	26	64	1.122	38	0.576	0.131	0.230	1.933
PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
psen2	7410	1327	415218	30	23	53	0.929	30	0.549	0.131	0.230	1.813
ncoa3	7410	1327	415218	11	10	21	0.368	11	0.506	0.131	0.230	1.629
tff1	7410	1327	415218	34	32	66	1.157	34	0.498	0.131	0.230	1.592
brca2	7410	1327	415218	14	14	28	0.491	14	0.482	0.131	0.230	1.526
erbb3	7410	1327	415218	17	17	34	0.596	17	0.482	0.131	0.230	1.526
prlr	7410	1327	415218	7	7	14	0.245	7	0.482	0.131	0.230	1.526
alpi	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
apeh	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
arx	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
bag3	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
bhlhe40	7410	1327	415218	2	2	4	0.070	2	0.482	0.131	0.230	1.526
ccnd2	7410	1327	415218	2	2	4	0.070	2	0.482	0.131	0.230	1.526
cdc34	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
cic	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
cited2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
cnksr2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
col4a2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
crmp1	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
csn2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
dap	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
denr	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
dnpep	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
dock2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
ecd	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
fgf3	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
gpr55	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
hoxc11	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
hsd17b2	7410	1327	415218	2	2	4	0.070	2	0.482	0.131	0.230	1.526
hsd17b7	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
hspa1b	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
il16	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
itpr3	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
kif5b	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
ksr2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
nell2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
pias3	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
rad52	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
serinc2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
sh2d3c	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
slc39a1	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
snai1	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
ssr3	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
stab2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
stra13	7410	1327	415218	2	2	4	0.070	2	0.482	0.131	0.230	1.526
thra	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
tk2	7410	1327	415218	3	3	6	0.105	3	0.482	0.131	0.230	1.526
ube2d2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
ube2l6	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
ube2s	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
ube3a	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
ube3b	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
uhrf1	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
usf2	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
utrn	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
vamp3	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
vamp8	7410	1327	415218	1	1	2	0.035	1	0.482	0.131	0.230	1.526
dnmt1	7410	1327	415218	3	4	7	0.123	3	0.411	0.131	0.230	1.216
nrg1	7410	1327	415218	6	8	14	0.245	6	0.411	0.131	0.230	1.216
tktl1	7410	1327	415218	3	4	7	0.123	3	0.411	0.131	0.230	1.216
rassf1	7410	1327	415218	24	33	57	0.999	24	0.404	0.131	0.230	1.183
wisp3	7410	1327	415218	5	7	12	0.210	5	0.399	0.131	0.230	1.164
cdh1	7410	1327	415218	7	10	17	0.298	7	0.394	0.131	0.230	1.143
ptp4a3	7410	1327	415218	7	10	17	0.298	7	0.394	0.131	0.230	1.143
ca12	7410	1327	415218	2	3	5	0.088	2	0.382	0.131	0.230	1.092
hk3	7410	1327	415218	6	9	15	0.263	6	0.382	0.131	0.230	1.092
hook3	7410	1327	415218	6	9	15	0.263	6	0.382	0.131	0.230	1.092
kcnn1	7410	1327	415218	2	3	5	0.088	2	0.382	0.131	0.230	1.092
myl9	7410	1327	415218	2	3	5	0.088	2	0.382	0.131	0.230	1.092
scgb3a1	7410	1327	415218	2	3	5	0.088	2	0.382	0.131	0.230	1.092
slit2	7410	1327	415218	2	3	5	0.088	2	0.382	0.131	0.230	1.092
slit3	7410	1327	415218	2	3	5	0.088	2	0.382	0.131	0.230	1.092

Table C.11. Breast cancer/sweat text-mining results and calculations.



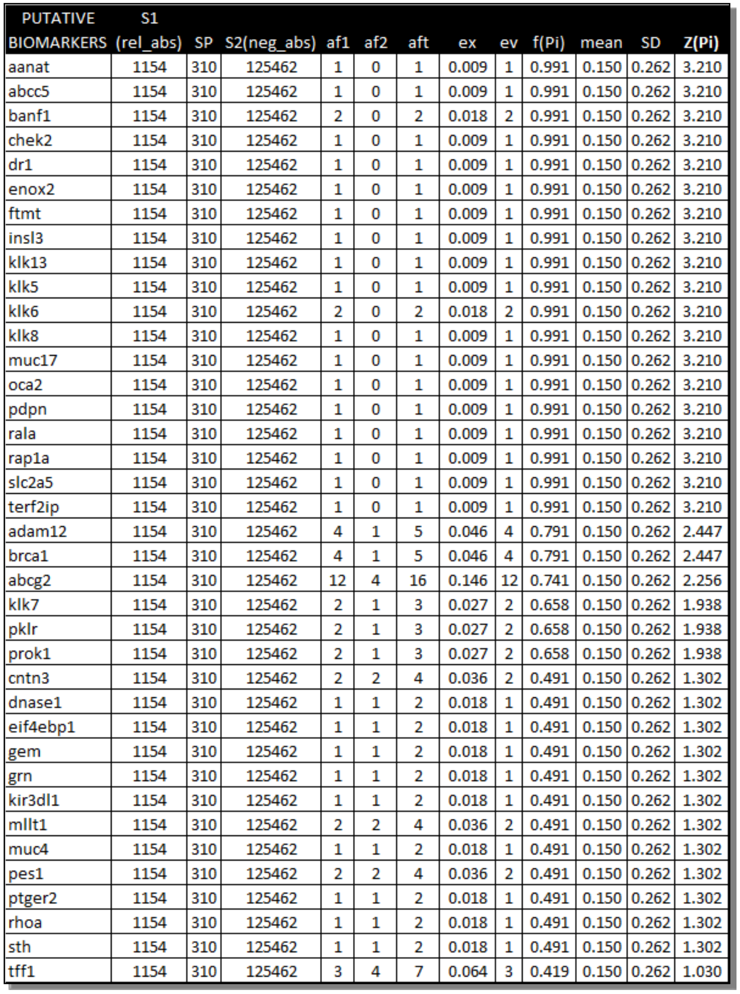
PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
acsm1	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
brca1	321	123	11079	2	0	2	0.056	2	0.972	0.360	0.370	1.655
brca2	321	123	11079	2	0	2	0.056	2	0.972	0.360	0.370	1.655
ca1	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
cacna1a	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
cdx2	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
ctnnbl1	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
cyp19a1	321	123	11079	2	0	2	0.056	2	0.972	0.360	0.370	1.655
dsg1	321	123	11079	2	0	2	0.056	2	0.972	0.360	0.370	1.655
dsg3	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
ebpl	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
erbb3	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
erbb4	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
fgf3	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
ftmt	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
mia	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
muc17	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
muc2	321	123	11079	2	0	2	0.056	2	0.972	0.360	0.370	1.655
muc4	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
muc7	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
nme1	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
pax5	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
pkd1	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
plin2	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
ppp1r14b	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
psen1	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
tsc2	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655
znf469	321	123	11079	1	0	1	0.028	1	0.972	0.360	0.370	1.655

Table C.12. Breast cancer/tears text-mining results and calculations.



PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
adm	40	26	11651	2	0	2	0.007	2	0.997	0.226	0.330	2.331
dym	40	26	11651	1	0	1	0.003	1	0.997	0.226	0.330	2.331
scgb2a1	40	26	11651	1	0	1	0.003	1	0.997	0.226	0.330	2.331

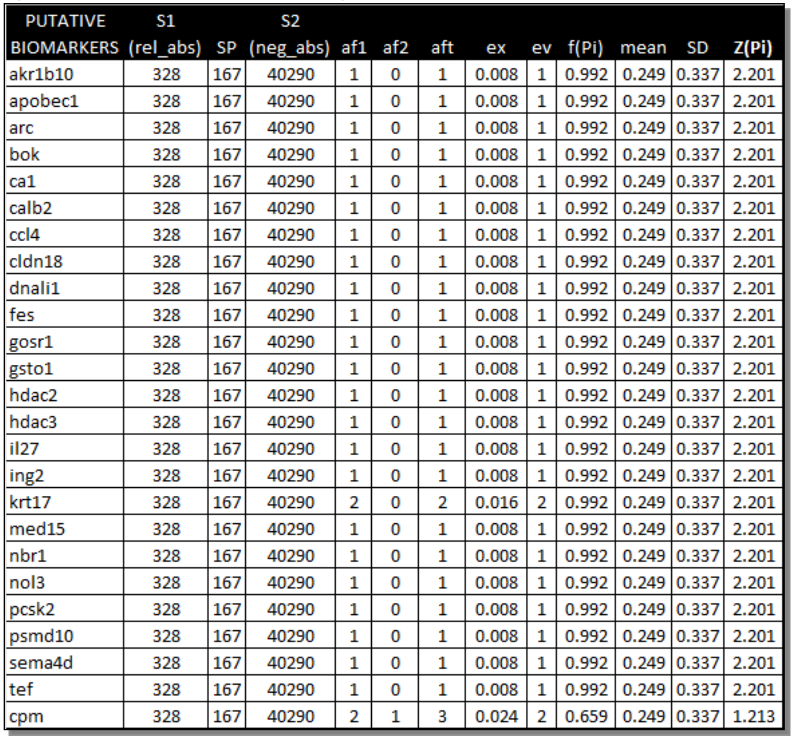
Table C.13. Breast cancer/urine text-mining results and calculations.



PUTATIVE	S1											
BIOMARKERS	(rel_abs)	SP	S2(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
aanat	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
abcc5	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
banf1	1154	310	125462	2	0	2	0.018	2	0.991	0.150	0.262	3.210
chek2	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
dr1	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
enox2	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
ftmt	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
insl3	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
klk13	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
klk5	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
klk6	1154	310	125462	2	0	2	0.018	2	0.991	0.150	0.262	3.210
klk8	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
muc17	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
oca2	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
pdpn	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
rala	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
rap1a	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
slc2a5	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
terf2ip	1154	310	125462	1	0	1	0.009	1	0.991	0.150	0.262	3.210
adam12	1154	310	125462	4	1	5	0.046	4	0.791	0.150	0.262	2.447
brca1	1154	310	125462	4	1	5	0.046	4	0.791	0.150	0.262	2.447
abcg2	1154	310	125462	12	4	16	0.146	12	0.741	0.150	0.262	2.256
klk7	1154	310	125462	2	1	3	0.027	2	0.658	0.150	0.262	1.938
pklr	1154	310	125462	2	1	3	0.027	2	0.658	0.150	0.262	1.938
prok1	1154	310	125462	2	1	3	0.027	2	0.658	0.150	0.262	1.938
cntn3	1154	310	125462	2	2	4	0.036	2	0.491	0.150	0.262	1.302
dnase1	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
eif4ebp1	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
gem	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
grn	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
kir3dl1	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
mllt1	1154	310	125462	2	2	4	0.036	2	0.491	0.150	0.262	1.302
muc4	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
pes1	1154	310	125462	2	2	4	0.036	2	0.491	0.150	0.262	1.302
ptger2	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
rhoa	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
sth	1154	310	125462	1	1	2	0.018	1	0.491	0.150	0.262	1.302
tff1	1154	310	125462	3	4	7	0.064	3	0.419	0.150	0.262	1.030

LUNG CANCER

Table C.14. Lung cancer/bile text-mining results and calculations.

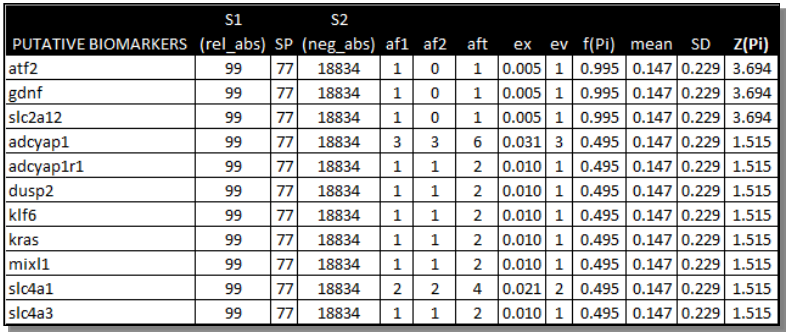


PUTATIVE BIOMARKERS	S1 (rel_abs)	SP	S2 (neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
akr1b10	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
apobec1	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
arc	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
bok	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
ca1	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
calb2	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
ccl4	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
cldn18	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
dnali1	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
fes	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
gosr1	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
gsto1	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
hdac2	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
hdac3	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
il27	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
ing2	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
krt17	328	167	40290	2	0	2	0.016	2	0.992	0.249	0.337	2.201
med15	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
nbr1	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
nol3	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
pcsk2	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
psmd10	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
sema4d	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
tef	328	167	40290	1	0	1	0.008	1	0.992	0.249	0.337	2.201
cpm	328	167	40290	2	1	3	0.024	2	0.659	0.249	0.337	1.213

Table C.15. Lung cancer/blood text-mining results and calculations.

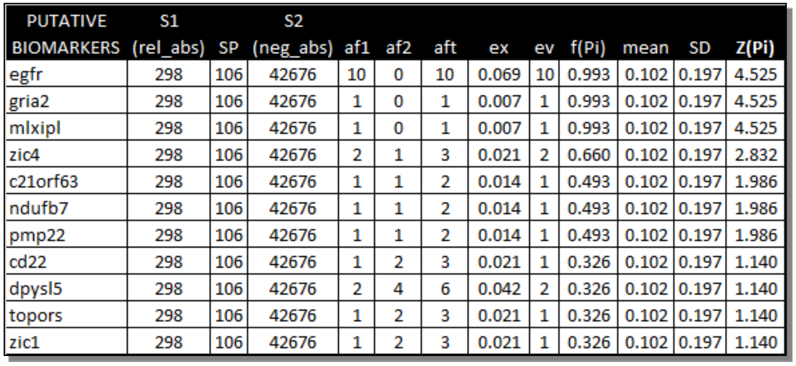
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
arl11	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
ash1l	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
atp6ap1	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
brsk2	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
btbd2	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
btbd3	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
c16orf80	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
card18	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
casc1	15710	1863	1522046	2	0	2	0.020	2	0.990	0.078	0.180	5.067
ccnb2	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
cdc45l	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
ciz1	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
cytsa	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
dll3	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
dpysl3	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
dync2h1	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
dyrk2	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
eml4	15710	1863	1522046	3	0	3	0.031	3	0.990	0.078	0.180	5.067
fam83a	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
galnt14	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
glra3	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
gnal	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
hif1an	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
kif18a	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
kif5a	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
klk12	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
lhx6	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
liph	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
nfkbiz	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
ociad2	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
ppfia1	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
ppp2r2a	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
rab23	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
rnf17	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
sox1	15710	1863	1522046	6	0	6	0.061	6	0.990	0.078	0.180	5.067
sox21	15710	1863	1522046	2	0	2	0.020	2	0.990	0.078	0.180	5.067
stk11ip	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
tmem189-ube2v1	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
tnfrsf1a	15710	1863	1522046	3	0	3	0.031	3	0.990	0.078	0.180	5.067
tomm34	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
tp53i11	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
tusc2	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
ubqln3	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
zic4	15710	1863	1522046	3	0	3	0.031	3	0.990	0.078	0.180	5.067
zmat3	15710	1863	1522046	1	0	1	0.010	1	0.990	0.078	0.180	5.067
xage1a	15710	1863	1522046	8	2	10	0.102	8	0.790	0.078	0.180	3.956
clca2	15710	1863	1522046	3	1	4	0.041	3	0.740	0.078	0.180	3.678
mpp2	15710	1863	1522046	3	1	4	0.041	3	0.740	0.078	0.180	3.678
zic2	15710	1863	1522046	3	1	4	0.041	3	0.740	0.078	0.180	3.678
mpp3	15710	1863	1522046	2	1	3	0.031	2	0.656	0.078	0.180	3.215
shox2	15710	1863	1522046	2	1	3	0.031	2	0.656	0.078	0.180	3.215
dpysl5	15710	1863	1522046	5	3	8	0.082	5	0.615	0.078	0.180	2.984
kras	15710	1863	1522046	8	5	13	0.133	8	0.605	0.078	0.180	2.930
actn4	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
anapc11	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
avpr1b	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
cxcl17	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
dlg2	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
dlg3	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
edil3	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
enc1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
enox2	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
flj11535	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
gas7	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
golga2	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
gpr87	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
hat1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
hmgb3	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
immp2l	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
irx1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
itga3	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
klk11	15710	1863	1522046	2	2	4	0.041	2	0.490	0.078	0.180	2.289
klk13	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
klk8	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
ldhd	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
lpar6	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
mycl1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
nfyb	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
nfyc	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
pak4	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
pbld	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
plxdc2	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
prmt6	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
pvrl4	15710	1863	1522046	2	2	4	0.041	2	0.490	0.078	0.180	2.289
recql	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
rgs11	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
rnf43	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
rpl19	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
rtel1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
siglec6	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
sit1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
slc6a20	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
slit1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
sox3	15710	1863	1522046	2	2	4	0.041	2	0.490	0.078	0.180	2.289
spanxc	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
sugt1	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
tsr2	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
vegfa	15710	1863	1522046	2	2	4	0.041	2	0.490	0.078	0.180	2.289
wnt7b	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
znf165	15710	1863	1522046	1	1	2	0.020	1	0.490	0.078	0.180	2.289
hnrnpa2b1	15710	1863	1522046	3	3	6	0.061	3	0.490	0.078	0.180	2.289
elavl4	15710	1863	1522046	17	18	35	0.358	17	0.475	0.078	0.180	2.210
cndp2	15710	1863	1522046	3	4	7	0.072	3	0.418	0.078	0.180	1.893
adam28	15710	1863	1522046	5	7	12	0.123	5	0.406	0.078	0.180	1.827
chrna3	15710	1863	1522046	2	3	5	0.051	2	0.390	0.078	0.180	1.734
chrna5	15710	1863	1522046	2	3	5	0.051	2	0.390	0.078	0.180	1.734
gadd45g	15710	1863	1522046	2	3	5	0.051	2	0.390	0.078	0.180	1.734
gstm4	15710	1863	1522046	2	3	5	0.051	2	0.390	0.078	0.180	1.734
ercc1	15710	1863	1522046	27	41	68	0.695	27	0.387	0.078	0.180	1.718
adc	15710	1863	1522046	3	5	8	0.082	3	0.365	0.078	0.180	1.595
akap3	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
arid1a	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
bub1	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
ccdc6	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
cep55	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
cnksr2	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
cspg4	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
cyp2a13	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
dleu7	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
glra1	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
hrg	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
imp2	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
kcnh4	15710	1863	1522046	4	8	12	0.123	4	0.323	0.078	0.180	1.364
kcnip4	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
ly6k	15710	1863	1522046	2	4	6	0.061	2	0.323	0.078	0.180	1.364
magea2	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
mlst8	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
mmp28	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
ms4a3	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
nmbr	15710	1863	1522046	2	4	6	0.061	2	0.323	0.078	0.180	1.364
pdgfc	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
plagl2	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
ptn	15710	1863	1522046	2	4	6	0.061	2	0.323	0.078	0.180	1.364
ptpn14	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
ptprg	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
rab27b	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
rgs19	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
scg3	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
slc30a2	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
ss18	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
stra13	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
stxbp5	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
tjp3	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
tpm4	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
ube2v1	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
vamp3	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
vash1	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
znf35	15710	1863	1522046	1	2	3	0.031	1	0.323	0.078	0.180	1.364
hdgf	15710	1863	1522046	3	7	10	0.102	3	0.290	0.078	0.180	1.178
l1cam	15710	1863	1522046	2	5	7	0.072	2	0.275	0.078	0.180	1.099
ntm	15710	1863	1522046	2	5	7	0.072	2	0.275	0.078	0.180	1.099
saa2	15710	1863	1522046	2	5	7	0.072	2	0.275	0.078	0.180	1.099

Table C.16. Lung cancer/breastmilk text-mining results and calculations.



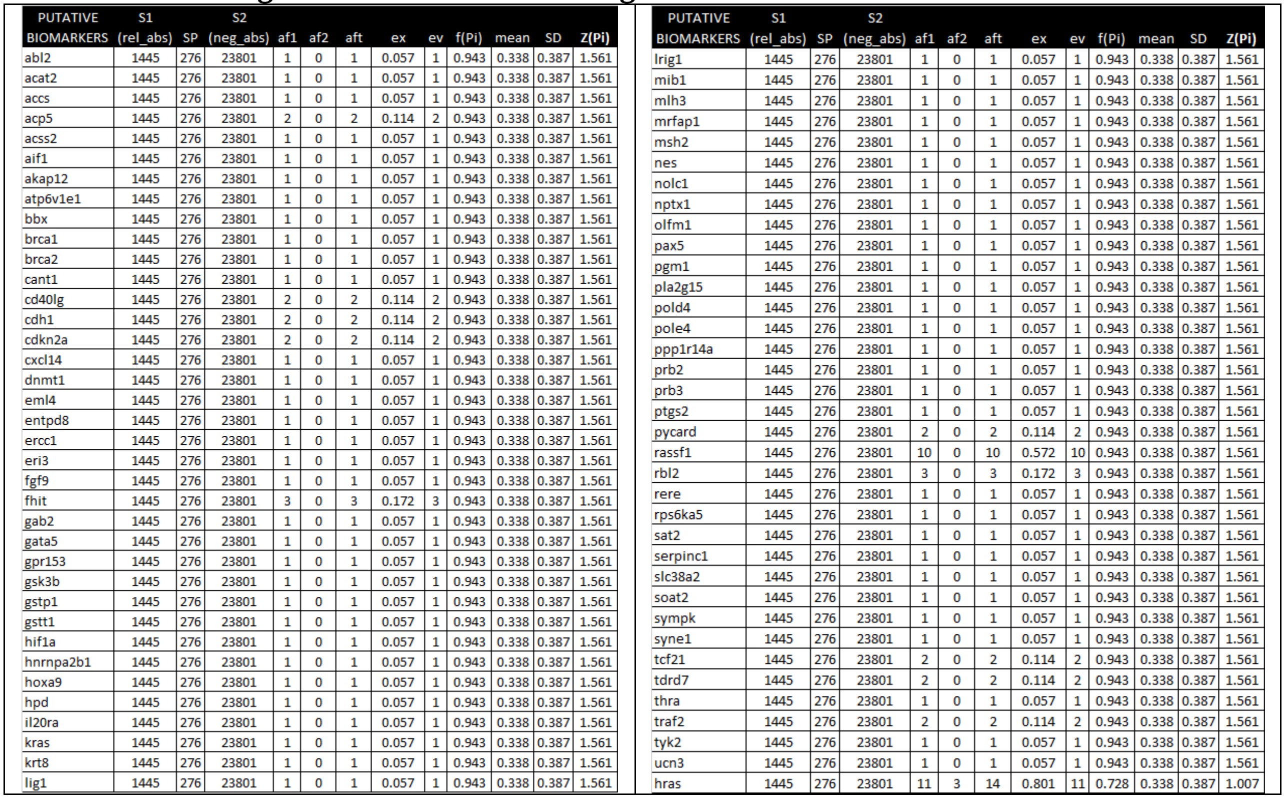
	S1		S2									
PUTATIVE BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
atf2	99	77	18834	1	0	1	0.005	1	0.995	0.147	0.229	3.694
gdnf	99	77	18834	1	0	1	0.005	1	0.995	0.147	0.229	3.694
slc2a12	99	77	18834	1	0	1	0.005	1	0.995	0.147	0.229	3.694
adcyap1	99	77	18834	3	3	6	0.031	3	0.495	0.147	0.229	1.515
adcyap1r1	99	77	18834	1	1	2	0.010	1	0.495	0.147	0.229	1.515
dusp2	99	77	18834	1	1	2	0.010	1	0.495	0.147	0.229	1.515
klf6	99	77	18834	1	1	2	0.010	1	0.495	0.147	0.229	1.515
kras	99	77	18834	1	1	2	0.010	1	0.495	0.147	0.229	1.515
mixl1	99	77	18834	1	1	2	0.010	1	0.495	0.147	0.229	1.515
slc4a1	99	77	18834	2	2	4	0.021	2	0.495	0.147	0.229	1.515
slc4a3	99	77	18834	1	1	2	0.010	1	0.495	0.147	0.229	1.515

Table C.17. Lung cancer/CSF text-mining results and calculations.



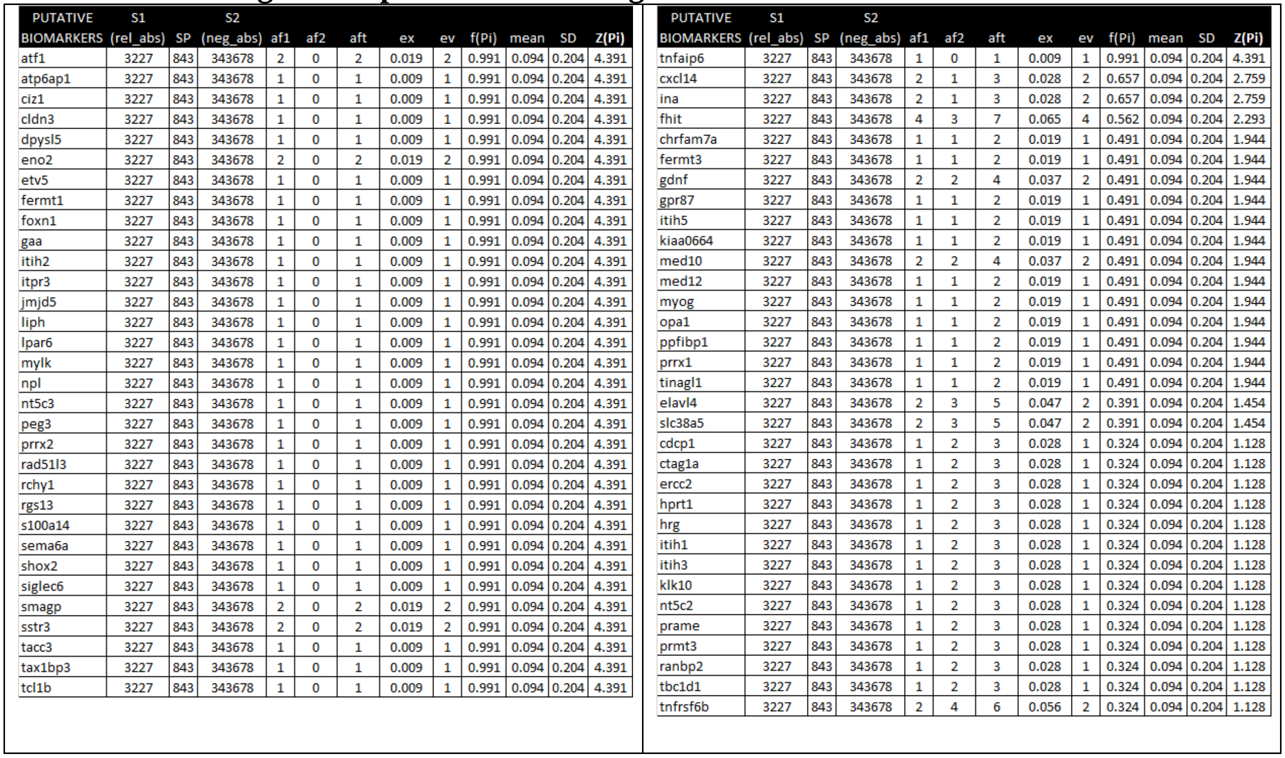
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
egfr	298	106	42676	10	0	10	0.069	10	0.993	0.102	0.197	4.525
gria2	298	106	42676	1	0	1	0.007	1	0.993	0.102	0.197	4.525
mlxipl	298	106	42676	1	0	1	0.007	1	0.993	0.102	0.197	4.525
zic4	298	106	42676	2	1	3	0.021	2	0.660	0.102	0.197	2.832
c21orf63	298	106	42676	1	1	2	0.014	1	0.493	0.102	0.197	1.986
ndufb7	298	106	42676	1	1	2	0.014	1	0.493	0.102	0.197	1.986
pmp22	298	106	42676	1	1	2	0.014	1	0.493	0.102	0.197	1.986
cd22	298	106	42676	1	2	3	0.021	1	0.326	0.102	0.197	1.140
dpysl5	298	106	42676	2	4	6	0.042	2	0.326	0.102	0.197	1.140
topors	298	106	42676	1	2	3	0.021	1	0.326	0.102	0.197	1.140
zic1	298	106	42676	1	2	3	0.021	1	0.326	0.102	0.197	1.140

Table C.18. Lung cancer/mucus text-mining results and calculations.



PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
abl2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
acat2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
accs	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
acp5	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
acss2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
aif1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
akap12	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
atp6v1e1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
bbx	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
brca1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
brca2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
cant1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
cd40lg	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
cdh1	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
cdkn2a	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
cxcl14	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
dnmt1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
eml4	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
entpd8	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
ercc1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
eri3	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
fgf9	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
fhit	1445	276	23801	3	0	3	0.172	3	0.943	0.338	0.387	1.561
gab2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
gata5	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
gpr153	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
gsk3b	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
gstp1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
gstt1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
hif1a	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
hnrnpa2b1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
hoxa9	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
hpd	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
il20ra	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
kras	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
krt8	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
lig1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
lrig1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
mib1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
mlh3	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
mrfap1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
msh2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
nes	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
nolc1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
nptx1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
olfm1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
pax5	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
pgm1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
pla2g15	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
pold4	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
pole4	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
ppp1r14a	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
prb2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
prb3	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
ptgs2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
pycard	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
rassf1	1445	276	23801	10	0	10	0.572	10	0.943	0.338	0.387	1.561
rbl2	1445	276	23801	3	0	3	0.172	3	0.943	0.338	0.387	1.561
rere	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
rps6ka5	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
sat2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
serpinc1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
slc38a2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
soat2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
sympk	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
syne1	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
tcf21	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
tdrd7	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
thra	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
traf2	1445	276	23801	2	0	2	0.114	2	0.943	0.338	0.387	1.561
tyk2	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
ucn3	1445	276	23801	1	0	1	0.057	1	0.943	0.338	0.387	1.561
hras	1445	276	23801	11	3	14	0.801	11	0.728	0.338	0.387	1.007

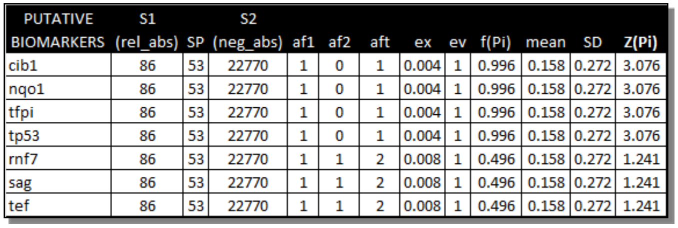
Table C.19. Lung cancer/plasma text-mining results and calculations.



PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
atf1	3227	843	343678	2	0	2	0.019	2	0.991	0.094	0.204	4.391
atp6ap1	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
ciz1	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
cldn3	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
dpysl5	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
eno2	3227	843	343678	2	0	2	0.019	2	0.991	0.094	0.204	4.391
etv5	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
fermt1	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
foxn1	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
gaa	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
itih2	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
itpr3	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
jmjd5	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
liph	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
lpar6	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
mylk	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
npl	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
nt5c3	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
peg3	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
prrx2	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
rad51l3	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
rchy1	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
rgs13	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
s100a14	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
sema6a	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
shox2	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
siglec6	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
smagp	3227	843	343678	2	0	2	0.019	2	0.991	0.094	0.204	4.391
sstr3	3227	843	343678	2	0	2	0.019	2	0.991	0.094	0.204	4.391
tacc3	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
tax1bp3	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
tcl1b	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391

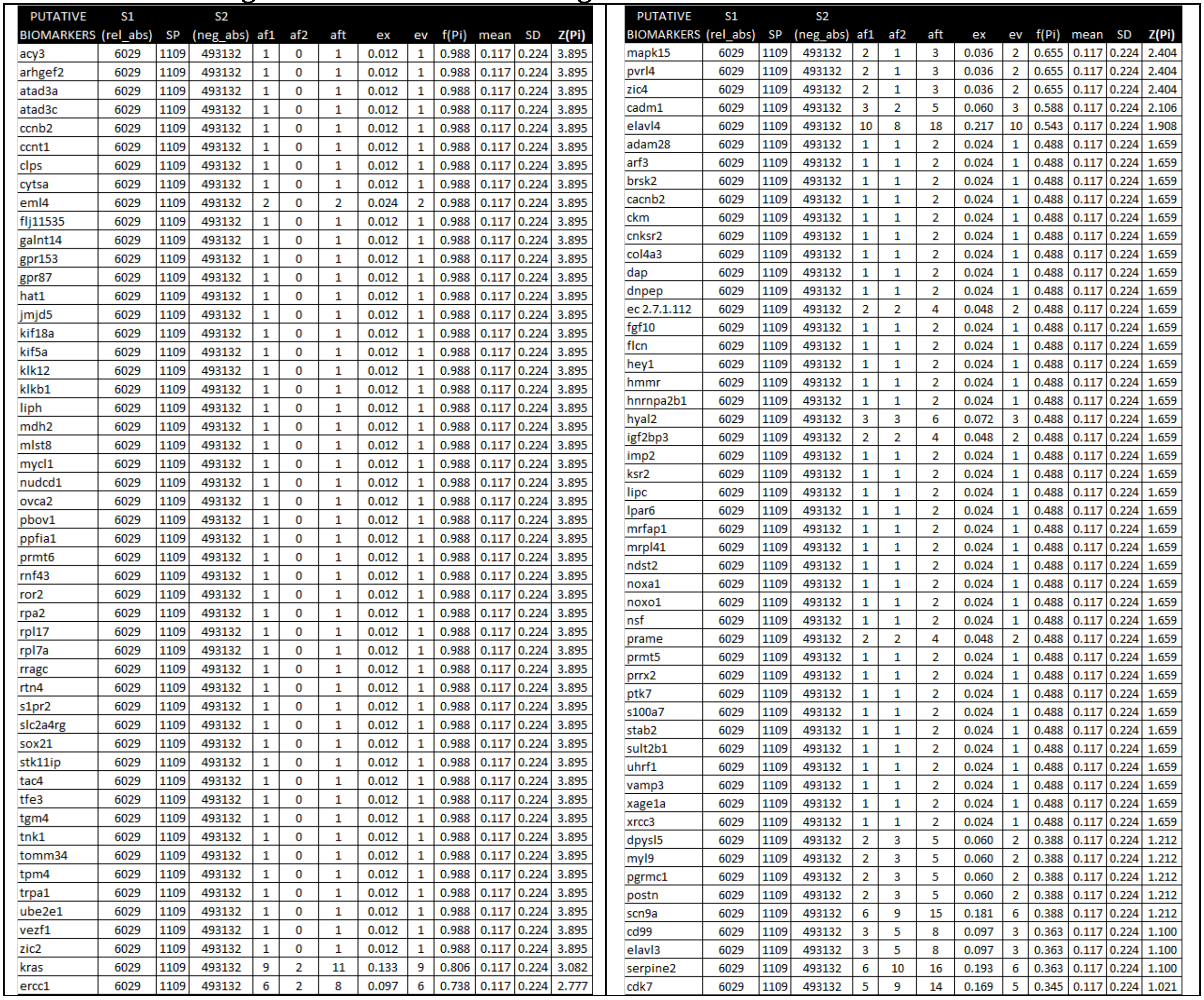
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
tnfaip6	3227	843	343678	1	0	1	0.009	1	0.991	0.094	0.204	4.391
cxcl14	3227	843	343678	2	1	3	0.028	2	0.657	0.094	0.204	2.759
ina	3227	843	343678	2	1	3	0.028	2	0.657	0.094	0.204	2.759
fhit	3227	843	343678	4	3	7	0.065	4	0.562	0.094	0.204	2.293
chrfam7a	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
fermt3	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
gdnf	3227	843	343678	2	2	4	0.037	2	0.491	0.094	0.204	1.944
gpr87	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
itih5	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
kiaa0664	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
med10	3227	843	343678	2	2	4	0.037	2	0.491	0.094	0.204	1.944
med12	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
myog	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
opa1	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
ppfibp1	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
prrx1	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
tinagl1	3227	843	343678	1	1	2	0.019	1	0.491	0.094	0.204	1.944
elavl4	3227	843	343678	2	3	5	0.047	2	0.391	0.094	0.204	1.454
slc38a5	3227	843	343678	2	3	5	0.047	2	0.391	0.094	0.204	1.454
cdcp1	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
ctag1a	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
ercc2	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
hprt1	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
hrg	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
itih1	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
itih3	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
klk10	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
nt5c2	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
prame	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
prmt3	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
ranbp2	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
tbc1d1	3227	843	343678	1	2	3	0.028	1	0.324	0.094	0.204	1.128
tnfrsf6b	3227	843	343678	2	4	6	0.056	2	0.324	0.094	0.204	1.128

Table C.20. Lung cancer/saliva text-mining results and calculations.



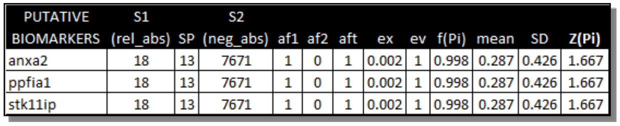
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
cib1	86	53	22770	1	0	1	0.004	1	0.996	0.158	0.272	3.076
nqo1	86	53	22770	1	0	1	0.004	1	0.996	0.158	0.272	3.076
tfpi	86	53	22770	1	0	1	0.004	1	0.996	0.158	0.272	3.076
tp53	86	53	22770	1	0	1	0.004	1	0.996	0.158	0.272	3.076
rnf7	86	53	22770	1	1	2	0.008	1	0.496	0.158	0.272	1.241
sag	86	53	22770	1	1	2	0.008	1	0.496	0.158	0.272	1.241
tef	86	53	22770	1	1	2	0.008	1	0.496	0.158	0.272	1.241

Table C.21. Lung cancer/serum text-mining results and calculations.



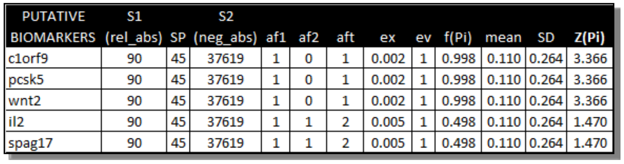
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
acy3	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
arhgef2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
atad3a	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
atad3c	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
ccnb2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
ccnt1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
clps	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
cytsa	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
eml4	6029	1109	493132	2	0	2	0.024	2	0.988	0.117	0.224	3.895
flj11535	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
galnt14	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
gpr153	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
gpr87	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
hat1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
jmjd5	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
kif18a	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
kif5a	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
klk12	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
klkb1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
liph	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
mdh2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
mlst8	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
mycl1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
nudcd1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
ovca2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
pbov1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
ppfia1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
prmt6	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
rnf43	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
ror2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
rpa2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
rpl17	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
rpl7a	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
rragc	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
rtn4	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
s1pr2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
slc2a4rg	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
sox21	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
stk11ip	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
tac4	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
tfe3	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
tgm4	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
tnk1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
tomm34	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
tpm4	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
trpa1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
ube2e1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
vezf1	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
zic2	6029	1109	493132	1	0	1	0.012	1	0.988	0.117	0.224	3.895
kras	6029	1109	493132	9	2	11	0.133	9	0.806	0.117	0.224	3.082
ercc1	6029	1109	493132	6	2	8	0.097	6	0.738	0.117	0.224	2.777
mapk15	6029	1109	493132	2	1	3	0.036	2	0.655	0.117	0.224	2.404
pvrl4	6029	1109	493132	2	1	3	0.036	2	0.655	0.117	0.224	2.404
zic4	6029	1109	493132	2	1	3	0.036	2	0.655	0.117	0.224	2.404
cadm1	6029	1109	493132	3	2	5	0.060	3	0.588	0.117	0.224	2.106
elavl4	6029	1109	493132	10	8	18	0.217	10	0.543	0.117	0.224	1.908
adam28	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
arf3	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
brsk2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
cacnb2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
ckm	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
cnksr2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
col4a3	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
dap	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
dnpep	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
ec 2.7.1.112	6029	1109	493132	2	2	4	0.048	2	0.488	0.117	0.224	1.659
fgf10	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
flcn	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
hey1	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
hmmr	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
hnrnpa2b1	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
hyal2	6029	1109	493132	3	3	6	0.072	3	0.488	0.117	0.224	1.659
igf2bp3	6029	1109	493132	2	2	4	0.048	2	0.488	0.117	0.224	1.659
imp2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
ksr2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
lipc	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
lpar6	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
mrfap1	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
mrpl41	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
ndst2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
noxa1	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
noxo1	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
nsf	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
prame	6029	1109	493132	2	2	4	0.048	2	0.488	0.117	0.224	1.659
prmt5	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
prrx2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
ptk7	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
s100a7	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
stab2	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
sult2b1	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
uhrf1	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
vamp3	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
xage1a	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
xrcc3	6029	1109	493132	1	1	2	0.024	1	0.488	0.117	0.224	1.659
dpysl5	6029	1109	493132	2	3	5	0.060	2	0.388	0.117	0.224	1.212
myl9	6029	1109	493132	2	3	5	0.060	2	0.388	0.117	0.224	1.212
pgrmc1	6029	1109	493132	2	3	5	0.060	2	0.388	0.117	0.224	1.212
postn	6029	1109	493132	2	3	5	0.060	2	0.388	0.117	0.224	1.212
scn9a	6029	1109	493132	6	9	15	0.181	6	0.388	0.117	0.224	1.212
cd99	6029	1109	493132	3	5	8	0.097	3	0.363	0.117	0.224	1.100
elavl3	6029	1109	493132	3	5	8	0.097	3	0.363	0.117	0.224	1.100
serpine2	6029	1109	493132	6	10	16	0.193	6	0.363	0.117	0.224	1.100
cdk7	6029	1109	493132	5	9	14	0.169	5	0.345	0.117	0.224	1.021

Table C.22. Lung cancer/SF text-mining results and calculations.



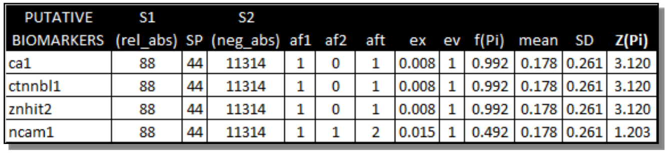
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
anxa2	18	13	7671	1	0	1	0.002	1	0.998	0.287	0.426	1.667
ppfia1	18	13	7671	1	0	1	0.002	1	0.998	0.287	0.426	1.667
stk11ip	18	13	7671	1	0	1	0.002	1	0.998	0.287	0.426	1.667

Table C.23. Lung cancer/stool text-mining results and calculations.



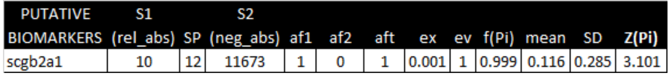
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
c1orf9	90	45	37619	1	0	1	0.002	1	0.998	0.110	0.264	3.366
pcsk5	90	45	37619	1	0	1	0.002	1	0.998	0.110	0.264	3.366
wnt2	90	45	37619	1	0	1	0.002	1	0.998	0.110	0.264	3.366
il2	90	45	37619	1	1	2	0.005	1	0.498	0.110	0.264	1.470
spag17	90	45	37619	1	1	2	0.005	1	0.498	0.110	0.264	1.470

Table C.24. Lung cancer/sweat text-mining results and calculations.



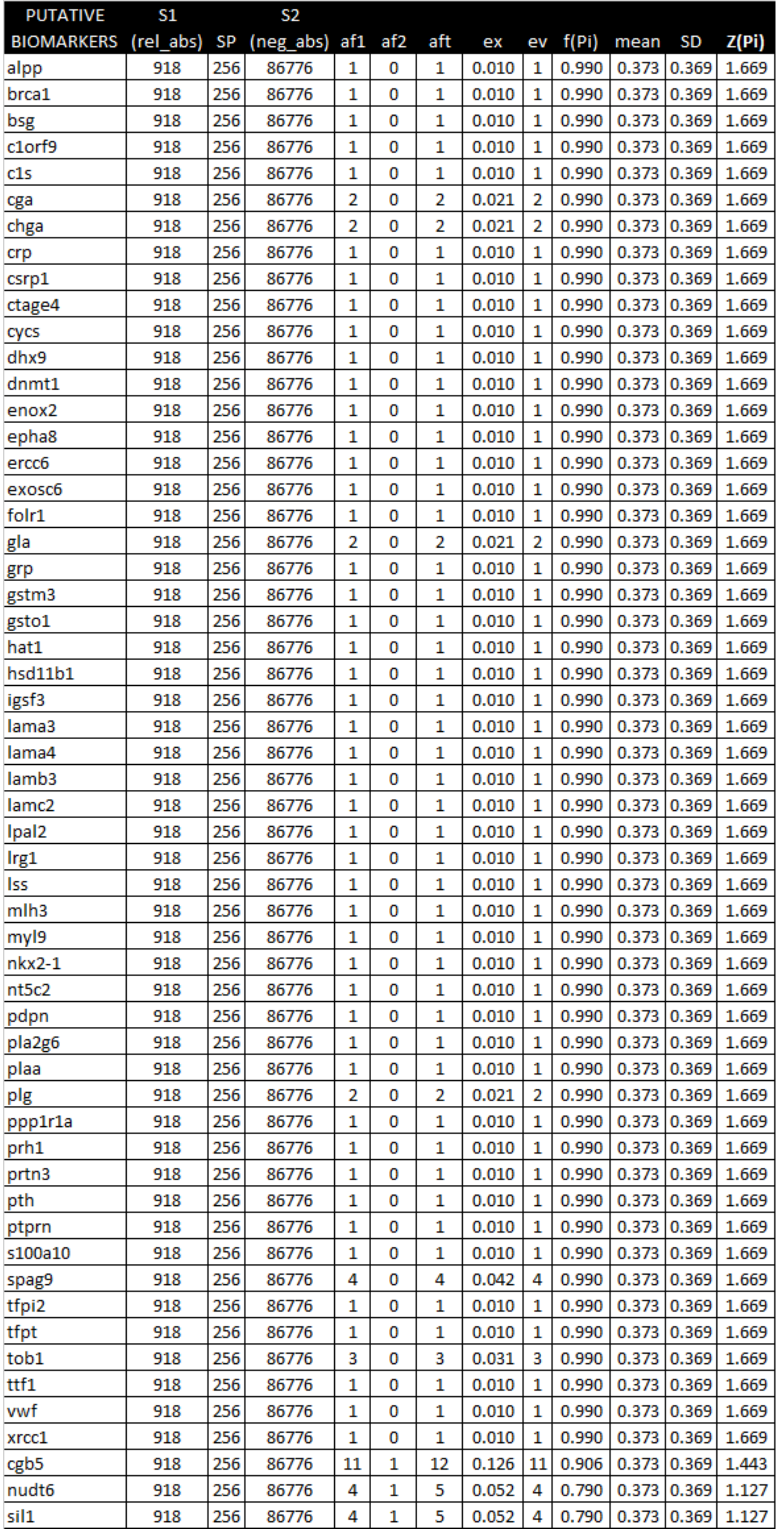
PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
ca1	88	44	11314	1	0	1	0.008	1	0.992	0.178	0.261	3.120
ctnnbl1	88	44	11314	1	0	1	0.008	1	0.992	0.178	0.261	3.120
znhit2	88	44	11314	1	0	1	0.008	1	0.992	0.178	0.261	3.120
ncam1	88	44	11314	1	1	2	0.015	1	0.492	0.178	0.261	1.203

Table C.25. Lung cancer/tears text-mining results and calculations.



PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	Z(Pi)
scgb2a1	10	12	11673	1	0	1	0.001	1	0.999	0.116	0.285	3.101

Table C.26. Lung cancer/urine text-mining results and calculations.



PUTATIVE	S1		S2									
BIOMARKERS	(rel_abs)	SP	(neg_abs)	af1	af2	aft	ex	ev	f(Pi)	mean	SD	**Z(Pi)**
alpp	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
brca1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
bsg	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
c1orf9	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
c1s	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
cga	918	256	86776	2	0	2	0.021	2	0.990	0.373	0.369	1.669
chga	918	256	86776	2	0	2	0.021	2	0.990	0.373	0.369	1.669
crp	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
csrp1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
ctage4	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
cycs	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
dhx9	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
dnmt1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
enox2	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
epha8	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
ercc6	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
exosc6	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
folr1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
gla	918	256	86776	2	0	2	0.021	2	0.990	0.373	0.369	1.669
grp	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
gstm3	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
gsto1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
hat1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
hsd11b1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
igsf3	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
lama3	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
lama4	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
lamb3	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
lamc2	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
lpal2	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
lrg1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
lss	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
mlh3	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
myl9	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
nkx2-1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
nt5c2	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
pdpn	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
pla2g6	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
plaa	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
plg	918	256	86776	2	0	2	0.021	2	0.990	0.373	0.369	1.669
ppp1r1a	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
prh1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
prtn3	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
pth	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
ptprn	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
s100a10	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
spag9	918	256	86776	4	0	4	0.042	4	0.990	0.373	0.369	1.669
tfpi2	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
tfpt	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
tob1	918	256	86776	3	0	3	0.031	3	0.990	0.373	0.369	1.669
ttf1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
vwf	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
xrcc1	918	256	86776	1	0	1	0.010	1	0.990	0.373	0.369	1.669
cgb5	918	256	86776	11	1	12	0.126	11	0.906	0.373	0.369	1.443
nudt6	918	256	86776	4	1	5	0.052	4	0.790	0.373	0.369	1.127
sil1	918	256	86776	4	1	5	0.052	4	0.790	0.373	0.369	1.127

MODELING RESULTS

Table C.27. Breast cancer modeling results

Data type	Comparison	Prior	Accuracy	SN	SP	Balanced Accuracy	Model Size	Union of variables used
Copy Number	Cancer vs. Normal	Uniform	100	100 100	100 100	100 100	2	WAS, KRTHB1, PJA1, AK125051
Copy Number	Cancer vs. Normal	Orig	100	100 100	100 100	100 100	1	WAS, AK125051 ,KRTHB1
Copy Number	Cancer vs. Normal	Ratio	100	100 100	100 100	100 100	2	WAS, KRTHB1 ,PJA1, AK125051
Methylation	Tumor vs. Normal	Ratio	96.37	100 0	0 100	50 50	5	CLDN19, KCNJ14, MUC15, TGFBR1, OR2H1, C1orf118, ITIH2, KCNC1, CLDN15, KCNJ14, SLFN3, C4orf8, DNASE1L2, MGC35048
Methylation	Tumor vs. Normal	Uniform	96.77	100 0	11.11 100	55.56 50	6	MUC15, TGFBR1, ADRA1A, KRTAP19-5, KCNJ14, CLDN19, C1orf177, MGC35048, C4orf8, C1orf118, CLDN15, PLXDC1, KCNC1, SLFN3, DNASE1L2
Methylation	Tumor vs. Normal	Orig	96.76	100 0	0 100	50 50	5	KCNJ14, ITIH2, MUC15, CLDN19, TGFBR1, MGC35048, C1orf177, KRTAP19-5, C4orf8, C1orf118, CLDN15
Methylation	Grade 1 vs 2	Ratio	71.21	86.79 0	15.39 100	51.09 50	9	C20orf177, CEACAM7, POT1, GJB6, CKMT2, TSPAN32, SH3BGRL3, FLJ44881, MUC17, HRC, C3orf22, THPO, KLHL6, CUEDC1, IL27, DYDC1, CPT1B, POP2, IBRDC1, PRODH2, IAPP, SEC61A2, LRMP, DOK5, PRKCDBP, TMC8, IL22, UTS2, FLJ90579, VWCE, IL10
Methylation	Grade 1 vs 2	Uniform	65.15	79.25 0	15.39 100	47.32 50	9	C20orf177, POP2, CKMT2, FLJ90579, IBRDC1, SH3BGRL3, FLJ44881, GJB6, THPO, ABCA3, CUEDC1, DTL, LRMP, DOK5, MUC17, HYI, CPT1B, FLJ42486, NALP8, PRODH2, TSPAN32, IAPP, SEC61A2, POT1, TMC8, NOX1, RASIP1, VWCE, C3orf22, ST6GALNAC3
Methylation	Grade 1 vs 2	None	20	1.89 100	100 1.89	50.94 50.94	10	POP2, LRMP, CKMT2, VWCE, GJB6, DTL, FLJ90579, NOX1, FLJ00060, CUEDC1, THPO, FLJ44881, SH3BGRL3, CEACAM7, FLJ42461, C3orf22, IAPP, C20orf177, INHBE,, CASP10, PROKR2, TSPAN32, PRODH2, DOK5, MFSD7, FLJ34922, PRKCDBP, C20orf177
Methylation	Grade N vs 1	Ratio	75.81	86.79 0	11.11 100	48.95 50	3	ZP4, GPR141, C1orf177, CLDN15, CHST3, C4orf8, FLJ10781, FOLR1, SLC4A11, SPARCL1, REM1
Methylation	Grade N vs 1	Orig	18.03	5.66 100	100 5.66	52.83 52.83	1	CHST3, DDAH2, C4orf8, FOLR1
Methylation	Grade N vs 1	Uniform	75.81	86.79 0	11.11 100	48.95 50	3	ZP4, CLDN19, ZBTB7B, CLDN15, CHST3, C4orf8, SPARCL1, FLJ10781, PDE9A, CYTL1, GPR141, REM1, SETBP1, HOM-TES-103
Methylation	Grade Nv2	Ratio	66.67	100 12.5	22.22 100	61.11 56.25	3	FLJ30058, LIMS3, DMBX1, PRKCB1, PCOLCE, SGCB, CLDN15
Methylation	Grade Nv2	Uniform	66.67	100 12.5	22.22 100	61.11 56.25	3	FLJ30058, LIMS3, DMBX1, PRKCB1, SGCB, PCOLCE, CLDN15
Methylation	Grade Nv2	Orig	100	100 100	100 100	100 100	1	LIMS, PRKCB1, FLJ30058, CLTB, ANGPTL2, LIMS3, CLDN15, PDPN

Methylation	GradeNv3	Orig	95.53	100 0	0 100	50 50	2	CLDN19, GPR132, CLDN15, C4orf8
Methylation	GradeNv3	Uniform	95.56	100 0	11.11 100	55.56 50	2	CLDN19, FLJ23657, CLDN15, C4orf8, GPR132
Methylation	GradeNv3	Ratio	99.44	100 100	100 99.42	100 99.71	3	KCNC1, CLDN19, C1S, C1orf118, C4orf8, GPR132, CLDN15
Microarray	Blood	Orig	57.025	100 3.7	3.7 100	51.85 51.85	15	AIF1, hCG2023505, LOC56181, ST8SIA4, CA1, CDK5R1, CD22, AJ223366.1, ZNF638, OR56B4, ACPT, P2RY14, KIAA0196, hCG1787898.2, hCG2007944, SPATA11, PF4V1, TBRG1, RPS23, HBQ1, SRRM2, hCG1642749.1, FBXO3, KIAA0196, TESK2, GGA2, hCG2014315, BCL2A1, DXYS155E, SCRN1, hCG1983348, USP10, RPS25, hCG1642170.3 , FLJ20160, ZNF638, Cep192, hCG2041718, ANXA3, hCG2041813
Microarray	Blood	Ratio	54.92	100 0	0 100	50 50	22	USP52, ZNF3, UCN, WIRE, NELL2, ZFP91, Z27499.1_CDS_1, ZMAT2, ZFP36L2, XM_373795, ZSWIM3, unk91, unk90, unk97, TBX21, unk59, PARC, ZNF638, ZNF652, LOC91526, ZCCHC14, ZC3HDC7, USP10, USP3, TM4SF13, DXYS155E, unk163, unk173, WSB2, TRAP1, ZNFN1A1, unk47, GFOD1, URP, UTRN, UGT2A1, TM4SF9, TRIM23, USP10, hCG1747327.2, ZF, RPS23, WARS, TUBB6, YWHAQ, SYK, TTBK1, unk57, SPN, UPK3B, UCP2, CA1, UNQ5783, ORM1, unk68, PPARA, SLC2A3, FLJ42953, unk113, C10orf33, IL2RB, NCOR2, UBAP2L
Microarray	Blood	Uniform	54.92	100 0	0 100	50 50	22	USP52, ZNF3, UCN, WIRE, NELL2, ZFP91, Z27499.1_CDS_1, ZMAT2, ZFP36L2, XM_373795, ZSWIM3, unk91, unk90, unk97, TBX21, unk59, PARC, ZNF638, ZNF652, LOC91526, ZCCHC14, ZC3HDC7, USP10, USP3, TM4SF13, DXYS155E, unk163, unk173, ZF, URP, UTRN, ZNFN1A1, unk47, WSB2, TRAP1, GFOD1, RPS23, hCG1747327.2, LOC221091, UGT2A1, UPK3B, TRIM23, TM4SF9, UNQ5783, UBXD1, UCP2, WSB2, ORM1, YWHAQ, CA1, unk68, PPARA, TRAP1, FLJ42953, ZNF3, C10orf33, SLC2A3, unk113, UBAP2L, TTBK1, unk58, TP53BP2, VDP, NCOR2, WARS, TM4SF13
Microarray	ER	Orig	90.2	100 37.5	37.5 100	68.75 68.75	9	CDK5R1, AJ223366.1, OR56B4, 438146_rc, ACPT, Cep192, P2RY14, HBQ1, hCG2007944, PTGER2, ST8SIA4, ZNF638, hCG17621.3, FUSIP1, YWHAQ, KIAA0196, ASAHL, TRAP1, hCG2015808, TSC, DDX46, hCG1787791.2, ELMO2

Microarray	ER	Ratio	98.08	100 100	88.89 100	94.44 100	8	ACPT, hCG2007944, AJ223366.1, OR56B4, 438146_rc, hCG1787898.2, SRRM2, CDK5R1, KIAA0196, hCG37981.3, ST8SIA4, TSC, P2RY14, IARS, hCG1789070.2, TBRG1, C10orf47, TRAP1, FBXO3, ZNF638, HBQ1, AF386301.1_CDS_1, SPATA11, TPSG1, MBC2, ZNF638, PLEK2, FUSIP1, LOC56181, hCG2014315, hCG1813545.1, Cep192, FUSIP1, hCG1644254.2, hCG1787791.2, hCG1747327.2, hCG1812838.1, hCG1642170.3, hCG1642749.1
Microarray	ER	Uniform	98.08	100 100	88.89 100	94.44 100	8	ACPT, hCG2007944, AJ223366.1, OR56B4, 438146_rc, hCG1787898.2, SRRM2, ST8SIA4, TSC, CDK5R1, hCG37981.3, KIAA0196, hCG1789070.2, IARS, P2RY14, TBRG1, C10orf47, ZNF638,HBQ1, Cep192, FBXO3, SPATA11, TRAP1, PLEK2, FUSIP1, TPSG1, MBC2, hCG2014315, LOC56181, hCG1813545.1, hCG1644254.2, hCG1747327.2, hCG1787791.2, hCG1812838.1, hCG1642170.3, hCG1642749.1
Microarray	Grade1v2	Ratio	61.54	100 6.67	6.25 100	53.13 53.33	10	hCG2040108, hCG40931.2, "hCG2041203, hCG2039305", "hCG20704.2, hCG2015359", "SNRPEL1, SNRPE", KIAA0020, Name, "hCG32985.2, hCG2042652", "hCG2001464.2, hCG2001453.1", "hCG27168.2, hCG2030721", "LOC440607, FCGR1A", hCG33299.3, hCG2026261, "hCG2041203, hCG2033271.2, hCG1989403, "hCG2032253, hCG1999251", "hCG40614.2, hCG1985370", ChGn, "hCG2038936, hCG2029987.1, hCG2003479", "hCG2040657, hCG1645925.2", unk179, hCG2042652", hCG38189.3, hCG1990955.1, "hCG2020044, hCG2043429", TOMM20, "hCG27618.3, hCG1640125.2", "hCG21570.3, hCG1783417.1", "hCG27168.2, SEC63, ZNF350, hCG1981858, hCG1989403", "hCG2033271.2, hCG40614.2,"hCG1642357.4,hCG28108.2", "hCG2017355,hCG1733583.1,hCG1983954.1", hCG16179.4, "hCG1984513,hCG2014440", "hCG1728885.2,hCG1739047.2"

Microarray	Grade1v2	Uniform	61.54	100 6.67	6.25 100	53.13 53.33	10	hCG2040108, hCG40931.2, "hCG2041203 ,hCG2039305", "hCG20704.2, hCG2015359", "SNRPEL1, SNRPE", KIAA0020, Name, "hCG32985.2, hCG2042652", "hCG2001464.2, hCG2001453.1", "hCG27168.2, hCG2030721", "LOC440607, FCGR1A", hCG33299.3, hCG2026261, hCG38189.3, hCG1990955.1, KIAA0746, "hCG2020044, hCG2043429", "hCG27618.3, hCG1640125.2", "hCG2040657, hCG1645925.2", unk179, hCG2030721", hCG40931.2", hCG14638.4, "hCG40614.2, hCG1985370", "hCG21570.3, hCG1783417.1", , SEC63, "hCG32985.2, ZNF350, hCG1981858, hCG2042652", hCG1989403", "hCG2033271.2, "hCG1642357.4, hCG28108.2", "hCG2017355, hCG1733583.1, hCG1983954.1", hCG16179.4, "hCG1984513, hCG2014440", "hCG1728885.2, hCG1739047.2"
Microarray	Grade1v2	Orig	60.53	95.65 6.67	6.67 95.65	51.16 51.16	12	hCG40614.2, hCG1985370, "LOC440607, FCGR1A", "SNRPEL1, SNRPE", hCG2041031, "hCG2015869, hCG2015868.1", "hCG27618.3, hCG1640125.2", "hCG27168.2, hCG2030721", "hCG32985.2, hCG2042652", "hCG33299.3, hCG2026261", "hCG20704.2, hCG2015359", "LOC283922, PDPR", "hCG2038936 ,hCG2029987.1, hCG2003479", ChGn, hCG2041203, hCG2039305, "hCG2040657, hCG1645925.2", "hCG38189.3, hCG1990955.1", ,hCG1646386.3, "hCG27168.2, "hCG2040108, hCG40931.2", "hCG21570.3, hCG1783417.1", "HIST2H2AA, HIST2H2AC", KIAA0020, , TXNRD2, hCG1746597.1, ZNF350, "hCG2033271.2, hCG1989403", RPS26, ,hCG1990955.1", hCG1985370", hCG2015359", "LOC440607, hCG14638.4, CX3CR1, "hCG2039500, hCG1737371.3", , "hCG2033271.2 ,hCG1989403"
Microarray	Grade1v3	Orig	42.11	4.35 100	100 4.35	52.17 52.17	10	FNBP4, TRPV4, SPATA5L1, MLL5, RAD21, hCG2039309.1, DNAJB9, hCG1747328.2, hCG17415.3, TMEM40, EIF4G3, hCG2014776, DPP9, DDX5, C19orf25, FBXO7, hCG2039497, hCG2023112.1, MGST2, hCG2010443, LOC55924, hCG24651.4, hCG2027440, PISD, hCG2039309.1, ANKRD11, SESTD1, RAB2B, ABCB7, SAP30, IPLA2(GAMMA), AGL, PISD
Microarray	Grade1v3	Uniform	71.8	100 33.33	31.25 100	65.63 66.67	10	FNBP4, FBXO7, TRPV4, hCG2039309.1, hCG1747328.2, RSN, RAD21, LOC55924, hCG17415.3, PISD, hCG2040593, DNAJB9, TMEM40, hCG17415.3, hCG2039497, HERC1, MLL5, unk133, DPP9, RAB2B, MGST2, FLJ11171, hCG2027440, ATR, ANKRD11, MLL5, HOM-TES-103, EIF4G3, hCG2014776

Microarray	Grade1v3	Ratio	79.49	100 53.33	50 100	75 76.67	11	FNBP4, MGST2, TRPV4, hCG2023112.1, hCG2039309.1, hCG1747328.2, ANKRD11, RSN, RAD21, hCG17415.3, PISD, TMEM40, DNAJB9, hCG2040593, FBXO7, hCG2039497, LOC55924, hCG2039309.1, FLJ11171, RAB2B, TMEM40, hCG2027440, ATR, MLL5, HOM-TES-103, DPP9, SESTD1, hCG2014776, EIF4G3
Microarray	Grade2v3	Orig	69.57	100 39.13	39.13 100	69.57 69.57	13	PTGS2, SRP14, CKLFSF8, CCDC5, SESN2, hCG2041826, TM4SF9, PF4, hCG2009487, POLR2A, RAB43, ITGB2, PSMB9, GPR155, hCG1820921.1, unk9, hCG2039161, hCG2010471.1, ZNF206, hCG1642482.3, CD79A, MGC29814, FHIT, CAMK2G, ADAR, ALOX5AP, DKFZp762O076, EDG8, hCG2042923, MTF1, FLJ23091, hCG22325.2, PAK1
Microarray	Grade2v3	Ratio	59.57	82.61 39.13	37.5 83.33	60.05 61.23	12	CAMK2G, MGC32065, PSMB9, PYGL, TM4SF9, ZNF206, hCG1781673.1, ITGB2, hCG2039161, GPR155, C20orf178, hCG2039498, hCG2041220, CD79A, CKLFSF8, TIZ, IRF7, MGC32065, hCG1782892.2, hCG2009487, hCG2041826, CCDC5, hCG16179.4, hCG20164.2, hCG1781894.2, SRP14, C1orf24, PF4, SESN2, FHIT, ProSAPiP2, ALOX5AP, PTGS2, hCG1642482.3, F13A1, AMPD2, HMGB1, ZNF206, EDG8, hCG26831.3, 413154_rc
Microarray	Grade2v3	Uniform	59.57	82.61 39.13	37.5 83.33	60.05 61.23	12	CAMK2G, MGC32065, PSMB9, PYGL, TM4SF9, ZNF206, hCG1781673.1, ITGB2, hCG2039161, GPR155, C20orf178, hCG2039498, hCG2041220, CD79A, CKLFSF8, TIZ, IRF7, hCG2009487, hCG2041826, hCG1782892.2, PFKFB3, hCG16179.4, PF4, SRP14, EDG8, PTGS2, C1orf24, FHIT, SESN2, ProSAPiP2, ALOX5AP, hCG1642482.3, AMPD2, F13A1k, 413154_rc, HMGB1, KLFSF8, CCDC5, hCG26831.3
Microarray	Menopause	Orig	29.41	2.70 100	100 2.70	51.35 51.35	9	SNRPD2, TRIM46, XPC, 41886, hCG1772363.3, SLC31A1, hCG1790688.1, hCG1820954.2, DEPC-1, NUSAP1, UNG, NTAN1, POR, FLJ10374, GTF2E2, STX17, ARID4B, SLC22A4, , LENG4, hCG1744783.2, hCG22538.3, hCG1790802.3
Microarray	Menopause	Uniform	26.92	0 100	100 0	50 50	11	SLC31A1, 41886, XPC, SNRPD2, hCG1790688.1, DEPC-1, SLC22A4, NTAN1, hCG1772363.3, hCG1820954.2, INCA, AGGF1, UNG, PKM2, PMM2, NUSAP1, STX17, hCG1820954.2, FLJ10374, hCG1991671.2, FLJ10374, hCG2040754, hCG22964.3, RARRES3, TMSB10, hCG1820528.1, POR, SIT, RASGRP4, BZW2, hCG22538.3, SCGB1A1, TRIM46, NUSAP1, ST3GAL5, STX17
Microarray	Menopause	Ratio	26.92	0 100	100 0	50 50	11	SLC31A1, 41886, XPC, SNRPD2, hCG1790688.1, DEPC-1, SLC22A4, NTAN1, hCG1772363.3, hCG1820954.2, NUSAP1, PMM2, PKM2, INCA, AGGF1, UNG, STX17, FLJ10374, NUSAP1, hCG1820954.2, POR, TRIM46, ARID4B, TMSB10, RASGRP4, LENG4,

								SIT, hCG1820528.1, INCA, SCGB1A1, hCG22538.3, BZW2, ST3GAL5
Protein	BM	Ratio	68.29	100 18.75	23.53 100	61.77 59.38	5	SMARCA1, MLLT6, UNK63, BZRAP1, FADD, HSPD1, IK, C17ORF65, CCR6, JPH3, EXOSC10, NARF, C4B
Protein	BM	Orig	60	100 0	0 100	50 50	8	EBNA1BP2, RNF130, CCR6, HDAC2, JPH3, TUBA1A, CCND1, ING2, FADD, HSPD1, IK, C17ORF65, TOMM70A, MRPL28, ING2, TERF2IP, UNK26, PKLR, DYNLL1, EGFL6, SAMD14, TANK, CCR6, CREBBP, XIRP1, FDFT1, NR4A3, XIRP1, EXOSC10, TUBA1A, MUC16, CCDC109B, C11ORF2, CCND1, SARNP, ZNF629, NUMA1
Protein	BM	Uniform	68.29	100 18.75	23.53 100	61.77 59.38	5	SMARCA1, NARF, MLLT6, BZRAP1, FADD, HSPD1, IK, C17ORF65, CCR6, JPH3, EXOSC10, C4B, COQ4, NAP1L4, HSPD1
Protein	NB	Ratio	56.76	100 0	5.88 100	52.94 50	1	RAB24, CREBBP, JPH3, MPHOSPH8, EBNA1BP2, FADD, UNK11, MLLT6, UNK63, BZRAP1
Protein	NB	Uniform	56.76	100 0	5.88 100	52.94 50	1	RAB24, CREBBP, JPH3, MPHOSPH8, EBNA1BP2, FADD, UNK11, C4B, UNK63, BZRAP1
Protein	NB	Orig	66.67	100 25	25 100	62.5 62.5	7	CCR6, JPH3, CAP1, CBX5, EBNA1BP2, TOMM70A, HSPD1, NOP2, CREBBP, FADD, RPS3A, IK, TERF2IP, ZNF48, ZNF238, ARHGEF18, SPAG7, TBC1D9, ZNF629, ZNF658B, XIRP1, PKLR, TANK, NOP2, YBX1, JPH3, UNK60, VPS13D
Protein	NM	Ratio	55.56	100 0	4.76 100	52.38 50	1	TANK, CCR6, JPH3, CAP1, CBX5, EBNA1BP2, FADD, ARPP21, HSPD1, MUC16
Protein	NM	Uniform	55.56	100 0	4.76 100	52.38 50	1	TANK, CCR6, JPH3, CAP1, CBX5, EBNA1BP2, FADD, ARPP21, HSPD1, MUC16
Protein	NM	Orig	50	91.67 0	0 91.67	45.83 45.83	7	CBX5, EBNA1BP2, FADD, CCR6, ARPP21, CREBBP, HSPD1, JPH3, CAP1, IK, EXOSC10, TERF2IP, CAP1, ARPC2, HDAC6, C11ORF2, LILRB1, CBX5, TANK, POLR2G, EXOSC10, YBX1, MRPL28, HDAC6, NPEPL1, GAPDH, RPL5, GSPT2
RT-PCR		Orig	88.89	100 66.67	66.67 100	83.33 83.33	4	CD68, FOXC2, PTPRC, TFRC, KLK3, JUN
RT-PCR		Ratio	89.29	100 66.67	70 100	85 83.33	5	CD68, TFRC, ABCB1, PTPRC, BCL2, AR, FOXC2, FOLH1, KLK3
RT-PCR		Uniform	89.27	100 66.67	70 100	85 83.33	5	CD68, TFRC, ABCB1, PTPRC, BCL2, AR, FOXC2, FOLH1, KLK3

Table C.28. Lung cancer modeling results

Data type	Comparison	Prior	Accuracy	SN	SP	Balanced Accuracy	Model Size	Union of variables used
ArrayCGH	RNA vs. DNA	Informed	50	100 0	0 100	50 50	2	unk96, ZNFN1A1, MCAM, PP, VTN, GAS7, unk574, AA454543, ORM1, unk493, KIAA0934, MSL3L1, AA460731, MRPL35
ArrayCGH	RNA vs. DNA	Uniform	47.06	100 0	0 100	50 50	3	unk96, ZNFN1A1, SFRP2, PP, VTN, IRF4, GAS7, AA454543, ORM1, TNNI3K, KIAA0934, MSL3L1, ARGBP2, F8A1
ArrayCGH	RNA vs. DNA	None	50	100 0	0 100	50 50	2	unk96, ZNFN1A1, MCAM, PP, VTN, GAS7, unk574, AA454543, ORM1, unk493, KIAA0934, MSL3L1, AA460731, MRPL35
Copy Number	Adeno vs. Squamous	Informed	64	100 10	14.29 100	57.14 55	11	MT1P3, FTO, CENTB2, merck-AX747832_at, merck-BC062771_at, merck-BM979827_at, merck-AJ420566_s_at, PXMP4, C1orf181, GLIS2, CENPO, LOC644285, FGFR1OP2, ESAM, merck-AL049252_a_at, merck-BQ446551_at, ALPK3, merck-AK024690_at, PRO0456, CGN, SLC28A3, SARS, POT1, merck-BU742340_at, LOC644246, merck-AK057683_at, CBFB-MYH11, SGCD
Copy Number	Adeno vs. Squamous	Uniform	62	100 5	9.5 100	54.76 52.5	10	MT1P3, KIAA0492, merck-AX747832_at, FGFR1OP2, merck-BC062771_at, merck-AL049252_a_at, merck-BM979827_at, C1orf181, GLIS2, LOC644246, CGN, SLC28A3, ESAM, SYDE2, POT1, merck-BU742340_at, CBFB-MYH11, merck-AJ420566_s_at, merck-AK057683_at, MEGF6, LOC154092, merck-BQ446551_at, HIF3A, CGN, SGCD, PXMP4, LOC644285, merck-AW418496_a_at

Copy Number	Adeno vs. Squamous	None	63.26	100 10	10 100	55 55	11	MT1P3, merck-BU742340_at, merck-AX747832_at, FGFR1OP2, merck-BC062771_at, merck-AL049252_a_at, merck-BM979827_at, PXMP4, ESAM, C1orf181, GLIS2, CGN, NUP50, SLC28A3, CBFB-MYH11, merck-AK057787_at, LOC145786, merck-AW444477_at, merck-AK057683_at, merck-AK001128_at, PRO0456, POT1, LONRF2, SGCD, SLC39A10, CENPO
Methylation	High vs. Control	Informed	68.97	100 30.77	35.71 100	67.85 65.39	7	DBF4, VKORC1, IMPDH2, ANP32B, EIF4A2, AC005011, MRPL24, PRIM1, H2AFV, MRPL37, MRPL24, CTNNAL1, RPL37A, HAUS1, COPS2, HNRNPA3 /// HNRNPA3P1, HMGN2, OAT, RPL29, TPD52L2, HSPD1, ZNF593, MEST, FN1, SLC3A2, C5orf32, LOC100510735 /// RPL29, HMGN2, UQCR10
Methylation	High vs. Control	Uniform	55.17	100 0	7.14 100	53.17 50	8	PRIM1, CLNS1A, TNNT1, RPL37A, VKORC1, AC005011, UBE2Q2, MRPL24, TMEM97, LGTN, OAT, HAUS1, COPS2, RPL37A, FOXM1, DBF4, RPLP0 /// RPLP0P6, EIF4A2, MYEOV2, CDKN1B, MRPL37, DH2, BOLA2 /// LOC440354 /// LOC595101, H2AFV, MYEOV2, MEST, CKMT1A /// CKMT1B, RAD21, ZNF593, SRI, IMPDH2, AC005011, HMGN2
Methylation	High vs. Control	None	85.71	100 69.23	69.23 100	84.62 84.62	9	PRIM1, H2AFV, HEBP2, VKORC1, YWHAB, MRPL24, RPL37A, ND2, FOXM1, ITGAV, MYL12A, HDAC1, UBE2E2, CYC1, ZNF721, ITGB3BP, SERINC3, METTL5, NDUFB9, CACYBP, FAM96A, MOBKL1A, STOM, PRSS3, DCBLD2, PPL, HN1, C14orf156, ITGB4, TMEM14C, AFFX-HUMGAPDH/M33197_5_at, TRNP1, JAG1, CBR1, TJP1, CD164, PRICKLE4 /// TOMM6, RTN4, ACTR3, PLSCR3, NOP56, LOC100506727, FAM127A, PRC1, AC004544
Methylation	High vs. Low	Informed	73.91	100 33.33	40 100	70 66.67	6	HN1, RRAS, ASNS, POLB, ITGAV, LOC100506727, CYC1, ZNF721, METTL5, ITGAV, PRSS3, YWHAB, XRCC6, PLSCR3, ATP6V1F, CDK11A, RPA3, UBE2E2, C14orf156, TMEM50A, C1orf103, RNASEK, KRT18, TIMP1, HDAC1, PPP3CA, GPRC5A, ASNS, CD164, C14orf156, PTTG1IP, UBE2S, CAPRIN2

Methylation	High vs. Low	Uniform	69.57	100 22.22	30 100	65 61.11	6	ACTR3, ZNF721, SLC38A2, ASNS, SHMT2, CES2, CD164, PRICKLE4 /// TOMM6, MYL12A, CYC1, METTL5, PRSS3, ITGB3BP, TMEM50A, HMGB2, ITGAV, AY094612, ATP6V1F, CDK11A, RPA3, UBE2E2, C14orf156, RRAS, ITGB4, PTTG1IP, PPP3CA, TIMP1, LOC100506727, HDAC1, UCHL1, GPRC5A, XRCC6, ZNF721, STUB1, NIT2, HN1, KRT18, BRD9, PLSCR3, BLCAP, XRCC6, COPS8
Methylation	High vs. Low	None	77.27	100 44.44	44.44 100	72.22 72.22	8	DCBLD2, HN1, MBOAT2, C14orf156, PRICKLE4 /// TOMM6, ITGAV, UBE2E2, ZNF721, MYL12A, HDAC1, CYC1, ITGB3BP, SERINC3, ZNF721, METTL5, NDUFB9, CACYBP, FAM96A, MOBKL1A, STOM, PRSS3, PPL, ITGAV, ITGB4, TMEM14C, AFFX-HUMGAPDH/M33197_5_at, TRNP1, JAG1, CBR1, TJP1, CD164, RTN4, ACTR3, PLSCR3, NOP56, LOC100506727, ND2, FAM127A, PRC1, AC004544
Methylation	Low vs. Control	Informed	48	66.67 11.11	30 93.75	48.33 52.43	8	DEPDC1B, DAP3, CCDC142 /// MRPL53, RSRC2, ADNP, NDUFB8, RPL18A /// RPL18AP3, MGST1, NME4, GAS5, EIF3K, CENPW, MGST1, NRAS, CUTA, LOC100499177, TMEM126A, SYPL1, FKBP1A, SRP14, AFFX-BioDn-5_at, AV724183, DLGAP5, RBX1, TOMM22, RAD51AP1, H2AFV, TRAPPC5, FIBP, SEC24B, NCAPD2, GNG5, CDCA4, JAG2, RSRC2, PTPRO, FOXM1, AFFX-HSAC07/X00351_3_at, PABPC1
Methylation	Low vs. Control	Uniform	76	100 33.33	40 100	70 66.67	7	CCDC142 /// MRPL53, AFFX-BioDn-5_at, NCAPD2, EIF4A2, NDUFB8, NRAS, H3F3B, GAS5, NDUFB11, PAICS, EIF3G, RSRC2, FOXM1, PABPC1 /// RLIM, LOC100505603 /// PNRC2, PHB, DLGAP5, DAP3, AV724183, MRPL51 /// SPTLC1, EIF3K, NRAS, H3F3B, H2AFV, DEPDC1B, AFFX-BioDn-5_at, PTPRO, GNAI3, ADNP, SRP14, LOC100499177, NCAPD2, XRCC6, RSRC2, LOC647979, RBX1, FZD6, GNG5, RAD51AP1, PSMD10, LUZP6 /// MTPN, ECEP55

Methylation	Low vs. Control	None	37.5	100 0	0 100	50 50	8	CCDC142 /// MRPL53, RSRC2, ADNP, AFFX-HUMGAPDH/M33197_3_at, DLGAP5, MRPL51 /// SPTLC1, RPL29, EI24, TOMM22, FKBP1A, YWHAB, HNRNPA1, LOC100499177, DAP3, XRCC6, C19orf53, EI24, CAP1, NRAS, ITGB1, EIF3K, TXNDC12, TXNDC17, NDUFB8, TMEM126A, LOC647979, AFFX-BioDn-5_at, KIAA1949, PTPRO, PSAT1, SRSF6, RPL28, AFFX-HUMGAPDH/M33197_3_at, PPP3CA, STARD4, RSRC2, TMEM147, CDCA4, XRCC6, GAS5, FKBP1A, CENPW
Microarray	Case vs. Control	Informed	50	93.75 1.37	12.16 100	52.96 50.69	20	FLJ20006, AP2M1, SLC25A46, GSR, WLS, C21orf59, FLJ10246, 41888, ASAP2, AL121916, FLJ20700, ZNF562, DKFZp547P082, AU147295, PRO1995, F13A1, TRIM68, AFFX-BioB-M, ING3, DDX27 /// SS18, KIAA1033, BE999967, VPREB3, PRO1995, F13A1, LTF, FLJ21272, NAE1, ING3, WBSCR22, RPL36, PEBP1, AL121916, PIGP, AU147295, NCL, PJA1, KLHL28, TNFAIP6, 41700, MTPAP, RPL27, AGFG1, RNASE6, ROGDI, GNA15, PTGDR, ZNF721, LONP2, FSTL1, DRAM1, RALGPS2, JAK2, NXT1, FLJ11786, PRPF19, CHIC2, HIRA, OGFOD1, C1QB, RPL3, NFATC2IP, ARF4, NEU1, DEFA1 /// DEFA1B /// DEFA3, RPL34, IGHA1 /// IGHA2 /// IGHG1 /// IGHG2 /// IGHG3 /// IGHM /// IGHV4-31 /// LOC100126583 /// LOC100290036, RCBTB2, COL4A3BP, PEBP1, TBX21, FLJ11786, TM7SF3, CXCL5,FSTL1,

Microarray	Case vs. Control	Uniform	55.2	100 5.48	6.76 100	53.38 52.74	18	PRO1995, FLJ20700, FLJ21272, FLJ20006, CHIC2, 41888, AL121916, FLJ10246, SLC25A46, RPL36, AU147295, RPL3, RNASE6, PRIM1, ASAP2, NFATC2IP, WLS, ING3, COL4A3BP, CNPY2, SRSF11, AFFX-BioB-M, IGHA1 /// IGHA2 /// IGHG1 /// IGHG2 /// IGHG3 /// IGHM /// IGHV4-31 /// LOC100126583 /// LOC100290036, NCL, DKFZp547P082, 41700, HLA-DQA1 /// HLA-DQA2, FLJ20700, CCDC90A, AP2M1, FLJ13197, TM7SF3, FLJ10246, SPARC, OGFOD1, 41700, ARF4, OSBPL10, AGFG1, RPL27, ROGDI, GNA15, PTGDR, ZNF721, LTF, MTPAP, LONP2, FSTL1, DRAM1, TMX2, ZNF562, C21orf59, FLJ21272, KIAA0182, ARHGEF18, SMOX, NKTR, AGFG1, RPS18, SORT1, C17orf60,FTSJD2, BBX, GPR89A /// GPR89B /// GPR89C, ANK1, JAK2, PIGP, NCL
Microarray	Case vs. Control	None	56.86	95 15.07	15.07 95	55.03 55.03	17	NKTR, DKFZp547P082, TMX2, C21orf59, 41700, PRO1995, FLJ20700, FLJ21272, FLJ20006, FLJ11786, EGF, AP2M1, AFFX-BioB-M, FLJ10246, 41888, SLC25A46, RNASE6, TM7SF3, KIAA0182, AU147295, FLJ13197, PEBP1, C21orf59, WLS, FLJ11786, TBX21, AL121916, NELL2, 41700, PJA1, FLJ11786, PIGP, ANK1, ING3, PLGLA /// PLGLB1 /// PLGLB2, BF984434, ASAP2, TRAF3IP3
Microarray	Adeno vs. SCC	Informed	97.53	100 83.33	75 100	87.5 91.67	8	PCBP1, TMBIM1, HSDL2, FLJ21272, IER3, DEFA1 /// DEFA1B /// DEFA3, AFFX-r2-Ec-bioC-3, FLJ23556, ATP6V1A, HEBP2, TBXAS1, AFFX-r2-Ec-bioB-3, CDC42EP3, AFFX-BioB-3, C7orf42, TM6SF1, ASGR1, DAZAP2, DPM1, RNF130, FBXO11, PRO1412, TIMM8B, BRP44L
Microarray	Adeno vs. SCC	Uniform	92.59	98.63 33.33	37.5 98.67	68.07 66	7	FLJ21272, IER3, ATP6V1A, BRP44L, DEFA1 /// DEFA1B /// DEFA3, PCBP1, TMBIM1, AFFX-BioC-5, AFFX-BioC-3, AFFX-r2-Ec-bioB-3, AFFX-r2-Ec-bioB-5, LMBRD1, CAB39, FLJ23556, PGCP, DAZAP2, KCNJ15, AFFX-r2-Ec-bioC-3, PPP3CA, DPM1, PRO1412, HADHB, AFFX-r2-Ec-bioC-5, AFFX-BioB-5, AFFX-BioDn-5, FBXO11

Microarray	Adeno vs. SCC	None	93.75	100 33.33	28.57 100	64.29 66.67	7	IER3, AFFX-BioB-3, AFFX-r2-Ec-bioB-3, AFFX-r2-Ec-bioB-5, CDC42EP3, TMBIM1, BRP44L, AFFX-r2-Ec-bioB-5, AFFX-r2-Ec-bioC-3, AZIN1, FLJ21272, AFFX-BioB-5, ATP6V1A, AFFX-BioDn-5, HSDL2, DEFA1 /// DEFA1B /// DEFA3, PRO1412, FLJ23556, NAPA, TBXAS1, AFFX-BioC-5, TFEC, FBXO11, DAZAP2, TM6SF1, ASGR1, PRO1412, BRP44L, DPM1, GSTT1, AZIN1, PCBP1, SLC5A5, PGCP, AFFX-r2-Ec-bioC-5, AFFX-BioB-M
Microarray	Smoking History: Former vs. Current	Informed	55	100 1.82	3.57 100	51.79 50.91	16	RTP4, OLFR89, FLJ23556, RTN3, SLC22A8, JUNB, COL9A2, EFNA3, PDLIM1, POLL, OSGEP, C7orf28B /// CCZ1, CLEC2D, TBC1D12, FLJ11117, NDUFB8, SFPQ, MFHAS1, OR7E37P, SOD3, RPS6KA1, RAPGEF2, HSAF000381, PXN, AFFX-M27830_5, SV2A, SFRS15, SFPQ, CRAT, HSAF000381, TMEM161A, UBE2D3, C16orf71, PXN, DKFZp547P082, UBR2, C6orf62, AU148154, AA654586, RAB14, PCNP, CHMP1B, SFMBT1, XPO1, MAPKAPK5, NRGN, OSGEP, LSR68, EML4, AW150065, DPM2, NUDT3, POLL, RBPJ, PPP4C, NUPL1, IGLC7 /// IGLV1-44, KIAA0317, SYNE1, FAM129A, MEF2A, TBC1D12, VCL, PXN, ANKRD28, NUSAP1, HNRNPA1 /// HNRNPA1L2 /// HNRNPA1P10 /// LOC728643, MEF2A, NAB1, **CCL5**, RPLP0 /// RPLP0P6, GGA1, ACTR2, **SOD2**, GPRC5C, PCBP1, KLF13,WDR19, CBY1

Microarray	Smoking History: Former vs. Current	Uniform	55	100 3.64	3.57 100	51.79 51.82	11	OLFR89, RTP4, LSM2, AA654586, HSAF000381, C7orf28B /// CCZ1, SLBP, RAPGEF2, EFNA3, SFPQ, FLJ23556, NOD2, MAPKAPK5, NUSAP1, HSAF000381, TBC1D12, BF448531, C16orf71, IGKV1-5, DKFZp547P082, DPM2, XCL1 /// XCL2, MRP63, AFFX-M27830_5, PPP4C, OSGEP, RPS6KA1, SOD2, RBM16, NDUFB8, PGAP3, LSR68, CEP57, IGK@ /// IGKC /// IGKV1-5, RPS24, CLEC2D, FLJ11117, NPHS2, AL080190, COL9A2, FANCG, CTTN, FLJ23556, FPR2, DPM2, AU148154, IGKV1D-8, RPLP0 /// RPLP0P6, AV742010, ANKRD28, LSR68, SYNE1, CRAT, FAR2, IGLL3P, KIAA0317, C17orf101, PELI1, CHMP1B, M85256, PCBP1, **SOD2**, TMEM161A, MFHAS1, **CCL5**, KLF13,WDR19, CBY1
Microarray	Smoking History: Former vs. Current	None	53.78	100 0	0 100	50 50	16	OLFR89, MAPKAPK5, OSGEP, C16orf71, LSR68, EML4, DKFZp547P082, RTP4, C14orf56, FLJ23556, HSAF000381, IGF2BP2, AFFX-M27830_5, CLEC2D, TBC1D12, FLJ11117, IGKV1D-8, RPS6KA1, EFNA3, NRGN, C7orf28B /// CCZ1, RTN3, CRAT, RFWD3, ANKRD28, SNRNP40, EML4, EHD3, GPRC5C, SFPQ, COL9A2, SOD2, LSM2, HSAF000381, LSR68, CLCN2, IGF2BP2, UBXN6, AL080190, CRAT, RFWD3, RPS4X /// RPS4XP6, XPO1

BIBLIOGRAPHY

32nd Annual CTRC-AACR San Antonio Breast Cancer Symposium. ***Sunday Morning Year-End Review***. Dec. 14, 2009.

Aarøe J, Lindahl T, Dumeaux V, Saebø S et al. **Gene expression profiling of peripheral blood cells for early detection of breast cancer.** *Breast Cancer Res* 2010, 12(1):R7.

Aboud OA, Weiss RH. **New opportunities from the cancer metabolome**. *Clin Chem.* 2013, 59(1):138-46.

Aceves C, Anguiano B, Delgado G. **Is iodine a gatekeeper of the integrity of the mammary gland?** *Journal of Mammary Gland Biology and Neoplasia* 2005, **10** (2): 189–196.

Ackermann M, Strimmer K. **A general modular framework for gene set enrichment analysis.** *BMC Bioinformatics* 2009, 10: 47.

Adamic LA, Wilkinson D, Huberman BA, and Adar E: **A literature based method for identifying gene-disease connections.** In Proceedings of the *IEEE Computer Society Bioinformatics Conference* 2002**, 1**:109-117.

Al-Mubaid H, Singh RK: **A new text mining approach for finding protein-to-disease associations.** *American Journal of Biochemistry and Biotechnology* 2005, **1**:145-152.

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. **FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes.** *Bioinformatics* 2004, **20**: 578–580.

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information.** *Bioinformatics* 2005, **21**: 2988–2993.

Alberg AJ, Samet JM (2010). ***Murray & Nadel's Textbook of Respiratory Medicine*** **(5th ed.).** Saunders Elsevier. ISBN 978-1-4160-4710-0.

Albert S, Gaudan S, Knigge H, et al. **Computer-assisted generation of a protein-interaction database for nuclear receptors**. *Mol Endocrinol.* 2003. **17**:1555-1567.

Alpaydin E. (2010) ***Introduction to machine learning*** **(2nd ed.).** MIT Press. ISBN-10: 0-262-01243-X, ISBN-13: 978-0-262-01243-0.

Alterovitz G, Xiang M, Liu J, Chang A, Ramoni MF: **System-wide peripheral biomarker discovery using information theory**. *Pacific Symposium on Biocomputing* 2008:231-242.

American Cancer Society. **Breast Cancer Facts & Figures 2005–2006**. Archived from the original on June 13, 2007. http://web.archive.org/web/20070613192148/http://www.cancer.org/downloads/STT/CAFF2005BrFacspdf2005.pdf.

American Institute for Cancer Research/ World Cancer Research Fund, **Food, Nutrition, Physical Activity and the Prevention of Cancer: a Global Perspective**, http://www.dietandcancerreport.org

Andrade MA, Valencia A. **Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families.** *Bioinformatics* 1998, **14**:600-607.

Andrade MA, Bork P. **Automated extraction of information in molecular biology**. *FEBS Letters* 2000. **476**:12-17.

Ante M, Wingender E, Fuchs M. **Integration of gene expression data with prior knowledge for network analysis and validation.** *BMC Res Notes* 2011, 4:520.

Aronson AR. **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program**, In: *American Medical Informatics Association*, 2001, pp. 17–21.

Aviel-Ronen S, Blackhall FH, Shepherd FA, Tsao MS. **K-ras mutations in non-small-cell lung carcinoma: a review.** *Clinical Lung Cancer* (Cancer Information Group) 2006, **8** (1): 30–38.

Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, et al. **GeneTrail – advanced gene set enrichment analysis.** *Nucleic Acids Res* 2007, 35: W186–W192.

Barry WT, Nobel AB, Wright FA. **Significance analysis of functional categories in gene expression studies: a structured permutation approach.** *Bioinformatics* 2005, 21: 1943–1949.

Barter RL, Schramm SJ, Mann GJ, Yang YH. **Network-based biomarkers enhance classical approaches to prognostic gene expression signatures.** *BMC Syst Biol. 2014;8 Suppl 4:S5.*

Bayes T, Price Mr. **An Essay towards solving a Problem in the Doctrine of Chances.** *Philosophical Transactions of the Royal Society 1763.* ***53****: 370–418.*

Beel J and Gipp B. **Google Scholar‘s Ranking Algorithm: An Introductory Overview.** In Birger Larsen and Jacqueline Leta, editors, Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI’09), volume 1, pages 230 – 241, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics.

Begg CB, Haile RW, Borg A, *et al.* **Variation of breast cancer risk among BRCA1/2 carriers**. *JAMA* 2008, **299** (2): 194–201.

Behera D, Balamugesh T. **Lung cancer in India**. *Indian Journal of Chest Diseases and Allied Sciences* (2004). **46** (4): 269–281.

Beissbarth T, Speed T. **GOstat: find statistically overrepresented gene ontologies within a group of genes.** *Bioinformatics* 2004, 20: 1464–1465.

Ben-Hur A, Horn D, Siegelmann H, Vapnik V. **Support vector clustering**. *Journal of Machine Learning Research* 2001, 2: 125–137.

Benjamini Y, Hochberg Y. **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society, Series B* 1995, **57** (1): 125–133.

Berriz GF, King OD, Bryant B, Sander C, Roth FP. **Characterizing gene sets with FuncAssociate.** *Bioinformatics* 2003, 19: 2502–2504.

Bigbee WL, Gopalakrishnan V, Weissfeld JL, Wilson DO, Dacic S, Lokshin AE, Siegfried JM. **A multiplexed serum biomarker immunoassay panel discriminates clinical lung cancer patients from high-risk individuals found to be cancer-free by CT screening**. *J Thorac Oncol*. 2012, **7**(4):698-708.

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, et al. **ClueGO: a Cytoscape plugin to decipher functionally grouped gene ontology and pathway annotation networks.** *Bioinformatics* 2009. 25: 1091–1093.

Blaschke C, Andrade MA, Ouzounis C, Valencia A. **Automatic extraction of biological information from scientific text: protein-protein interactions**. *ISMB*, 1999, **7**:60-67.

Boffetta P, Agudo A, Ahrens W*, et al*. **Multicenter case-control study of exposure to environmental tobacco smoke and lung cancer in Europe**. *Journal of the National Cancer Institute* (Oxford University Press) 1998, **90** (19): 1440–1450.

Boffetta P, Hashibe M, La Vecchia C, Zatonski W, Rehm J. **The burden of cancer attributable to alcohol drinking**. *International Journal of Cancer* 2006, **119** (4): 884–7.

Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ. **T-profiler: scoring the activity of predefined groups of genes using gene expression data.** *Nucleic Acids Res* 2005, 33: W592–W595.

Boser BE, Guyon IM, Vapnik V. **A training algorithm for optimal margin classifiers.** *In Fifth Annual Workshop on Computational Learning Theory,* Pittsburgh, 1992. ACM.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. **GO:TermFinder–open source software for accession gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes.** *Bioinformatics* 2004, 20: 3710–3715.

Brenner DR, Brennan P, Boffetta P, Amos CI, Spitz MR, Chen C, Goodman G, Heinrich J, Bickeböller H, Rosenberger A, Risch A, Muley T, McLaughlin JR, Benhamou S, Bouchardy C, Lewinger JP, Witte JS, Chen G, Bull S, Hung RJ. **Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls.** *Hum Genet.* 2013, 132(5):579-89.

Breslin T, Eden P, Krogh M. **Comparing functional annotation analyses with Catmap.** *BMC Bioinformatics* 2004, 5: 193.

Cancer Genome Atlas Network. **Comprehensive molecular portraits of human breast tumours.** *Nature* September 2012, Advanced online publication.

Carmona RH. 2006. **The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General.** U.S. Department of Health and Human Services. http://www.surgeongeneral.gov/library/secondhandsmoke. Secondhand smoke exposure causes disease and premature death in children and adults who do not smoke.

Carvalko JR, Preston K. **On Determining Optimum Simple Golay Marking Transforms for Binary Image Processing.** *IEEE Transactions on Computers* 1972. **21**: 1430–33.

Castillo-Davis CI, Hartl DL. **Genemerge - post-genomic analysis, data mining, and hypothesis testing.** *Bioinformatics* 2002, 19: 891–892.

Catelinois O, Rogel A, Laurier D, *et al.* **Lung Cancer Attributable to Indoor Radon Exposure in France: Impact of the Risk Models and Uncertainty Analysis.** *Environ. Health Perspect.* 2006, **114**(9): 1361–6.

Catlett NL, Bargnesi AJ, Ungerer S, Seagaran T, Ladd W, Elliston KO, Pratt D. **Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data.** *BMC Bioinformatics* 2013, 14:340.

Cavalieri E, Chakravarti D, Guttenplan J, *et al.* **Catechol estrogen quinones as initiators of breast and other human cancers: implications for biomarkers of susceptibility and cancer prevention**. *Biochimica et Biophysica Acta* 2006, **1766** (1): 63–78.

Centers for Disease Control (CDC) (1986) **1986 Surgeon General's report: the health consequences of involuntary smoking.** *MMWR. Morbidity and mortality weekly report,* **35** (50): 769–70.

Chang JT, Schutze H, Altman RB. **Creating an online dictionary of abbreviations from MEDLINE**. *J Amer Med Inform Assoc* 2002, **9**:612-620.

Chang JT, Schutze H, Altman RB. **GAPSCORE: finding gene and protein names ne word at a time**. *Bioinformatics* 2004. **20**:216-225.

Chang TW. **Binding of cells to matrixes of distinct antibodies coated on solid surface.** *J. Immunol. Methods* 1983, **65** (1-2): 217–23.

Chen G, Cairelli MJ, Kilicoglu H, Shin D, Rindflesch TC. **Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference.** *PLoS Comput Biol.* 2014, 10(6):e1003666.

Chen GK, Wei P, DeStefano AL. **Incorporating biological information into association studies of sequencing data.** *Genet Epidemiol.* 2011, 35 Suppl 1:S29-34.

Chen H, Goldberg MS, Villeneuve PJ. **A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases**. *Reviews on Environmental Health* 2008, **23** (4): 243–297.

Chen M, Cho J, Zhao H. **Incorporating biological pathways via a Markov random field model in genome-wide association studies**. *PLoS Genet.* 2011, 7(4):e1001353.

Chen X, Wang L. **Integrating biological knowledge with gene expression profiles for survival prediction of cancer.** *J Comput Biol.* 2009, 16(2):265-78.

Chiang JH, Yu HC, Hsu HJ. **GIS – a biomedical text-mining system for gene information discovery**. *Bioinformatics* 2004, **20**:120-121.

Chiu HF, Cheng MH, Tsai SS, *et al*. **Outdoor air pollution and female lung cancer in Taiwan.** *Inhalation Toxicology* 2006, 18 (13): 1025–1031.

Chlebowski RT, Blackburn GL, Thomson CA, *et al.* **Dietary fat reduction and breast cancer outcome: interim efficacy results from the Women's Intervention Nutrition Study**. *Journal of the National Cancer Institute* 2006, **98** (24): 1767–76.

Cohen AM. **Using symbolic network logical analysis as a knowledge extraction method on MEDLINE abstracts**. *BMC Bioinformatics* 2004, **in press**

Cohen AM, Hersh WR: **A survey of current work in biomedical text mining.** *Briefings in Bioinformatics* 2005, **6**:57-71.

Collaborative Group on Hormonal Factors in Breast Cancer. **Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease**. *Lancet* 2002, **360** (9328): 187–95

Collier N, et al. **Extracting the names of genes and gene products with a Hidden Markov Model**. In Proceedings of the *18th International Conference on Computational Linguistics*. 2000, 201-207.

Collins LG, Haines C, Perkel R, Enck RE. **Lung cancer: diagnosis and management.** *American Family Physician* (American Academy of Family Physicians) 2007, 75 (1): 56–63.

Cooper GF, Herskovits E. **A Bayesian method for the induction of probabilistic networks from data.** *Machine Learning*. 1992, 9: 309-347.

Cortes C, Vapnik V. **Support-vector networks.** *Machine Learning* 1995, **20** (3): 273.

Cox DR. **The regression analysis of binary sequences (with discussion).** *J Roy Stat Soc* 1958. **20**: 215–242.

Coyle YM, Minahjuddin AT, Hynan LS, Minna JD. **An ecological study of the association of metal air pollutants with lung cancer incidence in Texas.** *Journal of Thoracic Oncology* 2006, 1 (7): 654–661.

Cun Y, Fröhlich HF. **Prognostic gene signatures for patient stratification in breast cancer: accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions.** *BMC Bioinformatics* 2012, 13:69.

Daemen A, Signoretto M, Gevaert O, Suykens JA, De Moor B. **Improved microarray-based decision support with graph encoded interactome data.** *PLoS One* 2010, 5(4):e10225.

Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin B. **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.** *Nature Genet* 2002, 31: 19–20

Darnton AJ, McElvenny DM, Hodgson JT. **Estimating the number of asbestos-related lung cancer deaths in Great Britain from 1980 to 2000**. *Annals of Occupational Hygiene* 2006, **50** (1): 29–38.

Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK et al. **DNA methylation profiling reveals a predominant immune component in breast cancers.** *EMBO Mol Med* 2011, 3(12):726-41.

Defays D. **An efficient algorithm for a complete link method**. *The Computer Journal (British Computer Society)* 1977. **20** (4): 364–366.

Delaleu N, Immervoll H, Cornelius J, Jonsson R: **Biomarker profiles in serum and saliva of experimental Sjogren's syndrome: associations with specific autoimmune manifestations.** *Arthritis Research & Therapy* 2008, **10**:R22.

Department of Health. **Report of the Scientific Committee on Tobacco and Health**. March 1998. http://www.archive.official-documents.co.uk/document/doh/tobacco/contents.htm. Retrieved 2007-07-09.

Devereux TR, Taylor JA, Barrett JC. **Molecular mechanisms of lung cancer. Interaction of environmental and genetic factors**. *Chest* (American College of Chest Physicians) 1996, **109 (Suppl 3):** 14S–19S.

Deyati A, Younesi E, Hofmann-Apitius M, Novac N: **Challenges and opportunities for oncology biomarker discovery.** *Drug Discovery Today* 2012, **18**:614-624.

Di Renzo L, et al. (Italy) **Intake of red wine in different meals modulates oxidized LDL level, oxidative and inflammatory gene expression in healthy people: a randomized crossover trial.** *Oxid Med Cell Longev*. 2014; 2014:681318.

Ding J, Berleant D, Nettleton D, Wurtele E. **Mining Medline: abstracts, sentences, or phrases?** *Pac Symp. Biocomput*. 2002, 326-337.

Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, et al. **Improving gene set analysis of microarray data by SAM-GS**. *BMC Bioinformatics* 2007, 8: 242.

Donaldson I, Martin J, de Bruijn B, et al. **PreBIND and textomy – mining the biomedical literature for protein-protein interactions using a support vector machine**. *BMC Bioinformatics* 2003, **4**:11.

Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, et al. **MAPPFinder: using gene ontology and GenMAPP to create a global gene expression profile from microarray data.** *Genome Biol.* 2003, 4: R7.

Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. **Global functional profiling of gene expression.** *Genomics* 2003 81: 98–104.

Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. **A systems biology approach for pathway level analysis.** *Genome Research* 2007, 17: 1537-1545.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z. **agriGO: a GO analysis toolkit for the agricultural community.** *Nucleic Acids Res* 2010, 38: W64–W70.

Dunning AM, Healey CS, Pharoah PD, Teare MD, Ponder BA, Easton DF. **A systematic review of genetic polymorphisms and breast cancer risk**. *Cancer Epidemiology, Biomarkers & Prevention* 1999, **8** (10): 843–54.

Efron B, Tibshirani R. **On testing the significance of sets of genes.** *Ann Appl Stat* 2007, 1: 107–129.

Eliassen AH, Hankinson SE, Rosner B, Holmes MD, Willett WC. **Physical activity and risk of breast cancer among postmenopausal women**. *Arch. Intern. Med.* 2010, **170** (19): 1758–64.

EPA (2006). **Radiation information: radon**. http://www.epa.gov/rpdweb00/radionuclides/radon.html. Retrieved 2007-08-11
Eriksson G, et al. **Exploiting syntax when detecting protein names in text**. Proceedings of *Workshop on NLP in Biomedical Applications* 2002.

Exner HE, Hougardy HP. (1988) **Quantitative image analysis of microstructures**. DGM Informations-gesellschaft Verlag, Oberursel. ISBN 3-88355-132-5.

Fawcett T. **An introduction to ROC analysis.** *Pattern Recognition Letters* 2006, 27:861–874.

Feig SA, Hendrick RE. **Radiation risk from screening mammography of women aged 40–49 years**. *J Natl Cancer Inst Monogr* 1997, **22** (22): 119–24.

Ferlay J, Shin HR, Bray F, *et al*. **Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008**. *International Journal of Cancer* 2010, **127** (12): 2893–2917.

Feuk L., *et al*. **Structural variation in the human genome.** *Nature Reviews Genetics* 2006. **7**, 85–97.

Finkel J*, et al.* **Exploring the boundaries: gene and protein identification in biomedical text**. *BMC Bioinformatics* 2005. **6**:S5.

Fong KM, Sekido Y, Gazdar AF, Minna JD. **Lung cancer • 9: Molecular biology of lung cancer: clinical implications.** *Thorax* (BMJ Publishing Group Ltd.) 2003, 58 (10):

Forgy EW. **Cluster analysis of multivariate data: efficiency versus interpretability of classifications**. *Biometrics* 1965. **21**: 768–769.

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** Suppl 4:S9.

Freudenberg J, Propping P. **A similarity-based method for genome-wide prediction of disease-relevant human genes**. *Bioinformatics* 2002, **18 (Suppl 2)**:S110-S115.

Friedenson B. **Is mammography indicated for women with defective BRCA genes? Implications of recent scientific advances for the diagnosis, treatment, and prevention of hereditary breast cancer**. *MedGenMed* 2000, **2** (1): E9.

Friedenson B. **A theory that explains the tissue specificity of BRCA1/2 related and other hereditary cancers**. *J. Med. Med. Sci.* 2010, **1** (8): 372–384.

Friedenson B. **Preventing hereditary cancers caused by opportunistic carcinogens**. *J Med Med Sci* 2012, **3**: 160–178.

Friedman C, Kra P. Yu H, et al. **GENIES: a natural –language processing system for the extraction of molecular pathways from journal articles**. *Bioinformatics* 2001. **17 (Suppl.1),** S74-82.

Frijters R, Van Vugt M, Smeets R, Van Schaik R, De Vlieg J, Alkema W: **Literature mining for the discovery of hidden connections between drugs, genes and diseases.** *PLoS computational biology* 2010, **6**:e1000943.

Fukuda K, et al. **Toward information extraction: identifying protein names from biological papers**. In Proceedings of the *Pacific Symposium on Biocomputing* 1998. 707-718.

Gao K, Zhou H, Zhang L, Lee J, Zhou Q, Hu S, Wolinsky L, Farrell J, Eibl G, Wong D: **Systemic Disease-Induced Salivary Biomarker Profiles in Mouse Models of Melanoma and Non-Small Cell Lung Cancer.** *PLoS ONE* 2009, **4**:e5875.

Giordano SH, Cohen DS, Buzdar AU, Perkins G, Hortobagyi GN. **Breast carcinoma in men: a population-based study**. *Cancer* (2004, **101** (1): 51–7.

Glazko G, Emmert-Streib F (2009) **Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets.** *Bioinformatics* 25: 2348–2354.

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20**: 93–99.

Goetz T, von der Lieth C-W. **PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts**. *Nucleic Acids Research* 2005. **33**:W774-W778.

Gómez-Vela F & Díaz-Díaz N. **Gene network biological validity based on gene-gene interaction relevance.** *ScientificWorldJournal.* 2014, 2014:540679.

Gopalakrishnan V, Lustgarten JL, Visweswaran S, Cooper GF. **Bayesian rule learning for biomedical data mining.** *Bioinformatics.* 2010, 26 (5): 668-675.

Gorlova OY, Weng SF, Zhang Y et al. **Aggregation of cancer among relatives of never-smoking lung cancer patients.** *International Journal of Cancer* 2007, 121(1): 111–118.

Greene FL. (2002). ***AJCC cancer staging manual*****.** Berlin: Springer-Verlag. ISBN 0-387-95271-3.

Goeman JJ, Buhlmann P. **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics* 2007, 23: 980–987.

Guille A, Chaffanet M, Birnbaum D. **Signaling pathway switch in breast cancer.** *Cancer Cell International* 2013; 13:66.

Günther F, Pigeot I, Bammann K. **Artificial neural networks modeling gene-environment interaction.** *BMC Genet.* 2012, **13**:37.

Guo X, Liu R, Shriver CD, Hu H, Liebman MN. **Assessing semantic similarity measures for the characterization of human regulatory pathways.** *Bioinformatics* 2006, 22(8):967-73.
Hackshaw, AK, Law MR, Wald NJ. **The accumulated evidence on lung cancer and environmental tobacco smoke.** *British Medical Journal* 1997, 315(7114): 980–988.

Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J. **ProMiner: rule-based protein and gene entity recognition**. *BMC Bioinformatics* 2005. **6**:S14.

Harris TJR & McCormick F. **The molecular pathology of cancer.** *Nature Reviews Clinical Oncology* 2010, **7**, 251-265.

Henegar C, Cancello R, Rome S, Vidal H, Clement K, et al. **Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes.** *J Bioinform Comput Biol* 2006, **4**: 833–852.

Herbst RS, Heymach JV, Lippman SM. **Lung cancer**. *New England Journal of Medicine* 2008, **359** (13): 1367–1380.

Hill SM, Neve RM, Bayani N, Kuo WL, Ziyad S, Spellman PT, Gray JW, **Mukherjee S. Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology.** *BMC Bioinformatics* 2012, 13:94.

Hira ZM, Trigeorgis G, Gillies DF. **An algorithm for finding biologically significant features in microarray data based on a priori manifold learning**. *PLoS One.* 2014, 9(3):e90562.

Hirschman L, Park JC, Tsujii J, Wong L, and Wu CH: **Accomplishments and challenges in literature data mining for biology.** *Bioinformatics* 2002, **18:** 1553-1561.

Hirschman L, Colosimo M, Morgan A, Colombe J, Yeh A. **Task 1B: gene list task.** In Proceedings of the *Critical Assessment of Information extraction Systems in Biology (BioCreAtIvE) Workshop*, 2004.

Horn L, Pao W, Johnson DH. (2012). **Harrison's Principles of Internal Medicine** (18th ed.). McGraw-Hill. ISBN 0-07-174889-X.

Hotelling H. **Analysis of a complex of statistical variables into principal components.** *Journal of Educational Psychology* 1933, **24**, 417–441, and 498–520.

Hristovski D, Stare J, Peterlin B, Dzeroski S. **Supporting discovery in medicine by association rule mining in Medline and UMLS.** *Medinfo* 2001. **19**:1344-1348.

Hristovski D, Peterlin B, Mitchell JA, Humphrey SM: **Using literature-based discovery to identify disease candidate genes.** *International Journal of Medical Informatics* 2005, **74**:289-298.

Hua L, Zhou P. **Combining protein-protein interactions information with support vector machine to identify chronic obstructive pulmonary disease related genes.** *Mol Biol (Mosk).* 2014, 48(2):333-43.
Huang DW, Sherman BT, Lempicki RA. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res*. 2009, 37: 1–13.

Huang H, Wu X, Sonachalam M, Mandape SN, Pandey R, MacDorman KF, Wan P, Chen JY. **PAGED: a pathway and gene-set enrichment database to enable molecular phenotype discoveries.** *BMC Bioinformatics* 2012, 13 Suppl 15:S2.

Huerta AM, Salgado H, Thieffry D, Collado-Vides J. **RegulonDB: a database on transcriptional regulation in Escherichia coli.** *Nucleic Acids Res* 1998, 26: 55–59.

Hummel M, Meister R, Mansmann U. **GlobalANCOVA: exploration and assessment of gene group effects.** *Bioinformatics* 2008, 24: 78–85.

Hur J, Liu Z, Tong W, Laaksonen R, Bai JP. **Drug-induced rhabdomyolysis: from systems pharmacology analysis to biochemical flux.** *Chem Res Toxicol*. 2014, 27(3):421-32.

Husmeier D, Werhli AV. **Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with Bayesian networks.** *Comput Syst Bioinformatics Conf.* 2007, 6:85-95.

Hwang T, Atluri G, Xie M, Dey S, Hong C, Kumar V, Kuang R. **Co-clustering phenome-genome for phenotype classification and disease gene discovery.** *Nucleic Acids Res.* 2012, 40(19):e146.

Imaginis Corporation. **Breast Cancer: Statistics on Incidence, Survival, and Screening**. 2006. http://imaginis.com/breasthealth/statistics.asp. Retrieved 2006-10-09.

International Agency for Research on Cancer. **World Cancer Report**. 2008. http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdf. Retrieved 2011-02-26.

International Agency for Research on Cancer. **World Cancer Report**. 2008. http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900. Retrieved 2011-02-26.

Jemal A, Tiwari RC, Murray T, *et al.* **Cancer statistics, 2004**. *CA Cancer J Clin* 2004, **54** (1): 8–29.

Jenkinson G, Zhong X, Goutsias J. **Thermodynamically consistent Bayesian analysis of closed biochemical reaction systems.** *BMC Bioinformatics* 2010, 11:547.

Jensen LJ, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery.** *Nature Reviews Genetics* 2006, **7**:119-129.

Jia P, Zheng S, Long J, Zheng W, Zhao Z. **dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks.** *Bioinformatics* 2011, 27(1):95-102.

Jiang Z, Gentleman R. **Extensions to gene set enrichment.** *Bioinformatics* 2007, 23: 306–313.

Jin P, Han TH, Ren J, et al. **Molecular signatures of maturing dendritic cells: implications for testing the quality of dendritic cell therapies.** *J Transl Med.* 2010 Jan 15;8:4.

Jin S, Zou X. **Construction of the influenza A virus infection-induced cell-specific inflammatory regulatory network based on mutual information and optimization.** *BMC Syst Biol.* 2013, 7:105.

Johannes M, Brase JC, Fröhlich H, Gade S, Gehrmann M, Fälth M, Sültmann H, Beissbarth T. **Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients.** *Bioinformatics* 2010, 26 (17): 2136-2144.

Johannes M, Fröhlich H, Sültmann H, Beissbarth T. **pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery.** *Bioinformatics* 2011, 27(10):1442-3.

Jordan R, Visweswaran S, Gopalakrishnan V. **Semi-automated literature mining to identify putative biomarkers of disease from multiple biofluids.** *Journal of Clinical Bioinformatics* 2014, **4**; 13.

Joshi-Tope G, Vasrik I, Gopinath GR, Matthews L, Schmidt E, et al. **The genome knowledgebase: a resource for biologists and bioinformaticists.** *Cold Spring Harb Symp Quant Biol* 2003, 68: 237–243.

Joshi-Tope G, Gillespie M, Vasrik I, D'Eustachio P, Schmidt E, de Bone B, Jassal B, Gopinath GR, Wu GR, et al. **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Research* 2005, 33: D428-D432.

Kabir Z, Bennett K, Clancy L. **Lung cancer and urban air-pollution in Dublin: a temporal association?** *Irish Medical Journal* 2007, 100 (2): 367–369.

Kalager M, Haldorsen T, Bretthauer M, Hoff G, Thoresen SO, Adami HO. **Improved breast cancer survival following introduction of an organized mammography screening program among both screened and unscreened women: a population-based cohort study.** *Breast Cancer Res* 2009, 11(4):R44.

Kanehisa M, Goto S. **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, 28: 27–30.

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, et al. **The MetaCyc database.** *Nucleic Acids Res,* 2002 30: 59–61.

Kayaalp M et al. **Methods for accurate retrieval of MEDLINE citations in functional genomics** [online], http://trec.nist.gov/pubs/trec12/papers/nlm.genmics.pdf.

Kern JA, McLennan G. (2008). ***Fishman's Pulmonary Diseases and Disorders* (4th ed.).** McGraw-Hill. p. 1802. ISBN 0-07-145739-9.

Khatri P, Draghici S, Ostermeier GC, Krawetz SA. **Profiling gene expression using Onto-Express.** *Genomics* 2002, 79: 266–270.

Khatri P, Draghici S. **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, 21: 3587–3595.

Khatri P, Sellamuthu S, Malhotra P, Amin K, Done A, Draghici S. **Recent additions and improvements to the Onto-Tools.** *Nucleic Acid Research* 2005, 33:W762-W765.

Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, Tarca AL, Draghici S. **Onto-Tools: new additions and improvements in 2006.** *Nucleic Acids Research* 2007, 35:W206-W211.

Khatri P, Drăghici S, Tarca AL, Hassan SS, Romero R. **A system biology approach for the steady-state analysis of gene signaling networks.** *Proc 12th Iberoamerican Congress on Pattern Recognition, CIARP* 2007; Valparaiso, Chile.

Khatri P, Sirota M, Butte AJ. **Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.** *PLoS Comput Biol* 2012, 8(2): e1002375.

Kim DC, Yang CR, Wang X, Zhang B, Wu X, Gao J. **Discovery of lung cancer pathways using reverse phase protein microarray and prior-knowledge based Bayesian networks.** *Conf Proc IEEE Eng Med Biol Soc.* 2011, 2011:5543-6.

Kim I, Pang H, Zhao H. **Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes.** *Stat Med.* 2012, 31(15):1633-51.

Kim JD, Ohta T, Tateisi Y, Tsujii J. **GENIA corpus – a semantically annotated corpus for bio-text mining**. *Bioinformatics* 2003. **19 Suppl. 1**: i180-i182.

Kim JY, Shin KK, Lee AL, Kim YS, Park HJ, Park YK, Bae YC, Jung JS. **MicroRNA-302 induces proliferation and inhibits oxidant-induced cell death in human adipose tissue-derived mesenchymal stem cells.** *Cell Death and Disease* 2014. 5;e1385.

Kim SB, Yang S, Kim SK, Kim SC, Woo HG, et al. **GAzer: gene set analyzer.** *Bioinformatics (Oxford, England)* 2007, 23: 1697–1699.

Kim SY, Volsky DJ. **PAGE: parametric analysis of gene set enrichment.** *BMC Bioinformatics* 2005, **6**: 144.

King JY, Ferrara R, Tabibiazar R, Spin JM, Chen MM, Kuchinsky A, Vailaya A, Kincaid R, Tsalenko A, Deng DX, Connolly A, Zhang P, Yang E, Watt C, Yakhini Z, Ben-Dor A, Adler A, Bruhn L, Tsao P, Quertermous T, Ashley EA. **Pathway analysis of coronary atherosclerosis.** *Physiol Genomics* 2005, 23(1):103-18.
Kirouac DC, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger DA, Sorger PK. **Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks.** *BMC Syst Biol.* 2012, 6:29.

Kitaya K, Yasuo T, Yamaguchi T, et al. **Genes regulated by interferon-gamma in human uterine microvascular endothelial cells.** *Int J Mol Med.* 2007 Nov;20(5):689-97.

Klebanov L, Glazko G, Salzman P, Yakovlev A, Xiao Y. **A multivariate extension of the gene set enrichment analysis.** *J Bioinform Comput Biol* 2007, 5: 1139–1153.

Kong SW, Pu WT, Park PJ. **A multivariate approach for integrating genome-wide expression data and biological knowledge.** *Bioinformatics* 2006, 22: 2373–2380.

Krallinger M, Valencia A, and Hirschman L: **Linking genes to literature: text mining, information extraction, and retrieval applications for biology.** *Genome Biology* 2008, **9**(Suppl.2):S8.

Krämer A, Green J, Pollard J Jr, Tugendreich S. **Causal analysis approaches in Ingenuity Pathway Analysis.** *Bioinformatics* 2014, 30(4):523-30.

Küffner R, Fundel K, Zimmer R. **Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts.** *Bioinformatics* 2005, 21 Suppl 2:ii259-67.

Kumar MG, Patel NM, Nicholson AM, et al. **Reactive oxygen species mediate microRNA-302 regulation of AT-rich interacting domain 4a and C-C motif ligand 5 expression during transitions between quiescence and proliferation.** *Free Radic Biol Med.* 2012 Aug 15;53(4):974-82.

Lacroix M. **Significance, detection and markers of disseminated breast cancer cells**. *Endocrine-related Cancer* 2006, **13** (4): 1033–67.

Lau CS, Wong DT. **Breast cancer exosome-like microvesicles and salivary gland cells interplay alters salivary gland cell-derived exosome-like microvesicles in vitro.** *PLoS One.* 2012, 7(3):e33037.

Laurance, J. (2006). **Breast cancer cases rise 80% since Seventies**. *The Independent* (London). http://www.independent.co.uk/life-style/health-and-wellbeing/health-news/breast-cancer-cases-rise-80-since-seventies-417990.html. Retrieved 2006-10-09.

Lee Y, Wong D: **Saliva: An emerging biofluid for early detection of diseases.** *American Journal of Dentistry* 2009, **22**:241-248.

Leonard JE, Colombe JB, Levy JL: **Finding relevant references to genes and proteins in Medline using a Bayesian approach.** *Bioinformatics* 2002, **18**:1515-1522.

Li C, Li Y, Xu J, Lv J, Ma Y, Shao T, Gong B, Tan R, Xiao Y, Li X. **Disease-driven detection of differential inherited SNP modules from SNP network.** *Gene* 2011, 489(2):119-29.

Li H, Liu C: **Biomarker identification using text mining.** *Computational and Mathematical Methods in Medicine* 2012, 2012: 135780.

Li MX, Kwan JS, Sham PC. **HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis.** *Am J Hum Genet*. 2012, 91(3):478-88.

Liang KC, Patil A, Nakai K. **Discovery of Intermediary Genes between Pathways using Sparse Regression.** *PLOS One* 2015, **10**(9):e0137222.

Lin YC, Hsieh AR, Hsiao CL, Wu SJ, Wang HM, Lian IB, Fann CS. **Identifying rare and common disease associated variants in genomic data using Parkinson's disease as a model.** *J Biomed Sci*. 2014, 21:88.

Lindsay RK, Gordon MD. **Literature-based discovery by lexical statistics.** *J Amer Soc Information Sci* 1999. **50**:574-587.

Liu H, Friedman C. **Mining terminological knowledge in large biomedical corpora**. In Proceedings of the *8th Pacific Symposium on Biocomputing* 2003:415-426.

Lu Y, Liu PY, Xiao P, Deng HW. **Hotelling's T2 multivariate profiling for detecting differential expression in microarrays.** *Bioinformatics* 2005, 21: 3105–3113.

Liu Y, Maxwell S, Feng T, Zhu X, Elston RC, Koyutürk M, Chance MR. **Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from GWAS data.** *BMC Syst Biol.* 2012, 6 Suppl 3:S15.

Lloyd., S. P. **Least squares quantization in PCM**. *IEEE Transactions on Information Theory* 1982. **28** (2): 129–137.

Lustgarten JL, Kimmel C, Ryberg H, Hogan W. **EPO-KB: a searchable knowledge base of biomarker to protein links**. *Bioinformatics* 2008, 24:1418-1419.

Lustgarten JL, Visweswaran S, Bowser RP, Hogan WR, Gopalakrishnan V. **Knowledge-based variable selection for learning rules from proteomic data.** *BMC Bioinformatics* 2009, 10 (Suppl 9):S16.

Lyman GH, et al. **Impact of a 21-gene RT-PCR assay on treatment decisions in early-stage breast cancer: an economic analysis based on prognostic and predictive validation studies.** *Cancer* 2007, 109(6):1011-1018.

Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A. **Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations.** *PLoS Genet.* 2012, 8(5):e1002714.

Madigan MP, Ziegler RG, Benichou J, Byrne C, Hoover RN. **Proportion of breast cancer cases in the United States explained by well-established risk factors**. *Journal of the National Cancer Institute* 1995, **87** (22): 1681–5.

Magbanua MJ, Melisko M, Roy R, Sosa EV et al. **Molecular profiling of tumor cells in cerebrospinal fluid and matched primary tumors from metastatic breast cancer patients with leptomeningeal carcinomatosis.** *Cancer Res* 2013, 73(23):7134-43.

Marietta C, Thompson LH, Lamerdin JE, Brooks PJ. **Acetaldehyde stimulates FANCD2 monoubiquitination, H2AX phosphorylation, and BRCA1 phosphorylation in human cells in vitro: implications for alcohol-related carcinogenesis**. *Mutat. Res.* 2009, **664** (1-2): 77–83.

Martin D, Brun C, Remy E, Mouren P, Thieffry D, et al. **GOToolBox: functional analysis of gene datasets based on gene ontology.** *Genome Biol*. 2004, 5: R101.

Martin E, Kriegel HP, Sander J, Xu X (1996). **A density-based algorithm for discovering clusters in large spatial databases with noise.** *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).* AAAI Press. pp. 226–231. ISBN 1-57735-004-9.

Martin F, Thomson TM, Sewer A, Drubin DA, Mathis C, Weisensee D, Pratt D, Hoeng J, Peitsch MC. **Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks.** *BMC Syst Biol.* 2012, 6:54.

Martinez-Cruz La, Rubio a, Martinez-Chantar ML, Labarga a, Barrio I, et al. **GARBAN: genomic analysis and rapid biological annotation of cDNA microarray and proteomic data.** *Bioinformatics* 2003, 19: 2158–2160.

Mathiesen RR, Fjelldal R, Liestøl K, Due EU et al. **High-resolution analyses of copy number changes in disseminated tumor cells of patients with breast cancer.** *Int J Cancer* 2012, 131(4):E405-15.

McCulloch W, Pitts W. **A Logical Calculus of Ideas Immanent in Nervous Activity.** *Bulletin of Mathematical Biophysics* 1943, 5(4): 115–133.

McDonald DM, Chen H, Su H, Marshall BB. **Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser.** *Bioinformatics* 2004, 20(18): 3370-3378.

McDonald R, Pereira F. **Identifying gene and protein mentions in text using conditional random fields**. *BMC Bioinformatics* 2005, **6**, S6.

Medew J. (2010). **Study finds big risk of cancer in the family**. Sydney Morning Herald. http://www.smh.com.au/lifestyle/wellbeing/study-finds-big-risk-of-cancer-in-the-family-20100929-15xin.html.

Medina PP, Castillo SD, Blanco S, Sanz-Garcia M et al. **The SRY-HMG box gene, SOX4, is a target of gene amplification at chromosome 6p in lung cancer.** *Hum Mol Genet* 2009, 18(7):1343-52.

Merck Manual Professional Edition, Online edition. **Lung Carcinoma: Tumors of the Lungs**. http://www.merck.com/mmpe/sec05/ch062/ch062b.html#sec05-ch062-ch062b-1405. Retrieved 2007-08-15.

Mika S, Rost B. **Protein names precisely peeled off free text**. *Bioinformatics* 2004. **20**:i241-i247.

Milone DH, Stegmayer G, Lopez M, Kamenetzky L, Carrari F. **Improving clustering with metabolic pathway data.** *BMC Bioinformatics* 2014, **15**:101.

Minn AJ, Bevilacqua E, Yun J, Rosner MR. **Identification of novel metastasis suppressor signaling pathways for breast cancer.** *Cell Cycle* 2012, 11(13):2452-7.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat Genet* 2003, 34: 267–273.

Morris MK, Saez-Rodriguez J, Clarke DC, Sorger PK, Lauffenburger DA. **Training signaling pathway maps to biochemical data with constrained fuzzy logic: quantitative analysis of liver cell responses to inflammatory stimuli.** *PLoS Comput Biol.* 2011, 7(3):e1001099.

Muller HM, Kenny EE, Sternberg PW. **Textpresso: an ontology-based information retrieval and extraction system for biological literature.** *PloS Biol*. 2004. **2**:e309.

Namkung J, Raska P, Kang J, Liu Y, Lu Q, Zhu X. **Analysis of exome sequences with and without incorporating prior biological knowledge.** *Genet Epidemiol*. 2011, 35 Suppl 1:S48-55.

Narayanaswamy M, et al. **A biological named entity recognizer.** Proceedings of the *Pacific Symposium on Biocomputing* 2003. 427-438.

National Cancer Institute; **SEER stat fact sheets: Lung and Bronchus**. Surveillance Epidemiology and End Results. 2010

National Cancer Institute. M**ale Breast Cancer Treatment**. 2011. http://www.cancer.gov/cancertopics/pdq/treatment/malebreast/HealthProfessional. National Center for Biotechnology Information (US). Genes and Disease [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1998-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK22183/
National Health and Medical Research Council (1994). ***The health effects and regulation of passive smoking*****.** Australian Government Publishing Service. Archived from the original on September 29, 2007. http://www.obpr.gov.au/publications/submission/healthef/index.html.

National Research Center for Women & Families **2009 Update: When Should Women Start Regular Mammograms? 40? 50? And How Often is "Regular"?** November 2009. http://www.stopcancerfund.org/posts/211.

Neapolitan RE. **Learning Bayesian Networks.** (2004), Pearson Prentice Hall, ISBN 0-13-012534-2.

Nicholas B, Skipp P, Mould R, Rennard S, Davies DE, O'Connor CD, Djukanović R. **Shotgun proteomic analysis of human-induced sputum.** *Proteomics*. 2006, 6(15):4390-401.

Nolen BM, Lokshin AE. **The advancement of biomarker-based diagnostic tools for ovarian, breast, and pancreatic cancer through the use of urine as an analytical biofluid.** *Int J Biol Markers.* 2011, 26(3):141-52.

Novichkova S, Egorov S, Daraseila N: **MedScan, a natural language processing engine for MEDLINE abstracts.** *Bioinformatics* 2003, **19**:1699-1706.

O'Reilly KM, Mclaughlin AM, Beckett WS, Sime PJ. **Asbestos-related lung disease.** *American Family Physician* 2007, 75 (5): 683–688.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. **Kegg: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Research* 1999, 27:29-34.

Oumeraci T, Schmidt B, Wolf T, Zapatka M, Pich A, Brors B, Eils R, Fleischhacker M, Schlegelberger B, von Neuhoff N. **Bronchoalveolar lavage fluid of lung cancer patients: mapping the uncharted waters using proteomics technology.** *Lung Cancer.* 2011, 72(1):136-8.

Parikh A, Huang E, Dinh C, Zupan B, Kuspa A, Subramanian D, Shaulsky G. **New components of the Dictyostelium PKA pathway revealed by Bayesian analysis of expression data.** *BMC Bioinformatics* 2010, 11:163.

Park CY, Wong AK, Greene CS, Rowland J, Guan Y, Bongo LA, Burdine RD, Troyanskaya OG. **Functional knowledge transfer for high-accuracy prediction of under-studied biological processes.** *PLoS Comput Biol.* 2013, 9(3):e1002957.

Park YK, Kang TW, Baek SJ, Kim KI, Kim SY, Lee D, Kim YS. **CaGe: A Web-Based Cancer Gene Annotation System for Cancer Genomics.** *Genomics Inform*. 2012 Mar;**10**(1):33-39. Epub 2012 Mar 31.

Parker JS, et al. **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *Journal of Clinical Oncology* 2009, 27(8):1160-1167.

Patel KJ, Yu VP, Lee H, *et al.* **Involvement of Brca2 in DNA repair**. *Mol. Cell* 1998, **1** (3): 347–57. http://linkinghub.elsevier.com/retrieve/pii/S1097-2765(00)80035-0.

Patel S & Lyons-Weiler J. **caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer.** *Applied Bioinformatics* 2004, 3(1):49-62.

Pavlidis P, Qin J, Arango V, Mann J, Sibille E. **Using the Gene Ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex.** *Neurochem Res.* 2004, 29: 1213–1222.

Pearl J. **Probabalistic Reasoning in Intelligent Systems: Networks of Plausible Inference,** Morgan-Kaufmann, San Mateo, CA, 1998.

Peppercorn J. **Breast Cancer in Women Under 40**. *Oncology* 2009, **23** (6). http://www.cancernetwork.com/cme/article/10165/1413886.

Perez-Iratxeta C, Bork P, Andrade AM. **XplorMed: a tool for exploring MEDLINE abstracts**. *Trends Biochem Sci* 2001. **26**:573-575.

Perez-Iratxeta C, Bork P, Andrade MA. **Association of genes to genetically inherited diseases using data mining**. *Nature Genetics* 2002. **31**:316-319.

Perez-Iratxeta C, Wist M, Bork P, Andrade MA. **G2D: a tool for mining genes associated to disease**. *BMC Genetics* 2005. **6**:45.

Peto R, Lopez AD, Boreham J, *et al*. ***Mortality from smoking in developed countries 1950–2000: Indirect estimates from National Vital Statistics*.** 2006. Oxford University Press. ISBN 0-19-262535-7. /.

Pinkel D, Albertson DG. **Array comparative genomic hybridization and its applications in cancer.** *Nature Genetics* 2005, **37**:11-17.

Qaiser, BMM. (2012) **Principles and Practice of Chemotherapy**. Jaypee Brothers Medical Publishers, London. p.225.

Qin X, Xu H, Gong W, Deng W. **The Tumor Cytosol miRNAs, Fluid miRNAs, and Exosome miRNAs in Lung Cancer.** *Front Oncol.* 2015, **4**:357.

Qiu C, Wang J, Cui Q. **miR2Gene: pattern discovery of single gene, multiple genes, and pathways by enrichment analysis of their microRNA regulators.** *BMC Syst Biol.* 2011, **5** Suppl 2:S9.

Qui L, Ding L, Huang J, Wang D, Zhang J, Guo B**. Induction of copper/zinc-superoxide dismutase by CCL5/CCR5 activation causes tumor necrosis factor-α and reactive oxygen species production in macrophages.** *Immunology* 2009, 128:e325-e334.

Quinlan R. **Learning efficient classification procedures**. *Machine Learning: an artificial intelligence approach*, Michalski, Carbonell & Mitchell (eds.), Morgan Kaufmann, 1983, p. 463-482.

Quinlan JR. **Simplifying decision trees**. International Journal of Man-Machine Studies 1987, **27** (3): 221.

Rahnenführer J, Domingues FS, Maydt J, Lengauer T. **Calculating the statistical significance of changes in pathway activity from gene expression data.** *Stat Appl Genet Mol Biol* 2004, 3: Article 16.

Ramshankar V, Krishnamurthy A. **Lung cancer detection by screening – presenting circulating miRNAs as a promising next generation biomarker breakthrough.** *Asian Pac J Cancer Prev*. 2013, 14(4):2167-72.

Ren H, Francis W, Boys A, Chueh AC, Wong N, La P, Wong LH, Ryan J, Slater HR, Choo KH. **BAC-based PCR fragment microarray: high-resolution detection of chromosomal deletion and duplication breakpoints**. *Human Mutation* 2005, **25** (5): 476–82.

Rissanen J. **Modeling by shortest data description.** *Automatica*, 1978, 14 (5): 465-471.

Robinson MD, Grigull J, Mohammad N, Hughes TR. **FunSpec: a web-based cluster interpreter for yeast.** *BMC Bioinformatics* 2002, **3**: 35.

Rotunno M, Hu N, Su H, Wang C et al. **A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma.** *Cancer Prev Res (Phila)* 2011, 4(10):1599-608.

Salgia R, Skarin AT. **Molecular abnormalities in lung cancer**. *Journal of Clinical Oncology* 1998, **16** (3): 1207–1217.

Santoro E, DeSoto M, and Hong Lee, J. (2009). **Hormone Therapy and Menopause**. National Research Center for Women & Families. http://www.center4research.org/2010/03/hormone-therapy-and-menopause/.

Sariego J. **Breast cancer in the young patient**. *The American surgeon* 2010, **76** (12): 1397–1401.

Schick S, Glantz S. **Philip Morris toxicological experiments with fresh sidestream smoke: more toxic than mainstream smoke.** *Tobacco Control* 2005**, 14** (6): 396–404.

Schwartz AS, Hearst MA. **A simple algorithm for identifying abbreviation definitions in biomedical text**. In Proceedings of the *8th Pacific Symposium on Biocomputing* 2003, 451-462.

SEER data (SEER.cancer.gov) **Median Age of Cancer Patients at Diagnosis 2002-2003.**

Sekimizu T, Park HS, Tsujii J**. Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts**. In *Genome Inform Ser Workshop Genome Inform*. 1998. **9**:62-71.

Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.** *Bioinformatics* 2005, **21**:3191-3192.

Shah D, Wanchu A, Bhatnagar A. **Interaction between oxidative stress and chemokines: Possible pathogenic role in systemic lupus erythematosus and rheumatoid arthritis.** *Immunobiology* 2011, **216**:1010-1017.

Shames DS, Girard L, Gao B, Sato M et al. **A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies.** *PLoS Med* 2006, 3(12):e486.

Shinawi M, Cheung SW. **The array CGH and its clinical applications.** *Drug Discovery Today* 2008, 13:760-769.

Shojaie A, Michailidis G. **Analysis of gene sets based on the underlying regulatory network.** *J Comput Biol* 2009, 16: 407–426.

Sibson R. (1973). **SLINK: an optimally efficient algorithm for the single-link cluster method**. *The Computer Journal (British Computer Society)* 1973. **16** (1): 30–34.

Silver M, Janousova E, Hua X, Thompson PM, Montana G; **Alzheimer's Disease Neuroimaging Initiative. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression.** *Neuroimage.* 2012, 63(3):1681-94.

Smalheiser NR, Swanson DR. **Linking estrogen to Alzheimer’s disease: an informatics approach**. *Neurology* 1996, **47**:809-810.

Smalheiser NR, Swanson DR. **Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses**. *Comp. Meth. Prog. Biomed* 1998, **57**:149-153.

Solomonoff R. **An Inductive Inference Machine**. *IRE Convention Record, Section on Information Theory, Part 2*, 1957. pp., 56-62.

Sopori M. **Effects of cigarette smoke on the immune system.** *Nature Reviews. Immunology* 2002**. 2** (5): 372–7.

Srinivasan P: **Text mining: generating hypotheses from MEDLINE.** *Journal of the American Society for Information Science and Technology* 2004, **55**:396-413.

Stapley BJ, Benoit G. **Bibliometrics: information retrieval and visualization from co-occurrence of gene names in Medline abstracts**. *Pac. Symp. Biocomput*, 2000. **5**:529-540.
Starczynowski DT, Lockwood WW, Deléhouzée S, Chari R et al. **TRAF6 is an amplified oncogene bridging the RAS and NF-κB pathways in human lung cancer.** *J Clin Invest* 2011, 121(10):4095-105.

Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J. **Detecting gene relations from MEDLINE abstracts**. *Pac Symp Biocomputing* 2001. 483-495.

Stingo FC, Chen YA, Tadesse MG, Vannucci M. **Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes.** *Ann Appl Stat.* 2011, **5**(3):1978-2002.

Stoddard F,II, Brooks AD, Eskin BA, Johannes GJ. **Iodine alters gene expression in the MCF7 breast cancer cell line: evidence for an anti-estrogen effect of iodine**. *International journal of medical sciences* 2008, **5** (4): 189–96.

Strachan T, Read AP. ***Human Molecular Genetics***. Garland Science. 2010.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *PNAS USA* 2005, **102**: 15545–15550.

Subramanian J, Govindan R. **Lung cancer in never smokers: a review.** *Journal of Clinical Oncology* (American Society of Clinical Oncology) 2007, **25** (5): 561–570.

Sugimoto M, Wong DT, Hirayama A, Soga T, Tomita M. **Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles.** *Metabolomics.* 2010, 6(1):78-95.

Sulik GA. (2010). ***Pink Ribbon Blues: How Breast Cancer Culture Undermines Women's Health*.** USA: Oxford University Press. pp. 200–3. ISBN 0-19-974045-3.

Sun H, Fang H, Chen T, Perkins R, Tong W. **GOFFA: gene ontology for functional analysis–a FDA gene ontology tool for analysis of genomic and proteomic data.** *BMC Bioinformatics* 2006, 7 Suppl 2: S23.

Sun J, Zhao M, Jia P, Wang L, Wu Y, Iverson C, Zhou Y, Bowton E, Roden DM, Denny JC, Aldrich MC, Xu H, Zhao Z. **Deciphering Signaling Pathway Networks to Understand the Molecular Mechanisms of Metformin Action.** *PLoS Comput Biol.* 2015, 11(6):e1004202.

Swanson DR. **Fish oil, Raynaud's syndrome, and undiscovered public knowledge.** *Perspect Biol Med* 1986, **30**:7-18.

Swanson DR. **Migraine and magnesium: eleven neglected connections.** *Perspect. Biol. Med.* 1988, **31**:526-557.

Swanson DR. **Somatomedin C and arginine: Implicit connections between mutually isolated literatures.** *Perspect. Biol. Med*. 1990, **33**:157-186.
Swanson DR: **Medical literature as a potential source of new knowledge.** *Bulletin of the Medical Library Association* 1990, **78**:29-37.

Tan D, Zander DS. **Immunohistochemistry for Assessment of Pulmonary and Pleural Neoplasms: A Review and Update**. *Int J Clin Exp Pathol* 2008, **1** (1): 19–31.

Tanabe L, et al. **MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling**. *Biotechniques* 1999, **27**:1210-1217.

Tanabe L, Wilbur WJ. **Tagging gene and protein names in biomedical text.** *Bioinformatics* 2002, **18**:1124-1132.

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, et al. **A novel signaling pathway impact analysis (SPIA).** *Bioinformatics* 2009, 25: 75–82.

Tenenbaum JB, de Silva V, Langford JC. **A Global Geometric Framework for Nonlinear Dimensionality Reduction.** *Science* 2000, 290: 2319–2323.

Theruvathu JA, Jaruga P, Nath RG, Dizdaroglu M, Brooks PJ. **Polyamines stimulate the formation of mutagenic 1, N2-propanodeoxyguanosine adducts from acetaldehyde.** *Nucleic Acids Res.* 2005, **33** (11): 3513–20.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Res* 2003, 13: 2129–2141.

Thun MJ, Henley SJ, Burns D, Jemal A, Shanks TG, Calle EE. **Lung cancer death rates in lifelong nonsmokers**. *J. Natl. Cancer Inst.* 2006, **98** (10): 691–699.

Thun MJ, Hannan LM, Adams-Campbell LL, et al. Adami, Hans-Olov. ed. **Lung Cancer Occurrence in Never-Smokers: An Analysis of 13 Cohorts and 22 Cancer Registry Studies**. *PLoS Medicine* 2008, **5** (9): e185.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, et al. **Discovering statistically significant pathways in expression profiling studies.** *PNAS USA* 2005, **102**: 13544–13549.

Tiffin N, et al. **Integration of text- and data-mining using ontlogies successfully selects disease gene candidates**. *Nucleic Acids Research* 2005, **33**:1544-1552.

Travis WD. **Pathology of lung cancer.** *Clinics in Chest Medicine* 2002, **23** (1): 65–81.

Tredwell GD, Miller JA, Chow HH, Thompson PA, Keun HC. **Metabolomic characterization of nipple aspirate fluid by (1)H NMR spectroscopy and GC-MS.** *J Proteome Res.* 2014, 13(2):883-9.

Turner FS, Clutterbuck DR, Semple CA. **POCUS: mining genomic sequence annotation to predict disease genes**. *Genome Biol*. 2003, **4**(11):R75. Epub 2003 Oct 10.

Tyson DR, Ornstein DK. **Proteomics of cancer of hormone-dependent tissues**. *Adv Exp Med Biol.* 2008, 630:133-47.

van den Berg RA, Rubingh CM, Westerhuis JA, van der Werf MJ, Smilde AK. **Metabolomics data exploration guided by prior knowledge.** *Anal Chim Acta*. 2009, 651(2):173-81.

van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG. **A new web-based data mining tool for the identification of candidate genes for human genetic disorders**. *Eur. J. Hum. Genet.* 2003, **11**:57-63.

Vapnik V, Chervonenkis A. **A note on one class of perceptrons**. *Automation and Remote Control*, 1964. **25**.

Vaporciyan AA, Nesbitt JC, Lee JS, *et al*. (2000). *Cancer Medicine*. B C Decker. pp. 1227–1292. ISBN 1-55009-113-1.

Veenstra T, Conrads T, Hood B, Avellino A, Ellenbogen R, Morrison R: **Biomarkers: Mining the Biofluid Proteome.** *Molecular & Cellular Proteomics* 2005, **4**:409-418.

Venturi S. **Is there a role for iodine in breast diseases?** *The Breast* 2001, **10** (5): 379–382.

Walker SH, Duncan DB (1967). **Estimation of the probability of an event as a function of several independent variables.** *Biometrika* 1967, **54**: 167–178.

Wang K, Li M, Hakonarson H. **Analysing biological pathways in genome-wide association studies.** *Nat Rev Genet.* 2010, 11(12):843-54.

Wagner PD, Srivastava S. **New paradigms in translational science research in cancer biomarkers.** *Transl Res* 2012, **159**(4):343-353.

Weeber M, Klein H, Aronson AR, Mork JG, de Jong-van den Berg LT, Vos R. **Text-based discovery in biomedicine: the architecture of the DAD-system**. Proceedings of *AMIA Annual Fall Symosium* 2000, 903-907.

Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G. **Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide**. *J. Am. Med. Inform. Assoc*. 2003, **10**:252-259.

Weiss M, Hermsen M, Meijer G, Van Grieken N, Baak J, Kuipers E, Van Diest P. **Comparative genomic hybridization**. *Molecular Pathology* 1999, 52:243-251.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchecko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Research* 2007 Jan; **35(Database issue):**D5-12. Epub 2006 Dec 14.

Wilbur WJ, Yang Y. **An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts**. *Comput Biol Med* 1996. **26**:209-222.

WHO International Agency for Research on Cancer, **Tobacco Smoke and Involuntary Smoking.** *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* 83. 2002.

WHO international Agency for Research on Cancer **Press Release No. 180,** December 2007.

WHO, **Breast cancer: prevention and control**. http://www.who.int/cancer/detection/breastcancer/en/index1.html.

Xiong H. **Non-linear tests for identifying differentially expressed genes or genetic networks.** *Bioinformatics* 2006, 22: 919–923.

Wooster R, Weber BL. **Breast and ovarian cancer**. *N. Engl. J. Med.* 2003, **348** (23): 2339–47.

World Cancer Report, 2012

World Cancer Report, 2014

World Cancer Report, 2008

World Health Organization. **Gender in lung cancer and smoking research**. 2004. http://www.who.int/gender/documents/en/lungcancerlow.pdf. Retrieved 2007-05-26.

Wren JD, Bekeredjian R, Stewart JA, Shohet RV, and Garner HR: **Knowledge discovery by automated identification and ranking of implicit relationships.** *Bioinformatics* 2004, **20**: 389-398.

Xu X, Veenstra T: **Analysis of biofluids for biomarker research.** *Proteomics Clinical Applications* 2008, **2**:1403-1412.

Xuan W, et al. **GeneInfoMiner – a web server for exploring biomedical literature using batch sequence ID**. *Bioinformatics,* 2005, **21**:3452-3453.

Xuan W, Wang P, Watson SJ, Meng F: **Medline search engine for finding genetic markers with biological significance.** *Bioinformatics* 2007, **23**: 2477-2484.

Yager JD, Davidson NE. **Estrogen carcinogenesis in breast cancer**. *New Engl J Med* 2006, **354** (3): 270–82.

Yang JO, Charny P, Lee B, Kim S, Bhak J, Woo HG. **GS2PATH: a web-based integrated analysis tool for finding functional relationships using gene ontology and biochemical pathway data.** *Bioinformation* 2007, 2(5):194-6.

Yang P, Patrick E, Tan SX, Fazakerley DJ, Burchfield J, Gribben C, Prior MJ, James DE, Hwa Yang Y. **Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway.** *Bioinformatics* 2014, 30(6):808-14.

Ye J, Fang L, Zheng H, Zhang Y, Chen J, et al. **WEGO: a web tool for plotting GO annotations.** *Nucleic Acids Res* 2006, 34: W293–W297.

Yeh A, Hirschman L, Morgan A, Colosimo M. Task **1A: gene-related name mention finding evaluation.** In Proceedings of the *Critical Assessment of Information extraction Systems in Biology (BioCreAtIvE) Workshop***,** 2004.

Younesi E, Toldo L, Muller B, Friedrich CM, Novac N, Scheer A, Hofmann-Apitius M, Fluck J: **Mining biomarker information in biomedical literature.** *BMC Medical Informatics and Decision Making 2012*, **12**:148.

Yu H, Hripcsak G, Friedman C. **Mapping abbreviations to full forms in biomedical articles**. *J Amer. Med. Inform. Assoc*. 2002. **9**:262-272.

Yu H, Hatzivassiloglou V, Friedman C, et al. **Automatic extraction of gene and protein synonyms from MEDLINE and journal articles**. In Proceedings of the *AMIA Symposium* 2002, 919-923.

Yu H, Agichtein E. **Extracting synonymous gene and protein terms from biological literature**. *Bioinformatics* 2003, **19 Suppl. 1**: i340-i349.

Yuryev A. **Gene expression profiling for targeted cancer treatment.** *Expert Opin Drug Discov*. 2015, 10(1):91-9.

Zeeberg B, Feng W, Wang G, Wang M, Fojo A, et al. **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003. 4: R28.

Zeeberg B, Qin H, Narasimhan S, Sunshine M, Cao H, et al. **High-throughput gominer, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (cvid).** *BMC Bioinformatics* 2005, 6: 168.

Zhang B, Kirov S, Snoddy J. **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005, 33: W741–W748.

Zhang J, Ou JX, Bai CX. **Tobacco smoking in China: prevalence, disease burden, challenges and future strategies**. *Respirology* 2011, **16** (8): 1165–1172.

Zhao X, Zuo X, Qin J, Liang Y, Zhang N, Luan Y, Rao S. **A novel biological pathway expansion method based on the knowledge of protein-protein interactions.** *Yi Chuan*. 2014, 36(4):387-94.

Zhao Y, Chen MH, Pei B, Rowe D, Shin DG, Xie W, Yu F, Kuo L. **A Bayesian Approach to Pathway Analysis by Integrating Gene-Gene Functional Directions and Microarray Data.** *Stat Appl Genet Mol Biol.* 2013, **12**(3):393-412.

Zheng Q, Wang XJ. **GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis.** *Nucleic Acids Res* 2008. 36: W358–W363.

Zhou G, Zhang J, Su J, et al. **Recognizing names in biomedical texts: a machine learning approach**. *Bioinformatics* 2004**, 20**:1178-1190.

Zhou H, Zheng T. **Bayesian hierarchical graph-structured model for pathway analysis using gene expression data.** *Stat Appl Genet Mol Biol.* 2013, **12**(3):393-412.

Zhou M, Conrads T, Veenstra T: **Proteomics approaches to biomarker detection.** *Briefings in Functional Genomics and Proteomics* 2005, **4**:69-75.

Zhu D. **Semi-supervised gene shaving method for predicting low variation biological pathways from genome-wide data.** *BMC Bioinformatics* 2009, 10 Suppl 1:S54.

Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H: **Application of a new probabilistic model for mining implicit associated cancer genes from OMIM and Medline.** *Cancer Informatics* 2006, **2**:361-371.