GC 3 Final report—Tom Keating

**Mining Microarray Expression Data for Cancer Screening**

**I. Introduction and motivation**

Can we develop a test, which can accurately detect early-stage lung cancer with a simple blood sample? In this paper we analyze the 'TCGA - LUSC (Lung Cell Squamous Carcinoma)' Expression Profiling by Array Dataset’\* to see whether such a test is possible.

Lung cancer is one of the deadliest cancers in the world and LUSC is the one of the most common subtypes of lung cancer. The microarray measures the activity of genes in a sample by performing selective binding and measuring fluorescence of mRNA fragments (or gene 'transcripts': recall that DNA is transcribed into mRNA and released to the cytosome where it encounters ribosomes and is then translated into proteins). Using a microarray, rather than imaging, has advantages in terms of sensitivity and early detection: lung cancers begin small and may take months to years before becoming identifiable using x-rays or other imaging techniques. Obtaining tissue samples is invasive and impractical, whereas gene expression can be measured in peripheral blood samples.

The dataset consists of 551 patients. Every patient's array has 56,907 different transcripts. It is imbalanced, with 49 patients diagnosed as healthy and the rest having cancer (n=502). Each gene is labeled but is not annotated to describe its function; hence, a publicly available database of gene annotations will be added in Section 2. These annotations will be of great importance for selecting subsets of genes that are specific to lung cell activity, and whose functions can be analyzed in various ways to interpret the results from a biological standpoint as well as a purely analytic one.

With so many genes the natural question arises, how do we choose which ones? Why not use the whole array? We have to select a limited subset to avoid model overfitting, for one, and we need to consider cost as well. Microarrays, such as the whole genome array from Affymetrix, are over $500 each, whereas 96-well microtiter plates (Carolina Biosciences) are about $4 each. Simple microtiter plates are better suited for point of care testing: they can be used in conjunction with off-the-shelf equipment, and are practical and affordable. Using them would enable in-office blood draws and testing with rapid turnaround (less than one hour).

We hypothesize that the difference between expression in normal versus tumor type samples will provide enough signal to distinguish between the two types. We then refine the hypothesis and develop a binary classifier. Model performance will depend critically on

\*Available through https://portal.gdc.cancer.gov/

low false positive and false negative rates. The assay will also be designed for specificity, that is, to detect lung cancer but not other diseases or cancers: this is done to avoid false positives and to assist in differentiating between (primary) lung cancer and others. It is also designed this way because it is meant to be a screening tool for patients at higher risk for lung cancers, e.g. smokers, which allows some flexibility and guidance over gene selection.

**II. Data wrangling, cleaning, and a key discovery**

Our model depends on both the expression array dataset, and the gene annotations file. Each is wrangled and explored on its own, then combined for additional analysis.

**IIa. The LUSC expression array dataset**

The LUSC database was formatted using multiple delimiters and mixed datatypes, and initially suffered from readability problems. In addition, as we will see later, the expression data was raw (not log scaled) and in some instances had columns where expression levels were zero. These issues were addressed in the initial phase of the project.

A few experiments were run using different delimiter choices to import the data; once the data frame was readable using semicolon delimiters, patient identifier codes were stripped and substituted with numbers 1-551. (The patient codes were long alphanumeric strings. Having them in the dataset created long, highly cluttered column indices, so it was better to remove them). The new column headers, as well as the expression data itself, needed new names, so there were renamed (‘SampleType and ‘NAME’). To fix the issues with mixed data types, the sample type labels were switched from normal and tumor to 0 and 1, respectively, then the dataframe was converted to float values (which oddly took multiple attempts).

Upon taking an initial look at the data, certain questions immediately arose; in particular, whether any steps had been taken to normalize the data. At the outset, the dataset stated it had been normalized by TPM (transcripts per kilobase million) to make it interpretable across the samples, but that it needed further processing. Whether it had been normalized by taking the log of the intensity (a customary analysis step) was as yet unknown.

As a first step, the data was grouped by sample type for graphing. The graphs offered two immediate insights: 1) some of the expression levels were probably zero and 2) the distributions of normal versus tumor expression levels were different.

The latter point was crucially important. If we assumed there was no difference between the samples, the number of samples alone should have caused the variance to drop for tumor expression, because there were 502 tumor samples and only 49 normal. Instead, we see the opposite.

Secondly, with zero values for some samples and some genes present in the set, we needed to proceed with care in normalizing the data.

A blue and orange graph

Description automatically generated

Fig.1 A typical KDE plot of expression levels; normal is in blue, while tumor is in orange. Due to smoothing, some values appear to be less than zero. Note the difference in spread.

Sorting the data revealed many genes in normal tissue samples were inactive (expression level was zero)—because of this, it would create a lot of NANs when taking the log of the expression levels. We faced an important decision in terms of next steps:

One approach would be to simply discard or ignore genes with zero expression in normal tissue. It is possible that some of the genes active in cancerous samples may have nothing to do with cancer, but rather, reflect other biological processes inactive in healthy patients (e.g. inflammatory response). In other words, we don't want to select genes that would lead to false positives. (The issue of false positives and false negatives will be of enormous importance in evaluating the performance of the model).

The second, more neutral approach is to add a tiny offset (1x10-6) to all the expression data. It barely shifts the mean and has no effect on standard deviation. This approach avoids discarding genes early in the analysis and is a more conservative approach. (Besides, the practice of adding a tiny constant to the expression levels is common to most differential sequencing algorithms). In what follows we chose this latter approach to deal with the zeroes in expression data.

**IIb. Wrangling the gene annotations dataset and down-selecting targets**

Annotations of human genes were obtained from the following publicly-available source: <https://www.alliancegenome.org/downloads#gene-descriptions> which was found through this article: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7304461/> on automated gene annotation. The compressed, tab-delimited format was downloaded and unzipped for further processing. (While this sounds simple, a lot of experimentation had to take place before finding the best way to parse the data and load into a dataframe).

*Comments and rationale on next steps*

At this stage the two data sets included expression data for 57000 genes, and annotations for some 47000 of them\*. It was possible at this stage to forget ahead and merge the expression data with the annotations, but with so many predictors (and relatively few normal sample measurements), overfitting was almost guaranteed. Bio-science criteria needed to be developed for down-selecting the final set of genes for analysis.

The annotations alone were used to make these decisions. It was essential to keep measurement data (of tumors) out of the picture at this stage. Ultimately a few simple criteria were chosen for candidate genes:

1) Must be active in (and specific to) normal lung cell tissue. This meant the genes in normal lung tissue with zero expression were dropped, but also, the expression of genes that reflected background activity for cell types not specific to the lung was also used to eliminate certain targets and increase specificity.

2) Known biomarkers of disease should be excluded (to prevent false positives)

3) Genes implicated in cancer or carcinoma nonspecific to the lung were excluded.

\*Some genes are annotated ‘unknown’, and many ‘genes’ in the microarray set are regulatory factors, such as microRNAs, or pseudogenes, which do not code for proteins. This still leaves us with a rich set of genes for analysis.

The last decision was based on the project goal to increase the specificity of the test, to identify lung cancers only.

The use of expression data, while not used in connection with the selection criteria, was used in EDA for hypothesis testing and ‘sanity-checking’ as the analysis proceeded.

The selection criteria were implemented by searching for key words, and then dropping or masking out those genes containing them. As the size of the candidate set became reasonably small (less than 50 targets), the annotations were read in full to ensure that the selections met the criteria above.

As a result, the final candidate set contained 24 genes (later narrowed to 20, as explained in detail later). This number of genes was small enough that overfitting should not be an issue (tumor type samples numbered 502, so it was essential to get below 50 target genes; using only 20 was even better for practical reasons).

**III. Exploratory Data Analysis**

During data wrangling, it became clear that the normal and tumor genes showed unexpected changes in variance. A very useful explanation was found after a literature review. One source is quoted below:

“Variance (in expression) has been largely ignored because it has been considered solely in the context of experimental reproducibility, and therefore something that must be reduced. This was a reasonable bias in the early days of microarrays, but the robustness and reproducibility of the current generation of array platforms allows us to look at additional drivers of variance in gene expression studies. Increasingly there is evidence that biological sources of variation may play an important role in determining cellular and organismal phenotypes...”\*

Note that this article was written in 2011, and we have advanced a long way from the ‘early days’ of microarrays. Because we want to distinguish between populations of phenotypically different cell types, these built-in changes in variance were chosen as the key feature to use in our model. Moving forward with this hypothesis, we calculated the mean and standard deviations of normal and tumor cell types for all the genes in the array database. In addition, the difference in means for both types, divided by the normal variance, was calculated for each gene. This variable (‘Normstd’) was used for EDA and modified slightly in the final model. Using the final list of target genes, we plotted the absolute value of the mean expression differences divided by the normal sample standard deviation. The results are shown in Figure 2, below.

\*Ref: PLoS Genet. 2011 Aug; 7(8): e1002207. Published online 2011 Aug 11. “Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease”

A graph of numbers and letters

Description automatically generated with medium confidence

**Figure 2.** Absolute expression differences between normal and tumor samples, normalized by standard deviation of expression in normal. Many of these signals are very strong.

Figures 2 and 3 confirmed the hypothesis that there were strong differences in expression levels for the genes in our target set. In Figure 3, below, the plot of expression in the two tissue types (normal = 0) is quite dramatic, with the expression markedly different in tumor type tissue. In Figure 2, above, the absolute values are plotted (because in some instances, normal genes are being downregulated in the tumor cells).

A graph of a graph

Description automatically generated with medium confidence

**Figure 3**. The expression profiles for SIM2, a gene in our target set involved in regulation of transcription of RNA polymerase II and embryonic lung development. Note the sharp change in variance, with activity tightly controlled at low levels in normal tissue.

As shown in Figure 3, SIM2 is being activated in tumors, whereas others are being suppressed; this is true of several genes in the set. As an additional check we passed our gene list to the gene ontology resource (<https://geneontology.org/>), which produced a list of biological processes involved:

1. MutS-gamma complex: A heterodimer involved in the stabilization of DNA recombination intermediates, the promotion of crossover recombination, and the proper assembly of the synaptonemal complex in meiotic prophase nuclei.
2. PR-DUB complex: A multimeric protein complex that removes monoubiquitin from histone H2A.
3. Polycomb repressive deubiquitinase complex 7-dehydrocholesterol reductase activity.
4. Inferior colliculus development: The inferior colliculus (IC) (Latin, lower hill) is the principal midbrain nucleus of the auditory pathway and receives input from several more peripheral brainstem nuclei in the auditory pathway, as well as inputs from the auditory cortex.

A couple of these are unusual (the first, because one would not expect chromosomal crossover and meiosis to be happening in these tissues at all), and the last is strange for obvious reasons. The reasons behind this will be seen later as we delve into the specific annotations. Pathways 2 and 3 are involved in gene regulation, cell cycle progression and cellular identity.

A deep dive into the annotations will be performed to allay any concerns about lack of specificity. We should note that, even though we are striving to avoid non-specific gene activity, certain cellular processes necessary for survival are common to all cell types; we are weeding out the obvious cases but likely can not exclude every non-specific factor.

**IV. Preprocessing**

The preprocessing for this analysis is made somewhat simpler by the fact that none of our data are categorical (except for the output class), so the creation of dummy variables is unnecessary. As with most modeling efforts we intend to proceed with a train/test split, the key difference being that after creating the splits, we then have to take the training set for normal genes and calculate the standard deviation for each.

Earlier in the EDA phase we used the standard deviation for all normal genes, but in order to avoid data leakage, the calculation has to take place after the split. With fewer samples in the calculation, this will lead to slightly larger standard deviations, and slightly weaker features.

Once the training sample standard deviations are calculated, the expression data is divided by the normal standard deviation, so that each observation in the set (normal or tumor, for all target genes) is normed. During the EDA phase we looked at difference in expression from normal to tumor: this approach is not used for the model at all. We only use normed expression of the target gene set for our predictions, to avoid data leakage.

The training/test split was 0.33, since we had only 49 normal samples to begin with and using a test set of less than 16 normal samples was undesirable for model evaluation. Finally, the train/test had to be stratified to ensure that the proportion of normal and tumor samples was consistent between train and test splits.

The approach outlined above is fairly straightforward; that said, it has to be re-done any time the target gene list (and train/test split) is changed. Only one such change was made as modeling proceeded, but it might have been worthwhile to automate this part of the pre-processing, to speed up investigation of other target sets in the future.

**V. Model Development, Part 1.**

For the classifier we decided to begin with a logistic regression model, using the normalized train/test data for the 24-gene set determined earlier. Initial results, before optimization, were very encouraging, as shown in Figure 4, below.

A chart of a comparison of a color

Description automatically generated with medium confidence

**Figure 4.** Confusion matrix showing initial results for the logistic regression model. Overall accuracy was 96%, but the presence of false positive and negative was a concern.

Following these initial results, the feature importance was examined and compared with the earlier EDA results. Table 1, below, lists the feature importances. Because of the observed differences, it seemed worthwhile to do an exhaustive check of the target gene list, in case certain aspects of the genes were overlooked using the initial screening criteria. (Recall that the screening was done automatically based on key words, not by exhaustively reading through all the annotations). With a smaller set in hand, we now set about that task.

**Table 1**. Feature importances for initial target gene set.

Gene Importance

CCDC40 5.202279

TMEM38B 4.931180

CIC 2.930689

MSH5 2.868202

SDHC 2.844702

SIM2 2.780806

NAT2 1.302780

ZNF157 1.093700

STK40 0.995118

NKIRAS2 0.781099

SNHG20 0.585136

LAMA3 0.029760

ASXL1 -0.173980

DPPA4 -0.214109

MAN1A2 -0.841604

DPPA2 -1.016413

MAN2A1 -1.032632

TBX4 -1.099089

ATXN1L -1.483904

COPA -1.610834

MIR367 -1.684682

CFC1B -3.569829

NKIRAS1 -3.616764

FENDRR -5.074418

Because Table 1 shows these genes vary widely in importance, we wondered whether there were particular reasons having to do with gene function, and whether some genes really belong in our set. The gene functions and analysis are as follows:

**CCDC40**: This gene is involved in axonemal dynein complex assembly; determination of left/right symmetry; and lung development. Acts upstream of or within cilium assembly; flagellated sperm motility; and regulation of cilium beat frequency. Located in axoneme. It is implicated in primary ciliary dyskinesia 15.

Primary ciliary dyskinesia is a rare inherited disorder which can result in lung infections, as the cilia work to clear lung airways of debris and pathogens. Because it is rare, it makes sense to keep this gene in the set; moreover, ciliary dyskinesia could be secondary to lung cell squamous cell cancer.

**TMEM38B:** Predicted to enable potassium channel activity. Predicted to be involved in potassium ion transmembrane transport. Predicted to act upstream of or within several processes, including cellular response to caffeine; lung development; and regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion. Located in nucleus. Implicated in osteogenesis imperfecta type 14.

Comment: these are critically important cellular functions but do not appear linked to other disease, so this will be kept.

**CIC:** This gene is predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II transcription regulatory region sequence-specific DNA binding activity. Predicted to be involved in several processes, including learning or memory; regulation of DNA-templated transcription; and social behavior. Predicted to act upstream of or within lung alveolus development and negative regulation of transcription by RNA polymerase II. Located in nucleoplasm. Implicated in autosomal dominant intellectual developmental disorder 45.

Here there is some overlap in activity with unrelated tissue types but this gene will be kept.

**MSH5**: Gene is predicted to enable double-stranded DNA binding activity. Predicted to be involved in chiasma assembly. Predicted to act upstream of or within female gamete generation and homologous chromosome pairing at meiosis. Predicted to be located in synaptonemal complex. Predicted to be active in nucleus. Implicated in lung non-small cell carcinoma; primary ovarian insufficiency 13; and spermatogenic failure.

The presence of this gene in our set explains the first odd finding from the gene ontology search presented earlier. This gene also presents a small dilemma: non-small cell cancer is different from small-cell lung cancer. For now this gene is being kept because it is specific to the tissue type, and because it shares risk factors with small-cell lung cancer.

**SDHC**: Predicted to enable heme binding activity. Predicted to be involved in mitochondrial electron transport, succinate to ubiquinone. Predicted to be located in mitochondrial inner membrane. Implicated in Carney-Stratakis syndrome; gastrointestinal stromal tumor; lung non-small cell carcinoma; and paraganglioma.

Here we have a gene which would create false positives with GI stromal tumors and other diseases, so this will be removed from our final predictor set.

**SIM2**: Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II transcription regulatory region sequence-specific DNA binding activity. Predicted to be involved in regulation of transcription by RNA polymerase II. Predicted to act upstream of or within embryonic pattern specification; lung development; and negative regulation of transcription by RNA polymerase II. Located in nuclear body.

This gene can be safely included.

**NAT2:** Predicted to enable arylamine N-acetyltransferase activity. Predicted to be involved in xenobiotic metabolic process. Predicted to be located in cytosol. Implicated in several diseases, including autoimmune disease (multiple); carcinoma (multiple); contact dermatitis (multiple); leukemia (multiple); and lung disease (multiple).

This gene has to be excluded to minimize false positives from other diseases and cancers.

**ZNF157:** Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II cis-regulatory region sequence-specific DNA binding activity. Predicted to be involved in regulation of transcription by RNA polymerase II. Predicted to act upstream of or within lung alveolus development; mammary gland morphogenesis; and regulation of cell fate commitment. Predicted to be active in nucleus.

This gene will be kept in the final set. ‘Cell fate commitment’ can included cell programmed death (apoptosis) which can be triggered by malfunctioning cells.

**STK40**: Predicted to enable ATP binding activity; protein serine kinase activity; and protein serine/threonine kinase activity. Predicted to act upstream of or within several processes, including glycogen metabolic process; lung development; and respiratory system process. Located in cytosol and nucleoplasm.

No problems here, this will be kept in the final gene set.

**NKIRAS2**: Predicted to enable GTPase activating protein binding activity. Predicted to be involved in Ral protein signal transduction. Predicted to act upstream of or within several processes, including lung alveolus development; regulation of signal transduction; and surfactant homeostasis. Predicted to be located in cytoplasm.

No problems with this either. ‘Surfactant homeostatis’ would relate to the function of key compounds keeping lungs in the alveoli lubricated so that the alveoli can expand and contract smoothly during breathing. While disruptions can be non-specific to cancer, this gene will be kept as it is essential for normal lung function.

**SNHG20:** Predicted to be involved in cellular response to fatty acid; gene expression; and macrophage activation. Implicated in lung non-small cell carcinoma.

'Implicated' is ok; again, this gene could be involved in non-small cell rather than small-cell carcinoma, but will be kept.

**LAMA3**: Predicted to enable integrin binding activity. Predicted to be an extracellular matrix structural constituent. Involved in endodermal cell differentiation. Located in endoplasmic reticulum. Implicated in junctional epidermolysis bullosa and lung small cell carcinoma.

*Implicated* in the kind of carcinoma we are interested in, but unfortunately, it's of low predictive value. (JEB is a rare inherited disorder, and uncommon enough that it is safe to keep this gene).

**ASXL1**: Predicted to enable several functions, including nuclear retinoic acid receptor binding activity; peroxisome proliferator activated receptor binding activity; and transcription coactivator activity. Predicted to be involved in several processes, including negative regulation of fat cell differentiation; regulation of intracellular signal transduction; and response to retinoic acid. Predicted to act upstream of or within several processes, including heart morphogenesis; hematopoietic or lymphoid organ development; and lung saccule development. Part of PR-DUB complex. Implicated in SM-AHNMD; acute myeloid leukemia; chronic myelomonocytic leukemia; myelodysplastic syndrome; and myelofibrosis.

We have to remove this one, because it's implicated in other cancers.

**DPPA4:** Predicted to enable chromatin binding activity. Predicted to be involved in system development. Predicted to act upstream of or within lung-associated mesenchyme development. Predicted to be located in nucleoplasm. Predicted to be active in nucleus.

No problems with this gene.

**MAN1A2**: Predicted to enable mannosyl-oligosaccharide 1,2-alpha-mannosidase activity. Predicted to be involved in N-glycan processing. Predicted to act upstream of or within glycoprotein metabolic process; lung alveolus development; and respiratory gaseous exchange by respiratory system. Located in Golgi apparatus.

No problems with this gene either.

**DPPA2:** Predicted to enable chromatin binding activity. Predicted to be involved in system development. Predicted to act upstream of or within several processes, including epigenetic regulation of gene expression; lung-associated mesenchyme development; and positive regulation of stem cell proliferation. Located in nucleoplasm.

Likewise no problems here.

**MAN2A1**: Predicted to enable alpha-mannosidase activity. Predicted to be involved in N-glycan processing. Predicted to act upstream of or within several processes, including lung alveolus development; respiratory gaseous exchange by respiratory system; and retina morphogenesis in camera-type eye. Located in Golgi medial cisterna.

No problems.

**TBX4**: Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II cis-regulatory region sequence-specific DNA binding activity. Involved in embryonic hindlimb morphogenesis; embryonic lung development; and skeletal system morphogenesis. Predicted to be part of chromatin. Predicted to be active in nucleus. Implicated in arthropathy and ischiocoxopodopatellar syndrome.

Comments: ICPPS is a rare genetic disorder of the pelvis and feet, and for that reason we will keep this gene. Arthropathy is more common, so there might be some basis for throwing this gene out, but for now we will keep it.

**ATXN1L**: Predicted to enable DNA binding activity; POZ domain binding activity; and RNA binding activity. Predicted to be involved in learning or memory; regulation of DNA-templated transcription; and social behavior. Predicted to act upstream of or within several processes, including lung alveolus development; negative regulation of transcription by RNA polymerase II; and positive regulation of hematopoietic stem cell proliferation. Located in nucleolus and nucleoplasm.

This gene will be kept (and may explain the last unusual finding in the gene ontology).

**COPA:** Predicted to enable mRNA binding activity. Acts upstream of or within pancreatic juice secretion. Located in cytoplasm; extracellular space; and growth cone. Part of COPI vesicle coat. Implicated in autoimmune interstitial lung, joint, and kidney disease.

One might consider dropping COPA as well as TBX4 if there are some concerns about joint diseases contributing to false positives, but for now we will keep them in the set.

**MIR367**: Predicted to be involved in miRNA-mediated post-transcriptional gene silencing. Predicted to act upstream of or within cellular response to leukemia inhibitory factor and lung development. Predicted to be part of RISC complex.

Unfortunately this one has to be removed, as it is implicated in leukemia.

**CFC1B**: Predicted to enable activin receptor binding activity and nodal binding activity. Predicted to be involved in circulatory system development; nodal signaling pathway; and regionalization. Predicted to act upstream of or within several processes, including heart development; left lung morphogenesis; and spleen development. Predicted to be active in cell surface and extracellular region. Implicated in tetralogy of Fallot and visceral heterotaxy.

Comments: Tetralogy of Fallot and visceral heterotaxy are congenital heart defects; no need to discard this gene.

**NKIRAS1**: Predicted to enable GTPase activating protein binding activity. Predicted to be involved in Ral protein signal transduction. Predicted to act upstream of or within lung alveolus development; regulation of tumor necrosis factor-mediated signaling pathway; and surfactant homeostasis. Located in cytosol and endoplasmic reticulum.

This gene is safe to keep.

**FENDRR**: Predicted to enable core promoter sequence-specific DNA binding activity. Predicted to be involved in in utero embryonic development and lung development. Predicted to act upstream of or within several processes, including chromatin organization; embryonic lung development; and lateral mesoderm development.

This gene will be kept.

*Summary of annotation findings*

We will have to drop the following genes for our final analysis, as they are implicated in other cancers and might reduce specificity: MIR367, ASXL1, SDHC, NAT2.

**V. Model Development, part 2.**

We now have a reduced gene set; however, with such a small number of positives in the test set, it was necessary to check whether our initial results were (un)lucky by doing cross validation. We performed 5-fold CV on the dataset and found a .99 weighted f1 test result, with a standard deviation of 0.01. This meant the approach looks safe.

The next step was model tuning. A grid search using f1\_weighted to evaluate results was run with different values of C (learning rate) and regression penalty (ridge vs lasso). The result was a solver with these parameters: liblinear solver, max iterations = 500, C=0.1, and penalty=l2. Using the optimized model with the final gene set, we obtained the following results, listed in Table 2 and graphed as a confusion matrix in Figure 5:

**Table 2**. Classification Report for Test Data

Sample precision recall f1-score support

Normal 0.89 1.00 0.94 16

Tumor 1.00 0.99 0.99 166

accuracy 0.99 182

macro avg 0.94 0.99 0.97 182

weighted avg 0.99 0.99 0.99 182

A chart of different colored squares

Description automatically generated

**Figure 5.** Final model results: confusion matrix. The final, 20-gene set performed exceptionally well overall.

Given such remarkable performance, it did not seem necessary or worthwhile to spend time trying other binary classifiers. The only flaw here is the 2 false negative tumor samples; there were no false positives, which is excellent given the small number of actual normal samples. The final model parameters and gene targets will be provided in a .txt file in Github along with the code notebook and final report.

**VI. Summary and Conclusions, further work.**

While the model results appear excellent, it would be unwise to jump to conclusions and immediately suppose this sort of test can be commercialized in short order. Unfortunately, there would likely be a long process of validation, during which it is possible that, as ‘normal’ tissue samples are added, the gene list would have to be refined. (Part of validation would be to verify that peripheral blood samples would work in the place of tissue samples from biopsies, etc; in blood samples the concentration of relevant mRNAs could be lower as the tumor-related expression products would be more dispersed.)

Furthermore, ‘normal’ samples might include those from patients with other underlying conditions (e.g., joint arthropathy) which might also result in modifying the gene set, or perhaps even expanding it to a multiclass model. That said, with many more normal observations, we also expect the variance in normal gene activity to narrow, resulting in better predictors. So there is reason to believe this sort of test could be deployed, eventually.

Another consideration here is IP protection. Unfortunately, the US patent office restricts patents that cite gene sequences to three or less; it would be very expensive to patent enough sequences together as a panel for a diagnostic application such as this. With a large enough body of microarray data, one might contemplate building a model which treats the expression values as images, and use image classification to predict normal vs tumor (a feat which might get around the IP issues to some degree). Imagine, for instance, a multiclass microarray image analysis model that could also distinguish other types of disease. In any case, the LUSC database did not provide the data in image format (though it may not be impossible to recover), and many more normal images would have to be added before a model such as this would be feasible.

Finally, some additional work might be done to further subset the genes, perhaps to eliminate the COPA and TBX4 as noted in the report; this might be done in conjunction with some pre-processing automation, so that any further adjustments to the gene set could be rapidly evaluated. As it stands, though, the feasibility of this approach now seems very solid; hopefully, such tests will become more common in the future.