Deep Learning Image Analysis and Liquid Biopsy for Lung Cancer Detection

by Andrew Gao

Abstract

The American Cancer Society projects that over **135,000** Americans **will die** from lung cancer in 2020 (1). It is recognized that the earlier lung cancer is detected, the higher the survival rate and the lower the treatment cost. However, current detection methods, such as human interpreted imaging tests and tissue biopsies, are **inaccurate**, **potentially dangerous**, and **costly** (2). The purpose of this project is to identify lung cancer earlier and more accurately through a blood-based liquid biopsy test using epigenetic signatures of DNA methylation in lung cancer in conjunction with automated CT scan analysis. RNA sequencing was also attempted to augment the other methods.

A novel application of targeted bisulfite padlock probing for DNA methylation sequencing was applied to 43 plasma samples total, some with lung cancer. 8 samples were first tested as a pilot study. Data was analyzed using the Bioconductor software library. Novel RNA sequencing was attempted and multiple experiments were conducted, however, successful RNA sequencing was unattainable. Several deep learning architectures and modifications were tested on various datasets and preprocessing methods for the lung cancer classifying task.

Using Keras, deep learning architecture, and convolutional neural networks, an average accuracy of **93%** and AUC of **0.98** were achieved. RNA sequencing prep did not work but significant improvements were shown in RNA/cDNA quality after modifications. DNA methylation sequencing results were very promising and many statistically significant methylation signature differences were identified among lung cancer and control samples, especially for lung adenocarcinoma in particular.

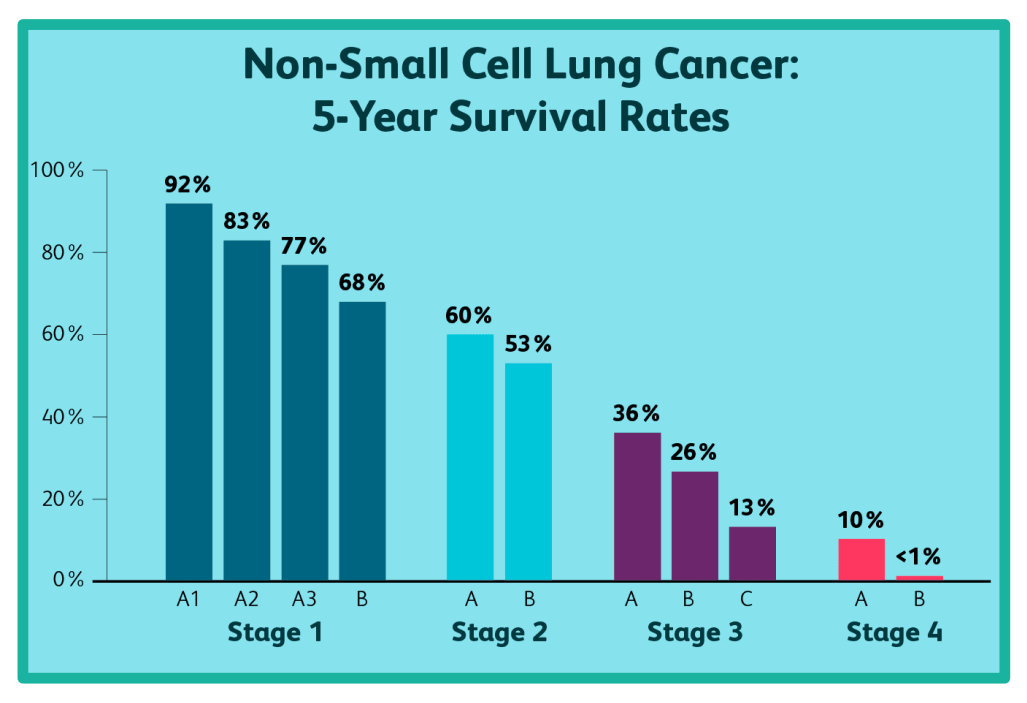


Figure 1:
Data from the American Cancer Society (2017) shows that early cancer treatment is paramount for higher survival rates. At late stages, the survival rate dips below 1% whereas at Stage 1, it can exceed 90%.

Citation: Rathod, Hetal. "Non-Small Cell Lung Cancer Survival Rates Graph," Verywell Health, 1 Nov. 2019, www.verywellhealth.com/thmb/TzvYcshmDJ84spQSe11rclY5XW0=/1500x0/filters:no_upscale():max_bytes(150000):strip_icc():format(webp)/lung-cancer-5yr-chart-04-5b8558d3c9e77c00253444cb.png. Accessed 16 Jan. 2020.

Introduction

Impetus:

I chose to research early lung cancer detection methods due to the **personal impact** it has had on me as well as millions around the world. My grandfather passed away from lung cancer a few years ago and I was devastated. The cancer would have been treatable had it been detected at an earlier stage. This inspired me to pursue cancer research, specifically in the area of early detection.

Cancer:

- Second leading cause of death (1 in 6 deaths) (3)
- 18.1 million cases diagnosed in 2018 (4)
- Lung cancer is one of the most prevalent cancers (4)
- 1/16 Americans will be diagnosed with lung cancer (5)
- Economic cost of lung cancer: 1.16 trillion USD (6)

Early cancer detection is highly correlated to survival. It is imperative that detection methods at early stages improve because current methods are inadequate.

Current Methods:

Tissue Biopsy:

(uses surgery/needle to collect tissue samples from lung)

- Invasive
- Expensive (average cost exceeds $14,000) (7)
- Risky
- Inaccurate (intratumoral heterogeneity) (8)
- Not repeatable
- Only a "snapshot" of the cancer

Human-Analyzed CT Scan:

- Time consuming analysis
- Difficult to diagnose lung nodules (for humans)
- >90% false positive rate for nodules (9)
- 48% stage 1 lung cancer detection rate (9)
- No genomic data, only phenotypic

Alternative Methods:

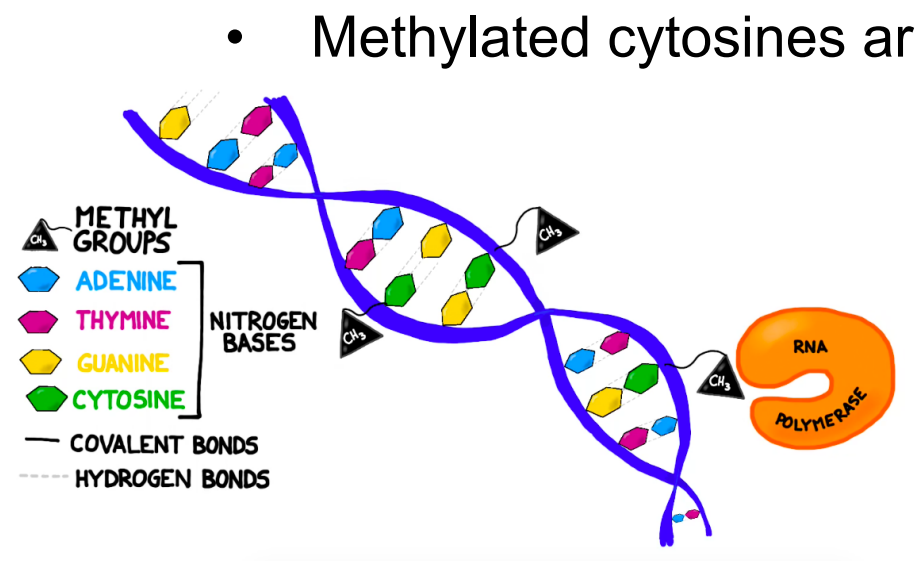
Liquid Biopsy:

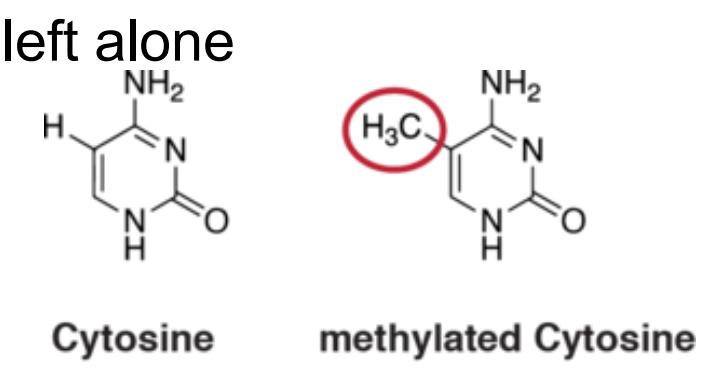
- Uses blood drawn, does not need to be from lung
- Repeatable
- No surgery or invasive procedures required
- Comprehensive picture of disease
- Can have faster result turnaround time (10)
- Accounts for intratumoral heterogeneity

Through apoptosis or necrosis, dying tumor cells release fragmented DNA into the bloodstream. Fragments from tumor cells are known as circulating tumor DNA (ctDNA. Usually, liquid biopsy works through using DNA sequencing.

DNA Methylation:

- When cells become tumor cells, epigenetic changes occur to the DNA: methylation
- Methylation: adding a methyl group (CH3) to cytosines
  - Silences the gene
  - Important in gene expression regulation, cell cycle, cell development, and more (11)
  - Used by cancer to shut off tumor suppressors and other important anti-cancer genes
  - Can be sequenced after sodium bisulfite treatment
    - Unmethylated cytosines are turned into uracil
    - Methylated cytosines are left alone





(**Left**) [BOGObiology]. (2017, October 11). Epigenetics: Nature vs.. [Video File]. Retrieved from https://youtu.be/Q8BMP6HDIco
(**Top**) Mariuswalter / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0) https://commons.wikimedia.org/wiki/File:DNA_methylation.png

Convolutional Neural Networks:

- Neural networks specialized for recognizing patterns in images (at a basic level of understanding)
- Could be trained to differentiate between cancerous lung nodules and benign lung nodules in CT scans
- Faster than humans
- Potentially more accurate than humans

Overall, targeted DNA methylation sequencing and automated deep learning nodule classification could improve the accuracy of lung cancer detection at early stages and also drive down costs across the board.

Hypothesis

If targeted DNA methylation sequencing and deep learning CT scan nodule classification are used, lung cancer can be reliably diagnosed at early stages.

CT Scan Materials

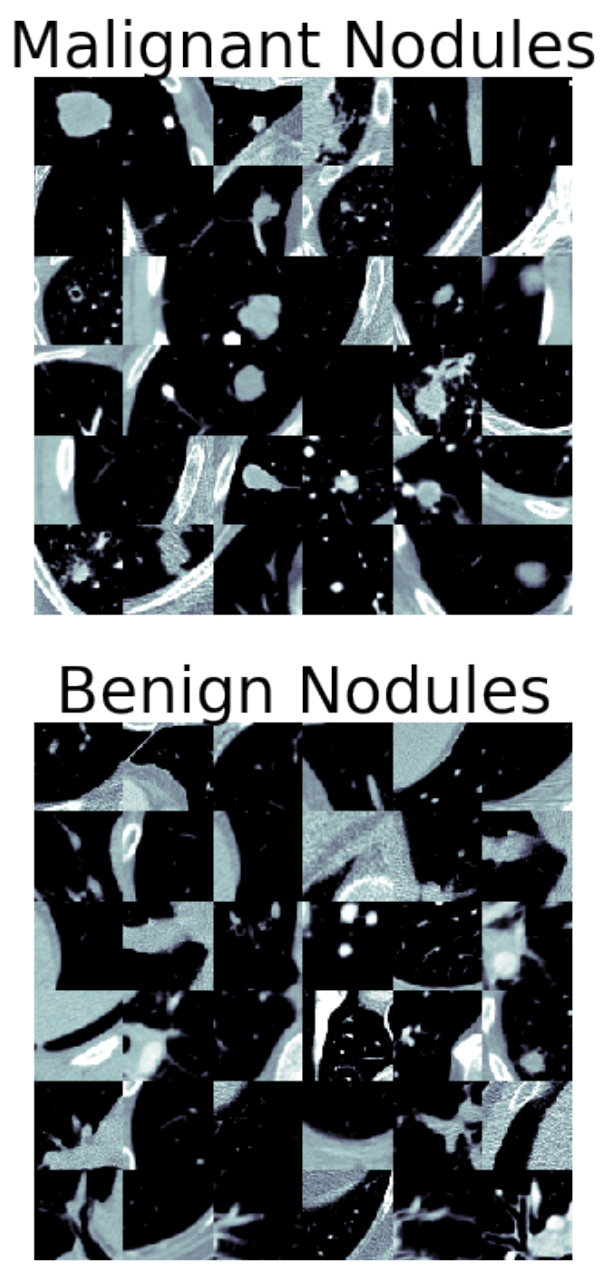
PyCharm Professional 2019
Python 3.7 programming language
Tensorflow
Pytorch
Keras
Densenet169, 201
Nvidia RTX 2070 GPU
Kaggle IDE
Sklearn
Online Nvidia K80 GPU

LUNA, LIDC-IDRI, and derived datasets (all freely available) (12,13,14,15)
Various Python Libraries and Packages
- Numpy
- SciITK
- Pillow

Slicer CT Scan Visualization Software
Source code from Kevin Mader (16)

CT Scan Methodology

An open source (License: Apache 2.0) code for classifying lung nodules was found online. The AUC metric and accuracy were both only 68% which is worse than a radiologist would be. Additionally, the code did not run anymore and had several errors due to deprecation. The code was 2 years old and had not been updated. However, I felt that the code had potential.



Examples of the nodule images

Strengths:

- Low-computational cost (can be run in browser)
  - Uses SqueezeNet (17)
- Relatively balanced dataset
- Accuracy greater than 50%

Weaknesses:

- Does not function anymore
- Accuracy lower than human, a little better than guessing
- Trained on CPU which is slower than using a GPU
- Parameters (epochs, batch size, learning rate, etc.) are default

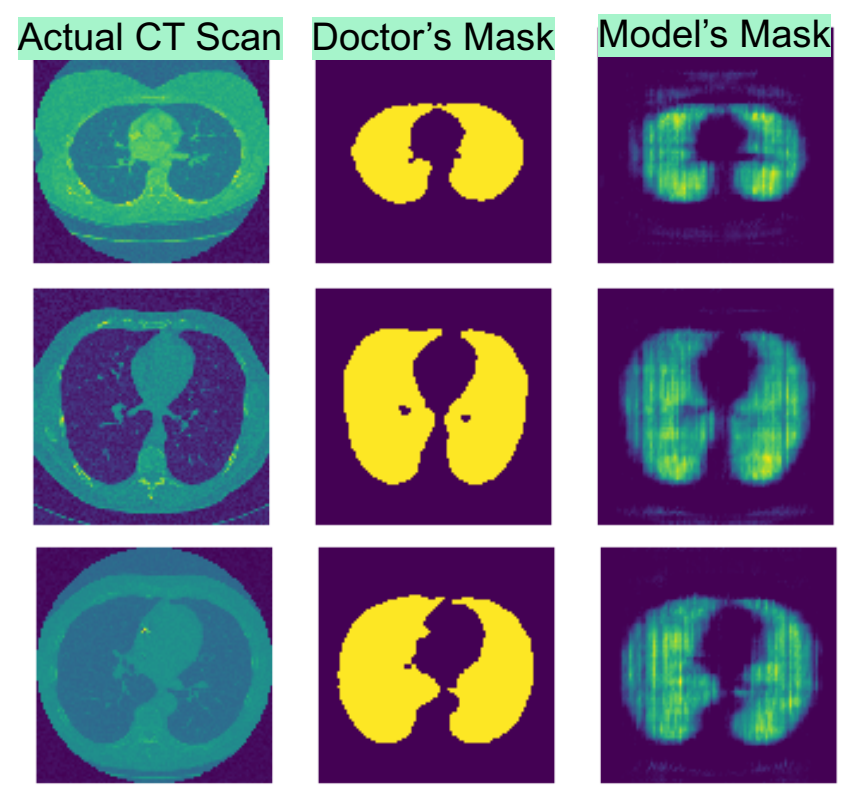
I decided to test how to improve the accuracy of the machine learning model. I conducted several trials, altering a variety of parameters, using a free online GPU provided by Kaggle. After some main parameters were selected, others such as learning rate, convolutional layer architecture, and loss function, were tweaked to improve metrics.

Parameters tested experimentally:

- Epochs
- Batch size
- Training to Test set size ratio

Other changes:

- Convolutional layer architecture
- Loss function
- Learning rate



Lung segmentation results. Each row is one sample and the radiologist and model generated masks, respectively.

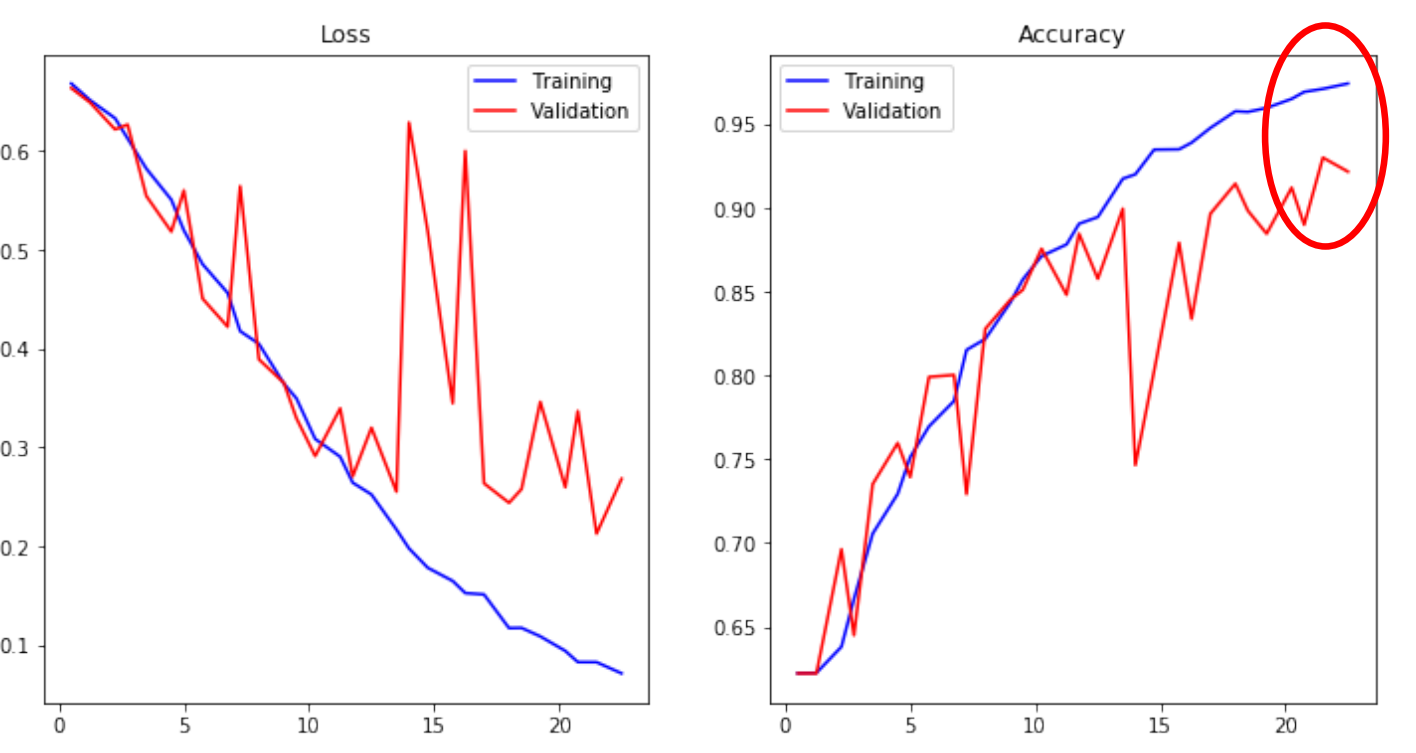
Dataset Information:

- Dataset: Benign: 4165 Malignant: 2526
- The dataset was processed from the LIDC-IDRI dataset by Kevin Mader (18). Each sample consists of a lung nodule image and a value of 0 or 1, depending on if it is benign or malignant.
- The ratio of benign to malignant nodules was 1.65:1.

Lung segmentation was also attempted. It consists of training a machine learning model to segment out the lung from the whole CT scan. This is the first step in using machine learning to diagnose lung cancer because the program needs to receive an image of just the region of interest, the lungs, with the chest cavity stripped away.

CT Scan Results

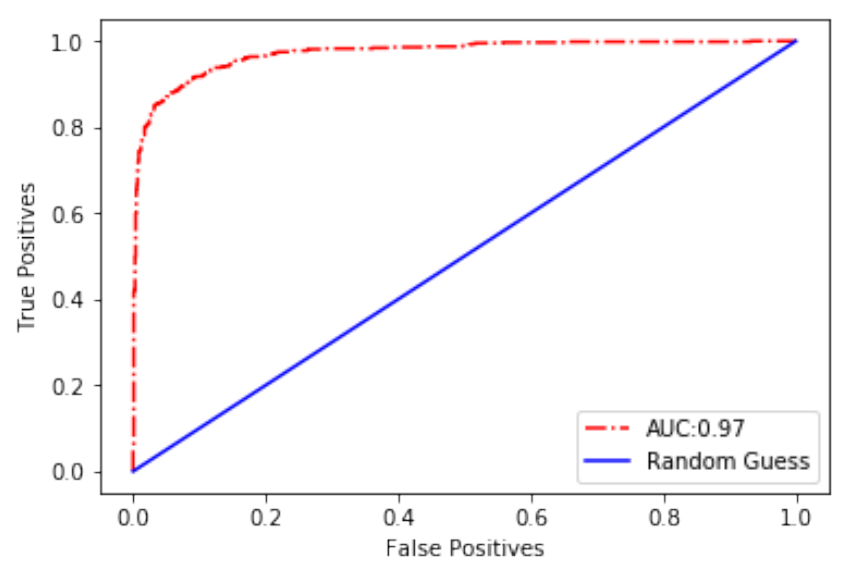
After extensive testing, the average AUC and accuracy without overfitting were 0.98 and 93% respectively.



Parameters:

Learning rate: 0.0012
Train to test ratio: 85:15
Epochs: 27
Loss function: binary cross entropy
Batch Size: 32
2 Fire Modules added

The selection of the final parameters was heavily guided by the results of experimentation, but adjustments were made holistically.



	Precision	Recall	F1 Score	Samples
Benign	94%	95%	94%	1041
Malignant	92%	89%	90%	632
	Accuracy			
Total	93%			1673

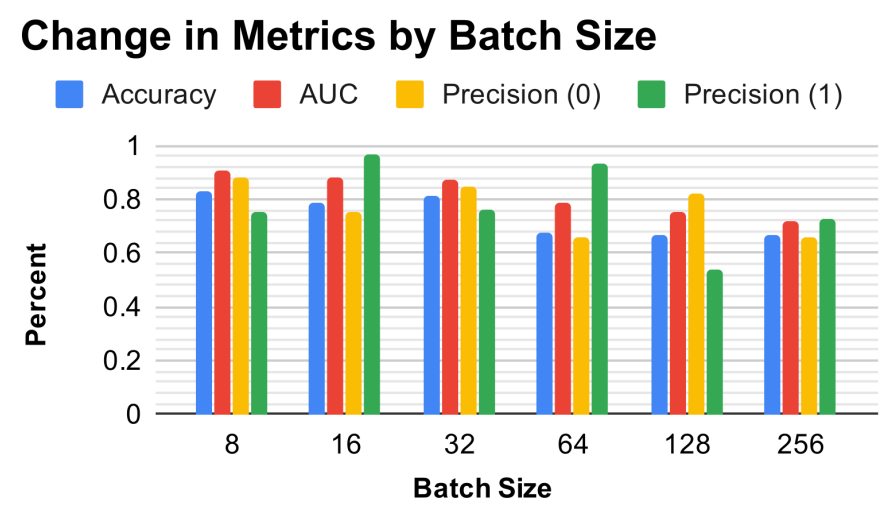
About:

Above are the results for the final model. There are graphs of loss vs. epochs, accuracy vs. epochs, AUC, and a table of the sklearn classification metrics.
Below are the experimental data after altering batch size, epochs, and the training to test size ratio.
A (0) or (1) in the legend indicates the metric from the "perspective" of the benign classification and malignant classification, respectively. For example, Precision (0) indicates the precision of the model for classifying benign nodules.
To the right is the confusion matrix for one of the runs of the final model.

	True Positive	True Negative
Predicted Positive	581	51
Predicted Negative	62	979

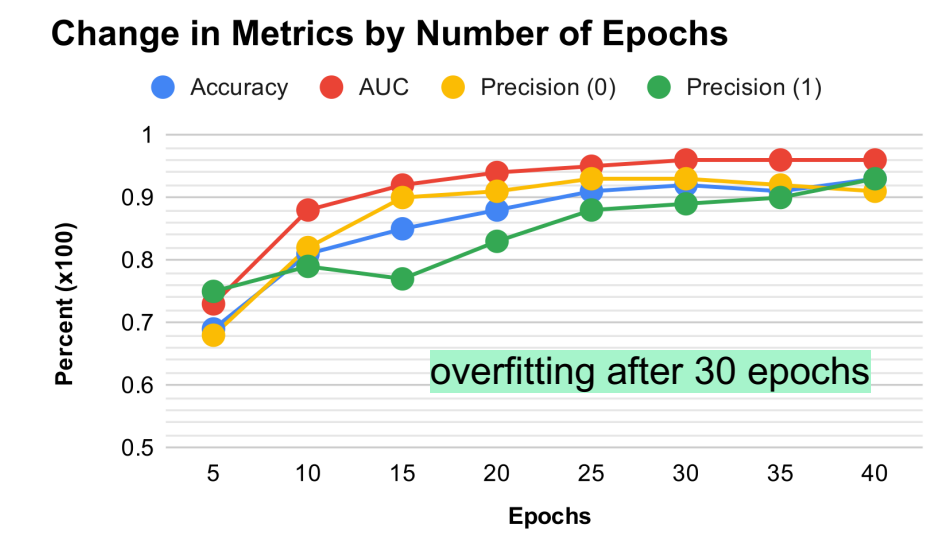
Batch Size:

Learning rate = 0.0001
Loss = categorical cross entropy
Train size = 0.75 of total
Epochs = 10



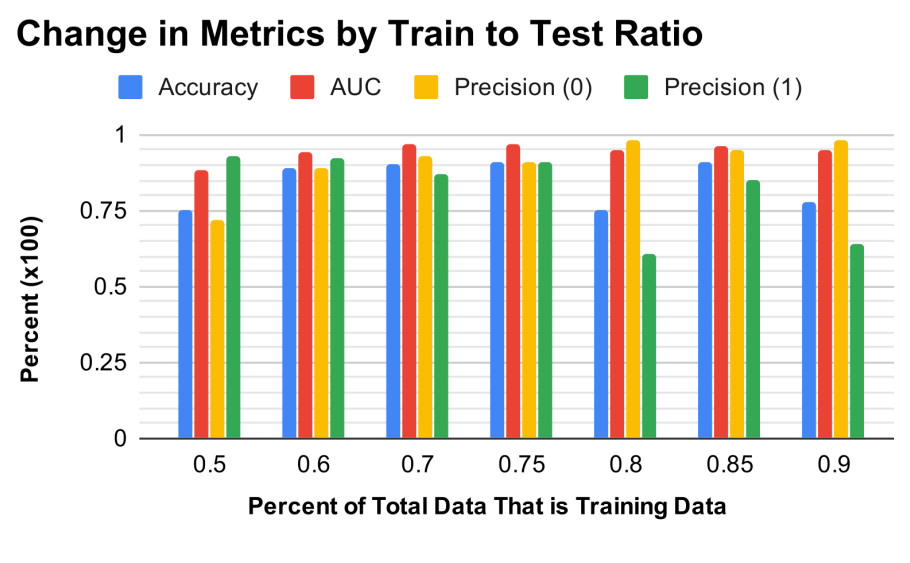
Epochs:

Learning rate = 0.0001
Loss = categorical cross entropy
Train size = 0.75 of total
Batch Size = 32



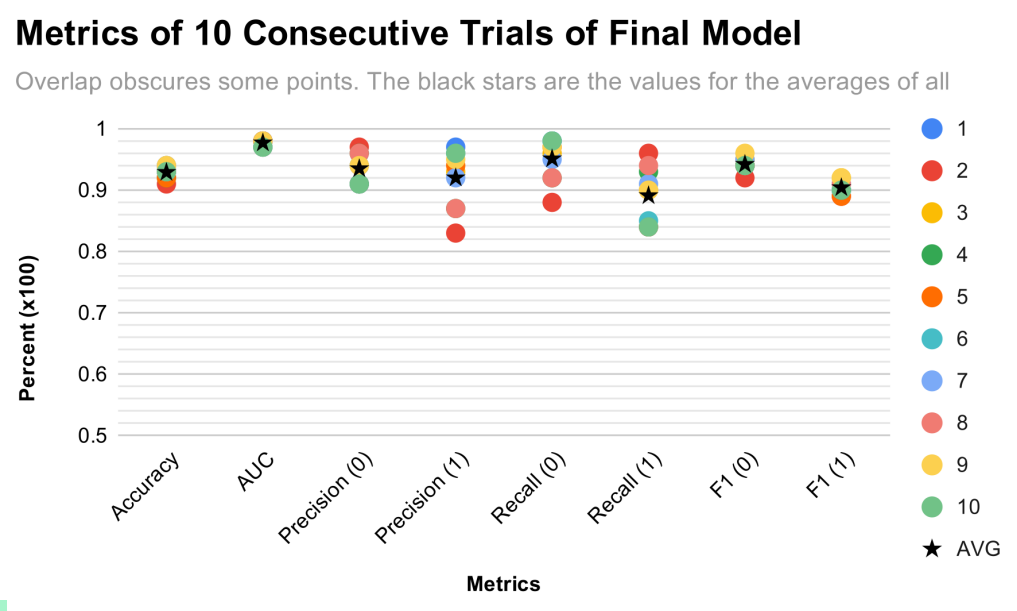
Training to Test Size Ratio:

Learning rate = 0.0001
Loss = categorical cross entropy
Train size = 0.75 of total
Epochs = 10



Final Model Testing:

Final model was tested 10 times to get a more accurate measure of performance.
AUC, F1, and Accuracy were very stable, while Precision and Recall were more varied.



Lung Segmentation:

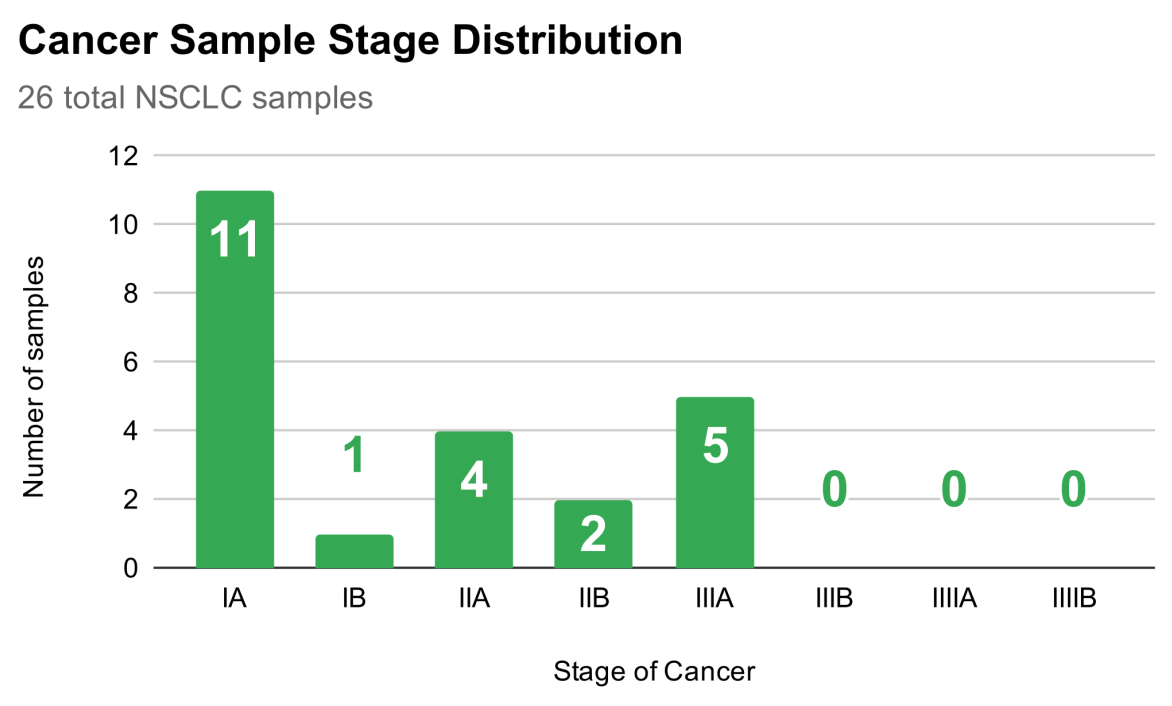
Validation Accuracy: 93.5%
Training Accuracy: 91.9%

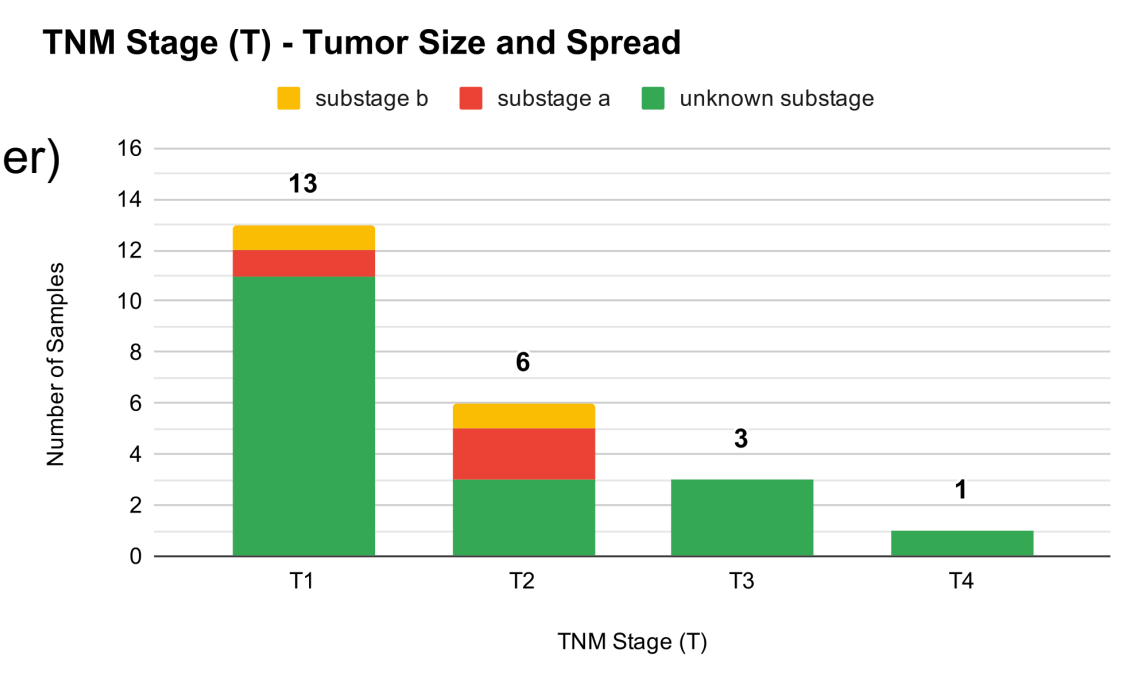
A sample result is shown to the left. The rightmost image is the mask that was generated by the machine learning model.

Liquid Biopsy Materials

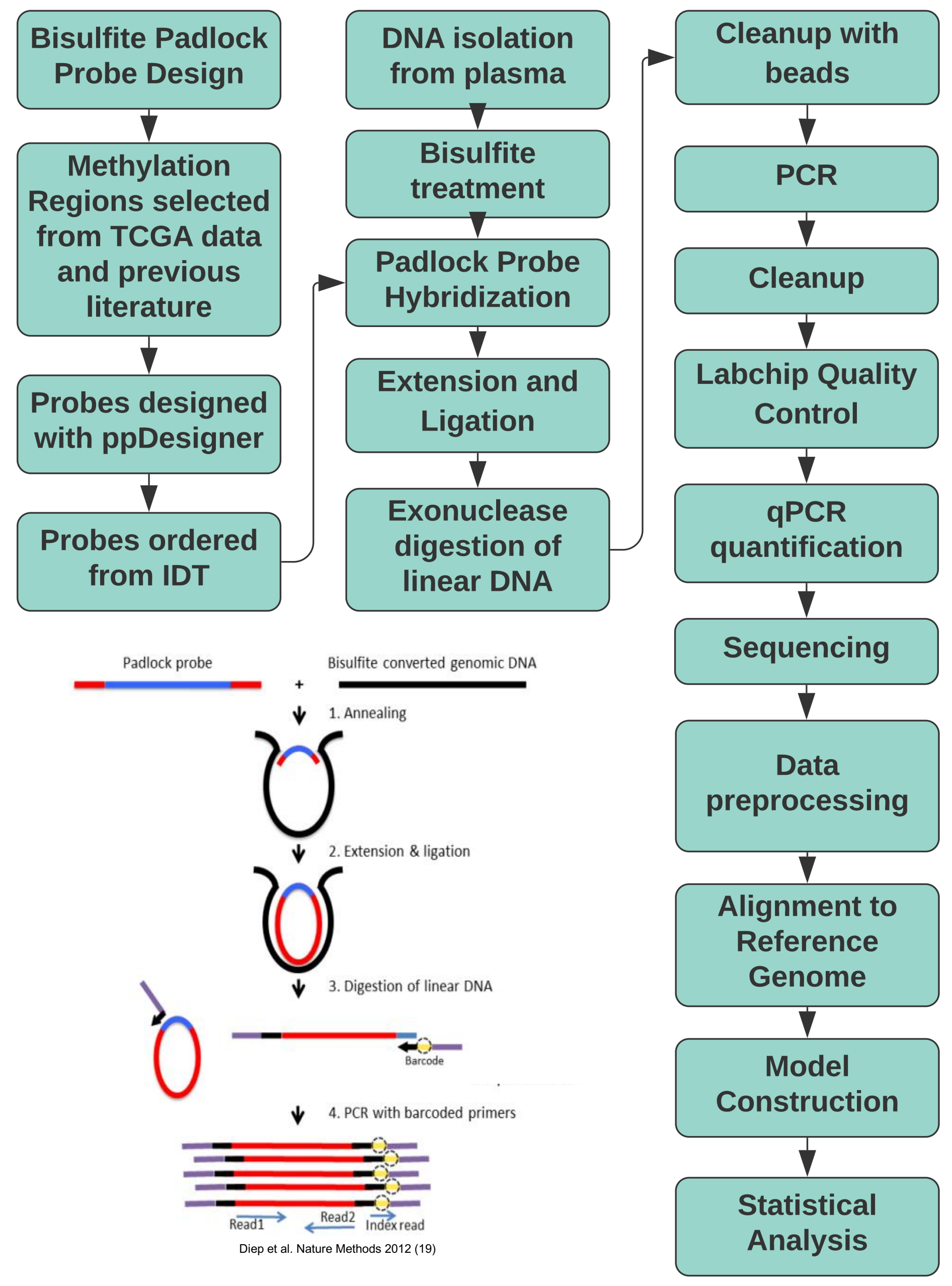
Computational:
BWA-meth
Star
Perl
Bowtie 2
ppDesigner
Bioconductor
R Studio
Python
UCSC Genome Browser

Wet Lab:
NSCLC samples: 23
Normal samples: 20
2000 Bisulfite Padlock Probes (IDT)
Bisulfite Conversion Kit (Thermofisher)
Plasma DNA extraction kit (Qiagen)
AmpureXP beads
NextSeq
LabChip GX
Qubit
Thermocycler
Centrifuge
Full list of other reagents in notebook





Liquid Biopsy Methodology


Padlock probe
Bisulfite converted genomic DNA
1. Annealing
2. Extension & ligation
3. Digestion of linear DNA
Barcode
4. PCR with barcoded primers
Read1
Read2
Index read

Bisulfite Padlock Probe Design → Methylation Regions selected from TCGA data and previous literature → Probes designed with ppDesigner → Probes ordered from IDT → Padlock Probe Hybridization

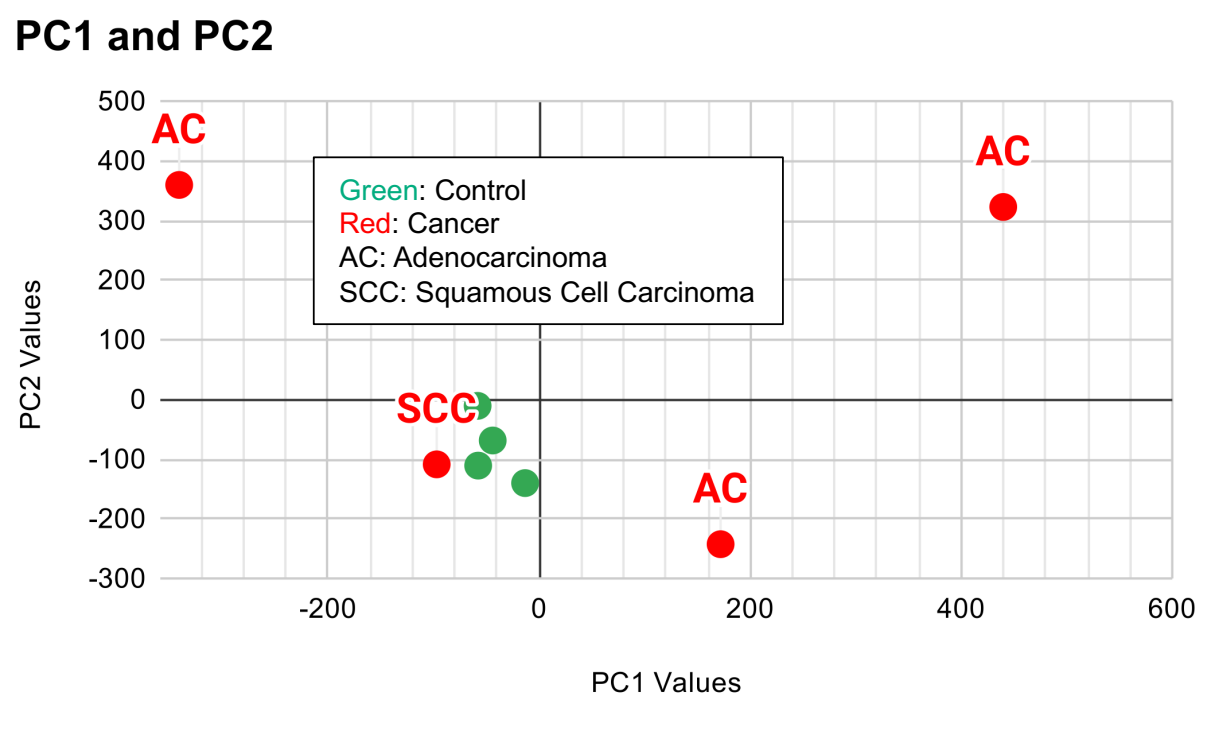
DNA isolation from plasma → Bisulfite treatment → Padlock Probe Hybridization → Extension and Ligation → Exonuclease digestion of linear DNA → Cleanup with beads

Cleanup with beads → PCR → Cleanup → Labchip Quality Control → qPCR quantification → Sequencing → Data preprocessing → Alignment to Reference Genome → Model Construction → Statistical Analysis

Diep et al. Nature Methods 2012 (19)

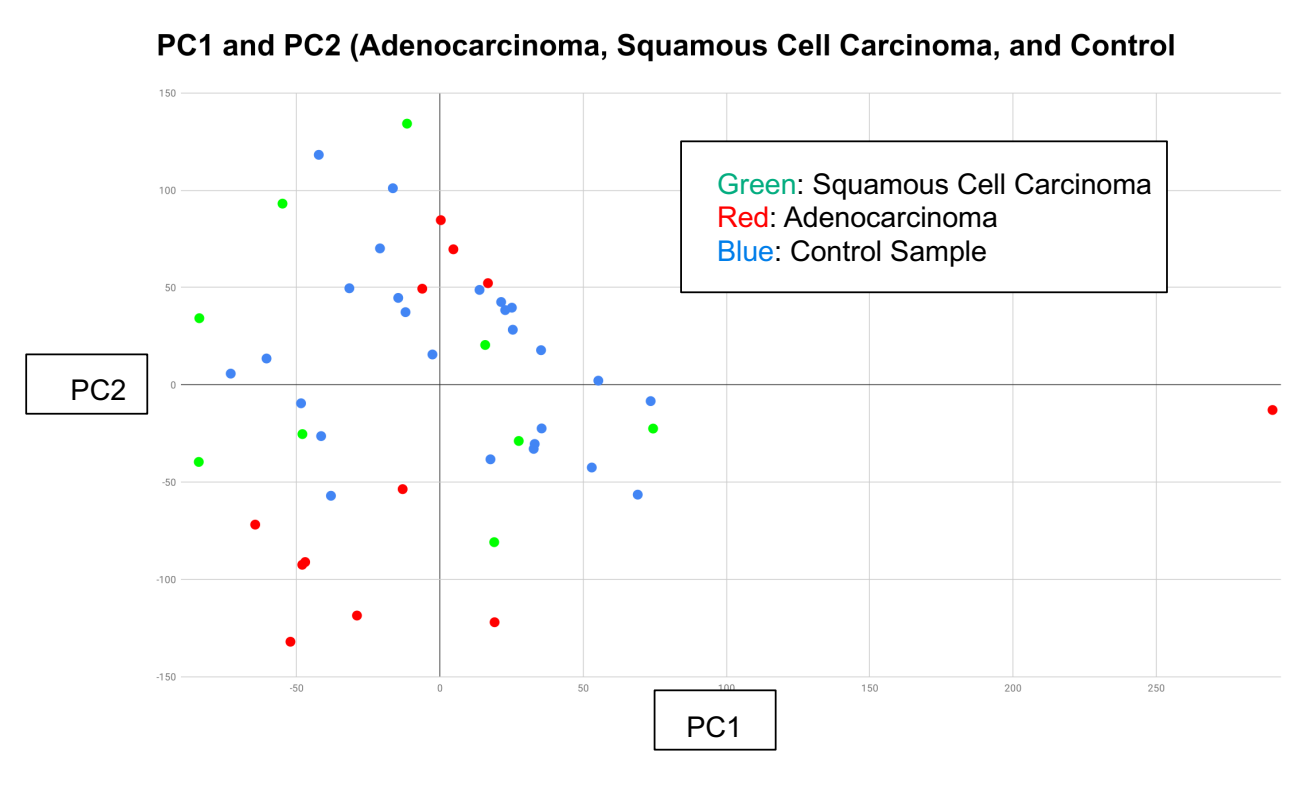
Liquid Biopsy Results

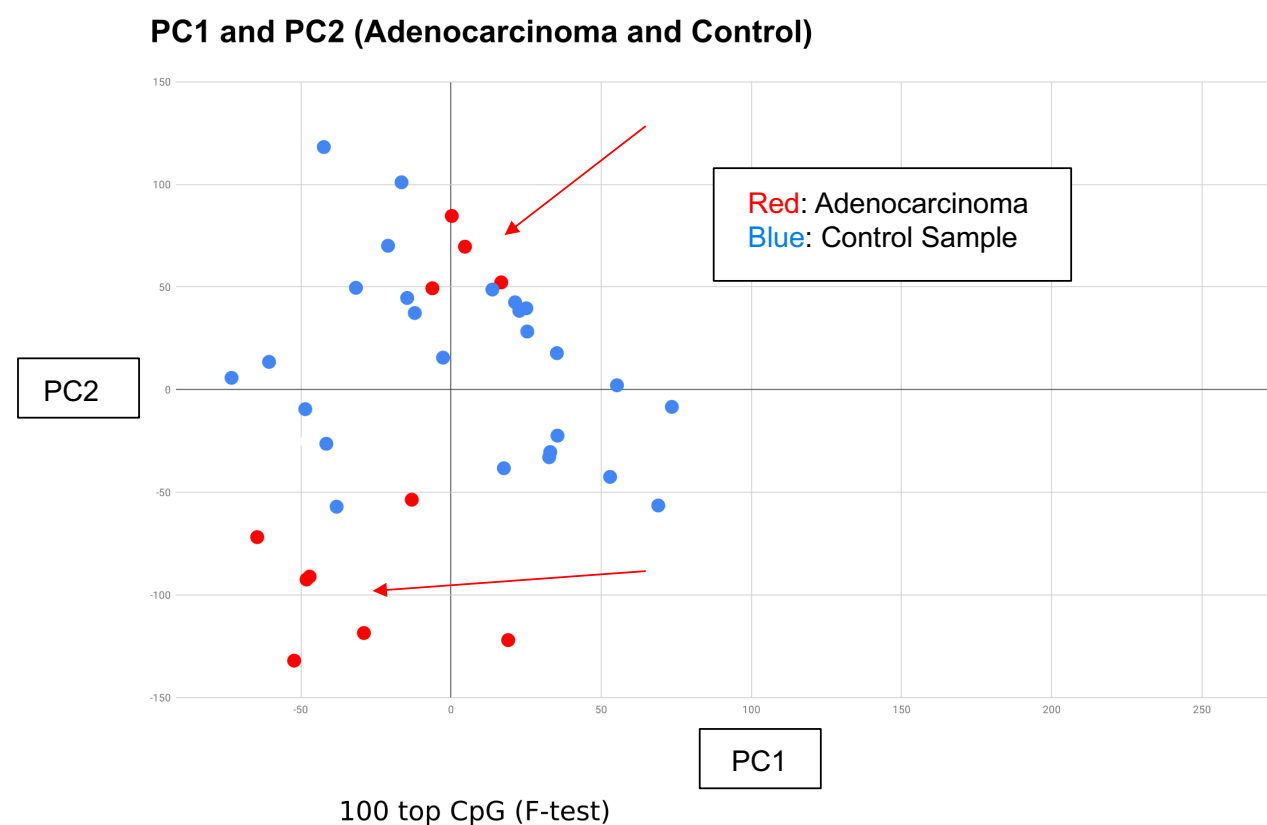
Pilot Study:



I first tested methylation sequencing on 4 lung cancer and 4 control samples. The results are shown to the left, plotted with Principal Component Analysis.
Except for one, the cancer samples are distinctly isolated from the control samples.
(explained in Conclusions)

Main Study:



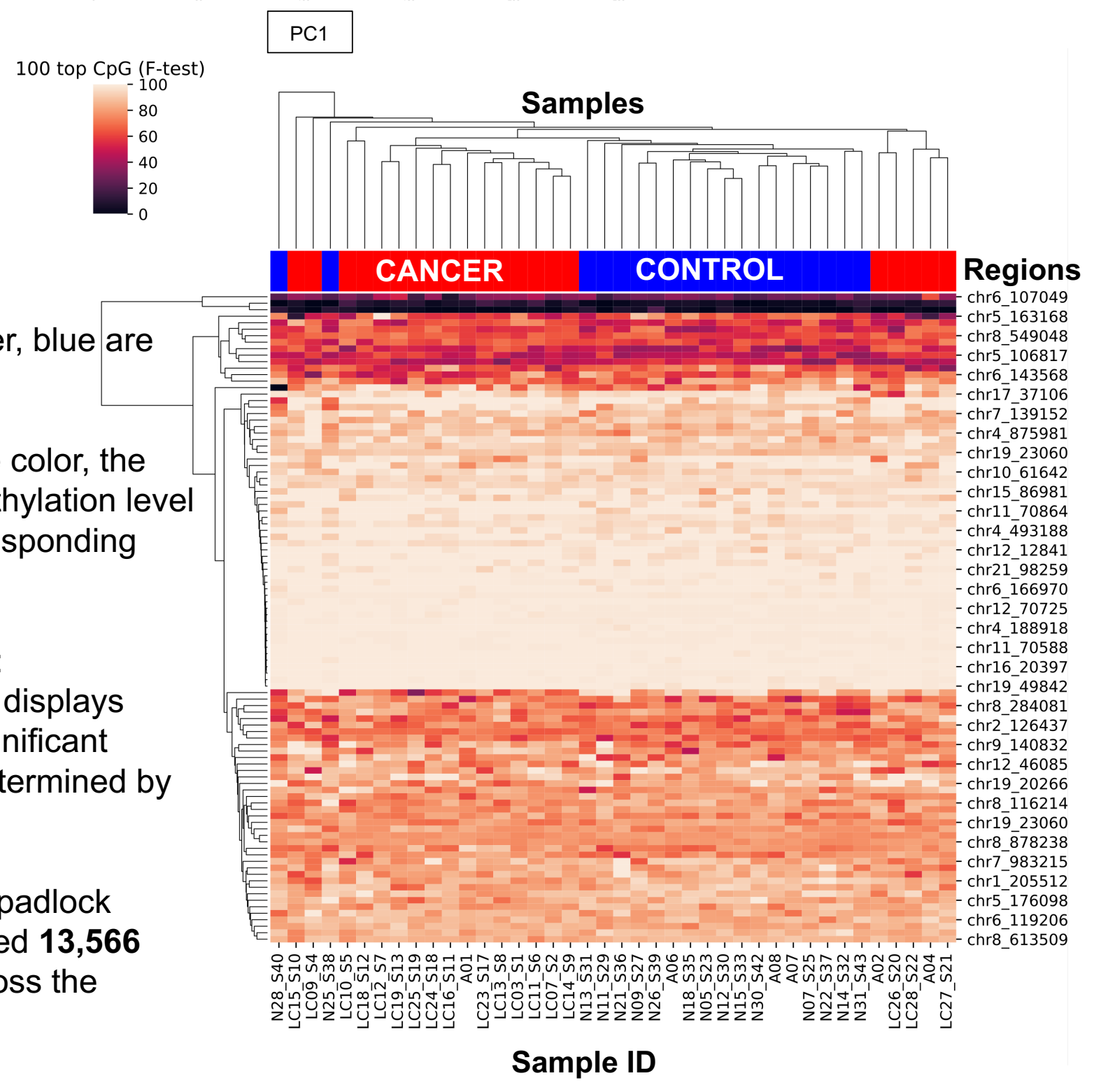


Principal Component Analysis was used again to graph the results of all 43 samples.

There is little separation between squamous cell carcinoma and control samples, similar to in the pilot experiment.

However, when graphed separately, a clear separation and clustering between adenocarcinoma and control samples appears.

Heatmap:



Red are cancer, blue are control.

The lighter the color, the higher the methylation level is for the corresponding region.

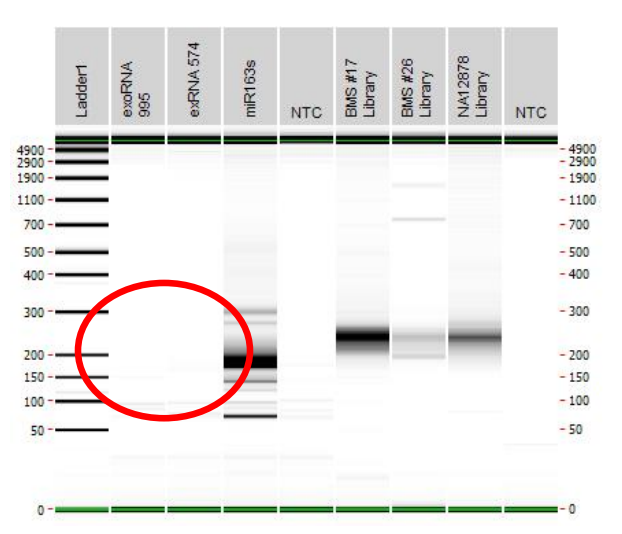
Statistical test:
Heatmap only displays statistically significant differences determined by the F-test.

Overall 2,000 padlock probes captured **13,566** CpG sites across the genome.

RNA Sequencing Attempts

RNA sequencing via cDNA synthesis was attempted several times with modifications, however it was not successful.

S574	2 ng/ul
S578	2 ng/ul
S995	too low
S996	too low



Analysis

CT Scan Analysis:

The CT scan analysis had impressive results: 93% accuracy and 0.97 AUC.
Some important considerations are:

- The dataset was imbalanced by 1.65:1 benign to malignant.
  - Little classification bias found
- Dataset only had around 7,000
  - Generally around 1,000 per class is the minimum

Graphs of log loss and accuracy over epochs indicate that there was little overfitting. The **statistical analysis** of precision, recall, and F1 did not indicate any concerning bias or overfitting.

Confusion matrix values for one run of the final model.

Sensitivity	Specificity	False Positive	False Negative	F1 Score	Accuracy	Precision
0.90	0.95	0.05	0.09	0.91	0.93	0.92

Average classification metrics for 10 runs of the final model

Accuracy	AUC	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1 (0)	F1 (1)
0.93	0.98	0.94	0.92	0.95	0.89	0.94	0.90

Lung segmentation also performed well and there was little evidence of overfitting. However, much more work is required.

Liquid Biopsy:

- Pilot study revealed that squamous cell carcinoma has very different methylation signatures than adenocarcinoma
  - The adenocarcinoma samples were far away on the PCA graph while the squamous cell carcinoma sample was clustered near the control samples
- There appear to be methylation differences depending on the location of the tumor. Tumors in the upper lobes displayed significantly different methylation levels.
- Control samples are "sandwiched" between tumor samples
  - This is good for classification
- Adenocarcinoma is separable from control with PCA
- Squamous cell carcinoma is difficult to separate from control with PCA
- There are 2 outlier control samples in the heatmap
- Heatmap shows that there are some regions on the genome that are differentially expressed in lung cancer
- Heatmap only shows **statistically significant** differentially methylated regions using the **F-test** with a critical value of 2.1242.

Strengths of the experimental design:

- Unlike other studies, almost every sample used for sequencing is early stage
- Even distribution of adenocarcinoma and squamous cell carcinoma
- Minimized batch effect and confounding due to balanced cancer and normal control sample order

Conclusions

The results indicate that both CT scan analysis with convolutional neural networks and liquid biopsy through targeted DNA methylation sequencing have potential for early lung cancer diagnoses. The use case to combine the two very different approaches is to first use liquid biopsy to screen for potential lung cancer patients and then confirm the diagnosis with CT scanning.

The nodule classification program had higher accuracy than radiologists, based on the reports of several studies. The dataset had nearly 7,000 samples which should be decent for generalizing and overfitting was not evident.

Several differentially methylated genomic regions were identified, as shown in the heatmap, and Principal Component Analysis (PCA) revealed that adenocarcinoma and control samples are separable and bounded. Out of over 13,000 methylation sites targeted, I identified the top 100 most significantly differentially methylated sites.

Targeted DNA methylation sequencing appears to have greater success for distinguishing between lung adenocarcinoma and normal controls than for distinguishing between squamous cell carcinoma and normal controls.

One unexpected finding I came to is that although they are subtypes of non-small cell lung cancer, adenocarcinoma and squamous cell carcinoma have very different epigenetic landscapes and methylation biomarkers should be identified for each separately. Another finding Is that methylation can vary based on tumor location in the lung.

Thus, I accept my hypothesis.

Implications

- More accurate early stage lung cancer detection
- Non-invasive and cost effective cancer detection
  - Application to other cancers
- Earlier diagnosis = higher survival rate
  - Hundreds of thousands of lives saved
- Applied to asthma diagnosis with neural networks using airway segmentation in CT scans

Future Research

- Find matching samples with both CT scans and blood
- Comparison of methylation changes at different stages
- Larger sample size
- Improve RNA sequencing method (attempted)
- Explore differences between upper and lower lobe tumors
- Explore differences between adenocarcinoma and squamous cell carcinoma