Cell-Free DNA methylation as a biomarker of malignant vs. benign lung nodules: Cross-validation of 2022 and 2023 patient cohorts

Siri Dandu (SEAS 2026)[1]; Prishna Martinez (COL 2025)[1]; Jacob E. Till, MD, PhD[1]; Zachariya Yazdani[1]; Melinda Yin, PhD[1]; Jeffrey Thompson, MD[2]; Erica L. Carpenter, MBA, PhD[1]

[1]Division of Hematology-Oncology, Department of Medicine, Perelman School of Medicine, University of Pennsylvania; [2]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, Perelman School of Medicine, University of Pennsylvania;

Background

- A lung nodule is a rounded lesion which is measured to be up to 3 cm in size. [1]
- Screening with Low-dose CT (LDCT) scans has shown to be effective in detecting potentially malignant lung nodules (≤ 3 cm diameter) and decreasing lung cancer mortality [1]
- Small nodule classification based on morphology alone can be difficult, leading to an "indeterminate" diagnosis [1]
- LDCT scans can lend false-positive results for indeterminate nodules and repeat screening may be required, increasing radiation exposure risk [17]
- Tissue biopsy is the current standard for molecular detection, however bronchoscopy is an invasive procedure and re-biopsies are not always possible [17]

cfDNA + ctDNA

- Cell-free DNA (cfDNA) can be found in the blood of both healthy individuals and patients with cancer, and is often found in higher concentrations in patients with cancer [20]
- A portion of cfDNA derived from circulating tumor cells in patients with cancer is known as circulating tumor DNA (ctDNA) [20]
- cfDNA can serve as a non-invasive clinical tool to dectect cancer [20]

cfDNA Methylation Analysis in Patients with Lung Nodules:

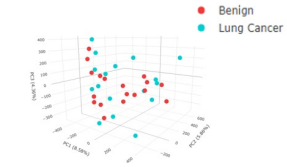
- CpG Methylation is the addition of a methyl group to the cytosine in a cytosine-guanine nucleotide sequence. This methylation tend to "silence" expression of a gene.
- Methylation signatures may help detect malignant versus benign lung tumors. Unmethylated CpG sites can be identified after treatment with the APOBEC enzyme which converts CpG to CpU, allowing us to determine which regions are methylated, a change detected by sequencing.
- Hypermethylatied sites at tumor suppressor gene promoters could indicate malignancy because they "silence" the mechanisms that are intended to suppress rapid growth of tumorous cells.
- cfDNA is a combination of extra-cellular DNA from various cell types. We can determine cellular contribution by referencing unique methylation signatures of individual cell types as corresponding to the amount of DNA in our sample: this is deconvolution.

Methods

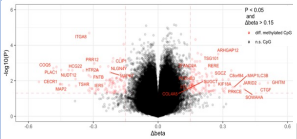
- For Cohort 2023, the cfDNA was extracted from plasma which was extracted from the blood (IRB protocol #824357) of 35 patients (n = 18 benign, n = 17 malignant).
- We used the QIAamp MinElute ccfDNA Mini Kit (#55204) to extract Cell-free DNA from the plasma and quantified that amount with a Qubit Fluorometric Quantification kit.
- We used the NEBNext® Enzymatic Methyl-Seq Kit (#E7120) to convert and amplify the extracted cfDNA (10 ng).
- Meythylation profiling on 100-500 ng of DNA was done at the University of Minnesota Genomics Center using the MethylationEPIC BeadChip (Infinium) 900K+ EPIC v2.0 Array.
- The Sensible Step-wise Analysis of DNA Methylation BeadChips (SeSAMe) package was used to conduct methylation data analysis in R.[7]
- A similar project was conducted in 2022, with a cohort of n=24 (n= 10 benign, n = 14 malignant). However, a EPIC v1.0 Array was used for methylation profiling. Supervised clustering based on differentially methylated regions (DMRs) based on that cohort were tested on this cohort and vice versa.

Results

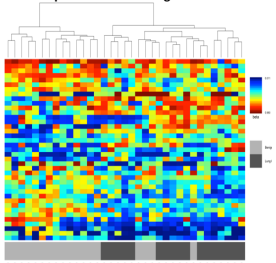
A. Unsupervised Principal Component Analysis of Cohort 2023



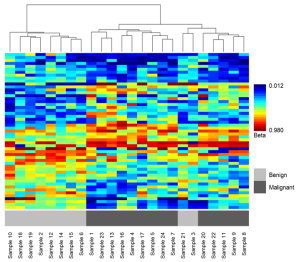
B. Volcano Plot of 2023 Cohort (189 DMRs)



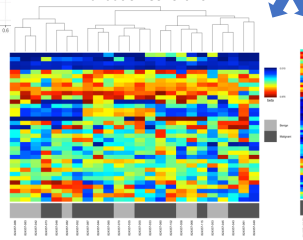
C. Supervised Clustering of Cohort 2023



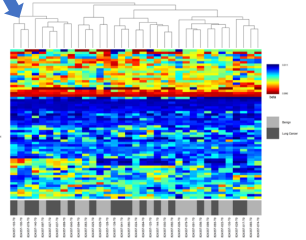
D. Supervised Clustering of Cohort 2022



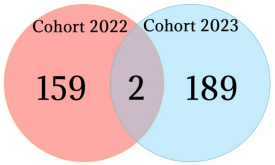
E. Validation Cohort 2022



F. Validation Cohort 2023



G. Venn Diagram



A. Top 3 principal components of the autosomal methylation beta values from the 2023 Cohort . There is no clear clustering based on malignant and benign classification.

B. Volcano plot showing 189 differentially methylated regions (DMRs) between malignant and benign patients determined by a linear model. Each of the loci had a difference in beta-value by at least 0.15, and an unadjusted p-value of less than 0.05.

C. Heatmap of the supervised hierarchical Euclidean clustering based on the 189 DMRs from the 2023 Cohort.

D. Supervised heatmap based on the 159 DMRs from the 2022 Cohort. These DMRs were found with similar methods as the 2023 Cohort.

E. Supervised heatmap of the 2022 Cohort clustered based on the 189 DMRs found in the 2023 Cohort.

F. Supervised heatmap of the 2023 Cohort clustered based on the 159 DMRs found in the 2022 Cohort.

G. Venn diagram of the overlap between the DMRs of the 2023 Cohort and 2022 Cohort. There were only 2 similar DMRs between the two cohorts.

Literature Review

Author (Year)	Genes	Test Set (n)	AUC for Test Set (95% CI)	Summary
Chen, et al. (2020)[8]	CDO1, SOX17, HOXA7	246	0.88 (0.84-0.93)	• 8 lung cancer specific genes tested, best performing panel identified • 246 patients with nodules, (163 M, 83 B)
Hulbert, et al. (2016)[9]	CDO1, SOX17, TAC1	210	0.77 (0.68-0.86)	• Tested 6 cancer-specific genes from The Cancer Genome Atlas • 210 patients with nodules (150 stages I/IIA, 60 B)
Qi, et al. (2021)[10]	Top 300 DMRs	97	0.96 (0.960-.97)	• Test utilized top 300 differentially methylated regions (DMRs) from whole genome cfMeDIP-seq • 7 patients without nodules, 23 benign nodules, 35 malignant nodules (≤3 cm), 32 tumors (>3 cm)
Wielscher, et al. (2015)[11]	HOXD10, PAX9, PTPRN2, STAG3	46	0.85 (0.72-0.95)	• Illumina methylation array identified DMRs of interest, 64 passed QC and were used to generate 4-gene model in a 204-patient cohort • 4-gene model was tested in a 46-patient cohort (23 M, 23 healthy)
Huang, et al. (2020)[12]	SHOX2, PTGER4	140, 30	0.86 (0.80-0.92)	• Methylation of SHOX2 and PTGER4 shown to be biomarkers of lung cancer in literature • 140 patients with nodules in first set (104 M, 36 B) • 30 patients in validation set (19 M, 11 B)
Liang, et al. (2019)[13]	9 DMRs	66	0.82 (0.70-0.93)	• Tissue-derived cancer-specific methylation markers from 230 samples, 9 markers selected with training set of 66 plasma samples • Independent test set of another 66 plasma samples (39 M, 27 B)
Gao, et al. (2015)[14]	APC, RASSF1A	89	0.81 (N/A)	• Tumor suppressor genes which are commonly hypermethylated in cancer patients • 89 patients with nodules (58 M, 31 B)
Vrba, et al. (2020)[15]	10 Marker Set	65	0.96 (0.91-1.0)	• MIR129-2, LINC01158, CCDC181, PRKCB, TBR1, ZNF781, MARCH11, VWC2, SLC9A3, HOXA7 • Subset best for NSCLC selected from 1,250 biomarkers discovered *in silico* • designed to detect 10 carcinoma types Panel tested on 65 patients (18 M, 47 healthy)

Discussion and Future Directions

- The supervised clustering of the DMRs for both the 2022 Cohort and 2023 Cohort worked well to separate malignant vs. benign patients. However, when cross-validating the 2022 Cohort data with the DMRs found for the 2023 Cohort, there is no clear separation established between malignant and benign patients. Similarly, when the 2023 Cohort was clustered using the DMRs found in the 2022 Cohort, there is also no clear separation between malignant and benign patients. This suggests that our 2022 and 2023 models might be overfit and not generalizable to new data.
- There were only 2 similar CpG sites between 2022 Cohort and the 2023 Cohort for differentially methylated regions. The EPIC v2.0 Array was used for the methylation profiling of the 2023 Cohort, while the EPIC v1.0 Array was used for the 2022 Cohort. Translatability of EPIC V1 and EPIC V2 data is still an active area of research and harmonization in a non-trivial task (CITATION).
- P-values, rather than adjusted p-values, were used for the differential methylation analysis because the size of the cohorts were not large enough for adjusted p-values to be significant.
- In the future, it could be beneficial to conduct differential methylation analysis on a combination of the two cohorts of patients in order to increase the sample size.

Acknowledgements

We would like to thank the entire Carpenter Lab for helping and supporting us through our research, especially Zach for extensively guiding us through the methylation data analysis and Dr. Till and Dr. Yin for helping us understand this project. We also wish to express our sincere gratitude to Dr. Carpenter and Dr. Thompson for the opportunity to work in their lab. Funding was provided from CURF through the Penn Undergraduate Research Mentoring Program.

References