Specific Aims

This proposal aims to identify unique mutational signatures, evolutionary patterns, and predictive biomarkers for treatment response and resistance in Black Americans with Non-Small Cell Lung Cancer (NSCLC) living in dense urban areas. Metastasis, the spread of cancer cells from the primary tumor to spatially separated sites, is the leading cause of cancer-related death. Understanding the evolutionary dynamics that lead to metastasis is vital for developing interventions to prevent and treat advanced stages of cancer. While significant progress has been made in understanding the early stages of tumor evolution, the genetic and evolutionary underpinnings of metastatic cancer remain less explored. This knowledge gap is particularly evident in the historically understudied Black American population.

By focusing on Black Americans, we seek to uncover molecular mechanisms of cancer that may be distinct from those observed in White individuals living in similar urban environments but not exposed to the same environmental and socioeconomic disadvantages. Black Americans often face higher exposure to air pollution and other environmental carcinogens due to systemic inequities in housing and urban planning. These factors, combined with unique genetic predispositions, may drive different mutational processes and cancer evolution trajectories.

This research is crucial for understanding the underlying mechanisms of NSCLC in Black Americans, which can lead to more personalized and effective treatment strategies. Furthermore, the findings from this study could inform public health policies aimed at reducing environmental risks and promoting equitable housing and environmental laws. Through innovative methodologies, we aim to narrow the gap in cancer research and enhance clinical outcomes for this underrepresented group.

This proposal builds on the pioneering work of leveraging whole genome sequencing to infer the evolutionary history of cancer. Advanced computational tools, such as the Gain Route Identification and Timing in Cancer (GRITIC) method, have been developed to time complex copy number gains and map the clonal evolution of tumors. These tools have provided critical insights into the subclonal architecture and evolutionary trajectories of various cancers, revealing pervasive intra-tumor heterogeneity and ordered paths of primary tumor evolution.

This project aims to revolutionize our understanding of NSCLC progression through cutting-edge advancements in modeling cancer evolution. By fulfilling these aims, the initiative is poised to bridge critical knowledge gaps in oncology, providing new analytical tools that map the evolution of cancer with unmatched precision. These advancements are expected to enhance the timing and specificity of treatment interventions and to lay the groundwork for tailored therapeutic strategies that are finely adjusted to the genetic and environmental profiles of individual patients, with special attention to the unique needs of Black communities. Ultimately, this research will be used to significantly improve early cancer detection, thereby transforming patient care and outcomes.

Aim 1: Preprocess raw data from the Black Americans with NSCLC cohort *Hypothesis: Comprehensive data preprocessing will transform the raw data from the Black Americans with NSCLC cohort into a high-quality dataset suitable for detailed mutational and evolutionary analysis.*

- Perform data cleaning and preprocessing by removing low-quality sequences, aligning the reads to the reference genome, and filtering data for further analysis with `GRITIC`
- Normalize the data to account for batch effects and other technical variances
- Detect and classify mutations (single nucleotide variants, insertions, deletions, etc.) using established bioinformatics pipelines to prepare the dataset for subsequent evolutionary analysis.

Aim 2: Conduct evolutionary analysis to establish a baseline for subsequent comparison *Hypothesis: Applying advanced computational tools will reveal unique cancer evolutionary patterns, precise mutation timings, and distinct subclonal populations specific to the Black Americans with NSCLC cohort, providing a critical baseline for future studies.*

- Apply `GRITIC` and complementary tools from Gerstung et al., such as `cancerTiming`, `MutationTimeR`, and `PhylogicNDT`, to map the clonal and subclonal evolution of tumors in this dataset
- Identify key mutation timings and evolutionary patterns and use as a baseline for further comparison and benchmarking

Aim 3: Develop innovative methods to map the paths of mutational processes throughout clonal evolution with improved mutation-time resolution *Hypothesis: Improving mutation-time resolution of GRITIC will refine the accuracy of mapping mutational processes and timing driver mutations and enhance our understanding of cancer evolution.*

- Establish high-resolution mutation-time techniques to precisely time mutational signature activities
- Utilize the high-resolution timing data as inputs to improve the accuracy of timing key driver mutations
- Develop accessible open-source software for use by the scientific community

Research Strategy

Significance

Lung Cancer and Environmental Exposures Lung cancer remains the leading cause of cancer-related mortality worldwide, with non-small cell lung cancer (NSCLC) accounting for 85% of lung cancer cases in the US.[1] Akin to tobacco smoking, exposure to the complex mixture of air pollution, particularly fine particulate matter ($PM_{2.5}$) and nitric oxide (NO), poses a major risk factor for developing lung cancer. In heavily polluted cities like Los Angeles, exposure to these pollutants significantly increases the risk of developing lung cancer.[2,3] In 2014, the Nurse's Health Study found that living within 200 meters of a highway and a $10\ \mu g/m^3$ increase in $PM_{2.5}$ levels were associated with an increased risk of lung cancer (HR = 1.57; 95% CI: 1.26, 1.77).[4] Furthermore, a 2019 meta-analysis estimated that a $10\ \mu g/m^3$ increase in $PM_{2.5}$ exposure in Europe and North America increased lung cancer risk by 25%.[5]

Despite the clear evidence linking air pollution exposure to elevated lung cancer risk, the precise molecular mechanisms by which these complex pollutant mixtures initiate and promote NSCLC remain poorly understood, representing a critical knowledge gap. This study will investigate lung cancer in African Americans/Blacks (Black Americans), an understudied group that exhibits a high prevalence of aggressive, early-onset tumors that are often driven by distinct molecular profiles like EGFR mutations.[6] Elucidating the environmental drivers and biological pathways of lung carcinogenesis in this subgroup could reveal novel diagnostic approaches.

Addressing Lung Cancer Inequities in Black Americans Although Black Americans have lower smoking rates compared to non-Hispanic Whites, they experience significantly higher lung cancer incidence and mortality rates, especially among men.[7–9] This disparity is striking, as Black Americans tend to initiate smoking later in life and consume fewer cigarettes compared to their White counterparts.[8,10] Black women, despite smoking fewer cigarettes, have the same or higher incidence of lung cancer as White women.

Current lung cancer screening guidelines based on pack-years and age[11] fail to adequately identify Black Americans at risk. Black Americans are diagnosed with lung cancer at a significantly younger age than Whites, often before reaching the screening threshold of 30 pack-years or age 55.[12] The molecular drivers underlying these aggressive, early-onset lung cancers in the Black population remain unclear. However, disparities in environmental exposures, particularly air pollution, are suspected to play a role.[6] Evidence shows that Black Americans are consistently exposed to significantly higher levels of $PM_{2.5}$ and NO compared to non-Hispanic Whites. This study will utilize a multi-regional cohort of non-smokers, former smokers, and smokers, to identify the molecular connections between air pollutants and lung cancer in Black Americans.

Furthermore, existing studies do not account for how social determinants of health in Black Americans may modulate susceptibility to cancers.[9] Addressing this gap is crucial for accurately assessing risk and developing prevention strategies in diverse populations.

Characterization of Environmental Exposure Outdoor air pollution, including $PM_{2.5}$, is classified as a Group 1 carcinogen by the International Agency for Research on Cancer (IARC).[13] Past studies demonstrate a clear link between residing near major roadways and an elevated risk of developing lung cancer.[13] Exhaust from combustion engines releases a mixture of carcinogenic compounds into the atmosphere near major roadways. These pollutants include polycyclic aromatic hydrocarbons (PAHs), nitrogen oxides, and toxic heavy metals such as arsenic, nickel, and lead.[14] Previous studies have attempted to map air pollution levels using census tract data. However, these methods only detect a limited subset of pollutants, failing to capture the full complexity of environmental pollutants. Moreover, existing research does not account for how rising global temperatures associated with climate change may alter the chemical composition and carcinogenic potency of air pollution over time. Another major shortcoming is the lack of integration of social determinants of health, such as obesity, diabetes, and chronic inflammatory conditions, which may exacerbate susceptibility to cancer.

Harnessing Advanced Computational Models for Precise Mutation Timing and Cancer Progression Analysis Gerstung *et al.* apply a suite of sophisticated computational tools including `cancerTiming`, `MutationTimeR`, `PhylogicNDT SinglePatientTiming`, and `PhylogicNDT LeagueModel`, to analyze cancer progression and mutation timing. Another recent advancement in genomic analysis is `GRITIC`, detailed in Baker *et al.* This method utilizes advanced computational techniques to precisely time copy number gains in clonal populations. The strength of `GRITIC` lies in its ability to accurately determine the sequential timing of these copy number gains.

However, `GRITIC` and similar methods face limitations, such as the computational intensity of Bayesian inferences and Markov Chain Monte Carlo (MCMC) approaches, which restrict their application in large-scale or real-time scenarios. Additionally, they often require assumptions about mutation rates and copy number states that may not hold in all scenarios. Currently, `GRITIC` requires DNA segments to have at least 20 single nucleotide variants (SNVs), and does not extend to systems with high copy number ($\geq\ 9$).[15]

Our proposed research aims to build upon and enhance the current GRITIC methodology by addressing these limitations. We will develop improved mutational signature activity timing approaches with higher resolution, enabling more precise tracking of mutagenic exposures over clonal evolutionary time. These refined timing methods will serve as priors for developing more accurate driver mutation timing approaches. By integrating temporally varying distributions of mutation signatures, we will significantly enhance the precision of mutation timing and the temporal ordering of key driver mutations.

Additionally, we are committed to creating open-source, user-friendly software tools to disseminate these advanced methodologies to the broader research community. By making these tools accessible, we aim to promote widespread adoption and foster collaborative efforts to advance cancer research. Through these improvements, our research will overcome the current limitations of GRITIC and similar methods, providing more precise and scalable approaches to studying cancer evolution and the impact of environmental factors on mutational processes.

Potential for Transformative Impact This study will employ advanced geospatial methods to quantify individual exposures to air pollutants in Black communities in LA, Chicago, New Orleans, Charlestown SC, Richmond VA, and Rochester NY. Crucially, it will integrate this environmental exposure data with social determinants of health and biological factors that modulate disease susceptibility in these communities. Black populations in LA have historically faced disproportionately higher exposure to air pollution due to factors like redlining, the placing of industrial facilities near their neighborhoods, and a lack of green spaces. Despite having some of the lowest rates of smoking in the US, LA suffers from some of the worst highway-generated air pollution. By precisely characterizing these elevated exposures and combining them with data on obesity, diabetes, chronic inflammation, and other risk factors prevalent in Black communities, the goal is to develop a comprehensive analysis that elucidates how environmental drivers interact synergistically with social and biological parameters to initiate and promote aggressive, early-onset NSCLC in this population.

This multidisciplinary approach, which combines external exposure assessments with internal susceptibility factors, is poised to provide novel mechanistic insights into the environmental carcinogenesis pathways that contribute to the excessive lung cancer burden observed in Black communities. By correlating precise air pollution exposure data with epidemiological cohorts and molecular tumor profiling from Black NSCLC patients, the research will forge a comprehensive model of how environmental toxins catalyze lung carcinogenesis amidst the backdrop of social and biological vulnerabilities in this underserved population. We anticipate that this innovative approach will provide new insights into the role of air pollution in the development of NSCLC among Black Americans. This will help us develop early detection strategies. This is increasingly vital as, despite overall declining lung cancer rates, the incidence of NSCLC among women of color is rising in LA and similar urban areas.

Innovation

1. Focus on an Understudied Population: Targeting the Black Americans with NSCLC cohort addresses a significant gap in cancer research. Understanding the distinct mutational signatures and evolutionary patterns in this population can lead to more personalized and effective treatment strategies, ultimately improving clinical outcomes and addressing health disparities.
2. Comprehensive Evolutionary Analysis: The use of advanced computational tools, including GRITIC and complementary methods from Gerstung *et al.*, to map the clonal and subclonal evolution of tumors is a significant innovation. This approach will provide a detailed understanding of the evolutionary dynamics and precise mutation timings specific to the Black Americans with NSCLC cohort.
3. High-Resolution Mutation Timing: Another major innovation is the development of methods for mapping the trajectories of mutational processes with improved mutation-time resolution. By establishing high-resolution techniques to precisely time mutational signature activities and refining the accuracy of driver mutation timing, this aim will significantly enhance our understanding of cancer evolution.
4. Correlate Environmental Toxin to Higher Rates of NSCLC: The study will identify molecular mechanisms correlated with exposure to environmental toxins, which is a highly impactful aspect of this research. Understanding these correlations can lead to insights into how environmental factors contribute to cancer development and progression in Black Americans with NSCLC.
5. Development of Open-Source Tools: Creating accessible, user-friendly software tools that will be disseminated to the research community ensures that the advanced methodologies developed in this study can be widely adopted and applied in other cancer research contexts. This commitment to open science and collaboration is crucial to community-wide innovation by promoting broader impact and advancements in the field.

Through these innovative approaches, this research will bridge critical knowledge gaps in oncology, particularly regarding the unique needs of Black communities. The insights gained from this study are expected to lead to more personalized and effective treatment strategies, improved early cancer detection, and ultimately, better patient care and outcomes.

Approach

Aim 1: Preprocess raw data from the Black Americans with NSCLC cohort

Introduction Aim 1 focuses on preprocessing the raw genomic data collected from Black Americans with NSCLC. This process is essential for ensuring the accuracy and reliability of subsequent analyses. Given the complexity and volume of genomic data, meticulous preprocessing is required to remove biases, correct errors, and prepare the data for detailed mutation and evolutionary analysis.

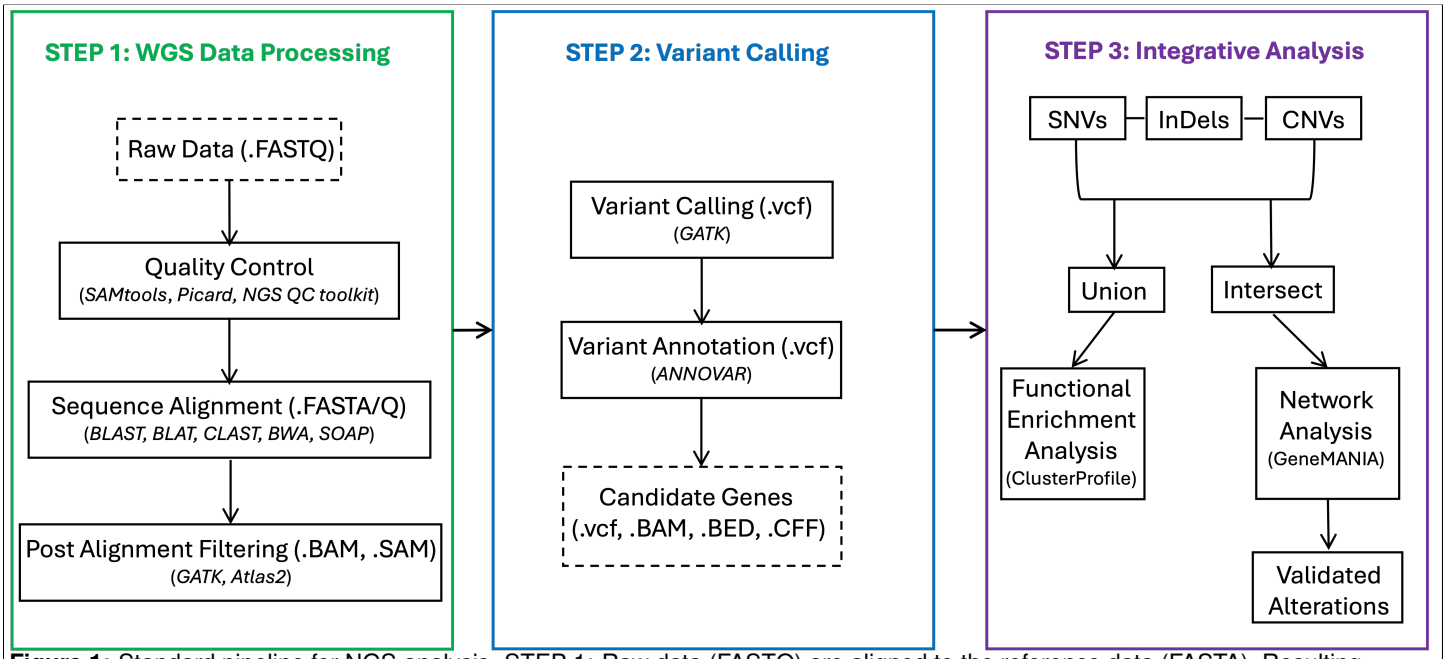

Figure 1: Standard pipeline for NGS analysis. STEP 1: Raw data (FASTQ) are aligned to the reference data (FASTA). Resulting alignments are stored in binary alignment map (BAM) file format. STEP 2: Sequence variants are identified and annotated using tools such as GATK and ANNOVAR. Candidate genes are stored in various file formats, *e.g.* .vcf, .BAM, .BED, .CFF. STEP 3: Candidate genes with SNVs, InDels, and CNVs are analyzed and validated using tools such as ClusterProfile and GeneMANIA.

Whole genome sequencing (WGS) is a comprehensive method for analyzing the entire genomic DNA of an organism. This technique allows for the identification of genetic variations, including single nucleotide variants (SNVs), insertions and/or deletions (InDels), and structural variations across the entire genome. Next-generation sequencing (NGS) refers to a collection of modern sequencing technologies that enable the rapid sequencing of large amounts of DNA. These technologies have revolutionized genomic research by providing high-throughput, accurate, and cost-effective sequencing solutions.[16] NGS allows for easy implementation of WGS, providing a powerful framework for uncovering the genetic underpinnings of diseases such as NSCLC.

Given the nature of genomic data, which often contains significant amounts of noise due to sequencing processes, a series of rigorous data cleaning protocols is commonly implemented. This includes removing low-quality sequences to prevent erroneous interpretations, aligning the raw reads to a reference genome to identify the genomic coordinates of the sequences, removing duplicate reads to avoid overrepresentation, and correcting sequencing errors to improve data accuracy. Additionally, it is crucial to account for mutations present in healthy cells to differentiate between somatic mutations and germline variants, thus focusing on cancer-specific changes.[16,17]

Once the data has been cleaned, the next step is normalization. Genomic data can be affected by various technical variances and batch effects that arise during sample collection, sequencing, and processing. These variances can obscure true biological signals if not properly addressed.[18] The normalization process involves adjusting for systematic biases and standardizing the data to ensure that it is comparable across different samples and experimental conditions. This is particularly important when dealing with large datasets collected from diverse sources, as it enhances the comparability of the data and ensures that any observed differences are due to biological variations rather than technical artifacts.

The final preprocessing step involves the detection and classification of mutations. This step is crucial for understanding the genetic underpinnings of NSCLC and involves identifying single nucleotide variants (SNVs), which are point mutations that can significantly impact gene function, as well as detecting insertions and deletions (indels) that result in structural

genomic changes. Additionally, it is essential to identify and analyze copy number variants (CNVs), which are large regions of the genome that have been duplicated or deleted. CNVs can have significant implications for gene expression and tumor progression. Using established bioinformatics pipelines to classify and annotate these genetic variations provides insights into their potential impact and relevance to cancer progression. By accurately detecting and classifying these mutations and CNVs, we can create a comprehensive genetic profile of NSCLC in Black Americans, which serves as a foundation for further evolutionary analyses[17].

We anticipate several challenges in this aim, including high inter-individual variability in mutation rates and evolutionary patterns. To address these challenges, we plan to incorporate a larger sample size to mitigate the impact of individual variability and provide more robust statistical power. Employing advanced statistical techniques will help ensure that our findings are generalizable and reliable. The computational demands of methods like GRITIC will be managed by leveraging high-performance computing resources, ensuring that analyses are conducted efficiently and accurately. Additionally, using standardized data formats and developing custom scripts for data merging and analysis will facilitate the seamless integration of results from different computational methods (Gerstung et al., 2020).

Rationale Aim 1 focuses on preprocessing the raw genomic data from Black Americans diagnosed with NSCLC, a demographic often underrepresented in cancer research. This step is crucial for establishing a reliable baseline for subsequent analyses. By implementing rigorous data cleaning and preprocessing protocols, we aim to remove low-quality sequences, align the reads to the reference genome, and filter data for further analysis using `GRITIC`. The resulting high-quality dataset will be essential for accurately mapping mutational signatures and evolutionary patterns specific to this population.

1.1. Perform data cleaning and preprocessing by removing low-quality sequences, aligning the reads to the reference genome, and filtering data for further analysis with `GRITIC` This stage involves collaborating with regional medical centers and cancer research networks to collect extensive genomic data from NSCLC patients, particularly those of Black American descent. The data will undergo rigorous cleaning and preprocessing steps to remove biases and errors. This includes aligning sequences to reference genomes, removing duplicate reads, correcting sequencing errors, and accounting for mutations in healthy cells. We will implement stringent quality control measures to validate the data's integrity and completeness, ensuring it meets the analytical requirements of the computational methods to be applied. This step is expected to be time-consuming, given the complexity of preprocessing genomic data.

1.2. Normalize the data to account for batch effects and other technical variances In this phase, we will normalize the dataset to account for batch effects and other technical variances. Normalization is crucial to ensure that the data is comparable across different samples and experiments. This involves adjusting for systematic biases that may arise during data collection and sequencing processes. We will employ robust statistical techniques to standardize the data, making it suitable for downstream analyses such as mutation detection and evolutionary trajectory modeling.

1.3. Detect and classify mutations (single nucleotide variants, insertions, deletions, etc.) using established bioinformatics pipelines to prepare the dataset for subsequent evolutionary analysis. After normalizing the data, we will detect and classify mutations, including single nucleotide variants (SNVs), insertions, and deletions, using established bioinformatics pipelines. These pipelines will enable us to systematically identify and annotate genetic variations within the dataset. The accurate detection and classification of mutations are critical for subsequent evolutionary analyses, as they provide the foundation for understanding the genetic dynamics and progression of NSCLC in this population.

Challenges & Alternative Approaches We anticipate challenges such as high inter-individual variability in mutation rates and evolutionary patterns. To address this, we plan to incorporate a larger sample size and employ robust statistical methods to ensure the generalizability of our findings. The computational demands of the `GRITIC` method will be managed by leveraging high-performance computing resources and optimizing algorithm parameters to enhance computational efficiency without compromising accuracy. We will also ensure seamless integration of results from different computational methods by using standardized data formats and developing custom scripts for data merging and analysis.

Aim 2: Conduct evolutionary analysis to establish a baseline for subsequent comparison

Rationale In the realm of cancer research, particularly in studies like those conducted by Gerstung *et al.*, traditional methods for modeling mutation timings often rely on statistical models. These models typically encompass a variety of approaches ranging from regression-based frameworks to simpler probabilistic models like Markov chains. These models are well-established for their interpretability and methodological transparency but may lack the nuanced capacity to handle complex, high-dimensional, and non-linear patterns that characterize genetic data and cancer progression.

Unlike conventional statistical models that may not capture complex dependencies, the LSTM component of the hybrid model can analyze sequential and temporal data over extended periods. This capability allows it to discern underlying patterns that are predictive of disease progression that might be overlooked by other methods. By integrating LSTM-derived insights with an HMM, the hybrid model enriches the probabilistic modeling of hidden states. The LSTM layer enables the interpretation of details and complexities within sequential data that may not be immediately apparent or directly measurable. Using the LSTM outputs in the HMM model improves the HMM's ability to predict state transitions and better capture the

complex interactions in the mutation processes.

Utilizing an LSTM enables the model to learn from and adapt to new data dynamically, a significant advantage over traditional models that may require static reconfiguration or retraining. The integration of LSTM with HMM not only enhances accuracy but also provides finer resolution in the timing of mutations, offering detailed insights that are crucial for effective treatment planning and personalized medicine.

2.1. Apply `GRITIC` and complementary tools from Gerstung et al., such as `cancerTiming`, `MutationTimeR`, and `PhylogicNDT`, to map the clonal and subclonal evolution of tumors in this dataset To harness the predictive power of advanced machine learning for understanding disease progression, we propose developing an LSTM network specifically designed to analyze biological sequences. The LSTM layer will extract temporal features and patterns that indicate shifts in disease progression. The LSTM will output detailed state prediction probabilities and feature vectors that encapsulate the dynamic characteristics of the disease, providing a rich dataset that reflects both the current state and the likely future states of the disease. These outputs from the LSTM will then be fed into the HMM. Specifically, the state prediction probabilities generated by the LSTM will be used as emission probabilities in the HMM. This step is expected to refine the HMM's ability to map observed data to the correct hidden states. Additionally, the feature vectors derived from the LSTM will serve as observations in the HMM. This integration enhances the HMM's capability to accurately model transitions between different disease states, leveraging the detailed context provided by the LSTM to improve both the accuracy and reliability of the disease progression model.

2.2. Identify key mutation timings and evolutionary patterns and use as a baseline for further comparison and benchmarking In this aim, we will implement and optimize both LSTM and HMM models within a unified framework, to ensure robust predictions of mutation events. To achieve this, we will initially pre-train the LSTM on a designated subset of the PCAWG dataset. This preliminary step is designed to stabilize the LSTM's feature extraction capabilities before it is fully integrated with the HMM. After we train the LSTM, we will use the outputs of the LTSM to dynamically adjust the parameters of the HMM. This adjustment process will utilize a combined training approach that integrates backpropagation for optimizing the LSTM alongside the Baum-Welch algorithm for refining the HMM. To validate the model's robustness, we will perform cross-validation to prevent overfitting and to ensure that the model generalizes effectively across different subsets of data. Lastly, we will continuously optimize the model parameters, focusing on improving key performance metrics such as accuracy, sensitivity, and specificity in predicting mutation timing. This thorough approach to training and validation aims to create a predictive model that is both precise and reliable in its application to real-world clinical data.

Challenges & Alternative Approaches In our approach to model complexity and data sparsity, it's important to note that HMMs are inherently proficient at handling sparse data, providing a robust fallback mechanism when large datasets are not available. This capability allows us to potentially minimize the reliance on extensive machine learning processes in scenarios where data is limited, while still maintaining the flexibility to fully leverage more complex machine learning techniques, such as those involving LSTM networks, as richer datasets become available. By designing the system with this dual capability, we ensure that the model remains effective under varied data conditions, offering both simplicity and adaptability. This approach not only safeguards against overfitting but also ensures that the model can dynamically scale its complexity based on the volume and detail of the data it processes.

Aim 3: Develop innovative methods to map the paths of mutational processes throughout clonal evolution with improved mutation-time resolution

Rationale Aim 3 of our project involves the crucial step of benchmarking the LSTM-HMM hybrid model against established computational methods such as those from Gerstung et al. and the GRITIC method. Furthermore, we plan to extend the application of this model to other types of cancer to assess its generalizability and efficacy across various oncological contexts. This approach is designed to solidify the model's robustness and adaptability and to potentially set a new standard in the precision modeling of cancer evolution.

3.1. Establish high-resolution mutation-time techniques to precisely time mutational signature activities We will start by implementing the LSTM-HMM model on the NSCLC dataset that has been analyzed in Aim 1 and Aim 2. The primary goal here is to directly compare the performance of our model against the outcomes derived from the Gerstung et al. and GRITIC methods. Performance metrics such as accuracy, computational efficiency, and adaptability will be meticulously documented. We will use statistical measures like the area under the ROC curve (AUC), confusion matrices for classification accuracy, and time-efficiency analyses to quantitatively assess each model's performance.

3.2. Utilize the high-resolution timing data as inputs to improve the accuracy of timing key driver mutations After benchmarking, we will apply the LSTM-HMM model to additional cancer datasets. These datasets will be selected based on their variance in diagnosis, treatment outcomes, and genetic diversity to ensure a comprehensive test of the model's applicability. For each new type of cancer analyzed, we will adjust and fine-tune the model parameters to accommodate different genetic signatures and mutation rates. This phase will help in identifying unique and shared mutation timings and evolutionary patterns across cancers, which could be crucial for understanding broad and specific pathways of oncogenesis.

3.3. Develop accessible open-source software for use by the scientific community Each application of the model will be followed by a detailed comparative analysis where results from the LSTM-HMM model will be juxtaposed against those obtained using traditional methods. This will allow us to evaluate the added value of our approach in terms of enhanced resolution, predictive accuracy, and the ability to adapt to different types of cancer data. Insights gained from these comparisons will be critical in demonstrating the model's effectiveness and could pave the way for its adoption in clinical settings.

Challenges & Alternative Approaches One anticipated challenge is the variance in data quality and completeness across different cancer datasets, which could affect model performance. To mitigate this, we will employ data augmentation techniques and sophisticated imputation methods to ensure data integrity. Additionally, the computational demands of applying LSTM-HMM to diverse and extensive datasets will be managed through the use of scalable cloud computing resources and optimizing the model's architecture for high-performance computing environments.

References

[1] Julian R. Molina, Ping Yang, Stephen D. Cassivi, Steven E. Schild, and Alex A. Adjei. “Non–Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship”. In: *Mayo Clinic proceedings. Mayo Clinic* 83.5 (May 2008). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2718421/, pp. 584–594. ISSN: 0025-6196.

[2] Christine D. Berg, Joan H. Schiller, Paolo Boffetta, Jing Cai, Casey Connolly, Anna Kerpel-Fronius, Andrea Borondy Kitts, David C.L. Lam, Anant Mohan, Renelle Myers, Tejas Suri, Martin C. Tammemagi, Dawei Yang, and Stephen Lam. “Air Pollution and Lung Cancer: A Review by International Association for the Study of Lung Cancer Early Detection and Screening Committee”. In: *Journal of Thoracic Oncology* 18.10 (Oct. 2023), pp. 1277–1289. ISSN: 15560864. DOI: 10.1016/j.jtho.2023.05.024.

[3] Yanqian Huang, Meng Zhu, Mengmeng Ji, Jingyi Fan, Junxing Xie, Xiaoxia Wei, Xiangxiang Jiang, Jing Xu, Liang Chen, Rong Yin, Yuzhuo Wang, Juncheng Dai, Guangfu Jin, Lin Xu, Zhibin Hu, Hongxia Ma, and Hongbing Shen. “Air Pollution, Genetic Factors, and the Risk of Lung Cancer: A Prospective Study in the UK Biobank”. In: *American Journal of Respiratory and Critical Care Medicine* 204.7 (Oct. 2021), pp. 817–825. ISSN: 1073-449X. DOI: 10.1164/rccm.202011-4063OC.

[4] Robin C. Puett, Jaime E. Hart, Jeff D. Yanosky, Donna Spiegelman, Molin Wang, Jared A. Fisher, Biling Hong, and Francine Laden. “Particulate Matter Air Pollution Exposure, Distance to Road, and Incident Lung Cancer in the Nurses’ Health Study Cohort”. In: *Environmental Health Perspectives* 122.9 (Sept. 2014), pp. 926–932. ISSN: 1552-9924. DOI: 10.1289/ehp.1307490.

[5] Hung-Ling Huang, Yung-Hsin Chuang, Tzu-Hsuan Lin, Changqing Lin, Yen-Hsu Chen, Jen-Yu Hung, and Ta-Chien Chan. “Ambient Cumulative PM2.5 Exposure and the Risk of Lung Cancer Incidence and Mortality: A Retrospective Cohort Study”. In: *International Journal of Environmental Research and Public Health* 18.23 (Nov. 2021), p. 12400. ISSN: 1661-7827. DOI: 10.3390/ijerph182312400.

[6] Anita Marcinkiewicz, Aleksandra Ochotnicka, Karolina Borowska-Waniak, Kinga Skorupińska, Dominik Michalik, and Maja Borowska. “THE IMPACT OF AIR POLLUTION ON THE OCCURRENCE OF LUNG CANCER: A LITERATURE REVIEW”. In: *Archiv Euromedica* 13.5 (Oct. 2023). ISSN: 2199885X. DOI: 10.35630/2023/13/5.507.

[7] Nadine Belony, Bing Ren, Phuc Pham, Matthew Gregory, Pablo E. Puente, Nazarius S. Lamango, Ite A. Offringa, and Yong Huang. “Abstract 4038: Study of Therapeutic Effects of Polyisoprenylated Cysteinyl Amide Inhibitors on Lung Cancer Cells of Black Patients Using 3D-Printed Alveolar Model”. In: *Cancer Research* 82.12_Supplement (June 2022), p. 4038. ISSN: 0008-5472. DOI: 10.1158/1538-7445.AM2022-4038.

[8] Jose Thomas Thaiparambil, Zheng Yin, and Randa El-Zein. “Abstract 1948: Novel Insights into Lung Cancer & Chronic Obstructive Pulmonary Disease Racial Disparities”. In: *Cancer Research* 83.7_Supplement (Apr. 2023), p. 1948. ISSN: 0008-5472. DOI: 10.1158/1538-7445.AM2023-1948.

[9] Celina I. Valencia, Francine C. Gachupin, Yamilé Molina, and Ken Batai. “Interrogating Patterns of Cancer Disparities by Expanding the Social Determinants of Health Framework to Include Biological Pathways of Social Experiences”. In: *International Journal of Environmental Research and Public Health* 19.4 (Feb. 2022), p. 2455. ISSN: 1661-7827. DOI: 10.3390/ijerph19042455.

[10] Karyn Hede. “Drilling Down to the Causes of Racial Disparities in Lung Cancer”. In: *JNCI: Journal of the National Cancer Institute* 102.18 (Sept. 2010), pp. 1385–1387. ISSN: 0027-8874. DOI: 10.1093/jnci/djq371.

[11] Rebecca Landy, Li C. Cheung, Corey D. Young, Anil K. Chaturvedi, and Hormuzd A. Katki. “Absolute Lung Cancer Risk Increases among Individuals with $>$15 Quit-Years: Analyses to Inform the Update of the American Cancer Society Lung Cancer Screening Guidelines”. In: *Cancer* 130.2 (Jan. 2024), pp. 201–215. ISSN: 0008-543X. DOI: 10.1002/cncr.34758.

[12] Rafael Meza, Jihyoun Jeon, Iakovos Toumazis, Kevin Ten Haaf, Pianpian Cao, Mehrad Bastani, Summer S. Han, Erik F. Blom, Daniel E. Jonas, Eric J. Feuer, Sylvia K. Plevritis, Harry J. de Koning, and Chung Yin Kong. “Evaluation of the Benefits and Harms of Lung Cancer Screening With Low-Dose Computed Tomography: Modeling Study for the US Preventive Services Task Force”. In: *JAMA* 325.10 (Mar. 2021), pp. 988–997. ISSN: 1538-3598. DOI: 10.1001/jama.2021.1077.

[13] Shilpa N. Gowda, Anneclaire J. DeRoos, Rebecca P. Hunt, Amanda J. Gassett, Maria C. Mirabelli, Chloe E. Bird, Helene G. Margolis, Dorothy Lane, Matthew R. Bonner, Garnet Anderson, Eric A. Whitsel, Joel D. Kaufman, and Parveen Bhatti. “Ambient Air Pollution and Lung Cancer Risk among Never-Smokers in the Women’s Health Initiative”. In: *Environmental Epidemiology* 3.6 (Oct. 2019), e076. ISSN: 2474-7882. DOI: 10.1097/EE9.0000000000000076.

[14] Xian-Jun Yu, Min-Jun Yang, Bo Zhou, Gui-Zhen Wang, Yun-Chao Huang, Li-Chuan Wu, Xin Cheng, Zhe-Sheng Wen, Jin-Yan Huang, Yun-Dong Zhang, Xiao-Hong Gao, Gao-Feng Li, Shui-Wang He, Zhao-Hui Gu, Liang Ma, Chun-Ming Pan, Ping Wang, Hao-Bin Chen, Zhi-Peng Hong, Xiao-Lu Wang, Wen-Jing Mao, Xiao-Long Jin, Hui Kang, Shu-Ting Chen, Yong-Qiang Zhu, Wen-Yi Gu, Zi Liu, Hui Dong, Lin-Wei Tian, Sai-Juan Chen, Yi Cao, Sheng-Yue Wang, and Guang-Biao Zhou. “Characterization of Somatic Mutations in Air Pollution-Related Lung Cancer”. In: *EBioMedicine* 2.6 (June 2015), pp. 583–590. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2015.04.003.

[15] Toby M. Baker, Siqi Lai, Tom Lesluyes, Haixi Yan, Annelien Verfaillie, Stefan Dentro, Andrew R. Lynch, Amy L. Bowes, Nischalan Pillay, Adrienne M. Flanagan, Charles Swanton, Maxime Tarabichi, and Peter Van Loo. *The History of Chromosomal Instability in Genome Doubled Tumors*. Oct. 2023. DOI: 10.1101/2023.10.22.563273.

[16] Daniel C. Koboldt. "Best Practices for Variant Calling in Clinical Sequencing". In: *Genome Medicine* 12.1 (Oct. 2020), p. 91. ISSN: 1756-994X. DOI: 10.1186/s13073-020-00791-w.

[17] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G. Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C. Boutros, David D. Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D. Morris, Paul T. Spellman, David C. Wedge, and Peter Van Loo. "The Evolutionary History of 2,658 Cancers". In: *Nature* 578.7793 (Feb. 2020), pp. 122–128. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1907-7.

[18] Andrew E. Jaffe, Thomas Hyde, Joel Kleinman, Daniel R. Weinbergern, Joshua G. Chenoweth, Ronald D. McKay, Jeffrey T. Leek, and Carlo Colantuoni. "Practical Impacts of Genomic Data "Cleaning" on Biological Discovery Using Surrogate Variable Analysis". In: *BMC Bioinformatics* 16 (Nov. 2015), p. 372. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0808-5.