Cell-free DNA (cfDNA) refers to fragmented DNA molecules found circulating freely in the bloodstream or other bodily fluids. Sequencing cfDNA molecules enables the non-invasive detection of cancer by identifying tumor-specific genetic mutations, such as through the use of GEMINI (GEnome-wide Mutational Incidence for Non-Invasive detection of cancer). [4] To better understand tumor mechanisms and subtypes, it is important to explore the relationship between somatic mutations and gene regulation in tumor versus healthy cells. Analyzing cfDNA methylation and histone modification patterns can help uncover this relationship. [12, 2]
In the proposed study, we seek to connect the 3D landscape of cfDNA with its somatic mutations by analyzing matched genetic, methylation, and histone modification data. Mutations are limited in that they cannot capture the tissue of origin, while epigenetic features overcome this limitation and have been shown to be stereotyped between tissues and higher in number of alterations compared to mutations. [12, 3, 5] We specifically create genetic and epigenetic profiles of breast cancer patients and apply machine learning models to predict breast cancer subtypes and infer regulatory mechanisms underlying these subtypes.

Aim 1: Determine the association of the cfDNA methylome and cancer subtypes

Recent studies have shown that methylation patterns are highly indicative of cell type of origin.[15] By analyzing methylation patterns of cfDNA, we can infer the tissue of origin of the tumor and apply this towards subtyping.

1A: Apply cfMethyl-seq [16] and whole-genome sequencing to a breast cancer cohort to capture genome-wide methylation profiles of CpG-rich regions and mutation profiles. We collect methylation data for both tumor and healthy samples to identify tumor-specific methylation changes. A cohort of 1000 breast cancer patients with various subtypes will be used.
1B: Prediction of cancer subtypes based on methylation and mutation features. Using the approach of GEMINI, we can calculate tumor-normal matched methylation scores over nonoverlapping genomic bins. We then will evaluate machine learning models in predicting cancer subtypes based on GEMINI-methylation features, GEMINI-mutation features, and a combination of both.
1C: Inferring driving gene expression programs from cfDNA-derived mutation and methylation signals. We will identify hypermethylated and hypomethylated activity in promoter and enhancer regions and associate these signals with gene expression programs in the tumor samples to determine regulatory mechanisms underlying each breast cancer subtype.

Aim 2: Predicting tumor transcriptome through cfDNA-derived mutations and histone modification signals in promoter and enhancer regions

Different cancer subtypes have distinct gene expression profiles. cfDNA is characterized by nucleosomes that protect it from nuclease-mediated degradation, preserving protein modifications such as histone modifications (HM). Combining cfDNA HM data, which is associated with gene expression programs of their cells of origin [14, 17, 2], with mutation profiles can provide a view of the gene expression landscape in a tumor.

2A: Collect cfCHiP-seq [2] to generate histone modification signals and mutation profiles. We will use the same cohort to generate H3K4me3 and H3K27ac signals, histone modifications that are associated with promoters and enhancers, respectively.
2B: Collect transcriptome data for tumor samples in the cohort. We will collect RNA-seq data for the tumor samples in the cohort to generate gene expression profiles.
2C: Build and evaluate machine learning models to predict gene expression profiles based on GEMINI-mutation features, H3K4me3 signals, and H3K27ac signals. We will evaluate the performance of machine learning models in predicting gene expression profiles based on GEMINI-mutation features, H3K4me3 signals, and H3K27ac signals.

The proposed approach will provide a comprehensive view of the relationship between somatic mutations and regulatory properties of DNA, specifically methylation patterns and histone modifications. This will enable the non-invasive subtyping of cancer based on cfDNA data, providing insights into the regulatory mechanisms underlying cancer subtypes and informing the development of targeted therapies.

Significance

Paired genetic and epigenetic data for non-invasive cancer subtyping

The proposed study seeks to provide a comprehensive view of the interplay between somatic mutations and regulatory properties of cfDNA, specifically methylation patterns and histone modifications. Existing studies have focused on either genetic or epigenetic aspects of cfDNA, and only recently have there been efforts in understanding how these modalities interact within the same sample. [3, 6] Using enriched cfDNA samples and parallel assaying, we generate paired genetic and epigenetic data that will help understand the correspondence between genetic mutations and the 3D landscape of DNA. This will enhance our ability in several cancer-related prediction tasks, such as predicting tumor behavior, tumor growth, therapy response, and metastatic potential. Moreover, this study may also motivate further work in integrating other omes of cfDNA, such as fragmentomics. [12]

Identification of significant regulatory regions

Regulatory regions, such as promoters and enhancers, are important for understanding oncogenic drivers and tumor suppressors. Profiling cfDNA for histone modifications and methylation patterns will bring us closer to identifying the regulatory patterns in a tumor, as genetic mutations alone do not provide a complete picture of the regulatory landscape. With the additional collection of tumor transcriptome data, we can thoroughly investigate gene regulation in a tumor solely based on the data from a remote blood draw. This paves the way for the construction of regulatory networks that can be used to predict tumor behavior based on cfDNA properties.

More granular subtyping for improved personalized medicine

While two different subtypes of cancer may share the same genetic mutation profiles, they may have distinct transcriptome and proteome profiles. By incorporating methylation and histone modification data, we may be able to distinguish more granular subtype differences that are not immediately apparent in genetic mutation profiles alone. [9] Discovering more granular subtypes can allow us to construct a hierarchy of cancer subtypes and potentially explain mechanisms of clonal evolution. Identifying these differences can help in the development of more targeted therapies.

Advancing non-invasive liquid biopsy diagnostics

The proposed approach may revolutionize cancer diagnostics by enabling the classification of a patient's subtype through a fast and non-invasive blood test. For heterogeneous cancers like breast cancer [8], this will be particularly impactful as it will allow for crucial targeted and personalized treatment plans. Current diagnosis methods often rely on invasive tissue biopsies, which may not be feasible for all patients due to the risks associated with surgery and the difficulty in obtaining tissue samples from certain parts of the body. cfMethyl-seq and cfChIP-seq not only offer non-invasive alternatives for arriving at similar diagnostic conclusions, but are also cost-effective and scalable. [16, 17] Furthermore, identification of significant associations between methylation, mutation, histone modification events with the transcriptome of different cancer subtypes can be used to create targeted biomarker assays, which can ultimately be validated both analytically - determining the precision and sensitivity of the assay - and clinically - determining the utility of the assay in informing medical decisions. [18]

Innovation

This study introduces a novel integrative approach by combining multiple cfDNA omics, including methylation, mutations, and histone modifications, from the same sample. This method reduces cost and time while enhancing the depth of analysis. By applying this workflow to heterogeneous cancer cohorts, we demonstrate its generalizability across cancer types and potential applicability to other diseases. Furthermore, the integration of machine learning models with our cfDNA-derived data enables real-time subtype predictions, with the potential for adoption into clinical workflows. Future extensions of this framework may incorporate longitudinal data to track tumor evolution and treatment responses, offering dynamic information concerning cancer progression and treatment efficacy.

Aim 1: Determine the association of the cfDNA methylome and cancer subtypes

Aim 1A: Generate paired cfDNA methylation, mutation, and tumor transcriptomic profiles in a breast cancer cohort

Rationale: Breast cancer is a highly heterogenous disease, where epigenetic changes, such as methylation, play a critical role in tumor development, progression, and subtype differentiation. [8] The integration of cfDNA methylation profiles with mutation profiles and a transcriptomic picture of the tumor offers a comprehensive framework identify mutations driving tumor initiation and progression and understand how aberrant methylation patterns can regulate tumor suppressor genes and oncogenes.

Methods

Construct a cfDNA integrative omics dataset using a 1000-patient breast cancer cohort: We will analyze a cohort of 1,000 breast cancer patients representing the four major subtypes: Luminal A, Luminal B, HER2-enriched, and Triple-negative. [8] Blood samples will be collected from each patient, and cfDNA will be extracted using the QIAamp Circulating Nucleic Acid Kit. Corresponding tissue biopsies will also be collected for bulk RNA sequencing to capture tumor transcriptomic data. To link cfDNA methylation and genome sequencing data accurately, each cfDNA molecule will be tagged with a unique molecular identifier (UMI). Tumor fraction in each cfDNA sample will be estimated using ichorCNA [1], and samples with low tumor fraction will be excluded from further analysis. cfDNA concentration and quality will also be validated using the Bioanalyzer 2100. [13] The extracted cfDNA will then be split for two complementary assays: cfMethyl-seq and next-generation whole-genome sequencing (WGS). cfMethyl-seq is cost-efficient and captures genome-wide CpG methylation with low DNA input, while WGS provides comprehensive genomic data for a multi-omics analysis of breast cancer subtypes.

Generation of paired GEMINI-methylation and GEMINI-mutation features: We apply the GEMINI method [4] to calculated tumor-normal matched methylation and mutation scores over nonoverlapping genomic bins. The generation of regional methylation and mutation features for J cancer patients and K healthy patients is defined as:

$$M_i = \frac{\sum_{j\in J} y_{ij}^T}{\sum_{j\in J} x_{ij}^T} - \frac{\sum_{k\in K} y_{ik}^N}{\sum_{k\in K} x_{ik}^N}, \quad G_i = \frac{\sum_{j\in J} z_{ij}^T}{\sum_{j\in J} x_{ij}^T} - \frac{\sum_{k\in K} z_{ik}^N}{\sum_{k\in K} x_{ik}^N} \tag{1}$$

where M_i and G_i are the methylation and mutation scores for bin i, respectively. y_{ij}^T and y_{ik}^N are the number of methylated sites in tumor and normal samples, respectively, for bin i. x_{ij}^T and x_{ik}^N are the total number of viable reads in tumor and normal samples, respectively, for bin i. z_{ij}^T and z_{ik}^N are the number of mutations in tumor and normal samples, respectively, for bin i. Having the same binned representation for both methylation and mutation data allows for direct comparison between the two modalities for downstream analyses.

Aim 1B: Prediction of cancer subtypes based on methylation and mutation features

Rationale: Following the generation of GEMINI-methylation and GEMINI-mutation features, we are able to determine the association of the cfDNA methylome, genetic mutations, and breast cancer subtypes. Using machine learning models, we can evaluate the predictive power of these features in distinguishing between the subtypes.

Methods

Unsupervised clustering of methylation and mutation features: To explore patterns in the cfDNA GEMINI features, unsupervised clustering will be performed on methylation and mutation data independently and in combination. Dimensionality reduction techniques, such as UMAP, will be used to visualize sample groupings and assess separation between tumor and normal samples, as well as among cancer subtypes. This analysis will help determine whether integrating methylation and mutation features improves the resolution of cancer subtyping.

Supervised classification of cancer subtypes: We will evaluate the performance of basic machine learning

models in predicting the breast cancer subtypes based on GEMINI-methylation features, GEMINI-mutation features, and a combination of both. We will use multinomial logistic regression as a baseline to predict subtypes and assess feature importance. More complex models, such as random forests and neural networks, will then be evaluated for improved predictive performance.

Aim 1C: Inferring driving gene expression programs from cfDNA-derived mutation and methylation signals

Rationale: With the availability of gene expression data in the tumor, we can validate the association between the methylation and mutation signals in the cfDNA and the gene expression programs in the tumor. Specifically, we can determine whether specific methylation events or genetic mutations are associated with the upregulation or downregulation of specific genes. This information can provide insights into the regulatory mechanisms underlying each breast cancer subtype.

Methods

Identify hypermethylated and hypomethylated activity in promoter and enhancer regions We will isolate the methylation signals in the promoter and enhancer regions of the genome, as these regions are known to be most relevant for gene expression regulation. We will then identify, for each cancer subtype, differentially methylated regions (DMRs) between the tumor and normal samples using Dispersion Shrinkage for Sequencing (DSS), a package for detecting differentially methylated regions from WGBS data. [7]

Associate regulatory regions' methylation and mutation signals with gene expression programs: We will correlate the methylation and mutation signals in the regulatory regions with their corresponding gene expression programs in the tumor samples. We will also validate specific DMRs using annotations from the MENT and PubMeth databases [18], which provide prior knowledge on the regulatory roles of specific methylation events in cancer.

Expected Results and Interpretation

We anticipate significant differences in methylation and mutation patterns between tumor and normal samples, as is consistent with hypermethylation and higher mutation frequences observed in tumors. Subtype-specific differences in these signals are also expected, reflecting distinct genetic and epigenetic profiles of each breast cancer subtype. In promoter and enhancer regions, we expect to observe hypermethylated regulatory elements to be associated with downregulated gene expression, while hypomethylated regions may be associated with upregulated gene expression. There may also be specific mutations in these regions that are associated with altered gene expression. Integrating methylation and mutation features is hypothesized to enhance machine learning classification performance [11], implying that the modalities provide complementary information. Similarly, dimensionality reduction should reveal separations between tumor vs. normal samples and among cancer subtypes as a result of distinct molecular signatures.

Potential Pitfalls and Alternative Approaches

One of the primary challenges in collecting matched genetic and epigenetic data is having a sufficient yield and quality of cfDNA to perform both assays. Without significant enrichment of cfDNA, it may be difficult to detect methylation signals and mutations in some samples. We chose the QIAamp Circulating Nucleic Acid Kitbecause it has been shown to have one of the highest yields of cfDNA compared to other experimental workflows.[13] In the event that insufficient cfDNA is obtained, we will consider using PCR-based methods, such as methylation-specific PCR, to enrich for the methylation signal. [10] For the machine learning experiments, we may observe a lack of clear separation between the cancer subtypes and consequently, a difficulty in predicting the subtypes. This could be due to overlapping biological characteristics, or more likely, noise in cfDNA data. To address this potential pitfall, we will consider applying various denoising approaches, including feature selection and batch effect correction. We can also adjust the bin size adopted for the GEMINI cfDNA features to reduce noise and redundancy.

References

[1] Viktor A. Adalsteinsson, Gavin Ha, Samuel S. Freeman, Atish D. Choudhury, Daniel G. Stover, Heather A. Parsons, Gregory Gydush, Sarah C. Reed, Denisse Rotem, Justin Rhoades, Denis Loginov, Dimitri Livitz, Daniel Rosebrock, Ignaty Leshchiner, Jaegil Kim, Chip Stewart, Mara Rosenberg, Joshua M. Francis, Cheng-Zhong Zhang, Ofir Cohen, Coyin Oh, Huiming Ding, Paz Polak, Max Lloyd, Sairah Mahmud, Karla Helvie, Margaret S. Merrill, Rebecca A. Santiago, Edward P. O'Connor, Seong H. Jeong, Rachel Leeson, Rachel M. Barry, Joseph F. Kramkowski, Zhenwei Zhang, Laura Polacek, Jens G. Lohr, Molly Schleicher, Emily Lipscomb, Andrea Saltzman, Nelly M. Oliver, Lori Marini, Adrienne G. Waks, Lauren C. Harshman, Sara M. Tolaney, Eliezer M. Van Allen, Eric P. Winer, Nancy U. Lin, Mari Nakabayashi, Mary-Ellen Taplin, Cory M. Johannessen, Levi A. Garraway, Todd R. Golub, Jesse S. Boehm, Nikhil Wagle, Gad Getz, J. Christopher Love, and Matthew Meyerson. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*, 8(1):1324, November 2017. Publisher: Nature Publishing Group.

[2] Sylvan C. Baca, Ji-Heui Seo, Matthew P. Davidsohn, Brad Fortunato, Karl Semaan, Shahabbedin Sotudian, Gitanjali Lakshminarayanan, Miklos Diossy, Xintao Qiu, Talal El Zarif, Hunter Savignano, John Canniff, Ikenna Madueke, Renee Maria Saliby, Ziwei Zhang, Rong Li, Yijia Jiang, Len Taing, Mark Awad, Cindy H. Chau, James A. DeCaprio, William D. Figg, Tim F. Greten, Aaron N. Hata, F. Stephen Hodi, Melissa E. Hughes, Keith L. Ligon, Nancy Lin, Kimmie Ng, Matthew G. Oser, Catherine Meador, Heather A. Parsons, Mark M. Pomerantz, Arun Rajan, Jerome Ritz, Manisha Thakuria, Sara M. Tolaney, Patrick Y. Wen, Henry Long, Jacob E. Berchuck, Zoltan Szallasi, Toni K. Choueiri, and Matthew L. Freedman. Liquid biopsy epigenomic profiling for cancer subtyping. *Nature Medicine*, 29(11):2737–2741, November 2023. Publisher: Nature Publishing Group.

[3] Fenglong Bie, Zhijie Wang, Yulong Li, Wei Guo, Yuanyuan Hong, Tiancheng Han, Fang Lv, Shunli Yang, Suxing Li, Xi Li, Peiyao Nie, Shun Xu, Ruochuan Zang, Moyan Zhang, Peng Song, Feiyue Feng, Jianchun Duan, Guangyu Bai, Yuan Li, Qilin Huai, Bolun Zhou, Yu S. Huang, Weizhi Chen, Fengwei Tan, and Shugeng Gao. Multimodal analysis of cell-free DNA whole-methylome sequencing for cancer detection and localization. *Nature Communications*, 14(1):6042, September 2023. Publisher: Nature Publishing Group.

[4] Daniel C. Bruhm, Dimitrios Mathios, Zachariah H. Foda, Akshaya V. Annapragada, Jamie E. Medina, Vilmos Adleff, Elaine Jiayuee Chiao, Leonardo Ferreira, Stephen Cristiano, James R. White, Sarah A. Mazzilli, Ehab Billatos, Avrum Spira, Ali H. Zaidi, Jeffrey Mueller, Amy K. Kim, Valsamo Anagnostou, Jillian Phallen, Robert B. Scharpf, and Victor E. Velculescu. Single-molecule genome-wide mutation profiles of cell-free DNA for non-invasive detection of cancer. *Nature Genetics*, 55(8):1301–1310, August 2023. Publisher: Nature Publishing Group.

[5] M. Cisneros-Villanueva, L. Hidalgo-Pérez, M. Rios-Romero, A. Cedro-Tanda, C. A. Ruiz-Villavicencio, K. Page, R. Hastings, D. Fernandez-Garcia, R. Allsopp, M. A. Fonseca-Montaño, S. Jimenez-Morales, V. Padilla-Palma, J. A. Shaw, and A. Hidalgo-Miranda. Cell-free DNA analysis in current cancer clinical trials: a review. *British Journal of Cancer*, 126(3):391–400, February 2022. Publisher: Nature Publishing Group.

[6] Pin Cui, Xiaozhou Zhou, Shu Xu, Weihuang He, Guozeng Huang, Yong Xiong, Chuxin Zhang, Tingmin Chang, Mingji Feng, Hanming Lai, and Yi Pan. Prediction of methylation status using WGS data of plasma cfDNA for multi-cancer early detection (MCED). *Clinical Epigenetics*, 16(1):34, February 2024.

[7] Hao Feng and Hao Wu. Differential methylation analysis for bisulfite sequencing using DSS. *Quantitative Biology (Beijing, China)*, 7(4):327–334, December 2019.

[8] Liantao Guo, Deguang Kong, Jianhua Liu, Ling Zhan, Lan Luo, Weijie Zheng, Qingyuan Zheng, Chuang Chen, and Shengrong Sun. Breast cancer heterogeneity and its implication in personalized precision therapy. *Experimental Hematology & Oncology*, 12(1):3, January 2023.

[9] Simon Heeke, Carl M. Gay, Marcos R. Estecio, Hai Tran, Benjamin B. Morris, Bingnan Zhang, Ximing Tang, Maria Gabriela Raso, Pedro Rocha, Siqi Lai, Edurne Arriola, Paul Hofman, Veronique Hofman, Prasad Kopparapu, Christine M. Lovly, Kyle Concannon, Luana Guimaraes De Sousa, Whitney Elisabeth Lewis, Kimie Kondo, Xin Hu, Azusa Tanimoto, Natalie I. Vokes, Monique B. Nilsson, Allison Stewart, Maarten Jansen, Ildikó Horváth, Mina Gaga, Vasileios Panagoulias, Yael Raviv, Danny Frumkin, Adam Wasserstrom, Aharona Shuali, Catherine A. Schnabel, Yuanxin Xi, Lixia Diao, Qi Wang, Jianjun Zhang, Peter Van Loo, Jing Wang, Ignacio I. Wistuba, Lauren A. Byers, and John V. Heymach. Tumor- and circulating-free DNA methylation identifies clinically relevant small cell lung cancer subtypes. *Cancer Cell*, 42(2):225–237.e5, February 2024. Publisher: Elsevier.

[10] Ja-Lok Ku, You-Kyung Jeon, and Jae-Gahb Park. Methylation-specific PCR. *Methods in Molecular Biology (Clifton, N.J.)*, 791:23–32, 2011.

[11] Norbert Moldovan, Ymke van der Pol, Tom van den Ende, Dries Boers, Sandra Verkuijlen, Aafke Creemers, Jip Ramaker, Trang Vu, Sanne Bootsma, Kristiaan J. Lenos, Louis Vermeulen, Marieke F. Fransen, Michiel Pegtel, Idris Bahce, Hanneke van Laarhoven, and Florent Mouliere. Multi-modal cell-free DNA genomic and fragmentomic patterns enhance cancer survival and recurrence analysis. *Cell Reports Medicine*, 5(1):101349, January 2024.

[12] Lucas Penny, Sasha C. Main, Steven D. De Michino, and Scott V. Bratman. Chromatin- and nucleosome-associated features in liquid biopsy: implications for cancer biomarker discovery. *Biochemistry and Cell Biology*, 102(4):291–298, August 2024. Publisher: NRC Research Press.

[13] Eleni Polatoglou, Zsuzsanna Mayer, Vida Ungerer, Abel J. Bronkhorst, and Stefan Holdenrieder. Isolation and Quantification of Plasma Cell-Free DNA Using Different Manual and Automated Methods. *Diagnostics*, 12(10):2550, October 2022.

[14] Ronen Sadeh, Israa Sharkia, Gavriel Fialkoff, Ayelet Rahat, Jenia Gutin, Alon Chappleboim, Mor Nitzan, Ilana Fox-Fisher, Daniel Neiman, Guy Meler, Zahala Kamari, Dayana Yaish, Tamar Peretz, Ayala Hubert, Jonathan E. Cohen, Azzam Salah, Mark Temper, Albert Grinshpun, Myriam Maoz, Samir Abu-Gazala, Ami Ben Ya'acov, Eyal Shteyer, Rifaat Safadi, Tommy Kaplan, Ruth Shemer, David Planer, Eithan Galun, Benjamin Glaser, Aviad Zick, Yuval Dor, and Nir Friedman. ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nature Biotechnology*, 39(5):586–598, May 2021. Publisher: Nature Publishing Group.

[15] Benjamin L. Spector, Lauren Harrell, Drinnan Sante, Gerald J. Wyckoff, and Laurel Willig. The methylome and cell-free DNA: current applications in medicine and pediatric disease. *Pediatric Research*, 94(1):89–95, July 2023. Publisher: Nature Publishing Group.

[16] Mary L. Stackpole, Weihua Zeng, Shuo Li, Chun-Chi Liu, Yonggang Zhou, Shanshan He, Angela Yeh, Ziye Wang, Fengzhu Sun, Qingjiao Li, Zuyang Yuan, Asli Yildirim, Pin-Jung Chen, Paul Winograd, Benjamin Tran, Yi-Te Lee, Paul Shize Li, Zorawar Noor, Megumi Yokomizo, Preeti Ahuja, Yazhen Zhu, Hsian-Rong Tseng, James S. Tomlinson, Edward Garon, Samuel French, Clara E. Magyar, Sarah Dry, Clara Lajonchere, Daniel Geschwind, Gina Choi, Sammy Saab, Frank Alber, Wing Hung Wong, Steven M. Dubinett, Denise R. Aberle, Vatche Agopian, Steven-Huy B. Han, Xiaohui Ni, Wenyuan Li, and Xianghong Jasmine Zhou. Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer. *Nature Communications*, 13(1):5566, September 2022. Publisher: Nature Publishing Group.

[17] Christoffer Trier Maansson, Peter Meldgaard, Magnus Stougaard, Anders Lade Nielsen, and Boe Sandahl Sorensen. Cell-free chromatin immunoprecipitation can determine tumor gene expression in lung cancer patients. *Molecular Oncology*, 17(5):722–736, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1878-0261.13394.

[18] Lijing Zhang and Jinming Li. Unlocking the secrets: the power of methylation-based cfDNA detection of tissue damage in organ systems. *Clinical Epigenetics*, 15:168, October 2023.