Cell-free DNA, or 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). [2] 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. [3]
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. [3] We specifically collect data from breast cancer patients, as breast cancer is a highly heterogenous disease with several subtypes that have distinct genetic and epigenetic profiles. Using their cfDNA genetic and epigenetic profiles, we apply machine learning models to predict breast cancer subtypes and infer regulatory mechanisms underlying these subtypes.

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.[5] By analyzing the methylation patterns of cfDNA, we can infer the tissue of origin of the tumor and apply this towards subtyping a tumor.

1. **Apply cfMethyl-seq [6] 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 2000 breast cancer patients with various subtypes will be used.

2. **Extract GEMINI-methylation features by comparing hyper-methylated regions in tumor vs. normal samples and compare with GEMINI mutation profiles.** Using the approach of GEMINI, we can calculate tumor-normal matched methylation scores over nonoverlapping genomic bins. We then will evaluate basic machine learning models in predicting cancer subtypes based on GEMINI-methylation features, GEMINI-mutation features, and a combination of both.

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

Different cancer subtypes have distinct gene expression profiles. Combining cfDNA histone modification data, which is associated with gene expression programs of their cells of origin [4, 7], with mutation profiles can provide a more comprehensive view of the gene expression landscape in a tumor.

1. **Collect cfCHiP-seq [1] and sequence a breast cancer cohort to generate histone modification signals and mutation profiles.** We will use a cohort of 2000 patients with various cancer types and subtypes to generate H3K4me3 and H3K27ac signals. H3K4me3 and H3K27ac are histone modifications that are associated with promoters and enhancers, respectively.

2. **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.

3. **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 basic 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 approach will provide a comprehensive view of the relationship between somatic mutations and regulatory properties of DNA, specifically methylation patterns and histone modifications. While many studies have focused on either genetic or epigenetic cfDNA samples, there is largely a lack of studies that utilize matched genetic and epigenetic data for the same samples. Using enriched cfDNA samples, we can generate paired genetic and epigenetic data that will help understand the correspondence between genetic mutations and the 3D landscape of DNA. This will help improve our ability in several cancer-related prediction tasks, such as predicting tumor behavior, tumor growth, therapy response, and metastatic potential. The proposed approach may also motivate further work in combining other omes of cfDNA, such as fragmentomics and nucleosomics.

Identification of significant regulatory regions

Regulatory regions are important for understanding driver mutations and gene expression programs in cancer. 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.

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. Identifying these differences can help in the development of more targeted therapies.

Advancing non-invasive liquid biopsy diagnostics

The proposed approach will revolutionize cancer diagnostics by enabling the classification of a patient's subtype through a fast and non-invasive blood test. For a heterogeneous cancers like breast cancer, this will be particularly impactful as it will allow for more 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.

Innovation

By combining the analyses of multiple omes of cfDNA, we are able to conduct multiple assays on the same sample, which is a novel approach in the field of cfDNA analysis. Our workflow in associating methylation, mutation, and histone modification data is applied towards a heterogeneous cancer cohort, however this approach can be extended to other cancer types and diseases. We also envision the potential for our machine learning models to be integrated into clinical workflows, enabling real-time predictions of cancer subtypes. Future work that adopts our approach may also consider assessing epigenetic and mutational signature changes over time, which can provide insights into tumor evolution and monitoring of treatment response.

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

Aim 1A: Apply cfMethyl-seq to a breast cancer cohort to capture genome-wide methylation profiles of CpG-rich regions and mutation profiles

Rationale

Breast cancer is highly heterogenous, where epigenetic changes, such as methylation, play a critical role in tumor development, progression, and subtype differentiation. cfMethyl-seq offers a non-invasive way to profile genome-wide methylation patterns in circulating tumor DNA. Using this method allows to simultaneously acquire genetic and methylation signals in a single workflow. Detailed mutation and methylation profiles will help identify genetic and epigenetic biomarkers associated with each cancer subtype.

Methods

Sequencing a 2000 patient breast cancer cohort: We will select a cohort of 2000 breast cancer patients spanning the four major subtypes: Luminal A, Luminal B, HER2-enriched, and Triple-negative. For each patient, we will draw blood samples and extract cfDNA using QIAamp Circulating Nucleic Acid Kit, a protocol shown to have high yield and quality of cfDNA. We will then tag each cfDNA molecule with a unique molecule identifier (UMI) to match corresponding cfMethyl-seq data with WGS data. Using ichorCNA, we will then estimate the tumor fraction in each sample and filter out samples with low tumor fraction. We also validate the concentration and quality of cfDNA using Bioanalyzer 2100. Samples are then split for cfMethyl-seq and next-generation whole-genome sequencing. Standard library preparation protocols are used for both assays.

Generation of GEMINI-methylation and GEMINI-mutation features: We apply the GEMINI method to calculated tumor-normal matched methylation and mutation scores over nonoverlapping genomic bins. The generation of methylation and mutation features is defined as:

$$M_i = \frac{\sum y_{ij}^T}{\sum x_{ij}^T} - \frac{\sum y_{ij}^N}{\sum x_{ij}^N}, \quad G_i = \frac{\sum z_{ij}^T}{\sum x_{ij}^T} - \frac{\sum z_{ij}^N}{\sum x_{ij}^N} \tag{1}$$

where M_i and G_i are the methylation and mutation scores for bin i, respectively. y_{ij}^T and y_{ij}^N are the number of methylated reads in tumor and normal samples, respectively, for bin i. x_{ij}^T and x_{ij}^N are the total number of reads in tumor and normal samples, respectively, for bin i. z_{ij}^T and z_{ij}^N are the number of mutations in tumor and normal samples, respectively, for bin i. cfMethyl-seq only captures CpG-rich regions, thus most of the bins may not have any methylation signal. Regardless, having the same binned representation for both methylation and mutation data allows for direct comparison between the two modalities.

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. Furthermore, we can explore the association of methylation events with specific genetic mutations, providing insights into the regulatory mechanisms underlying each subtype.

Methods

Correlation of methylation and mutation features with cancer subtypes: To determine the association of methylation and mutation features, we will calculate the correlation between all the features. Formally, for each cancer subtype $c \in \{1, 2, 3, 4\}$, we calculate the Pearson correlation between each methylation feature M_i and mutation feature G_i with the subtype label y_c.

Unsupervised clustering of methylation and mutation features: Let M be the matrix of methylation

features, of shape $n_M \times m$, where n_M is the number of samples with methylation signals and m is the number of bins. Similarly, let G be the matrix of mutation features, of shape $n_G \times m$, where n_G is the number of samples with mutation signals. To visualize groupings of samples based on these profiles, we apply UMAP to matrix M and G to reduce the dimensionality of the data to a 2D space. We can then plot the UMAP embeddings and color the points by tumor vs. normal status to see if there is a clear separation between the two groups. After filtering out normal samples, we can then color the points by cancer subtype to see if there is a clear separation between the subtypes. Finally, we can apply a similar analysis to the combined methylation and mutation features to see if the two modalities provide complementary information in distinguishing between the subtypes.

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 first will apply a simple multinomial logistic regression model to predict the subtypes based on the methylation and mutation features, which will serve as a baseline model. The benefit of using a logistic regression model is that it is interpretable and can provide insights into the importance of each feature in predicting the subtypes. We will then evaluate more complex models, including random forests and neural networks to see if they can improve the predictive performance.

Expected Results and Interpretation

We expect to find the methylation and mutation signals to have significant differences between the tumor and normal samples. This is expected as tumor samples are known to have hyper-methylation events and higher mutation frequencies compared to normal samples. We also expect to see differences in the methylation and mutation signals between the different cancer subtypes. Different subtypes of breast cancer are characterized by distinct genetic and epigenetic profiles, and we expect these differences to be reflected in the methylation and mutation signals. We also expect to see a correlation between some methylation events and specific genetic mutations. Some mutations, particularly in promoter and enhancer regions, affect the methylation patterns of nearby CpG sites. For the unsupervised clustering analysis, we expect to see a fairly clear separation between the tumor and normal samples, and a fairly clear separation between the different cancer subtypes. Because different subtypes are characterized by different genetic and epigenetic profiles, we expect these differences to be reflected in the UMAP embeddings for the corresponding feature sets. For the supervised classification task, we hypothesize that combining the methylation and mutation features will improve the predictive performance because the two modalities should provide complementary information in categorizing the subtypes.

Potential Pitfalls and Alternative Approaches

For Aim 1A, 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 Kit because it has been shown to have one of the highest yields of cfDNA compared to other methods. 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 signalprior to sequencing and the generation of GEMINI-methylation and GEMINI-mutation features. For Aim 1B, we may observe a lack of clear separation between the different cancer subtypes and consequently, the models may not perform well in predicting the subtypes. This could be due to overlapping biological characteristics, noise in cfDNA data, or the lack of sufficient discriminatory power in either feature set. Furthermore, the combined methylation and mutation features may introduce redundancy or noise, potentially degrading model performance instead of improving it. This could result in both unsupervised clustering and supervised classification failing to clearly differentiate subtypes. 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-methylation and GEMINI-mutation features to reduce noise and redundancy.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.