**Machine Learning Approaches to Determine Variant Origin in Liquid Biopsies**

**Jenna Canzoniero1,Daniel Rabizadeh, Archana Balan1, Jillian Phallen1, Blair V. Landon1, Lavanya Sivapalan1, Benjamin Green1, Zineb Belcaid1, Susan Scott1, Gavin Pereira1, Vincent Lam1, Christine L. Hann1, Wade T. Iams2, Christine M. Lovly2, Patrick M. Forde1,3, Victor E. Velculescu1,3, Robert B. Scharpf1,4 and Valsamo Anagnostou1,3\***

1The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

2 Division of Hematology-Oncology, Department of Medicine, Vanderbilt University Medical Center and Vanderbilt-Ingram Cancer Center, Nashville, TN, USA

3The Bloomberg-Kimmel Institute for Cancer Immunotherapy, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA­­­

4Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

**\* To whom correspondence should be addressed:**

Valsamo Anagnostou, MD, PhD

CRB2, Rm546, 1550 Orleans Street

Baltimore, MD 21287

Phone: 410-614-8948

Email: [vanagno1@jhmi.edu](mailto:vanagno1@jhmi.edu)

**Abstract**

**Introduction**: Targeted next-generation sequencing (NGS) of cell-free DNA in plasma, referred to as liquid biopsy, has become a valuable diagnostic tool in clinical oncology. However, detection of variants related to clonal hematopoiesis (CH) is a major confounder that significantly impairs the clinical utility of liquid biopsies. Here we developed a machine-learning model to determine tumor versus CH origin of variants identified in plasma-only NGS.

**Methods**: We assembled a training cohort of 426 variants identified by targeted deep plasma sequencing from 225 patients with stage I-IV breast, colorectal, esophageal, lung, and ovarian cancer, coupled with matched white blood cell (WBC) and tumor tissue NGS to allow determination of the reference origin for each plasma variant. We employed Extreme Gradient Boosting (XGBoost) to integrate mutant fragment, variant, gene, and patient level features to predict tumor versus CH plasma variant origin, evaluating the performance of this approach within the training cohort using 10-fold cross-validation. To externally validate, we applied the fixed model to an independent validation cohort 409 variants detected from 74 patients with breast, colorectal, and prostate cancer.

**Results**: Using serial plasma samples, we identified stable statistical measures of differences in fragment feature distributions between mutant and wild type fragments; these were subsequently incorporated into an XGBoost machine-learning model along with variant, gene and patient features to predict cellular origin of variants. Our model predicted variant origin with an AUC of 0.93788 (95% CI 0.85288-1) from 10-fold cross validation in the training cohort. The fixed model in the independent validation cohorts; the fixed model predicted plasma variant origin with accuracy 80.2% (p<2e-16 compared to null information rate=0.6112).

**Conclusions**: We developed a machine-learning model that integrates patient, gene, variant and fragment features to predict tumor versus CH origin of plasma variants across solid tumors and NGS sequencing platforms. The ability to identify bona fide tumor variants in plasma-only sequencing fills a critical need in the clinical implementation of liquid biopsy-guided cancer therapy by reducing misinterpretation due to CH contamination.

**Main**

Analysis of cell-free DNA (cfDNA) from liquid biopsy samples has emerged as an efficient, non-invasive method for multiple purposes including the early detection of cancers, tracking tumor heterogeneity and evolutionary dynamics to monitor disease progression [1], and detecting minimal residual disease to assess response to therapy [2,3]. cfDNA consists of short, double stranded DNA fragments released into circulation from cells undergoing apoptosis or necrosis. In healthy individuals, the majority of cfDNA is derived from hematopoietic cells with a small percentage derived from other organs [4, 5]. In cancer patients, a very small fraction of the cfDNA is also tumor derived (ctDNA). Due to the high percentage of hematopoietic and typically low concentrations of ctDNA within the population of cfDNA, a major challenge to analyzing liquid biopsy samples is to accurately determine the origin of the variants detected and avoid misclassification of clonal hematopoiesis (CH) derived variants as tumor derived [6, 7].

~~Clonal hematopoiesis of indeterminate potential (CHIP) is a phenomenon when a hematopoietic cell acquires a benign genetic mutation when dividing. These mutations often occur in leukemia-associated driver genes and may confer a fitness advantage to the cell, resulting in its disproportionate propagation and accumulation of the CH mutation~~ [Clonal hematopoiesis of indeterminate potential (CHIP) is a phenomenon where a hematopoietic cell acquires a benign genetic mutation in leukemia-associated driver genes, leading to a fitness advantage that results in the disproportionate propagation and accumulation of cells with the CH mutation.][8, 11]. Since the majority of cfDNA is derived from hematopoietic cells, these CHIP derived mutations may confound the interpretation of the cfDNA assay [9, 10]. Depending on the age and health condition of the patient, more than half of the mutations found in plasma cfDNA can also be detected in WBC sequencing. Furthermore, these CHIP mutations frequently occur in cancer-associated genes such as TP53, KRAS, or JAK2 [11, 12].

The misidentification of CHIP mutations as tumor derived can significantly impair the use of liquid biopsies for clinical prognosis such as in cancer screenings, selecting treatment options, or detection of minimal residual disease and monitoring treatment response. As mentioned above, healthy individuals derive the majority of their cfDNA from hematopoietic cells. If the patient is in early stages of cancer, the amount of ctDNA will be very small, and some CH mutations may be misclassified as tumor mutations, leading to an unreliable diagnosis [10]. Similarly, the consistent detection of CH mutations pre and post treatment may be incorrectly interpreted as presence of minimal residual disease and ineffective treatment response [7]. A recent study in patients with resectable gastric cancer has shown that without the removal of CH derived variants, ctDNA analysis could not predict risk of recurrence or survival post-operation [13]. A similar study in patients with small cell lung cancer also found that without eliminating CH variants [from the cfDNA variant calls], tracking circulating tumor burden and molecular response could not predict overall or progression free survival [14]. In order for ctDNA analysis to be meaningful, it is crucial to accurately differentiate between tumor and CH derived variants.

One way to circumvent this issue is to use a tumor-informed approach, which relies on the genotyping of the tumor tissue to correctly identify tumor derived variants. However, this method reintroduces the issue of needing an adequate tissue sample, which may be difficult to obtain and may not capture the genetic heterogeneity of the cancer. Another approach utilizes parallel sequencing of patient-matched WBCs to filter out CH variants from the cfDNA results, and can achieve comparable results to tumor-informed methods [15]. Although this approach bypasses the requirement for a tumor biopsy, extra time and costs are still needed to extract, sequence, and analyze the WBC DNA. Furthermore, most commercial targeted sequencing panels do not include matched WBC sequencing, which may result in a high margin of error when interpreting cfDNA data. As a result, there is a critical need to devise a strategy capable of differentiating ctDNA variants from biological background noise independent of tumor tissue and WBC sequencing analysis. This will not only improve the use of cfDNA for clinical prognosis, but also will allow retrospective studies to be performed on plasma samples without matched WBC sequencing.

The analysis of cfDNA fragmentation patterns, also known as fragmentomics, presents a potential way of determining variant origin without the use of parallel sequencing of matched tumor or WBC samples. Many studies have shown that when cfDNA is released into circulation, it retains certain fragmentomic and epigenomic features that are unique to its specific cell type due to the different release pathways and enzymes involved. Therefore, it is plausible that the origin of different variants can be determined based on the epigenetic features and fragmentomic patterns of the DNA fragment it is found on. Such features can include different fragment lengths, end motifs, and nucleosome footprints. [16, 17]. Numerous studies have consistently shown that ctDNA fragments tend to have shorter lengths than wild-type (WT) fragments [18, 19, 20]. Furthermore, fragments containing CH variants are similar in length to WT fragments, in contrast to the shorter fragments associated with tumor DNA [20, 21]. Other studies have identified particular sets of DNA end motifs and signature nucleosome footprints that can distinguish cfDNA molecules originating from hematopoietic cells compared to other types [22, 23]. Taken together, these patterns offer a promising methodology for distinguishing CH variants from tumor variants without need ~~of~~ [for] tumor ~~of~~ [or] WBC sequencing.

Here, we have developed a novel machine-learning model that accurately identifies tumor origin variants from CH associated variants using the fragmentomic features listed above. By integrating fragmentomic, variant, and patient level features, we can predict tumor vs. CH variant origin using plasma-only sequencing without additional WBC or tumor sequencing, which will enhance the precision and reliability of utilizing ctDNA for clinical prognosis purposes.

**Results**

**Creation of training cohort**

We hypothesized that integration of fragmentomic, variant, and patient level features into a machine learning model would enable accurate prediction of tumor versus CH origin for variants identified in plasma-only sequencing (Figure 1a). In order to develop this approach, we assembled sequencing data from 225 patients with stage I-IV breast, colorectal, esophageal, ovarian, non-small cell (NSCLC) and small-cell lung (SCLC) cancer, all of whom had undergone plasma cfDNA and WBC DNA sequencing, and 67.6% had undergone tumor tissue sequencing. By comparing across matched sequencing samples, we were able to identify a reference origin for 426 plasma variants, which became the training cohort (Figure 1b). Tumor origin variants had statistically significant higher variant allele frequency (VAF) compared to CH-origin variants (mean 0.084 versus 0.023, Welch two sample t-test p-value = 4.634e-07, Extended Data Fig.1). However, given significant overlap in distribution, VAF alone was a poor predictor of variant origin with AUC 0.61433 (95% confidence interval 0.55979 - 0.66886, Supplemental figure 1b).

**Identification of serially stable fragmentation statistics**

In order to facilitate the incorporation of fragmentomics into the development of machine learning model, we sought to identify summary statistics that could be used to describe fragmentation patterns in mutant versus non-mutant reads, regardless of variant origin. Both fragment length and location of fragment endpoints. Figures 2a-c show illustrative examples of potential fragmentation patterns. In figure 2a, which is CH-derived variant, the distribution of read length for mutant fragments overlaps that of wild-type fragments, and the location of fragment end-points relative to the location of the mutation is similar. In figure 2b, which is a tumor-derived variant, mutant read length is shorter than wild-type, resulting in a left-shift of the length cumulative density function and a shifting of the relative endpoint peaks closer together. In figure 2c, which is also a tumor-derived variant, mutant read length is similar to wild-type, as seen in overlapping length cumulative density, but the relative endpoint locations for mutant reads are shifted. Thus in order to capture multiple possible differences in fragmentation patterns, we needed to incorporate multiple summary statistics.

Applying the assumption [Given/ With the premise/Reasoning that ] that fragmentation summary statistics that were representative of biology would be consistent over time, we examined the stability of different fragmentation summary statistics in a cohort of variants that were detected in patients at baseline and follow-up timepoints. This serial cohort included 41 variants found in 29 patients with stage I-III colorectal, esophageal, or NSCLC who had serial cfDNA plasma sequencing performed while undergoing treatment as part of clinical trials (Figure 2c). Summary statistics were calculated by comparing mutant versus wild-type reads for a given patient-locus and did not take into account variant origin. Length statistics included F-statistic ratio of variances and T-statistic. Other statistics calculated related to length, endpoint location, and endpoint 4-mer motif are described in supplemental information. Each statistical value was calculated by comparing wild-type to mutant reads for each patient-locus at each time point independently. We then calculated the correlation of the statistical values between baseline and follow-up time points (Extended data fig 1)). Notably, the F-statistic and T-statistic were not well correlated over time (Figure 2d). Fragmentation summary statistics that were significantly correlated using Pearson correlation over time were then incorporated as features into the machine learning model (Methods).

**Model development and performance in the training cohort**

We then combined the fragmentomic features identified as stable over time in the serial cohort with variant and patient level features to develop a machine learning model to predict tumor vs CH origin. Variant level features included the variant allele frequency, the scaled variant allele frequency relative to other variants identified in a patient’s plasma, and variant and gene annotations of the reported hematologic versus solid tumor frequency in COSMIC. The patient level feature was age. As seen in the heatmap in Figure 3a, no single feature clearly distinguished plasma variant origin. However, when integrated together into a machine learning model using XGboost, we were able to predict variant origin with an AUC 0.93788 (95% confidence interval 0.85288-1, 10-fold cross validation repeated 10 times, Figure 3b). Model feature importance analysis revealed that the top five most important features included features from all three levels - fragment, variant, and patient (Figure 3c). The best performing model was locked.

**Model performance in the independent validation cohort**

In order to further assess our model in an independent validation cohort, we obtained data related to 2019 publication by Razavi et al. This cohort included patients with metastatic breast, prostate and NSCL cancer who had undergone sequencing of plasma, WBC, and tumor tissue to generate reference origin. We were able to obtain the plasma sequencing files and calculate model features for ~~409~~ [1,412] variants found in 74 patients (Figure 4a – Do we need to recreate this figure?). Again, as seen in the heatmap in Figure 4b, no single feature clearly distinguished plasma variant origin. [Similarly to the training cohort, no single feature was sufficient to distinguish the plasma variant origin by itself (Figure 4b). ] When we applied our fixed machine learning model, we were able to predict plasma variant origin with accuracy of 80.2% (p-value <2e-16 compared to null information rate of 0.6112). Accuracy across different tumor subtypes is shown in Table 1.

We then examined accuracy based on gene or gene set. As mutations in *TP53* are commonly found in both solid tumor and CHIP, making this gene particularly difficult for determining variant origin, we examined model performance in *TP53* variants (Figure 3c). For the 36 identified *TP53* mutations, the model accurately predicted origin 86.11% (p-value 0.083 based on null information rate of 0.75). Model performance in other gene sets including all gene except *DMT3A*, genes associated with CHIP, and genes where the reference origin was split between tumor and CH-derived is shown in Table 2.

**Longitudinal cohort**

**@@@ Cairo 5**

**Acknowledgements**

This work was supported in part by the US National Institutes of Health grants CA121113 (V.A. and V.V.), the Bloomberg-Kimmel Institute for Cancer Immunotherapy (V.A and P.M.F.), the Johns Hopkins SKCCC core grant NCI CCSG P30 CA006973, the Department of Defense Congressionally Directed Medical Research Programs grant CA190755 (V.A. and P.M.F.), the ECOG-ACRIN Thoracic Malignancies Integrated Translational Science Center grant UG1CA233259 (V.A., V.V.), the V Foundation (V.A. and V.V.) and the LUNGevity Foundation (V.A. and V.V.).

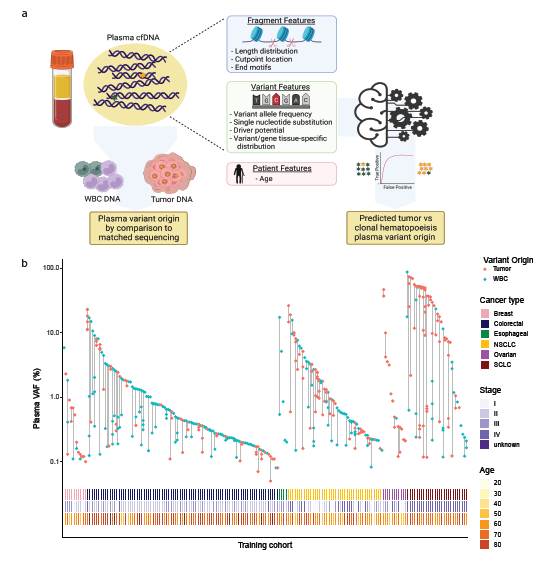
**Author Contributions**

V.A and J.C. conceived the study and contributed to the study design, data analysis, interpretation, and writing. A.B. contributed to data analysis.

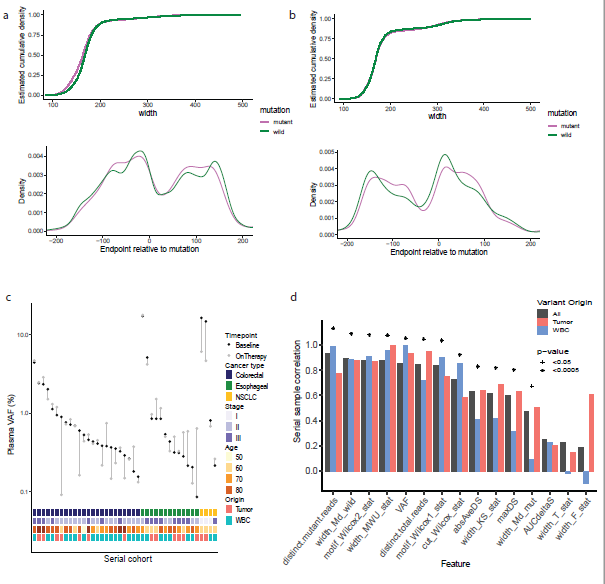
**Competing Interests**

**V.A.** receives research funding to Johns Hopkins University from Astra Zeneca, Personal Genome Diagnostics/Labcorp and Delfi Diagnostics and has received research funding to Johns Hopkins University from Bristol-Myers Squibb in the past 5 years. V.A is an inventor on patent applications (63/276,525, 17/779,936, 16/312,152, 16/341,862, 17/047,006 and 17/598,690) submitted by Johns Hopkins University related to cancer genomic analyses, ctDNA therapeutic response monitoring and immunogenomic features of response to immunotherapy that have been licensed to one or more entities. Under the terms of these license agreements, the University and inventors are entitled to fees and royalty distributions. **V.L.** has served in a consultant/advisory role for Takeda, Seattle Genetics, Bristol-Myers Squibb, AstraZeneca and Guardant Health and has received research funding from GlaxoSmithKline, BMS, Merck and Seattle Genetics. **P.M.F.** has received research funding to Johns Hopkins University from AstraZeneca, Bristol-Myers Squibb, Novartis, Corvus and Kyowa. He has also served as a consultant for Amgen, AstraZeneca, Bristol-Myers Squibb, Daiichi Sankyo, Genentech, G1 Therapeutics, Iteos, Janssen, Merck, Surface Oncology, Mirati, Novartis, Sanofi and as a DSMB member for Polaris and Flame Therapeutics. **V.E.V.** is a founder of Delfi Diagnostics, serves as on the Board of Directors and as a consultant for this organization, and owns Delfi Diagnostics stock, which is subject to certain restrictions under university policy. Additionally, Johns Hopkins University owns equity in Delfi Diagnostics. V.E.V. divested his equity in Personal Genome Diagnostics (PGDx) to LabCorp in February 2022. V.E.V. is an inventor on patent applications submitted by Johns Hopkins University related to cancer genomic analyses and cell-free DNA for cancer detection that have been licensed to one or more entities, including Delfi Diagnostics, LabCorp, Qiagen, Sysmex, Agios, Genzyme, Esoterix, Ventana and ManaT Bio. Under the terms of these license agreements, the University and inventors are entitled to fees and royalty distributions. V.E.V. is an advisor to Danaher, Takeda Pharmaceuticals, and Viron Therapeutics. All other authors declare no conflicts of interest.

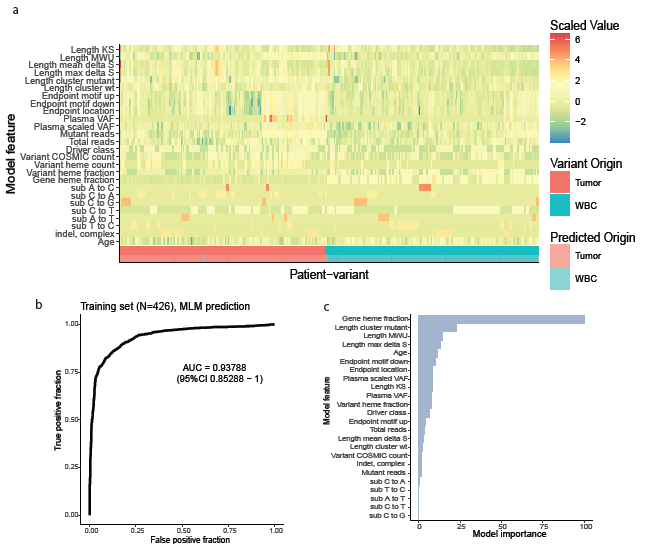
**Figures**



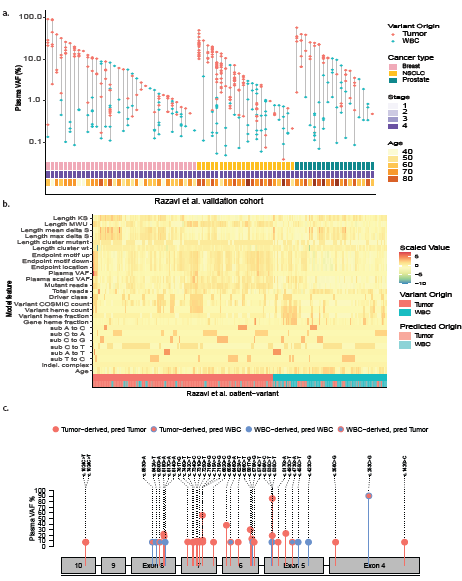
**Figure 1. Development of a machine learning algorithm to predict CH versus tumor origin for variants detected in plasma.** a. Schematic illustrating algorithm development. For variants detected in plasma, fragment features are extracted from sequencing data and combined with variant features and patient features into a machine learning model that predicts if each variant is tumor or CH derived. Model performance is assessed by comparison to the reference plasma variant origin which is determined by comparison to matched WBC and tumor tissue sequencing. b. Training cohort of 426 variants identified in 225 patients with stage I-IV breast, colorectal, esophageal, NSCLC, ovarian and SCLC. Each column represents a patient, with plasma VAF along the y-axis and points color-coded by reference variant origin (tumor vs WBC).



**Figure 2. Assessment of fragmentomic features.** a. Illustrative example of fragmentomic features from a patient with a tumor-derived plasma variant. In the top panel, fragment length density is right shifted (shorter) for mutant compared to wild-type fragments. This is also reflected in the bottom panel, where the fragment endpoints relative to mutation location (set at 0) are closer together in mutant versus wild-type fragments. b. An example from a different patient with a tumor-derived plasma variant that is not distinguishable by fragment length alone. Here, in the top panel, fragment length density is the same for both mutant and wild-type fragments. However, in the bottom panel, the locations of the relative fragment endpoints are shifted in mutant versus wild-type fragments. c. The serial cohort used to assess fragmentomic features over time. Each column represents a patient-variant, with points on the y-axis depicting VAF at baseline (black) and at follow-up (gray) time points. d. Pearson’s product moment correlation of fragmentomic feature values at baseline and follow-up timepoints. Features that were correlated over time with p-value <0.05 were subsequently used for development of the machine learning model.



**Figure 3. Machine learning model in the training cohort.** a. Heatmap with rows depicting the scaled values of features inputted into the machine learning algorithm and each column representing a single variant identified in a given patient. The second from the bottom row shows the reference variant origin where red is tumor-origin and blue is CH-origin. The bottom row shows the variant origin predicted by the machine learning algorithm. b. Receiver-operator curve (ROC) for the machine learning algorithm to predict variant origin from 10-fold cross-validation repeated 10 times. c. Feature importance in the machine learning model. AUC=area under ROC.



**Figure 4. Machine learning model performance in an independent validation cohort.** a. The independent validation cohort which was derived from Razavi et al. Each column represents a patient with variant allele frequency for variants detected in plasma in the y-axis. Plasma variants that were determined by matched sequencing to have reference origin of tumor are red, versus WBC blue. This cohort included metastatic patients with breast, NSCLC, and prostate cancer. b. For the validation cohort, heatmap with rows depicting the scaled values of features inputted into the fixed machine learning algorithm. Each column represents a single variant with reference and predictive variant origin in the second from bottom and bottom rows respectively. c. Model performance for plasma mutations found in *TP53* in the independent validation cohort. Points filled in red are tumor-derived by reference origin and filled in blue are WBC-derived by reference origin. Outlines are red if predicted tumor-derived by the machine learning model and blue if predicted WBC-derived by reference origin.

**Table 1. Performance of machine learning model in independent validation cohort by cancer type.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cancer type** | **Sensitivity** | **Specificity** | **Accuracy** | **p-value** |
| Overall  (N=409) | 0.868 | 0.6981 | 0.802 | <2e-16 |
| Breast  (N=155) | 0.8901 | 0.7031 | 0.8129 | 1.75E-09 |
| NSCLC  (N=172) | 0.8707 | 0.6786 | 0.8081 | 6.68E-05 |
| Prostate  (N=82) | 0.814 | 0.7179 | 0.7683 | 4.64E-06 |

**Table 2. Performance of machine learning algorithm in independent validation cohort by gene or gene sets.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Genes** | **Sensitivity** | **Specificity** | **Accuracy** | **p-value** |
| All excluding *DNMT3A*  (N=382) | 0.8715 | 0.6391 | 0.7906 | 2.19E-09 |
| *TP53*  (N=36) | 0.9259 | 0.6667 | 0.8611 | 0.08347 |
| CHIP-associated\*  (N=82) | 0.8919 | 0.8308 | 0.8529 | 4.64E-06 |
| Mixed origin\*\*  (N=71) | 0.9048 | 0.6897 | 0.8169 | 4.40E-05 |

\*CHIP-associated genes include *DNMT3A, TET2, ASXL1, PPM1D, TP53, JAK2, RUNX1, SF3B1, SRSF2, IDH1, IDH2, U2AF1, CBL, ATM, CHEK2*; \*\* Mixed origin genes include *TP53, ATM, KMT2D, PPM1D, GNAS*

**References**

References

1. Parikh AR, Leshchiner I, Elagina L, et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nat Med*. 2019;25(9):1415-1421. doi: 10.1038/s41591-019-0561-9 [doi]
2. Sivapalan L, Murray JC, Canzoniero JV, Landon B, Jackson J, Scott S, Lam V, Levy BP, Sausen M, Anagnostou V. Liquid biopsy approaches to capture tumor evolution and clinical outcomes during cancer immunotherapy. J Immunother Cancer. 2023 Jan;11(1):e005924. doi: 10.1136/jitc-2022-005924. PMID: 36657818; PMCID: PMC9853269.
3. Liu J, Chen X, Wang J, et al. Biological background of the genomic variations of cf-DNA in healthy individuals. *Ann Oncol*. 2019;30(3):464-470. doi: S0923-7534(19)31073-7 [pii]
4. Barbany, G, Arthur, C, Liedén, A, Nordenskjöld, M, Rosenquist, R, Tesi, B, Wallander, K, Tham, E. (Karolinska University Hospital Solna; Karolinska Institutet, Stockholm, Sweden). Cell-free tumour DNA testing for early detection of cancer – a potential future tool. *J Intern Med* (Review). 2019; 286: 118– 136.
5. Hu Z, Chen H, Long Y, Li P, Gu Y. The main sources of circulating cell-free DNA: Apoptosis, necrosis and active secretion. *Critical reviews in oncology/hematology*. 2021;157:103166. <http://dx.doi.org/10.1016/j.critrevonc.2020.103166.> doi:10.1016/j.critrevonc.2020.103166.
6. Rose Brannon, A., Jayakumaran, G., Diosdado, M. *et al.* Enhanced specificity of clinical high-sensitivity tumor mutation profiling in cell-free DNA via paired normal sequencing using MSK-ACCESS. *Nat Commun* **12**, 3770 (2021). <https://doi.org/10.1038/s41467-021-24109-5>
7. Chan HT, Nagayama S, Chin YM, et al. Clinical significance of clonal hematopoiesis in the interpretation of blood liquid biopsy. *Molecular oncology*. 2020;14(8):1719-1730. <https://pubmed.ncbi.nlm.nih.gov/32449983> [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7400786/.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7400786/) doi: 10.1002/1878-0261.12727.
8. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. *Science*. 2019;366(6465):eaan4673. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050831/.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050831/) Accessed Jun 24, 2023. doi: 10.1126/science.aan4673.
9. Abbosh C, Swanton C, Birkbak NJ. Clonal haematopoiesis: A source of biological noise in cell-free DNA analyses. *Ann Oncol*. 2019;30(3):358-359. doi: S0923-7534(19)31080-4 [pii].
10. Chan HT, Chin YM, Nakamura Y, Low S. Clonal hematopoiesis in liquid biopsy: From biological noise to valuable clinical implications. *Cancers (Basel)*. 2020;12(8):2277. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7463455/.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7463455/) Accessed Jun 24, 2023. doi: 10.3390/cancers12082277.
11. Razavi P, Li BT, Brown DN, et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat Med*. 2019;25(12):1928-1937. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7061455/.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7061455/) Accessed Jun 24, 2023. doi: 10.1038/s41591-019-0652-7
12. Hu Y, Ulrich BC, Supplee J, et al. False-positive plasma genotyping due to clonal hematopoiesis. *Clin Cancer Res*. 2018;24(18):4437-4443. doi: 10.1158/1078-0432.CCR-18-0143 [doi].
13. Leal A, van Grieken N,C.T., Palsgrove DN, et al. White blood cell and cell-free DNA analyses for detection of residual disease in gastric cancer. *Nature communications*. 2020;11(1):525. <https://pubmed.ncbi.nlm.nih.gov/31988276> [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6985115/.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6985115/) doi: 10.1038/s41467-020-14310-3.
14. Lavanya Sivapalan, Wade T. Iams, Zineb Belcaid, Susan C. Scott, Noushin Niknafs, Archana Balan, James R. White, Prasad Kopparapu, Christopher Cann, Blair V. Landon, Gavin Pereira, Victor E. Velculescu, Christine L. Hann, Christine M. Lovly, Valsamo Anagnostou; Dynamics of Sequence and Structural Cell-Free DNA Landscapes in Small-Cell Lung Cancer. Clin Cancer Res 15 June 2023; 29 (12): 2310–2323. <https://doi.org/10.1158/1078-0432.CCR-22-2242>
15. Chan HT, Nagayama S, Otaki M, Chin YM, Fukunaga Y, Ueno M, Nakamura Y and Low S-K (2023) Tumor-informed or tumor-agnostic circulating tumor DNA as a biomarker for risk of recurrence in resected colorectal cancer patients. Front. Oncol. 12:1055968. doi:10.3389/fonc.2022.1055968
16. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science*. 2021;372(6538). doi: 10.1126/science.aaw3616.
17. Oberhofer A, Bronkhorst AJ, Uhlig C, Ungerer V, Holdenrieder S. Tracing the Origin of Cell-Free DNA Molecules through Tissue-Specific Epigenetic Signatures. Diagnostics (Basel). 2022 Jul 29;12(8):1834. doi: 10.3390/diagnostics12081834. PMID: 36010184; PMCID: PMC9406971.
18. Mouliere F, Chandrananda D, Piskorz AM, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*. 2018;10(466):eaat4921. doi: 10.1126/scitranslmed.aat4921. doi: 10.1126/scitranslmed.aat4921 [doi].
19. Jiang, Peiyong, et al. “Lengthening and Shortening of Plasma DNA in Hepatocellular Carcinoma Patients.” *Proceedings of the National Academy of Sciences*, vol. 112, no. 11, 2015, https://doi.org/10.1073/pnas.1500076112.
20. Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. 2019;570(7761):385-389. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6774252/.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6774252/) Accessed Jun 24, 2023. doi: 10.1038/s41586-019-1272-6.
21. Marass F, Stephens D, Ptashkin R, et al. Fragment size analysis may distinguish clonal hematopoiesis from tumor-derived mutations in cell-free DNA. *Clinical chemistry (Baltimore, Md.)*. 2020;66(4):616-618. <https://www.ncbi.nlm.nih.gov/pubmed/32191320.> doi: 10.1093/clinchem/hvaa026.
22. Jiang P, Sun K, Peng W, et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov*. 2020;10(5):664. <http://cancerdiscovery.aacrjournals.org/content/10/5/664.abstract.> doi: 10.1158/2159-8290.CD-19-0622.
23. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*. 2016;164(0):57-68. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4715266/.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4715266/) Accessed Jun 24, 2023. doi: 10.1016/j.cell.2015.11.050.

**Methods**

*Cohorts*

The training cohort, used for refining and testing the machine learning model, consisted of 225 patients with stage I-IV breast, colorectal, esophageal, ovarian, non-small cell lung cancer or small cell lung cancer and was assembled from cohorts of patients enrolled in clinical trials or in an IRB-approved biorepository that have previously been described 1-6 . All patients were required to have matched WBC (buffy coat sequencing), and 152 patients (67.6%) also had matched tumor tissue sequencing. For patients who had serial cfDNA collections, only the baseline assessment was used in the training cohort.

The serial cohort included 29 patients with stage I-III colorectal, esophageal, or NSCLC who were undergoing treatment as part of a clinical trial. Patients were included in the serial cohort if they had two separate plasma samples with the cfDNA same variant. This cohort allowed for comparison of the same variant at separate time points and the tracking of any fluctuations of fragmentomic features throughout treatment. Of these patients, 20 patients had baseline samples that were also included in the training cohort.

The independent validation cohort included 124 patients with stage IV breast, NSCLC, or prostate cancer who had cfDNA, WBC (buffy coat) and tumor tissue obtained as part of a prospective observational study described by Razavi et al9. We included only patients that had both plasma variants with well-characterized reference origins (classified as WBC-matched, biopsy-matched, or IMPACT-bam-matched) and available age data. In total, 74 patients were included in the validation analysis. All patients in this cohort had blood drawn within 6 weeks of tumor biopsy.  No patients in this cohort were part of any of the 6 studies aggregated to create the training cohort, nor was any of the data used for informing the model parameters.

*Targeted error-correction sequencing and sequence data processing*

All the plasma cfDNA and matched buffy coat WBC samples used for the training and serial cohorts were sequenced using TEC-Seq protocol8. Mean depth of coverage for both cfDNA and WBC was 5000x, depending on the cohort the targeted panel included between 63-219 genes commonly mutated in cancer. Further details on protocol specifications, sample collection, sequencing parameters, and gene panels have been previously described. 1-6 For patients in the MEDOCC subcohort, tissue sequencing was performed using a targeted panel of cancer-associated genes at an average depth of 750x. For the remaining patients with available matched tumor tissue, whole exome sequencing was conducted on both tumor and normal tissue to a mean depth of 100x as previously described1-6. A total of 426 variants with definitive origins were identified and met the criteria for inclusion in the training cohort.

For the independent validation cohort plasma and buff coat samples were sequenced with the high intensity targeted cfDNA assay by GRAIL using a panel of 508 genes with >60,000x depth, while tumor and matched normal samples were sequenced using the MSK-IMPACT assay.9 We retrieved the processed plasma bam files from the European Genome-phenome Archive (accession number EGAS00001003755). The cfDNA variants were extracted from supplementary information included with the original publication, which included status of variant in matched tumor and buffy samples. Only mutations with origin categorized in the original publication (classified as phenotype WBC matched, biopsy matched, or IMPACT-bam matched) were included; variants classified as uncertain origin were discarded. Only SNVs and InDels were included in the final analysis, excluding TNPs and DNPs, resulting in a total of 1,418 variants.

*Evaluation of mutation origin*

For the training and serial cohorts, we developed a custom pipeline to determine the reference origin for each variant identified in plasma sequencing by comparison to the matched WBC and tissue sequencing. Following sample, quality control, trimming and alignment we ran X (*which tools/variant callers were used*) to identify somatic variants in our plasma sample. Next, the variants were classified into one of several origin categories: Tumor, likely tumor, WBC, likely WBC, germline or unknown. The COSMIC database was used to annotate each variant and mark cancer hotspots at positions with 25 or greater occurrences on COSMIC. Variants were classified as germline, if it did not meet the threshold to be considered a hotspot, and the variant allele frequency (VAF) was greater than 25% in both the patient’s plasma and WBC samples. Variants detected in plasma were classified as WBC origin if the same mutation was detected in WBC sequencing with at least 5 supermutant families. Given that clonal hematopoiesis is closely linked as a precursor to hematological cancers, for variants annotated as hematopoietic in COSMIC, we reduced the required threshold to 2-4 supermutant families to be classified as WBC origin. Additionally, we classified variants as WBC origin if found in WBC in only 2-4 supermutant families, but not detected in available tumor tissue sequencing for positions with at least 100x coverage of the reference base. Variants were classified as likely WBC origin if they did not fully meet the criteria for WBC origin classification but were either detected in WBC sequencing with 2-4 supermutant families, detected in 1 supermutant family at a position reported as a hematopoietic variant in COSMIC, or not found in the tumor tissue sequencing data. Variants detected in plasma were classified as tumor origin if they were also detected in tumor sequencing and found in no more than one supermutant family in WBC sequencing. Plasma variants were classified as likely tumor origin if they met at least two of the following criteria: no WBC supermutants with a read depth of at least 1000x at that locus, at least 10 reports in COSMIC where solid tumors had greater number of occurrences than hematologic malignancies, or at least two read families identified in tumor sequencing through manually investigation with IGV. Plasma variants that did not meet criteria for any of the above classifications were deemed of unknown origin and were excluded from further analysis.

For the independent validation cohort, we used the mutation origin for each variant described in the supplementary table 11 of Razavi et al9. Only mutations detected in patients with cancer and classified as phenotype WBC matched, biopsy matched, or IMPACT-bam matched were included, while variants of uncertain origin were excluded. For our purposes, biopsy matched and IMPACT-bam matched were grouped together as tumor origin.

*Longitudinal evaluation*

In the longitudinal cohort, 41 variants were identified at two separate timepoints in a given patient. Of these, only 23 variants that were detected in a sample collected at baseline were included in the training cohort. We set out to discover metrics from our fragment data that remained most consistent across sample collections from the same individual, which could be utilized to refine a machine learning to detect authentic biological signal. In contrast, metrics that vary significantly in different plasma samples are more likely attributed to technical sequencing artifacts, or possibly reflecting a phenotypic change in the disease, but that is less relevant to the goals of this project. For each variant and at each timepoint, fragmentomic summary statistics were calculated on reads at the given locus, including length distributions, fragment end motif patterns, and statistical comparison between mutant and wild-type reads. We determined variant origin by comparison to matched buffy or tumor sequencing with our criteria detailed above; however, variant origin was not considered in assessment of longitudinal stability of fragmentomic summary statistics. We evaluated the different metrics independent of whether or not the variant was WBC derived or tumor derived. After generating data for each timepoint, the correlation between summary statistics across timepoints was assessed using Pearson’s product-moment correlation on paired samples. The statistical tests whose correlation over time had p-value <0.05 were included in the development of machine learning model.

*Machine learning*

In each position containing a ctDNA variant, we analyzed all cfDNA fragments, including those of cancer origin and wild-type, to calculate the selected metrics on fragment lengths, fragment end motifs, and other relevant features. The parameters used to train our MLM model included the variant allele fraction in plasma samples, the total number of distinct reads, the number of distinct reads containing the mutation, and a scaled version of the plasma variant allele fraction. We also included gene-specific parameters to account for the broader variant context, that included the gene and whether or not the variant is a known driver mutation. Variants found withing genes highly recurrent in clonal hematopoiesis were flagged.

Fragment length statistics were represented by the median widths of mutant and wild-type fragments, respectively. Statistical comparisons between fragment widths were conducted using the Kolmogorov-Smirnov statistic and the Mann-Whitney U statistic. Additionally, we considered the absolute average difference score, which reflects the average absolute difference in sequencing depth between mutant and wild-type fragments, and the maximum difference score, indicating the maximum observed difference in sequencing depth.

Motif analysis parameters included the Wilcoxon statistics for two different motif comparisons and an assessment of the distribution of fragment end motifs using the Wilcoxon statistic. Motifs were determined from the 4 bases on each end of the fragment on the 5’ orientation. Variant signature and the specific nucleotides being substituted were used as well. Clinical parameters, including patient’s age, cancer, and stage, were also considered to provide a comprehensive view of the sample characteristics.

To further enhance the model’s predictive accuracy, we incorporated several parameters based on annotations from the COSMIC database. The fraction of hematological occurrences was calculated taking the total number of occurrences across all cancer types in COSMIC and the number of catalogued cases from only hematological cancers, which often contain variants that arise during clonal hematopoiesis. Lastly, for training and testing we relied on the origin classifications that we had determined with the matched tumor/WBC sequencing.

The machine learning algorithm integrates all the above data to create a score from 0 to 1 for each variant. Features are described in detail in supplemental table @@@. Input features are pre-processed with centering, scaling, and Yeo Johnson transformation, with all pre-processing done within each cross-validation fold. Output scores are binarized at a cutoff (default 0.5) into CH-predicted origin and tumor-predicted origin. The algorithm was implemented using XGBoost (version 1.6.0.1) and caret package (version 6.0-92) in R. Model training performed using 10-fold cross validation repeated 10 times, optimizing ROC. Feature importance was extracted using built-in caret feature importance function.

*Statistical Analyses*

All statistical analysis was conducted in R Version 4.2.1. Fragmentomic summary statistics are described in supplemental table. For serial cohort, correlation between the summary statistics across timepoints was determined using Pearson’s product moment correlation between paired samples.

*Data Availability*

Upload to EGA

*Code Availability*

All code that is necessary to run the machine learning model in R, as well as, additional statistical analyses are publicly available on GitHub (<https://github.com/dannyrabiz/cfDNA_CHIP_MLM/>). [we’ll change this link when we add it to MolecOnc github]