GMR
The Original

VARIANT ORIGIN PREDICTION FROM WITHIN-LOCUS CELL-FREE DNA FRAGMENTOMICS: A SELF-NORMALIZED, SINGLE-SAMPLE CLASSIFIER DISCRIMINATING TUMOR-SOMATIC FROM HEMATOPOIETIC VARIANTS

Rajesh Bhammar[1,2], Darshan Marjadi[1], Salil Vaniawala[2]

[1]Department of Biotechnology, Shree Ramkrishna Institute of Computer Education & Applied Sciences, Sarvajanik University, Surat, Gujarat, India
[2]SN Genelab Pvt. Ltd., Surat, Gujarat, India
***Corresponding author:** Rajesh Bhammar (rajesh@sngenelab.com)

ABSTRACT

Background. Determining the biological origin of a mutation detected in plasma cell-free DNA (cfDNA) — whether it is tumor-somatic or hematopoietic (constitutional germline or clonal hematopoiesis, CHIP) — is a shared, unsolved obstacle across liquid-biopsy applications [1, 2]. Misassignment of origin degrades the specificity of targeted-therapy variant classification, inflates false-positive calls in multi-cancer early detection (MCED), and confounds molecular residual disease (MRD) monitoring [2, 3]. Matched white blood cell (WBC) sequencing resolves origin but increases assay cost and is frequently unavailable [1]. We developed a within-locus fragmentomic classifier that infers variant origin from a single plasma sample, without a matched normal. The classifier reads the fragment-length contrast between the mutant (ALT) and reference (REF) alleles at the same genomic position, which serves as a self-normalizing signal.

Methods. In a retrospective single-center cohort of 226 histopathologically confirmed solid-tumor patients with matched buffy-coat germline and 35 healthy-donor plasma controls (73-gene hybrid-capture panel, ~20,000× pre-deduplication depth, unique molecular identifier consensus, hg19), we computed per-variant ALT-versus-REF fragment-length cumulative distribution functions (CDFs). From these we derived eight self-normalized within-locus features and trained an L2-regularized logistic regression classifier (somatic versus hematopoietic) with a pre-specified 0.50 decision threshold and a plasma variant allele frequency (VAF) $\geq$ 0.05 abstain floor. Performance was estimated by patient-grouped cross-validation and a held-out patient split, and calibration was assessed.

Results. Using 1,943 quality-controlled variants (338 tumor-somatic, 1,605 hematopoietic) from 249 individuals, the classifier reached an out-of-fold area under the receiver-operating-characteristic curve (AUC) of 0.920 (sensitivity 83.1%, specificity 86.8% at 0.50) and, on the held-out patient set, an AUC of 0.962 (sensitivity 93.6%, specificity 85.8%); specificity on healthy-control plasma was 88.9%. Calibration was good (Brier score 0.103). Tumor-somatic alleles were systematically enriched in short (< 150 bp) fragments relative to the same-locus reference, whereas germline and healthy-control alleles overlaid the ~167 bp mononucleosomal reference profile.

Conclusions. A self-normalized, within-locus fragmentomic classifier predicts cfDNA variant origin from a single plasma draw without matched WBC sequencing. It may help with targeted-therapy variant triage, suppressing hematopoietic false positives in early detection, and tumor-informed MRD tracking.

KEYWORDS: cell-free DNA, fragmentomics, variant origin prediction, clonal hematopoiesis, liquid biopsy, early cancer detection, minimal residual disease

INTRODUCTION

Circulating cell-free DNA (cfDNA) released into plasma carries genetic material from cells throughout the body, and the tumor-derived fraction (circulating tumor DNA, ctDNA) has become a versatile analyte for non-invasive oncology [4, 5]. A single mutation observed in plasma, however, does not reveal its origin [1]. The same nucleotide change may arise in the tumor, be inherited in the germline, or emerge in an expanding hematopoietic clone (CHIP); these origins carry opposite clinical implications yet are indistinguishable based on the variant call and its allele fraction alone [6]. A tumor-somatic variant may warrant clinical action, whereas the same change arising in the germline or in a hematopoietic clone does not [7]. Assigning per-variant origin may therefore help improve the sensitivity and specificity of non-invasive oncology assays, and it is central to several liquid-biopsy applications [1].

Targeted therapy: Clinical interpretation of a tumor variant follows structured frameworks — the AMP/ASCO/CAP tiering system for somatic variants [8], curated actionability knowledge bases such as OncoKB [9], and the ACMG/AMP criteria for germline classification [10]. The interpretation depends on knowing the variant's origin [11]. When the origin is uncertain, a germline polymorphism may be misidentified as a somatic

driver [11]. More seriously, a hematopoietic or germline variant in a gene such as TP53, ATM, or CHEK2 may be reported as a targetable tumor alteration and steer therapy inappropriately [12]. Blood-derived variants in DNA damage response genes are a documented and frequent source of such misclassification in plasma-only assays [12]. Population databases (gnomAD [13]) and matched-normal subtraction mitigate the germline case, but CHIP is not captured reliably by population frequency and, without matched WBC sequencing, remains a residual confounder that increases with age [1, 12].

Early cancer detection: Early diagnosis is among the strongest determinants of cancer survival: five-year survival is far higher for localized than for metastatic disease [14], yet diagnosis is often late in low- and middle-income settings [15]. Somatic mutations are among the most specific markers available for blood-based detection, because a tumor-derived mutation is, in principle, absent from normal tissue [16]. Beyond the small number of actionable drivers, broad sequencing panels can detect many somatic mutations per tumor, raising the chance of finding at least one tumor signal in plasma [17]. This gain has a cost: the larger the panel, the more germline and CHIP variants it also captures, and these are readily mistaken for tumor signal [12]. Blood-based MCED assays therefore face a specificity problem at population scale, where a false-positive result can trigger wasteful, potentially risky, and emotionally distressing confirmatory work-up in a person without cancer [18]. A method that labels an individual plasma variant as hematopoietic rather than tumor-derived can remove many of these false positives at the variant level, before they reach a sample-level call [19]. Such a method is directly relevant wherever large somatic panels are used for detection.

Molecular residual disease monitoring: MRD assays track ctDNA after curative-intent treatment to detect relapse months before imaging [20]. Their sensitivity depends on counting tumor-specific molecules at a very low tumor fraction, where a handful of misattributed hematopoietic reads can create a spurious positive and erode both specificity and lead time. CHIP remains the principal contaminant, and tumor-informed and tumor-agnostic MRD designs alike must actively distinguish tumor-derived from blood-derived variants to remain reliable [1]. Per-variant origin assignment is thus a prerequisite for trustworthy longitudinal monitoring.

The origin of an individual variant can be inferred from the plasma sample itself [4]. Cell-free DNA fragmentation is non-random in healthy individuals [21]. Nucleosomal protection during degradation yields a dominant ~167-bp mononucleosomal mode, and tumor-derived molecules are systematically shorter, with a principal mode around 134–144 bp [22]. Critically, CHIP- and germline-derived molecules are both hematopoietic in origin and track the ~167-bp profile, so fragment length separates tumor from hematopoietic DNA even though it cannot separate CHIP from germline [4].

At the same genomic locus, we compare the fragment-length distribution of molecules carrying the mutant (ALT) allele against those carrying the reference (REF) allele. Because both distributions are drawn from the same locus, the same library, and the same sample, the comparison is self-normalizing: locus-, GC-, and sample-level biases cancel, and what remains is the origin signal. A tumor-somatic allele is associated with short fragments while its same-locus reference is dominated by normal-length molecules, so the two distributions diverge. A germline or CHIP allele shares the hematopoietic length profile of its reference, so the two distributions coincide. This within-locus contrast is the basis of the classifier developed here, and it is intended to serve, from a single sample, all three applications above: variant triage for targeted therapy, removal of hematopoietic false positives in early detection, and tumor-specific filtering for residual disease monitoring.

Methods

Study design and sample inclusion

This was a retrospective, single-center, non-interventional analysis of residual plasma and matched buffy coat specimens from an institutional repository. Institutional ethical approval (Ref. RK/PHD/005) was obtained, and written informed consent was secured from all participants prior to collection. Eligible cases were histopathologically confirmed solid tumors in patients aged 18 years or older with no prior systemic therapy; healthy controls were identified retrospectively and de-identified. Samples were collected between May 2024 and December 2025. The cohort comprised 226 cancer patients (median age 64.4 years) and, for the fragmentomic origin analysis reported here, 35 healthy-donor plasma controls with targeted sequencing. The distribution of the cancer cohort by tumor type and stage is given in Table 1. This origin analysis draws on the same cohort and sequencing workflow as our companion multimodal early-detection study.

Sample processing and sequencing: Blood (8–10 mL) was collected in PAXgene Blood ccfDNA tubes, and plasma was separated by a two-step centrifugation and stored at −80°C. cfDNA was extracted on the Roche MagNA Pure 96 system and quantified on a Qubit fluorometer. Sequencing libraries were prepared with the Agilent SureSelect XT HS Low Input Library Preparation Kit for Illumina (10–50 ng cfDNA input); the workflow comprised end-repair and dA-tailing, ligation, AMPure XP bead purification, indexed PCR with dual-indexed primers, and a final AMPure XP purification. To preserve the short cfDNA fragments, the adapter concentration was titrated down (1:2), the ligation incubation was extended to 60 minutes, and the AMPure XP bead ratio was adjusted to 1.5×. Target enrichment used a custom Agilent SureSelect XT HS panel of biotinylated 120-mer probes covering all coding exons and flanking splice sites (±10 bp) of 73 recurrently mutated solid-tumor genes at 1× tiling density, yielding 1,933 target regions. Libraries were sequenced on an Illumina NovaSeq 6000 (2×150 bp) to a target mean pre-deduplication depth of ~20,000×. Matched germline DNA was extracted from the buffy coat (QIAamp DNA Blood Mini Kit), sheared on a Diagenode Bioruptor Pico, and prepared with the same capture panel.

Reads were base-quality trimmed with AGeNT (Q20), aligned to the hg19 reference with BWA-MEM v0.7.17, and collapsed into UMI consensus reads with AGeNT CReaK. Somatic variants were called from the consensus alignments with the Illumina DRAGEN caller configured for low-frequency cfDNA variants, and germline variants from the matched buffy coat were subtracted to define the blood-presence status of each plasma variant.

Per-variant within-locus fragmentomics: For every called variant, reads spanning the position were partitioned by the allele they carried, and the insert size of each fragment was recorded. We then constructed two empirical fragment-length cumulative distribution functions per variant: one over ALT-supporting fragments and one over same-locus REF-supporting fragments. Insertion and deletion lengths were corrected so that the size of the indel did not masquerade as a fragment-length shift. Variants entered the analysis only if they passed quality control: read depth ≥ 50, ALT allele depth ≥ 5, mapping and strand-quality thresholds, soft-clip fraction ≤ 0.10, not in a homopolymer or short tandem repeat artifact context, not homozygous (which removes the same-locus reference control), and with at least 20 ALT and 20 REF fragments contributing to the two curves.

The discriminative quantity is the signed within-locus short-fraction excess (the percentage of ALT fragments shorter than 150 bp minus the same percentage for REF fragments at that locus); a positive value indicates that the mutant allele is carried by shorter molecules than its reference control—the expected tumor signature [23]. From the paired CDFs we derived eight self-normalized features for the classifier: the signed short-fraction excess; the mutant and reference fractions below 100, 150, and 167 bp; the mononucleosomal-band fraction of mutant fragments; the signed maximum CDF separation; and the signed area between the ALT and REF CDFs.

Classifier development and validation: Variants were labeled from the matched-normal genotype as tumor-somatic (absent from adequately covered blood, gnomAD-rare, and not ClinVar-benign) or hematopoietic (present in blood, which pools constitutional germline with CHIP); dual-origin variants and low-VAF healthy-control calls were excluded from the hematopoietic class. The classifier was an L2-regularized logistic regression on the eight standardized within-locus features, with class weighting to balance the somatic and hematopoietic classes and a decision threshold pre-specified and locked at 0.50. To control class imbalance without distorting the operating point, hematopoietic negatives were VAF-matched to the somatic positives at a 3:1 ratio. This yielded a modeling set of 1,943 variants (338 tumor-somatic, 1,017 matched-normal hematopoietic, and 588 healthy-control) from 249 individuals: 214 cancer patients contributed variants, of whom 78 contributed somatic calls, together with 35 healthy donors. Discrimination was estimated by patient-grouped stratified five-fold cross-validation, so that no patient contributed variants to both training and held-out folds, and separately by a held-out 30% patient split with no patient shared with the training set. All preprocessing was fit within training folds only, and probabilistic calibration was assessed with the Brier score, the expected calibration error, and reliability diagrams.

Statistical analysis: Areas under the ROC curve were computed as the Mann–Whitney rank statistic. Sensitivity (positive percent agreement) and specificity (negative percent agreement) were evaluated at the locked 0.50 threshold. Analyses used Python (scikit-learn); per-variant fragment extraction used the validated in-house cfDNA fragmentomics pipeline.

Results

Cohort and variant set: The cancer cohort spanned the five most common solid tumors in the repository — lung (71), colorectal (42), ovary (35), prostate (32), and breast (28) — as well as 18 rarer tumors (Table 1). After quality control and within-locus labeling, the modeling set comprised 1,943 variants from 249 individuals: 338 tumor-somatic variants (from 78 cancer patients), 1,017 matched-normal hematopoietic variants, and 588 healthy-control variants. The removal of dual-origin contamination eliminated blood-present alleles that were partially tumor-derived, and the VAF floor removed low-frequency healthy noise, together yielding a cleaner hematopoietic negative class.

The within-locus fragmentomic origin signal: Across representative tumor-somatic variants, the mutant-allele fragment-length CDF was consistently left-shifted relative to the same-locus reference, and the corresponding insert-size density showed a clear excess of short (< 150 bp) mutant fragments (Figure 1). The effect was visible across driver and non-canonical genes alike (e.g., TP53, EGFR, KRAS, ALK, and APC) and across a range of plasma VAFs, confirming that the signal is a property of the molecules carrying the variant rather than of any single hotspot. In healthy-control plasma, the same construction produced overlapping ALT and REF curves and near-symmetric mirrored (butterfly) insert-size densities (Figure 2), the expected germline behavior, with correspondingly low predicted probabilities of somatic origin. These paired figures are the visual basis of the classifier: origin is read from the divergence, or coincidence, of the two same-locus distributions.

Figure 1. Representative tumor-somatic variants. For each of 30 variants, the within-locus fragment-length cumulative distribution of mutant (ALT, red) versus reference (REF, blue) fragments is shown, with the overlapping insert-size density as an inset; the dotted line marks 150 bp. Titles give gene, base change, sample, plasma VAF, and the model-predicted probability of somatic origin. Mutant alleles are systematically enriched in short fragments relative to the same-locus reference.

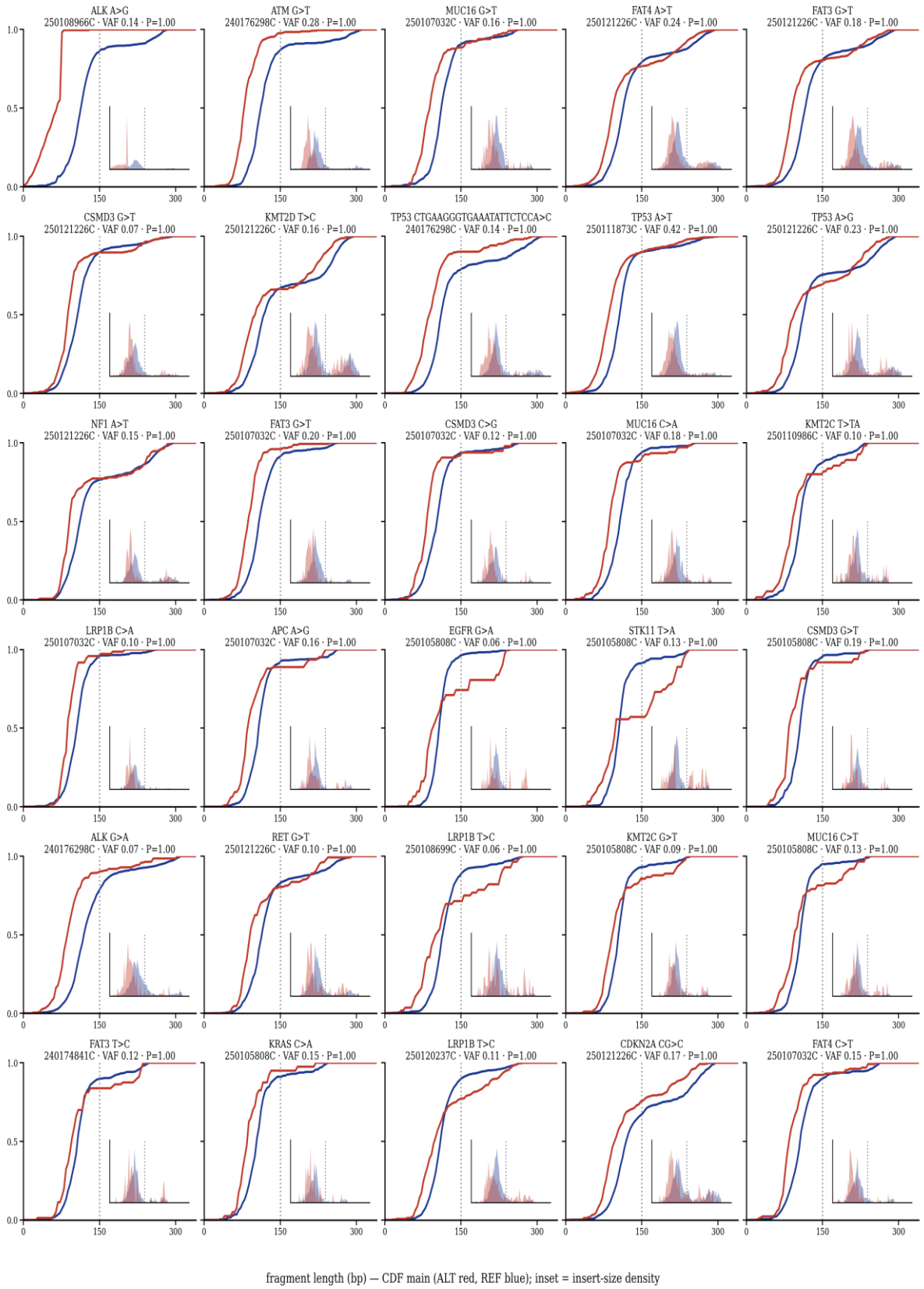
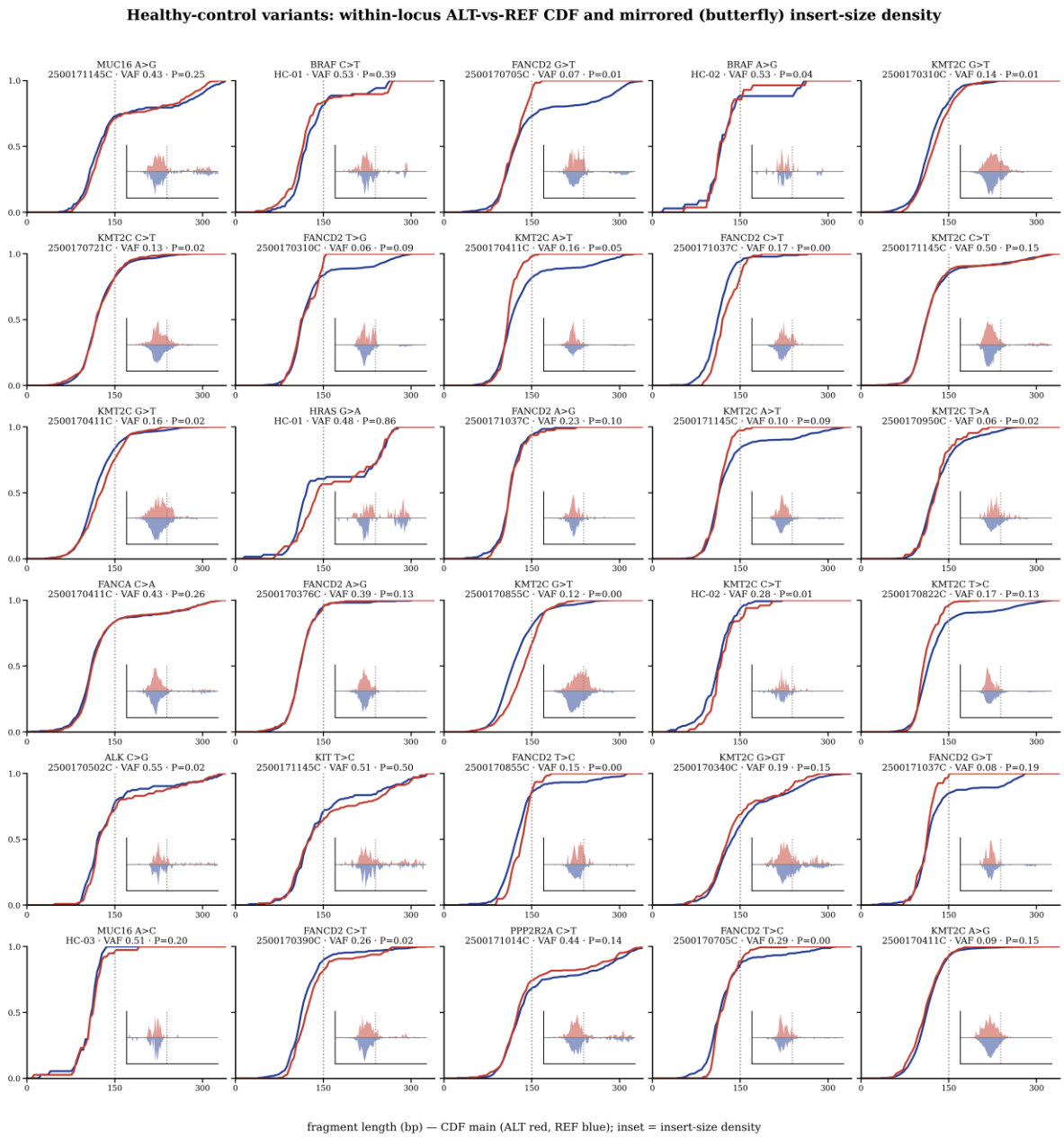
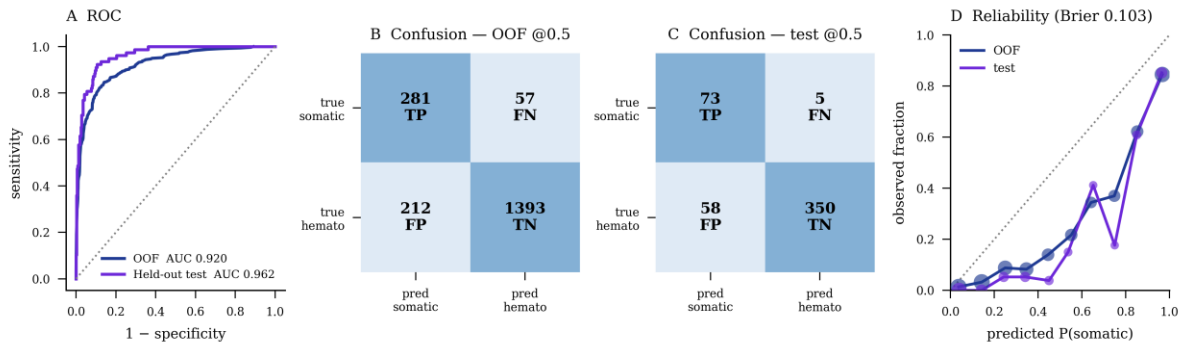


Figure 2. Healthy-control variants. For each of 30 healthy-donor plasma variants, the within-locus ALT-versus-REF CDF is shown with a mirrored (butterfly) insert-size density inset (ALT upward, REF downward). Mutant and reference distributions coincide and the butterfly densities are near-symmetric, the expected hematopoietic/germline profile; predicted probabilities of somatic origin are low.



Classifier performance and calibration: Trained on the eight within-locus features, the classifier discriminated tumor-somatic from hematopoietic variants with an out-of-fold AUC of 0.920 under patient-grouped cross-validation (sensitivity 83.1%, specificity 86.8% at the locked 0.50 threshold; 281 true-positive, 57 false-negative, 212 false-positive, and 1,393 true-negative variants). On the held-out patient split, in which no patient was shared with the training set, the AUC was 0.962 (sensitivity 93.6%, specificity 85.8%; Table 2, Figure 3A–C). Specificity evaluated on healthy-control plasma alone was 88.9%. Probabilistic outputs were well calibrated (Brier score 0.103; expected calibration error 0.130), with reliability curves close to the diagonal across the mid-range and only modest tail deviation (Figure 3D), supporting use of the outputs as probabilities and honoring the pre-specified 0.50 operating point.

Figure 3. Classifier performance. (A) ROC curves for patient-grouped out-of-fold cross-validation and the held-out patient test. (B, C) Confusion matrices at the locked 0.50 threshold for the out-of-fold and held-out test evaluations. (D) Reliability (calibration) diagram; marker size is proportional to bin count.



The predicted probabilities separated the classes as expected: tumor-somatic variants concentrated at high probabilities, whereas matched-normal hematopoietic and healthy-control variants fell below the 0.50 threshold (Figure 4).

Figure 4. Score distribution by class. Predicted probability of somatic origin for each class; each violin shows the full distribution, with the interquartile range as a thick bar and the median as a white dot. The dashed line marks the 0.50 threshold.

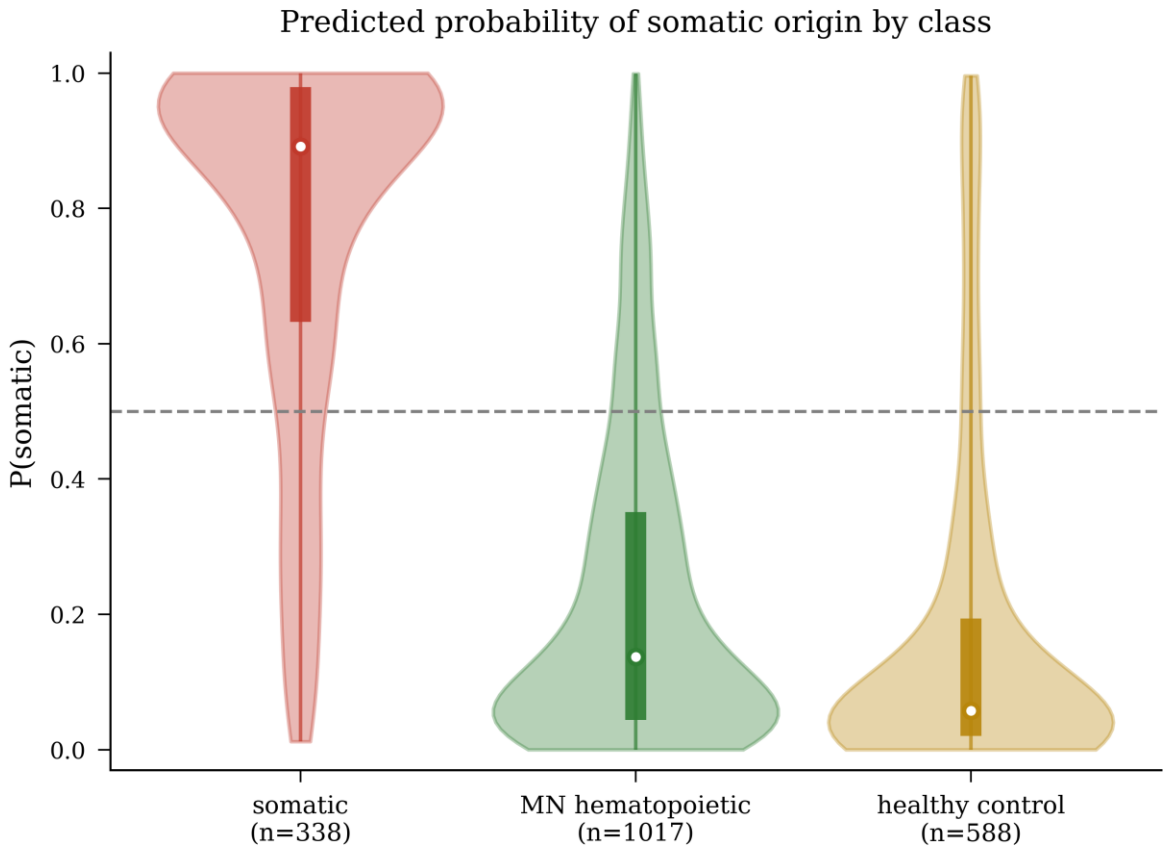


Table 1. Composition of the cancer cohort by tumor type and stage (n = 226).

Tumor Type	Stage I	Stage II	Stage III	Stage IV	Total	% III–IV
Lung	0	0	45	26	71	100.0%
Colorectal	6	10	17	9	42	61.9%
Ovary	0	0	19	16	35	100.0%
Prostate	2	17	11	2	32	40.6%
Breast	3	15	8	2	28	35.7%
Others	2	1	9	6	18	83.3%
Total	13	43	109	61	226	75.2%

Table 2. Classifier performance for tumor-somatic versus hematopoietic origin at the locked 0.50 threshold.

Evaluation	AUC	PR-AUC	Sensitivity	Specificity	TP	FN	FP	TN
Patient-grouped OOF (5-fold)	0.920	0.772	83.1%	86.8%	281	57	212	1393
Held-out patient test (30%)	0.962	0.844	93.6%	85.8%	73	5	58	350
Healthy control	—	—	—	88.9%	—	—	—	—

Calibration (out-of-fold): Brier score 0.103; expected calibration error 0.130. OOF, out-of-fold; PR-AUC, area under the precision–recall curve; TP/FN/FP/TN, true/false positive/negative variant counts.

CONCLUSION

We show that the biological origin of an individual cfDNA variant — tumor-somatic versus hematopoietic — can be predicted from a single plasma sample, without matched white blood cell sequencing, by contrasting the fragment-length distributions of the mutant and reference alleles at the same locus. The self-normalizing nature of this within-locus comparison is its central methodological advantage: because both distributions come from the same position, library, and sample, locus- and sample-level biases cancel and the residual signal reflects origin. The resulting classifier achieved an out-of-fold AUC of 0.920 and a held-out-patient AUC of 0.962 with well-calibrated probabilities.

This variant-level origin module speaks directly to the three challenges that motivated it. For targeted-therapy variant classification, it adds a sequence-independent axis that can support calling a non-canonical variant as tumor-derived, or flag a blood-derived variant in a DNA damage response gene before it is reported as actionable. For multi-cancer early detection, it removes hematopoietic false positives at the variant level, where they originate, complementing sample-level fragmentomic classifiers and the orthogonal mutation and copy-number channels of a multi-modal assay [17]. For molecular residual disease monitoring, it provides a filter to retain tumor-specific molecules at low tumor fraction. Consistent with prior work on variant origin prediction [24], origin can be assigned computationally, and the within-locus formulation offers a simple and interpretable way to do so.

Notable limitations qualify these findings. The cohort is retrospective and single-center. The 73-gene solid-tumor panel does not cover the canonical CHIP genes (DNMT3A, TET2, ASXL1), so the hematopoietic class is dominated by constitutional germline variants, and the classifier is best described as discriminating tumor-somatic from hematopoietic origin rather than resolving CHIP from germline, which fragment length cannot do [4]. Incorporation of fragment end motif and nucleosome-positioning features and fusion with allele fraction in copy number context are natural extensions expected to raise both sensitivity and specificity further. The within-locus variant-origin classifier described here is a practical, low-cost step toward resolving one of the central ambiguities of plasma-only liquid biopsy.

DECLARATIONS

Ethics approval and consent to participate

This is a retrospective study. Institutional ethical approval (Ref no: RK/PHD/005) was secured to use leftover samples from Shree Ramkrishna Institute of Computer Education & Applied Sciences, Sarvajanik University, Surat, Gujarat, India. Written informed consent was obtained from all participants prior to sample collection.

Consent for publication

Not applicable.

Availability of data and materials

The within-locus fragmentomics pipeline and classifier code, datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

RB and SV are affiliated with SN Genelab Pvt. Ltd. SV is the founder and RB is a salaried employee of the company. DM declares no competing interests. These affiliations did not influence the study design, analysis, or interpretation of results.

Funding

This research received no external funding.

Authors' contributions

RB, DM and SV designed the study. RB processed the samples, generated and analyzed the data. RB was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Abbreviations. cfDNA, cell-free DNA; ctDNA, circulating tumor DNA; CHIP, clonal hematopoiesis of indeterminate potential; VOP, variant origin prediction; CDF, cumulative distribution function; VAF, variant allele frequency; AUC, area under the curve; ROC, receiver operating characteristic; MCED, multi-cancer early detection; MRD, molecular residual disease; UMI, unique molecular identifier; WBC, white blood cell; OOF, out-of-fold.

REFERENCES

1. Razavi P, Li BT, Brown DN, Jung B, Hubbell E, Shen R, et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nature Medicine. 2019;25:1928–37. https://doi.org/10.1038/s41591-019-0652-7 .

2. Semenkovich NP, Szymanski JJ, Earland N, Chauhan PS, Pellini B, Chaudhuri AA. Genomic approaches to cancer and minimal residual disease detection using circulating tumor DNA. Journal for ImmunoTherapy of Cancer. 2023;11. https://doi.org/10.1136/jitc-2022-006284 .

3. Leal A, Grieken NCT van, Palsgrove DN, Phallen J, Medina JE, Hruban C, et al. White blood cell and cell-free DNA analyses for detection of residual disease in gastric cancer. Nature Communications. 2020;11:525. https://doi.org/10.1038/s41467-020-14310-3 .

4. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570:385–9. https://doi.org/10.1038/s41586-019-1272-6 .

5. Jamshidi A, Liu MC, Klein EA, Venn O, Hubbell E, Beausang JF, et al. Evaluation of cell-free DNA approaches for multi-cancer early detection. Cancer Cell. 2022;40:1537. https://doi.org/10.1016/j.ccell.2022.10.022 .

6. Keller L, Belloum Y, Wikman H, Pantel K. Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond. British Journal of Cancer. 2020;124:345–58. https://doi.org/10.1038/s41416-020-01047-5 .

7. Chan T, Nagayama S, Chin YM, Otaki M, Hayashi R, Kiyotani K, et al. Clinical significance of clonal hematopoiesis in the interpretation of blood liquid biopsy. Molecular Oncology. 2020;14:1719–30. https://doi.org/10.1002/1878-0261.12727 .

8. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. Journal of Molecular Diagnostics. 2016;19:4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002 .

9. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology. 2017;2017:1–16. https://doi.org/10.1200/po.17.00011 .

10. Richards S, Aziz N, Bale SJ, Bick D, Das S, Gastier-Foster JM, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015;17:405–24. https://doi.org/10.1038/gim.2015.30 .

11. Jones S, Anagnostou V, Lytle K, Parpart-Li S, Nesselbush M, Riley DR, et al. Personalized genomic analyses for cancer mutation discovery and interpretation. Science Translational Medicine. 2015;7. https://doi.org/10.1126/scitranslmed.aaa7161 .

12. Ptashkin R, Mandelker D, Coombs CC, Bolton KL, Yelskaya Z, Hyman DM, et al. Prevalence of Clonal Hematopoiesis Mutations in Tumor-Only Clinical Genomic Profiling of Solid Tumors. JAMA Oncology. 2018;4:1589. https://doi.org/10.1001/jamaoncol.2018.2297 .

13. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang QS, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43. https://doi.org/10.1038/s41586-020-2308-7 .

14. Connal S, Cameron JM, Sala A, Brennan PM, Palmer DS, Palmer JD, et al. Liquid biopsies: the future of cancer early detection. Journal of Translational Medicine. 2023;21:118. https://doi.org/10.1186/s12967-023-03960-8 .

15. Cazap E, Magrath I, Kingham TP, Elzawawy A. Structural Barriers to Diagnosis and Treatment of Cancer in Low- and Middle-Income Countries: The Urgent Need for Scaling Up. Journal of Clinical Oncology. 2015;34:14–9. https://doi.org/10.1200/jco.2015.61.9189 .

16. Hench IB, Hench J, Tolnay M. Liquid Biopsy in Clinical Management of Breast, Lung, and Colorectal Cancer. Frontiers in Medicine. 2018;5:9. https://doi.org/10.3389/fmed.2018.00009 .

17. Chen Y, Han T, Zhou Y, Mao B, Zhuang W. Efficacy comparison of targeted next-generation sequencing in the identification of somatic mutations in circulating tumor DNA from different stages of lung cancer. Neoplasma. 2019;66:652–60. https://doi.org/10.4149/neo_2018_181130n910 .

18. Snead CM, Zheng D, Ngo-Metzger Q, Gould MK. Multicancer Detection Assays: Promise and Potential Harms of a Novel Cancer Screening Tool. The Permanente Journal. 2026;30:84–90. https://doi.org/10.7812/tpp/25.075 .

19. Brannon AR, Jayakumaran G, Diosdado M, Patel J, Razumova A, Hu Y, et al. Enhanced specificity of clinical high-sensitivity tumor mutation profiling in cell-free DNA via paired normal sequencing using MSK-ACCESS. Nature Communications. 2021;12:3770. https://doi.org/10.1038/s41467-021-24109-5 .

20. Tie J, Wang Y, Tomasetti C, Li L, Springer S, Kinde I, et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Science Translational Medicine. 2016;8. https://doi.org/10.1126/scitranslmed.aaf6219 .

21. Stejskal P, Goodarzi H, Srovnal J, Hajdúch M, Veer L van 't, Magbanua MJM. Circulating tumor nucleic acids: biology, release mechanisms, and clinical relevance. Molecular Cancer. 2023;22:15. https://doi.org/10.1186/s12943-022-01710-w .

22. Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza RM, Baker DN, et al. Fragment Length of Circulating Tumor DNA. PLoS Genetics. 2016;12. https://doi.org/10.1371/journal.pgen.1006162 .

23. Sánchez C, Roch B, Mazard T, Blache P, Dache ZAA, Pastor B, et al. Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. JCI Insight. 2021;6. https://doi.org/10.1172/jci.insight.144561 .

24. Brown DW, Sun D, Fine AD, He S, McDevitt MR, Pontbriand KM, et al. Fragmentomic-based algorithm to computationally predict tumor-somatic, germline, and clonal hematopoiesis variant origin in liquid biopsy. The Journal of Liquid Biopsy. 2025;9:100311. https://doi.org/10.1016/j.jlb.2025.100311 .