Title

Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer.

Journal

Science Translational Medicine, 15(678)

Authors

Budhraja, Karan
McDonald, Bradon
Stephens, Michelle
et al.

Publication Date

2023-01-11

DOI

10.1126/scitranslmed.abm6863

Peer reviewed

HHS Public Access

Author manuscript

Sci Transl Med. Author manuscript; available in PMC 2023 July 11.

Published in final edited form as:

Sci Transl Med. 2023 January 11; 15(678): eabm6863. doi:10.1126/scitranslmed.abm6863.

Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer

Karan K. Budhraja[1,†], **Bradon R. McDonald**[1,†], **Michelle D. Stephens**[1,†], **Tania Contente-Cuomo**[2], **Havell Markus**[3], **Maria Farooq**[4], **Patricia F. Favaro**[1], **Sydney Connor**[5], **Sara A. Byron**[2], **Jan B. Egan**[6], **Brenda Ernst**[6], **Timothy K. McDaniel**[2,‡], **Aleksandar Sekulic**[6], **Nhan L. Tran**[6], **Michael D. Prados**[7], **Mitesh J. Borad**[6], **Michael E. Berens**[2], **Barbara A. Pockaj**[6], **Patricia M. LoRusso**[8], **Alan Bryce**[6], **Jeffrey M. Trent**[2], **Muhammed Murtaza**[1,*]

[1]Department of Surgery and Center for Human Genomics and Precision Medicine, University of Wisconsin–Madison; Madison, WI 53705, USA.

[2]Translational Genomics Research Institute, Phoenix, AZ 85004, USA.

[3]Pennsylvania State University, Hershey, PA 17033, USA.

[4]Department of Medicine, The University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA.

[5]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218 USA.

[6]Mayo Clinic, Scottsdale, AZ 85259, USA.

[7]Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA 94143, USA

[8]Yale Cancer Center, New Haven, CT 06520, USA.

Abstract

Genome-wide fragmentation patterns in cell-free DNA (cfDNA) in plasma are strongly influenced by cellular origin due to variation in chromatin accessibility across cell types. Such differences between healthy and cancer cells provide the opportunity for development of novel cancer

*Corresponding author: murtaza@surgery.wisc.edu.
†These authors contributed equally to this work.
‡Current address: Delfi Diagnostics, 1810 Embarcadero Road, Suite 100. Palo Alto, CA 94303, USA

Author Contributions: KKB, BRM and MM conceptualized and designed the study. KKB, BRM, HM, and MM developed methods. SB, JBE, BE, TKM, AS, NLT, MDP, MJB, MEB, BAP, PML, AB, and JMT designed and conducted prospective clinical studies. MDS, TCC, MF, PFF, and SC generated data. KKB, BRM, and HM analyzed sequencing data. KKB, BRM, MDS, and MM interpreted data. KKB, BRM, MDS, and MM wrote the paper with assistance from TCC and JMT. All authors approved the final manuscript.

Competing Interests: KKB, BRM, HM, and MM are inventors on patent applications covering technologies described here including patent application number PCT/US20/41469, titled “Methods of detecting disease and treatment response in cfDNA”. MM has consulted for AstraZeneca, Bristol Myers Squibb, Castle Biosciences, currently consults for Translational Genomics Research Institute (TGen), serves on the scientific advisory board of and holds stock options for PetDx. BRM, TCC, TKM, and MM are inventors on patent applications submitted by TGen related to cancer genomics and cell-free DNA analyses that have been licensed to Exact Sciences, under terms reviewed and approved by TGen. TKM is an employee and shareholder of Delfi Diagnostics, serves on the scientific advisory board of Deepcell and Omniome. All other authors declare that they have no competing interests.

Data and materials availability: Plasma sequencing data generated for this study is available in dbGaP (phs003170.v1.p1). Normalization tables, recurrently protected region maps, and code used for data analysis are available in Zenodo (10.5281/zenodo.7402091). All other data associated with this study are present in the paper or the Supplementary Materials.

diagnostics. Here, we investigated whether analysis of cfDNA fragment end positions and their surrounding DNA sequences reveals the presence of tumor-derived DNA in blood. We performed genome-wide analysis of cfDNA from 521 samples and analyzed sequencing data from an additional 2,147 samples, including healthy individuals and patients with 11 different cancer types. We developed a metric based on genome-wide differences in fragment positioning, weighted by fragment length and GC-content (information-weighted fraction of aberrant fragments, iwFAF). We observed that iwFAF strongly correlated with tumor fraction, was higher for DNA fragments carrying somatic point mutations, and was higher within genomic regions affected by copy number amplifications. We also calculated sample-level means of nucleotide frequencies observed at genomic positions spanning fragment ends. Using a combination of iwFAF and 9 nucleotide frequencies from 3 positions surrounding fragment ends, we developed a machine-learning model to differentiate healthy individuals from cancer patients. We observed an area under the receiver operative characteristic curve (AUC) of 0.91 for detection of cancer at any stage and an AUC of 0.87 for detection of stage I cancer patient samples. Our findings remained robust with as few as 1 million fragments analyzed per sample, demonstrating that analysis of fragment ends can become a cost-effective and accessible approach for cancer detection and monitoring.

One-sentence summary:

Analyzing the positioning and nucleotide frequency at plasma DNA fragment ends can improve cancer detection.

Introduction

Analysis of plasma DNA has enabled novel diagnostic approaches in prenatal (1), transplant (2), and cancer medicine (3). For patients with cancer, detection and quantification of tumor-derived plasma DNA has mostly relied on analysis of somatic genetic alterations. To improve early detection of cancer, tumor DNA can be detected using targeted analysis of genomic loci that are recurrently affected by somatic alterations in cancer (4). However, there are few such genes and genomic loci that are recurrently altered in most patients across cancer types. In addition, genes that are commonly mutated in cancer (such as *TP53* and *KRAS*) can also be affected in non-malignant conditions, such as clonal hematopoiesis of indeterminate potential, limiting the application of targeted gene resequencing for multi-cancer early detection (5). An alternative approach is to analyze plasma DNA methylation, which captures a combination of tissue-specific and cancer-specific alterations in DNA methylation (6). Using this approach, early detection of cancer relies on identifying unusual patterns of methylation in plasma DNA driven by DNA shed from tissue types that do not routinely contribute DNA into the bloodstream (7). In patients with early-stage disease at presentation or in pre-symptomatic patients where early detection of cancer is relevant, the fraction of tumor-derived DNA in plasma is very low (often 0.1% or below). Consequently, plasma DNA must be obtained from multiple tubes of blood to provide sufficient amounts for targeted resequencing and achieve required depth of sequencing coverage to detect mutations or methylation signatures. These requirements can be costly and limit the development, evaluation, and implementation of early detection using plasma DNA analysis in large clinical and epidemiological studies.

Some have attempted to circumvent this challenge by trading the depth of molecular analysis for the breadth of analysis of the genome and by leveraging genomic features that capture how DNA shed from different cell types is processed and fragmented in blood. Fragmentation characteristics of cell-free DNA (cfDNA) are not random and reflect chromatin accessibility in the cells that contribute such DNA into plasma (8). DNA fragments from genomic loci bound by nucleosomes or other proteins are protected from degradation in plasma (9). Nucleosome positioning and chromatin accessibility vary between cell types and in different cell states (10). Consequently, when DNA from a cancer cell is shed into plasma, the protected fragments may differ in genomic position relative to the majority of cfDNA in plasma, which is derived from peripheral blood cells (11).

Analysis of the relative density of short and long fragments across the genome can capture differences in chromatin accessibility (12). This approach measures short fragment density in each of approximately 500 fixed windows of 5 megabases across the genome, requires 1 – 2× whole genome sequencing and uses a machine-learning model to help distinguish patients with cancer from healthy individuals. Additionally, 4-base-pair (bp) sequence motifs within cfDNA fragment ends can be tissue-specific, potentially due to variation in molecular pathways that drive DNA shedding and degradation in plasma (13). Quantification of individual fragments that carry specific subsets of sequence motifs can help identify plasma DNA samples from patients with cancer.

Here, we evaluated the hypothesis that fragmentation breakpoints from tumor-derived DNA in plasma can serve as a cancer biomarker, using an approach called genome-wide analysis of fragment ends (GALYFRE). Unlike earlier studies that relied on differences in fragment lengths across genomic regions or on differences in sequence motifs in individual cfDNA fragments, GALYFRE aggregates genomic positioning of breakpoints across all sequenced fragments in a sample. For each sample, we quantified fragments that break in genomic regions protected from degradation in cfDNA from healthy individuals, and adjusted for fragment length and GC-content. Additionally, we measured the mean nucleotide frequencies at positions adjacent to fragment ends. Through analysis of more than 2000 samples from patients with cancer, we showed that measurement of aberrant fragmentation is a potential biomarker to distinguish blood samples from patients with cancer and healthy individuals.

Results

Measurement and comparison of aberrant fragmentation in healthy individuals and patients with cancer

To evaluate whether genomic positioning of fragment ends in plasma DNA was different between cancer patients and healthy individuals, we first inferred a map of genomic regions recurrently protected from degradation using the Windowed Protection Score from Snyder *et al* (8). Using whole-genome sequencing of plasma DNA from 17 healthy individuals, we identified 12.7 million recurrently protected regions (RPRs) across the genome, with a median length of 39 bp and spanning a total of 504.7 megabases (fig. S1). The median density of RPRs was 4,754 per megabase with minor variations across chromosomes (fig. S1). A bootstrap analysis performed by removing one healthy sample demonstrated

reproducibility of the RPR map with a minimum of 81.6% similarity between any two iterations (fig. S2). We developed a metric for fragment position aberrancy by quantifying fragments that intersected RPRs. Fragments with one or both end positions within RPRs were identified as aberrant and those that spanned the length of the RPRs were identified as non-aberrant. In plasma samples from healthy individuals, we found that fragment length and GC-content influenced the probability of aberrancy (fig. S3). Therefore, each fragment's contribution to this metric was normalized based on these factors. This resulted in a single information-weighted fraction of aberrant fragments (iwFAF) value for each sample.

To ensure our findings would be generalizable across cancer types, disease stages, and pre-analytical factors (such as differences in sample processing or sequencing instrument), we performed genome-wide analysis of fragment ends using data that we generated from 521 sequencing libraries as well as published sequencing data from 2,147 plasma DNA samples (12, 14–16). Overall, these 2,668 plasma samples represented 286 healthy individuals, 994 patients with cancer (across 11 cancer types), and 103 individuals with other non-malignant disease.

In sequencing data that we generated and analyzed, compared to 24 plasma samples from healthy individuals, mean iwFAF was higher in 47 samples from patients with early-stage breast cancer ($P = 2.20 \times 10^{-4}$), in 39 samples from patients with cholangiocarcinoma ($P = 1.01 \times 10^{-9}$), in 45 samples from patients with glioblastoma ($P = 2.27 \times 10^{-4}$), and in 261 samples from patients with melanoma ($P = 2.11 \times 10^{-4}$; Fig. 1A, table S1, table S2, data file S1, data file S2, and data file S3). In published datasets, we similarly found that mean iwFAF was higher across multiple cancer types, compared to corresponding healthy cohorts (Fig. 1A, table S2, data file S4, and data file S5). When iwFAF was compared across three independent sets of healthy individuals, no significant difference was observed ($P = 0.437$, one-way ANOVA). In addition, plasma samples from 67 patients with chronic hepatitis B without cirrhosis and from 36 patients with hepatitis B-associated liver cirrhosis were not distinguishable from corresponding healthy individuals. However, plasma samples from patients with hepatocellular carcinoma showed higher iwFAF compared to healthy individuals ($P = 2.86 \times 10^{-6}$; Fig. 1A).

Comparison between aberrant fragmentation and tumor fraction in cfDNA

To compare iwFAF to fraction of tumor DNA in plasma, we measured tumor fraction using analysis of copy number aberrations in patients with advanced cancer (15, 17). Across 938 samples with at least 3% tumor fraction from patients with cancer and 24 samples from healthy controls, iwFAF was strongly correlated with tumor fraction (Spearman's $\rho = 0.77$, $P = 4.66 \times 10^{-190}$; Fig. 1B). To ascertain whether aberrant DNA fragments in plasma were disproportionately contributed by the tumor, we focused our analysis on plasma DNA samples with high tumor fraction from patients with metastatic melanoma (18). Tumor contribution to plasma DNA from different genomic regions is influenced by copy number. If aberrant DNA fragments are more likely to be tumor-derived, we expect iwFAF to be higher for genomic loci affected by copy number amplification. In 27 plasma samples with at least 20% tumor fraction from 14 patients with metastatic melanoma, we found higher iwFAF in regions affected by copy number gain compared to regions unaffected by copy

number change or those affected by copy number loss (Fig. 1C and fig. S4; $P < 1 \times 10^{-24}$ for both comparisons, Mann-Whitney U test). To further assess the tumor specificity of aberrant DNA fragments in plasma, we performed deep whole-genome sequencing at greater than 375× coverage for plasma samples from two patients with metastatic melanoma with tumor fractions of 36% and 39%. We evaluated DNA fragments at positions with tumor-specific single-nucleotide variants. In both plasma samples, we found mutated fragments were more likely to be aberrant compared to non-mutated fragments (Fig. 1D and table S3; $P = 3.6 \times 10^{-4}$ and $P = 1.6 \times 10^{-15}$, two-proportion Z-test).

In longitudinal plasma DNA samples from patients with metastatic melanoma, changes in iwFAF were consistent with changes in tumor fraction (Fig. 2A, Fig. 2B, fig. S5, and data file S3). In patients with glioblastoma, we compared iwFAF with tumor fraction in plasma DNA measured using targeted digital sequencing (TARDIS) (19, 20). We found that longitudinal changes in iwFAF were consistent with changes in tumor fraction, even though measured tumor fraction ranged from 0.01% to 1.2% (Fig. 2C, Fig. 2D, and data file S6). We compared the difference in iwFAF and tumor fraction between any two consecutive samples where both had quantifiable tumor fraction. Changes in iwFAF and tumor fraction were correlated in patients with melanoma and glioblastoma (data file S7, Spearman's $\rho = 0.68$ and 0.67, $P = 1.02 \times 10^{-9}$ and 3.47×10^{-3}, respectively). Because the range of calculated iwFAF is narrow (0.59 to 0.68), we scaled iwFAF between 0 and 1, such that 0 represented iwFAF in a healthy sample and 1 represented the highest iwFAF measured in a cancer sample. Changes in scaled iwFAF and tumor fraction were comparable in magnitude for patients with metastatic melanoma (fig. S6).

Evaluation of potential pre-analytical confounders affecting measurement of iwFAF

To measure the contribution of potential confounders to analysis of fragment ends, we performed multiple comparisons across demographics, sample processing conditions, and replicate sequencing runs. In plasma DNA samples from 196 samples from healthy volunteers, we found no significant difference in iwFAF across 4 age groups (<50 years, 50–54 years, 55–59 years, and ≥ 60 years; $P = 0.573$, one-way ANOVA; fig. S7 and data file S8). In the same dataset, we observed no significant difference in iwFAF between male and female healthy individuals ($P = 0.310$; fig. S7 and data file S8). We collected three matched samples using different blood collection tube types (EDTA, PAXgene, and HAEM-Lok) from 24 healthy individuals. We extracted plasma DNA and prepared sequencing libraries independently from each of the three samples. iwFAF was strongly correlated between these replicates and no significant difference was observed across the three tube types (pairwise Spearman's ρ 0.93 to 0.96, $P = 0.95$, one-way ANOVA; fig. S8 and data file S9). For 24 patients with early-stage breast cancer, we extracted DNA using two different methods from matched plasma aliquots from the same blood tube and prepared independent sequencing libraries. iwFAF was strongly correlated between the two measurements (Spearman's $\rho = 0.90$, $P = 2.8 \times 10^{-9}$). In paired comparison, iwFAF was significantly lower for plasma DNA extracted using the Qiagen spin-column method compared to MagMax magnetic beads ($P = 5.9 \times 10^{-4}$, paired T-test; fig. S9 and data file S9). For 41 plasma DNA samples from patients with metastatic melanoma, we prepared libraries and generated sequencing data using two different Illumina sequencing platforms (NextSeq 500 and NovaSeq 6000).

iwFAF was strongly correlated between the two measurements (Spearman's $\rho = 0.99$, $P = 2.9 \times 10^{-40}$). In paired comparisons, iwFAF was significantly lower from sequencing data generated on the NovaSeq compared to NextSeq ($P = 1.06 \times 10^{-24}$, paired T-test; fig. S10 and data file S9). However, the effect sizes observed across DNA extraction methods and replicate sequencing runs (Cohen's d of 0.226 and 0.116, respectively) suggested these factors make very small contributions (if any) to changes in iwFAF, far less than the magnitude of differences observed between cancer patients and healthy individuals (table S2).

To evaluate whether iwFAF is an indirect measure of short plasma DNA fragment proportion, we compared iwFAF with the fraction of short fragments (defined as less than 150 bp) across 196 samples from healthy individuals, and found a modest positive correlation (Spearman's $\rho = 0.435$, $P = 1.8 \times 10^{-10}$; fig. S11A). We further compared iwFAF with plasma DNA concentration across 174 samples from healthy volunteers and found a weak negative correlation (Spearman's $\rho = -0.28$, $P = 1.6 \times 10^{-4}$; fig. S11B). Because total plasma DNA concentration is higher on average in patients with cancer compared to healthy individuals (21), differences in plasma DNA concentration across samples are unlikely to explain the observed increase in iwFAF in patients with cancer.

Evaluation of differences in genomic positioning of plasma DNA fragments by measuring nucleotide frequencies at fragment ends

Calculation of iwFAF relies on inferred RPR maps and hence this approach excludes any fragments that do not intersect a known RPR. This approach limits the proportion of informative data to the annotated region of the genome. In 2,489 samples analyzed in this study, a mean of 84.1% of fragments were used in iwFAF calculations. To maximize utilization of all available data from each sample independent of available genomic annotation, we developed a complementary method to evaluate differences in genomic positioning of plasma DNA fragments that does not rely of annotation for genomic features such as RPRs. Using sequencing reads aligned to the reference genome sequence, we calculated nucleotide frequencies for each position 10 bp upstream and downstream of both fragment ends, averaged across all fragments for each sample (22). This results in 168 measurements of mean nucleotide frequency per sample (4 nucleotides × 21 genomic loci × 2 fragment ends; Fig. 3A). We performed multidimensional scaling and compared the first two dimensions of mean nucleotide frequencies with iwFAF for samples from two cohorts of patients with metastatic cancers. Absolute values for correlation between the second dimension of nucleotide frequencies at fragment ends and iwFAF were 0.62 ($P = 3.18 \times 10^{-60}$) and 0.59 ($P = 4.40 \times 10^{-27}$) for patients with breast cancer and prostate cancer, respectively (table S4). To identify specific nucleotide positions that may capture differences in fragment end positioning, we calculated the correlation between iwFAF and each nucleotide frequency for both cohorts (Fig. 3B, fig. S12). Some positions, such as the second and third base on the inside of fragments (positions 1, 1', and 2, 2'), showed stronger correlation compared to others, such as the first base inside the fragment end (positions 0, 0') or the fourth base inside the fragment end (positions 3, 3'). We selected positions with a summed nucleotide frequency correlation coefficient of at least 1.0 in both cohorts. To adjust for internal correlation between nucleotide frequencies, we performed multivariate linear

regression to predict iwFAF using the 64 nucleotide frequencies at these 16 positions (Fig. 3C). Nine nucleotide frequencies with the highest adjusted mean coefficient magnitudes were located at only 3 positions across both fragment ends: position −1 (first position outside the left fragment end) and positions 1’ and 2’ (second and third positions inside the right fragment end).

Development of a machine-learning model to distinguish plasma from patients with cancer and healthy individuals

To evaluate whether genome-wide analysis of fragment ends enables detection of cancer, we trained random-forest machine-learning models to distinguish between plasma samples from patients with cancer and healthy individuals. Samples from patients with non-malignant disease were excluded. To avoid overfitting our classification model, we restricted the analysis to the earliest available plasma sample for each patient, generally obtained at enrollment in the clinical study (table S5). Cross-validation analysis was averaged over 100 runs, using 80% of samples for training and 20% for testing in each iteration, split proportionally for each cohort (23). This dataset and partitioning strategy were used for all following analyses.

A linear model trained using iwFAF as a single feature showed an area under the receiver operating characteristics curve (AUC) value of 0.78 (Fig. 4A and table S6). A similar model trained using fraction of short fragments as a single feature showed an AUC value of 0.65 (fig. S13 and table S6). We tested a model based on the 9 nucleotide frequencies most correlated with iwFAF and found an AUC value of 0.89 (Fig. 4A). For our final classification model, GALYFRE, we incorporated iwFAF together with the 9 most correlated nucleotide frequencies. Empirical evaluation showed that at higher model depths, the difference in mean AUC between training and validation data increased, suggesting the potential for over-fitting (fig. S14 and table S7). Based on this observation, we limited model depth to 5. GALYFRE achieved an averaged AUC value of 0.91 (Fig. 4A). To further validate the performance of this approach, we repeated model training while holding out 20% of the samples, randomly chosen and excluded from training and testing during cross-validation. Runs performed with hold-out data achieved a similar averaged AUC value of 0.91 (fig. S15 and table S6). As expected, classification performance was influenced by cancer stage, with AUC values of 0.87 for patients with Stage I cancer to 0.91 for patients with Stage IV cancer (Fig. 4B). Performance also varied across cancer types (fig. S16 and fig. S17). AUC values for 6 of 10 tested cancer types were greater than 0.9 with the lowest AUC of 0.82 observed for patients with ovarian cancer (fig. S16). At 95% specificity, mean sensitivity across 100 cross-validation runs was 66.9% (95% CI 66.1% to 67.8%) across all cancer types, highest at 94.3% and 90.8% for patients with glioblastoma and cholangiocarcinoma and lowest at 45.5% and 53.8% for patients with ovarian and breast cancer (table S8). To evaluate the relative contribution of iwFAF and each nucleotide frequency in GALYFRE, we calculated Shapley values and found that iwFAF was the most informative feature (fig. S18) (24). The next most informative nucleotide frequency represented the first position outside the left fragment end.

To estimate the minimum sequencing depth for GALYFRE, we sub-sampled the data to simulate low-depth sequencing. First, we generated 10 independent replicates across 1000 depths from 105 samples. Calculated iwFAF values were highly reproducible, with coefficient of variation ranging from 0.027% at 10 million fragments to 0.11% at one million fragments (fig. S19). We then randomly selected a dataset of one million fragments per sample. With GALYFRE trained on this low-depth dataset, the averaged AUC value was 0.91 (fig. S20 and table S6).

Discussion

Our results demonstrated that across multiple cancer types, the positions of tumor-derived plasma DNA fragment ends diverge from those of background DNA fragments contributed by peripheral blood cells. We leveraged this observation and showed proof-of-principle results that analysis of fragment end positions and their adjacent sequences can be useful as a biomarker for cancer detection. Our approach used a machine-learning model trained on only 10 features derived from genome-wide assessment of fragment end positions, and we showed that this approach has potential relevance for earlier detection of multiple cancer types, including those with no established methods for screening, such as cholangiocarcinoma, pancreatic cancer, and gastric cancer.

Earlier studies of fragmentation patterns in circulating tumor DNA evaluated local differences in average fragment size in windows across the genome as an approach for cancer detection (12) or used fragment size to improve sensitivity for detection of somatic genomic alterations (25, 26). Analyses at the individual fragment level identified genomic loci (27) and nucleotide motifs (13, 28) preferentially found in DNA shed from liver cells and found liver-derived DNA was higher in patients with hepatocellular carcinoma. One study reported biased representation of nucleotide frequencies at plasma DNA fragment ends (29). In contrast to these studies, we found sample-level aggregated measurement of fragment end positions can serve as a biomarker for detection of multiple cancer types. Compared to earlier reports in which machine learning was utilized for cancer detection using plasma DNA analysis (based on fragment sizes or methylation marks), our approach achieves comparable classification performance despite reliance on a much simpler model with a limited number of features (6, 12). Combined with our reliance on a random forest model (compared to gradient boosted trees or neural networks), the use of a simpler, low feature model reduces the likelihood for overfitting to confounders (such as technical and pre-analytical differences between sample sets) and aids interpretability (30). When machine-learning models use thousands of features to discriminate between cancer and healthy samples, it is often unclear what specific biological features drive performance (31). Our approach combines a genome-wide metric of differences in fragment end positions and mean nucleotide frequencies from 3 loci surrounding fragment end positions. The three loci that drive performance in our approach include the first base on the outside and the second and third bases on the inside of fragment ends. Unlike earlier approaches that use 4 bp end motifs from sequenced fragments (13), our model does not rely on the first or the fourth bases on the inside of fragment ends. In our analysis, the nucleotide frequency at the first base across fragments did not contribute to classification between cancer patients and healthy individuals, potentially because it is primarily driven by enzymatic preference

for plasma DNA degradation that affects fragments from peripheral blood cells and tumor cells equally (32), instead of differences in fragment end positions driven by variation in chromatin accessibility across cells.

The classification performance of GALYFRE in patients with glioblastoma is particularly surprising, given how challenging circulating tumor DNA detection has been for these patients using mutation-based assays (33). A potential explanation is that, unlike mutation-based assays, analysis of fragment ends leverages differences between cell-free DNA shed from peripheral blood cells compared with cell-free DNA from a combination of malignant cells and microenvironment cells constituting the tumor. Hence, GALYFRE may perform better for cancers originating in tissues that rarely contribute cell-free DNA into plasma within healthy individuals. However, this also indicates a potential limitation that aberrant fragmentation patterns in plasma may not be specific to cancer and may arise from unexpected tissue contributions in plasma due to other systemic or acute conditions including pregnancy and transplant (13). Because analysis of fragmentation patterns is likely to be less cancer-specific than analysis of somatic mutations, it is even more relevant to delineate the effects of biological and technical pre-analytical factors. In our approach, we observed that the fraction of aberrant fragments was affected by differences in fragment size and GC content. After applying an approach that we developed to normalize the contribution of these factors, we observed that iwFAF was not significantly affected by age or gender and remained consistent across replicate analyses of DNA extraction methods, blood collection tubes, and sequencing runs. In particular, we did not find elevated iwFAF in plasma samples from patients with hepatitis (with or without liver cirrhosis), whereas a higher iwFAF was observed in patients with hepatocellular carcinoma. These observations suggested that increases in iwFAF and aberrant fragmentation are greater in magnitude in patients with cancer compared to patients with non-malignant inflammatory conditions, potentially due to higher rates of cell-free DNA shedding from tumors.

Although our current results for classification between patients with cancer and healthy individuals are encouraging, there are multiple limitations of this study and potential opportunities for further improvement. In the current study, cases and controls were not matched for age, sex, or co-morbidities. To develop this approach further for early cancer detection, a larger reference dataset is needed, including healthy individuals across age, sex, and a wide range of co-morbidities, as well as plasma samples obtained in patients with non-malignant acute and chronic inflammatory conditions. Each patient’s results may then be obtained when they are unaffected by acute illness, compared with matched reference samples and interpreted in the appropriate clinical context. Although multiple cancer types are represented in our results, the number of cases for some cancer types were small (such as lung cancer or glioblastoma) or unevenly distributed across different disease stages (such as melanoma). In addition, performance of the machine-learning model may benefit from incorporation of demographic information and co-morbidities, together with fragmentation patterns (34). Our approach can also be improved and characterized further through evaluation of alternative library preparation approaches, such as single-stranded DNA sequencing, or alternative approaches for machine learning and through analysis of additional numbers of samples from patients across disease stages for each cancer type to increase accuracy of cancer detection. Data from specific cancer types may be useful

to predict tumor type for plasma samples from cancer patients, through either selection of the most informative genomic regions to calculate iwFAF, and by identifying cancer type-specific nucleotide frequencies at fragment ends.

When tumor fractions in plasma DNA samples are low, such as in patients with early-stage cancers or in mid-treatment samples, precision in measurement of tumor fraction using mutation-based assays is limited by the number of somatic mutations and the amount of DNA analyzed (19). Because GALYFRE relies on aggregate analysis of DNA fragments from across the genome, we found high precision in iwFAF across repeated samples with a coefficient of variation of just 0.1% when 1 million fragments were resampled. In serial samples from patients with metastatic melanoma and glioblastoma, we found that changes in iwFAF were correlated with changes in tumor fraction based on somatic genomic alterations, although the magnitude of corresponding changes was much smaller in iwFAF. Overall, iwFAF values in our study spanned a narrower range compared to tumor fraction measurements. When iwFAF was scaled, serial changes in the two measurements became more comparable in patients with metastatic melanoma, but not in patients with glioblastoma, suggesting that future studies should assess if appropriate quantitative scaling of iwFAF values specific to each cancer type are needed to apply this approach for monitoring of treatment response. In addition to quantitative precision, we found that the performance of GALYFRE for cancer detection using just 1 million fragments per sample parallels published methods(12, 13). Because GALYFRE requires a limited depth of sequencing and low amount of input DNA to achieve reproducible performance for cancer detection and quantification of tumor fraction, we predict that such data can be obtained from small volumes of blood or dried blood spots (35) and that reaction volumes for sequencing library preparation can be reduced to lower assay costs.

In summary, we developed an approach for analysis of plasma DNA fragment end positions and showed that the results of this analysis hold potential as a biomarker for cancer diagnostics. The simplicity of our approach, as well as the small amount of plasma DNA and sequencing data required, can increase access to blood-based cancer detection and monitoring, particularly for resource-constrained health systems. Our results serve as an encouraging proof-of-principle, but additional case-control studies are needed to establish quantitative thresholds for both early detection and monitoring treatment response in patients with cancer. Once such thresholds are identified, prospective evaluation of real-world diagnostic performance in clinical cohorts will be required.

Materials and Methods

Study design

The aim of this study was to investigate differences in fragmentation patterns in plasma DNA between patients with cancer and healthy individuals. Whole genome sequencing data was generated from plasma DNA samples. Additional sequencing data from FinaleDB was used in the analysis(14). Computational methods and models were developed retrospectively using a combination of these datasets. Different computational approaches to identify and quantify differences in fragment end characteristics were evaluated retrospectively. Prior power analysis, randomization, or blinding was not performed for the clinical study.

Patient plasma sample collection and processing

Healthy volunteers were enrolled at the Translational Genomics Research Institute in Phoenix, AZ, and blood samples were collected under protocol numbers 20142638 and 20181812, approved by the Western Institutional Review Board (IRB). Blood and tissue samples from patients with melanoma were collected at the Mayo Clinic in Arizona under protocol number 16–001453 and within a multi-center clinical trial (NCT02094872) under protocol number 20140190 approved by the Western IRB(18). Blood samples from patients with breast cancer were collected at the Mayo Clinic in Arizona under protocol number 14–006021, from patients with glioblastoma within a clinical trial (NCT02060890) at the University of California San Francisco, in California under protocol number 20141201 approved by Western IRB (20), and from patients with cholangiocarcinoma at the Mayo Clinic in Arizona under protocol number 12–004713. All patients provided informed consent. For a subset of patients with cancer, multiple blood samples were collected including at presentation and during treatment and follow-up.

Blood samples were collected in K_2 EDTA tubes. Plasma was separated within 3 hours of venipuncture by centrifugation at 820*g* for 10 minutes, followed by a second centrifugation at 16,000*g* for 10 minutes. One milliliter aliquots of plasma were stored at −80°C until DNA extraction. In a subset of healthy individuals, additional matched blood samples were also collected in PAXgene cell-free DNA tubes (Qiagen) and HAEM-Lok tubes (DeltaDNA Biosciences), and a comparison of iwFAF across blood tubes was performed.

DNA was extracted using either the MagMAX Cell-Free DNA Isolation Kit (Thermo Fisher Scientific) or QIAamp Circulating Nucleic Acid Kit (Qiagen) from 1 to 4 ml plasma. Cell-free DNA was quantified prior to library preparation using the Qubit dsDNA HS assay (Thermo Fisher Scientific), cell-free DNA ScreenTape analysis on the TapeStation 4200 (Agilent), or using an in-house digital PCR assay(36). Whole genome sequencing libraries were prepared from plasma DNA using ThruPLEX Plasma-Seq or Tag-seq library preparation kits (Takara).

External data

Fragment end positions and clinical annotation for an additional 1,798 samples from cancer patients, 103 samples from patients with non-malignant disease, and 246 samples from healthy individuals were obtained from FinaleDB(14). These data were aggregated from three previously published studies(12, 15, 16). BEDTools v2.29.0 (37) was used for all associated analyses.

Tumor fraction determination

Tumor fraction was inferred through copy number analysis (CNA) of plasma DNA using HMMcopy and ichorCNA v0.3.2(15, 17, 38–40). Because the reported limit of detection for ichorCNA is 3% tumor fraction, any samples with ichorCNA-inferred tumor fractions below this threshold were excluded from tumor fraction correlation analyses.

Tumor fraction in plasma samples from patients with glioblastoma was measured using targeted digital sequencing (TARDIS)(19). Briefly, patient-specific somatic mutations

were identified by analyzing exome sequencing data from tumor biopsies and germline DNA. Clonal mutations were selected as targets for amplicon sequencing, adjusting for copy number aberrations in the tumor genome and overall tumor purity. Target-specific multiplexed primers were designed and evaluated for in vitro performance using control DNA samples. TARDIS sequencing libraries were prepared and sequenced on a NovaSeq 6000 (Illumina). Sequencing data were analyzed to evaluate targeted genomic loci and quantify circulating tumor DNA detection in each sample. Tumor fraction was calculated as the mean of all measured variant allele fractions.

Identifying recurrently protected regions

A map of recurrently protected regions (RPRs) was inferred from 17 plasma samples from heathy individuals (sequenced to mean coverage of 40×, range = 28× to 48×), using the Windowed Protection Score (WPS) peak-calling method described by Snyder *et. al.*(8). This was performed using 120 – 180 bp fragments across chromosomes 1 – 22. The EXTREGION parameter used for each of these chromosomes was the entire length of the chromosome. Read start sites were extracted using the recommended boundary of 200 bp inwards from EXTREGION (to guarantee inclusion of these 120 – 180 bp reads), resulting in REGION parameter value starting at 201 bp and ending at 200 bp less than the chromosome size. The recommended window size of 120 bp was used to calculate WPS values.

To evaluate the robustness of our RPR map, we generated a series of bootstrapped maps by removing one of the 17 healthy samples in each case. The total numbers of base pairs of intersections and unions of these maps were then used to compute the Jaccard similarity for each pair of RPR maps. This evaluation was also performed for the number of RPRs identified. Finally, the mean score assigned to RPRs in the intersecting region of each pair of maps was calculated (fig. S2).

Analysis of aberrant fragment end positions

A fragment position aberrancy metric was developed by quantifying fragments intersecting RPRs. Using the RPR map, cell-free DNA fragments were identified as aberrant if one or both ends were located within a protected region. Non-aberrant fragments were identified as those spanning the length of a protected region. Fragments that had no intersection with any RPR were excluded. We found that the probability of a given fragment being aberrant was influenced by fragment size and GC content (fig. S3). Because fragment-size distribution and GC-content distribution can be influenced by pre-analytical factors, we normalized for these fragment features. Each fragment's contribution to the final metric was weighted based on the probability of the fragment being aberrant or non-aberrant given its size and GC content. These probabilities were calculated using healthy samples from this study, Cristiano *et al.(12)*, and Jiang *et al.* (16) separately. For our dataset, we used 16 samples from healthy individuals (37× total genomic coverage)(9). For the external datasets, 30% of the healthy samples, with an equal number of male and female samples were combined and used for probability calculations including 64 samples from Cristiano *et al.* and 10 samples from Jiang *et al.* (total genomic coverage of 495× and 62×, respectively). Healthy samples used for normalization were excluded from downstream analyses of iwFAF and

model cross-validation. For each dataset, a normalization table was generated by calculating the probability of a fragment being aberrant given its size (P_{size}) and GC content (P_{GC}).

The following equations 1, 2, and 3 were used to calculate a sample-wide weighted aberrant fragmentation metric (iwFAF):

$$W_{aberrant} = \sum_{n=0}^{n} log_2\left(\frac{1}{P_{size} * P_{GC}}\right) \quad (1)$$

$$W_{non-aberrant} = \sum_{n=0}^{n} log_2\left(\frac{1}{(1-P_{size}) * (1-P_{GC})}\right) \quad (2)$$

$$iwFAF = \frac{W_{aberrant}}{W_{aberrant} + W_{non-aberrant}} \quad (3)$$

Analysis of fragment end position nucleotide frequencies

Positions from 10 bp upstream to 10 bp downstream of each fragment end base were considered, for a combined 21 bases on each fragment end. For each sample, mean base frequencies at each position were calculated for all fragments, based on the nucleotide from the hg19 reference genome, using homerTools v4.11(22). Each sample was represented by a vector of length 168 (21 positions × 2 fragment ends × 4 bases).

To infer the utility of nucleotide frequency for tumor detection, nucleotide frequencies for samples with at least 3% tumor fraction from patients with cancer were reduced to two dimensions using multidimensional scaling. Pairwise distances between points in the 168-dimensional vector space were calculated using cosine distance. This was done separately for four cancer types: cholangiocarcinoma, melanoma, breast cancer, and prostate cancer. Spearman correlations with tumor fraction and iwFAF were then evaluated.

To reduce the influence of pre-analytical factors on measured nucleotide frequencies and to enrich for tumor-derived signal, we analyzed the correlation of each of the 168 nucleotide frequency values with iwFAF. Position-wise cumulative correlation magnitudes were calculated for metastatic prostate cancer ($n = 553$) and metastatic breast cancer ($n = 948$), both from the Adalsteinsson *et. al.* dataset(15). Positions with a cumulative correlation greater than 1.0 for each cancer type were then identified, and their intersection was used to select 16 informative positions. The corresponding 64 (16 positions × 4 bases) nucleotide frequencies were then used to make generalized linear models (GLMs) to predict iwFAF for each of the two cancer types. The absolute values of coefficients were calculated for each GLM, and the mean was taken for each of the 64 nucleotide frequencies. The 9 nucleotides with the greatest coefficient magnitudes were then selected for GALYFRE.

Cancer sample classification model

Using GALYFRE, a combination of iwFAF and the 9 selected nucleotide frequency features, we built a random forest model to distinguish between samples from healthy individuals and patients with cancer. The data used for building this model were limited to one sample per patient (the earliest time point available for each), to avoid potential signal leakage between training and validation data. Samples from 196 healthy individuals and 465 patients with cancer across 10 cancer types (table S5) were stratified by cancer type (and a single stratification for healthy samples) and split into 80% training and 20% validation data. Such stratified splits ensure that training and validation data have similar representation leading to improved generalization on validation data(23).

To address imbalance in representation by cancer type, training data samples were upsampled at random to generate a uniform number of samples for each cancer type that was equal to the number of healthy samples. A minimum of one copy per sample was included. To address class imbalance, the total number of healthy samples in the training data was then upsampled at random so that number of healthy samples and cancer samples was equal. This results in a default classification accuracy of 50% for the binary classifier for training data. Such resampling was not done for validation or hold-out data.

A random forest classifier using 100 decision trees was trained and evaluated over 100 runs using 1000 activation thresholds uniformly distributed between 0 and 1. This binary classifier was trained using a label of 0 for healthy samples and 1 for samples from patients with cancer. Each decision tree considered a random selection of 3 features for a random sample of 70% of the training data. The decision trees were subject to a maximum depth of 5 and a minimum leaf size of 5 observations for pre-pruning. Decision trees were not subject to purity-based post-pruning. During training, for all 100 runs the depths of trees for the learned model were 5.

For interpretation of the learned models, Shapley values (24) were calculated for the features of each model using the training data for reference. A combination of training and validation data was used to calculate Shapley values using 100 Monte Carlo simulations. Feature importance was calculated using mean magnitude of Shapley effect in the binary classification dataset (fig. S18).

For more rigorous evaluation of GALYFRE, 20% of the data was held out before any training. The remaining data were split into 80% for model learning and 20% for model validation. For 100 runs, the learned model was evaluated on both the run’s validation data and the hold-out data. This was averaged over 10 repetitions (fig. S15).

To further analyze classifier models for overfitting, we evaluated the impact of tree depth on classifier performance based on 100 runs (fig. S14).

Statistical analysis

Statistical analyses were performed using Julia and Python(41, 42). Significance values of differences between two iwFAF distributions were evaluated using the t-test and Cohen’s d effect size. Statistical significance between distributions of iwFAF in copy number

loss, neutral, or gain regions was calculated using the Mann-Whitney U test. Correlations were calculated using Spearman's ρ. To compute the statistical significance of correlation, the correlation coefficients were first converted to a t-statistic and then P-value was calculated based on population size. Matched samples were compared using the paired t-test. Comparison of iwFAF between mutated and non-mutated DNA fragments within a plasma sample was performed using the two-proportion Z-test. One-way ANOVA was performed to evaluate differences in iwFAF between 3 or more groups. All reported P-values are two-sided; P-values below 0.05 were considered statistically significant. Bonferroni correction was performed for multiple comparison testing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We would like to thank B. Moore, D. Metz, and S. Buchholtz at TGen, and the volunteers and patients who participated in this study. Editorial services were provided by Nancy R. Gough (BioSerendipity, LLC).

Funding:

Supported by funding from the Ben and Catherine Ivy Foundation to MM, JMT and SC, from the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under award number 1U01CA243078–01A1 to MM and 1R01CA223481–01 to MM, and by a Stand Up To Cancer (SU2C) – Melanoma Research Alliance Melanoma Dream Team Translational Cancer Research Grant (#SU2C-AACR-DT0612) to JMT and PML. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research (AACR).

References and Notes

1. Wong FCK, Lo YMD, Prenatal Diagnosis Innovation: Genome Sequencing of Maternal Plasma. Annual Review of Medicine 67, 419–432 (2016).
2. Burnham P, Khush K, De Vlaminck I, Myriad Applications of Circulating Cell-Free DNA in Precision Organ Transplant Monitoring. Annals of the American Thoracic Society 14, S237–S241 (2017). [PubMed: 28945480]
3. Van Der Pol Y, Mouliere F, Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. Cancer Cell 36, 350–368 (2019). [PubMed: 31614115]
4. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A, Hruban RH, Wolfgang CL, Goggins MG, Dal Molin M, Wang TL, Roden R, Klein AP, Ptak J, Dobbyn L, Schaefer J, Silliman N, Popoli M, Vogelstein JT, Browne JD, Schoen RE, Brand RE, Tie J, Gibbs P, Wong HL, Mansfield AS, Jen J, Hanash SM, Falconi M, Allen PJ, Zhou S, Bettegowda C, Diaz LA Jr., Tomasetti C, Kinzler KW, Vogelstein B, Lennon AM, Papadopoulos N, Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 359, 926–930 (2018). [PubMed: 29348365]
5. Hu Y, Ulrich BC, Supplee J, Kuang Y, Lizotte PH, Feeney NB, Guibert NM, Awad MM, Wong KK, Janne PA, Paweletz CP, Oxnard GR, False-Positive Plasma Genotyping Due to Clonal Hematopoiesis. Clin Cancer Res 24, 4437–4443 (2018). [PubMed: 29567812]
6. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Consortium C, Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol 31, 745–759 (2020). [PubMed: 33506766]
7. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, Samet Y, Maoz M, Druid H, Arner P, Fu KY, Kiss E, Spalding KL, Landesberg G, Zick A, Grinshpun A, Shapiro AMJ, Grompe M, Wittenberg AD, Glaser B, Shemer R, Kaplan T, Dor Y, Comprehensive human cell-type

methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun 9, 5068 (2018). [PubMed: 30498206]

8. 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 164, 57–68 (2016). [PubMed: 26771485]
9. Markus H, Zhao J, Contente-Cuomo T, Stephens MD, Raupach E, Odenheimer-Bergman A, Connor S, McDonald BR, Moore B, Hutchins E, McGilvrey M, De La Maza MC, Van Keuren-Jensen K, Pirrotte P, Goel A, Becerra C, Von Hoff DD, Celinski SA, Hingorani P, Murtaza M, Analysis of recurrently protected genomic regions in cell-free DNA found in urine. Science Translational Medicine 13, eaaz3088 (2021). [PubMed: 33597261]
10. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, Dewitt WS, Lee C, Regalado SG, Read DF, Steemers FJ, Disteche CM, Trapnell C, Shendure J, A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174, 1309–1324.e1318 (2018). [PubMed: 30078704]
11. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, Chan W-K, Ma ESK, Chan SL, Cheng SH, Chan RWY, Tong YK, Ng SSM, Wong RSM, Hui DSC, Leung TN, Leung TY, Lai PBS, Chiu RWK, Lo YMD, Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proceedings of the National Academy of Sciences 112, E5503–E5512 (2015).
12. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, Jensen SØ, Medina JE, Hruban C, White JR, Palsgrove DN, Niknafs N, Anagnostou V, Forde P, Naidoo J, Marrone K, Brahmer J, Woodward BD, Husain H, Van Rooijen KL, Ørntoft M-BW, Madsen AH, Van De Velde CJH, Verheij M, Cats A, Punt CJA, Vink GR, Van Grieken NCT, Koopman M, Fijneman RJA, Johansen JS, Nielsen HJ, Meijer GA, Andersen CL, Scharpf RB, Velculescu VE, Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019). [PubMed: 31142840]
13. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, Heung MMS, Xie T, Shang H, Zhou Z, Chan RWY, Wong J, Wong VWS, Poon LC, Leung TY, Lam WKJ, Chan JYK, Chan HLY, Chan KCA, Chiu RWK, Lo YMD, Plasma DNA end motif profiling as a fragmentomic marker in cancer, pregnancy and transplantation. Cancer Discovery, CD-19–0622 (2020).
14. Zheng H, Zhu MS, Liu Y, FinaleDB: a browser and database of cell-free DNA fragmentation patterns. Bioinformatics 37, 2502–2503 (2021). [PubMed: 33258919]
15. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, Gydush G, Reed SC, Rotem D, Rhoades J, Loginov D, Livitz D, Rosebrock D, Leshchiner I, Kim J, Stewart C, Rosenberg M, Francis JM, Zhang C-Z, Cohen O, Oh C, Ding H, Polak P, Lloyd M, Mahmud S, Helvie K, Merrill MS, Santiago RA, O'Connor EP, Jeong SH, Leeson R, Barry RM, Kramkowski JF, Zhang Z, Polacek L, Lohr JG, Schleicher M, Lipscomb E, Saltzman A, Oliver NM, Marini L, Waks AG, Harshman LC, Tolaney SM, Van Allen EM, Winer EP, Lin NU, Nakabayashi M, Taplin M-E, Johannessen CM, Garraway LA, Golub TR, Boehm JS, Wagle N, Getz G, Love JC, Meyerson M, Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications 8, (2017).
16. Jiang P, Chan CWM, Chan KCA, Cheng SH, Wong J, Wong VW-S, Wong GLH, Chan SL, Mok TSK, Chan HLY, Lai PBS, Chiu RWK, Lo YMD, Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proceedings of the National Academy of Sciences 112, E1317–E1325 (2015).
17. Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K, Chin SF, Turashvili G, Hirst M, Caldas C, Marra MA, Aparicio S, Shah SP, Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res 22, 1995–2007 (2012). [PubMed: 22637570]
18. LoRusso PM, Sekulic A, Sosman JA, Liang WS, Carpten J, Craig DW, Solit DB, Bryce AH, Kiefer JA, Aldrich J, Nasser S, Halperin R, Byron SA, Pilat MJ, Boerner SA, Durecki D, Hendricks WPD, Enriquez D, Izatt T, Keats J, Legendre C, Markovic SN, Weise A, Naveed F, Schmidt J, Basu GD, Sekar S, Adkins J, Tassone E, Sivaprakasam K, Zismann V, Calvert VS, Petricoin EF, Fecher LA, Lao C, Eder JP, Vogelzang NJ, Perlmutter J, Gorman M, Manica B, Fox L, Schork N, Zelterman D, DeVeaux M, Joseph RW, Cowey CL, Trent JM, Identifying treatment options for

BRAFV600 wild-type metastatic melanoma: A SU2C/MRA genomics-enabled clinical trial. PLoS One 16, e0248097 (2021). [PubMed: 33826614]

19. McDonald BR, Contente-Cuomo T, Sammut S-J, Odenheimer-Bergman A, Ernst B, Perdigones N, Chin S-F, Farooq M, Mejia R, Cronin PA, Anderson KS, Kosiorek HE, Northfelt DW, McCullough AE, Patel BK, Weitzel JN, Slavin TP, Caldas C, Pockaj BA, Murtaza M, Personalized circulating tumor DNA analysis to detect residual disease after neoadjuvant therapy in breast cancer. Science Translational Medicine 11, eaax7392 (2019). [PubMed: 31391323]
20. Byron SA, Tran NL, Halperin RF, Phillips JJ, Kuhn JG, de Groot JF, Colman H, Ligon KL, Wen PY, Cloughesy TF, Mellinghoff IK, Butowski NA, Taylor JW, Clarke JL, Chang SM, Berger MS, Molinaro AM, Maggiora GM, Peng S, Nasser S, Liang WS, Trent JM, Berens ME, Carpten JD, Craig DW, Prados MD, Prospective Feasibility Trial for Genomics-Informed Treatment in Recurrent and Progressive Glioblastoma. Clin Cancer Res 24, 295–305 (2018). [PubMed: 29074604]
21. Meddeb R, Dache ZAA, Thezenas S, Otandault A, Tanos R, Pastor B, Sanchez C, Azzi J, Tousch G, Azan S, Mollevi C, Adenis A, El Messaoudi S, Blache P, Thierry AR, Quantifying circulating cell-free DNA in humans. Scientific Reports 9, (2019).
22. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK, Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell 38, 576–589 (2010). [PubMed: 20513432]
23. Wan N, Weinberg D, Liu T-Y, Niehaus K, Ariazi EA, Delubac D, Kannan A, White B, Bailey M, Bertin M, Boley N, Bowen D, Cregg J, Drake AM, Ennis R, Fransen S, Gafni E, Hansen L, Liu Y, Otte GL, Pecson J, Rice B, Sanderson GE, Sharma A, St J. Tang John, C., Tzou A, Young L, Putcha G, Haque IS, Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. BMC Cancer 19, (2019).
24. Štrumbelj E, Kononenko I, Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41, 647–665 (2014).
25. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, Mair R, Goranova T, Marass F, Heider K, Wan JCM, Supernat A, Hudecova I, Gounaris I, Ros S, Jimenez-Linan M, Garcia-Corbacho J, Patel K, Østrup O, Murphy S, Eldridge MD, Gale D, Stewart GD, Burge J, Cooper WN, Van Der Heijden MS, Massie CE, Watts C, Corrie P, Pacey S, Brindle KM, Baird RD, Mau-Sørensen M, Parkinson CA, Smith CG, Brenton JD, Rosenfeld N, Enhanced detection of circulating tumor DNA by fragment size analysis. Science Translational Medicine 10, eaat4921 (2018). [PubMed: 30404863]
26. Mouliere F, Robert B, Arnau Peyrotte E, Del Rio M, Ychou M, Molina F, Gongora C, Thierry AR, High fragmentation characterizes tumour-derived circulating DNA. PLoS One 6, e23418 (2011). [PubMed: 21909401]
27. Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, Wong J, Wong VWS, Chan HLY, Chan KCA, Lo YMD, Chiu RWK, Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci U S A 115, E10925–E10933 (2018). [PubMed: 30373822]
28. Chen L, Abou-Alfa GK, Zheng B, Liu JF, Bai J, Du LT, Qian YS, Fan R, Liu XL, Wu L, Hou JL, Wang HY, PreCar T, Genome-scale profiling of circulating cell-free DNA signatures for early detection of hepatocellular carcinoma in cirrhotic patients. Cell Res 31, 589–592 (2021). [PubMed: 33589745]
29. Chandrananda D, Thorne NP, Bahlo M, High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. BMC Med Genomics 8, 29 (2015). [PubMed: 26081108]
30. Liu H, in Encyclopedia of Machine Learning, Sammut C, Webb GI, Eds. (Springer US, Boston, MA, 2010), pp. 402–406.
31. James G, Witten D, Hastie T, Tibshirani R, An introduction to statistical learning (Springer, 2013), vol. 112.
32. Serpas L, Chan RWY, Jiang P, Ni M, Sun K, Rashidfarrokhi A, Soni C, Sisirak V, Lee WS, Cheng SH, Peng W, Chan KCA, Chiu RWK, Reizis B, Lo YMD, Dnase1l3 deletion causes aberrations in

length and end-motif frequencies in plasma DNA. Proc Natl Acad Sci U S A 116, 641–649 (2019). [PubMed: 30593563]

33. Muller Bark J, Kulasinghe A, Chua B, Day BW, Punyadeera C, Circulating biomarkers in patients with glioblastoma. Br J Cancer 122, 295–305 (2020). [PubMed: 31666668]
34. Tanos R, Tosato G, Otandault A, Al Amir Dache Z, Pique Lasorsa L, Tousch G, El Messaoudi S, Meddeb R, Diab Assaf M, Ychou M, Du Manoir S, Pezet D, Gagniere J, Colombo PE, Jacot W, Assenat E, Dupuy M, Adenis A, Mazard T, Mollevi C, Sayagues JM, Colinge J, Thierry AR, Machine Learning-Assisted Evaluation of Circulating DNA Quantitative Analysis for Cancer Screening. Adv Sci (Weinh) 7, 2000486 (2020). [PubMed: 32999827]
35. Heider K, Wan JCM, Hall J, Belic J, Boyle S, Hudecova I, Gale D, Cooper WN, Corrie PG, Brenton JD, Smith CG, Rosenfeld N, Detection of ctDNA from Dried Blood Spots after DNA Size Selection. Clin Chem 66, 697–705 (2020). [PubMed: 32268361]
36. Markus H, Contente-Cuomo T, Farooq M, Liang WS, Borad MJ, Sivakumar S, Gollins S, Tran NL, Dhruv HD, Berens ME, Bryce A, Sekulic A, Ribas A, Trent JM, LoRusso PM, Murtaza M, Evaluation of pre-analytical factors affecting plasma DNA analysis. Sci Rep 8, 7375 (2018). [PubMed: 29743667]
37. Quinlan AR, Hall IM, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). [PubMed: 20110278]
38. Chen S, Zhou Y, Chen Y, Gu J, fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890 (2018). [PubMed: 30423086]
39. Li H, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997, (2013).
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome S Project Data Processing, The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]
41. Bezanson J, Edelman A, Karpinski S, Shah VB, Julia: A fresh approach to numerical computing. SIAM review 59, 65–98 (2017).
42. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy C, Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 352 (2020). [PubMed: 32094914]

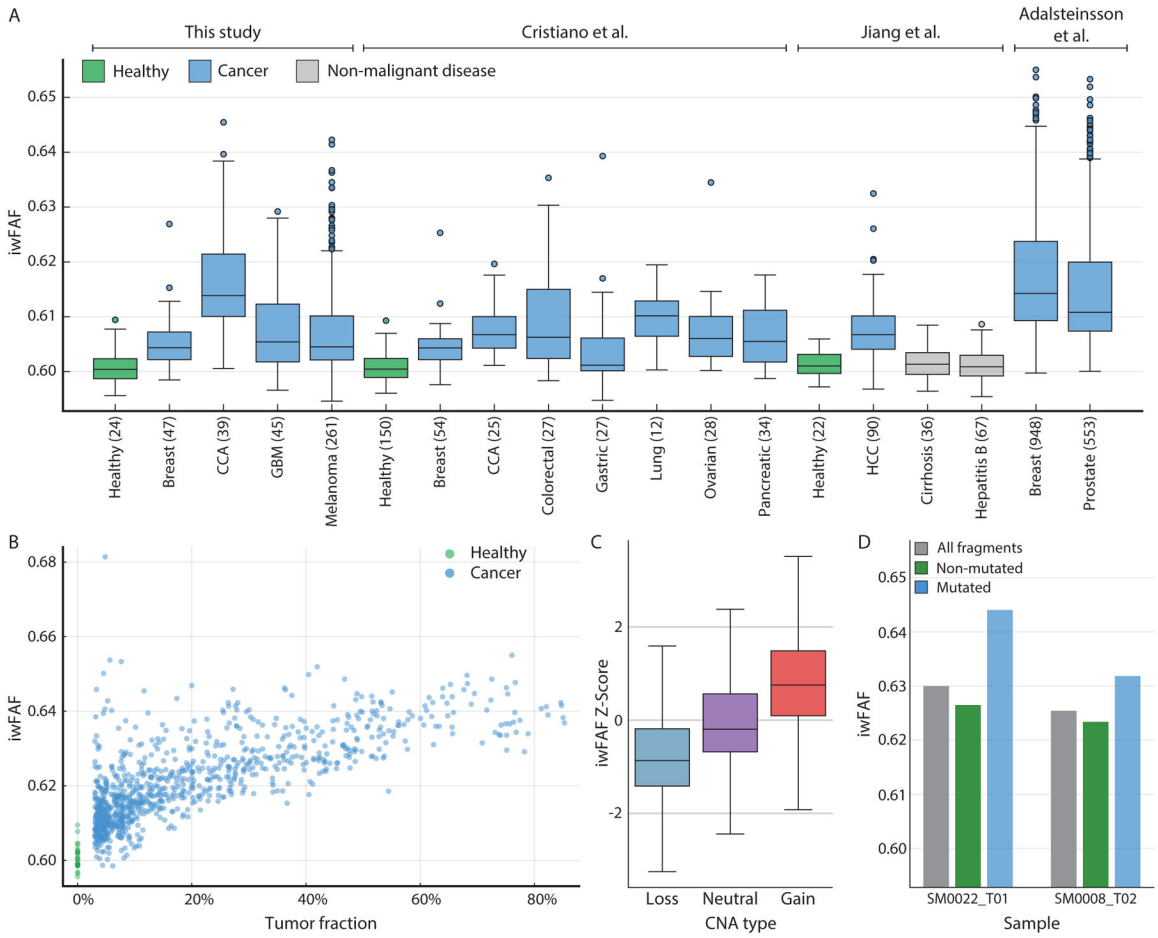

Fig. 1. Evaluation of information-weighted fraction of aberrant fragments (iwFAF) using plasma DNA whole-genome sequencing.

(A) Boxplots showing distributions of iwFAF values in plasma DNA of healthy individuals (green), patients with cancer (blue), and patients with non-malignant disease (gray) from four studies including the current study(12, 15, 16). The number of samples included in each category are indicated in parentheses. Each group of cancer patients and patients with non-malignant disease were compared to corresponding group of healthy individuals. P values for pairwise comparisons are reported in table S2. Two outliers (iwFAF of 0.6812 and 0.6814) were removed from the plot to improve visualization, both samples from patients with metastatic breast cancer in the Adalsteinsson *et al.* dataset(15). Abbreviations: CCA, cholangiocarcinoma; GBM, glioblastoma; HCC, hepatocellular carcinoma. (**B)** Scatterplot comparing tumor fraction with iwFAF in 938 samples from patients with cancer (blue) and 24 samples from healthy individuals (green). Plasma samples with at least 3% tumor fraction measured using ichorCNA were included in this comparison. Tumor fraction and iwFAF were strongly correlated (Spearman's $\rho = 0.77$, $P = 4.66 \times 10^{-190}$). (**C)** Boxplots show distribution of iwFAF z-scores in regions with copy number loss, neutral, or gain across 27 samples with at least 20% tumor fraction from patients with metastatic melanoma. Z-scores were calculated using the mean and standard deviation of copy number neutral regions from each patient. (**D)** Bar charts showing iwFAF values calculated from fragments overlapping tumor-specific single-nucleotide variants in plasma samples from two patients

with metastatic melanoma. iwFAF was calculated from all fragments (gray), fragments carrying the tumor-specific allele (blue), and fragments carrying the wild-type allele (green). iwFAF values for mutated fragments were significantly higher than mutated fragments ($P = 1.6 \times 10^{-15}$ and $P = 3.6 \times 10^{-4}$, two-proportion Z-test).

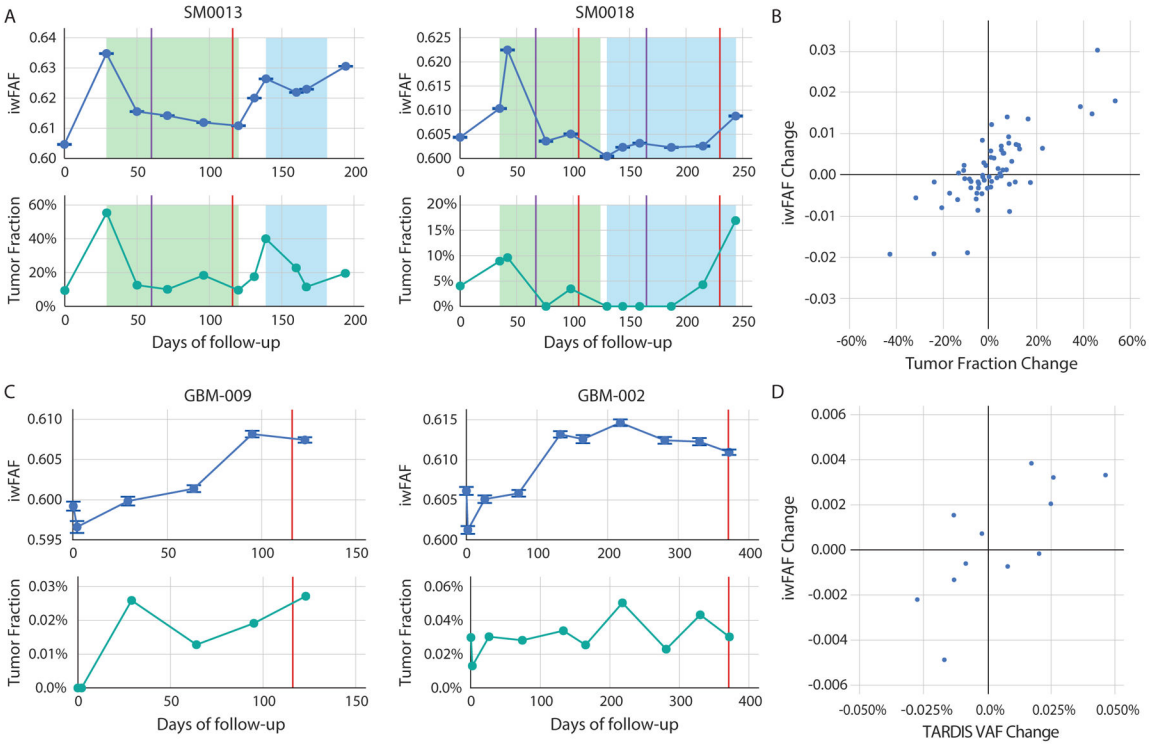

Fig. 2. Comparison of tumor fraction and iwFAF in longitudinal samples from patients with cancer.

(A) iwFAF values (upper graphs) and tumor fractions inferred using ichorCNA (lower graphs) plotted for longitudinal plasma samples from two patients with metastatic melanoma. Green and blue shaded regions indicate courses of treatment. Vertical lines indicate response measured by imaging (RECIST): Purple indicates stable disease and red indicates progressive disease. Standard deviations were calculated for each iwFAF measurement based on the number of sequenced fragments and corresponding observed standard deviation in resampling experiments from control samples. (**B)** Scatterplot comparing change in iwFAF with change in ichorCNA tumor fraction between 63 pairs of samples with measurable tumor fraction obtained from 13 patients with metastatic melanoma (Spearman's $\rho = 0.68$, $P = 1.02 \times 10^{-9}$). (**C)** iwFAF values (upper graphs) and tumor fractions determined using TARDIS (lower graphs) plotted for longitudinal plasma samples from two patients with glioblastoma. Vertical red lines indicate clinical disease progression. (**D)** Scatterplot comparing change in iwFAF with change in TARDIS tumor fraction between 17 pairs of samples with measurable tumor fraction from three patients with glioblastoma (Spearman's $\rho = 0.67$, $P = 3.47 \times 10^{-3}$). Five outliers were excluded from the plot shown to improve visualization, with iwFAF change between timepoints of 5.878×10^{-3}, -1.950×10^{-4}, -4.223×10^{-3}, 6.784×10^{-3}, and 7.348×10^{-3} corresponding to tumor fraction changes of 1.1617×10^{-2}, -1.0003×10^{-2}, -1.288×10^{-3}, 6.3×10^{-5}, and 5.7×10^{-5} (data S7).

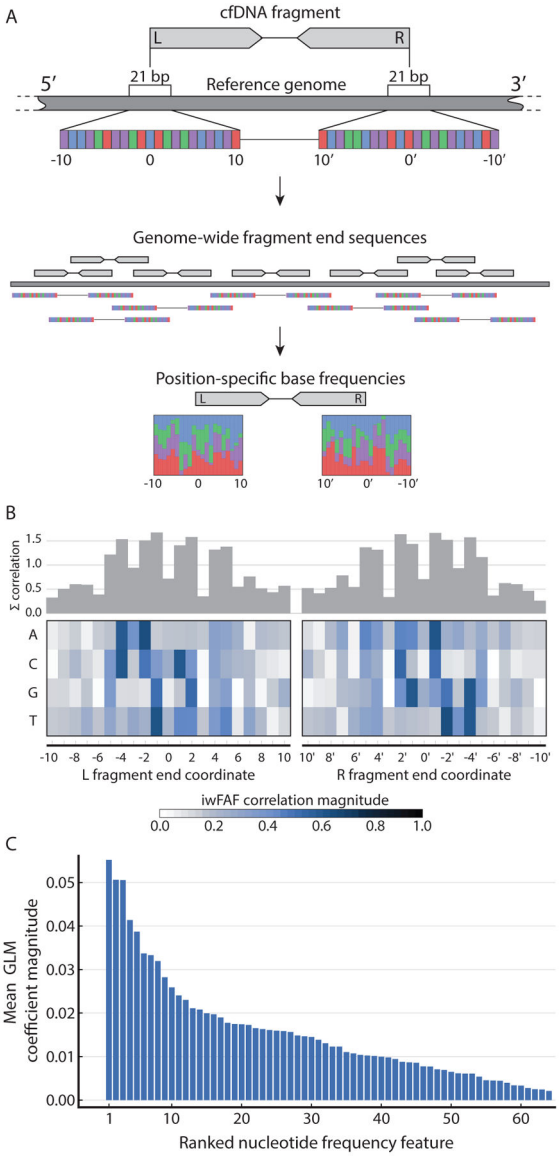

Fig. 3. Analysis of nucleotide frequencies from genomic loci spanning fragment ends.

(A) Schematic of fragment end nucleotide frequency calculations used in GALYFRE. Nucleotide frequencies were measured for positions 10 bp inside (positions 1 to 10) and outside (positions −1 to −10) of each fragment end base (position 0), on the left and right side of each fragment separately. We calculated the frequency of each nucleotide at each position, across all aligned fragments in each plasma DNA sample. (**B)** Heatmap showing the magnitude of correlation between iwFAF and each nucleotide frequency at each position. Frequencies were calculated using 948 samples from 400 patients with metastatic breast cancer. Darker colors indicate a stronger correlation (range of magnitudes of correlation values: 0.003 to 0.66). The sum of correlations at each position is shown in gray above the heatmap. (**C)** Mean adjusted magnitude of regression coefficients obtained from a generalized linear model predicting iwFAF from 64 nucleotide frequencies.

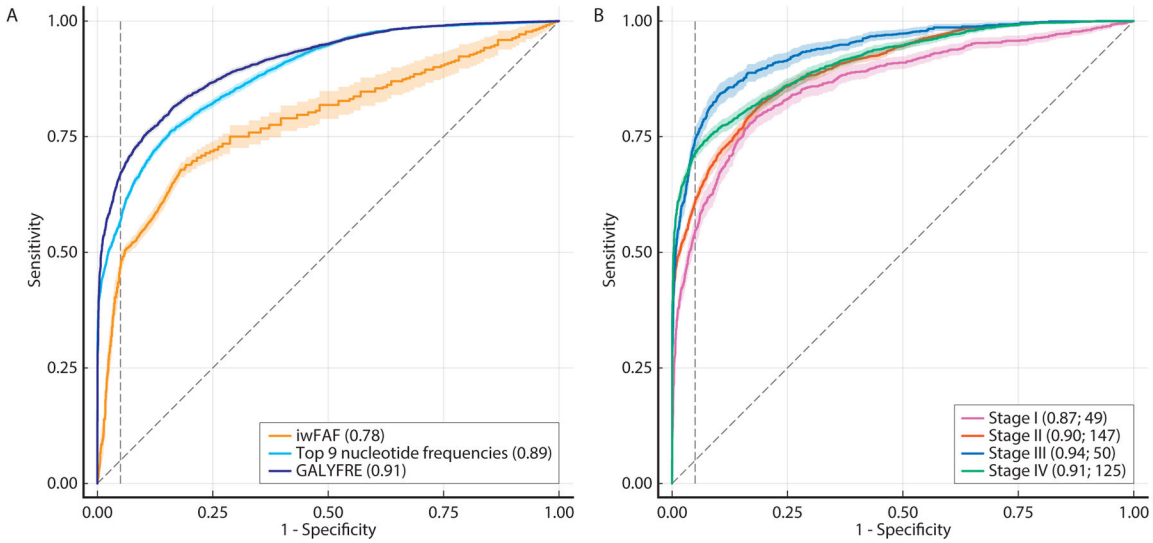

Fig. 4. Diagnostic performance for cancer detection using genome-wide analysis of fragment ends.

Results from a random forest classifier (GALYFRE) trained to distinguish cancer patients from healthy individuals, using iwFAF and nucleotide frequencies at fragment ends in plasma whole genome sequencing data. Training and cross-validation were performed using samples from 196 healthy individuals and 465 patients with cancer, representing 10 cancer types. (**A**) Overall performance from patient samples found in this study, Cristiano *et al.*, and Jiang *et al.* combined based on iwFAF alone, the set of 9 nucleotide frequencies, and the combination of the two (GALYFRE). (**B**) GALYFRE performance by disease stage. Performance by tumor type and by stage within each tumor type is shown in fig. S16 and fig. S17 and sensitivity values at 95% specificity are recorded in table S8.