PNAS RESEARCH ARTICLE | MEDICAL SCIENCES COMPUTER SCIENCES

Fragmentation signatures in cancer patients resemble those of patients with vascular or autoimmune diseases

Samuel D. Curtis[a,b,c,d], Tingshan Liu[e,f], Yuxin Bai[e,f], Yuxuan Wang[b,c,d], Sambit Panda[e,f], Adam Li[g], Haoyin Xu[e,f], Eliza O'Reilly[h], Lisa Dobbyn[b,c,d], Maria Popoli[b,c,d], Janine Ptak[b,c,d,i], Natalie Silliman[b,c,d,h], Chris Thoburn[b], Jeanne Tie[j,k,l], Peter Gibbs[j,l,m], Lan T. Ho-Pham[n,o], Bich N. H. Tran[p], Thach S. Tran[p,q], Tuan V. Nguyen[p,q,r,s,t], Maximilian F. Konig[c,u,v], Michelle Petri[u], Antony Rosen[u], Christopher A. Mecoli[b,u], Ami A. Shah[u], Frits Mulder[w], Nick van Es[w], PLATO-VTE Study Group[1], Chetan Bettegowda[b,c,d,x], Kenneth W. Kinzler[b,c,d], Nickolas Papadopoulos[b,c,d], Joshua T. Vogelstein[e,f], Bert Vogelstein[b,c,d,i,2], and Christopher Douville[b,c,d,y,z,2]

Affiliations are included on p. 11.

Contributed by Bert Vogelstein; received December 26, 2024; accepted July 9, 2025; reviewed by Dan Landau and Viktor Adalsteinsson

Multiple case-controlled studies have shown that analyzing fragmentation patterns in plasma cell-free DNA (cfDNA) can distinguish individuals with cancer from healthy controls. However, there have been few studies that investigate various types of cfDNA fragmentomics patterns in individuals with other diseases. We therefore developed a comprehensive statistic, called fragmentation signatures, that integrates the distributions of fragment positioning, fragment length, and fragment end-motifs in cfDNA. We found that individuals with venous thromboembolism, systemic lupus erythematosus, dermatomyositis, or scleroderma have cfDNA fragmentation signatures that closely resemble those found in individuals with advanced cancers. Furthermore, these signatures were highly correlated with increases in inflammatory markers in the blood. We demonstrate that these similarities in fragmentation signatures lead to high rates of false positives in individuals with autoimmune or vascular disease when evaluated using conventional binary classification approaches for multicancer earlier detection (MCED). To address this issue, we introduced a multiclass approach for MCED that integrates fragmentation signatures with protein biomarkers and achieves improved specificity in individuals with autoimmune or vascular disease while maintaining high sensitivity. Though these data put substantial limitations on the specificity of fragmentomics-based tests for cancer diagnostics, they also offer ways to improve the interpretability of such tests. Moreover, we expect these results will lead to a better understanding of the process—most likely inflammatory—from which abnormal fragmentation signatures are derived.

cell-free DNA | cancer screening | fragmentomics | rheumatology | autoimmune diseases

Significance

The knowledge that abnormal cell-free DNA (cfDNA) fragmentation patterns are found in patients without cancer will guide the development of more effective molecular diagnostics and encourage more research into the use of cfDNA in individuals with cancer as well as in those with autoimmune or vascular diseases. The decomposition of cfDNA fragmentation patterns into Fragmentation Signatures reveals a fundamental relationship among abnormal fragmentation signatures, plasma cfDNA concentration, and an increase in inflammatory plasma proteins. The purposeful incorporation of patients with nonmalignant diseases during test development represents a major advance for improved multicancer earlier detection (MCED). From a basic science perspective, our findings indicate that a shared inflammatory process is likely responsible for the abnormal fragmentation patterns in cancer and other diseases.

The use of cell-free DNA (cfDNA) to assist with the diagnosis of cancer has a long and venerable history. More than 40 y ago, researchers showed that the concentration of cfDNA in the blood of cancer patients was higher than in healthy controls (1). Later studies showed that this increase was not specific for cancer, and that elevations in blood cfDNA concentrations occur in other states, including exercise, trauma, cardiovascular disease, sepsis, aseptic inflammation, autoimmune disease, and viral infections (2–5). However, in the early 90's, researchers demonstrated that mutations present in cancer cells can provide highly specific biomarkers for cancer (6, 7). These studies stimulated efforts to use genetic alterations such as mutations and aneuploidy, as well as epigenetic alterations such as DNA methylation, in a variety of clinical samples, including plasma, serum, sputum, Pap Smears, cerebrospinal fluid, saliva, and urine (8–17). Such analyses are referred to as "liquid biopsies," emphasizing their noninvasive nature compared to conventional biopsies. There are now thousands of examples of the productive use of such liquid biopsies to assist in the diagnosis of patients with cancer or suspected cancer, and several of these tests are offered commercially, some with Food and Drug Administration approval.

More recently, the analysis of fragmentation patterns in cfDNA, called fragmentomics, has emerged as a promising approach for the evaluation of liquid biopsies (18, 19). In healthy individuals, cfDNA fragmentation patterns have a characteristic pattern that includes a length distribution consistent with the wrapping of DNA around nucleosomes, an end-motif pattern indicative of digestion from specific nucleases, and a genomic positioning pattern that represents a footprint of the chromatin proteins bound to nuclear DNA within cells (20, 21). Individuals with cancer frequently have alterations to their cfDNA fragmentation patterns, including alterations to fragment positioning, fragment length, fragment end-motifs patterns, and repetitive elements (22–28). Multiple case-controlled, retrospective studies have demonstrated the potential of fragmentomics

[1]A complete list of the PLATO-VTE Study Group can be found in the *SI Appendix*.

[2]To whom correspondence may be addressed. Email: vogelbe@jhmi.edu or cdouvil1@jhmi.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2426890122/-/DCSupplemental.

Published August 20, 2025.

for multicancer earlier detection (MCED) and minimal residual disease monitoring.

However, the mechanisms responsible for fragmentomic patterns in the cfDNA from cancer patients, and whether these occur exclusively in individuals with cancer, are still unclear. We here report observations, initiated serendipitously, that demonstrate these patterns are not specific to cancer patients and can arise in the absence of any neoplastic cells.

Results

Background. This study was initiated with an analysis of plasma cfDNA from patients with unprovoked venous thromboembolism (VTE). In the absence of known risk factors, such as postsurgery, a substantial fraction of such patients are found to have cancer (29, 30). We hoped to use cfDNA to reveal which patients with an unprovoked VTE are likely to have cancer and thereby could benefit from further imaging studies and earlier detection of a previously undiagnosed malignancy. We used a targeted mutation panel to search for mutations in the cfDNA, and whole genome sequencing to search for aneuploidy. This study is still ongoing, but during its course, we evaluated the proportion of an end-motif signature (called MendSeqS) that we had been investigating for potential incorporation into a screening test for cancer (24). To our surprise, we found that a high proportion of VTE patients had a MendSeqS pattern that was indistinguishable from that in cancer patients, even though only a small proportion of the VTE group had cancer. This stimulated us to evaluate other fragmentation patterns in VTE patients, as well as to evaluate whether patients with other illnesses might have similar fragmentation patterns.

cfDNA Fragmentation Patterns in Healthy Individuals and Patients with Cancer, Autoimmune, or Vascular Diseases. To evaluate fragmentation patterns, we performed shallow whole-genome sequencing (~1×) on the cfDNA from 941 plasma samples from 882 individuals as follows:

Group IA – controls with no known history of cancer, used for normalization (n = 130)

Group IB – controls with no known history of cancer, used for validation (n = 255)

Group IIA – cancer <30 d prior to surgical excision with a high plasma tumor fraction (>10%, n = 69)

Group IIB – cancer <30 d prior to surgical excision with undetectable plasma tumor fraction, (n = 142)

Group III – unprovoked VTE <30 d following diagnosis who did not develop cancer within 2 y of the thromboembolic event (VTE; n = 75)

Group IV – current diagnosis of systemic lupus erythematosus (SLE; n = 21)

Group V – current diagnosis of dermatomyositis (DM; n = 143).

Group VI – current diagnosis of systemic sclerosis (SSc; n = 106)

Previous studies have demonstrated that alterations to cfDNA fragmentation patterns are highly correlated with the fraction of molecules in the plasma that are tumor-derived (25, 31). Upon the finding that nonmalignant conditions such as VTE shared similar alterations, we aimed to better evaluate the effect of circulating tumor DNA (ctDNA) on fragmentation patterns. Using a measurement of aneuploidy [ichorCNA (32)] we selected two nonoverlapping groups of patients with cancer: (IIA) those with a high plasma tumor fraction (>10%) and (IIB) those with undetectable plasma tumor fraction (i.e., tumor fraction < 3%). Aneuploidy is widely appreciated to be exquisitely specific for cancer (33). We also used ichorCNA to confirm that all individuals without a diagnosis of cancer, including those with vascular or autoimmune diseases, had no detectable aneuploidy in their plasma (Dataset S1). For statistical rigor, the controls were also divided into two nonoverlapping groups: Group IA, a "Normalization Control Group" (n = 130) and Group IB, an independent "Validation Control Group" (n = 255). Group IA was used to convert all variables into z-scores. Group IB was used to compare controls with patients harboring various diseases in an unbiased manner (*Methods*).

Fragment End-Positions. The ends of cfDNA fragments are not randomly distributed throughout the genome. In other words, the fragment end-positions are different from what would be expected if nuclear DNA is mechanically sheared or digested by nonspecific nucleases. Rather, cfDNA fragment end positions appear to reflect the chromatin state of the cells from which the DNA originated (34–37). Numerous recurrently protected regions (RPRs) are found in the cfDNA in healthy individuals, resulting in a relative decrease of fragments whose ends are located within these regions (25). These regions are thought to be protected by chromatin proteins in the normal leukocytes of these patients, as leukocytes contribute the vast majority of cfDNA to the plasma of healthy individuals as well as those with cancer (38).

Budhraja et al. generated a map of RPRs using high depth sequencing of cfDNA from healthy individuals and demonstrated that individuals with cancer have a significant increase in the frequency of fragments whose end-positions map within the RPRs (25). We analyzed the frequency of fragment end-positions within RPRs in our WGS data (*Methods*) and confirmed Budhraja's results. Individuals with cancer from a variety of tumor types had major increases in the representation of these fragment ends while control samples showed no such increase (Fig. 1*A*). We also observed increases in fragment-ends within RPRs in individuals with VTE, SLE, DM, and scleroderma (Fig. 1*A*). The fragment end-positions were remarkably similar in the plasma of patients with cancer or the other diseases, with a peak at the center of the RPR and valleys at ±100 bp from the center (Fig. 1*A*).

To generate a univariate biomarker derived from fragment end-positions, Budhraja et al. described a metric called information-weighted Fraction of Aberrant Fragments (iwFAF). This statistic reflects the fraction of fragments that have fragment ends within RPRs, weighted by the length and GC content of the fragments. We confirmed that this metric separated cancer patients from controls in our cohort (Fig. 1*B*). Similar to the findings from Budhraja et al., we observed that the iwFAF was increased in cancer patients with high plasma tumor fraction (Group IIA) compared to those with low plasma tumor fraction (Group IIB; $P < 0.001$) (Fig. 1*B*). The difference between iwFAF was observed in cancers with high plasma tumor fractions, regardless of cancer type (*SI Appendix*, Fig. S1). Furthermore, we also observed increases in iwFAF in individuals with VTE ($P < 0.001$), SLE ($P < 0.001$), DM ($P < 0.001$), and to a lesser extent, scleroderma ($P = 0.0017$) (Fig. 1*B*).

Fragment Length Patterns. The first identified fragmentomic biomarker for cancer was observed over 25 y ago as a change in the distribution of fragment lengths of cfDNA (39). Since then, the fragment length of cfDNA molecules has often been used as a biomarker for cancer screening and monitoring (23, 40–43). Individuals with cancer have been shown to have at least three types of alterations in their fragment length patterns: i) an increase in the proportion of short, subnucleosomal fragments (<160 bp), ii) a 10 bp periodicity in the fragmentation patterns of these short fragments, and iii) a decrease in the proportion of nucleosomal

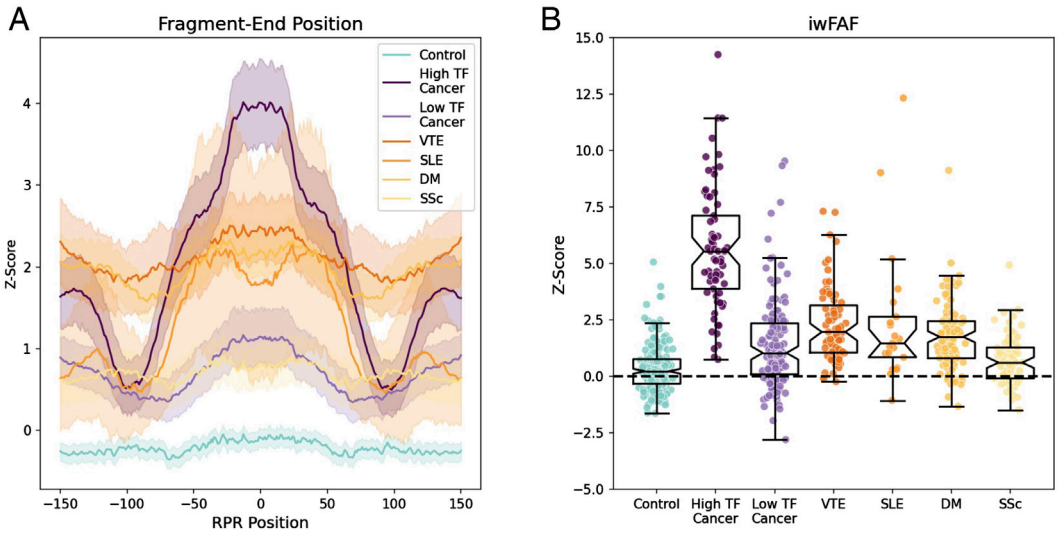

Fig. 1. Fragment end-positioning in RPRs. (*A*) Fragment end-position z-scores. Solid lines represent the mean z-score for each group while shaded regions represent the 95% CI. The z-scores were calculated using the distribution of fragment-ends in the Normalization Control Group (Group IA) and the blue represents the distribution of z-scores in the Validation Control Group (Group IB). "TF" = plasma tumor fraction. (*B*) Box plots of the information-weighted Fraction of Aberrant Fragments (iwFAF) value. The dotted black line represents the mean values of the Validation Control Group (Group IB). Asterisks indicate statistical significance differences compared to Group IB ($P < 0.001$). Horizontal lines on boxplots represent 1st, 2nd, and 3rd quartiles. Notches on boxplots represent 95% CI of the median (median ± 1.57 × IQR/$n^{0.5}$). Whiskers represent 1.5*IQR. Notches One low tumor fraction cancer outlier in *B* (z-score>20) was removed to enhance visibility. Fragment end-positioning variables and iwFAF for all patients are provided in Dataset S2.

fragments (160 to 180 bp) compared to healthy controls. In Fig. 2, we plot the distribution of fragments with lengths between 80 bp and 225 bp, representing the great majority (>90%) of the total cfDNA fragments in plasma. Intriguingly, we found that in individuals with VTE, DM, or scleroderma, the fragmentation patterns were similar to those of cancer patients (Fig. 2*A*). Specifically, there was an increase in the proportion of subnucleosomal fragments, a clear 10- bp periodicity among these short fragments, and an increase in the proportion of nucleosomal fragments (Fig. 2*A*; also see *SI Appendix*, Fig. S2, which describes the entire distribution of fragments, including those larger than 225 bp).

As with RPRs, it is useful to be able to define a single variable that captures the essence of the length patterns. One such variable, described by Christiano et al. (23), is the fragment length ratio (FLR), defined as the ratio of short (100 to 150 bp) to long (151 to 220 bp) fragments. We found that the plasma of cancer patients with a high plasma tumor fraction (Group IIA) had a significant increase in the FLR, in agreement with Christiano et al. (23) ($P < 0.00001$; Fig. 2 *B* and *D*). However, cancer patients with a low plasma tumor fraction (Group IIB) did not have a significant increase in FLR ($P = 0.22$, Fig. 2 *B* and *D* and *SI Appendix*, Fig. S1). The FLR was considerably higher in patients VTE, SLE, DM, and scleroderma than it was in cancer patients with low plasma tumor fractions (Group IIB, $P < 0.001$, Fig. 2 *B* and *D*).

Fragment End-Motif Patterns. One frequently used fragmentation-based biomarker used to distinguish patients with cancer from healthy individuals is the nucleotide motif at the 5′ end of plasma DNA fragments (24, 44–46). We analyzed the four bases (tetramer) at the 5′ end of each cfDNA molecule for each of the 256 possible tetramers and then converted end-motif frequencies into z-scores based on the Normalization Control Group (Group IA; *Methods*). As expected, we found major differences in end-motif frequencies between the Validation Control Group (Group IB of heat map in Fig. 3*A*) vs. cancer patients with high tumor fraction in their plasma (Group IIA in Fig. 3*A*). The same tetramers were correspondingly elevated or depressed in frequency in the cancer patients with low tumor fraction in their plasma (Group IIB in Fig. 3*A*), but to a lesser extent than in the cancer patients with high tumor fraction in their plasma. Strikingly, the same tetramers that were elevated in frequency in the cancer patients were often also elevated in patients with VTE, SLE, SSc, and DM (Fig. 3 *A* and *B*). The same was true for tetramers whose frequencies were depleted in cancer patients compared to controls (blue bars in Fig. 3*A*). To quantify the effects shown in the heat map, we correlated the mean end-motif z-scores of the 256 tetramers in the various groups of patients. We found that the individual tetramer frequencies in patients with VTE ($R = 0.88$, $P < 0.00001$), SLE ($R = 0.475$, $P < 0.00001$), DM ($R = 0.76$, $P < 0.00001$), and SSc ($R = 0.72$, $P < 0.00001$) were highly correlated with those of patients with cancers having high tumor fractions (Fig. 3*B*). Fragment end-motif z-scores were uniformly greater in patients with VTE, SLE, and DM patients than they were in cancer patients with low tumor plasma fractions ($P < 0.00001$, Fig. 3*B*).

Jiang et al. described a single variable that captures the major characteristics of the motif patterns, called Motif Diversity Score (MDS) (46). We found, as expected, that the plasma of cancer patients with high tumor plasma fractions (>10%) had a significant increase in the MDS compared to the Validation Control Group ($P < 0.001$; Fig. 3*C*). Cancer patients with a lower plasma tumor fraction did not have a significant increase in MDS compared to the controls (Fig. 3*C*). In contrast, individuals with VTE ($P < 0.001$), SLE ($P < 0.001$), DM ($P < 0.001$), and scleroderma ($P < 0.001$) had a significant increase in the MDS compared to the control group (Fig. 3*C*).

Fragmentation Signatures. We next sought to further explore the nature of the fragmentation patterns in these various disease states in an unbiased manner. For this purpose, we evaluated all three of the fragmentation patterns described above (fragment end-positions, fragment lengths, and fragment end-motifs) using Principal Component Analysis (PCA). PCA is a dimension reduction technique, like that of Non-Matrix Factorization (NMF), which has in the past been used to define mutational signatures rather than fragmentation signatures (47). PCA provides the optimal linear reconstruction of the data in terms of minimizing the squared error and is guaranteed to find the global minimum, whereas NMF is not. While supervised approaches rely on finding disease-specific signatures, unsupervised approaches such as PCA operate without

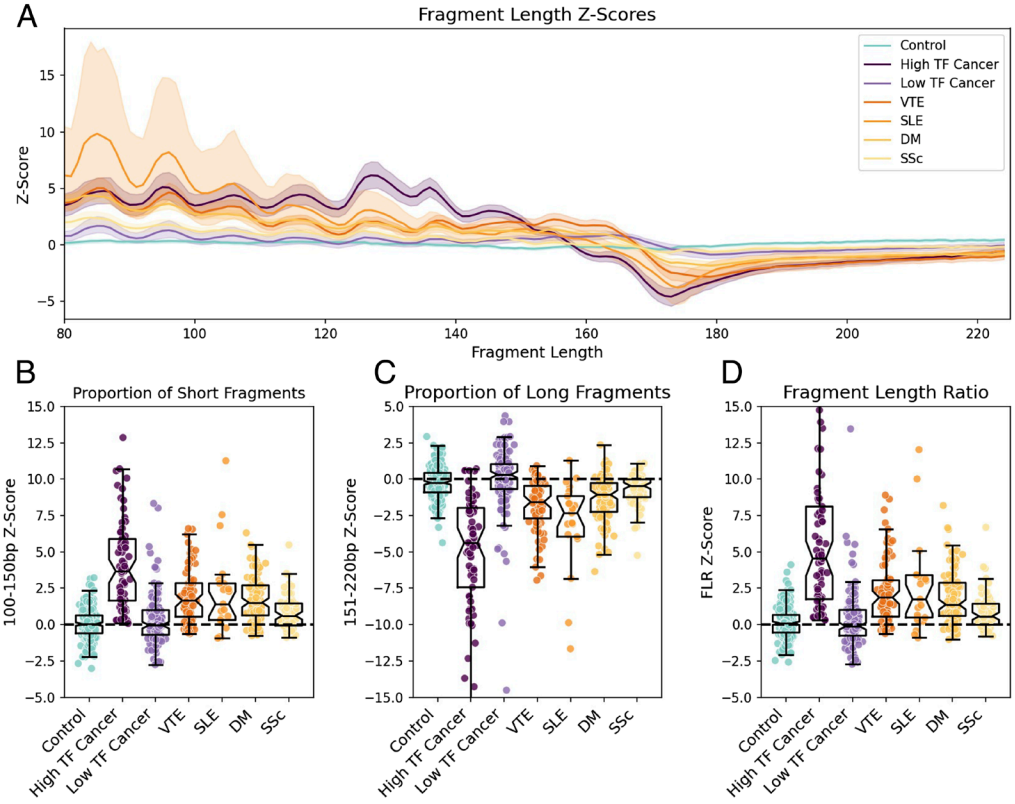

Fig. 2. Fragment length distributions (*A*) Fragment length z-scores. Solid lines represent the mean z-score for each disease group. The z-scores were calculated on the basis of the distribution of lengths in the control group used for normalization (Group IA), and the blue represents the distribution of lengths of the Validation Control Group (Group IB), with the mean of Group IB depicted by the horizontal dashed line. TF = plasma tumor fraction. (*B–D*) Box plots of the proportion of short fragments (100 to 150 bp) (*B*), the proportion of long fragments (151 to 220) (*C*), and the ratio of short fragments over long fragments (*D*). The dotted black line represents the mean values of the Validation Control Group (Group IB). Horizontal lines on boxplots represent 1st, 2nd, and 3rd quartiles. Notches on boxplots represent 95% CI of the median (median ± 1.57 × IQR/$n^{0.5}$). Whiskers represent 1.5*IQR. In (*B–D*) 11, 8, and 14 outlying points, respectively, in the patients in the high plasma tumor fraction group of cancer patients (Group IIA) are not shown, so as to more conservatively visualize the differences between patients with disease and controls. Fragment length ratios for all patients are provided in Dataset S3.

any knowledge of disease group labels, allowing the algorithm to identify variance-driving components based solely on the intrinsic properties and relationships of the variables. For each fragmentation pattern, we analyzed the minimum number of principal components that were required to capture more than 90% of the explained variance—resulting in two, two, and eight principal components for end-positions, lengths, and end-motifs, respectively (Fig. 4). These 12 principal components comprise what we call *fragmentation signatures*.

To better understand the representation of components within fragmentation signatures we analyzed the singular values, eigenvalues, and eigenvectors (loadings) for each of the 12 principal components. Our analysis demonstrated that the 1st component (highest explained variance) of each fragmentation pattern was highly correlated with the fragmentation patterns (z-scores) found in cancer patients with high plasma tumor fractions ($P < 0.0001$, *SI Appendix*, Fig. S3 and Dataset S6). Similarly, we found that the 1st principal component for each fragmentation pattern was significantly elevated in cancer patients with high plasma tumor fractions as well as those with autoimmune or vascular disease ($P < 0.0001$). Interestingly, we found that some of the components with less explained variance appeared to represent information that was specific to certain disease groups. For example, fragment end-position PC2 (FP PC2) was related to an increase in fragment-ends within the center of the RPR and was significantly elevated only in cancer patients with high plasma tumor fractions ($P < 0.0001$). Furthermore, fragment end-motif PC3 (FM PC3) was related to an increase in T-motifs and was significantly decreased only in individuals with autoimmune or vascular conditions ($P < 0.0001$).

To determine how these fragmentation signatures were distributed within our cohort, we performed clustering using AutoGMM on these twelve components. AutoGMM is an automated Gaussian Mixture Model (GMM) framework that streamlines clustering by automatically optimizing critical hyperparameters (48). We applied AutoGMM to our fragmentation signatures and identified two major clusters—Cluster 1 and Cluster 2 (Fig. 4*A*). We found that the principal components with the highest explained variance were most informative for distinguishing between the two clusters, whereas those with low variance were relatively similar between clusters (Fig. 4*B* and Dataset S5). The clustering analysis revealed a striking pattern: Cluster 1 contained nearly all control samples (97%), while Cluster 2 contained most cancer patients with high plasma tumor fraction (75%) (Fig. 4*C* and Dataset S5). Interestingly, Cluster 2 contained only a small portion (20%) of cancer patients with low plasma tumor fractions but a majority (146/280) of patients with other nonmalignant conditions—64% of VTE cases, 63% of DM cases, 38% of SLE cases, and 25% of SSc cases (Fig. 4*C*).

As a "sanity check," we determined whether the patients in Cluster 1 had different individual fragmentomics patterns, rather than fragmentation signatures, than those in Cluster 2. Indeed, the fragment end-positions, fragment lengths, and fragment end-motifs of samples in Cluster 1 were all different from those in Cluster 2 (Fig. 4 *D–F*). To ensure that these differences were not driven by specific disease groups we evaluated the same fragmentation patterns within each disease group and found similar differences in fragmentation patterns between clusters (*SI Appendix*, Fig. S5). Similarly, we investigated whether Clusters 1 and 2 separated patients with or without various diseases when using aggregate (rather than individual) metrics of fragmentation patterns. As noted earlier in this paper, iwFAF, FLR, and MDS are aggregate metrics of RPR fragment end-positions, fragment lengths, and fragment end-motifs, respectively. Across all diseases, patients in Cluster 2 had higher iwFAF, higher FLR, and higher MDS scores than patients in Cluster 1 ($P < 0.001$; *SI Appendix*, Fig. S6).

Relationship between Fragmentation Signatures, cfDNA Concentrations, and Circulating Inflammatory Markers. With the knowledge that fragmentation signatures are often increased in patients with autoimmune diseases, we determined whether plasma markers for inflammation were correlated with these signatures. We were able to analyze the concentration of 17 plasma

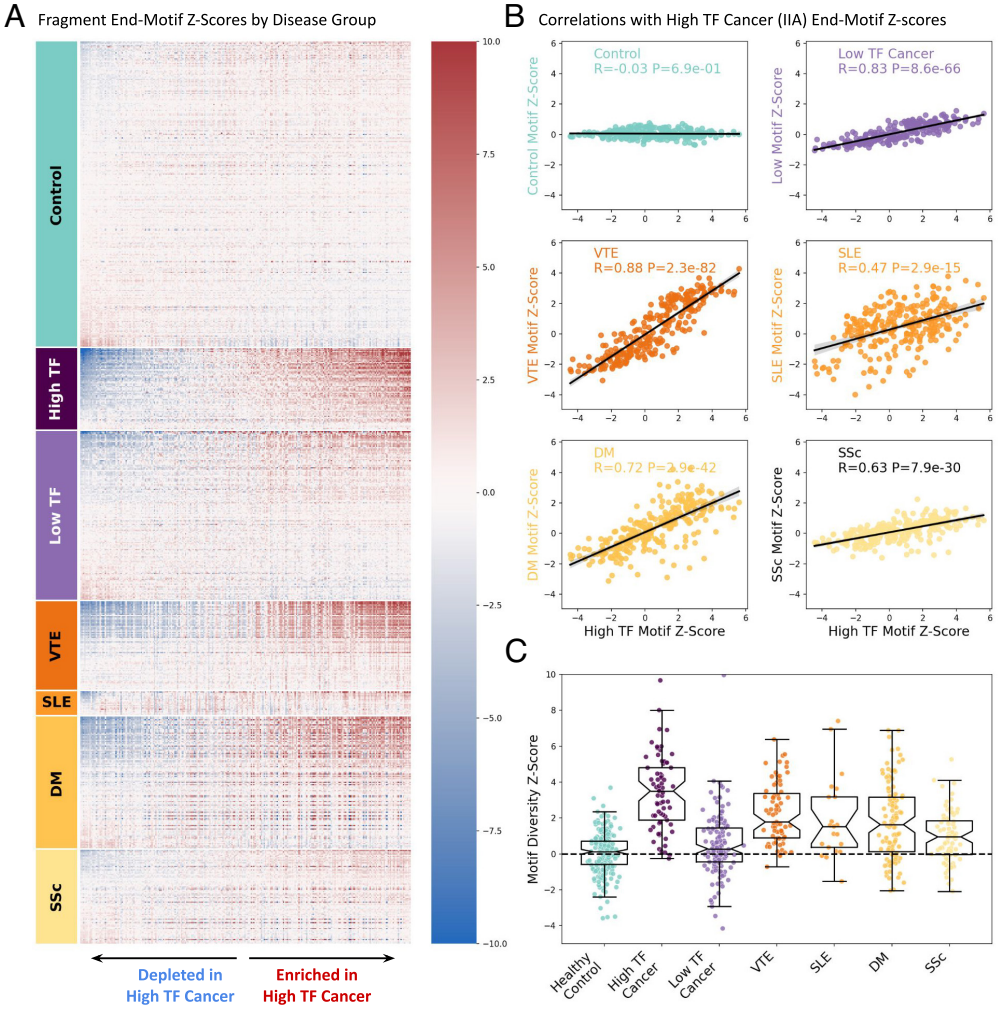

Fig. 3. Fragment-end motifs. (*A*) Fragment-end motif z-scores for each sample separated by sample group. Motifs are sorted by the mean z-score in High TF Cancers. Samples are grouped on the *y*-axis according to the disease group. Order of samples within the disease group (*y*-axis) is sorted based on the z-score of the CCCT end-motif. (*B*) Correlation between fragment-end motif z-scores in individuals with cancers having high plasma tumor fraction (*x*-axis) and the other groups. Each data point represents the mean z-score among samples for a single tetramer end-motif, so there are 256 data points for each group. (*C*) Motif Diversity Score (MDS) for each sample stratified by sample type. Horizontal lines on boxplots represent 1st, 2nd, and 3rd quartiles. Notches on boxplots represent 95% CI of the median (median ± 1.57 × IQR/$n^{0.5}$). Whiskers represent 1.5*IQR. Fragment end-motif variables and MDS for all patients are provided in Dataset S4.

proteins in 838 of the 882 (95%) patients included in this study (Dataset S7). These proteins were originally chosen because the literature indicated that they were often elevated in patients with cancer (8). Seven of the 17 proteins were very highly correlated with an abnormal fragmentation signature: GDF-15, OPG, IL-8, myeloperoxidase(MPO), HGF, OPN, TIMP-1, and NSE ($P < 0.0001$, Fig. 5 *A* and *B*). Importantly, all of these seven were classic markers for inflammation (49), and none were specific biomarkers for cancer, such as CEA, CA19-9, AFP, or CA125. The effect sizes of these correlations were large, with *P*-values $< 10^{-5}$ after correcting for false discovery using the Benjamini–Hochberg heuristic. To control for the possibility that these correlations are driven by individual disease groups, we stratified the samples by disease group and found that the relationship between fragmentation signatures and inflammatory markers remained consistent across disease groups (*SI Appendix*, Fig. S7).

The concentration of cfDNA has been demonstrated to be an activator of innate immunity and is often higher in cancer patients and individuals with autoimmune or vascular disease compared to healthy controls (1, 38, 50). In light of these observations we evaluated whether there was a correlation between fragmentation signatures and cfDNA concentration. Indeed, our clustering of fragmentation signatures revealed a significant difference in cfDNA concentrations between samples in cluster 1 and cluster 2 (Fig. 5*C*). When we analyzed each patient group separately, this relationship was also significant ($P < 0.005$) in several patient groups. To further characterize these relationships, we examined the relationship between cfDNA concentration and three fragmentation metrics (iwFAF, FLR, and MDS). Each metric showed significant positive correlations ($P < 0.05$) with plasma cfDNA levels in many patient groups (*SI Appendix*, Fig. S8).

Supervised Learning to Distinguish Malignant from Nonmalignant Disease Using Fragmentation Signatures and Plasma Proteins. We next aimed to determine whether a supervised learning method could more effectively distinguish autoimmune or vascular disease from cancer. We utilized a newly described AI (a.k.a. machine learning) algorithm called MIGHT (Curtis et al., in press at PNAS) that has major advantages over other commonly used classification approaches. Using a bootstrapping methodology, MIGHT incorporates canonical cross-validation into the learning process to train, calibrate, and estimate posteriors without the need for external datasets. This methodology has been shown to provide more reliable and accurate estimates of sensitivity and specificity compared to other state of the art algorithms (Curtis et al., in press at PNAS). Estimates generated by MIGHT are universally consistent, meaning that asymptotically they achieve the optimal result, regardless of the underlying distributions of the variables, and do not rely on the assumption that the variables are distributed in any specific fashion, such as linearly, among the cases (e.g., cancer) and controls (e.g., healthy individuals).

We first applied MIGHT to estimate the sensitivity and specificity of fragmentation signatures in a classical MCED setting, wherein a model is trained to classify individuals as either cancer or noncancer (e.g., binary classification). We trained a MIGHT model using the Fragmentation Signatures derived from a cohort of healthy

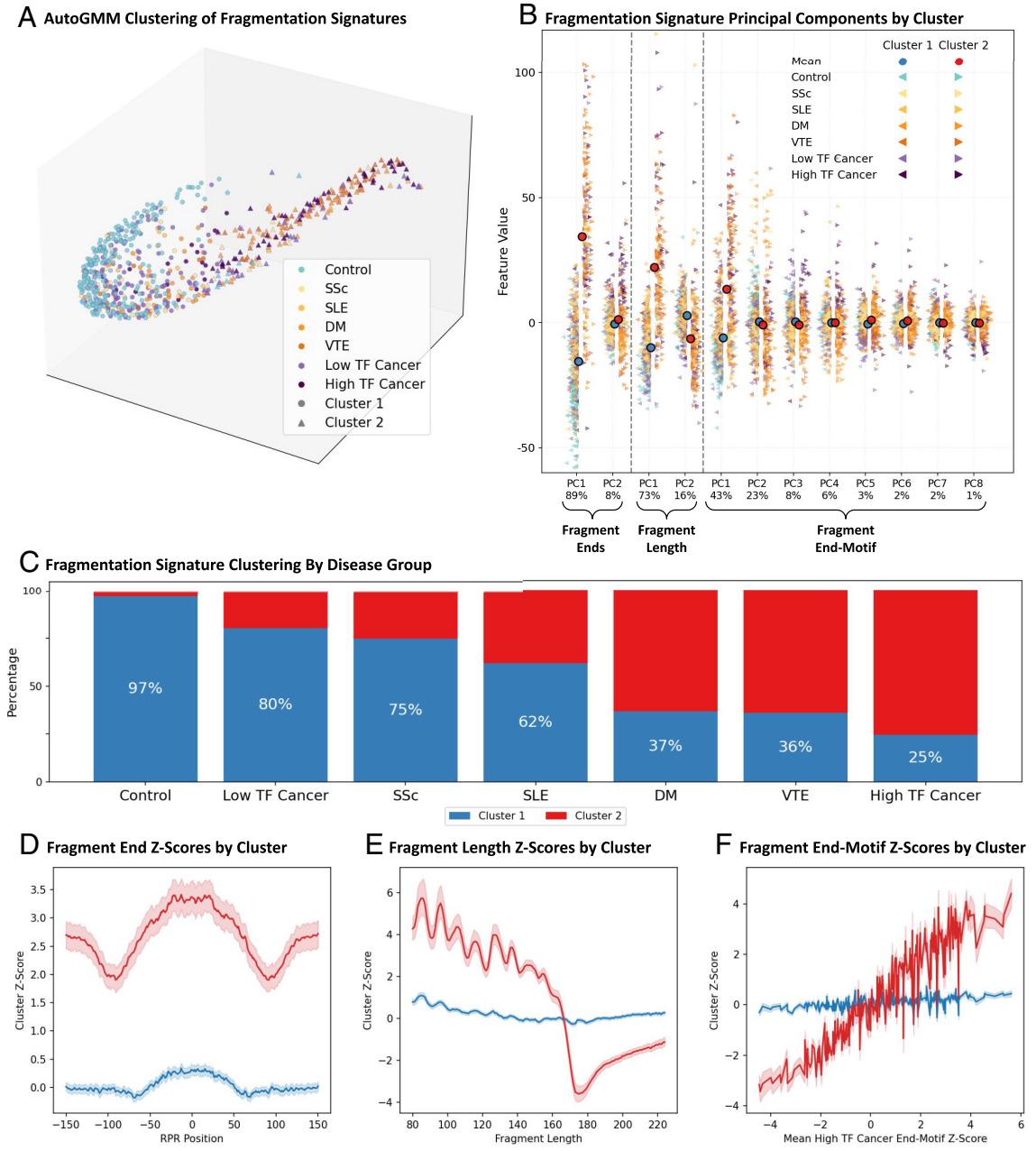

Fig. 4. Automated Gaussian mixture model clustering of cfDNA fragmentation signatures. (*A*) unsupervised AutoGMM clustering based on 12 principal components derived from variations in fragmentation patterns, including fragment end-positions, fragment lengths, and fragment end-motifs. (*B*) Strip plots illustrating the distribution of each principal component between clusters 1 and 2. Blue and red circles indicate the mean value for clusters 1 and 2, respectively. (*C*) Proportion of samples in each disease group assigned to Cluster 1 (blue) and 3 (red). (*D–F*) Mean fragment end-positions, fragment lengths, and fragment end-motif z-scores for samples in Cluster 1 (blue) and 2 (red). (*F*) Z-Scores are sorted based on the mean z-score in high tumor fraction cancers (IIA). Principal components and cluster assignment for all patients are provided in Dataset S5.

controls (n = 255) along with low and high tumor fraction cancers from the pancreas, lung, colon, breast, liver, esophagus, and stomach (n = 193). We subsequently applied the PCA and classification models, including predetermined thresholds for 98% specificity, to our training data and a nonoverlapping cohort of individuals diagnosed with autoimmune or vascular disease (n = 286) (*SI Appendix*, Fig. S9*A*). We observed high sensitivity at a predefined specificity of 98% (S@98) in our training data with sensitivities of 67% and 26% for high and low tumor fraction cancers, respectively. However, we observed high rates of false positives in individuals with VTE (48%), SLE (57%), DM (50%), and SSc (14%) (*SI Appendix*, Fig. S9*B*). In a MIGHT model integrating both fragmentation signatures and our panel of 17 plasma proteins, we observed increased sensitivity for high and low tumor fraction cancers but similarly high rates of false positives in individuals with autoimmune or vascular disease (*SI Appendix*, Fig. S9*C*).

To improve the specificity of fragmentation signatures, we hypothesized that including patients with autoimmune or vascular disease in the training set could allow MIGHT to learn specific variables that could differentiate these patients from those with cancer. To test this hypothesis, we trained a MIGHT model on the fragmentation signatures derived from the same cohort of low and high tumor fraction cancers alongside a "noncancer" cohort including healthy controls as well as those with autoimmune or vascular diseases (*SI Appendix*, Fig. S10*A*). In this scenario, the S@98 was 53% and 23% for high and low tumor fraction cancers, respectively (*SI Appendix*, Fig. S10*B*). Adding plasma proteins to this model increased S@98 to 66% and 53% for high and low

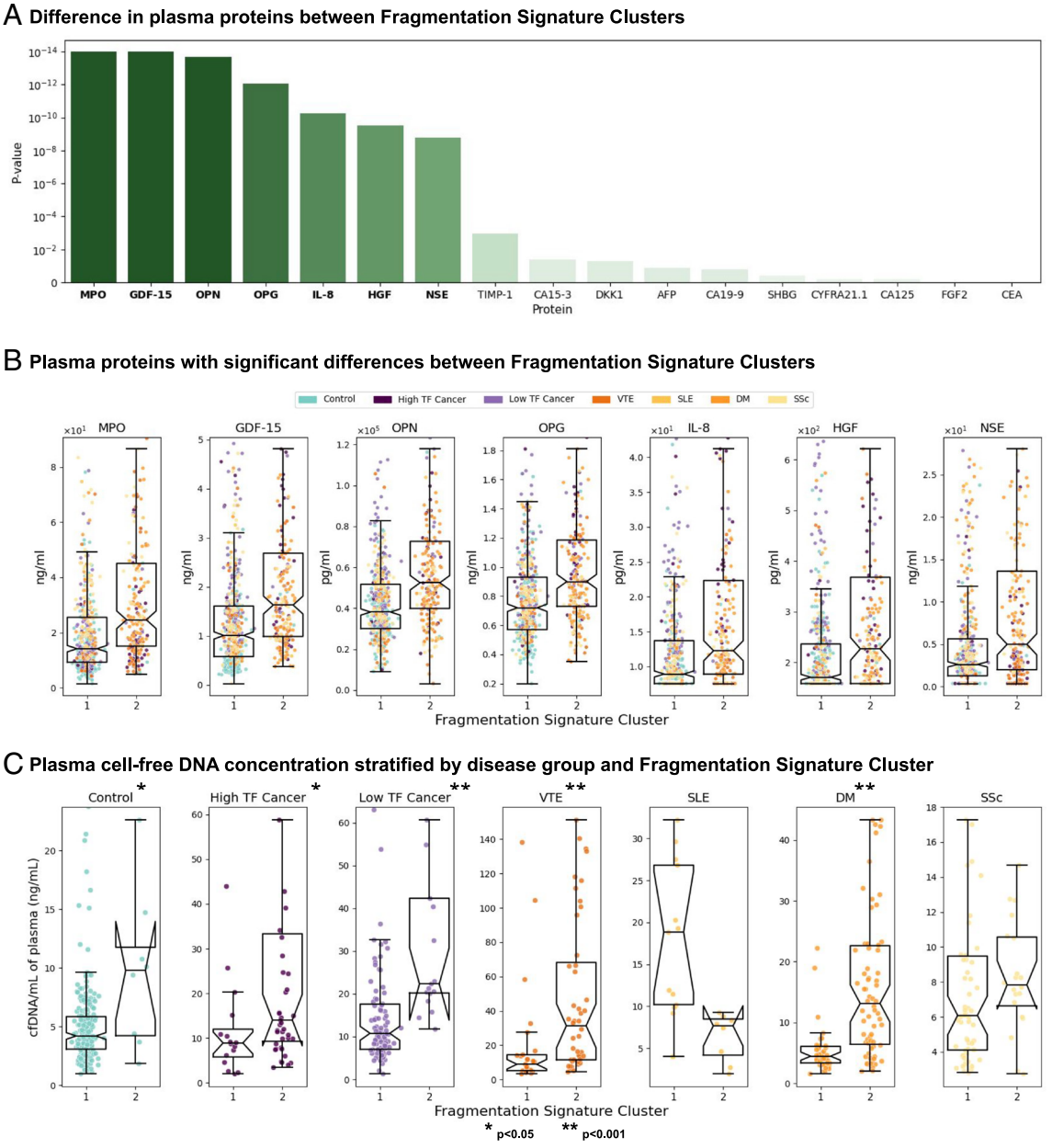

Fig. 5. Comparison of inflammatory biomarkers between Fragmentation Signature Clusters. (*A*) One-sided Mann–Whitney *P*-values comparing plasma proteins between individuals in Clusters 1 and 2. *P*-values were corrected using Benjamini–Hochberg FDR adjustment. (*B*) Box and strip plots illustrating the distribution of plasma proteins with significant differences between Fragmentation Signature Clusters 1 and 2 ($P < 0.0001$). (*C*) Comparison of plasma cfDNA concentration between Fragmentation Signature Clusters 1 and 2 stratified by disease group. Outlier points > IQR are not shown in these graphs so as to more conservatively visualize the differences, but all values are provided in Dataset S7. *P*-values for comparisons between Fragmentation Signature Clusters were performed using one-sided Mann–Whitney *U* tests. * indicates $P < 0.05$. ** indicates $P < 0.001$. Protein concentrations for all patients are provided in Dataset S7.

tumor fractions, respectively (*SI Appendix*, Fig. S10*C*). In sum, adding patients with autoimmune or vascular diseases to the training set did indeed decrease the fraction of patients with autoimmune or vascular diseases who were falsely classified as having cancer (thereby increasing specificity), but also decreased the fraction of patients with cancer who were truthfully classified as having cancer (thereby decreasing sensitivity); compare *SI Appendix*, Fig. S10*C* with *SI Appendix*, Fig. S9*C*).

Our analysis of fragmentation signatures and circulating proteins revealed significant similarity between individuals with cancer and those with autoimmune and vascular disease. However, we also identified specific fragmentation signature components that may be able to stratify patients with autoimmune or vascular disease from healthy controls and patients with cancer. We therefore generalized MIGHT to enable it to work on multiclass data, rather than merely two-class. We then proved that this multiclass MIGHT is also a universally consistent estimator of S@98 (*SI Appendix*, *Methods*, Theorem 1). We then could evaluate whether a multiclass model trained to distinguish three classes—healthy controls, patients with autoimmune or vascular diseases, and patients with cancer—could learn subtle differences in the fragmentation signatures that the binary approach was unable to detect. For clarity, the samples chosen for this multiclass model development (Fig. 6*A*) were identical to those used to develop the two-class model (*SI Appendix*, Fig. S10*A*). Using this approach, we observed that patients with cancer were classified as well as they were with the conventional approach (~78% and ~60% for high- and low-tumor fraction patients in both conventional and three-class models, respectively. However, patients with autoimmune or vascular diseases were much less frequently falsely

classified as harboring cancer (compare Fig. 6*C* to *SI Appendix*, Fig. S9*C*). For example, 40% of patients with VTE and 44% of patients with DM were wrongly classified as having cancer with a conventional model (*SI Appendix*, Fig. S9*C*) while only 6.7% of patients with VTE and 0.9% of patients with DM were wrongly classified as having cancer with the multiclass model (Fig. 6*C*).

Discussion

In the past decade, fragmentomic biomarkers have been increasingly investigated in case-controlled studies as high sensitivity, high specificity biomarkers for cancer screening and monitoring (17, 18, 41). Alterations in the fragmentation patterns have also been shown to be highly enriched in ctDNA, allowing for both in-vitro and in-silico enrichment of ctDNA to enhance the detection of orthogonal biomarkers such as somatic mutations and copy number alterations (51, 52). Our results confirm previous observations (17, 18, 50, 51) documenting the ability to use fragmentomics of cfDNA to distinguish patients with and without cancer. However, our observations of abnormal fragmentation patterns in individuals with VTE, SLE, DM, and SSc unequivocally document that these fragmentomic biomarkers are not specific for cancer.

Our observations are consistent with previous studies that found abnormal fragmentation patterns in conditions other than cancer. Chan et al. previously observed that individuals with SLE had significant alterations to fragment lengths, specifically an increase in the proportion of short fragments (53). Similarly, Zhu et al. recently identified abnormal fragment lengths and end-motifs that were associated with alterations to postprandial metabolic and immune states (54). Though all these and previous observations motivate serious concern for the ability of fragmentomics to distinguish patients with cancer from those without cancer in MCED testing, we showed that this concern is partially mitigated by our development of a multiclass approach to MCED that demonstrates improved specificity for patients with autoimmune or vascular disease while maintaining high sensitivity for patients with cancer (Fig. 6*C*).

The results of the current study are analogous to the historical results on cfDNA concentrations. As mentioned in the introduction, it was originally thought that high cfDNA concentrations were a specific marker for cancer. However, it was later found that high cfDNA concentrations were not specific for cancer, as they were elevated in numerous other circumstances (55–58). Now, the results reported in this paper show that fragmentation patterns, originally suggested to be a specific biomarker for cancer, are actually not specific for cancer. Moreover, we observed that individuals with abnormal fragmentation signatures (Cluster 1) had significantly higher cfDNA concentrations compared to those with normal signatures (Cluster 2), suggesting a possible relationship between the two biomarkers (Fig. 5*C*). Similarly, we observed a significant positive correlation between plasma cfDNA concentrations and plasma cfDNA fragmentation metrics (iwFAF, FLR, MDS) within the four disease groups that had a significant increase in cfDNA concentration (high tumor fraction cancer, low tumor fraction cancer, VTE, and DM) (Fig. 5 and *SI Appendix*, Fig. S8). Interestingly, we observed a slight negative correlation between fragmentation metrics and cfDNA concentrations in healthy individuals, a finding that is in line with previous studies (59).

The most parsimonious explanation of our data is that the same inflammatory process found in patients with autoimmune and vascular diseases is also found in cancer patients, and that this process is responsible for the abnormal fragmentation signatures. This explanation is supported by the fact that a wide range of inflammatory markers, including circulating proteins and cfDNA itself, are elevated in patients with abnormal fragmentation signatures (Fig. 5). Recent studies have demonstrated that increased concentrations of cfDNA directly activate innate immunity (60, 61) and are associated with disease activity and markers of inflammation in a range of conditions including cancer, VTE, SLE, DM, and SSc (50, 59, 62–64). Previous studies have come to similar conclusions, proposing that inflammatory processes such as necroptosis, NETosis, or other methods of phagocytosis may be responsible for the increase in cfDNA, and possibly for abnormal fragmentation patterns, in a variety of conditions (45, 65–67). To reconcile the observations that abnormal fragmentation signatures are both highly correlated with the tumor fraction but not tumor-specific, we propose that the same phagocytic mechanism is responsible for both the release of ctDNA into the plasma and the generation of abnormal fragmentation patterns. This hypothesis is supported by previous studies demonstrating that necrotic tumor cells release DNA mainly through phagocytosis (68). Interestingly, MPO was the most elevated protein in individuals with abnormal fragmentation signatures (Fig. 5 and *SI Appendix*, Fig. S4). MPO is the most abundant protein within neutrophil-extracellular traps (NETs) and directly generates reactive oxygen species (ROS) upon neutrophil activation (65). Five of the six (OPN, GDF-15, IL-8, OPG, NSE) remaining inflammatory proteins associated with abnormal fragmentation signatures are similarly induced by ROS and related to increased activity of the innate immune response (49, 69, 70). This theory is also consistent with the historic idea that cancer represents an unhealed wound, first introduced by Virchow in the 19th century (71). This interpretation is supported by many studies indicating that high levels of circulating inflammatory markers are commonly observed in cancer patients (72, 73). The data in Fig. 5 add to this body of literature by showing strikingly high correlations between typical markers of inflammation and an abnormal fragmentation signature.

An alternative explanation of our data is that there are two different processes responsible for the abnormal fragmentation signatures, one occurring within the neoplastic cells of the cancer cells and one occurring in the absence of any cancer in the body. For example, it is possible that inflammatory cells degrade DNA in the same way as neoplastic cells, leading to similar fragmentation signatures. We believe this potential explanation is unlikely because it has been shown that most of the cfDNA in cancer patients is derived from leukocytes, not neoplastic cells (38). A more tenable explanation is that the same type of cells (e.g., monocytes or granulocytes) within a tumor degrade their own cellular DNA or DNA from other cells that they digest, then release this DNA into the circulation as part of NETosis or related inflammatory processes. Many cancers are known to harbor relatively high numbers of inflammatory cells, including macrophages, consistent with the "unhealed wound" concept (74). This explanation is supported by the seminal observation that plasma DNA fragments derived from CAR19 T-cells or derived from lymphomas of patients treated with these CAR19 T-cells have similar size distributions (75).

There are several practical implications of our data. They place limits on the specificity of fragmentomics-based methods for cancer detection because the fragmentation signals are not specific for cancer. The prevalence of autoimmune diseases is high, affecting ~5 to 10% of the population, and is increasing (76–79). One could exclude all patients with known vascular or autoimmune diseases from fragmentomics-based testing, but this would not be ideal. Notably, a subset of patients with vascular or autoimmune diseases are at high risk of having an undiagnosed cancer so cancer screening in this population is particularly worthwhile (29, 63, 80–82). Indeed, the rationale for the current study was our serendipitous observation that patients with VTE had abnormal fragmentation patterns. A second reason is that some patients with autoimmune diseases are not yet aware that they have the disease as the presentation is often

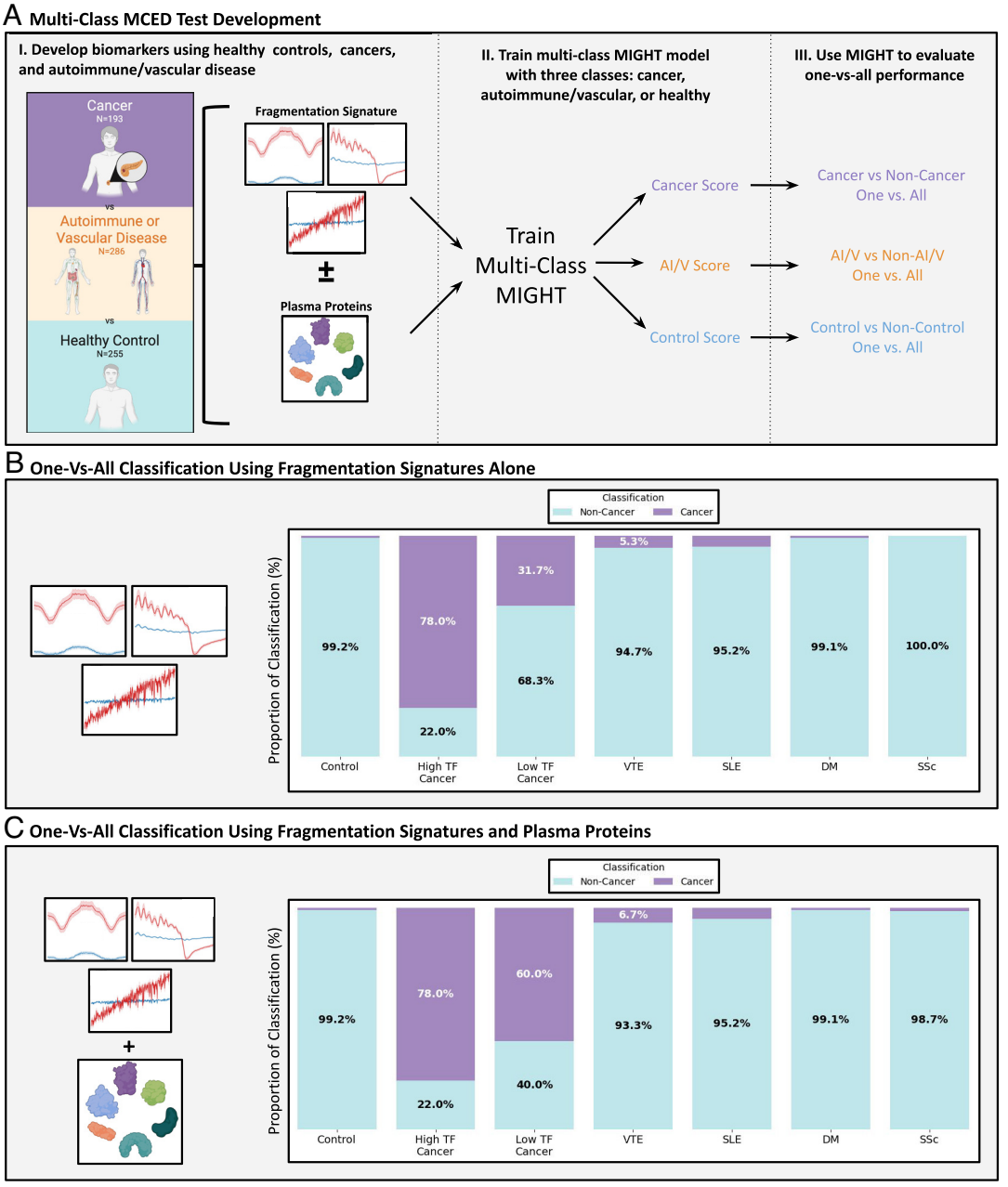

Fig. 6. Multiclass MCED test development. (*A*) Development of a multiclass MIGHT model using three classes: i) individuals with no diagnosis of disease (healthy controls; n = 255) ii) individuals with autoimmune or vascular disease (n = 286) and iii) individuals with cancer (n = 193). (*B*) One-vs.-all performance of multiclass MIGHT using fragmentation signatures alone. (*C*) One-vs.-all performance of multiclass MIGHT using fragmentation signatures and plasma proteins. Multiclass MIGHT posteriors for all patients are provided in Dataset S8.

nonspecific or disease can be low grade and symptoms can be intermittent. On the other hand, it is possible that such patients may have less abnormal fragmentomics signatures than those we studied, who were already diagnosed with these diseases. Conversely, it is also possible that other diseases with inflammatory components, including various infectious diseases or allergic conditions, have abnormal fragmentomicsignatures,leadingtoadditionalpositiveresultsinfragmentomics-based tests for the presence of cancer. These are all important questions for future investigation.

On the bright side, the knowledge that patients with vascular and autoimmune diseases have fragmentation patterns similar to those of advanced cancer patients offers clues for both basic science and clinical medicine. This knowledge could lead to a better appreciation of the cellular and biochemical mechanisms responsible for the generation of cfDNA in both healthy and disease states, which are currently poorly understood. And though the fragmentation patterns are similar among cancer patients and those with other diseases, they are not identical, and vary with the type of noncancerous disease. It is possible that AI-based investigation could distinguish the fragmentation signatures found in cancer patients from those with other diseases, thereby restoring the specificity of such tests. And viewed from the perspective of the autoimmune or vascular disease patient rather than from that of the cancer patient, it is conceivable that fragmentation patterns could improve early diagnosis of autoimmunity or vascular disease, help assess the effects of treatment, distinguish one type of autoimmune disease from another, or identify distinct disease subsets.

There are of course limitations to the current study. Though MIGHT does not require an external validation set to obtain confidence limits about its predictions within the sampled population (Curtis et al., in press at PNAS), no algorithm, including MIGHT, can exclude confounding variables or that the same results would be

obtained if other populations, library preparation methods, or DNA purification methods were studied (83). Moreover, the current sample size for some disease groups (e.g., SLE, n = 21) was small and future studies should make use of larger cohorts to document reproducibility. In addition, the reported relationships between cfDNA fragmentation and inflammatory conditions/markers are associations at this stage, and a causal relationship remains to be established. Finally, while known batch effects were removed (*Methods*) the effect of unknown batch effects may still affect the results. On the other hand, our use of multiple nonoverlapping cohorts of healthy controls, normalization of known preanalytic batch effects, use of unsupervised methods for dimension reduction and clustering, and high-powered supervised methods such as MIGHT do lend confidence in the results presented within this study.

Methods

Experimental Study Design. This study was approved by the Institutional Review Boards for Human Research at Johns Hopkins Medical Institutes and other participating institutions in compliance with the Health Insurance Portability and Accountability Act. No proper sample size was calculated; samples were chosen on the basis of availability. All individuals participating in the study provided written consent. Blood was collected in Streck tubes or in Ethylenediaminetetraacetic acid (EDTA) tubes, and plasma separated from cells within 2 d or 2 h, respectively. Plasma was purified using the BioChain Cell-free DNA Extraction Kit (Cat X K5011625). All patients were deidentified, and patients are not known to anyone outside the research group. Demographics for the individuals are included in Dataset S1. Certain data from individuals from the control groups, as well as from the cancer groups, IIB, and VII, were included in a recent study to develop MIGHT (manuscript submitted to PNAS), though with different study goals.

Plasma samples from adult participants with unprovoked VTEs were collected within 10 d of the VTE event. Patients were eligible if they were aged ≥40 y and had a first episode of symptomatic, objectively confirmed, unprovoked VTE, i.e., lower-extremity deep vein thrombosis and/or pulmonary embolism. VTE was considered unprovoked if it was not related to pregnancy or puerperium, recent immobilization for ≥3 d (<3 mo), recent surgery (<3 mo), recent hospitalization (<3 mo), known genetic or acquired thrombophilia, or use of systemic estrogen therapy. Exclusion criteria were a known malignancy in the previous 5 y and enrollment >10 d after the VTE event. Patients with suspected cancer at presentation were only allowed to participate if the cancer had not yet been objectively confirmed by histology or cytology. All participants provided written informed consent prior to enrollment. All VTE patients assessed in our study were not diagnosed with cancer within a follow-up period of at least 2 y following the thromboembolic event.

Adult patients with SLE met either the revised American College of Rheumatology (ACR) criteria for SLE or the 2012 Systemic Lupus International Collaborating Clinics classification criteria (84). Patients ≥18 y of age with SLE were recruited from the Johns Hopkins Lupus Clinic at scheduled outpatient visits or, if hospitalized, the inpatient services at The Johns Hopkins Hospital.

Adult participants with scleroderma or DM were recruited from the Johns Hopkins Scleroderma Center Research Registry and the Johns Hopkins Myositis Research Registry. Participants in the scleroderma registry had features concerning for scleroderma either defined by 2013 ACR/EULAR classification criteria, 3 of 5 CREST (calcinosis, Raynaud's, esophageal dysmotility, sclerodactyly, telangiectasia) criteria, definite Raynaud's phenomenon, abnormal nailfold capillaries, and a scleroderma specific autoantibody, or a high titer scleroderma autoantibody. All DM patients met ACR/EULAR 2017 Idiopathic Inflammatory Myositis Classification Criteria.

Whole Genome Sequencing. We previously developed a library preparation workflow that can efficiently recover input DNA fragments and simultaneously incorporate double-stranded molecular barcodes (85). In brief, libraries were prepared with cfDNA using an Accel-NGS 2S DNA Library Kit (Swift Bio- sciences, 21024) with the following critical modifications: 1) DNA was pretreated with 3 U of USER enzyme (New England BioLabs, M5505L) for 15 min at 37 °C to excise uracil bases; 2) the SPRI bead/PEG NaCl ratios used after each reaction were 2.0×, 1.8×, 1.2×, and 1.05× for end repair 1, end repair 2, ligation 1, and ligation 2, respectively; 3) a custom 50 μM 3′ adapter was substituted for reagent Y2 and 4) a custom 42 μM 5′ adapter was substituted for reagent B2. Libraries were subsequently PCR amplified in 50-μL reactions using primers targeting the ligated adapters. The following reaction conditions were used: 1× NEBNext Ultra II Q5 Master Mix (New England BioLabs, M0544L), 2 μM universal forward primer and 2 μM universal reverse primer. Libraries were PCR-amplified according to the following protocol: 98 °C for 20 s, then eight cycles of 98 °C for 10 s, 65 °C for 75 s, and hold at 4 °C. The products were purified with 1.8× SPRI beads (Beckman Coulter, B23317) and eluted in EB buffer (Qiagen). Whole genome libraries were sequenced with paired-end 2× 100 bp sequencing on either a HiSeq 4000 or NovaSeq 6000 to a median depth of 26.9M read pairs (IQR 23.5–30.5).

Bioinformatic Pipeline. Fastq files were demultiplexed using a custom script that utilized index sequences added during library preparation. Demultiplexed read 1 and read 2 fastq files were trimmed using a custom script to remove 27 base oligonucleotides added during library preparation. Trimmed sequences were then aligned to the hg19 genome with bowtie2 (86) using end-to-end alignment. After alignment, UID duplicates were removed using a custom script. Picard AddOrReplaceReadGroups (87) was used to add read groups. Samtools flagstat (88) was used to evaluate alignment. Binary Alignment Map (BAM) files were converted to bed format using bedtools (89). Custom scripts for the analysis of fragmentation patterns were all written using python 3.9.12. All scripts are available from the authors upon request. Samples that had multiple technical replicates were consolidated into a single sample by taking the average of each variable between all replicates.

Quality Control. Each sample was evaluated based on library DNA concentration, read alignment metrics, GC content of the sequenced molecules, and total molecules. Any samples with less than 4 ng/μL of DNA, greater than 2.5% singletons, less than 80% of reads mapped, less than 80% of reads properly paired, less than 43% GC content, greater than 48% GC content, or less than five million usable molecules were removed from analysis. A total of 24 samples, representing 2.3% of the processed samples, were removed. Only properly paired reads with a MAPQ>30 mapped to autosomal chromosomes were used for the analysis of fragmentation patterns.

Preanalytic Conditions. Variations in preanalytic conditions including blood collection tubes, blood processing, blood storage, and DNA extraction have been shown to have significant effects on the analysis of cfDNA (90). When comparing samples collected in EDTA or Streck tubes we observed significant differences in fragment end-positions, lengths, and end-motifs (*SI Appendix*, Fig. S11). Samples in EDTA or Streck tubes also had differences in the time from blood collection to plasma separation (2 d or 2 h, respectively). To account for confounders associated with tube type or time to plasma separation we separated the normalization cohort into two nonoverlapping groups: one containing healthy controls processed in EDTA tubes (n = 62) and another containing healthy controls processed in Streck tubes (n = 68). Z-scores for samples in other groups were calculated using the normalization cohort corresponding to the tube in which the sample was processed. Information for the tube in which each sample was processed and subsequently normalized can be found in Dataset S1.

Fragment Length Analysis. Fragment length was extracted from the BAM files using the TLEN alignment field. Only fragments between 70 bp and 500 bp were analyzed. Fragment length frequencies were calculated as the count of each individual length divided by the total number of fragments of length 70 to 500 bp. The proportion of short and long fragments were calculated as the sum of frequencies of fragments 100 to 150 bp and 151 to 220 bp, respectively. The FLR was calculated as the proportion of short fragments divided by the proportion of long fragments.

Fragment End-Motif Analysis. Fragment start position, end position, and strandedness (±) were extracted from the fragment BED file. The full nucleotide sequence of each read pair was then extracted from the hg19 reference genome using bedtools nuc (89). Orientation of 5′ and 3′ of each fragment was inferred using the strandedness of each molecule. Fragments that aligned to the nonreference (−) strand of the hg19 reference genome were reverse complemented. The final four bases (tetramer) were then extracted for both the 5′ and 3′ ends. After analyzing all fragments the genome-wide frequencies of the 5′ and 3′ end-motifs were calculated by dividing by the count of each motif by the total number of fragments analyzed. Due to end-repair of the 3′ end during library preparation, the average frequency between the 5′ end-motif and reverse complement of the 3′ end-motif was used as the final frequency.

Fragment End-Position Analysis. A BED file for RPRs was downloaded from Budhraja et al. Bedtools intersect (v2.30.0) was then used to intersect the sample fragment bed file with each RPR, requiring only a single base position of

overlap between the molecule and the genomic loci to be included. To calculate the fragment end-position variables, we first determined the central base position of each genomic locus. If there was an even number of bases in the locus, the central position was rounded up. Each bond between the bases was considered as a possible breakpoint (e.g. counting the number of phosphodiester bonds, not nucleotides). For each genomic locus, we analyzed positions −150 to +150 from the central position, where position -1 is the bond between the central nucleotide and the nucleotide directly upstream. The number of fragment-ends at that position was divided by the number of fragments that overlapped that position (e.g., had coverage at the nucleotide upstream and downstream of that position). In total, 300 possible positions were evaluated for each locus.

ichorCNA. ichorCNA version 0.3.2 was downloaded from the GitHub repository https://github.com/broadinstitute/ichorCNA. Wig files were generated using readCounter with arguments –window 5000000 –quality 30. CreatePanelOfNormals.R was used to generate a panel of normals (n = 124). The "normal" initialization parameters selected were c(0.95, 0.99, 0.995, 0.999) and the ploidy initialization parameter was 2.

Fragmentation Signatures. For the analysis of fragmentation signatures, PCA was first performed on z-scores for each fragmentation pattern independently using *sklearn.decomposition.PCA(n_components=0.9)*. Resulting principal components for each fragmentation pattern were then consolidated into a single matrix. AutoGMM from the graspologic package was applied to perform model-based clustering on the PCs of fragmentation data. It automates hyperparameter selection by iterating over combinations of candidate parameters: the number of clusters (we set both min_components=2 and max_components=2), covariance structure (default: all types, e.g., spherical, diagonal, tied, full), and initialization methods (default: k-means++ and random). The optimal model was selected via the Bayesian information criterion, balancing goodness-of-fit and model complexity. We retained the default tolerance (tol=1e-3) and maximum iterations (max_iter=100), ensuring convergence. This approach accommodates nonspherical cluster geometries in the PC-reduced space while automatically inferring cluster numbers, mitigating biases from manual parameter tuning.

Statistical Analyses. All statistics were generated using python version 3.9.12 and *scipy* version 1.13.1. One-sided tests were performed for comparisons of cfDNA concentration and plasma protein levels using *scipy.stats.mannwhitneyu(alternative='less')*. All other statistics generated were with two-sided Mann–Whitney *U* tests. Correlation coefficients were calculated using *scipy.stats.pearsonr*.

MIGHT. MIGHT was installed from GitHub (https://docs.neurodata.io/treeple/dev/install.html) using treeple version 0.9.1. All MIGHT analyses were run using the following parameters: *est=HonestForestClassifier(n_estimators=n_estimators, max_samples=1.6, max_features='sqrt', bootstrap=True, stratify=True, n_jobs=10, random_state=9515, honest_prior="ignore," honest_method='apply', honest_fraction=0.367, kernel_method=True, tree_estimator=ObliqueDecisionTreeClassifier(feature_combinations=1.5))*.

Evaluation of Plasma Proteins. The Bioplex 200 platform (Biorad, Hercules CA) was used to determine the concentration of multiple target proteins in the plasma samples. Luminex bead-based immunoassays (Millipore, Billerica NY) were performed following the manufacturers protocols and concentrations were determined using five parameter log curve fits (using Bioplex Manager 6.0) with vendor provided standards and quality controls. The HCC BP1MAG-58K panel was used to detect AFP, CA125, CA15-3, CA19-9, CEA, HGF, OPN, CYFRA21.1, IL-8, and FGF2. The HCMBMAG-22K panel was used to detect GDF-15, NSE, OPG, and DKK1. The HCC BP3MAG-58K panel was used to detect MPO and SHBG. The HTMP1MAG-54K panel was used to detect TIMP-1.

Data, Materials, and Software Availability. Anonymized genetic sequencing data have been deposited in EGA (https://ega-archive.org/studies/EGAS00001008004) (91).

ACKNOWLEDGMENTS. The authors thank Shervin Tabrizi for his participation as a reviewer. This study was supported by NIH grant R21NS113016 (C.B.), NIH grant RA37CA230400 (C.B.), NIH grant U01CA230691 (C.B. and N.P.), Oncology Core CA 06973 (K.W.K., B.V., N.P.), The Virginia and D.K. Ludwig Fund for Cancer Research (C.B., K.W.K., B.V., N.P., and C.D.), Commonwealth Fund (N.P.), Thomas M Hohman Memorial Cancer Research Fund (C.B.), The Sol Goldman Sequencing Facility at Johns Hopkins (B.V.), The Conrad R. Hilton Foundation (K.W.K., N.P., and B.V.), Benjamin Baker Endowment 80049589 (Y.W. and C.D.), NIH grant Ovarian Cancer SPORE DRP 80057309 (C.D.), Swim Across America (C.D.), Burroughs Wellcome Career Award for Medical Scientists (C.B.), Thomas M Hohman Memorial Cancer Research Fund (C.B)., NIH grant U01CA230691 (C.B.), and National Health and Medical Research Council Investigator Grant APP1194970 (J.T.). C.D. would like to acknowledge funding from Swim Across America. M.F.K. was supported by National Institutes of Health/National Institute of Allergy and Infectious Diseases grant 1R21AI176764-01, the Rheumatology Research Foundation Investigator Award, the Harrington Discovery Institute Scholar-Innovator Award, the Jerome L. Greene Foundation, the Cupid Foundation, and the Stephen & Renee Bisciotti Foundation.

Author affiliations: [a]Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [b]Department of Oncology, the Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [c]The Ludwig Center for Cancer Genetics and Therapeutics, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [d]The Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [e]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218; [f]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218; [g]Department of Computer Science, Columbia University, New York, NY 10027; [h]Department of Applied Mathematics and Statistic, Johns Hopkins University, Baltimore, MD 21205; [i]The HHMI, Chevy Chase, MD 20815; [j]Division of Personalized Oncology, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; [k]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC 3011, Australia; [l]Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, VIC 3010, Australia; [m]Department of Medical Oncology, Western Health, Melbourne, VIC 3021, Australia; [n]BioMedical Research Center, Pham Ngoc Thach University of Medicine, Ho Chi Minh City 72510, Vietnam; [o]Clinical Genetics Research Group, Saigon Precision Medicine Research Center, Ho Chi Minh City 72512, Vietnam; [p]Saigon Precision Medicine Research Center, Ho Chi Minh City 72512, Vietnam; [q]School of Biomedical Engineering, University of Technology, Sydney, NSW 2007, Australia; [r]Tâm Anh Research Institute, Ho Chi Minh City 721000, Vietnam; [s]Centre for Health Technologies, University of Technology, Sydney, NSW 2007, Australia; [t]School of Population Health, University of New South Wales, Kensington, NSW 2003, Australia; [u]Division of Rheumatology, Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21218; [v]Institute for NanoBioTechnology, The Johns Hopkins University, Baltimore, MD 21218; [w]Department of Vascular Medicine, Amsterdam Cardiovascular Sciences, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam 1105 AZ, the Netherlands; [x]Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [y]Department of Medicine, Johns Hopkins Medical Institutes, Baltimore, MD 21205; and [z]Division of Quantitative Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Author contributions: S.D.C., N.P., B.V., and C.D. designed research; S.D.C., T.L., Y.B., Y.W., L.D., M.P., J.P., N.S., C.T., and B.V. performed research; S.D.C., T.L., Y.B., Y.W., S.P., A.L., H.X., E.O., J.T., P.G., L.T.H.-P., B.N.H.T., T.S.T., T.V.N., M.F.K., M.P., A.R., C.A.M., A.A.S., F.M., N.v.E., P.-V.S.G., C.B., K.W.K., N.P., J.T.V., and C.D. contributed new reagents/analytic tools; S.D.C., T.L., Y.B., and B.V. analyzed data; and all authors wrote the paper.

Reviewers: D.L., Cornell University; and V.A., Broad Institute.

Competing interest statement: J.T. has a speaking, advisory, and consultancy with Haystack Oncology, Amgen, Novartis, AstraZeneca, Merck Serono, Merck Sharp & Dohme, Beigene, Pierre Fabre, Bristol Myers Squibb, Gilead, and Daiichi Sankyo. K.W.K., B.V., and N.P. are founders of Thrive Earlier Detection, an Exact Sciences Company. K.W.K., N.P., and C.D. are consultants to Thrive Earlier Detection. K.W.K. and N.P. are consultants to Neophore. K.W.K., B.V., N.P., and C.D. hold equity in Exact Sciences. K.W.K., B.V., and N.P. are founders of and own equity in Haystack Oncology & ManaT Bio. B.V. is a consultant to and holds equity in Catalio Capital Management. C.B. is a co-founder of OrisDx. C.B. and C.D. are co-founders of Belay Diagnostics. K.W.K., B.V., and N.P., hold equity in and are consultants to CAGE Pharma. The companies named above, as well as other companies, have licensed previously described technologies related to the work described in this paper from Johns Hopkins University. C.B., K.W.K., N.P., B.V., and C.D., are inventors on some of these technologies. Licenses to these technologies are or will be associated with equity or royalty payments to the inventors as well as to Johns Hopkins University. Patent applications on the work described in this paper may be filed by Johns Hopkins University. The terms of all these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies.

1. S. A. Leon, B. Shapiro, D. M. Sklaroff, M. J. Yaros, Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res.* **37**, 646–650 (1977).
2. M. Fleischhacker, B. Schmidt, Circulating nucleic acids (CNAs) and cancer–A survey. *Biochim. Biophys. Acta* **1775**, 181–232 (2007).
3. L. Kananen *et al.*, Circulating cell-free DNA level predicts all-cause mortality independent of other predictors in the Health 2000 survey. *Sci. Rep.* **10**, 13809 (2020).
4. O. Fridlich *et al.*, Elevated cfDNA after exercise is derived primarily from mature polymorphonuclear neutrophils, with a minor contribution of cardiomyocytes. *Cell. Rep. Med.* **4**, 101074 (2023).
5. J. Yi *et al.*, Increased plasma cell-free DNA level during HTNV infection: Correlation with disease severity and virus load. *Viruses* **6**, 2723–2734 (2014).
6. D. Sidransky *et al.*, Identification of p53 gene mutations in bladder cancers and urine samples. *Science* **252**, 706–709 (1991).
7. D. Sidransky *et al.*, Identification of ras oncogene mutations in the stool of patients with curable colorectal tumors. *Science* **256**, 102–105 (1992).
8. J. D. Cohen *et al.*, Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).

9. M. C. Liu, G. R. Oxnard, E. A. Klein, C. Swanton, M. V. Seiden, Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
10. L. Raman, A. Dheedene, M. De Smet, J. Van Dorpe, B. Menten, WisecondorX: Improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res.* **47**, 1605–1614 (2019).
11. Y. Wang *et al.*, Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas. *Sci. Transl. Med.* **7**, 293ra104 (2015).
12. C. Douville *et al.*, Seq-ing the SINEs of central nervous system tumors in cerebrospinal fluid. *Cell Rep. Med.* **4**, 101148 (2023).
13. S. U. Springer *et al.*, Non-invasive detection of urothelial cancer through the analysis of driver gene mutations and aneuploidy. *Elife* **7**, e32143 (2018).
14. I. Kinde *et al.*, Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci. Transl. Med.* **5**, 167ra164 (2013).
15. T. H. T. Cheng *et al.*, Noninvasive detection of bladder cancer by shallow-depth genome-wide bisulfite sequencing of urinary cell-free DNA for methylation and copy number profiling. *Clin. Chem.* **65**, 927–936 (2019).
16. C. Douville *et al.*, Assessing aneuploidy with repetitive element sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 4858–4863 (2020).
17. A. J. Widman *et al.*, Ultrasensitive plasma-based monitoring of tumor burden using machine-learning-guided signal enrichment. *Nat. Med.* **30**, 1655–1666 (2024).
18. A. R. Thierry, Circulating DNA fragmentomics and cancer screening. *Cell Genom.* **3**, 100242 (2023).
19. S. C. Ding, Y. M. D. Lo, Cell-free DNA fragmentomics in liquid biopsy. *Diagnostics (Basel)* **12**, 978 (2022).
20. M. W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, J. Shendure, Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
21. D. S. C. Han *et al.*, The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Human Genet.* **106**, 202–214 (2020).
22. N. Umetani *et al.*, Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J. Clin. Oncol.* **24**, 4270–4276 (2006).
23. S. Cristiano *et al.*, Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
24. S. D. Curtis *et al.*, Identifying cancer patients from GC-patterned fragment ends of cell-free DNA. medRxiv [Preprint] (2022). https://www.medrxiv.org/content/10.1101/2022.08.02.22278319v1 (Accessed 3 August 2022).
25. K. K. Budhraja *et al.*, Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer. *Sci. Transl. Med.* **15**, eabm6863 (2023).
26. C. Douville *et al.*, Machine learning to detect the SINEs of cancer. *Sci. Transl. Med.* **16**, eadi3883 (2024).
27. Q. Zhou *et al.*, Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209852119 (2022).
28. A. V. Annapragada *et al.*, Genome-wide repeat landscapes in cancer and cell-free DNA. *Sci. Transl. Med.* **16**, eadj9283 (2024).
29. F. I. Mulder *et al.*, Venous thromboembolism in cancer patients: A population-based cohort study. *Blood* **137**, 1959–1969 (2021).
30. F. I. Mulder *et al.*, Platelet RNA sequencing for cancer screening in patients with unprovoked venous thromboembolism: A prospective cohort study. *J. Thromb. Haemost.* **21**, 905–916 (2023).
31. F. Mouliere *et al.*, Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).
32. V. A. Adalsteinsson *et al.*, Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
33. L. M. Sack *et al.*, Profound tissue specificity in proliferation control underlies cancer drivers and aneuploidy patterns. *Cell* **173**, 499–514.e23 (2018).
34. K. Sun *et al.*, Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res.* **29**, 418–427 (2019).
35. M. S. Esfahani *et al.*, Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* **40**, 585–597 (2022).
36. P. Ulz *et al.*, Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat. Commun.* **10**, 4666 (2019).
37. H. Markus *et al.*, Analysis of recurrently protected genomic regions in cell-free DNA found in urine. *Sci. Transl. Med.* **13**, 581 (2021).
38. A. K. Mattox *et al.*, The origin of highly elevated cell-free DNA in healthy individuals and patients with pancreatic, colorectal, lung, or ovarian cancer. *Cancer Discov.* **13**, 2166–2179 (2023).
39. M. B. Giacona *et al.*, Cell-free DNA in human blood plasma: Length measurements in patients with pancreatic cancer and healthy controls. *Pancreas* **17**, 89–97 (1998).
40. J. E. Medina *et al.*, Early detection of ovarian cancer using cell-free DNA fragmentomes and protein biomarkers. *Cancer Discov.* **15**, 105–118 (2025).
41. I. van't Erve *et al.*, Cancer treatment monitoring using cell-free DNA fragmentomes. *Nat. Commun.* **15**, 8801 (2024).
42. P. Jiang *et al.*, Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1317–E1325 (2015).
43. F. Mouliere, N. Rosenfeld, Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3178–3179 (2015).
44. L. Serpas *et al.*, Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 641–649 (2019).
45. Z. Zhou *et al.*, Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220982120 (2023).
46. P. Jiang *et al.*, Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov.* **10**, 664–673 (2020).
47. L. B. Alexandrov *et al.*, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
48. T. L. Athey, T. Liu, B. D. Pedigo, J. T. Vogelstein, Autogmm: Automatic and hierarchical gaussian mixture modeling in python. arXiv [Preprint] (2019). https://arxiv.org/abs/1909.02688 (Accessed 12 August 2021).
49. M. Scatena, L. Liaw, C. M. Giachelli, Osteopontin: A multifunctional molecule regulating chronic inflammation and vascular disease. *Arterioscler. Thromb. Vasc. Biol.* **27**, 2302–2309 (2007).
50. P. Mondelo-Macía, P. Castro-Santos, A. Castillo-García, L. Muinelo-Romay, R. Diaz-Peña, Circulating free DNA and its emerging role in autoimmune diseases. *J. Pers. Med.* **11**, 151 (2021).
51. C. T. Maansson *et al.*, In vitro size-selection of short circulating tumor DNA fragments from late-stage lung cancer patients enhance the detection of mutations and aneuploidies. *J. Liquid Biopsy* **4**, 100141 (2024).
52. H. Markus *et al.*, Refined characterization of circulating tumor DNA through biological feature integration. *Sci. Rep.* **12**, 1928 (2022).
53. R. W. Y. Chan *et al.*, Plasma DNA aberrations in systemic lupus erythematosus revealed by genomic and methylomic sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5302–E5311 (2014).
54. Z. Zhu *et al.*, Dynamic profiling of cell-free DNA fragmentation uncovers postprandial metabolic and immune alterations. *Human Genomics* **19**, 27 (2025).
55. S. B. Brusca *et al.*, Plasma cell-free DNA predicts survival and maps specific sources of injury in pulmonary arterial hypertension. *Circulation* **146**, 1033–1045 (2022).
56. K. Cano-Gamez *et al.*, The circulating cell-free DNA landscape in sepsis is dominated by impaired liver clearance. bioRxiv [Preprint] (2025). https://www.biorxiv.org/content/10.1101/2025.02.17.638622v1 (Accessed 25 February 2025).
57. O. Fridlich *et al.*, Elevated cfDNA after exercise is derived primarily from mature polymorphonuclear neutrophils, with a minor contribution of cardiomyocytes. *Cell Rep. Med.* **4**, 101074 (2023).
58. T. R. Kolarova, H. S. Gammill, J. L. Nelson, C. M. Lockwood, R. Shree, At preeclampsia diagnosis, total cell-free DNA concentration is elevated and correlates with disease severity. *J. Am. Heart Assoc.* **10**, e021477 (2021).
59. Y. Malki *et al.*, Analysis of a cell-free DNA-based cancer screening cohort links fragmentomic profiles, nuclease levels, and plasma DNA concentrations. *Genome Res.* **35**, 31–42 (2025).
60. M. Korabecna *et al.*, Cell-free DNA in plasma as an essential immune system regulator. *Sci. Rep.* **10**, 17478 (2020).
61. L. K. M. Lam *et al.*, DNA binding to TLR9 expressed by red blood cells promotes innate immune activation and anemia. *Sci. Transl. Med.* **13**, eabj1008 (2021).
62. B. Duvvuri, C. Lood, Cell-free DNA as a biomarker in autoimmune rheumatic diseases. *Front. Immunol.* **10**, 502 (2019).
63. J. Jee *et al.*, DNA liquid biopsy-based prediction of cancer-associated venous thromboembolism. *Nat. Med.* **30**, 2499–2507 (2024).
64. Y. Wang *et al.*, Clinicopathological and circulating cell-free DNA profile in myositis associated with anti-mitochondrial antibody. *Ann. Clin. Transl. Neurol.* **10**, 2127–2138 (2023).
65. B. Pastor *et al.*, Association of neutrophil extracellular traps with the production of circulating DNA in patients with colorectal cancer. *iScience* **25**, 103826 (2022).
66. A. Paunel-Görgülü *et al.*, cfDNA correlates with endothelial damage after cardiac surgery with prolonged cardiopulmonary bypass and amplifies NETosis in an intracellular TLR9-independent manner. *Sci. Rep.* **7**, 17421 (2017).
67. M. Wang *et al.*, Biomarkers of peripheral blood neutrophil extracellular traps in the diagnosis and progression of malignant tumors. *Cancer Med.* **13**, e6935 (2024).
68. J.-J. Choi, C. F. Reich III, D. S. Pisetsky, The role of macrophages in the in vitro generation of extracellular DNA from apoptotic and necrotic cells. *Immunology* **115**, 55–62 (2005).
69. A. M. di Candia, D. X. de Avila, G. R. Moreira, H. Villacorta, A. S. Maisel, Growth differentiation factor-15, a novel systemic biomarker of oxidative stress, inflammation, and cellular aging: Potential role in cardiovascular diseases. *Am. Heart J. Plus* **9**, 100046 (2021).
70. L. E. DeForge *et al.*, Regulation of interleukin 8 gene expression by oxidant stress. *J. Biol. Chem.* **268**, 25568–25576 (1993).
71. M. Deyell, C. S. Garris, A. M. Laughney, Cancer metastasis as a non-healing wound. *Br. J. Cancer* **124**, 1491–1502 (2021).
72. D. Il'yasova *et al.*, Circulating levels of inflammatory markers and cancer risk in the health aging and body composition cohort. *Cancer Epidemiol. Biomarkers Prev.* **14**, 2413–2418 (2005).
73. T. H. Nøst *et al.*, Systemic inflammation markers and cancer incidence in the UK Biobank. *Eur. J. Epidemiol.* **36**, 841–848 (2021).
74. V. Thorsson *et al.*, The immune landscape of cancer. *Immunity* **51**, 411–412 (2018).
75. B. J. Sworder *et al.*, Determinants of resistance to engineered T cell therapies targeting CD19 in large B cell lymphomas. *Cancer Cell* **41**, 210–225.e5 (2023).
76. N. Conrad *et al.*, Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: A population-based cohort study of 22 million individuals in the UK. *Lancet* **401**, 1878–1890 (2023).
77. M. Murray, Guest blog: A major health crisis: The alarming rise of autoimmune disease. (2024). https://nationalhealthcouncil.org/blog/a-major-health-crisis-the-alarming-rise-of-autoimmune-disease/. Accessed 10 October 2024.
78. A. H. Abend *et al.*, Estimation of prevalence of autoimmune diseases in the United States using electronic health record data. *J. Clin. Invest.* **135**, e178722 (2025).
79. S. M. Hayter, M. C. Cook, Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun. Rev.* **11**, 754–765 (2012).
80. A. A. Shah, L. Casciola-Rosen, A. Rosen, A. Review: Cancer-induced autoimmunity in the rheumatic diseases. *Arthritis Rheumatol.* **67**, 317–326 (2015).
81. A. A. Shah, A. Rosen, L. Hummers, F. Wigley, L. Casciola-Rosen, Close temporal relationship between onset of cancer and scleroderma in patients with RNA polymerase I/III antibodies. *Arthritis Rheum.* **62**, 2787–2795 (2010).
82. M. Gayed, S. Bernatsky, R. Ramsey-Goldman, A. Clarke, C. Gordon, Lupus and cancer. (2009).
83. H. Wang *et al.*, A standardized framework for robust fragmentomic feature extraction from cell-free DNA sequencing data. *Genome Biol.* **26**, 141 (2025).
84. M. Petri *et al.*, Derivation and validation of the systemic lupus international collaborating clinics classification criteria for systemic lupus erythematosus. (2012).
85. J. D. Cohen *et al.*, Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands. *Nat. Biotechnol.* **39**, 1220–1227 (2021).
86. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
87. Broad Institute, Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. (2019). https://broadinstitute.github.io/picard/. Accessed 19 March 2019.
88. H. Li *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
89. A. R. Quinlan, I. M. Hall, BEDtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
90. V. Ungerer, J. A. Bronkhorst, S. Holdenrieder, Preanalytical variables that affect the outcome of cell-free DNA measurements. *Crit. Rev. Clin. Lab. Sci.* **57**, 484–507 (2020).
91. S. Curtis *et al.*, Fragmentation signatures in cancer patients resemble those of patients with vascular or autoimmune diseases. European Genome-phenome Archive. https://ega-archive.org/studies/EGAS00001008004. Deposited 29 July 2025.