Machine learning to detect the SINEs of cancer

Abstract

We previously described an approach called RealSeqS to evaluate aneuploidy in plasma cell-free DNA (cfDNA) through the amplification of ~350,000 repeated elements with a single primer. We hypothesized that an unbiased evaluation of the large amount of sequencing data obtained with RealSeqS might reveal other differences between plasma samples from patients with and without cancer. This hypothesis was tested through the development of a novel machine-learning approach called Alu Profile Learning Using Sequencing (A-PLUS) and its application to samples from 5108 individuals, 2037 with cancer and the remainder without cancer. Samples from cancer patients and controls were pre-specified into four cohorts used for: 1) model training, 2) analyte integration and threshold determination, 3) validation, and 4) reproducibility. A-PLUS alone provided a sensitivity of 40.5% across 11 different cancer types in the Validation Cohort, at a specificity of 98.5%. Combining A-PLUS with aneuploidy and 8 common protein biomarkers detected 51% of 1167 cancers at 98.9% specificity. We found that part of the power of A-PLUS could be ascribed to a single feature – the global reduction of AluS sub-family elements in the circulating DNA of cancer patients. We confirmed this reduction through the analysis of another independent dataset obtained with a very different approach (whole genome sequencing). The evaluation of Alu elements therefore has the potential to enhance the performance of several methods designed for the earlier detection of cancer.

Commented [CD1]: Not a good idea to use median specificity when prostate, kidney and head and neck are all new to the validation set and all very low. Median sensitivity is not very impressive unless you restrict to the original cancer types

Introduction

Alu's are short interspersed nuclear elements (SINEs) of ~ 300 bp, with more than 1 million copies spread throughout the genome [1]. Their role in biology and evolution is an ongoing area of research, but some elements have already been shown to be involved in the regulation of tissue-specific genes. In cancer cells, they participate in structural changes, probably through homologous recombination given their widespread distribution throughout the genome and highly similar sequences [2 3]. Moreover, Alu's are hypomethylated early during tumor progression [4 5 6 7 8 9 10], and this feature has been incorporated into methods for the earlier detection of cancer through plasma cell-free DNA (cfDNA) analysis [11] . Alu's also reflect the altered fragmentation patterns found in cfDNA in cancer patients: one of the first plasma multi-cancer biomarkers used qPCR to calculate the ratio of short and long Alu segments [12 13 14].

Whole genome sequencing (WGS) has been widely employed in recent blood-based multi-cancer earlier detection assays. WGS should in theory allow evaluation of Alu elements, but predictive algorithms often discard them as a result of bioinformatic challenges stemming from their resemblance to each other and difficulties in mapping them unambiguously [15]. Even with the inclusion of mappable Alu elements, shallow WGS is inefficient to optimally evaluate Alu elements because they represent only a small fraction of the genome ~11% [1].

We have previously developed an approach, called RealSeqS, to specifically amplify Alu sequences [16]. RealSeqS offers advantages over WGS, including a simpler workflow that does not require library construction, a reduced requirement for input DNA, faster computational analysis, and higher sequencing coverage at individual Alu loci. Specifically, the RealSeqS workflow uses a single-primer pair to concomitantly amplify ~350,000 Alu elements. For an equivalent sequencing depth, RealSeqS achieves ~28-fold greater coverage of the Alu elements it amplifies than achievable with WGS at an equivalent sequencing depth, enabling improved predictive modeling.

As noted above, there is much precedent for Alu sequence elements being especially prone to epigenetic changes in various cancers. Epigenetic changes include those involving methylation and chromatin fragmentation patterns (as reviewed in [17]). We therefore hypothesized that the representation of specific Alu elements might be different in the cell-free DNA (cfDNA) of plasma of patients with cancer than in normal controls. Because there are so many Alu elements in the genome, an evaluation of this hypothesis required machine learning tools. Here, we report and test a machine-learning based approach, called Alu Profile Learning Using Sequencing (A-PLUS), to distinguish individuals with cancer from those without cancer on the basis of the representation of Alu elements in their cfDNA.

Main

Rationale and background of the assay

During the development and implementation of RealSeqS, we observed substantial differences in read depth at specific loci. These loci did not appear to correlate with cancer-specific copy number variations, which was the original intent of RealSeqS, or regions of high technical variability in the non-cancer controls. We hypothesized that an unbiased supervised machine-learning method might be able to select cancer-specific Alu element representations from RealSeqS data and be used to provide a metric in addition to aneuploidy for cancer patient classification.

The detection of cancer in asymptomatic patients, which is the primary goal of multi-cancer earlier detection tests, requires very high specificity. Designing a highly specific machine-learning algorithm to predict cancer status from the~350,000 features assessed in RealSeqS sequencing data thereby poses technical challenges. First, the selected features must be empiric and solely derived from the sequencing data. Unlike the evaluation of aneuploidy, or of mutations, methylation, or other epigenetic changes, we did not know (and still do not know) why certain Alu elements are more represented than others in the cfDNA from cancer patients. Presumably, these differences result from nucleases or chromatin structure characteristics that are different in cancer cells from those in normal cells, but this is speculative. We also do not know the cell types of origin of differently represented Alu loci in the cfDNA. They could be from neoplastic cells, from non-neoplastic cells of the same organ surrounding the cancer cells that have been destroyed by the cancer, or from one or more types of leukocytes. Note that leukocytes are the major source of cfDNA in patients with or without cancer [18].

Commented [CD2]: Add austin's paper

Other challenges facing the development of a highly-specific ML algorithm are more general than those noted above. ML models built on thousands of features often unintentionally result in predictions based on confounding variables such as ethnicity, sex, sample processing, or batch effects at any one of the experimental procedures used to obtain the final data rather than based on attributes of cancer per se [19]. Learning and integrating features optimally requires more training samples than the number of available features and this is impossible from a logistical standpoint when there are 350,000 features and limited research resources. This problem is often referred to as the curse of dimensionality (d>>n) [20]. Even under the best circumstances, ML models often do not reliably classify samples when tested on data from cohorts independent of those used for training, even when the ML model is based on multiple folds of cross-validation [21].

To address the challenges listed above, we incorporated several principles into the development of A-PLUS, as summarized in Fig. 1. First, we attempted to identify and eliminate confounding loci associated

with technical noise, ethnicity, sex, and batch differences. Second, we reduced the number of features from 350,000 using Principal Component Analysis (PCA). Third, we used an order of magnitude more samples (thousands rather than hundreds) than typically used in initial studies on new tests of cfDNA performance. Fourth, we divided samples into four pre-specified and non-overlapping cohorts to avoid over-fitting: Cohort 1 was used to choose features and train the ML model; Cohort 2 was used to establish thresholds for scoring samples as positive or negative; Cohort 3 was used to independently test (validate) the ML model based on Cohort 1 and the thresholds based on Cohort 2; and Cohort 4 was used to evaluate reproducibility of the scoring system.

Technical nuances underlying the development of A-PLUS are detailed in Methods and the code is publicly available (add link). In the remainder of this section, we discuss results related to the four cohorts described above.

Cohort 1: A-PLUS Feature selection and model training

Cohort 1 consisted of 459 samples previously analyzed for aneuploidy [16] using RealSeqS and additional 250 samples from controls of Vietnamese, Han Chinese, South Asian, and Native American/Inuit ethnicities not represented in our previous publications. These additional samples were included because inherited polymorphisms within Alu's could alter alignments and the subsequent red depth representation of Alu loci. Of the total 709, 400 samples were from patients without cancer and 309 from cancer patients (Table 1). The samples from patients included those with cancers of the breast, colorectal, esophagus, lung, liver, pancreas, ovary, and stomach. Slightly less cancer than control samples were purposefully used because we valued specificity over sensitivity; cancer samples erroneously classified as controls were deemed less harmful to performance of the final classifier than the reverse. Sample demographics are listed in Supplementary Table 1.

Important elements of the training included normalization of read depths and the removal of amplicons with insufficient coverage, removal of amplicons that were unstable based on T-tests (Methods). After these steps, there were 121,197 loci of the original 350,000 that remained. Principal Component Analysis (PCA) was then used to reduce dimensionality. Finally, a Support Vector Machine (SVM) was used to identify the 60 top PCA components. This feature number (60) was ~10% of the total number of unique patients in the training set which we considered a reasonable compromise to cope with the p>>n conundrum [20]. Performance was not assessed in Cohort 1. The non-cancer samples were used to generate a euploid reference panel to for aneuploidy calls (Materials and Methods). Both the cancers and non-cancer samples were used to generate and optimize model building.

A criticism of our prior study that introduced RealSeqS was the use of exclusion criteria. To reduce the chance of over-fitting and clearly establish metrics prior to evaluation of samples, we used the previously published metrics and thresholds for inclusion. We also assessed the presence of large molecular weight DNA as previously described. The metrics excluded 3.1% of Cohort 1 samples. We report A-PLUS scores and GAS for all excluded samples (Additional File XX). We felt the use of pre-analytic metrics is necessary given the numerous providers over the time span of ~6 years. We hope that automation in a CLIA laboratory would further reduce the number of samples flagged for consideration.

Cohort 2: Determination of thresholds

Cohort 2 included samples from 707 cancer patients (total of 852 samples) and 1049 controls without cancer (total of 1402 samples). ~~The use of more than one independent sample from the same individual helped us assess the stability of the assay, as replicates were generally made with different aliquots of plasma, different batches of PCR, and different sequencing runs.~~ Cancers included those from colon, esophagus, stomach, breast, colorectum, lung, ovary, and pancreas. The A-PLUS score corresponding to 99% specificity among the control samples was 0.28. At this this threshold, a median sensitivity of 60% was observed across the 8 different cancer types. Samples from patients with cancers of the esophagus and stomach had the highest sensitivities (86%) and samples from breast cancer patients had the lowest of 33% (Fig 2A).

We then generated global aneuploidy scores (GAS) for Cohort 2 samples with the same RealSeqS data used to generate A-PLUS scores. The GAS uses a different machine learning technique to generate a single score that reflects gains or losses of 39 chromosome arms, focusing on those that are typically observed in cancers [22]. A GAS threshold of >0.64 yielded a 99% specificity in the Cohort 2 control samples and a median sensitivity of 19% in the samples from cancer patients . The highest sensitivities at 99% specificity were achieved for cancers of the esophagus and liver. In the 687 cancers scoring negatively (i.e., below the 99%-specificity threshold) in the GAS assay, 318 (46%) scored positively (i.e., above the 99% specificity threshold) in the A-PLUS assay. Conversely, 81% of the cancer samples that scored positively in GAS also scored positively in A-PLUS (Fig. 2D) Supplementary Table 2)

Next, we compared A-PLUS sensitivity to a panel of 8 protein markers previously shown to be useful for cancer detection when employed at high thresholds (OPN, HGF, AFP, CA125, CA15-3, CEA, CA19-9, TIMP-1; Methods).

To integrate these 8 protein values into a single score, we used Logistic Regression to generate a protein score. Protein values <98th percentile in the control samples were set to zero to minimize the possibility that predictions would be based on technical noise or batch effects and thereby reduce overfitting. None of the proteins should be depleted in cancer. Performance for this protein score was assessed using 10 fold cross validation and a "protein score" threshold of >0.73 was selected in order to generate 99% specificity. This protein score produced a sensitivity of 54% in the cancers (Supplementary Table 2). The highest sensitivities for proteins were achieved for cancers of the liver and stomach. In the 535 cancers scoring negatively in the proteins assay, 44% scored positively in A-PLUS (Fig. 2D).

We used Logistic Regression to integrate A-PLUS and GAS with the proteins into a multi-analyte classifier. Like the protein score, feature values <98th percentile in the control samples were set to zero. Performance for this multi-analyte classifier was assessed using 10-fold cross validation and a classifier threshold of >0.87 was selected in order to generate 99% specificity. This threshold produced a median sensitivity of 72% (Fig XX; Supplementary Table 2). The highest sensitivities were achieved for cancers of the esophagus and liver. The performance for the multi-analyte classifier is higher than any individual analyte without sacrificing specificity. Using the pre-defined inclusion metrics from above, we flag 117 samples—5.2% (5.1% cancers and 5.4% non-cancers). These samples were still assessed for protein biomarkers without issue and classified in the multi-analyte classifier with the corresponding A-PLUS and GAS scores set to 0. All scores are reported in these flagged samples are available in Additional File XXX.

Cohort 3: Independent validation

Commented [CD3]: Removed and repeats added to Cohort 4

Commented [CD4]: (add column to Supplementary Table 1 indicating which samples had scores previously reported for GAS and proteins).

Commented [CD5]: I don't think the ROC curves don't add much

Receiver Operating Characteristic (ROC) curves for Cohort 2 based on A-PLUS alone, GAS alone, proteins alone, and their combination are shown in Fig. 3A.

Cohort 3 samples were from 2960 individuals, including 1167 patients with cancers of eleven types: Breast, Colorectum, Esophagus, Head and Neck, Kidney, Lung, Ovary, Pancreas, Prostate, Stomach, and Uterus (Table 1 and Supplementary Table 1). None of these patients had been described in prior publications. The cohort also included 1793 control samples from patients without known cancers; of these, A-PLUS scores were newly derived for all of them but GAS scores and protein scores on these patients have been published

The 99% thresholds defined by Cohort 2 controls were used to assess sensitivities in Cohort 3 for each of the assays described above. For A-PLUS alone, a median sensitivity of 27% was achieved in Cohort 3 (Supplementary Table 1, Supplementary Fig. 1A), while 1.5% of samples from controls were misclassified (i.e., specificity of 98.7%). As with Cohort 2, the highest sensitivity was observed in samples from patients with cancers of the esophagus (84%) and the lowest sensitivities were in the cancers not represented in Cohort 2—kidney, prostate, uterus, and head and neck. Breast continued to have low sensitivity. The sensitivities in the seven cancer types that were evaluated in both Cohorts 2 and 3 were similar but did exhibit some notable differences (Fig. 43AA). The specificity observed between the retrospectively defined 99% and the observed specificity in Cohort 3 was not statistically significant (p>0.05 Two portion Z-test). The sensitivities observed between Cohorts 2 and 3 for breast, esophagus, ovarian, pancreatic cancer show statistically no difference (P>0.05 Two proportion Z-test). Colorectal and stomach cancers were nominally significant (Lung p=0.046 and Stomach p=0.038—Two proportion Z-test). Lung (Cohort 2--54% vs Cohort 3--27%) exhibits very notable differences. The drop in sensitivity maybe attributed to histology differences between the cohorts. Cohort 2 lung cancers are predominantly squamous lung cancers (XX%) while Cohort 3 lung cancers are almost exclusively adenocarcinomas (XX% vs XX%). Other groups have reported higher sensitivities in squamous lung cancers compared to adenocarcinoma and squamous cell carcinomas [23].These data confirmed that the features and machine learning algorithms used to develop A-PLUS generalized to an independent dataset (Fig. 4D).

Overall, the sensitivities of aneuploidy alone (Fig. 4B) as well as proteins alone (Fig. 4C) were also similar in Cohorts 2 and 3 (Supp Fix). The multi-analyte test incorporating A-PLUS, aneuploidy, and proteins achieved a median of 37% sensitivity at 98.9% specificity in Cohort 3 using the thresholds determined in Cohort 2 (Fig. 4D). Among the cancers common to Cohorts 2 and 3, the median sensitivity was 75%. Specificity remains very high in our validation Cohort which is a major concern in screen especially when the underlying model uses machine-learning methods and several different analytes. Similar to A-PLUS, the multi-analyte classifier sensitivities for colorectal, stomach and lung were statistically significant between cohorts (P<0.05 Two proportion Z-test).

~~The additional~~ We graphically depicted the overlap in analytes (A-PLUS, GAS, PROT) at the pre-defined threshold as a Venn diagram in (Fig. 3E~~XX~~). Notably, A-PLUS made a greater contribution to positive calls than aneuploidy or proteins. A-PLUS detected 41@@% of the samples that were not detected by either aneuploidy or proteins (Supplementary Fig. 1D).

The cancers represented in Cohort 3 were relatively early in the sense that none had any distant metastatic lesions evident at presentation (i.e., none were Stage IV). When categorized by stage, the sensitivities for all analytes and the muti-analyte classifier increased with stage~~.~~ . There were four cancer types (derived from Head and Neck, Kidney, Prostate, or Uterus) in Cohort 3 that were not represented in either the cohort used for training (Cohort 1) or threshold definition (Cohort 2). Even so,

Commented [CD6]: No Samples in cohort 3 were previously reported

(add column to Supplementary Table 1 indicating which samples had scores previously reported for GAS and proteins).

Commented [CD7]: Two Proportion Z-Test

Commented [CD8]: Missing histology for some samples. Need to confirm after getting lung histology.

Commented [CD9]: Cite grail's study

Commented [CD10]: Violin plots of the data from each assay in Cohort 3 are plotted in Supplementary Fig. 2A, B, and C.

Commented [CD11]: This is not correct. A venn diagram does not do this. Some samples positive from one analyte may not be positive from the combined classifier.

We can illustrate the overlap at the pre specified thresholds

Formatted: Not Highlight

Formatted: Not Highlight

20% to 30% of these cancers could be detected with the multi-analyte test (Supplementary Fig. 4). Using the pre-defined inclusion metrics from above, we flag 57 samples—1.9% (0.9% cancers and 3.5% non-cancers. We note this is lower than the previous Cohorts.

Cohort 4: Reproducibility

The technical reproducibility of the A-PLUS and GAS assays (both based on RealSeqS sequencing data) were evaluating in 544 individuals (419 Non-cancers and 125 cancers) from Cohort 2 patients and 1142 individuals (1121 Non-cancers and 21 cancers) from Cohort 3 patients.

Separate aliquots of purified DNA (i.e., separate template molecules) from the same plasma sample were independently amplified using the single RealSeqS primer pair and the PCR products sequenced. In all samples, the sequencing was done on different days.

In this cohort, 54 individuals had at least 1 of the pairs flagged using the pre-defined metrics and 32% having both flagged. No flagged samples with either draw 1 or draw 2 were assessed for score reproducibility.

The scores in were highly correlated. Using the thresholds defined by Cohort 2, 95.7% of the 1632 pairs scored concordantly (either positively or negatively) for A-PLUS (Supp Fig. 5A). Of the 70 discordant samples, 23 scored just below the threshold in one of the two DNA aliquots (i.e., between the scores required for 98% and 99% specificity) and above the score required for 99% specificity in the other aliquot. Of the 70 discordant pairs, 29 are non-cancer false positive that do not replicate in the other draw. With GAS, 99.3% pairs were concordant (Fig. 5B). Of the 11 discordant, 6 are false positives that are not observed in the other repeat.

Alu subsets

We next asked whether there was a subset of the Alu loci included in the A-PLUS heuristic that were particularly important for its success in distinguishing samples from controls and cancer patients. While small numbers of AluY loci were enriched in samples from cancer patients, the most striking observation was a global reduction in read depth across all AluS loci. There are 686,962 AluS loci distributed thoughout the genome, and the AluS sub-family is younger than the AluY subfamily [24 25]. Using only a single feature - the average normalized read depth of AluS elements (herein dubbed AluS-Rep) from RealSeqS data - without any machine learning algorithms samples from cancer patients could be distinguished from those of controls (AUC = 0.70; Fig. 6A, B). A-PLUS scores and the proportional representation of AluS elements were inversely correlated (-0.19; p<2.2e-16 via Pearson's).

Commented [CD12]: BV said: Chris: is this true?)

CD: yes

We wondered whether the reduced representation of AluS elements in the cfDNA from cancer patients was the result of some unknown bias in amplification efficiency or sequencing generated from the RealSeqS approach. To answer this question, we evaluated WGS data from a publicly available dataset that included 266 cfDNA samples from cancer patients (25 bile duct; 54 breast; 22 colorectum; 27 stomach; 76 lung; 26 ovary; 35 pancreatic) and 260 samples from controls without cancer [26] (Supplementary Table 2). Samples from cancer patients indeed had a global reduction of AluS compared to controls (p<2.2e-16, one sided t-test; Fig. 6C;). Moreover, this single feature could distinguish samples from cancer patients and controls with an AUC of 0.84 in ROC analysis (Supp Fig. XX).

Finally, we evaluated whether AluS-Rep in WGS data could be combined with the evaluation of aneuploidy in the same dataset. We used the WisecondorX (Materials and Methods) to detect aneuploidy and counted the number of aberrant (z>5 or z<-5 WiseCondorX default settings) regions throughout the genome in the FinaleDB dataset. We combined scores in a model agnostic fashion with a Boolean OR. AluS<0.657 OR Copy Number Alterations >3 was sufficient for a positive call. At a specificity of >98%, the addition of AluS analysis enhanced sensitivity of detection of cancers from 38% with aneuploidy alone and 36% with AluS alone, to 62% in combination (Supplementary Table 2). Similar to our results, liver (84%), stomach (74%), and colorectal (70%) cancers had the highest sensitivities.

Discussion

The results described above show that the evaluation of the representation of SINEs can significantly add to the power of aneuploidy to detect cancers. In RealSeqS data, the A-PLUS algorithm considerably enhanced sensitivity over that achieved for aneuploidy alone at matched specificities (Fig. 4). We discovered that part of the power of A-PLUS was derived from the global reduction in one Alu sub-family (AluS) (Fig. 6). Though a single feature (AluS-Rep from RealSeqS data) could be used as a stand-alone classifier (Fig. 6B), it was not as powerful as A-PLUS, which uses 95,116 AluS features and an additional 16,702 AluY and 9,373 AluJ features (Additional Data file 1). The under-representation of AluS in cfDNA from cancer patients was supported and extended by the evaluation of WGS data (Fig. 6C, D). This single feature (AluS-Rep from WGS data) could be used to increase the performance of a classifier based on WGS copy number analysis, without any additional wet bench experiments (Fig. 6D).

One of the strengths of our study was its independent cohort design. Cohort 1 was used for training, Cohort 2 to establish thresholds for scoring a sample as positive, and Cohort 3 used to evaluate sensitivity at a pre-determined specificity. And Cohort 4 (different experiments from the same individuals as Cohort 3) was used to assess reproducibility. Numerous problems with machine-learning algorithms that limit their application to other datasets have been highlighted in the literature [19]. Moreover, it is now generally recognized that cross-validation, though an effective approach when the number of samples is limiting, is not as reliable for predicting performance as a completely independent dataset [21]. It took several years (~6 years) for us to acquire the 7130 samples evaluated in this study, but we felt it was critical to have a sufficient number of samples to evaluate performance in independent datasets.

Another strength of our study was that one of its major new findings – the reduced representation of AluS elements in the cfDNA of cancer patients - could be independently confirmed using a completely independent experimental approach (WGS) performed in a variety of laboratories, on samples distinct from those processed in our laboratory (Fig. 6).

One of the weaknesses of our study is that all RealSeqS experiments were performed in our laboratory. We are confident that other, future samples evaluated in our laboratory, using identical methods for blood collection, sample storage, DNA purification, PCR-mediated amplification, and sequencing, will perform similarly based on the comparison between Cohorts 2 and 3. However, we cannot be confident that other laboratories that perform similar experiments will achieve the same performance. It is conceivable that small differences in any of the experimental procedures used could impact performance, and these can confound analysis.

Another weakness of our study is that A-PLUS is empirical. Alu element representation is unequivocally different in the cfDNA of cancer patients than in normal individuals, but we don't know why. The usual suspects are differences in chromatin structure or nucleases in neoplastic cells vs non-neoplastic cells. However, we are not sure that the observed differences in Alu element representation arise from the neoplastic cells themselves or other cells within or outside of tumors, such as WBCs[17 27].

Regardless of mechanism, our study shows that Alu element representations in general, and AluS subfamily elements in particular, are altered in the cfDNA of patients with many different cancer types. Future investigation of the mechanisms underlying their altered representation will be facilitated by their abundance in the genome and their similar secondary structures. At the practical level, it will be informative to determine whether Alu representation can add sensitivity to other features obtained through WGS data, such as fragment sizes, end motifs, or chromatin accessibility [28 29], as well as to assays of mutation or DNA methylation.

Materials and Methods

Patient Samples

This study was approved by the Institutional Review Boards for Human Research at Johns Hopkins Medical Institutes in compliance with the Health Insurance Portability and Accountability Act. All individuals participating in the study provided written informed consent. Plasma was purified from XXX healthy individuals and XXX patients with cancer using a BioChain Cell-free DNA Extraction Kit (Cat X K5011625). All patients were de-identified and patients are not known to anyone outside the research group. Demographics for the individuals in the study are included in Supplementary Table 1.

RealSeqS Experimental Protocol

A detailed experimental protocol for RealSeqS is listed in the Supporting Appendix of Douville et al. 2020. Briefly, PCR was performed in 25 uL reactions containing 7.25 uL of water, 0.125 uL of each primer, 12.5 uL of NEBNext Ultra II Q5 Master Mix (New England Biolabs cat # M0544S), and 5 uL of DNA. Eight independent reactions were performed in ~0.1 ng to 0.25 ng of DNA. A second round of PCR was then performed to add dual indexes to each PCR product prior to sequencing. The second round of PCR was performed in 25 uL reactions containing 7.25 uL of water, 0.125 uL of each primer, 12.5 uL of NEBNext Ultra II Q5 Master Mix (New England Biolabs cat # M0544S), and 5 uL of DNA containing 5% of the PCR product from the first round. Amplification products from the second round were purified with AMPure XP beads (Beckman cat # a63880), as per the manufacturer's instructions, prior to sequencing. As noted above, each sample was amplified in eight independent PCRs in the first round. Each of the eight independent PCRs was then re-amplified using index primers in the second PCR round. The sequencing reads from the 8 replicates were summed for the bioinformatic analysis but could also be assessed individually for quality control purposes. All oligonucleotides were purchased from IDT (Coralville, Iowa).

Sequence Analysis

Massively parallel sequencing was performed using a Hiseq4000. During the first round of PCR, degenerate bases at the 5' end of one of the primers were used as molecular barcodes (unique identifiers, UIDs) to uniquely label each DNA template molecule. This ensured that each DNA template molecule was counted only once, as described in (2). In all instances in this paper, the term "reads"

refers to uniquely identified reads (UIDs). Depending on the experiment, each read was sequenced on average 1.1 times. An average of XX million reads per sample (IQR XXXX M to XXX M) was assessed. If multiple reads had the same UID, we required at least 50% of the reads to map to the same genomic location. Reads with the same UID, but with discordant genomic locations were discarded from analysis. The alignment pipeline is available at ((https://zenodo.org/record/3656943).

Alu Profile Learning Using Sequencing Model Building

Alu Profile Learning Using Sequencing (A-PLUS) is a supervised machine learning approach to identify differences in normalized read depth for RealSeqS loci between non-cancer and cancer cell samples. To build A-PLUS, we employed the following steps:

1. Assemble a diverse and balanced training set of non-cancer and non-metastatic cancer samples. Alu SINEs are known to have ethic specific single nucleotide polymorphisms that could alignment and potentially alter the normalized read depth representation of various RealSeqS loci. We wanted to limit the potential for possible confounders to impact predictions. Our training set consisted of XXX previously published non-cancer and XX cancer samples (breast, colorectum, esophagus, lung, liver, pancreas, ovary, and stomach). We added XXX non-cancer samples not previously published to expand the ethnic representation of samples. The additional unpublished samples included XX Vietnamese, XX Han Chinese, XX South Asian, and XXX Native American/Inuit samples. Only samples with sufficient read depth were considered. A full table of sample demographics is included in XX.
2. Repeat RealSeqS for samples with sufficient remaining DNA. We performed RealSeqS on XX of the XXX total samples.
3. Normalize read depth for all training samples. To do this, divide a sample's autosomal loci by its total autosomal coverage and multiply by 10,000,000. This normalizing step allows all samples to be compared against each other regardless of differences in total coverage.
4. Perform Amplicon Selection.
    a. Remove loci with insufficient coverage. We only considered loci with an average of XX normalized reads in our training set. After applying this filter, only XXX loci remained.
    b. Perform the paired T-test for samples with a second available assay. Loci that were statistically significant ($p<0.05$) were discarded. After applying this filter, only XXX loci remained.
5. Perform principal component analysis (PCA) on the XXX loci. PCA was performed using the prcomp function in R version 3.4.
6. Generate a support vector machine (SVM) using 60 principal components from Step 5 as predictive features. The number of components (n=60) was based on ~10% of the total number of samples in the Training set (~600).

Detection of Aneuploidy

We have previously described our algorithm to detect the presence of aneuploidy in amplicon sequencing in Douville et al. 2018 and Douville et al. 2020. Our approach uses the normalized read counts of 500-kb intervals across the genome and performs a "within-sample" comparison. Outlier intervals and germline copy number variance are filtered. The remaining intervals are aggregated across the chromosome arm and a statistical significance is calculated. The 39 non-acrocentric chromosome arm statistical significances are then used as predictive features in a supervised machine learning model.

A support vector machine (SVM) [30] was trained on XXX normal putatively euploid non-cancer samples and XXX aneuploid cancer samples to discriminate between technical and biological noise in the general population and the presence of the most common aneuploidies in cancer. The model was built using R with the e1071 library [27]. The model generates the Global Aneuploidy Score (GAS) which is ranges from 0 to 1 with the higher the score the higher likelihood that the sample is aneuploid.

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, Bilerica NY) were performed following the manufacturers protocols and concentrations were determined using 5 parameter log curve fits (using Bioplex Manager 6.0) with vendor provided standards and quality controls. The HCCBP1MAG-58K panel was used to detect OPN, HGF, AFP, CA125, CA15-3, CEA, and CA19-9. The HTMP1MAG-54K panel was used to detect TIMP-1.

Multi-analyte Classifier

Using XXX non-cancer samples and XXX cancer samples (Block 2), we generated a multi-analyte classifier with A-PLUS, GAS, and 8 proteins (OPN, HGF, AFP, CA125, CA15-3, CEA, CA19-9, TIMP-1) as predictive features. To ensure no overlap with the samples used to generate GAS and A-PLUS models samples from Block 1 were not used in training. All 10 predictive features are elevated in cancer. Feature values <95th percentile in the non-cancer Block 2 samples were set to 0 to ensure the multi-analyte classifier would not base predictions from technical noise of batches and reduce possible overfitting. We used logistic regression from the XXX R library package. The feature coefficients are listed in Supplementary Table XX.

Data Availability.

Concise summaries of sequencing data and predictions are provided in Additional Files 3 and 4. The code and sample files used in the study are available at (https://zenodo.org/record/3656943).

1. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12**, 1–12 (2011).
2. Deininger, P. L. & Batzer, M. A. Alu Repeats and Human Disease. *Mol. Genet. Metab.* **67**, 183–193 (1999).
3. Pascarella, G. Non-allelic homologous recombination of Alu and LINE-1 elements generates somatic complexity in human genomes.
4. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **4**, 143–153 (2004).
5. Rodriguez, J. *et al.* Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. *Nucleic Acids Res.* **36**, 770–784 (2008).
6. Daskalos, A. *et al.* Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *Int. J. Cancer* **124**, 81–87 (2009).
7. Cho, N.-Y. *et al.* Hypermethylation of CpG island loci and hypomethylation of LINE-1 and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features. *J. Pathol.* **211**, 269–277 (2007).
8. Choi, I.-S. *et al.* Hypomethylation of LINE-1 and Alu in well-differentiated neuroendocrine tumors (pancreatic endocrine tumors and carcinoid tumors). *Mod. Pathol.* **20**, 802–810 (2007).
9. Richards, K. L. *et al.* Genome-Wide Hypomethylation in Head and Neck Cancer Is More Pronounced in HPV-Negative Tumors and Is Associated with Genomic Instability. *PLOS ONE* **4**, e4941 (2009).
10. Hunt, K. V. *et al.* scTEM-seq: Single-cell analysis of transposable element methylation to link global epigenetic heterogeneity with transcriptional programs. *Sci. Rep.* **12**, 5776 (2022).
11. Zhou, Q. *et al.* Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proc. Natl. Acad. Sci.* **119**, e2209852119 (2022).

12. Agostini, M. *et al.* Circulating Cell-Free DNA: A Promising Marker of Pathologic Tumor Response in Rectal Cancer Patients Receiving Preoperative Chemoradiotherapy. *Ann. Surg. Oncol.* **18**, 2461–2468 (2011).
13. Mead, R., Duku, M., Bhandari, P. & Cree, I. A. Circulating tumour markers can define patients with normal colons, benign polyps, and cancers. *Br. J. Cancer* **105**, 239–245 (2011).
14. Iqbal, S. *et al.* Circulating cell-free DNA and its integrity as a prognostic marker for breast cancer. *SpringerPlus* **4**, 265 (2015).
15. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
16. Douville, C. *et al.* Assessing aneuploidy with repetitive element sequencing. *Proc. Natl. Acad. Sci.* **117**, 4858–4863 (2020).
17. Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616 (2021).
18. Lui, Y. Y. *et al.* Predominant Hematopoietic Origin of Cell-free DNA in Plasma and Serum after Sex-mismatched Bone Marrow Transplantation. *Clin. Chem.* **48**, 421–427 (2002).
19. Moser, T., Kühberger, S., Lazzeri, I., Vlachos, G. & Heitzer, E. Bridging biological cfDNA features and machine learning approaches. *Trends Genet.* **0**, (2023).
20. Verleysen, M. & François, D. The Curse of Dimensionality in Data Mining and Time Series Prediction. in *Computational Intelligence and Bioinspired Systems* (eds. Cabestany, J., Prieto, A. & Sandoval, F.) 758–770 (Springer, 2005). doi:10.1007/11494669_93.
21. Wan, N. *et al.* Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer* **19**, 832 (2019).
22. Douville, C. *et al.* Detection of aneuploidy in patients with cancer through amplification of long interspersed nucleotide elements (LINEs). *Proc. Natl. Acad. Sci.* **115**, 1871–1876 (2018).

23. Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat. Commun.* **12**, 5060 (2021).
24. Jurka, J. & Smith, T. A fundamental division in the Alu family of repeated sequences. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 4775–4778 (1988).
25. Willard, C., Nguyen, H. T. & Schmid, C. W. Existence of at least three distinct Alu subfamilies. *J. Mol. Evol.* **26**, 180–186 (1987).
26. Zheng, H., Zhu, M. S. & Liu, Y. FinaleDB: a browser and database of cell-free DNA fragmentation patterns. *Bioinformatics* **37**, 2502–2503 (2021).
27. Sworder, B. J. *et al.* Determinants of resistance to engineered T cell therapies targeting CD19 in large B cell lymphomas. *Cancer Cell* **41**, 210-225.e5 (2023).
28. Mouliere, F. *et al.* Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. *EMBO Mol. Med.* **10**, e9323 (2018).
29. Ding, S. C. & Lo, Y. M. D. Cell-Free DNA Fragmentomics in Liquid Biopsy. *Diagnostics* **12**, 978 (2022).
30. Pisner, D. A. & Schnyer, D. M. Chapter 6 - Support vector machine. in *Machine Learning* (eds. Mechelli, A. & Vieira, S.) 101–121 (Academic Press, 2020). doi:10.1016/B978-0-12-815739-8.00006-7.

1. F. Mouliere *et al.*, Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med* **10**, (2018).
2. R. Saiki *et al.*, Combined landscape of single-nucleotide variants and copy number alterations in clonal hematopoiesis. *Nat Med* **27**, 1239-1249 (2021).