Published in final edited form as:

Sci Transl Med. 2024 January 24; 16(731): eadi3883. doi:10.1126/scitranslmed.adi3883.

Machine learning to detect the SINEs of cancer

Christopher Douville[1,2,3,4,5,*,†], **Kamel Lahouel**[6,7,8,9,†], **Albert Kuo**[2,4,5,9], **Haley Grant**[2,3,5,9], **Bracha Erlanger Avigdor**[2,3,4,5], **Samuel D. Curtis**[2,3,4,5], **Mahmoud Summers**[2,3,4,5], **Joshua D. Cohen**[2,3,4,5], **Yuxuan Wang**[2,3,4,5], **Austin Mattox**[2,3,4,5], **Jonathan Dudley**[2,3,4,5,10], **Lisa Dobbyn**[2,3,4,5], **Maria Popoli**[2,3,4,5], **Janine Ptak**[2,3,4,5,11], **Nadine Nehme**[2,3,4,5], **Natalie Silliman**[2,3,4,5,11], **Cherie Blair**[2,3,4,5,11], **Katharine Romans**[2,3,4,5], **Christopher Thoburn**[4], **Jennifer Gizzi**[4], **Robert E. Schoen**[12,13], **Jeanne Tie**[14,15,16], **Peter Gibbs**[14,15,17], **Lan T. Ho-Pham**[18,19], **Bich N. H. Tran**[20], **Thach S. Tran**[20,21], **Tuan V. Nguyen**[20,21,22,23,24], **Michael Goggins**[2,3,4,10,25], **Christopher L. Wolfgang**[26], **Tian-Li Wang**[10,27], **Ie-Ming Shih**[10,27], **Anne Marie Lennon**[2,4,5,25,28], **Ralph H. Hruban**[2,10], **Chetan Bettegowda**[2,3,4,5,29], **Kenneth W. Kinzler**[2,3,4,5], **Nickolas Papadopoulos**[2,3,4,5], **Bert Vogelstein**[2,3,4,5,11], **Cristian Tomasetti**[6,7,8,9,*]

[1]Division of Quantitative Sciences, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA.

[2]Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA.

[*]Corresponding author. cdouvil1@jhmi.edu (C.D.); ctomasetti@coh.org (C.T.).

[†]These authors contributed equally to this work.

Author contributions: C.D., K.L., N.P., K.W.K., B.V., and C. Tomasetti conceptualized the study. C.D., K.L., N.P., K.W.K., B.V., and C. Tomasetti designed the methodology. C.D., K.L., A.K., H.G., B.E.A., S.D.C., M.S., and C. Tomasetti performed the computer programming and algorithm development aspects. J.D.C., Y.W., A.M., J.D., L.D., M.P., J.P., N.N., and N.S. performed the bench experiments needed for investigation. C. Blair, K.R., C. Thoburn, J.G., R.E.S., J.T., P.G., L.T.H.-P., B.N.H.T., T.S.T., T.V.N., M.G., C.L.W., T.-L.W., I.-M.S., A.M.L., R.H.H., C. Bettegowda, K.W.K., N.P., and B.V. acquired samples needed for investigation. C.D., K.L., N.P., K.W.K., B.V., and C. Tomasetti performed the visualization aspects of the study. C.D. performed the project administration. C.D., K.L., N.P., K.W.K., B.V., and C. Tomasetti were responsible for supervision, writing the original draft, and reviews and edits.

Competing interests: Under a license agreement between Exact Sciences Corp. and the Johns Hopkins University, C. Tomasetti and the University are entitled to royalty distributions. C. Tomasetti has patent applications for intellectual property related to cancer early detection. C. Tomasetti is a member of the Scientific Advisory Board of PrognomiQ Inc. and an advisor for Haystack Oncology. C. Tomasetti is also a paid consultant for the Rising Tide Foundation and Bayer AG. B.V., K.W.K., 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. B.V., K.W.K., N.P., and C.D. hold equity in Exact Sciences. N.P. and C.D. are consultants to Thrive Earlier Detection. B.V., K.W.K., J.D.C., and N.P. are founders of and own equity in Haystack Oncology and ManaT Bio. K.W.K. and N.P. are consultants to Neophore. K.W.K., B.V., and N.P. hold equity in and are consultants to CAGE Pharma. B.V. is a consultant to and holds equity in Catalio Capital Management. C. Bettegowda is a consultant to Depuy-Synthes, Bionaut Labs, Haystack Oncology, and Galectin Therapeutics. C. Bettegowda is a cofounder of OrisDx. C.B. and C.D. are cofounders of Belay Diagnostics. 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. B.V., K.W.K., N.P., C. Bettegowda, 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 and 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. R.E.S. received research support from Immunovia, Freenome, and Exact Sciences. J.T. has a speaking, advisory, and consultancy with Haystack Oncology, Amgen, Novartis, AstraZeneca, Merck Serono, Merck Sharp & Dohme, Beigene, Pierre Fabre, Bristol Mysers Squibb, Gilead, and Daiichi Sankyo. The following patent applications are associated with this manuscript "Assessing and treating cancer" (63/538,719 Provisional) and "Rapid aneuploidy detection" (US20220259668A1 Pending).

Data and materials availability: RealSeqS data generated in this study can be requested upon approval at the European Genome Archive (EGA) (https://ega-archive.org/studies/EGAS00001007169). All predictions and feature values are available in tables S2 and S3 (tables S1 to S8 are available in data file S1). The code is available at https://zenodo.org/record/8225868.

[3]Ludwig Center, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA.

[4]Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA.

[5]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

[6]Center for Cancer Prevention and Early Detection, City of Hope, Duarte, CA 91010, USA.

[7]Center for Cancer Prevention and Early Detection, City of Hope, Division of Mathematics for Cancer Evolution and Early Detection, Department of Computational and Quantitative Medicine, Beckman Research Institute, City of Hope, Duarte, CA 91010, USA.

[8]Division of Integrated Cancer Genomics, Translational Genomics Research Institute, Phoenix, AZ 85004, USA.

[9]Department of Biostatistics, Johns Hopkins University School of Public Health, Baltimore, MD 21205, USA.

[10]Department of Pathology, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA.

[11]Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA.

[12]Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA.

[13]Department of Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA.

[14]Division of Personalized Oncology, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia.

[15]Department of Medical Oncology, Melbourne, VIC 3000, Australia.

[16]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC 3011, Australia.

[17]Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, VIC 3010, Australia.

[18]BioMedical Research Center, Pham Ngoc Thach University of Medicine, Ho Chi Minh City 72510, Vietnam.

[19]Clinical Genetics Research Group, Saigon Precision Medicine Research Center, Ho Chi Minh City 72512, Vietnam.

[20]Saigon Precision Medicine Research Center, Ho Chi Minh City 72512, Vietnam.

[21]School of Biomedical Engineering, University of Technology Sydney, NSW 2007, Australia.

[22]Tâm Anh Research Institute, Ho Chi Minh City, Vietnam.

[23]Centre for Health Technologies, University of Technology, NSW 2007, Australia.

[24]School of Population Health, University of New South Wales, NSW 2003, Australia.

[25]Department of Medicine, Johns Hopkins Medical Institutes, 733 N. Broadway, Baltimore, MD 21205, USA.

[26]Department of Surgery, NYU Langone, New York City, NY 11209, USA.

[27]Department of Gynecology and Obstetrics, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA.

[28]Department of Surgery, Johns Hopkins Medical Institutes, 733 N. Broadway, Baltimore, MD 21205, USA.

[29]Department of Neurosurgery, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA.

Abstract

We previously described an approach called RealSeqS to evaluate aneuploidy in plasma cell-free DNA 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 machine learning approach called Alu Profile Learning Using Sequencing (A-PLUS) and its application to 7615 samples from 5178 individuals, 2073 with solid cancer and the remainder without cancer. Samples from patients with cancer and controls were prespecified into four cohorts used for model training, analyte integration, and threshold determination, validation, and 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 eight common protein biomarkers detected 51% of the 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 subfamily elements in the circulating DNA of patients with solid cancer. We confirmed this reduction through the analysis of another independent dataset obtained with a different approach (whole-genome sequencing). The evaluation of Alu elements may therefore have the potential to enhance the performance of several methods designed for the earlier detection of cancer.

INTRODUCTION

Alu elements are short interspersed nuclear elements (SINEs) of ~300 base pairs, with more than 1 million copies spread throughout the human 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). Moreover, Alu elements are hypomethylated early during tumor progression (3–9), and this feature has been incorporated into methods for the earlier detection of cancer through plasma cell-free DNA (cfDNA) analysis (10). Alu elements also reflect the altered fragmentation patterns found in the cfDNA of patients with cancer: One of the first plasma multicancer biomarkers used quantitative polymerase chain reaction (qPCR) to calculate the ratio of short and long Alu segments (11–13).

Whole-genome sequencing (WGS) has been widely used in recent blood-based multicancer earlier detection assays (14–16). 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 (17). Even with the inclusion of mappable Alu elements, shallow WGS is inefficient to optimally evaluate Alu elements because they represent only a small fraction (~11%) of the genome (1).

We previously developed an approach, called RealSeqS, to specifically amplify Alu sequences (18). RealSeqS offers advantages over WGS, including a simpler workflow that does not require library construction and has 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 is 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 [reviewed in (19)]. We therefore hypothesized that the representation of specific Alu elements might be different in the cfDNA of plasma 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 (ML) tools. Here, we report and test a ML-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.

RESULTS

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 alterations, which was the original intent of RealSeqS, or regions of high technical variability, in the noncancer controls. We hypothesized that an unbiased supervised ML 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 classification.

The detection of cancer in asymptomatic patients, which is the primary goal of multicancer earlier detection tests, requires very high specificity. Designing a highly specific ML 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 patients with cancer. 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 nonneoplastic cells of the same organ surrounding the cancer cells that have been destroyed because of the cancer environment, or from one or more types of leukocytes. Note that leukocytes are the major source of cfDNA in patients with or without cancer (20).

Other challenges facing the development of a highly specific ML cancer detection 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 on attributes of cancer per se (14, 21–24). 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) (25). 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.

Cross-validation is frequently used to assess classifier performance but can generate overly optimistic estimates when the ML model accidentally learns a confounding variable that “leaks” between samples in the training and testing folds (14, 26, 27). Adjustments to traditional *k*-fold cross-validation, including nested, *k*-batch, ordered *k*-batch, or balanced *k*-batch cross-validation, can eliminate knowledge leak (14, 28), but with high-dimensional data such as ours, feature selection is one of the keys to success. With cross-validation, each iteration yields a different set of features. In practice, a diagnostic test must use a prespecified threshold to determine whether the sample is positive. Selecting a threshold becomes even more challenging when optimizing for the high-specificity requirements needed in multicancer early detection. Nested cross-validation is one potential solution where half of the folds are used for model construction and half for threshold determination, but this still will not eliminate the potential for knowledge leak across folds. Even a subtle drift in model prediction values near the threshold could result in a much higher-than-expected number of positive tests when applied to large numbers of individuals in an asymptomatic cohort.

To address the challenges listed above, we incorporated several principles into the development of A-PLUS. 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 components 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 prespecified and nonoverlapping cohorts to minimize overfitting: 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 (Fig. 1).

The prespecified study design is relatively simple, is easily understood, and offers many advantages evaluating a high-dimensional ML model but is not without drawbacks. The primary concern is that any approach designed on a limited number of samples will be underpowered compared with a model using our entire dataset. To date, many of the proposed multicancer early detection models rely on thousands of predictive features but have only been trained on a few hundred to a thousand samples (14, 15, 29). Given how widespread this issue is, we asked not what the most powerful model is but whether we can structure our study to evaluate and generate realistic performance estimates for a complex ML model with limited training examples. Technical nuances underlying the development of A-PLUS are detailed in Materials and Methods, and the code is publicly available at https://zenodo.org/record/8225868. 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 566 individuals (715 total samples) from 354 individuals without cancer (400 total samples) and 202 patients with solid cancers (315 total samples). Individuals with multiple samples are always from the same time point and considered technical replicates. Cohort 1 has 459 samples previously analyzed for aneuploidy (18) using RealSeqS (18) and an additional 250 samples from controls of Vietnamese, Han Chinese, South Asian, or Native American/Inuit ethnicities not included in previous studies. These additional samples were included because inherited polymorphisms within Alu elements could alter alignments and the subsequent read depth representation of Alu loci. Slightly fewer 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 table S1 in data file S1.

Important elements of the training included normalization of read depths and the removal of amplicons with insufficient coverage and removal of amplicons that were unstable based on *t* tests. After these steps, there were 121,197 loci of the original 350,000 that remained. PCA was then used to reduce dimensionality. Last, a support vector machine (SVM) was used to identify the 60 first 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 d>>n conundrum (25). The noncancer samples were used to generate a euploid reference panel for aneuploidy calls. Both the cancers and noncancer samples were used to generate and optimize model building. Performance was not assessed in cohort 1.

To reduce the chance of overfitting and establish quality metrics before evaluation of samples, we used the previously published metrics and thresholds for inclusion of patients rather than exclude any participants because of any metric related to A-PLUS performance. Our previously introduced quality control metrics were designed to identify technical outliers (18). The first metric identified samples with unusual GC bias. Next, we used two different size metrics to identify the presence of large molecular weight DNA. Large molecular weight DNA is frequently attributed to genomic DNA from blood cells lysed during collection or storage rather than from cfDNA. Contamination from genomic DNA can limit the sensitivity of an assay given that it is unlikely to be derived from the tumor. These previously defined criteria excluded 3.1% of cohort 1 samples, and their global aneuploidy scores (GAS) and A-PLUS scores are listed in table S2 for transparency. We excluded a small fraction of patients in cohorts 2, 3, and 4 using identical criteria (tables S2 and S3).

Cohort 2: Analyte integration and determination of thresholds

Cohort 2 included samples from 704 patients with solid tumors (total of 852 samples) and 958 control individuals without cancer (total of 1402 samples). None had metastases at the time of the blood collection, and as with cohort 1, 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 in cohort 2 (fig. S1A). At this threshold, samples from patients with cancers of the esophagus and stomach had the highest sensitivities [86% confidence interval (CI): 68 to 96% and 87% CI: 73 to 94%, respectively], and samples from patients with breast cancer had the lowest (34% CI: 28 to 40%; Fig. 2A and tables S3 and S4).

We also investigated whether the number of principal components would affect performance. An A-PLUS model using the first 60 principal components outperformed models with 15, 30, 90, and 240 principal components. Across cohort 2, 51.5% of cancers were detected using a model built with 60 principal components, whereas 36.5% of cancers were detected with a model built on 15, 46.0% with 30, 49.9% with 90, and 49.2% with 240 based on a threshold that produces >99% specificity. We also tested whether a random forest model would outperform the SVM model we used, with all other parameters identical. The SVM outperformed the random forest model (51.5 versus 39.8% at 99% specificity). A full list of the principal components is provided in table S2.

We then generated GAS scores for cohort 2 samples with the same RealSeqS data used to generate A-PLUS scores. The GAS uses a different ML technique to generate a single score that reflects gains or losses of 39 chromosome arms, focusing on those arms that are typically altered in cancers (30). A GAS threshold of >0.64 yielded a 99% specificity in the cohort 2 control samples (fig. S1B). The highest sensitivities at this 99% specificity were achieved for cancers of the esophagus and liver (43% CI: 26 to 62% and 37% CI: 17 to 61%, respectively) and the lowest in breast (6% CI: 4 to 10%; Fig. 2, A and B, and tables S3 and S4). In the 687 cancers scoring negatively (that is, below the 99% specificity threshold) in the GAS assay, 318 (46%) scored positively (that is, above the 99% specificity threshold) in the A-PLUS assay (Fig. 2C and tables S3 and S4). Conversely, 81% of the cancer samples that scored positively in GAS also scored positively in A-PLUS. No A-PLUS–positive noncancer samples scored positively in GAS (Fig. 2D).

Next, we evaluated a panel of eight protein markers previously shown to be useful for cancer detection when used at high thresholds (OPN, HGF, AFP, CA125, CA15–3, CEA, CA19–9, and TIMP-1). To integrate these eight protein values into a single score, logistic regression with binary classification was used to generate a protein score. Performance was assessed using 10-fold cross-validation and a protein score ("PROT") of >0.73 generated 99% specificity in the samples from patients without cancer (fig. S1C). The highest sensitivities at this threshold were achieved for cancers of the liver and stomach (84% CI: 60 to 96% and 69% CI: 54 to 81%, respectively) and the lowest in breast (18% CI: 14 to 23%; tables S3 and S4). In the 535 cancers scoring negatively with the protein assay, 44% scored positively with A-PLUS.

Logistic regression was then used to integrate A-PLUS and GAS with the proteins into a multi-analyte classifier (table S5). Performance was assessed using 10-fold cross-validation and a threshold of >0.87 generated 99% specificity (fig. S1D). The highest sensitivities were achieved for patients with cancers of the esophagus and liver (90% CI: 72 to 97% and 90% CI: 66 to 98%, respectively) and the lowest in patients with breast cancers (32%; tables S3 and S4). The sensitivity of the multi-analyte classifier was equivalent or higher than any individual analyte in every cancer type and maintained high specificity.

Cohort 3: Independent validation

Cohort 3 samples were from 2960 individuals, including 1167 patients with solid tumors of 11 types: breast, colorectum, esophagus, head and neck, kidney, lung, ovary, pancreas, prostate, stomach, and uterus (Table 1 and table S1). None of these samples had been previously analyzed with RealSeqS in prior publications.

The 99% thresholds defined by cohort 2 were used to assess performance in cohort 3 for each of the assays described above (table S3). For A-PLUS, the specificity observed in cohort 3 (98.5%) was slightly less than the 99% expected from cohort 2 ($P = 0.03$, one-sided two-sample Z test for equality of proportions). In the seven cancer types that were evaluated in both cohorts 2 and 3, the cancer-type sensitivities were similar, with the exception of lung, where the sensitivities were 54 and 27%, respectively (Fig. 3A and tables S3 and S4). This decrease in cohort 3 sensitivity for lung cancers could not be attributed to differences in cancer subtype between the two cohorts (table S1). The sensitivities and specificities of aneuploidy alone as well as proteins alone were similar in cohorts 2 and 3 (tables S3 and S4). The multi-analyte test incorporating A-PLUS, aneuploidy, and proteins detected more than two-thirds of cancers for each of the following organs: esophagus, pancreas, ovary, stomach, and colorectum, at an observed specificity of 98.9% (Fig. 3B and tables S3 and S4). The sensitivities of the two other cancer types evaluated in both cohorts 2 and 3 were lower (37% CI: 29 to 46% and 26% CI: 16 to 38% in lung and breast, respectively). Four cancer types (head and neck, kidney, prostate, and uterus) were represented in cohort 3 but not cohort 1 or 2. Although no algorithmic training was performed on these cancer types, the sensitivity for their detection ranged from 17 to 36% (Fig. 3C). Cancers of the liver were not available in cohort 3, although the sensitivity for their detection in cohort 2 was high (90%).

The overlap in analyte detections (A-PLUS, GAS, and PROT) at the predefined thresholds is depicted as a Euler diagram in Fig. 3 (D and E). A-PLUS made a greater contribution to positive calls than aneuploidy or proteins, and A-PLUS detected 41% of the samples that were not detected by either aneuploidy or proteins.

Most cancers represented in cohort 3, and all the samples in cohorts 1 and 2, were relatively early in the sense that few had any distant metastatic lesions evident at the time of sample acquisition. However, there were 15 cases (0.6% of 2349), all ovarian and all in cohort 3, that were later restaged to IV. When these were removed, the sensitivity for ovarian cancer detection decreased from 75 to 73%. The sensitivity was 72% for ovarian in cohort 2, which did not contain any stage IV cancers. When categorized by stage I, II, or III, the average sensitivities for all cancers using each individual analyte or the multi-analyte classifier increased with stage in both cohorts 2 and 3 (table S4). With the multi-analyte classifier, for example, sensitivity averaged 57% in stage I, 60% in stage II, and 65% in stage III.

Cohort 4: Reproducibility

The technical reproducibility of the A-PLUS and GAS assays (both based on RealSeqS sequencing data) was evaluated in 1686 individuals (1539 without cancer and 147 with cancer) from cohort 2 or cohort 3. The cancers were predominantly breast [*n* = 98 (67%)], colorectum [*n* = 17 (12%)], and lung [*n* = 11 (7%)], with the remainder [*n* = 21 (14%)] from various other solid tumors. The samples were all technical repeats taken from the same patient at the same time point. When an individual had three or more technical replicates available, only the first two replicates were considered; all replicate scores are provided in tables S2 and S3. Separate aliquots of purified DNA (comprising separate template molecules) from the same plasma sample were independently amplified using the single RealSeqS primer pair, and the PCR products were sequenced. The amplification and sequencing of each DNA sample from the same patient were performed on different days. Of the 1686 pairs, quality control failed for at least one of the two samples in 54 (3.2%) of the cases.

Using the thresholds defined by cohort 2, 95.8% of the 1632 pairs scored concordantly (either positively or negatively) with a Cohen’s kappa of 0.56 (95% CI: 0.47 to 0.65) for A-PLUS (Fig. 4 and table S6). The imperfect concordance of A-PLUS reflects the balance between specificity and sensitivity. The specificity was purposefully set to be very high (99%), realizing that this could limit sensitivity in light of experimental variation between different assays on the same samples. Thus, the difference between specificities of the replicates, at the same preset threshold, was only 1.2%, whereas the difference between sensitivities of the replicates was more than 10× higher (18%) (Fig. 4 and table S6). Of the 70 discordant samples, 23 scored just below the threshold in one of the two DNA aliquots (that is, between the scores required for 98 and 99% specificity) and above the score required for 99% specificity in the other aliquot.

With GAS, 99.3% pairs were concordant (Fig. 4B) with a Cohen’s kappa of 0.66 (CI: 0.48 to 0.85). The lower sensitivity of GAS in cohort 4 compared with cohort 3 was due to the fact that most (67%) of the cancer cases in cohort 4 were derived from patients with breast cancers, whereas only 6% of the cancers in cohort 3 were derived from the breast. Breast cancers had the lowest GAS in cohort 3.

Alu subsets

We next asked whether there was a subset of the Alu loci included in the A-PLUS heuristic that was particularly important for its success in distinguishing samples from individuals with or without cancer. The most notable observation was a global reduction in read depth across all AluS loci. The AluS subfamily is the largest subfamily and older than AluJ but younger than AluY (31, 32). Using only a single feature—the average normalized read depth of AluS elements (herein dubbed AluS fraction) from RealSeqS data, without any ML algorithms at all—samples from patients with cancer could be distinguished from those of controls [$P < 2 \times 10^{-176}$, one-sided t test; area under the curve (AUC) = 0.70; Fig. 5A]. A-PLUS scores and the proportional representation of AluS elements were inversely correlated (−0.19; $P < 2.2 \times 10^{-16}$ via Pearson's). We noted that the third principal component from the A-PLUS model was associated with the global change in read depth across AluS, with a Pearson correlation coefficient of 0.72. Given that the AluJ fraction, AluS fraction, and AluY fraction add to 1, a change in AluS naturally produces a change in the other fractions. The decrease in AluS fraction for cancer samples is associated with increases in the AluJ fraction ($P < 4 \times 10^{-38}$, one-sided t test; AUC = 0.46; fig. S2A) and the AluY fraction ($P < 5 \times 10^{-100}$, one-sided t test; AUC = 0.66; fig. S2B) fraction. Statistically significant differences in AluS fractions were observed in all of the cancer types evaluated [$P < 2 \times 10^{-16}$, analysis of variance (ANOVA)].

We then asked whether this observation could have been affected by the filtering used for A-PLUS feature selection. We repeated this analysis across all Alu elements in RealSeqS rather than only those used to build the model. Again, a decrease in the AluS fraction was observed in the cancer samples ($P < 3 \times 10^{-98}$, one-sided t test). Analogously, increases in the AluJ and AluY fractions ($P < 7 \times 10^{-47}$ and $P < 0.0002$, one-sided t tests) were observed in the cancer samples. A full list of the unfiltered fractions is included in table S7.

We wondered whether the reduced representation of AluS elements in the cfDNA from patients with cancer was the result of some unknown bias in amplification efficiency or sequencing generated from the RealSeqS approach, which used a single primer pair. To address this question, WGS data from our recently published study that included 65 cfDNA samples from individuals without cancer and 58 samples from patients with cancer were evaluated (55 colorectum, 2 lung, and 1 pancreas) (table S8) (33). Cancer samples had a global reduction of AluS representation compared with controls ($P < 2.6 \times 10^{-5}$, one-sided t test; Fig. 5B). Moreover, this single feature (AluS fraction, defined as the number of AluS reads divided by the number of all Alu reads) could distinguish samples from patients with cancer and controls with an AUC of 0.73 in receiver operating characteristic (ROC) analysis. The AluS fraction value corresponding to >99% specificity among the control samples was 0.6546. At this threshold, 36% of the samples from patients with cancers were positive. The increased representation of AluJ ($P > 0.05$, one-sided t test) and AluY ($P < 1 \times 10^{-5}$) was also observed in WGS data from patients with cancer (fig. S2, C and D).

Next, we evaluated a publicly available dataset that included 266 cfDNA samples from patients with cancer (25 bile duct, 54 breast, 22 colorectum, 27 stomach, 76 lung, 26 ovary, and 35 pancreatic) and 260 samples from individuals without cancer (34). Again, cancer samples had a global reduction of AluS ($P < 4 \times 10^{-43}$, one-sided t test; Fig. 5C). Using the preselected threshold for the AluS fraction, 21% of cancers were detected at a specificity of 98.8% (Table 2 and table S8). There was also a statistically significant difference among cancer types ($P < 0.0006$, ANOVA). An increased representation of AluJ ($P < 3 \times 10^{-16}$, one-sided *t* test; fig. S2E) and AluY ($P < 0.02$, one-sided *t* test; fig. S2F) was also observed in the publicly available WGS dataset.

Last, we evaluated whether AluS fraction in WGS data could be combined with the evaluation of aneuploidy in the same dataset. WisecondorX was used to count the number of aberrant chromosomal regions throughout the genome in the FinaleDB dataset. Aneuploidy was defined by the WisecondorX default settings. The scores for aneuploidy and AluS fraction were combined with a Boolean OR to avoid the possible pitfalls of ML algorithms mentioned above. The addition of AluS analysis enhanced the sensitivity of detection of cancers from 42% with aneuploidy alone and 21% with AluS alone to 52% in combination, with an aggregate specificity of 98.5% (Table 2). Similar to the A-PLUS results, patients with cancers of the stomach (67%) and colorectum (65%) had the highest sensitivities.

DISCUSSION

The results described above show that the evaluation of the representation of SINEs can 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. We found that part of the power of A-PLUS was derived from the global reduction in one Alu subfamily (AluS). Although a single feature (AluS fraction from RealSeqS data) could be used as a standalone classifier, it was not as powerful as A-PLUS, which uses 95,116 AluS features and an additional 16,702 AluY and 9373 AluJ features. The underrepresentation of AluS in cfDNA from patients with cancer was supported and extended by the evaluation of WGS data. This single feature (AluS fraction 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.

One of the strengths of our study was its independent cohort design. Cohort 1 was used for training, cohort 2 was used to establish thresholds for scoring a sample as positive, and cohort 3 was used to evaluate sensitivity at a predetermined specificity. Cohort 4 (comprising different experiments from the same individuals as cohort 3) was used to assess reproducibility. Numerous problems with ML algorithms that limit their application to other datasets have been highlighted in the literature (24). Moreover, it is now generally recognized that cross-validation, although an effective approach when the number of samples is limiting, is not as reliable for predicting performance as a completely independent dataset (14, 26, 27). Despite its advantages, this study design does have drawbacks. Models constructed on rigid prestructured smaller cohorts will inherently have less statistical power than if we had used the entire dataset and assessed performance using nested cross-validation. Another strength of our study was that one of its major findings—the reduced representation of AluS elements in the cfDNA of patients with cancer—could be confirmed using a completely independent experimental approach (WGS) on samples distinct from those processed in our laboratory.

One limitation of our study is that we only evaluated solid tumors. We have yet to evaluate whether our findings extend to hematologic malignancies or other types of solid tumors that are less common than those studied here. Another weakness is that all RealSeqS experiments were performed in our laboratory. We are confident that 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 will achieve the same performance. It is conceivable that small differences in any of the experimental procedures used could affect performance, and these can confound analysis. Another limitation of our study is that A-PLUS is empirical. Alu element representation is unequivocally different in the cfDNA of patients with cancer than in normal individuals, but we do not know why. The usual suspects are differences in chromatin structure or nucleases in neoplastic cells versus nonneoplastic cells. However, we are not sure that the observed differences in Alu element representation arise from the neoplastic cells themselves or from other cells within or outside of tumors, such as white blood cells (19, 35).

Regardless of the molecular basis for the observed differences, 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 sequences and 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 (36, 37), as well as to assays of mutation or DNA methylation.

MATERIALS AND METHODS

Study design

This study was designed to assess the potential of Alu elements to enhance methods designed for the earlier detection of cancer. 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 informed consent. Plasma was purified from 3105 individuals without cancer and 2037 patients with solid cancer 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 in the study are included in table S1. Samples from patients with cancer and cancer-free controls were prespecified into four cohorts used for model training, analyte integration and threshold determination, validation, and reproducibility.

RealSeqS experimental protocol

A detailed experimental protocol for RealSeqS is listed in the supporting appendix of Douville *et al.* (18). Briefly, PCR was performed in 25 µl of reactions containing 7.25 µl of water, 0.125 µl of each primer, 12.5 µl of NEBNext Ultra II Q5 Master Mix (New England Biolabs, catalog no. M0544S), and 5 µl of DNA. Eight independent reactions were performed in ~0.1 to 0.25 ng of DNA. A second round of PCR was then performed to add dual indexes to each PCR product before sequencing. The second round of PCR was performed in 25-µl reactions containing 7.25 µl of water, 0.125 µl of each primer, 12.5 µl of NEBNext Ultra II Q5 Master Mix (New England Biolabs, catalog no. M0544S), and 5 µl 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, catalog no. a63880), as per the manufacturer's instructions, before sequencing. As noted above, each sample was amplified in eight independent PCRs in the first round. Each of the eight independent PCRs was then reamplified using index primers in the second PCR round. The sequencing reads from the eight replicates were summed for the bioinformatic analysis but could also be assessed individually for quality control purposes. The forward (cgacgtaaaacgacggccagtNNNNNNNNNNNNNNNNGGTGAAACCCCGTCTCTACA) and reverse (cacacaggaaacagctatgaccatgCCTCCTAAGTAGCTGGGACTACAG) RealSeqS primers were generated at Integrated DNA Technologies.

Sequence analysis

Massively parallel sequencing was performed using 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 (38). In all instances in this paper, the term "reads" refers to UIDs. Depending on the experiment, each read was sequenced on average 1.1 times. An average of 10.4 million reads per sample (interquartile range, 7.9 to 12.7 million) were 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. Only samples with sufficient read depth (>2.5 million) were included.

A-PLUS model building

A-PLUS is a supervised ML approach to identify differences in normalized read depth for RealSeqS loci between noncancer and cancer cell samples. To build A-PLUS, we used the following steps:

1. Assemble a diverse and balanced training set of noncancer and nonmetastatic cancer samples. Alu SINEs are known to have ethnic-specific single-nucleotide polymorphisms that could affect alignment and potentially alter the normalized read depth representation of various RealSeqS loci. We wanted to limit the potential for possible confounders to affect predictions. Our training set consisted of 134 previously published noncancer and 198 solid tumor samples (breast, colorectum, esophagus, lung, liver, pancreas, ovary, and stomach) from stages I, II, and III (table S1). For the noncancers with self-reported ancestry

information, the cohort consisted of 41 African/Black, 24 Hispanic, 33 white/Caucasian, and 27 multiple ancestries. For the cancers, the self-reported ancestry consisted of 65 Asian, 3 African/Black, and 120 Caucasian/white. We added 264 noncancer samples and 13 cancer samples not previously published to diversify the ancestry representation of samples. The additional unpublished samples with ancestry information included 20 Vietnamese, 2 Pacific Islander, 59 Black/African, 35 Asian, 3 Native American/Inuit samples, and 136 Caucasian/white. The unpublished cancers with ancestry information included 2 Asian and 10 Caucasian/white samples.

2. The 99 patients with cancer were selected for the training set because each had a second technical replicate.

3. All samples were normalized for total read depth, that is, the sample's Alu loci were divided by its total autosomal coverage.

4. Perform amplicon feature selection.

   a. Remove sex chromosome loci. This removed 28,399 loci on chrX (AluJ = 5090; AluS = 18,111; AluY = 2905; other loci that are not Alu SINEs = 2293). This removed 2406 loci on chrY (AluJ = 311; AluS = 1495; AluY = 457; and other loci that are not Alu SINEs = 143).

   b. Remove loci with insufficient coverage. We only considered loci with an average of one normalized read in our training set and removed 445,711 loci (AluJ loci = 93,030; AluS loci = 242,659; AluY loci = 33,803; other loci that are not Alu SINEs = 76,219). After applying this filter, only 299,473 loci remained. Note that in theory RealSeqS could amplify 745,184 repetitive elements, but typically only 350,000 Alu SINEs were observed.

   c. Perform a paired *t* test for the 99 patients with cancer with a technical replicate. Loci that were statistically significant ($P < 0.01$) were discarded (AluJ loci = 16,377; AluS loci = 109,728; AluY loci = 27,389; other loci that are not Alu SINEs = 5). After applying these filters, only 145,974 loci of the original 745,184 loci remained. We acknowledge that a statistically significant difference in paired samples could still be informative if the effect size were greater than the technical variability. This filtering step was a conservative choice and likely removed informative loci. Given the large number of loci, we prioritized stability over power.

   d. Perform the Kolmogorov-Smirnov test on samples from different ancestries in the samples from individuals without cancer. Loci that were statistically significant ($P < 0.01$) were discarded (AluJ loci = 1559; AluS loci = 18,286; AluY loci = 4932; other loci that are not Alu SINEs = 0). After applying this filter, 121,197 loci remained. The remaining loci consisted of 9373 AluJ loci; 95,116 AluS loci; 16,702 AluY loci; and 6 loci that were not Alu SINEs. A full list of A-PLUS loci is listed in data file S2. A list of filtered loci is also provided. Genomic coordinates are in hg19.

5. Perform PCA on the 121,197 loci. PCA was performed using the prcomp function in R version 3.4.

6. Generate an SVM using the first 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). We used the e1071 (v1.7–9) library with a radial basis kernel, cost = 1, and nu = 0.5.

Detection of aneuploidy

An algorithm to detect the presence of aneuploidy from amplicon sequencing data has been described in our previous publications (18, 30). This approach uses the normalized read counts of 500-kb intervals across the genome and performs a “within-sample” comparison. Outlier intervals were bioinformatically identified and filtered. The remaining intervals were aggregated across the chromosome arm, and a circular binary algorithm is performed (39). Any aberrant segment <5 MB was assumed to be a germline copy number variation and bioinformatically removed. The statistical significance was then calculated for the chromosome arm without the aberrant <5-MB segments. The 39 non-acrocentric chromosome arm statistical significances were then used as predictive features in a supervised ML model. An SVM (40) was trained on 363 presumably euploid samples from individuals without cancer, from 128 samples of patients with cancer, and from 648 in silico–generated aneuploid samples derived from the samples of individuals without cancer. The model was built using R v3.4 with the e1071 library (41) and generates a GAS that ranges from 0 to 1. Several studies have evaluated copy number alterations in clonal hematopoiesis of indeterminate potential (CHIP) in large population databases (42–44). We initially did not explicitly account for the possibility of CHIP, but many of the candidate regions are <5 MB and would already be excluded using our previously described bioinformatic filters. On the basis of these CHIP studies, we instituted one additional bioinformatic filter to limit the possibility that a positive GAS was derived from CHIP. CHIP samples most frequently exhibit only one aneuploid arm and are limited to a small number of candidate arms (34). Any aneuploid sample (positive GAS) with only one aneuploid arm from chr 5q, 13q, 14q, and 20q was therefore bioinformatically labeled as CHIP and the GAS set to 0 (table S3).

Evaluation of plasma proteins

The Bioplex 200 platform (Bio-Rad) was used to determine the concentration of eight proteins in the plasma samples. Luminex bead– based immunoassays (Millipore) were performed following the manufacturer’s protocols, and concentrations were determined using five-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 1402 noncancer samples and 852 cancer samples (cohort 2), a multi-analyte classifier using A-PLUS scores, GAS, and the eight proteins described above was generated. The protein values were supplemented with 126 negative controls from Luminex bead–based immunoassays (table S3). Some samples were not assessed for proteins. To account for missing feature values in these samples, we randomly sampled and assigned values from our prior studies of individuals without cancers to these samples. The missing values are labeled in table S3. Feature values <98th percentile in the noncancer cohort 2 samples were set to 0 to reduce possible overfitting to features with low scores. We used logistic regression in R with the feature coefficients listed in table S5. Cross-validation was performed by randomly partitioning the dataset into 10 equally sized folds. We used the same seed to preserve the random partitions. Then, a model was generated on nine folds and tested on the one fold that was withheld. This protocol was repeated for the remaining nine folds. At the conclusion of cross-validation, each sample was withheld once, and the scores on the withheld fold were recorded.

External validation using WGS data

Samples from (15) were downloaded from FinaleDB on 19 February 2023. These included 266 samples from patients with cancers of the bile duct, breast, colorectum, stomach, lung, ovary, and pancreas and 260 samples from individuals without cancer. Molecules that mapped to autosomal Alu elements were retrieved using overlapSelect from the kent src tools package using a threshold = 1 (the entire read had to fall within the Alu element). The AluS fraction was calculated as the fraction of AluS molecules divided by the total molecules in Alu elements. We applied WisecondorX using default parameters and only considered autosomal chromosomes for the evaluation of aneuploidy in these samples (45). Default thresholds ($z > 5$ or $z < -5$) were used, and >3 autosomal copy number alterations were used as a threshold for positivity.

Statistical analysis

The two-sample Z test for equality of proportions was used to determine whether the sensitivity and specificity of cohort 2 versus cohort 3 were different. The two-sample Z test for equality of proportions assumes sufficient sample size in both populations, and this was assessed and satisfied before being used. Differences were considered statistically significant when $P < 0.05$. Cohen's kappa statistic was used to assess whether concordance between technical replicates was different from random chance. Error bars represent confidence intervals as calculated by the Wilson score interval (46, 47). Pearson's correlation coefficient was used to compare A-PLUS scores to the AluS fraction. Cohort sample size was not selected for statistical power but on the basis of sample availability, and no samples previously reported were used in the validation set (cohort 3).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding:

This study was supported by NIH grant U01CA271884 (C. Tomasetti), NIH grant R21NS113016 (C. Bettegowda), NIH grant RA37CA230400 (C. Bettegowda), NIH grant U01CA230691 (N.P. and C. Bettegowda), Oncology Core CA 06973 (B.V., K.W.K., and N.P.), the Virginia and D.K. Ludwig Fund for Cancer Research (B.V., K.W.K., N.P., C. Bettegowda, and C.D.), the John Templeton Foundation 62818 (C. Tomasetti), Commonwealth Fund (N.P.), Thomas M. Hohman Memorial Cancer Research Fund (C. Bettegowda), Alex's Lemonade Stand Foundation (C. Bettegowda), the Sol Goldman Sequencing Facility at Johns Hopkins (B.V.), Conrad R. Hilton Foundation (N.P., B.V., and K.W.K.), Benjamin Baker Endowment 80049589 (C.D. and Y.W.), NIH grant Ovarian Cancer SPORE DRP 80057309 (C.D.), Swim Across America (C.D.), Burroughs Wellcome Career Award for Medical Scientists (C. Bettegowda), Thomas M. Hohman Memorial Cancer Research Fund (C. Bettegowda), NIH grant U01CA230691 (C. Bettegowda), NIH grant 5P50CA062924–22 (A.M.L.), and NHMRC Investigator Grant APP1194970 (J.T.).

REFERENCES AND NOTES

1. Deininger P, Alu elements: Know the SINEs. Genome Biol. 12, 236 (2011). [PubMed: 22204421]
2. Deininger PL, Batzer MA, Alu repeats and human disease. Mol. Genet. Metab. 67, 183–193 (1999). [PubMed: 10381326]
3. Feinberg AP, Tycko B, The history of cancer epigenetics. Nat. Rev. Cancer 4, 143–153 (2004). [PubMed: 14732866]
4. Rodriguez J, Vives L, Jordà M, Morales C, Muñoz M, Vendrell E, Peinado MA, Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. Nucleic Acids Res. 36, 770–784 (2008). [PubMed: 18084025]
5. Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R, Kotsinas A, Gorgoulis V, Field JK, Liloglou T, Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int. J. Cancer 124, 81–87 (2009). [PubMed: 18823011]
6. Cho N-Y, Kim B-H, Choi M, Yoo E, Moon K, Cho Y-M, Kim D, Kang G, 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). [PubMed: 17139617]
7. Choi I-S, Estecio MRH, Nagano Y, Kim DH, White JA, Yao JC, Issa J-PJ, Rashid A, Hypomethylation of LINE-1 and Alu in well-differentiated neuroendocrine tumors (pancreatic endocrine tumors and carcinoid tumors). Mod. Pathol. 20, 802–810 (2007). [PubMed: 17483816]
8. Richards KL, Zhang B, Baggerly KA, Colella S, Lang JC, Schuller DE, Krahe R, 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). [PubMed: 19293934]
9. Hunt KV, Burnard SM, Roper EA, Bond DR, Dun MD, Verrills NM, Enjeti AK, Lee HJ, scTEM-seq: Single-cell analysis of transposable element methylation to link global epigenetic heterogeneity with transcriptional programs. Sci. Rep. 12, 5776 (2022). [PubMed: 35388081]
10. Zhou Q, Kang G, Jiang P, Qiao R, Lam WJ, Yu SC, Ma M-JL, Ji L, Cheng SH, Gai W, Epigenetic analysis of cell-free DNA by fragmentomic profiling. Proc. Natl. Acad. Sci. U.S.A. 119, e2209852119 (2022).
11. Agostini M, Pucciarelli S, Enzo MV, Del Bianco P, Briarava M, Bedin C, Maretto I, Friso ML, Lonardi S, Mescoli C, Toppan P, Urso E, Nitti D, 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). [PubMed: 21416156]
12. Mead R, Duku M, Bhandari P, Cree IA, Circulating tumour markers can define patients with normal colons, benign polyps, and cancers. Br. J. Cancer 105, 239–245 (2011). [PubMed: 21712823]
13. Iqbal S, Vishnubhatla S, Raina V, Sharma S, Gogia A, Deo SSV, Mathur S, Shukla NK, Circulating cell-free DNA and its integrity as a prognostic marker for breast cancer. SpringerPlus 4, 265 (2015). [PubMed: 26090312]
14. Wan N, Weinberg D, Liu T-Y, Niehaus K, Ariazi EA, Delubac D, Kannan A, White B, Bailey M, Bertin M, Boley N, Bowen D, Cregg J, Drake AM, Ennis R, Fransen S, Gafni E, Hansen

L, Liu Y, Otte GL, Pecson J, Rice B, Sanderson GE, Sharma A, St J. John C. Tang A. Tzou L. Young G. Putcha IS Haque, Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. BMC Cancer 19, 832 (2019). [PubMed: 31443703]

15. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, Jensen SØ, Medina JE, Hruban C, White JR, Palsgrove DN, Niknafs N, Anagnostou V, Forde P, Naidoo J, Marrone K, Brahmer J, Woodward BD, Husain H, van Rooijen KL, Ørntoft M-BW, Madsen AH, van de Velde CJH, Verheij M, Cats A, Punt CJA, Vink GR, van Grieken NCT, Koopman M, Fijneman RJA, Johansen JS, Nielsen HJ, Meijer GA, Andersen CL, Scharpf RB, Velculescu VE, Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019). [PubMed: 31142840]
16. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, Bruhm DC, Niknafs N, Ferreira L, Adleff V, Chiao JY, Leal A, Noe M, White JR, Arun AS, Hruban C, Annapragada AV, Jensen SØ, Ørntoft M-BW, Madsen AH, Carvalho B, de Wit M, Carey J, Dracopoli NC, Maddala T, Fang KC, Hartman A-R, Forde PM, Anagnostou V, Brahmer JR, Fijneman RJA, Nielsen HJ, Meijer GA, Andersen CL, Mellemgaard A, Bojesen SE, Scharpf RB, Velculescu VE, Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat. Commun. 12, 5060 (2021). [PubMed: 34417454]
17. Treangen TJ, Salzberg SL, Repetitive DNA and next-generation sequencing: Computational challenges and solutions. Nat. Rev. Genet. 13, 36–46 (2012).
18. Douville C, Cohen JD, Ptak J, Popoli M, Schaefer J, Silliman N, Dobbyn L, Schoen RE, Tie J, Gibbs P, Goggins M, Wolfgang CL, Wang T-L, Shih I-M, Karchin R, Lennon AM, Hruban RH, Tomasetti C, Bettegowda C, Kinzler KW, Papadopoulos N, Vogelstein B, Assessing aneuploidy with repetitive element sequencing. Proc. Natl. Acad. Sci. U.S.A. 117, 4858–4863 (2020). [PubMed: 32075918]
19. Lo YMD, Han DSC, Jiang P, Chiu RWK, Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science 372, eaaw3616 (2021).
20. Lui YY, Chik K-W, Chiu RW, Ho C-Y, Lam CW, Lo YD, Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clin. Chem. 48, 421–427 (2002). [PubMed: 11861434]
21. Tinker AV, Boussioutas A, Bowtell DDL, The challenges of gene expression microarrays for the study of human cancer. Cancer Cell 9, 333–339 (2006). [PubMed: 16697954]
22. Ransohoff DF, Gourlay ML, Sources of bias in specimens for research about molecular markers for cancer. J. Clin. Oncol. 28, 698–704 (2010). [PubMed: 20038718]
23. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA, Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739 (2010). [PubMed: 20838408]
24. Moser T, Kühberger S, Lazzeri I, Vlachos G, Heitzer E, Bridging biological cfDNA features and machine learning approaches. Trends Genet. 39, 285–307 (2023). [PubMed: 36792446]
25. Verleysen M, François D, in Computational Intelligence and Bioinspired Systems, Lecture Notes in Computer Science, Cabestany J, Prieto A, Sandoval F, Eds. (Springer, 2005), pp. 758–770.
26. Marée R, The need for careful data collection for pattern recognition in digital pathology. J. Pathol. Inform. 8, 19 (2017). [PubMed: 28480122]
27. Bussola N, Marcolini A, Maggio V, Jurman G, Furlanello C, Pattern recognition, in ICPR International Workshops and Challenges, Lecture Notes in Computer Science, Del Bimbo A, Cucchiara R, Sclaroff S, Farinella GM, Mei T, Bertini M, Escalante HJ, Vezzani R, Eds. (Springer International Publishing, 2021), pp. 167–182.
28. Parvandeh S, Yeh H-W, Paulus MP, McKinney BA, Consensus features nested cross-validation. Bioinformatics 36, 3093–3098 (2020). [PubMed: 31985777]
29. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, Oxnard GR, Klein EA, Smith D, Richards D, Yeatman TJ, Cohn AL, Lapham R, Clement J, Parker AS, Tummala MK, McIntyre K, Sekeres MA, Bryce AH, Siegel R, Wang X, Cosgrove DP, Abu-Rustum NR, Trent J, Thiel DD, Becerra C, Agrawal M, Garbo LE, Giguere JK, Michels RM, Harris RP, Richey SL, McCarthy TA, Waterhouse DM, Couch FJ, Wilks ST, Krie AK, Balaraman R, Restrepo A, Meshad MW, Rieger-Christ K, Sullivan T, Lee CM, Greenwald DR, Oh W, Tsao C-K, Fleshner N, Kennecke HF, Khalil MF, Spigel DR, Manhas AP, Ulrich BK, Kovoor PA, Stokoe C, Courtright JG, Yimer

HA, Larson TG, Swanton C, Seiden MV, Cummings SR, Absalan F, Alexander G, Allen B, Amini H, Aravanis AM, Bagaria S, Bazargan L, Beausang JF, Berman J, Betts C, Blocker A, Bredno J, Calef R, Cann G, Carter J, Chang C, Chawla H, Chen X, Chien TC, Civello D, Davydov K, Demas V, Desai M, Dong Z, Fayzullina S, Fields AP, Filippova D, Freese P, Fung ET, Gnerre S, Gross S, Halks-Miller M, Hall MP, Hartman A-R, Hou C, Hubbell E, Hunkapiller N, Jagadeesh K, Jamshidi A, Jiang R, Jung B, Kim T, Klausner RD, Kurtzman KN, Lee M, Lin W, Lipson J, Liu H, Liu Q, Lopatin M, Maddala T, Maher MC, Melton C, Mich A, Nautiyal S, Newman J, Newman J, Nicula V, Nicolaou C, Nikolic O, Pan W, Patel S, Prins SA, Rava R, Ronaghi N, Sakarya O, Satya RV, Schellenberger J, Scott E, Sehnert AJ, Shaknovich R, Shanmugam A, Shashidhar KC, Shen L, Shenoy A, Shojaee S, Singh P, Steffen KK, Tang S, Toung JM, Valouev A, Venn O, Williams RT, Wu T, Xu HH, Yakym C, Yang X, Yecies J, Yip AS, Youngren J, Yue J, Zhang J, Zhang L, Zhang LQ, Zhang N, Curtis C, Berry DA, Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann. Oncol. 31, 745–759 (2020). [PubMed: 33506766]

30. Douville C, Springer S, Kinde I, Cohen JD, Hruban RH, Lennon AM, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R, Detection of aneuploidy in patients with cancer through amplification of long interspersed nucleotide elements (LINEs). Proc. Natl. Acad. Sci. U.S.A. 115, 1871–1876 (2018). [PubMed: 29432176]

31. 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). [PubMed: 3387438]

32. Willard C, Nguyen HT, Schmid CW, Existence of at least three distinct Alu subfamilies. J. Mol. Evol. 26, 180–186 (1987). [PubMed: 3129565]

33. Curtis SD, Summers M, Cohen JD, Wang Y, Nehme N, Popoli M, Ptak J, Sillman N, Dobbyn L, Buchanan A, Tie J, Gibbs P, Ho-Pham LT, Tran BNH, Zhou S, Bettegowda C, Lennon AM, Hruban RH, Kinzler KW, Papadopoulos N, Vogelstein B, Douville C, Identifying cancer patients from GC-patterned fragment ends of cell-free DNA. medRxiv 22278319 (2022). 10.1101/2022.08.02.22278319.

34. Zheng H, Zhu MS, Liu Y, FinaleDB: A browser and database of cell-free DNA fragmentation patterns. Bioinformatics 37, 2502–2503 (2021). [PubMed: 33258919]

35. Sworder BJ, Kurtz DM, Alig SK, Frank MJ, Shukla N, Garofalo A, Macaulay CW, Shahrokh Esfahani M, Olsen MN, Hamilton J, Hosoya H, Hamilton M, Spiegel JY, Baird JH, Sugio T, Carleton M, Craig AFM, Younes SF, Sahaf B, Sheybani ND, Schroers-Martin JG, Liu CL, Oak JS, Jin MC, Beygi S, Hüttmann A, Hanoun C, Dührsen U, Westin JR, Khodadoust MS, Natkunam Y, Majzner RG, Mackall CL, Diehn M, Miklos DB, Alizadeh AA, Determinants of resistance to engineered T cell therapies targeting CD19 in large B cell lymphomas. Cancer Cell 41, 210–225.e5 (2023). [PubMed: 36584673]

36. Mouliere F, Mair R, Chandrananda D, Marass F, Smith CG, Su J, Morris J, Watts C, Brindle KM, Rosenfeld N, Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. EMBO Mol. Med. 10, e9323 (2018). [PubMed: 30401727]

37. Ding SC, Lo YMD, Cell-free DNA fragmentomics in liquid biopsy. Diagnostics 12, 978 (2022). [PubMed: 35454026]

38. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B, Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci. U.S.A. 108, 9530–9535 (2011). [PubMed: 21586637]

39. Olshen AB, Venkatraman ES, Lucito R, Wigler M, Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557–572 (2004). [PubMed: 15475419]

40. Pisner DA, Schnyer DM, in Machine Learning, Mechelli A, Vieira S, Eds. (Academic Press, 2020), pp. 101–121.

41. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C-C, Lin C-C, Meyer MD, Package 'e1071.' R J. (2019).

42. Terao C, Suzuki A, Momozawa Y, Akiyama M, Ishigaki K, Yamamoto K, Matsuda K, Murakami Y, McCarroll SA, Kubo M, Loh P-R, Kamatani Y, Chromosomal alterations among age-related haematopoietic clones in Japan. Nature 584, 130–135 (2020). [PubMed: 32581364]

43. Saiki R, Momozawa Y, Nannya Y, Nakagawa MM, Ochi Y, Yoshizato T, Terao C, Kuroda Y, Shiraishi Y, Chiba K, Tanaka H, Niida A, Imoto S, Matsuda K, Morisaki T, Murakami Y, Kamatani Y, Matsuda S, Kubo M, Miyano S, Makishima H, Ogawa S, Combined landscape of single-nucleotide variants and copy number alterations in clonal hematopoiesis. Nat. Med. 27, 1239–1249 (2021). [PubMed: 34239136]

44. Loh P-R, Genovese G, Handsaker RE, Finucane HK, Reshef YA, Palamara PF, Birmann BM, Talkowski ME, Bakhoum SF, McCarroll SA, Price AL, Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature 559, 350–355 (2018). [PubMed: 29995854]

45. Raman L, Dheedene A, De Smet M, Van Dorpe J, Menten B, WisecondorX: Improved copy number detection for routine shallow whole-genome sequencing. Nucleic Acids Res. 47, 1605–1614 (2019). [PubMed: 30566647]

46. Wilson EB, Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc. 22, 209–212 (1927).

47. Newcombe RG, Interval estimation for the difference between independent proportions: Comparison of eleven methods. Stat. Med. 17, 873–890 (1998). [PubMed: 9595617]

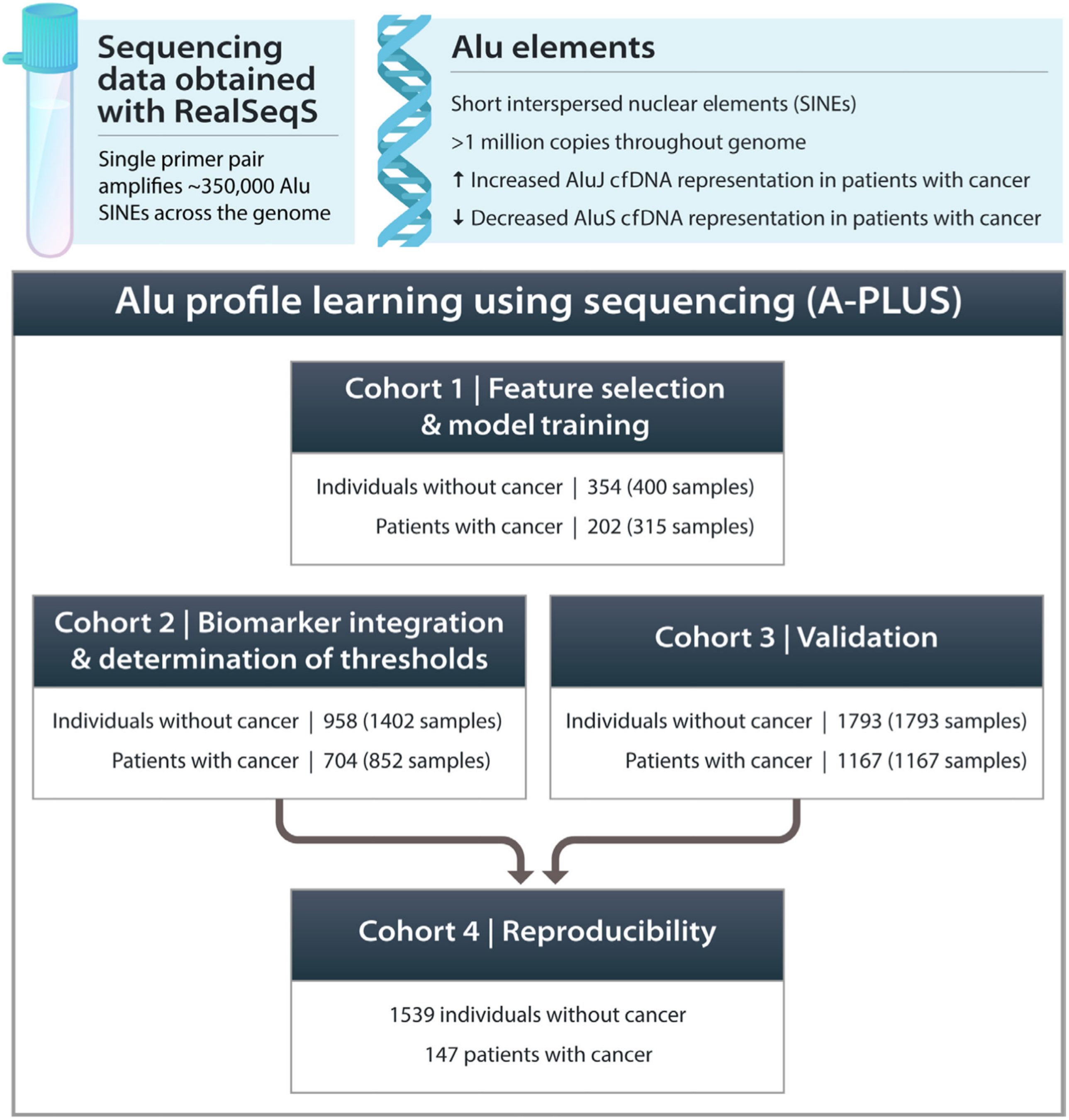
Sequencing data obtained with RealSeqS
Single primer pair amplifies ~350,000 Alu SINEs across the genome
Alu elements
Short interspersed nuclear elements (SINEs)
>1 million copies throughout genome
↑ Increased AluJ cfDNA representation in patients with cancer
↓ Decreased AluS cfDNA representation in patients with cancer
Alu profile learning using sequencing (A-PLUS)
Cohort 1 | Feature selection & model training
Individuals without cancer | 354 (400 samples)
Patients with cancer | 202 (315 samples)
Cohort 2 | Biomarker integration & determination of thresholds
Individuals without cancer | 958 (1402 samples)
Patients with cancer | 704 (852 samples)
Cohort 3 | Validation
Individuals without cancer | 1793 (1793 samples)
Patients with cancer | 1167 (1167 samples)
Cohort 4 | Reproducibility
1539 individuals without cancer
147 patients with cancer

Fig. 1. Study overview.

This study evaluated the potential of Alu elements to enhance methods designed for the earlier detection of cancer. We introduce an ML model called Alu-Profile Learning Using Sequencing (A-PLUS). The model uses differences in the representation of Alu elements in cell-free DNA (cfDNA) to predict cancer status. Samples were prespecified into four distinct cohorts used for model training, analyte integration and threshold determination, validation, and reproducibility.

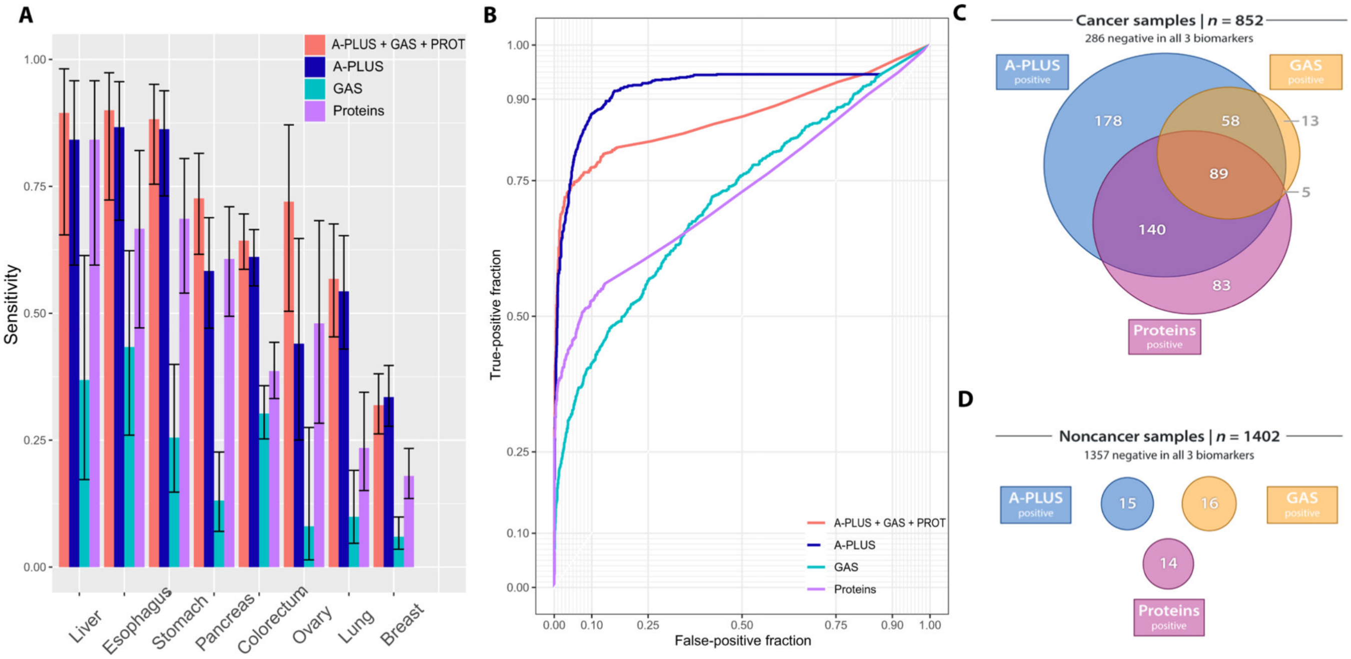

Fig. 2. Detection of cancer in cohort 2 plasma samples.

(**A**) Evaluation of the A-PLUS, GAS, and PROT individual performances as well as that of the multi-analyte classifier (A-PLUS + GAS + PROT). Sensitivities were calculated at 99% specificity in each case. Error bars represent 95% CI. (**B**) ROC curve for the samples in cohort 2. (**C**) Euler diagram of the overlap in biomarker predictions in cohort 2 cancer samples. (**D**) Euler diagram of the overlap in biomarker predictions in cohort 2 noncancer samples.

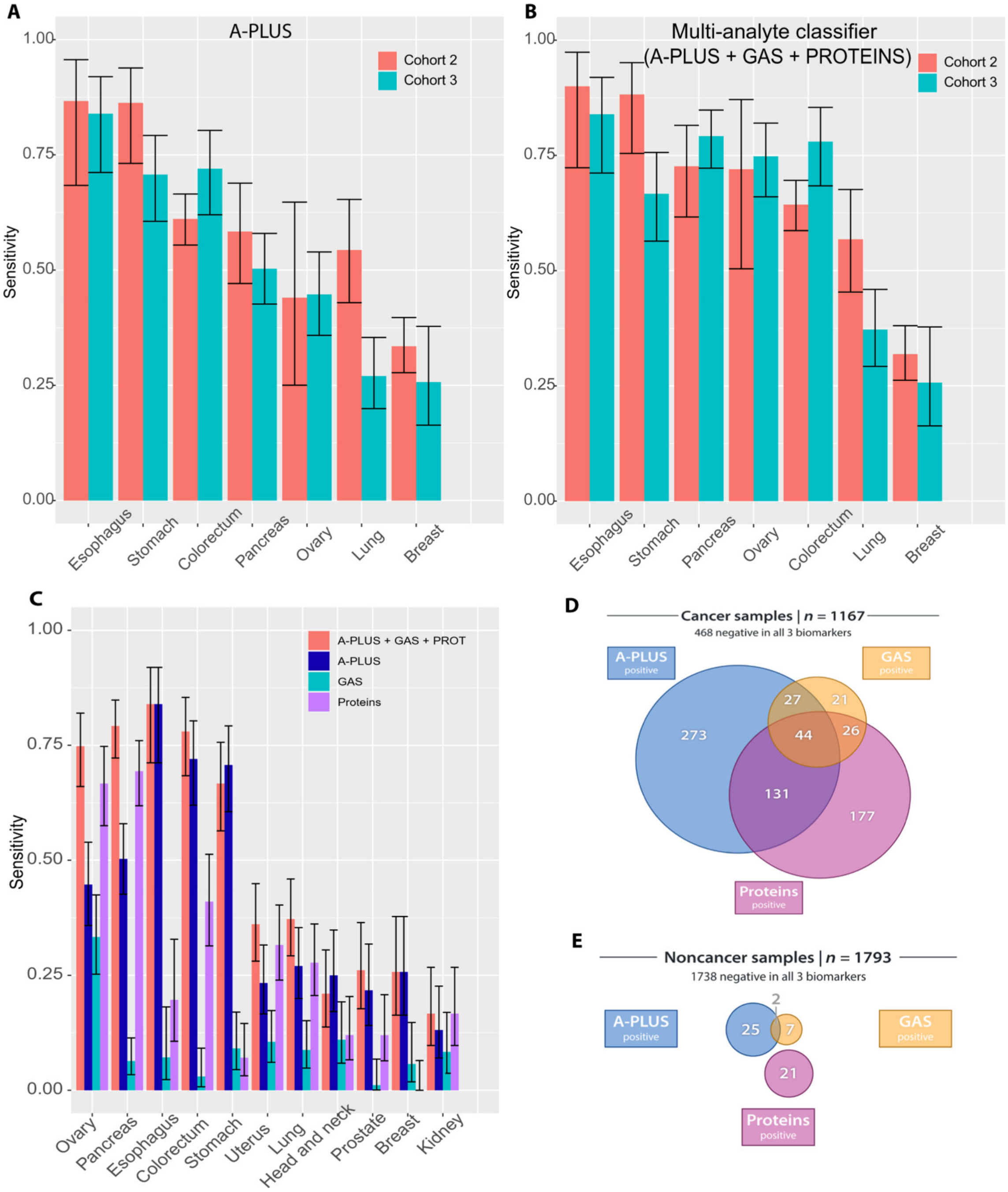

Fig. 3. Detection of cancer in cohort 3 plasma samples.

(**A**) Comparison of cohort 2 and cohort 3 performance for A-PLUS predictions. Only cancer types common to both cohorts are depicted. (**B**) Comparison of cohort 2 and cohort 3 performance for the multi-analyte classifier (A-PLUS + GAS + PROT). Only cancer types common to both cohorts are depicted. (**C**) Full performance metrics for the individual assays and multi-analyte classifier in cohort 3. (**D**) Euler diagram of the overlap of biomarker predictions for cohort 3 cancer samples. (**E**) Euler diagram of the overlap of biomarker predictions in cohort 3 noncancer samples.

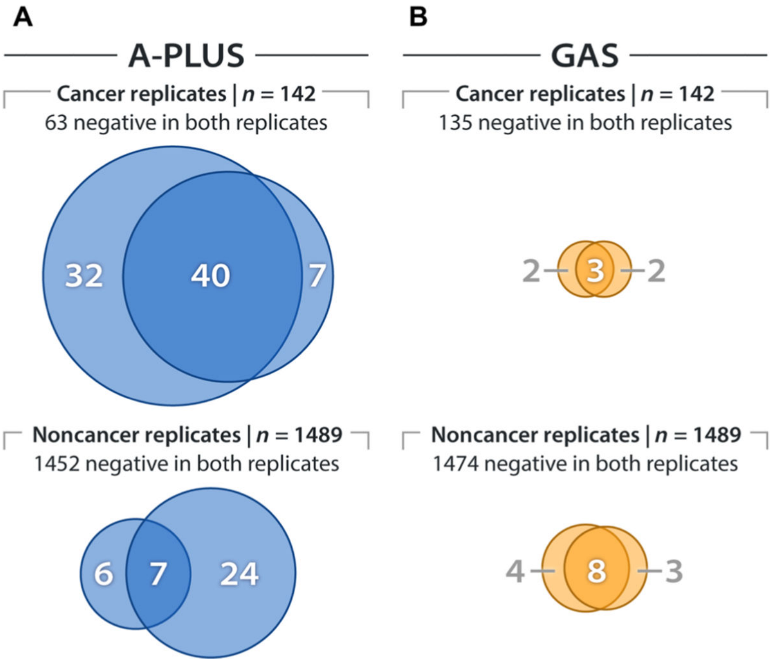

Fig. 4. Reproducibility of biomarker predictions in technical replicates.

The concordance of positive and negative predictions is depicted with Euler diagrams. (**A**) A-PLUS concordance. (**B**) GAS concordance.

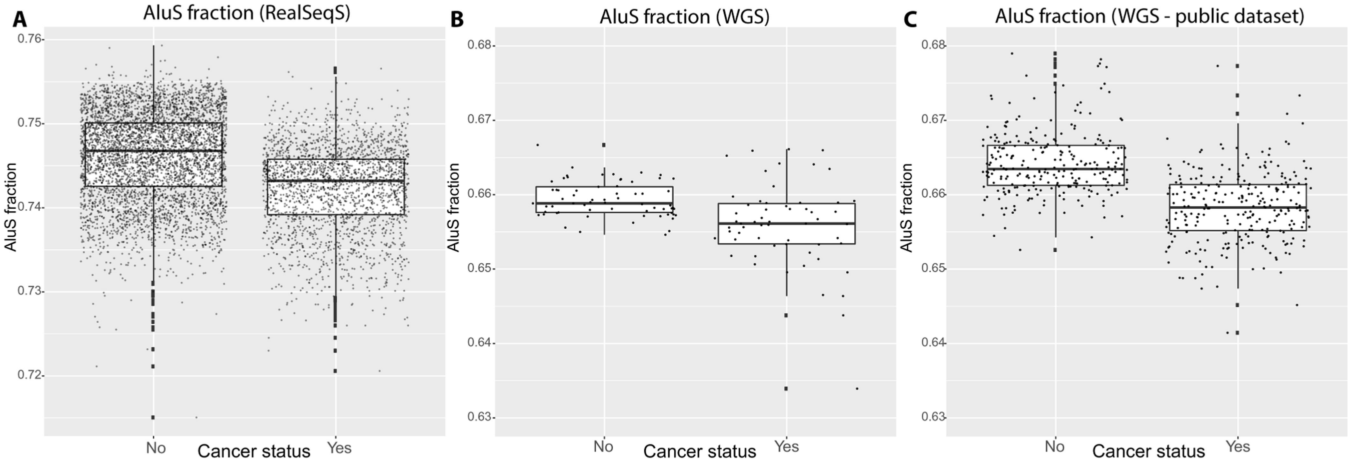

Fig. 5. Cancer samples exhibit a global reduction in sequencing depth of AluS when compared with the total depth of the Alu elements (AluS fraction).

(**A**) AluS fraction of RealSeqS data from samples used in the current study (cohorts 1, 2, 3, and 4). (**B**) AluS fraction for WGS samples. (**C**) AluS fraction for a publicly available WGS dataset. There were 312,138 AluJ loci, 686,962 AluS loci, and 143,178 AluY loci distributed throughout the genome. RealSeqS, however, only amplified a subset of the total Alu loci (AluJ—125,797; AluS—485,614; AluY—86,252), hence the substantial difference between the AluS fractions between RealSeqS and WGS samples.

Table 1.

Number of samples evaluated with RealSeqS.

	Noncancer	Breast	Colorectum	Esophagus	Head and neck	Kidney	Liver	Lung	Ovary	Pancreas	Prostate	Stomach	Uterus
Cohort 1	400	35	69	22	0	0	28	52	22	30	0	57	0
Cohort 2	1402	251	311	30	0	0	19	81	25	84	0	51	0
Cohort 3	1793	70	100	56	100	84		137	119	173	92	99	133

Table 2.

Evaluation of AluS fraction and aneuploidy in WGS data.

	Aneuploidy	**AluS fraction**	**Combined**	***n* =**
Breast	22%	13%	35%	54
Colorectum	61%	26%	65%	23
Liver	36%	32%	56%	25
Lung	50%	20%	59%	76
Ovary	38%	27%	50%	26
Pancreas	37%	9%	40%	35
Stomach	59%	41%	67%	27
Noncancers (false-positive rate)	0.4%	1.2%	1.5%	260