PNAS RESEARCH ARTICLE | MEDICAL SCIENCES
STATISTICS

Minimizing and quantifying uncertainty in AI-informed decisions: Applications in medicine

Samuel D. Curtis[a,b,c,d,e,1], Sambit Panda[f,g,1], Adam Li[h,1], Haoyin Xu[f,g], Yuxin Bai[f,g], Itsuki Ogihara[f,g], Eliza O'Reilly[i], Yuxuan Wang[b,c,d,e], Lisa Dobbyn[b,c,d,e], Maria Popoli[b,c,d,e], Janine Ptak[b,c,d,e,j], Nadine Nehme[b,c,d,e], Natalie Silliman[b,c,d,e,j], Jeanne Tie[k,l,m], Peter Gibbs[k,m,n], Lan T. Ho-Pham[o,p], Bich N. H. Tran[q], Thach S. Tran[q,r], Tuan V. Nguyen[q,r,s,t,u], Ehsan Irajizad[v,w], Michael Goggins[b,d,x,y], Christopher L. Wolfgang[z], Tian-Li Wang[x,aa], Ie-Ming Shih[x,aa], Amanda Fader[aa], Anne Marie Lennon[bb,cc], Ralph H. Hruban[b,x], Chetan Bettegowda[b,c,d,e,dd], Lucy Gilbert[ee], Kenneth W. Kinzler[b,c,d,e], Nickolas Papadopoulos[b,c,d,e], Bert Vogelstein[b,c,d,e,j,2], Joshua T. Vogelstein[f,g,2], and Christopher Douville[b,c,d,e,y,ff,2]

Affiliations are included on p. 10.

Contributed by Bert Vogelstein; received December 2, 2024; accepted July 8, 2025; reviewed by Danilo Bzdok and Alain R. Thierry

AI is now a cornerstone of modern dataset analysis. In many real world applications, practitioners are concerned with controlling specific kinds of errors, rather than minimizing the overall number of errors. For example, biomedical screening assays may primarily be concerned with mitigating the number of false positives rather than false negatives. Quantifying uncertainty in AI-based predictions, and in particular those controlling specific kinds of errors, remains theoretically and practically challenging. We develop a strategy called multidimensional informed generalized hypothesis testing (MIGHT) which we prove accurately quantifies uncertainty and confidence given sufficient data, and concomitantly controls for particular error types. Our key insight was that it is possible to integrate canonical cross-validation and parametric calibration procedures within a nonparametric ensemble method. Simulations demonstrate that while typical AI based-approaches cannot be trusted to obtain the truth, MIGHT can be. We apply MIGHT to answer an open question in liquid biopsies using circulating cell-free DNA (ccfDNA) in individuals with or without cancer: Which biomarkers, or combinations thereof, can we trust? Performance estimates produced by MIGHT on ccfDNA data have coefficients of variation that are often orders of magnitude lower than other state of the art algorithms such as support vector machines, random forests, and Transformers, while often also achieving higher sensitivity. We find that combinations of variable sets often decrease rather than increase sensitivity over the optimal single variable set because some variable sets add more noise than signal. This work demonstrates the importance of quantifying uncertainty and confidence—with theoretical guarantees—for the interpretation of real-world data.

predictive modeling | hypothesis testing | cancer screening | biomedical assays | biomarkers

Significance

The method developed in this paper, multidimensional informed generalized hypothesis testing (MIGHT), addresses a fundamental, underappreciated problem in AI when applied to big data: How do we confidently quantify the amount of predictive information in large sets of variables? Unlike commonly used AI approaches, simulations and theoretical results show that the estimates generated using MIGHT are guaranteed to converge to the truth and are highly reproducible across repetitions, particularly in settings with high dimensionality and low sample sizes. Comparisons between algorithms on real-world data demonstrate higher reliability of MIGHT estimates compared to state-of-the-art algorithms. The application of MIGHT to circulating cell-free DNA from 900 individuals with and without cancer introduces a framework for the development and evaluation of biomedical assays.

With data consisting of many variables, AI tools, such as deep neural networks or support vector machines (SVMs), are often employed for analysis in fields ranging from astronomy to zoology (1). Many of these tools have decades of theoretical development and real-world applications to justify our trust in them for predicting various outcomes given a set of variables (2). For example, in neuroscience, AI tools might predict the presence or absence of Alzheimer's disease (outcome) given variables derived from MRI. In genomics, the variables could be derived from the DNA or RNA sequences of a patient's tissue, and the outcome could be whether the patient has cancer.

Many real-world applications, however, require more than merely predictive accuracy. Consider developing a biomedical screen for a disease. The vast majority of individuals who get screened will not have the disease, even for relatively common diseases such as cancer. If the screen results in too many false positives (that is, incorrectly identifies people without the disease as having the disease, called low "specificity"), too many individuals will require further examination, including invasive assays. Thus, biomarkers for screening purposes must have very high specificity to be clinically useful, even if that means a decrease in correctly identifying cases (sensitivity, or true positive rate), and with it, a corresponding decrease in overall accuracy. Therefore, the development of an effective biomedical assay for screening may aim to optimize sensitivity at high specificity, such as sensitivity at 98% specificity, called S@98 hereafter (3).

Given high-dimensional data, one can leverage AI to estimate quantities of interest, such as sensitivity at a given specificity. To do so, an AI-based classifier is typically trained on data from a cohort of patients. That classifier provides a score for each sample, and a threshold for assigning each sample to the positive or negative class. Varying the threshold

[1] S.D.C., S.P., and A.L. contributed equally to this work.

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

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2424203122/-/DCSupplemental.

Published August 20, 2025.

yields a trade-off between fewer false positives and fewer false negatives. To maximize accuracy, one value of the threshold is chosen, but that threshold does not, in general, have high specificity (which is required in certain applications). The curve characterizing all possible trade-offs is called the receiver operating characteristic (ROC) curve (4). Ideally, the ROC curve generated from the training set accurately estimates the ROC curve for the entire population, not just the individuals in the training cohort. In the literature, ROC curves are typically described as properties of a classifier. Our perspective is that classifiers *estimate* an ROC curve that characterizes the *population* of interest. And we desire that our estimate satisfies the basic desirable properties of estimators, for example, that they are accurate with low variance.

Many classifiers have theoretical guarantees that their estimates of one point on the ROC is accurate (5–7). However, very few classifiers have theoretical guarantees that they can accurately estimate the entire ROC curve, which includes the sensitivity at all possible specificities (8). Moreover, empirically, AI algorithms are typically not well calibrated, i.e., the likelihood of reporting that an individual is positive is equal (calibrated) to the true probability that the individual is actually positive (9). This means that the ROC curves from AI algorithms are often inaccurate (10–12). To mitigate this issue, practitioners often use “calibration” techniques (9, 13). However, such calibration techniques typically lack theoretical guarantees that they converge to the truth. And, while they often empirically perform reasonably well in low-dimensional settings, in the high-dimensional settings of interest in modern datasets, they can be relatively inaccurate. Because of these inaccuracies and uncertainties, collecting an independent cohort of patients to validate the estimates derived from the training cohort is mandated. However, collection and analysis of additional patients and controls is costly and infeasible in many situations. Moreover, there is no theoretical guarantee that the results on the validation cohort are any more accurate than the results on the training cohort—or than on a third independent cohort—though the estimates on these other cohorts will not be overfit. This fundamental issue has contributed to the perceived crisis in scientific reproducibility surrounding AI-based predictions in medicine (14–17).

Once an estimate of a given statistic is obtained, it is also important to know whether that estimate is significantly different from what would be expected by “chance alone.” One could, in principle, simply run a permutation test to obtain a *P*-value. However, using a permutation test on statistics derived using standard AI methods lacks theoretical guarantees (18). Finally, investigators often want to know whether additional biomarkers could improve the sensitivity, specificity, or other performance metrics, and whether their estimated improvements are larger than one would expect by chance (19, 20). Again, one can combine multiple sets of biomarkers in various ways (21) but testing whether there is significant improvement lacks theoretical justification. All these gaps in theoretical understanding and justification of the existing AI toolkit for estimating medically relevant quantities limits the trustworthiness of existing tools. To address these gaps, we developed multidimensional informed generalized hypothesis testing (MIGHT).

Results

Simulations Demonstrating the Value of MIGHT. Suppose we wish to develop a screening test for cancer based on biomarkers containing many variables. Further, suppose that the true distribution of one of the assessed variable sets is standard normal (i.e., Gaussian) for the individuals without cancer (controls). In the cancer patients (cases), a subset looks just like the controls, while the distribution is shifted to the right for others (Fig. 1*A*) Finally, assume that all other of the assessed variables are identically distributed in cases and controls (Fig. 1*B*).

Fig. 1*C* shows the true ROC curve (black) for this simulation, which entails 256 patients and 4,096 variables, numbers that are typical in biomedical datasets used for initially exploring biomarkers. MIGHT’s estimate of the ROC curve (red) closely matched the truth, even though only one of the 4,096 variables contains any signal for cancer (see below for details on how MIGHT works). In contrast, the estimates from other machine learning approaches, after calibration, including random forest (RF, blue), nonlinear SVM (green), logistic regression (LR, orange), and k-nearest neighbors (kNN, brown), were all far from the truth (see *SI Appendix*, Algorithm 5 for details on other algorithms, all use default settings from scikit-learn).

We next considered sensitivity at a high specificity, e.g., 98% specificity (S@98), as a statistic of interest for screening purposes (22). We choose this metric because in clinical screenings, especially for detecting rare but serious conditions such as cancer, the goal is to identify as many true cases as possible without falsely alarming healthy patients. Among these, high specificity is essential, and sensitivity at a fixed high specificity (e.g., 98%) becomes a critical performance measure. Medically relevant datasets generally contain a large number of relatively uninformative (noisy) variables, and a crucial property of any estimate (such as S@98) based on such datasets is robustness against many noisy variables. MIGHT is largely insensitive to thousands of noisy variables, whereas other algorithms demonstrate a drastic performance drop when even tens of noisy variables are included (Fig. 1*D*). Another important property of an algorithm is that its estimates empirically converge to the truth as the sample size increases [sometimes this property is referred to as the minimal requirement for an estimator (23)]. MIGHT indeed converges to the truth in this setting, in fact, quite quickly (Fig. 1*E*), meaning that relatively few clinical samples are needed to achieve high accuracy. RF also converges, but more slowly, whereas the other algorithms do not seem to converge to the truth (Fig. 1*E*). The simulations in Fig. 1 illustrate the situation when variables are associated with cancer. When there is no association of cancer with any of the variables, MIGHT, like the other algorithms, accurately reported this fact (*SI Appendix,* Fig. S1).

Simply having an estimate of a statistic is insufficient for testing hypotheses. For example, suppose that MIGHT estimates that the S@98 of a specific variable set is 0.4. How likely is it that an estimate of S@98 is greater than 0.4 by chance alone? When there are thousands of variables in a set, this could certainly occur by chance, the so-called “curse of dimensionality” (24, 25). To evaluate MIGHT’s ability to reject the null hypothesis (i.e., that there is no relationship between the variables and the outcome), we computed its power in various settings. The power of an algorithm is its probability of correctly rejecting the null hypothesis when the null hypothesis is false. For the S@98 statistic, MIGHT achieved nearly perfect power (1.0) with only 256 samples, even when there were a relatively large number (4,096) of variables (Fig. 1 *F* and *G*). In contrast, other algorithms’ power dropped precipitously as more noisy variables were added and converged to a power of 100% substantially more slowly, if at all.

Another important property of AI algorithms is their false positive rate. Specifically, when there is no relationship between the variables and the outcome, i.e., the null hypothesis should not be rejected. A test is called valid if its rejection rate when the null hypothesis is true is less than or equal to the significance threshold (e.g., <0.05). Reassuringly, MIGHT as well as the other

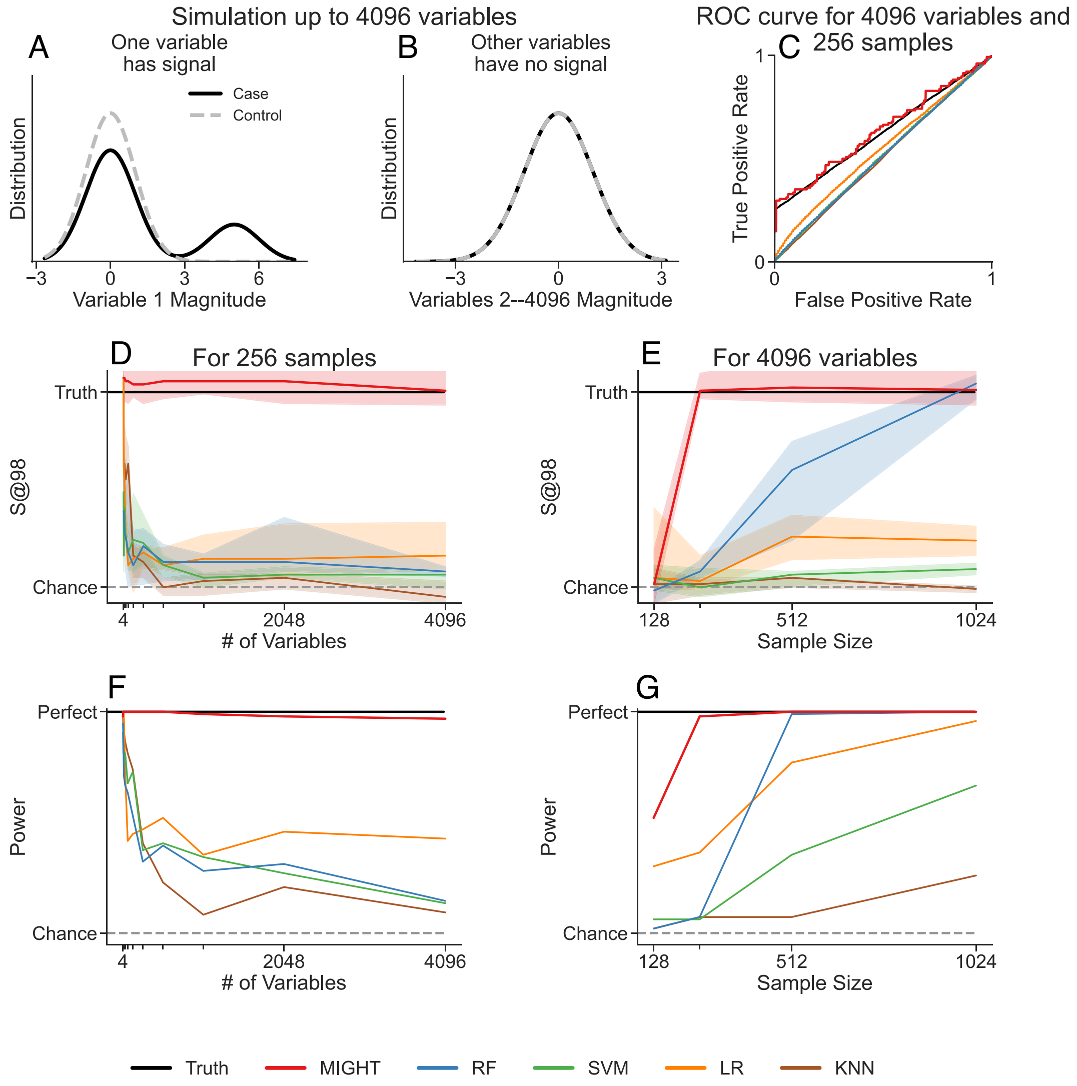


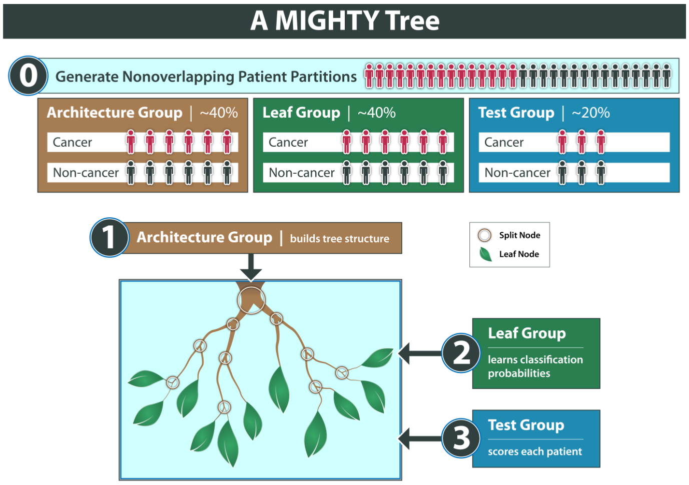
Fig. 1. Performance of MIGHT and conventional algorithms on simulated datasets. (*A* and *B*) Distributions for both case (black) and control (dotted gray), with the first variable shown in (*A*) and all other variables shown in (*B*). (*C*) Optimal ROC curve and estimated ROC curve for each algorithm using 4,096 variables and 256 samples. (*D*) Sensitivity at 98% Specificity (S@98) as a function of the number of variables using 256 samples. (*E*) S@98 as a function of sample size using 4,096 variables. (*F*) Power as a function of the number of variables using 256 samples. (*G*) Power as a function of sample size using 4,096 variables. The classifiers are MIGHT, random forest (RF), support vector machine (SVM), logistic regression (LR), and KNN.

algorithms evaluated in this study, had this property (*SI Appendix*, Fig. S1).

How MIGHT Works.

Computing a test statistic with MIGHT. The conventional heuristic for estimating test statistics such as sensitivity and specificity is to first train some AI procedure (e.g., support vector machine or deep neural network) on a subset of the data ("training set"). Second, use a held-out dataset ("testing set") to "calibrate" the classifier, which means to modify the output of the AI with the goal that the likelihood of reporting that an individual is positive is equal to (calibrated to) the true probability that the individual is actually positive. There are multiple standard empirical procedures for achieving this calibration, including isotonic regression and LR (9) [via Platt scaling (9, 13)]. Then, use a third nonoverlapping dataset ("validating set") to calculate the statistic of interest, such as S@98. Fourth, repeat the above procedure several times, each time with different data in each of the three subsets, and average the results ("cross-validation") (26). The primary issue of concern with this procedure is that it lacks theoretical guarantees that the estimated statistics converge to the truth. This means that a user does not know when it does (or does not) yield trustworthy estimates.

MIGHT uses this type of conventional heuristic (27), but with important modifications (Fig. 2 and *SI Appendix* for details). First, MIGHT constructs a single decision tree on a randomly chosen group of patients from the cohort (the Training Set in Fig. 2). MIGHT then uses an independent group of patients from the same cohort (the Calibrating Set in Fig. 2) to estimate the likelihood of each individual being positive (or not) (25, 28–32). Classical decision trees then essentially ignore the remaining samples (called "out of bag"). MIGHT instead uses the remaining samples (Validating Set in Fig. 2) as validation data for that particular tree. Each decision tree in MIGHT follows the standard process of training, calibration, and validation, but does so using our bagging strategy in place of traditional cross-validation. This is an efficient use of the available data because each tree uses every sample in the cohort rather than a subset of samples. We dub this type of decision tree a "MIGHTY Tree". For the final estimate of the test statistic, MIGHT generally uses 100,000 decision trees, called a MIGHTY Forest. We use so many trees because the S@98 estimate critically depends on the threshold chosen on


A MIGHTY Tree
0 Generate Nonoverlapping Patient Partitions
Architecture Group | ~40%
Cancer
Non-cancer
Leaf Group | ~40%
Cancer
Non-cancer
Test Group | ~20%
Cancer
Non-cancer
1 Architecture Group | builds tree structure
Split Node
Leaf Node
2 Leaf Group learns classification probabilities
3 Test Group scores each patient

```
Step 0: Generate non-overlapping patient sets
    $\mathcal{D}^{\mathrm{T+C}} \leftarrow$ Bootstrap($\mathcal{D}_n, \beta$)
    $\mathcal{D}^{\mathrm{Val}} \leftarrow \mathcal{D}_n \setminus \mathcal{D}^{\mathrm{T+C}}$
    $\mathcal{D}^{\mathrm{Cal}} \leftarrow$ RandomSplit($\mathcal{D}^{\mathrm{T+C}}, s$)
    $\mathcal{D}^{\mathrm{Train}} \leftarrow \mathcal{D}^{\mathrm{T+C}} \setminus \mathcal{D}^{\mathrm{Cal}}$
Step 1: Build tree structure
    tree $\leftarrow$ FitDecisionTree($\mathcal{D}^{\mathrm{Train}}, \Theta$)
Step 2: Learn classificiation probabilities
    tree $\leftarrow$ FitPosteriors(tree, $\mathcal{D}^{\mathrm{Cal}}$)
Step 3: Score each patient
    posterior $\leftarrow$ PredictPosterior(tree, $\mathcal{D}^{\mathrm{Val}}$)
```

Fig. 2. Schematic of MIGHT. In step 0, the samples are separated into three nonoverlapping groups, which are used for the purposes indicated in Steps 1, 2, and 3, corresponding to the train, test, and validation steps in classical cross-validation. As shown by the pseudocode, MIGHT uses bagging, rather than cross-validation. These steps can be repeated any number of times, regardless of the number of samples, whereas cross-validation is limited by the number of samples. See more details in *SI Appendix*, Algorithm 1.

only the positive cases, and so the variance of that threshold is quite high. Because of the efficiency of the code we developed, evaluating 100,000 decision trees is possible even when relatively small computation resources are available.

Computing a P-value with MIGHT. Suppose we want to test whether a given variable set contains any informative signal versus the null hypothesis that it contains none. One way to evaluate this is by computing a p-value, which quantifies how likely it is to observe the given data purely due to random chance. MIGHT enables computation of a P-value by incorporating permutations of sample labels (18, 33–35). Most permutation procedures using classifiers require training thousands of classifier iterations, which is computationally inefficient. We therefore devised an algorithm that only requires training one additional classifier, making it thousands of times more efficient (36).

The classical approach to performing a permutation test using classification algorithms proceeds as follows. 1) Train a classifier on the true data, and obtain an observed test statistic, such as S@98. 2) Permute the labels for each sample to remove any association between the labels and the variable set. 3) Train a classifier on these permuted data, and compute a test statistic using the permuted data. 4) Repeat steps 2 and 3 a thousand times to yield the distribution of the test statistic under the "null" hypothesis: that there is no association between the labels and the variable set. 5) Compare the observed test statistic to the null distribution of the test statistic; the P-value is the fraction of null statistics that are more extreme than the observed statistic. This procedure can be effective, however, it requires 1,000× more compute time than estimating the test statistic, because it requires training the classifier for each permutation. We use a modified procedure based on (36), that only requires training a single additional classifier.

As before, 1) train a classifier on the true data to obtain a test statistic, 2) permute the labels for each sample, and 3) train a classifier on the permuted data. However, we now redefine the observed test statistic to be the difference between the test statistic obtained on the true data, and the test statistic obtained on the permuted data. When there is a real signal in the nonpermuted data, we expect this difference to be large. 4a) Instead of retraining a new classifier, we simply permute the trees across the two classifiers. This yields two new forests, each with about half their trees learned on permuted data, and half on the true data. If the true data and permuted data are the same in distribution, then the resulting trees will also be about the same in distribution. Again, we compute the difference between the test statistics computed from each of these new forests. 4b) Repeat step 4 a thousand times, each time constructing a pair of new forests from the existing trees, to obtain a null distribution of the test statistic. 5) This step is the same as above, we compare the observed test statistic (the difference between the true forest and the null forest's test statistic), with the null distribution of the test statistic (obtain by permuting trees across forests); and the P-value is the fraction of null statistics that are more extreme than the observed statistic. See *SI Appendix*, Algorithm 3 for details. Additionally, we can also test whether one variable set is better than another variable set (details in *SI Appendix*, section B.5).

Theoretical guarantees for MIGHT. A suite of theoretical guarantees for MIGHT is provided in the *SI Appendix*. Here, we briefly state the key assumptions and results. For each patient, we observe a set of d variables x and a binary class label y. We assume that each data pair (x, y) is an independent sample of a set of random variables (X, Y) with unknown distribution (this is known as the iid assumption). An estimator of a given quantity related to the data distribution is consistent if it converges in probability to the truth as the sample size grows. The first result is that MIGHT yields a consistent estimate of the true posterior probabilities under certain conditions on the splitting procedure for each tree and mild assumptions on the data distribution. Let $\eta(x)$ be the true (i.e., population) probability that x is in class 1 (in this case, that x is a cancer patient), and let $\eta_n(x)$ be MIGHT's estimate of $\eta(x)$ from n data samples.

Lemma 1. *Under the setting and assumptions of SI Appendix, section A, $\eta_n(x)$ converges in probability to $\eta(x)$.*

Given a pointwise consistent estimator of the posterior probabilities, we next obtain that MIGHT produces a consistent estimate of sensitivity at any given specificity.

Theorem 1. *Under the setting and assumptions of SI Appendix, section A, we have that MIGHT's estimate of S@r converges in probability to the true population S@r, for any r.*

A hypothesis test is consistent if it will reject a false null with probability converging to one as the number of data samples increases. In other words, its power converges to one for any fixed significance level (33, 35, 37). Of note, while the literature includes many permutation tests (38), proving that a particular permutation procedure yields a consistent test is relatively rare (39, 40). Our final result states that a test for independence using a MIGHT test statistic is consistent.

Theorem 2. *Consider the permutation hypothesis test described in SI Appendix, section A.2, where the null hypothesis is X and Y are independent. For a significance threshold α in the interval (0,1), assume the number of permutations satisfies $M \geq 1/\alpha - 1$. Let the setting and assumptions of Section A in the SI hold and additionally assume that the trees are grown to depth k such that k grows to infinity as the sample size increases. Then, for a distribution of (X, Y) that satisfies the alternative hypothesis, the power converges to one as the number of data samples grows.*

The combination of the empirical results depicted in Fig. 1 and strong theoretical guarantees for MIGHT motivated exploration of its utility to experimental data.

Application of MIGHT to Experimental Data. The evaluation of ccfDNA from plasma, often called liquid biopsies, has been used for purposes ranging from noninvasive prenatal detection of genetic abnormalities in a fetus to cancer screening and monitoring (41–45). Because >50 million ccfDNA fragments are assessed in each patient—each with a unique sequence of ~160 base pairs (bp)—the resultant data incorporate an immense number of variables. Moreover, the distribution of many variables in the cancer patients is likely non-Gaussian and nonlinearly related to the corresponding distribution in the patients without cancer. One approach to analyzing raw ccfDNA sequencing data involves generating a set of informative variables by extracting features such as fragment length and end motifs. Multiple such variable sets can be constructed through different preprocessing and analytical pipelines. Determining whether any particular strategy yields a highly informative variable set—one that can serve as a clinically useful biomarker—remains an active and important area of scientific research (22, 46, 47).

We applied MIGHT to data on ccfDNA fragments purified from the plasma of 102 patients with cancers of the pancreas, colon, breast, liver, ovary, lung, esophagus, stomach, or kidney, and 250 patients without known cancer. To maximize the signal-to-noise ratio, only patients with advanced cancers were assessed (48–50). An average of 25.8 million fragments were collected from each plasma sample [interquartile range (IQR) 22.1 M—29.6 M]. We chose to analyze 44 variable sets, many of which have been analyzed in prior publications (41, 50–55). Each variable set contains between 3 and 15,370 variables. We are searching for the optimal predefined variable set, rather than seeking to select subsets of variables from within a predefined set. We focus on the S@98 statistic, because sensitivity at high specificity is a crucial determinant of the utility of liquid biopsies (48, 56, 57).

The reliability of a given algorithm's estimate of S@98, or any other statistic, is crucial for using such an algorithm in real-world data. Thus, for each of the 44 variable sets, we run, i.e., repeat, each algorithm 10 times on the same data. Because each algorithm has some degree of randomness (due to randomly sampling which data are used for training, testing, or validating), repeats of the identical algorithm on the identical samples will yield a different S@98 for each run. We found that the variability of MIGHT as quantified by the coefficient of variation (CoV) of the S@98 was often lower than 0.02 and as low as 0.004 (Fig. 3). This implies that if we estimate that S@98 was 44% on a single run of MIGHT, we expect other runs of MIGHT to also yield an estimate of between 43% and 45%. In contrast, every other algorithm tested tended to have CoV values that were about 10 times—and sometimes 100 times—higher than MIGHT's. This means that if, for example, SVM estimated S@98 of 44% on one run, on another run, it might be as high as 49% or as low as 39%. Importantly, MIGHT was not only more reliable than other algorithms but was often more sensitive than other algorithms. The S@98 for the best performing variable set, Wise-5, was 72% using MIGHT. The best performing algorithm other than MIGHT on this variable set, TabPFN (58), achieved an S@98 of 71%. However, the CoV values for the S@98 were 0.004 and 0.046 for MIGHT and TabPFN, respectively. We similarly measured the IQR and SD of each algorithm and found that MIGHT often had values that were one to two orders of magnitude lower than every other algorithm tested (SI Appendix, Fig. S4 A and B). The low reliability of these other algorithms rendered them of limited utility for estimating S@98 or for ranking the sensitivity of different variable sets. We therefore proceeded to use MIGHT to compare the performance of each of these 44 variable sets, grouped in classes defined by their biochemical nature (Fig. 3).

Aneuploidy-based variable sets, reflecting abnormal chromosome numbers in cancer cells (59–65), achieved the highest performance, with S@98 up to 0.72, correctly identifying 73 of 102

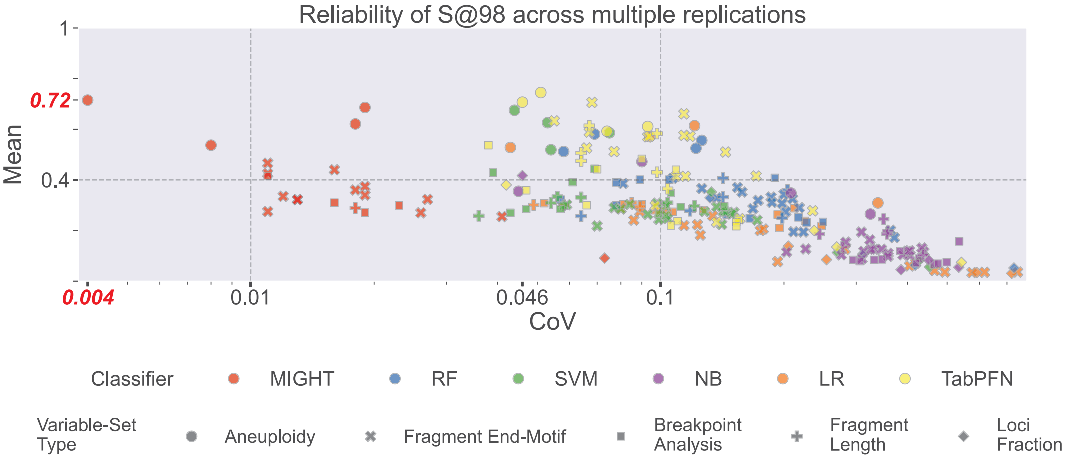


Fig. 3. Variation of S@98 estimates achieved with various classifiers. For each classifier, 10 iterations of the identical data from Cohort 1 were performed on each of 44 variable sets, with numbers of variables ranging from 3 to 15,370. Each dot represents a different one of the 44 variable sets. The classifiers were MIGHT, RF, SVM, Naive Bayes (NB), LR, and TabPFN. Highest sensitivity was observed with MIGHT on Wise-5, with a mean S@98 of 0.72. MIGHT posteriors for all variable sets and all patients in Cohort 1 are available in Dataset S3.

cancer samples while misclassifying only 5 of 250 normal samples (Fig. 3). Fragment End-Motifs, representing sequences at ccfDNA fragment ends (66, 67), yielded fifteen different variable sets (*SI Appendix*, Figs. S5 and S6) with varying numbers of variables (4 to ~16,000). Three motif-based sets (outside pentamers, tetramers, and hexamers) achieved S@98 estimates above 0.4. Breakpoint analysis, examining fragment end-positioning within regions with differential chromatin structure between cancer and healthy cells (54, 63, 66–70), yielded eleven variable sets based on breakpoint frequencies within recurrently protected regions (RPRs) (66), cis-regulatory elements, and repetitive elements, with four sets achieving S@98 > 0.4. Fragment length analysis, one of the earliest discriminating variables for cancer detection (50, 57, 71–74), showed that individual base-pair resolution (70 to 499 bp fragments, 430 variables) achieved S@98 of 0.45, while binning into larger intervals reduced performance. A commonly used ratio-based approach comparing short to long fragments across 5 Mb genomic intervals yielded an S@98 of 0.26 (75). Finally, loci fraction variable sets based on the relative abundance of repeated sequence families and other functionally important genomic elements (76) all achieved S@98 values lower than 0.4.

We applied our permutation-based approach for calculating a *P*-value to all 44 variable sets from whole genome sequencing of plasma ccfDNA. The number of variables in these sets differed by more than 10,000-fold (from 3 to ~16,000), and the S@98 of these sets differed by nearly sixfold (0.09 to 0.72). Nevertheless, for all 44 variable sets, MIGHT's estimates of S@98 were different from those predicted by chance alone ($P < 0.0001$). This provided confidence that even the sets with large numbers of variables were performing considerably better than expected if none of the variables within them, alone or in combination, were related to cancer. We also confirmed that the single best variable set, Wise5, achieves a significantly higher S@98 than any of the other variable sets (*SI Appendix*, Algorithms 3 and 4).

CoMIGHT For Evaluating More Than One Variable Set. Once MIGHT is used to discover that a variable set is associated with an outcome, a naturally arising question is whether another variable set adds to this association. This situation is particularly challenging when the number of variables in one set is vastly different from the number of variables in the other set. To address this question, we developed a variation of MIGHT, called CoMIGHT (for Combined MIGHT) to simultaneously evaluate multiple variable sets.

Computing a test statistic with CoMIGHT. Evaluation of multiple variable sets is often termed multiview or multimodal learning (19, 21). To learn with multiple variable sets, one could train AI on one set, train another AI on the other set, and then combine answers. This approach suffers when the variables in each set alone provide little information about outcome but the combination of the two sets provides a large amount of information. Alternatively, one could combine multiple variable sets into a single ensemble and ignore which variable comes from which set. When the number of variables in one set is much larger than the other, the signal from the variable set with fewer variables could be swamped by the noise in the variable set with more variables. CoMIGHT mitigates this effect by balancing both sets in a way that does not allow any node of the tree to use a large number of variables from one set and zero variables from the second set. This is achieved through differential (stratified) sampling from each variable set at each node of the tree (*SI Appendix*, section 2.B). This stratified approach is possible due to the nature of the random tree construction common in RFs and gradient boosting trees, but not as easily implemented in other algorithmic approaches, such as deep learning.

Simulations were performed to determine whether a second variable set improves S@98, by comparing the S@98 estimates with and without a second set (*SI Appendix*, Fig. S2 *A* and *B*). To simulate a particularly challenging situation, we evaluated the case in which only one of the variables in each set contributes any useful information. Even when the variable sets contain vastly different numbers of variables, S@98 estimates from CoMIGHT remain high even with thousands of uninformative variables from the second set, unlike other approaches (*SI Appendix*, Fig. S2 *C* and *D*). Moreover, even with so many uninformative variables, CoMIGHT converges to the truth with a relatively small number of samples (*SI Appendix*, Fig. S2*E*). The power of CoMIGHT to reject a false null hypothesis, like that of MIGHT, is considerably greater than that of conventional algorithms and remains relatively high, regardless of the number of variables or samples (*SI Appendix*, Fig. S2 *F* and *G*) (77). In addition to the examples provided by these simulated datasets, mathematical proofs that CoMIGHT is a universally consistent estimator are provided in *SI Appendix*, Theorem 3.

Computing a P-Value with CoMIGHT. As with MIGHT, these results brought up a basic statistical question: Does the inclusion of other variable sets modify the S@98 score of one variable set alone more than expected by chance? In statistics, this is known as a model selection problem. To answer this question, we developed a permutation-based approach for CoMIGHT that was analogous to that described for MIGHT. Its key principle was that only the variables in the second variable set, rather than all the variables or the labels of the samples, were permuted. Simulations illustrate that the power of CoMIGHT approaches one, whereas other algorithms failed to achieve such power at sample sizes typical of experimental data (number of samples < 400, *SI Appendix*, Fig. S3). Moreover, CoMIGHT was the only algorithm that proved valid where there is no relationship between variables and outcome (*SI Appendix*, Fig. S3).

Theoretical guarantees for CoMIGHT. In CoMIGHT, the question is whether an additional variable set adds any signal. To theoretically address this question, we enrich the theoretical framework mentioned above. Specifically, we assume there exists an additional variable set, Z, such that the data triple (x,y,z) is independently sampled from the random variable set (X,Y,Z). Perhaps the most natural statistic to consider in this scenario is conditional mutual information, which quantifies the amount of uncertainty between X and Y when conditioned on Z. In the *SI Appendix*, we prove that, under the same assumptions we used for MIGHT, that CoMIGHT's estimate of mutual information and conditional mutual information are consistent, that is, they both converge to the truth (Theorems 3 and 4, respectively).

Theorem 3. *Under the setting and assumptions of SI Appendix, section A, we have that MIGHT's estimate of mutual information* $I_n(X;Y)$ *converges in probability to the true population mutual information,* $I(X;Y)$*, as* $n \to \infty$.

Theorem 4. *Under the setting and assumptions of SI Appendix, section A, and the mutual information estimates using two honest forests built to Specification 1 for* $I_n([X, Z]; Y)$ *and* $I_n(Z; Y)$*, the conditional mutual information estimate* $I_n(X; Y \mid Z) = I_n([X, Z]; Y) - I_n(Z; Y)$ *is consistent. That is,* $I_n(X; Y \mid Z)$ *convergences in probability to* $I_n(X; Y|Z)$ *as* $n \to \infty$.

We then prove that under certain conditions, a positive mutual information implies S@r is greater than chance.

Lemma 2. *If the population ROC curve is concave, then* $I(X; Y) > 0$ *if and only if* $S@r > 1 - r/100$ *for all* $r \in [0, 100]$.

Moreover, if S@r is greater than chance, then so is mutual information. This collection of theoretical results connects mutual information to S@r, and motivates using CoMIGHT to estimate S@r for these purposes.

Application of CoMIGHT to Experimental Data. CoMIGHT was used to determine whether combining any variable set with the best performing variable set (Wise-5) would increase S@98 over the maximum achieved with a single variable set. We used a classical forward insertion approach, with a slight modification. Typically, forward insertion is used to determine whether a single variable improves performance; here, we inserted an entire variable set, with anywhere between 3 and 15,370 variables. We combined each of the 40 variable sets that are not based aneuploidy with Wise-5. Fig. 4 These combinations never increased S@98 above MIGHT's Wise-5 estimate of 0.72, indicating that these other variable sets added more noise than signal once aneuploidy was taken into account (Fig. 4).

MIGHT and CoMIGHT for the Detection of Early Cancers of the Breast and Pancreas. As an additional example of the value of MIGHT, we then addressed an important question in multicancer early detection: Can the identical variable sets be used to detect cancers derived from different tissue types? Previous studies have addressed this question, with somewhat conflicting results (48, 49, 75, 78). There are at least two possible explanations that could explain how different variables could have higher performance in one cancer type than in another cancer type. First, it is possible that the algorithms were not as well-designed for some cancer types as they were for others. Second, it is possible that the particular algorithm used to predict performance was not the issue, but rather that there were different amounts or different characteristics of the DNA released into the circulation from different cancer types. Which of these explanations is most likely can best be addressed by a method that is universally consistent, such as MIGHT.

To inform these explanations with MIGHT, we chose a cohort of 549 individuals without cancer, 126 patients with Stage II breast cancer, and 125 patients with Stage II pancreatic cancer (Dataset S4). Stage II cancers, when detected early enough, offer a higher possibility of cure than later stage cancers. In each cancer type, we assessed the 44 variable sets described above. Through CoMIGHT analysis, we found that the information provided about cancer status from the plasma of breast cancer patients was uniformly less than that from the plasma of pancreatic cancer patients (Fig. 5*A* and Dataset S5). Moreover, as with MIGHT, CoMIGHT was much more reliable than the other algorithms, with CoMIGHT achieving lower CoV for nearly every variable set than *any* of the variable sets for the other algorithms, while also typically achieving as high or higher S@98 and Fig. 5*C* and Dataset S6. The implications of these results for multicancer early detection are discussed below.

Discussion

As documented by theory as well as simulated and experimental data, the MIGHT suite of algorithms provides a broadly applicable approach to judge the amount of signal provided by a variable set. MIGHT-derived statistics such as S@98 can be directly applied to predictive modeling. When typically evaluating AI methods, benchmark data are used to compare algorithms, and this can be done because the ground truth—which samples are positive or negative—is available. However, when evaluating AI methods to estimate statistics such as S@98, there are no existing satisfactory benchmark datasets, because the probability of any sample being positive or negative is not available in real world data. This is also true when evaluating the power of an approach. For these reasons, we leverage theory and simulations to build evidence that our algorithms are trustworthy.

Accurately assessing signal in high-dimensional data is notoriously difficult due to the well-known curse of dimensionality and the associated bias–variance tradeoff. In contrast, a notable strength of MIGHT is its ability to maintain high predictive accuracy even when only a small number of informative variables are embedded within a large set of uninformative ones. This robustness—demonstrated in our experiment (Fig. 1), sets MIGHT apart from conventional methods that typically struggle under such conditions. This is particularly important when a variable set including only a few variables provides more information about a state than variable sets which each contain orders of magnitude more variables. CoMIGHT enables one to determine whether a second (or third,...) variable set adds any value at all. CoMIGHT revealed that in one cohort, combining variable sets always resulted in *lower* sensitivity rather than retaining the same sensitivity or higher sensitivity, because the other variable sets added more noise than signal. These results were consistent across algorithms, indicating that it was not the case that CoMIGHT was unable to find additional signal that was present, yet other algorithms were; rather, that MIGHT essentially found all the signal present across any of the variable sets (in the set of algorithms we considered). These kinds of results are informative for machine learning as they directly focus on the combination of variable sets providing high signal and low noise, with minimal bias and regardless of the nature of the distribution of the variables.

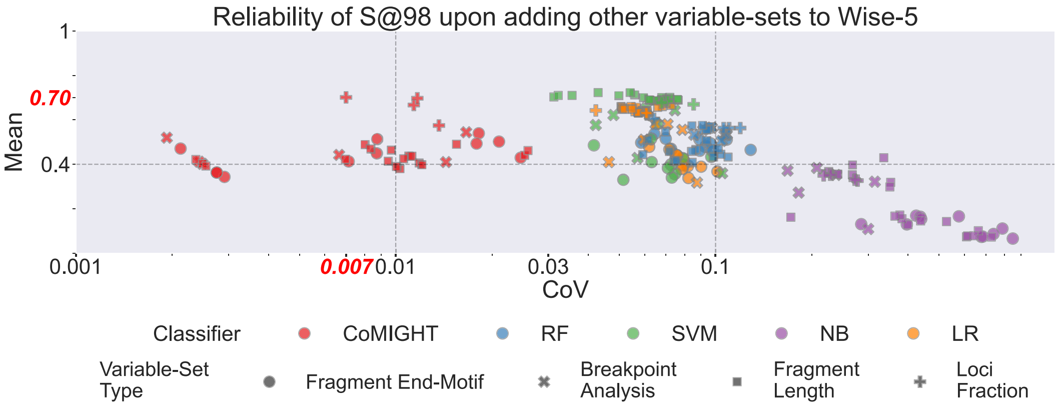


Fig. 4. Performance of CoMIGHT on Cohort 1. Relationship between the CoV and the mean S@98 across five classifiers upon adding other 43 variable sets to Wise-5. CoMIGHT classifier achieved S@98 at 0.70 with a CoV of 0.007, whereas the best other algorithm on multiple variable sets only achieved an S@98 of 0.70 with a CoV of 0.03, which is greater than fourfold worse than MIGHT.

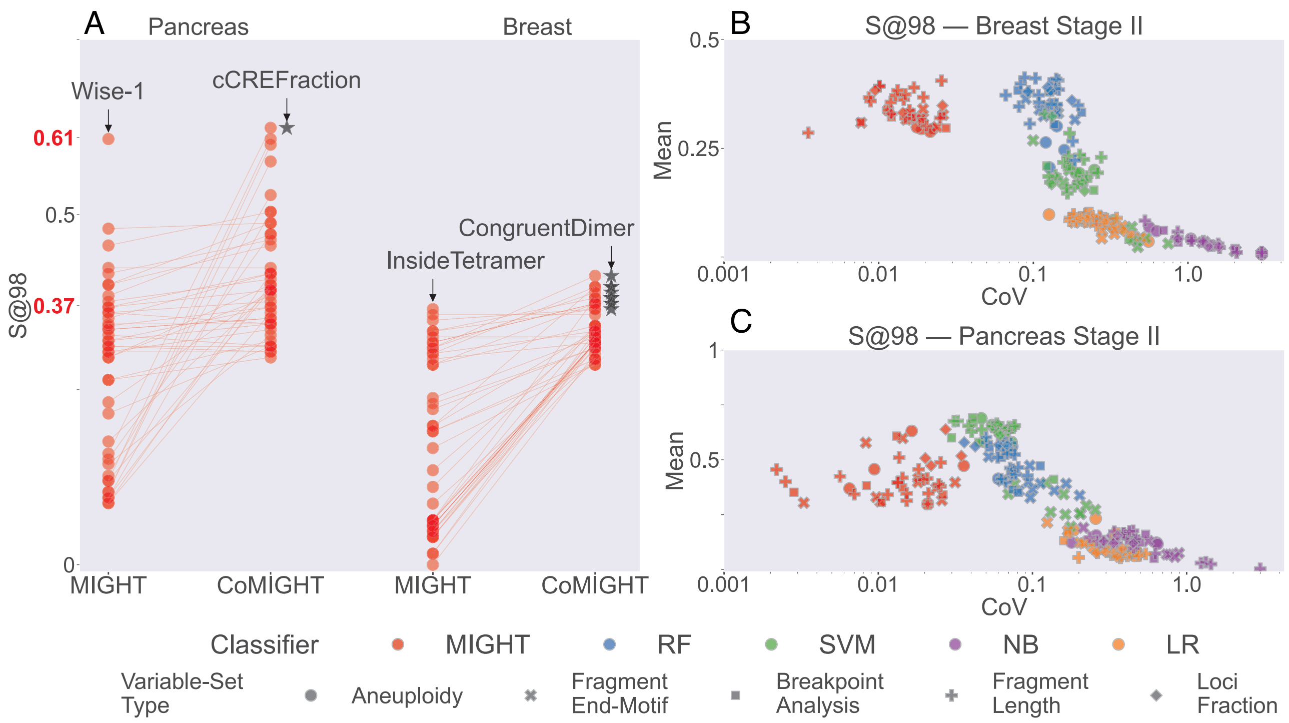


Fig. 5. Comparison of Stage II breast and pancreatic cancers. (*A*) MIGHT analysis of the 44 variable sets on pancreas and breast cancer. CoMIGHT analysis of adding the top variable set, to other 43 variable sets in either breast or pancreatic cancers. Variable sets that had statistically significant ($P < 0.005$) increases in S@98 over the best single variable set are labeled with stars. (*B*) Relationship between the CoV and the mean S@98 across five classifiers upon adding other 43 variable sets to InsideTetramer on breast cancer stage II. (*C*) Relationship between the CoV and the mean S@98 across five classifiers upon adding other 43 variable sets to Wise-1 on pancreas cancer stage II. S@98 estimates for MIGHT and CoMIGHT are available in Dataset S5. MIGHT and CoMIGHT posteriors for patients with pancreatic and breast cancer are available in Datasets S6 and S7, respectively.

Another useful property of MIGHT is its ability to analyze multiple cancer types. A second cohort revealed that machine learning algorithms based on the 44 sets of variables listed in Dataset S2 are unlikely to make the sensitivity for detection of patients with Stage II breast cancers as high as that of patients with Stage II pancreatic cancers. Interestingly, in this cohort, CoMIGHT revealed that certain pairs of variable sets had more signal than the top single variable set, suggesting that in early stage cancers, multiple types of variable sets may be valuable. To raise the sensitivity of an assay for breast cancer detection, therefore, investigators could add different variable sets from the DNA sequencing data, or from different analytes, such as DNA methylation, RNA, proteins, or metabolites. These other variables could be evaluated by MIGHT in the same way as described above, and their ability to add to these DNA-sequence based variables could be assessed via CoMIGHT.

Previous studies reporting whole genome sequencing of plasma ccfDNA have provided a plethora of variable sets that have been incorporated into various algorithms and predictive assays (41, 50–52). It has therefore been challenging to estimate which of these variable sets are most useful. The challenge arises in part because the amplification of the original template molecules and the other techniques required to determine the sequences of ccfDNA are performed differently in different laboratories, complicating comparison. A commonly encountered scenario is that an algorithm performs well in a first independent cohort, but not in a second cohort assembled by other researchers (due to "batch effects" or "Out-of-Distribution" data) (29, 31, 76, 79, 80). This circumstance contributes to the ongoing debate about the application of AI algorithms to life and death decisions that often are made in medical practice. The debate is fueled by the fact that no rigorously defined statistical approach has been available to compare sets of variables with respect to certain critical endpoints, such as sensitivity at high specificity. MIGHT addresses this latter issue and can be applied to multiple fields of scientific research. An interesting direction for future work is to generalize MIGHT to be able to handle batch effects and out-of-distribution data.

To facilitate building on this work, we have made all the code for these methods available at https://treeple.ai/ and have deposited the data in the European Genome-Phenome Archive under EGA00001007763.

Materials and Methods

Experimental Study Design. This study was approved by the Institutional Review Boards for Human Research at Johns Hopkins Medical Institutes and other participating institutions in compliance with the Health Insurance Portability and Accountability Act. No proper sample size was calculated. Samples were chosen on the basis of availability to be representative of multiple tumor types and a diverse range of tumor stages. All individuals participating in the study provided written consent. All analyses were retrospective in nature. Blood was collected in Streck tubes and plasma was purified from 678 individuals without cancer and 354 patients with solid cancer using the BioChain ccfDNA Extraction Kit (Cat X K5011625) within 2 d. All patients were deidentified, and patients are not known to anyone outside the research group. Demographics for the individual are included in Datasets S1 and S4.

Whole Genome Sequencing. We developed a library preparation workflow that can efficiently re-cover input DNA fragments and simultaneously incorporate double-stranded molecular barcodes (48). In brief, libraries were prepared with ccfDNACell using an Accel-NGS 2S DNA Library Kit (Swift Bio- sciences, 21024) with the following critical modifications: 1) DNA was pretreated with 3 U of USER enzyme (New England BioLabs, M5505L) for 15 min at 37 °C to excise uracil bases; 2) the SPRI bead/PEG NaCl ratios used after each reaction were 2.0×, 1.8×, 1.2×, and 1.05× for end repair 1, end repair 2, ligation 1, and ligation 2, respectively; 3) a custom 50 μM 3′ adapter was substituted for reagent Y2 and 4) a custom 42 μM 5′ adapter was substituted for reagent B2. Libraries were subsequently PCR amplified in 50-μL reactions using primers targeting the ligated adapters. The following reaction conditions were used: 1× NEBNext Ultra II Q5 Master Mix (New England BioLabs, M0544L), 2 μM universal forward primer and 2 μM universal reverse primer. Libraries were amplified with 7 or 11 cycles of PCR, depending on how many experiments were planned, according to the following protocol: 98 °C for 30 s, cycles of 98 °C for 10 s, 65 °C for 75 s and hold at 4 °C. If seven cycles were used, the libraries were amplified in single 50-μL reactions. If 11 cycles were used, the libraries were divided into eight aliquots and amplified in eight 50-μL reactions, each supplemented with an additional 0.5U of Q5 Hot Start High-Fidelity DNA Polymerase (New England BioLabs, M0493L), 1 μL of 10 mM dNTPs (New England BioLabs, N0447L) and 0.4 μL of 25 mM MgCl$_2$ solution (New England BioLabs, B9021S). The products were purified with 1.8× SPRI beads (Beckman Coulter, B23317) and eluted in EB buffer (Qiagen). Whole genome libraries were sequenced with paired-end 2× 100 bp sequencing on either a HiSeq 4000 or NovaSeq 6000 to a median depth of 26.6 M read pairs, or 1.35× (IQR 23.6 M-30.1 M).

Bioinformatic Pipeline. Fastq files were demultiplexed using a custom script that utilized index sequences added during library preparation. Demultiplexed read 1 and read 2 fastq files were trimmed using a custom script to remove 27 base oligonucleotides added during library preparation. Trimmed sequences were then aligned to the hg19 genome with bowtie2 using end-to-end alignment (81). After alignment, UID duplicates were removed using a custom script. Picard AddOrReplaceReadGroups was used to add read groups (82).

Quality Control. Samtools flagstat (83) was used to evaluate sequencing quality. Any samples that had greater than 2.5% singletons, less than 80% of reads mapped, or less than 80% of reads properly paired were removed from further analysis. Only molecules that were mapped to autosomal chromosomes, were properly paired, had a MAPQ > 30, and were between 70 to 500 bp in length were used in all analyses.

Fragment Length Analysis. Fragment length was extracted from the BAM files using the TLEN alignment field. Fragments for each sample were binned into either 1, 5, 10, 15, or 20 length bins. For example, fragment lengths of 70 to 74 is one 5-length bin, 70 to 79 is one 10-length bin, and 70 to 89 is one 20-length bin.

Fragment length ratios across the genome were calculated using a method inspired by a previous study (75). Individual fragments within autosomes were first binned into nonoverlapping 100 kb bins. Using the 124 control samples in the panel of normal (Dataset S1) we evaluated the mean coverage and CoV for each 100 kb bin. Any bin that had mean coverage below the 10th percentile or above the 99th percentile was removed. Any bin that had a CoV greater than the 90th percentile was removed. To correct for the possible influence of GC content on fragment lengths, we applied a LOESS regression with span of 0.75 on the relationship between average fragment GC and coverage calculated for each of the remaining 100 kb bins. Separate LOESS regression models were performed for short (100 to 150) and long (151 to 220) to account for possible differences in GC effects by fragment length. Fragment length frequencies were then normalized by subtracting the LOESS predicted value from the observed frequency, resulting in residuals for short and long fragments that were uncorrelated with GC content. Normalized fragment length frequencies were returned to the original scale by adding the median genome-wide short and long coverage to the normalized values. Finally, GC-normalized counts for each 100 kb bin were aggregated into 577 nonoverlapping 5 Mb bins to reduce dimensionality and noise.

Fragment End-Motif Analysis. Fragment start position, end position, length, and strandedness (±) were extracted from the bam file and converted into bed format. The start and end position of each fragment was then extended by 10 positions to evaluate the sequences upstream and downstream of each fragment-end. The full nucleotide of each fragment and the 10 bases upstream and downstream of the fragment was then extracted from the hg19 reference genome using bedtools nuc (84). Orientation of 5′ and 3′ of each fragment was inferred using the strandedness of each molecule. Fragments that aligned to the nonreference (−) strand of the hg19 reference genome were reverse complemented. The frequency of 5′ end-motifs, 3′ end-motifs, and fragment lengths were calculated by dividing by the count of each motif/length by the total number of fragments analyzed. Due to end-repair of the 3′ end during library preparation, the average frequency between the 5′ end-motif and reverse complement of the 3′ end-motif was used as the final frequency.

When evaluating end-motifs we analyzed "inside," "outside," and "congruent" motifs. "Inside" motifs are defined as the observed fragment-end motif of the sequenced molecule (*SI Appendix*, Figs. S5 and S6). Outside motifs are defined as the motifs upstream or downstream of the fragment-end, representative of the motif on the opposite side of the cleavage site. Congruent motifs are half outside and half inside. For example, the "congruent hexamer" is defined as the trimer upstream of the 5′ fragment-end plus the 5′ trimer motif of the sequenced molecule. When evaluating end-motifs, it is possible to analyze any given sized sequence such as monomers, dimers, or trimers, tetramers, pentamers, or hexamers. All the aforementioned sizes were analyzed for both inside and outside motifs, while only the even sized motifs (dimers, tetramers, hexamers) were analyzed for congruent motifs.

Breakpoint Analysis. BED files for ENCODE V3 candidate cis-regulatory elements and UCSC RepeatMasker repetitive elements were downloaded from the UCSC Genome Browser (85, 86). A bedfile for GM12878 A/B genome compartments was downloaded from Xiong and Ma (87). A BED file for RPRs was downloaded from Budhraja et al. (68). Bedtools intersect (v2.30.0) was then used to intersect the fragment bed file with each genomic element, requiring only a single base position of overlap between the molecule and the genomic loci to be included. Each region that is found to intersect with the chosen loci analyzed for fragment breakpoints.

To calculate the breakpoint variables, we first determined the central base position of each genomic loci (*SI Appendix*, Fig. S7). If there was an even number of bases in the loci the central position was rounded up. Each bond between the bases was considered as a possible breakpoint (e.g., counting the number of phosphodiester bonds, not nucleotides). For each genomic loci we analyzed positions −150 to +150 from the central position, where position −1 is the bond between the central nucleotide and the nucleotide directly upstream. To calculate the breakpoint ratio, the number of fragment-ends at that position was divided by the number of fragments that overlapped that position (e.g., had coverage at the nucleotide upstream and downstream of that position). In total, 300 possible breakpoints were evaluated for each loci.

Genomic Fraction Analysis. BAM files were intersected with autosomal loci from bed files for four different families of genomic loci: Alu elements, LINEs, Replication timing compartments, and cis-regulatory elements. Using bedtools intersect (v2.30.0), molecules from each sample were intersected with each family of genomic loci, requiring only a single base position of overlap between the molecule. If a single molecule overlapped with multiple loci of the same subfamily (e.g., multiple AluY repeats) it was only counted a single time. The number of fragments intersecting each of the different subfamilies of the target loci (e.g., L1/L2/L3 for LINEs, or AluJ/S/Y for Alus) was counted. Finally, the number of fragments in each subfamily was divided by the total number of fragments intersecting all subfamilies of the genomic loci (e.g., AluJ divided by AluJ+AluY+AluS).

Aneuploidy Analysis. WiseCondorX (64) version 1.2.4 was downloaded using conda with the following command: conda install -f -c conda-forge -c bioconda wisecondorx. A panel of normals was generated from 124 noncancer samples (Dataset S1). WiseCondorX was used to predict copy number changes in 1, 5, and 10 Mb bin sizes based on the panel of normals.

ichorCNA Analysis. ichorCNA (65) version 0.3.2 was downloaded from the GitHub repository https://github.com/broadinstitute/ichorCNA. Wig files were generated using readCounter with arguments–indow 5000000 –quality 30. CreatePanelOfNormals.R was used to generate a panel of normals (n = 124; Dataset S1). The "normal" initialization parameters selected were c(0.95, 0.99, 0.995, 0.999), the ploidy initialization parameter was 2, and only calls on the autosomal chromosomes were generated.

Data, Materials, and Software Availability. Anonymized genome sequencing data have been deposited in European Genome Archive (https://ega-archive.org/studies/EGAS00001007763) (88).

ACKNOWLEDGMENTS. We would also like to thank Mike Powell, Carey Priebe, Tim Coleman, Lucas Mentch, Juan M. Lavista Ferres, Ryan Hausen, David Donoho, Giles Hooker, Jason Klusowski, and Samir Hanash for helpful discussions. This study was supported by NIH grant R21NS113016 (C.B.), NIH grant RA37CA230400 (C.B.), NIH grant U01CA230691 (C.B. and N.P.), Oncology Core CA 06973 (K.W.K., N.P., and B.V.), The Virginia and D.K. Ludwig Fund for Cancer Research (C.B., C.D., K.W.K., N.P., and B.V.), Commonwealth Fund (N.P.), Thomas M Hohman Memorial Cancer Research Fund (C.B.), The Sol Goldman Sequencing Facility at Johns Hopkins (B.V.), The Conrad R. Hilton Foundation (K.W.K., N.P., and B.V.), Benjamin Baker Endowment 80049589 (Y.W. and C.D.), NIH grant Ovarian Cancer SPORE DRP 80057309 (C.D.), Swim Across America (C.D.), JHTV Innovation Grant (C.D.) Burroughs Wellcome Career Award for Medical Scientists (C.B.), Thomas M Hohman Memorial Cancer Research Fund (C.B.), NIH grant U01CA230691 (C.B.), 5P50CA062924-22 (A.M.L.), and NHMRC Investigator Grant APP1194970 (J.T.). C.D. would like to acknowledge funding from Swim Across America. S.P., H.X., Y.B., I.O., and J.T.V. would like to acknowledge generous funding from Collaborative Research: Transferable, Hierarchical, Expensive, Optimal, Robust, Interpretable Networks (NSF 20-540), NSF CAREER (NSF 17-537), DARPA's Lifelong Learning Machines program through contract FA8650- 18-2-7834, NSF award DMS-1921310, NIH award T32GM119998, and Microsoft Research. A.L. would like to acknowledge funding from the NSF Computing Innovation Fellowship (#2127309).

Author affiliations: [a]Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [b]Department of Oncology, the Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [c]The Ludwig Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [d]The Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [e]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287; [f]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218; [g]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218; [h]Department of Computer Science, Columbia University, New York, NY 10027; [i]Department of Applied Mathematics and Statistic, Johns Hopkins University, Baltimore, MD 21205; [j]The HHMI, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [k]Division of Personalized Oncology, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; [l]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC 3011, Australia; [m]Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, VIC 3010, Australia; [n]Department of Medical Oncology, Western Health, Melbourne, VIC 3000, Australia; [o]BioMedical Research Center, Pham Ngoc Thach University of Medicine, Ho Chi Minh City 72510, Vietnam; [p]Clinical Genetics Research Group, Saigon Precision Medicine Research Center, Ho Chi Minh City 72512, Vietnam; [q]Saigon Precision Medicine Research Center, Ho Chi Minh City 72512, Vietnam; [r]School of Biomedical Engineering, University of Technology, Sydney, NSW 2007, Australia; [s]Tâm Anh Research Institute, Ho Chi Minh City 721000, Vietnam; [t]Centre for Health Technologies, University of Technology, Sydney, NSW 2007, Australia; [u]School of Population Health, University of New South Wales, Kensington, NSW 2003, Australia; [v]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030; [w]Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX 77030; [x]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [y]Department of Medicine, Johns Hopkins Medical Institutes, Baltimore, MD 21205; [z]Department of Surgery, New York University Langone, New York City, NY 11209; [aa]Department of Gynecology and Obstetrics, Johns Hopkins Medical Institutions, Baltimore, MD 21287; [bb]Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA 15213; [cc]Department of Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, PA 15213; [dd]Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [ee]Department of Oncology, McGill University Health Centre, Montreal, QC H4A 3J1, Canada; and [ff]Division of Quantitative Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205

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

Reviewers: D.B., McGill University; and A.R.T., Institut Recherche en Cancérologie de Montpellier, INSERM U1194.

Competing interest statement: J.T. has a speaking, advisory, and consultancy with Haystack Oncology, Amgen, Novartis, AstraZeneca, Merck Serono, Merck Sharp & Dohme, Beigene, Pierre Fabre, Bristol Myers Squibb, Gilead, and Daiichi Sankyo. K.W.K., N.P., and B.V. are founders of Thrive Earlier Detection, an Exact Sciences Company. K.W.K., N.P., and C.D. are consultants to Thrive Earlier Detection. K.W.K. and N.P. are consultants to Neophore. K.W.K., N.P., B.V., and C.D. hold equity in Exact Sciences. K.W.K., N.P., and B.V. are founders of and own equity in Haystack Oncology & ManaT Bio. B.V. is a consultant to and holds equity in Catalio Capital Management. C.B. is a co-founder of OrisDx. C.B. and C.D. are co-founders of Belay Diagnostics. K.W.K., N.P., and B.V. hold equity in and are consultants to CAGE Pharma. The companies named above, as well as other companies, have licensed previously described technologies related to the work described in this paper from Johns Hopkins University. C.B., K.W.K., N.P., and B.V., and C.D. are inventors on some of these technologies. Licenses to these technologies are or will be associated with equity or royalty payments to the inventors as well as to Johns Hopkins University. Patent applications on the work described in this paper may be filed by Johns Hopkins University. The terms of all these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies. One of the reviewers of this study is at McGill alongside L.G., but there was no collaboration or association between them.

1. M. Hardt, B. Recht, *Patterns, Predictions, and Actions: Foundations of Machine Learning* (Princeton University Press, 2022).
2. B. Yu, K. Kumbier, Veridical data science. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 3920–3929 (2020).
3. S. M. Ghasem *et al.*, A novel method to guide biomarker combinations to optimize the sensitivity. bioRxiv [Preprint] (2024). https://www.biorxiv.org/content/10.1101/2024.04.12.589302v1 (Accessed 15 April 2024).
4. L. E. Dodd, M. S. Pepe, Partial AUC estimation and regression. *Biometrics* **59**, 614–623 (2003).
5. L. Devroye, L. Gyorfi, A. Krzyzak, G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Stat.* **22**, 1371–1385 (1994).
6. A. Farago, G. Lugosi, Strong universal consistency of neural network classifiers. *IEEE Trans. Inf. Theory* 39, 1146–1151 (1993).
7. G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9**, 2015–2033 (2008).
8. G. Lugosi, M. Pawlak, On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Trans. Inf. Theory* **40**, 475–481 (1994).
9. A. Niculescu-Mizil, R. Caruana, "Predicting good probabilities with supervised learning" in *Proceedings of the 22nd International Conference on Machine Learning* (ACM Press, 2005), pp. 625–632.
10. F. M. Ojeda *et al.*, Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Stat. Med.* **42**, 5451–5478 (2023).
11. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks. arXiv [Preprint] (2017). https://arxiv.org/abs/1706.04599 (Accessed 8 November 2024).
12. J. Nixon *et al.*, Measuring calibration in deep learning. arXiv [Preprint] (2019). https://arxiv.org/abs/1904.01685 (Accessed 8 November 2024).
13. J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* **10**, 61–74 (2000).
14. A. Brodeur, A. Dreber, F. Hoces de la Guardia, E. Miguel, Replication games: How to make reproducibility research more systematic. *Nature* **621**, 684–686 (2023).
15. P. Diaba-Nuhoho, M. Amponsah-Offeh, Reproducibility and research integrity: The role of scientists and institutions. *BMC Res. Notes* **14**, 451 (2021).
16. B. Recht, R. Roelofs, L. Schmidt, V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds. (Proceedings of Machine Learning Research (PMLR), 2019), 09–15 Jun 2019., pp. 5389–5400.
17. D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, V. Stodden, Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* **11**, 8–18 (2009).
18. P. I. Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses* (*Springer Series in Statistics*, Springer, ed. 3, 2004).
19. A. Sharma, A. Kumar, H. Daume, D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space" in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012), pp. 2160–2167.
20. E. Irajizad *et al.*, A blood-based metabolomic signature predictive of risk for pancreatic cancer. *Cell Rep. Med.* **4**, 9 (2023).
21. R. Perry *et al.*, mvlearn: Multiview machine learning in python. arXiv [Preprint] (2020). https://arxiv.org/abs/2005.11890 (Accessed 8 November 2024).
22. D. S. Micalizzi, L. V. Sequist, D. A. Haber, Deploying blood-based cancer screening. *Science* **383**, 368–370 (2024).
23. P. J. Bickel, K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics* (Prentice Hall, ed. 2, 2001).
24. R. Bellman, R. Kalaba, A mathematical theory of adaptive control processes. *Proc. Natl. Acad. Sci. U.S.A.* **45**, 1288–1290 (1959).
25. A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees. *Biometrics* **40**, 874 (1984).
26. A. Blum, A. Kalai, J. Langford, "Beating the hold-out: bounds for K-fold and progressive cross-validation" in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, (ACM, 1999), pp. 203–208.
27. I. Ghosal, Y. Zhou, G. Hooker, The infinitesimal jackknife and combinations of models. arXiv [Preprint] (2022). https://arxiv.org/abs/2209.00147 (Accessed 8 November 2024).
28. S. Athey, G. Imbens, Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7353–7360 (2016).
29. S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).
30. S. Athey, J. Tibshirani, S. Wager, Generalized random forests. *Ann. Statist.* **47**, 1148–1178 (2019).
31. S. R. Künzel, J. S. Sekhon, P. J. Bickel, B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 4156–4165 (2019).
32. E. Scornet, Random forests and kernel methods. *IEEE Trans. Inf. Theory* **62**, 1485–1500 (2016).
33. L. Mentch, G. Hooker, Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17**, 1–41 (2016).
34. D. Cevid, L. Michel, J. Näf, P. Bühlmann, N. Meinshausen, Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *J. Mach. Learn. Res.* **23**, 1–79 (2022).
35. G. J. Szekely, M. L. Rizzo, *The Energy of Data and Distance Correlation* (Chapman and Hall/CRC, ed. 1, 2023).
36. T. Coleman, W. Peng, L. Mentch, Scalable and efficient hypothesis testing with random forests. *J. Mach. Learn. Res.* **23**, 1–35 (2022).
37. D. S. Watson, M. N. Wright, Testing conditional independence in supervised learning algorithms. *Mach. Learn.* **110**, 2107–2129 (2021).
38. T. Hsing, S. Attoor, Dougherty relation between permutation-test p values and classifier error estimates. *Machine Learning* 52, 11–30 (2003).
39. I. Kim, A. Ramdas, A. Singh, L. Wasserman, Classification accuracy as a proxy for two-sample testing. *Ann. Stat.* **49,** 411–434 (2021).

40. D. Lopez-Paz, M. Oquab, Revisiting classifier two-sample tests. arXiv [Preprint] (2025). https://arxiv.org/abs/1610.06545 (Accessed 14 October 2024).
41. Y. M. D. Lo, D. S. C. Han, P. Jiang, R. W. K. Chiu, Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372,** eaaw3616 (2021).
42. S. C. Ding, Y. M. D. Lo, Cell-free DNA fragmentomics in liquid biopsy. *Diagnostics (Basel)* **12**, 978 (2022).
43. K. R. M. van der Meij *et al.*, TRIDENT-2: National implementation of genome-wide non-invasive prenatal testing as a first-tier screening test in the Netherlands. *Am. J. Hum. Genet.* **105**, 1091–1101 (2019).
44. A. Zviran *et al.*, Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat. Med.* **26**, 1114–1124 (2020).
45. P. Jiang *et al.*, Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov.* **10**, 664–673 (2020).
46. T. Moser, S. Kühberger, I. Lazzeri, G. Vlachos, E. Heitzer, Bridging biological cfDNA features and machine learning approaches. *Trends Genet.* **39**, 285–307 (2023).
47. J. C. M. Wan *et al.*, Liquid biopsies come of age: Towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).
48. J. D. Cohen *et al.*, Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
49. M. C. Liu *et al.*, Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
50. F. Mouliere *et al.*, High fragmentation characterizes tumour-derived circulating DNA. *PLoS One* **6**, e23418 (2011).
51. A. R. Thierry, S. El Messaoudi, P. B. Gahan, P. Anker, M. Stroun, Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev.* **35**, 347–376 (2016).
52. L. D. Maxim, R. Niebo, M. J. Utell, Screening tests: A review with examples. *Inhal. Toxicol.* **26**, 811–828 (2014).
53. A. R. Thierry, Circulating DNA fragmentomics and cancer screening. *Cell Genom.* **3**, 100242 (2023).
54. K. Sun *et al.*, Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res.* **29**, 418–427 (2019).
55. H. Markus *et al.*, Refined characterization of circulating tumor DNA through biological feature integration. *Sci. Rep.* **12**, 1928 (2022).
56. A. M. Lennon *et al.*, Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science* **369**, eabb9601 (2020).
57. A. Jamshidi *et al.*, Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell* **40**, 1537–1549.e12 (2022).
58. N. Hollmann *et al.*, Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).
59. F. Mitelman, Cancer: Chromosomal abnormalities. *eLS* 1–9 (2017).
60. H. Rajagopalan, C. Lengauer, Aneuploidy and cancer. *Nature* **432**, 338–341 (2004).
61. U. Ben-David, A. Amon, Context is everything: Aneuploidy in cancer. *Nat. Rev. Genet.* **21**, 44–62 (2020).
62. K. A. Knouse, T. Davoli, S. J. Elledge, A. Amon, Aneuploidy in cancer: Seq-ing answers to old questions. *Annu. Rev. Cancer Biol.* **1**, 335–354 (2017).
63. C. Douville *et al.*, 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).
64. L. Raman, A. Dheedene, M. De Smet, J. Van Dorpe, B. Menten, WisecondorX: Improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res.* **47**, 1605–1614 (2019).
65. V. A. Adalsteinsson *et al.*, Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
66. K. K. Budhraja *et al.*, Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer. *Sci. Transl. Med.* **15**, eabm6863 (2023).
67. Z. Zhou *et al.*, Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220982120 (2023).
68. H. Markus *et al.*, Analysis of recurrently protected genomic regions in cell-free DNA found in urine. *Sci. Transl. Med.* **13**, eaaz3088 (2021).
69. M. W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, J. Shendure, Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
70. P. Deininger, Alu elements: Know the SINEs. *Genome Biol.* **12**, 236 (2011).
71. P. Jiang *et al.*, Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1317–E1325 (2015).
72. H. R. Underhill *et al.*, Fragment length of circulating tumor DNA. *PLoS Genet.* **12**, e1006162 (2016).
73. N. Umetani *et al.*, Increased integrity of free circulating DNA in sera of patients with colorectal or periampullary cancer: Direct quantitative PCR for ALU repeats. *Clin. Chem.* **52**, 1062–1069 (2006).
74. F. Mouliere *et al.*, Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).
75. S. Cristiano *et al.*, Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
76. C. Douville *et al.*, Machine learning to detect the SINEs of cancer. *Sci. Transl. Med.* **16**, eadi3883 (2024).
77. G. Hooker, L. Mentch, S. Zhou, Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Stat. Comput.* **31**, 82 (2021).
78. C. Douville *et al.*, Assessing aneuploidy with repetitive element sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 4858–4863 (2020).
79. J. T. Leek *et al.*, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
80. E. W. Bridgeford *et al.*, Batch effects are causal effects: Applications in human connectomics. bioRxiv [Preprint] (2023). https://www.biorxiv.org/content/10.1101/2021.09.03.458920v4 (Accessed 8 November 2024).
81. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
82. Picard, A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. https://broadinstitute.github.io/picard/. Accessed 6 June 2025.
83. H. Li *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
84. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
85. ENCODE Project Consortium *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
86. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org. Accessed 6 June 2025.
87. K. Xiong, J. Ma, Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat. Commun.* **10**, 5069 (2019).
88. S. Curtis, *et al.*, Minimizing and quantifying uncertainty in AI-informed decisions- applications in medicine. *European Genome-Phenome Archive.* https://ega-archive.org/studies/EGAS00001007763. Deposited 29 July 2025.