Published online 27 October 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 4 **1**
https://doi.org/10.1093/nargab/lqab092

Ideafix: a decision tree-based method for the refinement of variants in FFPE DNA sequencing data

Maitena Tellaetxe-Abete [1,2,*], Borja Calvo[2] and Charles Lawrie[1,3,4,*]

[1]Molecular Oncology Group, Biodonostia Health Research Institute, Paseo Doctor Begiristain, 20014 Donostia/San Sebastian, Spain, [2]Intelligent Systems Group, Computer Science Faculty, University of the Basque Country, Paseo Manuel Lardizabal, 20018 Donostia/San Sebastian, Spain, [3]Ikerbasque, Basque Foundation for Science, 48009 Bilbao, Spain and [4]Radcliffe Department of Medicine, University of Oxford, Oxford, OX3 9BQ, UK

Received November 09, 2020; Revised September 14, 2021; Editorial Decision September 20, 2021; Accepted September 29, 2021

ABSTRACT

Increasingly, treatment decisions for cancer patients are being made from next-generation sequencing results generated from formalin-fixed and paraffin-embedded (FFPE) biopsies. However, this material is prone to sequence artefacts that cannot be easily identified. In order to address this issue, we designed a machine learning-based algorithm to identify these artefacts using data from >1 600 000 variants from 27 paired FFPE and fresh-frozen breast cancer samples. Using these data, we assembled a series of variant features and evaluated the classification performance of five machine learning algorithms. Using leave-one-sample-out cross-validation, we found that XGBoost (extreme gradient boosting) and random forest obtained AUC (area under the receiver operating characteristic curve) values >0.86. Performance was further tested using two independent datasets that resulted in AUC values of 0.96, whereas a comparison with previously published tools resulted in a maximum AUC value of 0.92. The most discriminating features were read pair orientation bias, genomic context and variant allele frequency. In summary, our results show a promising future for the use of these samples in molecular testing. We built the algorithm into an R package called Ideafix (DEAmination FIXing) that is freely available at https://github.com/mmaitenat/ideafix.

INTRODUCTION

Next-generation sequencing (NGS) is rapidly being adopted as the standard genetic diagnostic technique used in clinics. NGS methods can be used to diagnose hereditary disorders by germline mutation detection or increasingly to detect one or more somatic mutations for cancer diagnosis. Although fresh-frozen (FF) tissue is optimal for molecular techniques, it is rarely routinely available in clinical settings due to high storage costs and maintenance concerns. Instead, most routine pathology laboratories process and store tissue samples by formalin fixation and paraffin embedding (FFPE), as this preserves tissue and cellular morphology while allowing the samples to be stored at room temperature. However, due to the fixation process itself along with storage and extraction procedures, DNA obtained from FFPE samples suffers from significant levels of fragmentation, denaturation, cross-linking and chemical modifications, all of which can contribute to sequence artefacts. Even when such DNA is repaired using enzymatic treatment, the levels of sequence artefacts in NGS are much higher in FFPE than in FF material (1). The presence of artefacts is particularly challenging in cancer diagnosis, where the detection of single nucleotide changes can dictate treatment choice.

One of the most prevalent artefacts in FFPE material is deamination of cytosine residues to uracil, which, as a consequence of successive PCR amplification rounds, results in the C > T (or G > A antisense strand) variant (2,3). As a consequence, the general consensus is to discard C:G > T:A variants when the variant allele frequency (VAF) is <5% or <10%, as FFPE-associated artefacts have been reported to be present only at low levels (4–7). However, this approach is suboptimal as it prevents the detection of potentially clinically relevant low-frequency variants and limits the use of samples with low tumour content (8). Moreover, FFPE-associated artefacts have also been observed at frequencies >10% (9). There have been several attempts to address this issue that have focused mainly on sample preparation and repair, and include pretreatment of samples with uracil-DNA glycosylase (4), duplex sequencing (10), molecular tagging (11) or the preferential use of some DNA extraction kits over others (12). However, these approaches are far from perfect and often require additional costs and infrastructure beyond those available in most routine clinical diagnostic libraries. Moreover, these approaches cannot

*To whom correspondence should be addressed. Tel: +34 943 006071; Email: maitena.tellaeche@biodonostia.org
Correspondence may also be addressed to Charles Lawrie. Tel: +34 943 006138; Email: charles.lawrie@biodonostia.org

be applied retrospectively to the wealth of FFPE NGS data that already exist.

From a computational point of view, several strategies have been employed to deal with FFPE-associated sequence artefacts. Some authors have opted for rule-based systems for filtering the initial set of candidate variants. For example, Yost *et al.* controlled for formalin-induced deaminations by testing the allele frequency of each variant against the global nucleotide mismatch rate for that specific substitution using a binomial test, and then removing variants with allele frequencies not significantly different from those calculated (13). Similarly, Kerick *et al.* applied variant site-sequencing depth cut-offs in their filtering approach (14). The FIREVAT algorithm uses a series of filters including VAF, mapping quality and reference and alternate allele depths, to identify and remove artefacts. The cut-off values of these filters are found through a genetic algorithm and guided by mutational signatures (15).

In addition, some variant calling algorithms have been developed to identify suspicious variants in low-quality or FFPE sequencing data. The cisCall algorithm, for example, applies successive filters, such as the removal of groups of calls with extremely low VAFs and larger than expected groups of errors occurring close to each other, in order to identify different types of variants. All the filters are supported by statistical tests and internal controls (16). However, the performance is only evaluated on variants with a VAF >5%. Frampton *et al.* developed a Bayesian variant calling method that incorporates prior information about specific mutations related to a particular cancer, in addition to applying common quality filters (17). However, this tool is not publicly available and, in common with cisCall, the Frampton algorithm discards all variants with a VAF <5%. Another algorithm, LoLoPicker, estimates site-specific background error rates using a panel of control FFPE samples, which it then incorporates into a hypothesis test in order to identify variants (18), whereas the Pisces variant caller includes a module to minimize thermal damage and FFPE deaminations by recalibrating variant quality based on deviations from average mutation rate of each possible mutation type (19).

FFPE-associated deaminations, in common with artefacts such as oxidative damage of DNA, occur on only one strand of the original DNA template, resulting in an orientation bias between the first and second reads when paired-end sequencing is performed (20). This characteristic allows for the generation of an imbalance metric to quantify such damage. This metric is incorporated in several variant refinement tools, including GATK4, which contains the FilterByOrientationBias and LearnReadOrientationModel modules that filter out substitution artefacts that arise before the sequencing process on only one strand (21). However, the former has been kept as an experimental tool in favour of LearnReadOrientationModel by the GATK4 developers. This tool uses a Bayesian probabilistic model of single nucleotide substitutions occurring with orientation bias for each trinucleotide. Strand orientation bias is also used in the SOBDetector program, but instead of directly filtering calls, it calculates this bias score for every variant and adds the value to the variant calling (vcf) file for manual screening (22).

All of the algorithms developed to date are univariate methods. The Ideafix (deamination fixing) algorithm described in this paper is the first to use machine learning methods to tackle this problem. Using machine learning multivariate methods has the advantage over univariate methods that multiple descriptors can be tested simultaneously so that relationships between them can be exploited.

In this research, we assembled a collection of variant descriptors and evaluated the performance of five supervised learning algorithms for the classification of >1 600 000 variants, including both formalin-induced cytosine deamination artefacts and non-deamination variants, in order to arrive at our final model, Ideafix. Furthermore, we used an independent validation set to compare the performance of the Ideafix algorithm with three existing approaches and found our approach to generate better results.

MATERIALS AND METHODS

Building datasets

Data retrieval. Exome-sequencing data (fastq files) from 27 matched FFPE and FF samples were retrieved from the European Nucleotide Archive (accession number: SRP044740). The samples came from 13 different tumour specimens and were sequenced using Illumina technology. These data were used for training the classification models. In addition to this dataset, we used another two datasets as independent validation datasets. The first one consisted of whole exome sequencing (WES) data (fastq files) from matched FFPE and FF samples biopsied from two colon and two liver cancer samples and was downloaded from the European Genome-Phenome Archive (accession number: EGAD00001004066) (12). The second one corresponded to WES data from 16 matched FFPE and FF samples from gastro-oesophageal tumours and was downloaded from the European Nucleotide Archive (accession number: PRJEB44073) (23).

Preprocessing and filtering. Sequencing reads were first subjected to quality and adaptor sequence trimming using the Trimmomatic tool (24) and resulting reads were aligned to the hg19 reference genome using the BWA aligner (v0.7.17-r1188) (25). The Mutect2 (v4.0.8.1) variant calling algorithm was subsequently run in two different modes (26): first in tumour-only mode, i.e. individually on each FFPE and FF sample, and second in tumour/normal mode, using each FFPE sample as the tumour sample and its corresponding FF sample assigned as the normal sample. Single nucleotide polymorphisms were annotated against dbSNP database (v151) using SnpSift (v4.3t) (27,28) and variants falling adjacent to, or inside, homopolymer regions were annotated using VariantAnnotator-HomopolymerRun and the vcfpolyx utility from Jvarkit, respectively (26,29). vcf files were then filtered according to the following criteria: (1) to be a C:G > T:A single nucleotide variant; (2) to fulfil the PASS filter of the Mutect2 algorithm; (3) not to be supported only by unpaired reads; and (4) to display a VAF <30%. The latter threshold was chosen as we observed that >90% of deamination artefacts were lower than this value (Figure 1).

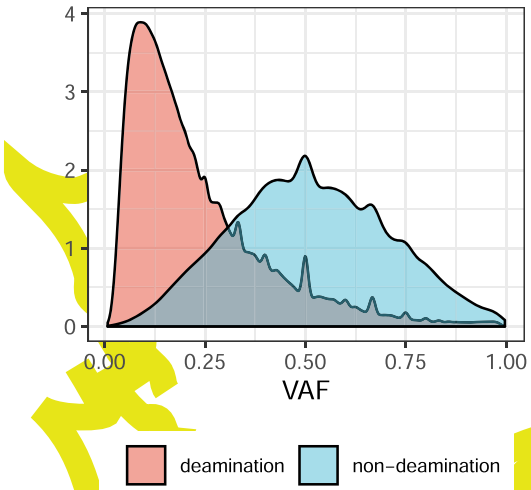

Figure 1. VAF distribution for deaminations and non-deaminations. While non-deaminations are symmetrically distributed around 0.5, deaminations are skewed towards low frequency values. Attending only to this feature, 0.3 is the threshold that would minimize the error in the identification and thus we limited the classification to variants below that threshold.

Variant labelling. Each variant was classified as either deamination if the variant was identified as a formalin-induced cytosine deamination or a non-deamination if the variant was germline, somatic or systematic sequencing error that had not been correctly filtered. Variants were considered to be non-deaminations if identified in both matched FF and FFPE samples by the Mutect2 algorithm running in tumour-only mode (30). This is so because any variant present in the FF tissue will very unlikely be in the same position as a formalin-induced deamination in the FFPE sample. Variants identified in the FFPE sample but not in the corresponding FF sample (tumour/normal mode) were classified as deaminations (Supplementary Figure S2).

In order to assess the deamination labelling approach we used, we compared with GATK's LearnReadOrientationModel tool (21). It should be noted that this tool looks for any read pair orientation bias and works in the absence of a paired FF sample. As variants that pass all quality filters and are present in more than one sample are more than likely non-deaminations, we only considered deaminations for this test. For the assessment, we specifically compared variants labelled by us as deaminations with C:G > T:A variants with VAFs <30% marked only with the *read_orientation_artifact* filtermark by LearnReadOrientationModel tool.

Descriptors. Predicting variables used for model training are described below and summarized in Table 1. These include a series of variables capturing formalin-induced cytosine deamination mechanisms that have already been described in the literature (4,5,20,31,32), and also others that we hypothesized could result in differences between artefactual deaminations and non-artefactual changes. These descriptors are grouped according to their concept or mechanism and are further described in the 'Feature Descriptors' section in the Supplementary Data.

Allele frequency. As deamination of cytosine to uracil only occurs post-mortem, these artefacts are not naturally replicated and are only amplified after being handled in the lab, such as in the process of library preparation for sequencing. This results in such artefacts being commonly observed at low frequencies (4,5), as opposed to real variants, which are not limited to specific frequencies. Hence, we incorporated VAF-related descriptors in our model. As VAF values reported by Mutect2 are not raw values but instead correspond to probabilistic guesses about the real frequency, we assembled five related variant frequency descriptors: the theoretical VAF reported by GATK, the total number of alternate and reference bases observed at site and the ratios between the number of alternate and reference bases observed and the total reads at site. These latter ratios were incorporated to aid in linear models non-capable of calculating such descriptors with potential to predict deaminations.

Read pair orientation bias. Sequencing by Illumina paired-end technology comprises a series of PCR amplification steps that result in the sequence of the template strand always being output in the first read (R1), and its reverse complement in the second read (R2). Mutations occurring *in vivo* are transmitted through replication mechanisms when dividing, so that in the case of C > T mutations, some child cells will also carry C > T mutations, while others will show G > A changes, depending on the strand used as the template for cell replication. Then, the template to be sequenced could contain any of C > T or G > A changes, and hence, the two changes would interchangeably be present in R1 and R2. In turn, deaminations occur in dead tissue, and due to the lack of active DNA replication mechanisms, they are only present on the DNA strand where the change happened and do not propagate to G > A changes. Hence, if C > T mutations occur in R1 and R2 only contains G > A complement changes, they are likely to be caused by deamination leading to orientation bias (Supplementary Figure S1). We defined a metric, FDeamC (fraction of cytosine deamination artefacts), based on the guanine oxidation measure FoxoG (fraction of guanine to 8-oxoguanine oxidation artefacts) defined by Costello *et al.* (33), as follows:

$$\text{FDeamC} = \begin{cases} \dfrac{n_a^{F1R2}}{n_a^{F1R2} + n_a^{F2R1}}, & \text{if} \quad C > T \\ \dfrac{n_a^{F2R1}}{n_a^{F1R2} + n_a^{F2R1}}, & \text{if} \quad G > A \end{cases} \tag{1}$$

where

n_a^{F1R2} = number of read pairs supporting the alternate allele in which the first read in the pair aligns to the forward strand (F1), and the second read to the reverse strand (R2)

n_a^{F2R1} = number of read pairs supporting the alternate allele in which the first read in the pair aligns to the reverse strand (R1), and the second read to the forward strand (F2)

Table 1. Variable descriptors included in the models and their source, grouped by their underlying concept or mechanism

Concept	Descriptor	Source
Allele frequency	VAF	Present in vcf file
	Number of reference bases	Present in vcf file
	Normalized number of reference bases	Present in vcf file
	Number of alternate bases	Present in vcf file
	Normalized number of alternate bases	Present in vcf file
Read pair orientation bias	FDeamC	Calculation shown in the 'Feature Descriptors' section in the Supplementary Data
	SOB	Calculation shown in the 'Feature Descriptors' section in the Supplementary Data
Genomic context and annotation in population databases	Reference allele	Extraction from reference genome using samtools
	Base one position before	Extraction from reference genome using samtools
	Base two positions before	Extraction from reference genome using samtools
	Dinucleotide before	Extraction from reference genome using samtools
	Base one position after	Extraction from reference genome using samtools
	Base two positions after	Extraction from reference genome using samtools
	Dinucleotide after	Extraction from reference genome using samtools
	Homopolymer length	Extraction from reference genome using VariantAnnotator-HomopolymerRun and vcfpolyx utility from Jvarkit
	isSNP	Presence checked in dbSNP using SnpSift
Fragment length	Fragment length	Present in vcf file
Median distance from end of read	Median position from read end	Present in vcf file
	Normalized median position from read end	Present in vcf file
Base and mapping quality	Base quality	Present in vcf file
	Base quality fraction	Present in vcf file
	Mapping quality	Present in vcf file
Strand bias (SB)	SB-GATK	Calculation shown in the 'Feature Descriptors' section in the Supplementary Data
	SB-GUO	Calculation shown in the 'Feature Descriptors' section in the Supplementary Data

We also used a similar but not equivalent metric to account for this bias, which was defined by Diossy *et al.* (22):

$$\mathrm{SOB} = \frac{n_a^{\mathrm{F1R2}} - n_a^{\mathrm{F2R1}}}{n_a^{\mathrm{F1R2}} + n_a^{\mathrm{F2R1}}} \quad (2)$$

FDeamC is defined in the [0–1] range, whereas SOB score lies in the [−1, 1] interval. It follows from both formulas that formalin-induced deaminations will show in both cases extreme values (close to 0 or 1 for FDeamC, 1 or −1 for SOB), whereas naturally occurring variants will have values close to 0.5 and 0 for FDeamC and SOB, respectively.

Genomic context and annotation in population databases. We hypothesized that cytosine genomic context may be affecting deamination probability and therefore created a descriptor consisting of a pentanucleotide sequence with two flanking bases on either side of the identified variant. We also included the length of the homopolymer surrounding the variant locus (0 for non-homopolymer, non-zero positive number when flanked by a homopolymer) and whether the mutation had been annotated as being a polymorphism in dbSNP database.

Median distance from end of read. Deaminations have been shown to be enriched at the ends of molecules, due to an increased sensitivity to deaminate of overhanging ends (31,32). In contrast, true variants should be distributed uniformly along the whole piece of DNA irrespective of context. In order to account for this factor, a descriptor showing the median distance from the end of the read of all the reads supporting the alternate allele, and its normalized version calculated by dividing this number by the fragment length, was added to the model. It should be noted that end of read does not always correspond to end of original molecule as DNA is subjected to artificial fragmentation ahead of being sequenced.

Base and mapping quality. In order to account for possible variability between FFPE and fresh samples, we incorporated the median Phred base quality score, the median read mapping quality value and the ratio between median Phred base quality of the alternate allele and median Phred base quality of the reference allele.

Fragment length. In order to account for the possibility that artefactual deaminations are more common in shorter fragments than longer ones, as a result of formalin producing cross-linking that causes DNA fragmentation and deaminations nearby, we included a variable that contains

the median length of the fragments carrying the variant alternate allele.

Strand bias. SB has been previously described as being associated with sequencing data postprocessing and genome context (34). We left it to the learning algorithm to determine whether this anomaly is also associated with formalin-induced artefacts. Different mathematical definitions have been described for capturing this bias (21,35). We considered two of them in our model, SB-GATK and SB-GUO, which are further described in the 'Feature Descriptors' section in the Supplementary Data.

Damage assessment

We assessed DNA damage using four distinct indicators: (1) median library fragment length; (2) mean library duplication level; (3) coefficient of variation of library depth of coverage; and (4) GIV score of C:G > T:A changes. The median fragment length value provides a measure of the DNA fragmentation level. When DNA is very fragmented, the amount of amplifiable template is small, and therefore, duplicate DNA fragments are created. When these fragments are sequenced, differences in their relative quantities lead to uneven depth of coverage. The GIV score is a measure of global read pair orientation bias in a given library and scores >2 are indicative of extensive damage. For the case of C:G > T:A changes, we calculate it as follows:

$$\mathrm{GIV} = \frac{\frac{v^{R1}_{C>T} + v^{R2}_{G>A}}{b^{R1}_{C} + b^{R2}_{G}}}{\frac{v^{R1}_{G>A} + v^{R2}_{C>T}}{b^{R1}_{G} + b^{R2}_{C}}} \quad (3)$$

where

$v^{R1}_{C>T}$ = number of C > T variants in R1
$v^{R2}_{G>A}$ = number of G > A variants in R2
$v^{R2}_{C>T}$ = number of C > T variants in R2
$v^{R1}_{G>A}$ = number of G > A variants in R1
b^{R1}_{C} = number of C nucleotides in R1
b^{R2}_{G} = number of G nucleotides in R2
b^{R1}_{G} = number of G nucleotides in R1
b^{R2}_{C} = number of C nucleotides in R2

Thus, as a general rule, the more damaged the sample, the more variable the depth of coverage, the lower the median fragment length and the greater the library duplication level and the C:G > T:A GIV score. Fragment length data were obtained directly from the bam files, depth of coverage information using samtools depth (30) and GIV scores using Damage-Estimator (20). Library duplication levels were calculated using FastQC and the mean value between R1 and R2 was considered (36).

Experimental design

Learning algorithms. Five well-established supervised learning algorithms were evaluated for the model: logistic regression (logReg), random forest (RF), extreme gradient boosting (XGBoost), naive Bayes (NB) and artificial neural network (ANN). The selection was made so that algorithms belonging to several families were considered: regression algorithms (logReg), Bayesian networks (NB), artificial neural network algorithms (ANN) and decision tree algorithms (XGBoost, RF). All the analyses were carried out in R and details regarding the packages and parameters used are described in Supplementary Table S1. It should be noted that ANN and XGBoost are designed to work only with numeric data, so for these two algorithms, categorical variables were one-hot feature encoded; i.e. each *n*-valued categorical variable was converted into *n* binary variables, one for each category.

Evaluation strategy. In order to choose the best classification algorithm as well as measure the quality of the final model, we adopted a leave-one-sample-out cross-validation performance assessment scheme. For each sample, this strategy divides the data into two parts: a test set that holds the observations corresponding to that sample and a training set that includes the variants in the rest of the samples.

To evaluate the performance of the classifiers, we used the value of the area under the receiver operating characteristic (ROC) curve (AUC), as this metric is independent of the classification threshold and is not so much affected by possible class imbalances (37).

Three different performance estimators were calculated: resubstitution (RE), cross-validation (CVE) and sample-out (SOE) AUC estimators. RE was obtained by both training and evaluating a classifier using the same dataset, i.e. each of the training sets. CVE was obtained by following 10 times 10-fold stratified cross-validation scheme within each training set. SOE was calculated by fitting the model to the training data and evaluating it in the test set. As the dataset contained data from 27 different samples, 27 estimations of this group of three estimators were obtained. Whereas looking at the gap between RE and CVE helps us find out whether there is room for a better model fit, CVE and SOE by themselves provide us with an estimation of the performance of the classifier in unseen instances, i.e. its generalization ability. Among these two, CVE is the most positively biased one as both training and test datasets share variants from the same samples. Additionally, SOE, for which the evaluation is made in a completely new sample, gives us a sample-aware performance estimation. Supplementary Figure S3 shows a graphical summary of this evaluation scheme.

To further test the model's generalization performance, the final model was evaluated in two independent datasets (colon and liver dataset and gastro-oesophageal dataset) corresponding to new subjects and different cancer types again by means of the AUC (Supplementary Figure S4).

Feature analysis. We analysed discriminatory power of the features in two different ways.

First, we carried out a univariate analysis by calculating the AUC value for continuous valued features and the normalized mutual information (NMI) for categorical descriptors (38). In brief, for calculating the AUC value of each continuous feature, we ranked the observations according to their value for that feature. Then, we set a threshold equal to the first value, classified the samples with values below that threshold in one class and those with values above that

threshold in the other class, and calculated the true positive and false positive rates of this classification. We repeated the same procedure using all the feature values as thresholds for the classification. After considering all the possible thresholds, we depicted the ROC curve and calculated the area underneath. It should be taken into account that the AUC is an effective metric for individual feature classification performance only when that feature follows a unimodal distribution for each of the classes. For more complex distributions such as multimodal ones, it may provide an overly pessimistic value. In the case of discrete variables, this method cannot be applied. As an alternative, we calculated the NMI value between each of those features and the class vector, quantifying this way the amount of information obtained about the deamination status by knowing that descriptor. The normalization we adopted here works by dividing the mutual information value by the mean value of the class and feature entropies. The choice of a normalized version of the mutual information over the regular mutual information was motivated by the fact that the upper bound of the latter depends on the number of possible values for the variable. As this number varies between features, so does the range of possible values they can take, making any comparison between features difficult.

After we had trained the classification models, we did a second predictive power assessment of the features through the best models' built-in feature importance metrics. These were decision tree-based models, with feature importance assessment being based on the improvement in the classification after using each feature as a decision node in the trees (39).

Performance comparison with other tools. The performance of the final model was compared with one of the state-of-the-art techniques, SOBDetector (22), and two common variant refinement practices: filtering variants by depth of coverage and filtering by VAF. For the depth of coverage and VAF filtering, we depicted the ROC curves and calculated their AUC values in the validation set following the same procedure we employed for the univariate feature analysis. SOBDetector was run using default parameters and the ROC curve and AUC value of the posterior probabilities that the variant is an artefact were calculated.

RESULTS

Sample damage analysis

We computed four FFPE damage indicators: the coefficient of variation of depth of coverage, the mean library duplication level, the median fragment length and the GIV score of C:G > T:A changes.

For the breast dataset, 20 out of the 27 (74%) FFPE samples displayed GIV scores >2 for C:G > T:A changes (median value = 2.98), whereas all the FF samples had values <1 (median value = 0.975). In addition, FF samples had larger median fragment sizes (median fragment size = 199 bp) and lower library duplication levels (median mean duplication level = 22%) compared to FFPE samples (median fragment size = 118 bp, median mean duplication level = 85.7%). The variability of the fragment size was larger in FFPE samples (range 100–138 bp) than FF samples (range 189–208) and duplication levels were also more variable in FFPE samples (range 26.5–94.5%) than in FF samples (range 19.7–28.6%). The coefficient of variation of coverage depth for FF samples had a median value of 1.08 (range 1.03–1.27), while for FFPE samples the median value was 1.33 (range 1.15–3.82). Likewise, median depth of coverage values was in the range 37–48X for FF samples, while FFPE samples showed diverse median values, ranging from 0X, in the case of samples with a great number of non-covered exons, to 172X (Table 2 and Figure 2). In summary, while FF samples showed no damage signs, FFPE samples were damaged in varying levels.

Regarding the colon and liver dataset, FFPE samples showed much less damage than the FFPE samples in the breast dataset (median GIV score = 0.89, median fragment length = 152, median mean duplication level = 18.6 and median coefficient of variation value of depth of coverage = 1.10, while for the breast dataset, median GIV score = 2.98, median fragment length = 118, median mean duplication level = 85.7 and median coefficient of variation value of depth of coverage = 1.33). Moreover, the four damage indicators showed close values for FFPE and FF samples (Table 2 and Supplementary Figure S7). Similarly, the gastro-oesophageal dataset also showed more limited damage levels than the breast dataset for all the indicators except for the median fragment length (median GIV score = 0.887, median fragment length = 114, median mean duplication level = 42.5 and median coefficient of variation value of depth of coverage = 0.823), and again for indicators other than the fragment length, the values for FF and FFPE were close to each other (Table 2 and Supplementary Figure S8).

Variant labelling

Identified deamination and non-deamination variants for the breast dataset averaged 62 154 and 388 per sample, with a total of 1 594 078 deamination variants and 10 102 non-deamination variants (Table 3 and Supplementary Table S3). Deamination to non-deamination per sample ratios ranged between 64.2 and 263, with an average value of 172, showing that this is a highly unbalanced problem and that this imbalance varies significantly between samples (Figure 3).

A comparison in the labelling approach that we took with the LearnReadOrientationModel tool of the GATK program demonstrated that we identified more deaminations (1 594 078 compared to 1 510 878). From these, 1 388 914 deaminations were commonly identified by both methods (87% concordant). Hence, 121 964 and 205 164 deaminations were non-overlapping to LearnReadOrientationModel and our approach, respectively (Supplementary Figure S9). All of the 121 964 deamination variants called only by LearnReadOrientationModel were present in the vcf files we obtained from tumour-only mode variant calling on the FFPE samples and passed the Mutect2 PASS filtering step, but were not present in the files generated by tumour/normal variant calling in the FFPE/FF sample pairs. Specifically, 50 813 (44%) of those variants had been called in tumour/normal mode but were not fulfilling PASS filtering and other 7007 additional variants (6%) had also been called in tumour-only mode variant calling on the

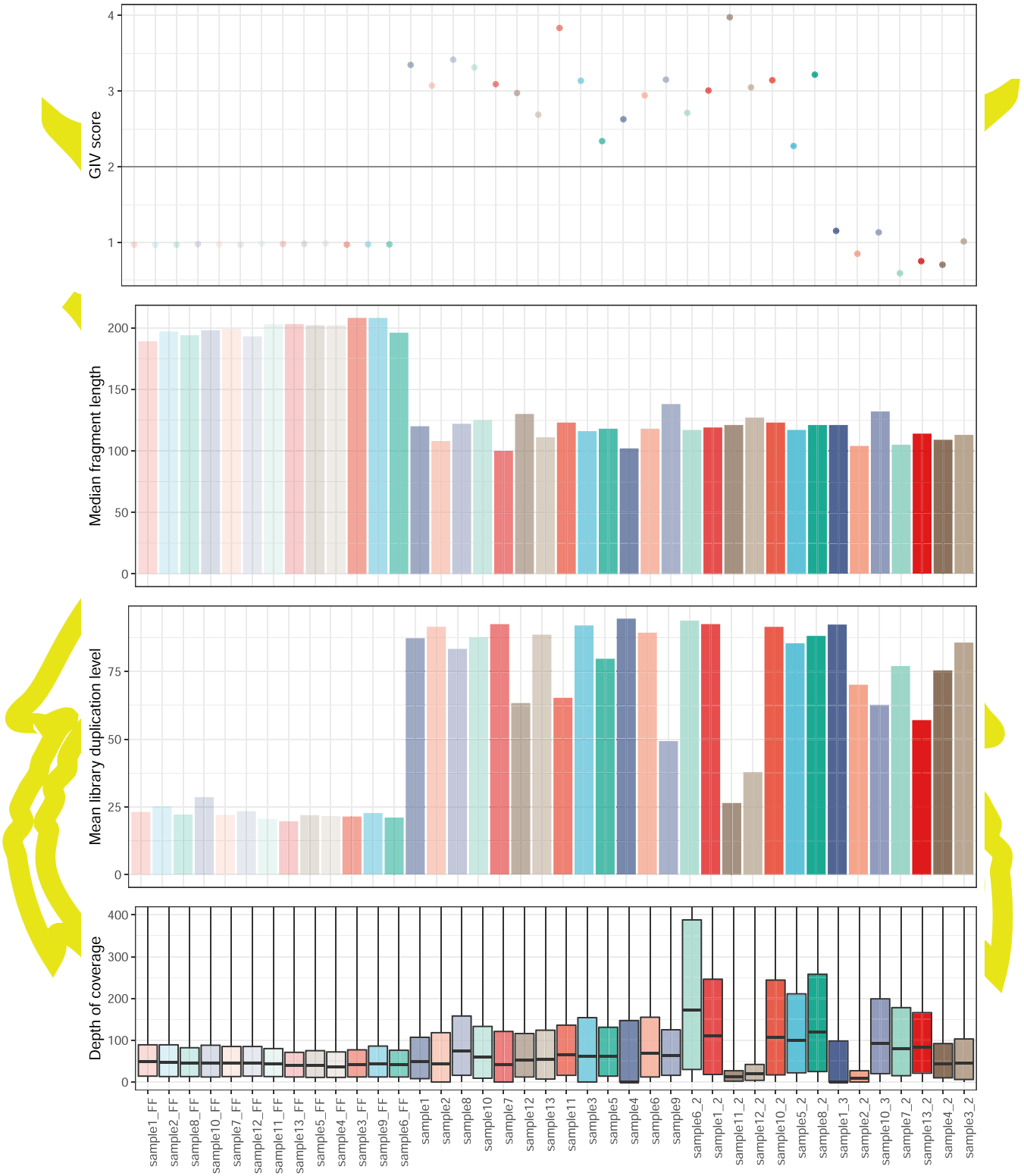

Figure 2. Damage analysis on breast dataset. All FF samples show GIV score values <2, suggesting an absence of C:G > T:A orientation bias in the libraries, while the majority of FFPE samples present scores above this threshold, indicating extensive presence of this bias. Differences are also observed regarding median fragment length and mean library duplication level, where FFPE samples present smaller fragments and larger duplication levels. Finally, depth of coverage is also rather uniform for FF samples, while it shows greater variability for FFPE specimens. Note that depth of coverage subfigure is zoomed to the boxes. Note also that GIV scores were calculated using all variants and not only those with a VAF <30%.

Table 2. Median and range values of four FFPE damage indicators calculated for both FF and FFPE samples in all the datasets

		FF	FFPE
Coverage depth coefficient of variation (X)	Breast dataset	1.08 (1.03–1.27)	1.33 (1.15–3.82)
	Colon and liver dataset	0.70 (0.69–0.71)	1.10 (1.08–1.15)
	Gastro-oesophageal dataset	0.698 (0.671–0.877)	0.823 (0.736–0.989)
Mean library duplication level (%)	Breast dataset	22 (19.7–28.6)	85.7 (26.5–94.5)
	Colon and liver dataset	20.3 (18.3–23.6)	18.6 (14.7–25.8)
	Gastro-oesophageal dataset	38.3 (30.4–43.3)	42.5 (31.8–69.4)
Median fragment length (bp)	Breast dataset	199 (189–208)	118 (100–138)
	Colon and liver dataset	163 (160–166)	152 (146–153)
	Gastro-oesophageal dataset	160 (116–176)	114 (86–137)
GIV score of C:G > T:A changes	Breast dataset	0.975 (0.968–0.986)	2.98 (0.594–3.98)
	Colon and liver dataset	1.02 (1.01–1.02)	0.89 (0.85–0.92)
	Gastro-oesophageal dataset	0.994 (0.887–1.01)	0.887 (0.560–0.996)

FFPE and FF samples show contrasting values in the breast dataset; instead, much smaller differences are observed in the colon and liver and gastro-oesophageal datasets, especially for the GIV score and mean library duplication level. Numbers in the table suggest larger damage levels in the breast samples than in the other two datasets.

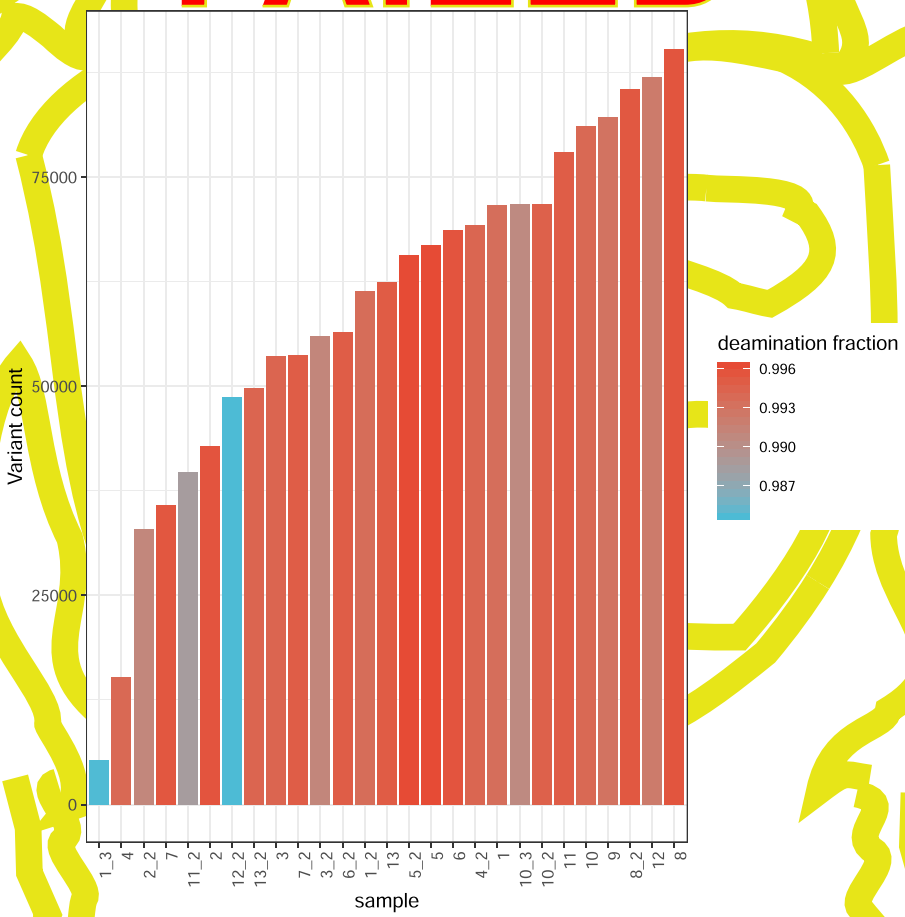

Figure 3. Per sample number of variants in the breast dataset. The colours indicate the deamination fractions of all the variants of that sample. As can be seen, the number of variants is very variable between samples, ranging from 5305 to 90 377. The fraction of deaminations on these variants is always >0.985, showing that this is a highly unbalanced problem. At the same time, this also means that the number of both deamination and non-deamination variants is very variable between samples, ranging from 5224 to 89 972 in the former case, and from 81 to 746 in the latter case. Samples with the same number before the underscore in their name correspond to the same specimen.

Table 3. Dataset summary: number of samples and C:G > T:A variants with a VAF <30% (deaminations and non-deaminations) in each dataset

Attribute	Breast dataset	Colon and liver dataset	Gastro-oesophageal dataset
No. of samples	27	4	16
No. of total deaminations	1 594 078	9994	446 878
No. of total non-deaminations	10 102	3157	18 793

FF samples. This suggests that these variants could not confidently be called differentially to the FF sample, and hence, they were highly likely labelled as deaminations by LearnReadOrientationBias incorrectly.

We hypothesize that the 205 164 variants that were not picked up by LearnReadOrientationModel did not display read orientation bias, which is the filter used by this algorithm.

Univariate feature analysis

Table 4 shows the AUC or NMI value for each variable included in the model. The SOB (AUC = 0.91) and FDeamC (AUC = 0.86) variables showed the most promising predictive capacity, followed by allele frequency-related features and the count of reference allele bases, whose AUC values ranged around 0.79 and 0.71, respectively. SB-GATK had an AUC of 0.74, whereas SB-GUO, the number of alternate allele bases and those features related to fragment length, base and mapping quality and position on read had AUC values between 0.51 and 0.58. For discrete variables, isSNP showed the largest classification power (NMI = 0.33), whereas the rest of descriptors tested had more lower values in the range of 1.4e-6 to 7.7e-4.

An important point to bear in mind with regard to the classification power of isSNP feature is that in our dataset it is confounded with the classification label. For this reason, and in order to avoid a confounding effect that would mislead classification performance evaluation, as argued in the 'Discussion' section, we decided to remove the isSNP descriptor from the datasets.

Model evaluation

The median CVE and SOE values of the models tested ranged from 0.81 to 0.89 (Table 5), with XGBoost (CVE = 0.89, SOE = 0.87) and RF (CVE = 0.88, SOE = 0.87) models performing best. These two algorithms were followed by ANN and logReg, and NB yielded the worst classification results.

In addition, the distribution of the three estimators across samples showed that RE and CVE had little variability between cases, as expected since a similar set of samples was used in all cases (Figure 4). Further analysis of the data looking at each fold shows a lack of variation between folds too, except for a few outliers in ANN (Figure 5). In contrast, SOE varied within a broader range: for all the algorithms, a difference >0.2 was observed from the sample with the lowest AUC to the one with the largest AUC.

Notably, for all the algorithms, there were samples in which the SOE was larger than the CVE, and even larger

Table 4. NMI or AUC values for the variables included in the model, ordered in a decreasing importance order

Feature	Type of variable	NMI [0–1]	AUC [0.5–1]
SOB	Real valued	-	0.91
FDeamC	Real valued	-	0.86
Normalized number of reference bases	Real valued	-	0.79
Normalized number of alternate bases	Real valued	-	0.79
VAF	Real valued	-	0.78
SB-GATK	Real valued	-	0.74
Number of reference bases	Real valued	-	0.71
SB-GUO	Real valued	-	0.58
Fragment length	Real valued	-	0.54
Base quality	Real valued	-	0.54
Median position from read end	Real valued	-	0.54
Number of alternate bases	Real valued	-	0.54
Base quality fraction	Real valued	-	0.53
Normalized median position from read end	Real valued	-	0.52
Mapping quality	Real valued	-	0.51
isSNP	Categorical	0.33	-
Base one position before	Categorical	7.7e−04	-
Base one position after	Categorical	7.1e−04	-
Dinucleotide before	Categorical	4.4e−04	-
Dinucleotide after	Categorical	4.3e−04	-
Homopolymer length	Integer	4.2e−04	-
Base two positions after	Categorical	8.1e−05	-
Base two positions before	Categorical	8.1e−05	-
Reference allele	Categorical	1.4e−06	-

See the 'Descriptors' section in main text and the 'Feature Descriptors' section in the Supplementary Data for further details on the descriptors.

Table 5. Median AUC CVE, RE and SOE estimations for each algorithm

	CVE	RE	SOE
logReg	0.845	0.847	0.828
NB	0.823	0.824	0.808
ANN	0.876	0.894	0.862
RF	0.884	1.000	0.867
XGBoost	0.894	0.943	0.873

RF and XGBoost show the largest CVE, RE and SOE values, followed by ANN, logReg and NB, in that order. In XGBoost, and especially RF, RE is well above CVE and SOE. logReg, NB and ANN show closer or non-existent gaps between the estimators, especially for CVE and RE. Median SOE is in all five algorithms the lowest estimator among the three.

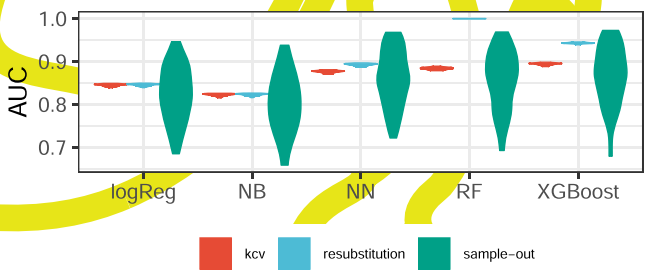

Figure 4. RE, CVE and SOE values for each algorithm. Note that for CVE, the median AUC of all folds for each sample is shown. This means that, first, these densities are proportional to variability across samples and, second, any differences between folds cannot be observed from this representation (Figure 5 gives this information). CVE and RE show very low variability; this is indeed an expected behaviour as a similar set of samples was used for calculating each of those values. For all the models, there is a difference >0.2 between the lowest and largest SOE values, showing that performance is very dependent on sample particularities.

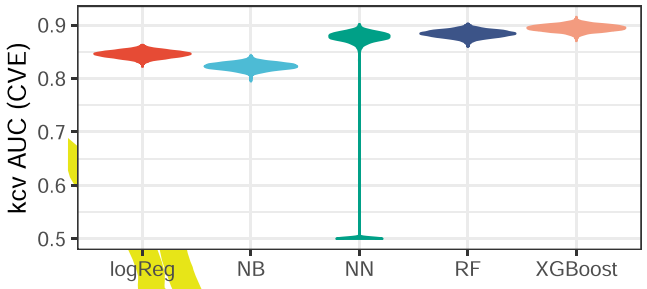

Figure 5. Ten-fold cross-validation AUC values, i.e. CVE values, for each algorithm. XGBoost, RF and ANN show the largest values, followed by logReg and NB, in that order. ANN density is more spread than the other four due to the presence of certain outliers; this indicates that the performance of this model is less stable across instances than the rest.

than RE, despite median SOE across samples being for the five algorithms the lowest estimation value among the three.

Model refinement

In order to test the choice of model parameters, we looked at differences between CVE and RE. In the case of logReg, NB and ANN, no substantial gap existed between CVE and RE (Figure 4 and Table 5). However, for both XGBoost (median RE = 0.94, median CVE = 0.89, median SOE = 0.87) and RF (median RE = 1, median CVE = 0.88, median SOE = 0.87), RE was well above CVE and SOE. In order to identify better hyperparameter configurations that could allow a better adjustment to the data in the case of these two latter algorithms, especially in RF (Figure 6), we carried out random grid searching for the number of learnt trees, the number of descriptors used to learn each tree and the maximum depth of the trees. No clear performance improvement was observed, suggesting a lack of room for further refinement (data not shown).

Final model and validation on external datasets

To further validate the approach, the best algorithms (XGBoost and RF) were used to create models with the data from the breast dataset and were evaluated on two independent validation datasets. The first validation dataset consisted of four FFPE–FF paired colon and liver tumour samples and contained a total of 9994 deaminations and 3157 non-deaminations with a VAF <30%, labelled following the same procedure. While still showing an imbalance between the classes (~3:1), this was much less pronounced than that in the breast cancer dataset. The second dataset consisted of 16 FFPE–FF paired gastro-oesophageal cancer samples with a total of 446 878 deaminations and 18 793 non-deaminations with a VAF <30% and again labelled the same way. The imbalance in this dataset was more pronounced than that in the colon and liver dataset (~24:1) but still less than that in the breast dataset (Table 3 and Supplementary Table S3). Resulting AUC values for the colon and liver dataset were 0.9643 and 0.9541 for XGBoost and RF, respectively; in the gastro-oesophageal dataset, AUC values for XGBoost and RF were 0.9639 and 0.9567, respectively. By way of illustration, we also calculated additional performance metrics that are dependent on a classification threshold. For doing so, we chose the classification threshold that would minimize the number of false positives while ensuring a minimum value for the sensitivity of 0.99. This was done using the data on the training set. It is important to note here that these results that are dependent on a threshold cannot be extrapolated to other scenarios since the criterion for selecting the threshold may not be suitable for them. The metrics for this threshold, considering the deamination as the positive class, are depicted in Table 6. Except for the negative predictive value (NPV) and, to a lesser extent, the specificity, values for both models are very similar. Of note is the fact that small improvements in the positive predictive value (PPV) (in the colon and liver dataset, the values were 0.9419 for XGBoost and 0.9494 for RF; in the gastro-oesophageal dataset, values were 0.9886 and 0.9903, respectively) are highly detrimental to the NPV (in the colon and liver dataset, specificity values were 0.8810 for XGBoost and 0.7965 for RF; in the gastro-oesophageal dataset, values were 0.7471 and 0.5776, respectively).

Performance comparison with state-of-the-art techniques

We compared the performance of our model with other existing approaches using the validation sets described earlier. Both XGBoost (AUC = 0.9643 for the colon and liver dataset, AUC = 0.9639 for the gastro-oesophageal dataset) and RF models (AUC = 0.9541 for the colon and liver dataset, AUC = 0.9567 for the gastro-oesophageal dataset) outperformed the rest of the models and showed almost identical ROC curves (Figure 7 and Table 7). In the colon and liver dataset, SOBDetector and filtering by VAF produced close to each other results, with AUC values slightly above 0.92. However, in the gastro-oesophageal dataset, while filtering by VAF kept its performance, SOBDetector fell to an AUC value of 0.7745. Filtering by depth of coverage was the worst solution (AUC = 0.688 for the colon and liver dataset, AUC = 0.7588 for the gastro-oesophageal dataset).

Feature analysis by built-in metrics

Features ranked by their predictive capacity by both RF and XGBoost built-in feature importance metrics are presented in Table 8. An important thing to bear in mind here is the way each algorithm manages correlated features: while RF gives similar importance to correlated features (for instance, the scaled importance of FDeamC and SOB was 1 and 0.71, respectively; the scaled importance of the VAF, normalized number of alternate bases and normalized number of reference bases was 0.59, 0.54 and 0.53, respectively), XGBoost favours one over the rest (FDeamC and SOB had scaled importance values of 1 and 0.12, respectively; VAF, normalized number of alternate bases and normalized number of reference bases had values of 0.051, 0.63 and 0, respectively). Thus, for this feature analysis we only focused on the ranking provided by the RF algorithm, as in the case of XGBoost we ran the risk of underrating a variable when it was simply correlated with another one.

In this RF ranking, FDeamC showed the highest classification power, far above the second one, which was SOB (scaled importance score of 0.71). The median variant position and the genomic context were quite relevant too (score

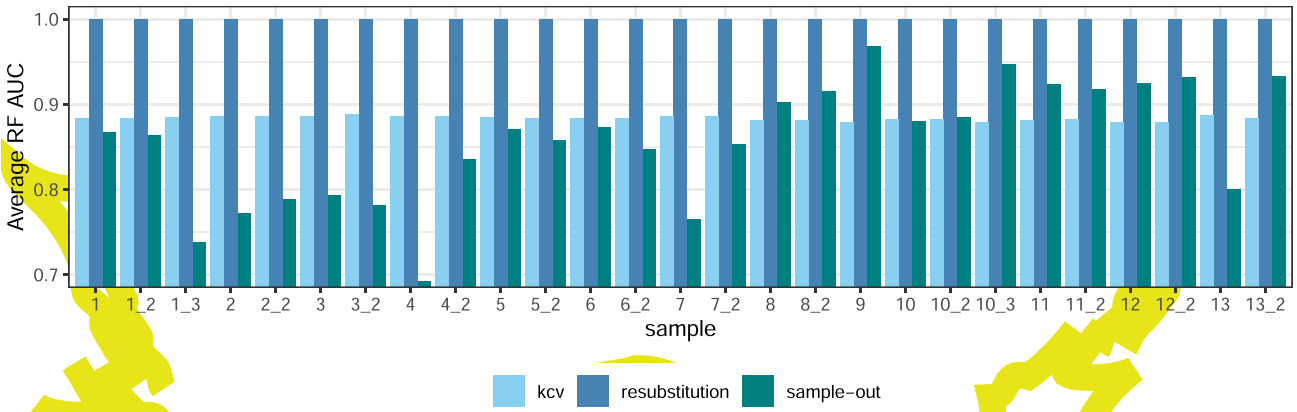

Figure 6. RE, CVE and SOE values for RF. RE is equal to 1 in all cases and a gap exists between this and the rest of estimators, suggesting room for model improvement. For the majority of the samples ($n = 17$), CVE is above SOE; still there are cases in which SOE is larger. This is surprising because CVE being larger than SOE was expected as instances from the same sample are shared between the training and test sets in the former. Samples with the same number before the underscore in their name correspond to the same specimen.

Table 6. Performance of the final XGBoost and RF models on the validation sets taking the deamination as the positive class

Validation set	Algorithm	AUC	F_1	Accuracy	Sensitivity	Specificity	PPV	NPV
Colon and liver	XGBoost	0.9643	0.9535	0.9284	0.9654	0.8115	0.9419	0.8810
	RF	0.9541	0.9406	0.9106	0.9320	0.8429	0.9494	0.7965 ¿¿
Gastro-oesophageal	XGBoost	0.9639	0.9891	0.9791	0.9896	0.7295	0.9886	0.7471
	RF	0.9567	0.9832	0.9680	0.9763	0.7715	0.9903	0.5776 ¿¿

Both models show AUC values close to 1 in the two datasets and almost identical values for the performance metrics that depend on a threshold, except for the NPV. Besides, the NPV itself and the specificity are the metrics with the lowest values, which suggests that, with this threshold, the major limitation is the detection of the true negatives, i.e. the non-deaminations.

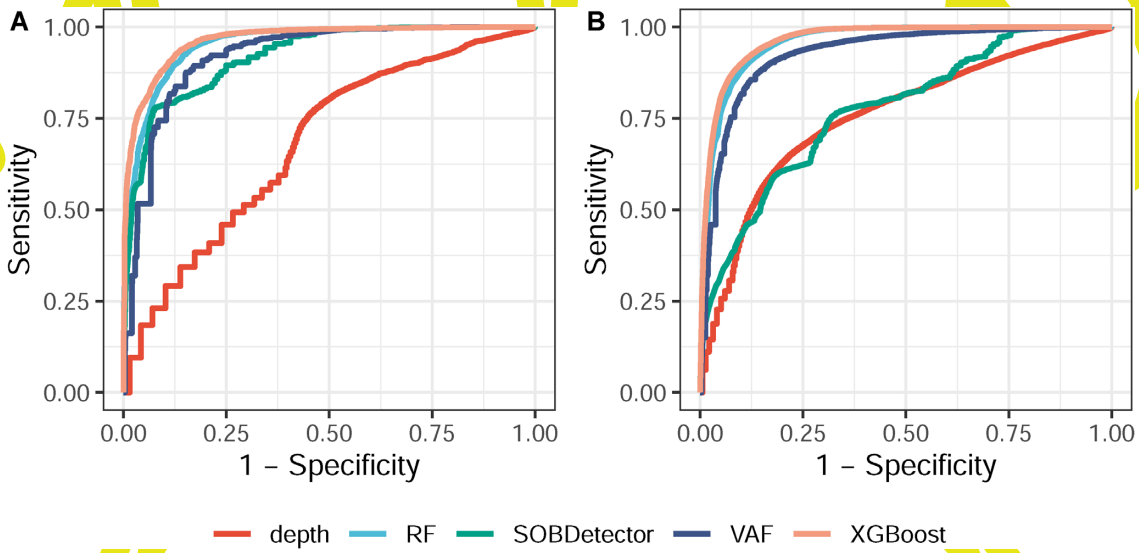

Figure 7. ROC curves for the five tested approaches on the colon and liver (**A**) and gastro-oesophageal (**B**) validation sets. For both datasets, XGBoost and RF outperform the other three approaches. In the colon and liver dataset, filtering by VAF and SOBDetector perform similarly and slightly worse than our models. In the gastro-oesophageal dataset, instead, filtering by VAF keeps its good performance, while SOBDetector performs much poorly. Filtering by depth of coverage is the worst-performing approach in both cases. AUC values are shown in Table 7.

Table 7. AUC values for the five tested models on the two validation sets

Validation set	XGBoost	RF	SOBDetector	VAF	Depth
Colon and liver	0.9643	0.9541	0.9216	0.9260	0.6883
Gastro-oesophageal	0.9639	0.9567	0.7745	0.9285	0.7588

XGBoost, RF and filtering by VAF all score above 0.92 in both datasets, while SOBDetector achieves those performance levels only in the colon and liver dataset. In both sets, XGBoost and RF show the best performance, with all AUC values above 0.95. Filtering by depth of coverage is several points below in all cases.

Table 8. Importance values and ranks provided by RF and XGBoost built-in metrics for the features in the models, ordered in decreasing importance order according to the RF model

Feature	Scaled importance (RF)	Scaled importance (XGBoost)	Rank (RF)	Rank (XGBoost)
FDeamC	1	1	1	1
SOB	0.71	0.12	2	3
Median position from read end	0.67	0.065	3	11
Dinucleotide after	0.65	0.0058 (CA)	4	27 (CA)
Dinucleotide before	0.63	0.0056 (CC)	5	28 (CC)
VAF	0.59	0.051	6	13
SB-GATK	0.57	0.09	7	8
Normalized number of alternate bases	0.54	0.63	8	2
Normalized number of reference bases	0.53	0	9	65
SB-GUO	0.51	0.11	10	4
Base quality fraction	0.51	0.055	11	12
Normalized median position from read end	0.50	0.10	12	6
Base quality	0.43	0.020	13	17
Number of reference bases	0.41	0.077	14	10
Number of alternate bases	0.39	0.028	15	15
Fragment length	0.39	0.10	16	7
Base two positions before	0.25	0.0044 (T)	17	31 (T)
Base two positions after	0.24	0.0069 (G)	18	25 (G)
Base one position after	0.18	0.090 (G)	19	9 (G)
Base one position before	0.17	0.11 (C)	20	5 (C)
Homopolymer length	0.13	0.021	21	16
Mapping quality	0.11	0.019	22	18
Reference allele	0.10	0.030 (C)	23	14 (C)

The importance values presented here have been scaled to have a maximum value of 1; i.e. the top feature has an importance value of 1 and the rest are scaled as a function of it. In the case of XGBoost, categorical features had been one-hot encoded, creating a much larger number of descriptors. For the sake of readability and simplification, we only report the importance and rank values of the top-scoring feature within each broken down categorical feature in XGBoost, which is shown between parentheses in the corresponding row.

of 0.67 for the median position; scores of 0.65 and 0.63 for the dinucleotides immediately after and before the variant, respectively). VAF and related features (i.e. the normalized numbers of alternate and reference bases) were also among the top features. This is consistent with the VAF distribution in our data whereby deaminations increased at near 0% VAF values, and non-deaminations were more common as we neared VAFs of 30% (Figure 1). SB-related features were also relatively high in the ranking and thus showed a degree of contribution to the model (scaled importance scores of 0.57 for SB-GATK and 0.51 for SB-GUO). Similar importance scores were achieved by the base quality fraction and the normalized median position on the read. The remaining features showed a score <0.50, meaning that their importance was less than one-half of that of FDeamC. These features included the base quality, the number of reference and alternate bases, the fragment length, the nucleotides one and two genomic positions before and after the variant, the length of the homopolymer they were in, their mapping quality and the reference allele. In general, importance scores showed a gradual decrease, with the exception of the first and second features, which were separated by a gap of 0.29, and reached a minimum value of 0.10, which corresponded to the feature representing the reference allele.

As said, the analysis of the ranking produced by XGBoost was not so straightforward, but it is safe to say that, although evident differences exist between the two rankings, they both agreed in highlighting the major relevance of those features related to read pair orientation bias and VAF and, to a certain extent, of those related to the SB and to the genomic context.

Implementation of the tool

We compiled the best two models trained on the breast dataset (i.e. XGBoost and RF) in an R package named Ideafix. Ideafix uses a vcf file generated with Mutect2 (either in tumour-only mode or in tumour/normal mode) as an input. It requires variant calling to be carried out using the Mutect2 algorithm as some of the model descriptors are specific to this tool. This variant caller was moreover chosen because it is publicly available, and it is one of the most widely used and best rated tools (40,41). Ideafix generates a new text file (the original vcf file or a new tab-separated file) indicating deamination probability and vari-

ant class (deamination or non-deamination) for each C:G > T:A variant with a VAF <30%. It should be noted that Ideafix has been designed to be used with data generated from Illumina paired-end technology, so its performance with data generated from other technologies such as Ion Torrent has yet to be evaluated.

DISCUSSION

In this study, we have explored the use of machine learning-based approaches for the identification of formalin-induced cytosine deaminations. Specifically, we have formulated this question as a fully supervised classification problem in which we try to classify low variant frequency (VAF <30%) variants into deaminations and non-deaminations. To our knowledge, this is the first time such methodology has been used for this purpose.

Given a set of variants found in an FFPE specimen, the algorithm we propose is able to identify the artefactual C:G > T:A changes with an estimated AUC >0.95. Other benchmarked methods scored AUC values between 0.92 and 0.93 at best. Moreover, existing methods are all univariate, whereas the Ideafix algorithm takes a multivariate approach that allows to exploit relationships between the descriptors.

Importantly, we carried out validation in multiple tumour types (breast, colon, liver, gastric, oesophageal) obtained, processed and sequenced in multiple centres and, as we have shown, displaying different levels of DNA damage. Predictably, the best results were obtained with samples that were less degraded, namely those in the colon and liver cancer and gastro-oesophageal cancer datasets. The fact that none of the variants in these samples were used for training the final model confirms that DNA damage levels greatly affect predictability. Figure 4 suggests that some samples are easier to deal with than others as well; here, the largest SOE value is >0.2 points above the lowest SOE value for all the models, and the SOE value of certain samples is larger than the CVE and even the RE. At the same time, all these observations suggest an absence of overfitting in the models.

The performance of our model was compared to three other approaches: SOBDetector, filtering variants by depth of coverage and filtering by VAF. The criteria we followed to select between existing tools or practices (Supplementary Table S2) are described next. First of all, comparing our refinement model against variant calling methods would have been unfair to the latter as even if their aim is the same, they solve different problems. Our refinement model assesses a list of already preprocessed variants that form up the complete set of existing variants—deaminations and non-deaminations—that is thereafter used for evaluation. Variant calling tools, on the other hand, are run on raw bam files, and thus they evaluate the entire coverage of the sequenced region, i.e. the exome in this case, for possible variants, which in turn expands beyond the collection of mutations to be evaluated. Among the variant refinement tools designed for FFPE data, we ruled out the two GATK features—LearnReadOrientationModel, because we already used it as an assessment of the variant labels, and FilterByOrientationBias, due to it being a beta feature. Between the remaining, we prioritized tool availability and user-friendliness, so that the final selection ended up being the aforementioned.

A comparison between five different classification algorithms demonstrated that decision tree-based models (XGBoost and RF) performed best. This most probably reflects their capacity to work well under default learning parameters, to pick up relationships between features, an effective handle of non-informative descriptors, outliers and class imbalance, and their treatment of bias and variance through the averaging of more constrained classifiers. It is also noteworthy that our dataset contains a quite large number of correlated features, e.g. FDeamC and SOB or all allele frequency-related features, and that decision tree algorithms are by nature robust to this multicollinearity (42).

As already stated, XGBoost and RF were the best-performing models with almost identical performance as can be concluded from the overlap of their ROC curves (Figure 7). Still, the particular classification threshold we set (the one that minimizes the number of false positives while keeping a minimum sensitivity value of 0.99) has a different impact on the two models; i.e. it provides different costs to false positive and false negative errors. This is observed particularly in the sensitivity and specificity values: while RF shows better specificity, XGBoost has a greater sensitivity. Likewise, it is noticeable that a minimal improvement in the PPV is heavily damaging to the NPV. We believe this is due to the number of positives (deaminations) being larger than the number of negatives (non-deaminations). We used similar but not identical rules to set the threshold for both models, so it is yet to be clarified whether this and not the model itself could be the reason of observing such differences.

Although decision tree models are black box models and their interpretation is difficult, some insight into the cytosine deamination process can be gained from them and from the univariate feature analyses. All the three feature importance analyses agreed in highlighting the relevance of read pair orientation bias and allele frequency for identifying deaminations. This comes as no surprise as there are a large number of works describing read pair orientation bias in formalin-induced cytosine deaminations, and indeed, several of the existing tools and filters to deal with data from FFPE tissues build upon those features (21,22). We hypothesized the possible influence of some new elements for deamination identification, such as the composition of flanking bases, median length of the fragments that carry the base change, base quality and position in the genomic fragment. While the relevance of most of them was shown to be limited, the genomic context and the variant position in the fragment seemed to be useful.

Interestingly, setting aside read pair orientation bias and allele frequency, RF and XGBoost differed to a certain extent in the importance given to the features. We believe one important reason behind these differences is the way the two models deal with correlated features. In the RF algorithm, two correlated features have an equivalent probability of being chosen to split a branch on a tree, as trees are independently grown on resampled datasets, and hence, the importance of the property they reflect is diluted. That is not the case for XGBoost, in which trees are created sequentially and each one learns on the observations that the previous tree has not been able to classify. As a consequence, when

two correlated features are present, if one of them has been used as a split in a previous tree, new trees will likely lay aside those features and build upon unrelated ones. Indeed, this behaviour is reflected in our results, as RF gave similar importance values to sets of correlated features such as FDeamC and SOB, VAF, normalized number of alternate bases and normalized number of reference bases, or SB-GATK and SB-GUO, while XGBoost did not. In any case, the fact that XGBoost is able to identify correlated features allows to use reduced collections of features. This is an interesting point, especially for occasions where obtaining the values of some features involves hard work.

We would also like to raise the issue of removing the isSNP descriptor from the model. As mentioned earlier, this feature is confounded with the classification label (deamination or non-deamination) in these datasets, due to the following. The procedure we followed to label the variants into deaminations and non-deaminations was based on their presence in one or both datasets of FFPE–FF paired datasets, respectively (see the 'Variant labelling' subsection in the 'Materials and Methods' section for further details). Thus, the non-deamination label was primarily given to both germline variants and systematic errors. While other non-systematic errors exist and thus could be labelled as deaminations, it is possible though less likely that a germline variant is not detected in both datasets or that a considerable number of them get affected by the deamination phenomenon and thus get the deamination label. As a consequence, the isSNP feature gets significantly confounded with the deamination status and this could result in models that rely very much on this feature for the classification. Although this situation is not spurious, it could impact the classification of some problem sets, such as those in which we are interested in separating somatic variants from deaminations, in a negative way. This is so because instead of recognizing deaminations, the model would be identifying germline variants to a great extent, which would be of no interest in such case, and at the same time, performance estimations we report here would then be overestimations of the real performance. Directly discarding these variants that were true for the isSNP feature seemed incorrect since, as we already stated, deaminations can also affect these positions. This fact was reinforced by the presence of such events in our dataset (data not shown). Another reason against acting this way is the existing evidence stating that filtering variants based solely on their position is inaccurate and leads to incorrect results (43). As a consequence, removing the isSNP descriptor from the datasets seemed to be the best option.

Class imbalance was a major challenge faced with the data, especially in the breast cancer dataset, which contained a set of artefacts between 64.2 and 263 times greater than the collection of real mutations. This suggests that samples in this dataset were much more degraded than those in the other two datasets and, indeed, damage analysis results confirmed this observation. This imbalance reflects the difference in the timing of the two mutation processes: while the real mutation burden of a biopsy will remain intact over time, cytosine deamination is a cumulative process that will not cease and that is dependent on factors such as storage time and time in contact with formaldehyde. Results show that, even if this imbalance is high, good performance can be obtained. In spite of the proficient performance, this class asymmetry opens the door to use other learning paradigms, such as one-class classification, that could better adapt to the present scenario.

CONCLUSIONS AND FUTURE WORK

In this work, we present an effective algorithm to identify cytosine deaminations within a list of low-frequency variants based on a collection of >20 descriptors. This method can hence be applied to filter variant calls from FFPE specimens and helps to confidently use these samples for molecular testing.

Our tool has several advantages over other existing approaches that deal with formalin-induced artefacts. When compared to wet-lab procedures, our algorithm is a non-invasive and cost-effective technique and can be applied retrospectively to existing FFPE material already sequenced. Compared to computational approaches, Ideafix not only outperforms them, but also does not automatically remove identified artefacts that could result in the generation of clinically important false negative results (15,19,21). Instead, we chose to allow the user to decide what to do with the class probability information, for example whether certain variants warrant further testing if they occur in a clinically relevant manner, such as mutations in the EGFR gene in lung cancer patients that dictate TKI treatment. Additionally, unlike other methodologies that require multiple filtering steps (13) and format conversion, the Ideafix algorithm is fully automatic.

This work has several limitations that could be improved in the future, including the lack of the true class labels of the data through independent validation of variants in FFPE material by other techniques such as qRT-PCR. We mitigated this risk by comparing FFPE specimens with matched FF samples from the same tissues. Even then we are aware that some variants classified as deaminations could be subclonal mutations only present in the FFPE portion of the tumour tissue or other non-systematic artefacts missed by quality filters. Encouragingly, when we compare the labels with those of a second deamination identification approach, we obtain a large concordance, which reinforces the labelling technique used. A second limitation is the sequencing data this method can be applied to. This is so because the top contributing feature in the final models, i.e. FDeamC, is only computable in DNA data sequenced with Illumina paired-end technology. Thus, while it is technically possible to apply the algorithm to data generated by other systems, e.g. Ion Torrent, the performance of our approach should be reevaluated.

Future work to improve this package will include testing and possible incorporation of new descriptors in the model such as the genomic complexity of the variant region, since less overlapping between SNVs detected in FFPE and FF samples has been reported in low-complexity regions than in reliable regions (44). Additionally, the number of PCR cycles used for template amplification could also be incorporated in the model, as this would probably refine any allele frequency value biased by this technique. Besides, we propose addressing the problem from alternative learning scenarios, instead of formulating the problem as fully su-

pervised. For instance, we could try to learn on probabilistic labels (45), i.e. to learn the classifier from a dataset with probabilities associated with the labels, or to approach the problem as a weak supervision problem (46). These strategies provide an alternative natural fit to the problem, as this is a problem where the majority of the labels will always be uncertain and no ground truth exists. One-class classification schemes could be suitable due to the inherent imbalance in these datasets.

One of the most interesting scenarios to apply this methodology in is somatic mutation detection, so we built our algorithm upon cancer DNA sequencing data. In fact, intratumour heterogeneity has largely been overlooked mainly due to technical difficulties, but in the recent decades more and more attention has been brought to it and now there is evidence that subclonal mutations play key roles in disease progression and response to therapy (8,47–49). Given that most biopsies are routinely collected as FFPE specimens for histopathological review, if precision medicine is to become part of routine healthcare, procedures able to confidently distinguish low-frequency somatic mutations from formalin-induced changes should be implemented. In any case, besides cancer somatic mutation detection, this method can be applied to any tissue sample of a heterogeneous cell population from FFPE data.

DATA AVAILABILITY

Ideafix is available at https://github.com/mmaitenat/ideafix. Used public data are available at https://www.ebi.ac.uk/ena/browser/view/SRP044740, https://www.ebi.ac.uk/ega/studies/EGAS00001002631 and https://www.ebi.ac.uk/ena/browser/view/PRJEB44073.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We are thankful to Dr Eric Bonnet and his team, Dr Irene Chong and Dr Syed Haider for providing the colon and liver cancer and gastro-oesophageal datasets. We also thank Dr Lorea Manterola for insightful discussions on formalin-induced deamination mechanisms that aided in the selection of model descriptors.

FUNDING

Departamento de Educación, Universidades e Investigación of the Basque Government [PRE_2019_2_0211 to M.T.A]; Ikerbasque, Basque Foundation for Science [to C.L.]; Starmer–Smith Memorial Fund [to C.L.]; Ministerio de Economía, Industria y Competitividad (MINECO) of the Spanish Central Government [to C.L., PID2019-104966GB-I00 to B.C.]; ISCIII and FEDER Funds [PI12/00663, PIE13/00048, DTS14/00109, PI15/00275 and PI18/01710 to C.L.]; Departamento de Desarrollo Económico y Competitividad and Departamento de Sanidad of the Basque Government [to C.L.]; Asociación Española Contra el Cancer (AECC) [to C.L.]; Diputación Foral de Guipuzcoa (DFG) [to C.L.]; Departamento de Industria of the Basque Government [ELKARTEK Programme, project code: KK-2018/00038 to C.L., ELKARTEK Programme, project code: KK-2020/00049 to B.C., IT-1244-19 to B.C.].

Conflict of interest statement. None declared.

REFERENCES

1. Haile,S., Pandoh,P., McDonald,H., Corbett,R.D., Tsao,P., Kirk,H., MacLeod,T., Jones,M., Bilobram,S., Brooks,D. *et al.* (2017) Automated high throughput nucleic acid purification from formalin-fixed paraffin-embedded tissue samples for next generation sequence analysis. *PLoS One*, **12**, e0178706.
2. Hofreiter,M., Jaenicke,V., Serre,D., Haeseler,A.V. and Pääbo,S. (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.*, **29**, 4793–4799.
3. Chen,G., Mosier,S., Gocke,C.D., Lin,M.-T. and Eshleman,J.R. (2014) Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol. Diagn. Ther.*, **18**, 587–593.
4. Do,H. and Dobrovic,A. (2015) Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil-DNA glycosylase. *Oncotarget*, **3**, 546–558.
5. Wong,S.Q., Li,J., Tan,A.Y., Vedururu,R., Pang,J.-M.B., Do,H., Ellul,J., Doig,K., Bell,A., McArthur,G.A. *et al.* (2014) Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Med. Genomics*, **7**, 23.
6. Wong,S.Q., Fellowes,A., Doig,K., Ellul,J., Bosma,T.J., Irwin,D., Vedururu,R., Tan,A.Y., Weiss,J., Chan,K. *et al.* (2015) Assessing the clinical value of targeted massively parallel sequencing in a longitudinal, prospective population-based study of cancer patients. *Br. J. Cancer*, **112**, 1411–1420.
7. Lupini,L., Bassi,C., Mlcochova,J., Musa,G., Russo,M., Vychytilova-Faltejskova,P., Svoboda,M., Sabbioni,S., Nemecek,R., Slaby,O. *et al.* (2015) Prediction of response to anti-EGFR antibody-based therapies by multigene sequencing in colorectal cancer patients. *BMC Cancer*, **15**, 808.
8. Shin,H.-T., Choi,Y.-L., Yun,J.W., Kim,N.K., Kim,S.-Y., Jeon,H.J., Nam,J.-Y., Lee,C., Ryu,D., Kim,S.C. *et al.* (2017) Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nat. Commun.*, **8**, 1377.
9. Ivanov,M., Laktionov,K., Breder,V., Chernenko,P., Novikova,E., Telysheva,E., Musienko,S., Baranova,A. and Mileyko,V. (2017) Towards standardization of next-generation sequencing of FFPE samples for clinical oncology: intrinsic obstacles and possible solutions. *J. Transl. Med.*, **15**, 22.
10. Schmitt,M.W., Kennedy,S.R., Salk,J.J., Fox,E.J., Hiatt,J.B. and Loeb,L.A. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. U.S.A.*, **109**, 14508–14513.
11. Kinde,I., Wu,J., Papadopoulos,N., Kinzler,K.W. and Vogelstein,B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 9530–9535.
12. Bonnet,E., Moutet,M.-L., Baulard,C., Bacq-Daian,D., Sandron,F., Mesrob,L., Fin,B., Delépine,M., Palomares,M.-A., Jubin,C. *et al.* (2018) Performance comparison of three DNA extraction kits on human whole-exome data from formalin-fixed paraffin-embedded normal and tumor samples. *PLoS One*, **13**, e0195471.
13. Yost,S.E., Smith,E.N., Schwab,R.B., Bao,L., Jung,H., Wang,X., Voest,E., Pierce,J.P., Messer,K., Parker,B.A. *et al.* (2012) Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.*, **40**, e107.
14. Kerick,M., Isau,M., Timmermann,B., Sültmann,H., Herwig,R., Krobitsch,S., Schaefer,G., Verdorfer,I., Bartsch,G., Klocker,H. *et al.* (2011) Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genomics*, **4**, 68.

15. Kim,H., Lee,A.J., Lee,J., Chun,H., Ju,Y.S. and Hong,D. (2019) FIREVAT: finding reliable variants without artifacts in human cancer samples using etiologically relevant mutational signatures. *Genome Med.*, **11**, 81.
16. Kato,M., Nakamura,H., Nagai,M., Kubo,T., Elzawahry,A., Totoki,Y., Tanabe,Y., Furukawa,E., Miyamoto,J., Sakamoto,H. *et al.* (2018) A computational tool to detect DNA alterations tailored to formalin-fixed paraffin-embedded samples in cancer clinical sequencing. *Genome Med.*, **10**, 44.
17. Frampton,G.M., Fichtenholtz,A., Otto,G.A., Wang,K., Downing,S.R., He,J., Schnall-Levin,M., White,J., Sanford,E.M., An,P. *et al.* (2013) Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.*, **31**, 1023.
18. Carrot-Zhang,J. and Majewski,J. (2017) LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. *Oncotarget*, **8**, 37032.
19. Dunn,T., Berry,G., Emig-Agius,D., Jiang,Y., Lei,S., Iyer,A., Udar,N., Chuang,H.-Y., Hegarty,J., Dickover,M. *et al.* (2019) Pisces: an accurate and versatile variant caller for somatic and germline next-generation sequencing data. *Bioinformatics*, **35**, 1579–1581.
20. Chen,L., Liu,P., Evans,T.C. and Ettwiller,L.M. (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752–756.
21. Van der Auwera,G.A., Carneiro,M.O., Hartl,C., Poplin,R., del Angel,G., Levy-Moonshine,A., Jordan,T., Shakir,K., Roazen,D., Thibault,J. *et al.* (2013) From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, **43**, 11.10.1–11.10.33.
22. Diossy,M., Sztupinszki,Z., Krzystanek,M., Borcsok,J., Eklund,A.C., Csabai,I., Pedersen,A.G. and Szallasi,Z. (2021) Strand Orientation Bias Detector to determine the probability of FFPE sequencing artifacts. *Brief. Bioinform.*, https://doi.org/10.1093/bib/bbab186.
23. Chong,I.Y., Starling,N., Rust,A., Alexander,J., Aronson,L., Llorca-Cardenosa,M., Chauhan,R., Chaudry,A., Kumar,S., Fenwick,K. *et al.* (2021) The mutational concordance of fixed formalin paraffin embedded and fresh frozen gastro-oesophageal tumours using whole exome sequencing. *J. Clin. Med.*, **10**, 215.
24. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
25. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
26. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
27. Ruden,D.M., Cingolani,P., Patel,V.M., Coon,M., Nguyen,T., Land,S.J. and Lu,X. (2012) Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.*, **3**, 35.
28. Sherry,S.T., Ward,M.-H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
29. Lindenbaum,P. (2015) JVarkit: Java-based utilities for bioinformatics. https://doi.org/10.6084/m9.figshare.1425030.v1.
30. Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.
31. Briggs,A.W., Stenzel,U., Johnson,P.L., Green,R.E., Kelso,J., Prüfer,K., Meyer,M., Krause,J., Ronan,M.T., Lachmann,M. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. U.S.A.*, **104**, 14616–14621.
32. Lindahl,T. and Nyberg,B. (1972) Rate of depurination of native deoxyribonucleic acid. *Biochemistry*, **11**, 3610–3618.
33. Costello,M., Pugh,T.J., Fennell,T.J., Stewart,C., Lichtenstein,L., Meldrim,J.C., Fostel,J.L., Friedrich,D.C., Perrin,D., Dionne,D. *et al.* (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.*, **41**, e67.
34. Li,H. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843–2851.
35. Guo,Y., Cai,Q., Samuels,D.C., Ye,F., Long,J., Li,C.-I., Winther,J.F., Tawn,E.J., Stovall,M., Lähteenmäki,P. *et al.* (2012) The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation. *Mutat. Res.*, **744**, 154–160.
36. Andrews,S. (2010) FastQC: a quality control tool for high throughput sequence data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
37. Bradley,A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.*, **30**, 1145–1159.
38. Witten,I.H., Frank,E., Hall,M.A. and Pal,C.J. (2016) In: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, MA.
39. Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM KDD'16)*, NY, pp. 785–794.
40. Bian,X., Zhu,B., Wang,M., Hu,Y., Chen,Q., Nguyen,C., Hicks,B. and Meerzaman,D. (2018) Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics*, **19**, 429.
41. Cai,L., Yuan,W., Zhang,Z., He,L. and Chou,K.-C. (2016) In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci. Rep.*, **6**, 36540.
42. Breiman,L. (2017) In: *Classification and Regression Trees*. Routledge, London.
43. Hiltemann,S., Jenster,G., Trapman,J., van der Spek,P. and Stubbs,A. (2015) Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Res.*, **25**, 1382–1390.
44. Robbe,P., Popitsch,N., Knight,S.J., Antoniou,P., Becq,J., He,M., Kanapin,A., Samsonova,A., Vavoulis,D.V., Ross,M.T. *et al.* (2018) Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genet. Med.*, **20**, 1196–1205.
45. Jin,R. and Ghahramani,Z. (2003) Learning with multiple labels. In: *Advances in Neural Information Processing Systems 15*. MIT Press.
46. Hernández-González,J., Inza,I. and Lozano,J.A. (2016) Weak supervision and other non-standard classification problems: a taxonomy. *Pattern Recogn. Lett.*, **69**, 49–55.
47. Landau,D.A., Carter,S.L., Stojanov,P., McKenna,A., Stevenson,K., Lawrence,M.S., Sougnez,C., Stewart,C., Sivachenko,A., Wang,L. *et al.* (2013) Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, **152**, 714–726.
48. Nadeu,F., Clot,G., Delgado,J., Martín-García,D., Baumann,T., Salaverria,I., Beà,S., Pinyol,M., Jares,P., Navarro,A. *et al.* (2018) Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia*, **32**, 645.
49. Mroz,E.A. and Rocco,J.W. (2013) MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol.*, **49**, 211–215.