Published online 30 May 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 2 **1**
https://doi.org/10.1093/nargab/lqad056

RFcaller: a machine learning approach combined with read-level features to detect somatic mutations

Ander Díaz-Navarro [1], Pablo Bousquets-Muñoz [1], Ferran Nadeu[2,3], Sara López-Tamargo[1], Silvia Beà[2,3,4], Elias Campo[2,3,4] and Xose S. Puente[1,3,*]

[1]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, 33006 Oviedo, Spain, [2]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain, [3]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), 28029 Madrid, Spain and [4]Hospital Clinic de Barcelona, Universitat de Barcelona, 08036 Barcelona, Spain

Received October 13, 2022; Revised May 11, 2023; Editorial Decision May 22, 2023; Accepted May 23, 2023

ABSTRACT

The cost reduction in sequencing and the extensive genomic characterization of a wide variety of cancers are expanding tumor sequencing to a wide number of research groups and the clinical practice. Although specific pipelines have been generated for the identification of somatic mutations, their results usually differ considerably, and a common approach is to use several callers to achieve a more reliable set of mutations. This procedure is computationally expensive and time-consuming, and it suffers from the same limitations in sensitivity and specificity as other approaches. Expert revision of mutant calls is therefore required to verify calls that might be used for clinical diagnosis. This step could take advantage of machine learning techniques, as they provide a useful approach to incorporate expert-reviewed information for the identification of somatic mutations. Here we present RFcaller, a pipeline based on machine learning algorithms, for the detection of somatic mutations in tumor–normal paired samples that does not require large computing resources. RFcaller shows high accuracy for the detection of substitutions and insertions/deletions from whole genome or exome data. It allows the detection of mutations in driver genes missed by other approaches, and has been validated by comparison to deep and Sanger sequencing.

INTRODUCTION

During the last decade, the introduction of next-generation sequencing (NGS) has transformed the study of cancer, with the identification of hundreds of novel alterations driving tumor transformation (1). Major international cancer projects such as the International Cancer Genome Consortium (2) and The Cancer Genome Atlas (TCGA) (3) have expanded the repertoire of genes mutated in cancer, as well as the biological processes involved in it (4–7). The continuous reduction in sequencing costs, together with the clinical significance of certain mutations for prognosis or treatment decisions, has transformed the use of NGS from large sequencing consortia to small-sized laboratories and clinical centers. However, the utility of NGS relies on the availability of somatic mutation calling pipelines with enough sensitivity to detect somatic mutations, and high specificity to prevent the calling of artifacts or germline variants as mutations.

Somatic single nucleotide variants (SSNVs) and small insertions/deletions (indels) constitute the most abundant type of mutation in tumor genomes, and different tools have been developed in order to call somatic mutations from tumor–normal paired samples. Most state-of-the-art variant callings are based on traditional statistical methods, such as CaVEMan (8), Mutect2 (9), MuSE (10), Strelka2 (11), Pindel (12) or SMuFin (13), among others. However, there is no full consensus on the mutations detected by each caller, usually sharing 77% of mutations between two independent callers, a percentage that is reduced as the number of programs increases (5). These differences are mainly due to the ability of each program to deal with the tumor heterogeneity and purity, normal contamination, sequencing and mapping artifacts, coverage and different tool configurations. Due to the fact that each caller usually detects specific false positive mutations, some collaborative projects such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) (5) or the TCGA Pan-Cancer Atlas MC3 (14) do not use a single caller but a combination of algorithms, keeping the intersection between them as the set of mutations that is more reliable. Despite the utility of this multi-pipeline approach to generate a consensus set of mutations, this strategy has a very large computational cost, demanding large servers and consuming up to days for the analysis of a single case.

*To whom correspondence should be addressed. Tel: +34 985105027; Email: xspuente@uniovi.es

In addition to classical statistics-based approaches, during the last few years there has been an expansion in the use of machine learning strategies for different purposes (15–19), including the development of new variant calling tools. The initial use of these methods was mainly focused on refinement, taking a list of potential variants extracted with other pipelines to filter and select a final set of mutations. However, these approaches still have a negative influence on computing time and reproducibility. On the other hand, recently developed pipelines use other machine learning approaches (20,21) or even neural networks (22,23) to directly perform variant calling for somatic mutations, although in some cases the computational cost is too high or the installation requirements too demanding for a medium-sized laboratory or institution.

Here, we describe RFcaller, an accurate, fast, light computational requirement and easy-to-use tool that uses read-level features together with machine learning strategies to identify somatic mutations (SSNVs and indels) from normal–tumor paired samples. Our pipeline has been trained for whole genome sequencing (WGS) data and its results have been compared with those obtained by the PCAWG, being very similar to those resulting by combining several tools.

MATERIALS AND METHODS

Selection of somatic mutations

For the development of the algorithms, two different set of mutations were used, a training set and a testing set. To build them, we extracted with bcftools (24) all possible somatic mutations from four WGS mantle cell lymphoma samples sequenced at 30× coverage (M032 and M439 for training; M065 and M431 for testing). For the initial training, previously published mutations (25) were defined as true positive mutations. With each iteration, all discordant calls were manually reviewed by three experts, through visual inspection (Supplementary Table S1), and the database was updated accordingly. This procedure resulted in the identification of novel *bona fide* mutations that would constitute false negatives in the initial set, as well as the rejection of certain mutations, such as artifacts or germline mutations present in the original dataset, that would represent false positives. After nine rounds of training the algorithms and curating the set of mutations, all discordant variants had already been examined, which allowed us to obtain a reliable dataset for training and testing the final version of the algorithms.

Algorithm training

To train the algorithms, we used the training set that contained 66 096 potential SSNVs (Supplementary Table S2) and 931 indels (Supplementary Table S3) for which read-level features were previously extracted (Supplementary Table S4). These data were used as input by TPOT (v0.11.1) (26), with the default configuration of the *TPOTRegressor* function, to find the best pipeline to train the regression algorithms. As a result, an extremely randomized tree 'Extra-Tree' regressor for SSNVs and a random forest regressor for indels were built. Both algorithms use a prediction model based on decision trees with some minor differences. For example, although both select the features randomly for each branch, these are split by the optimal cut point in a random forest, but randomly in an 'Extra-Tree'. Additionally, a random forest algorithm draws observations with replacement, while an 'Extra-Tree' extracts them without replacement. These differences allow 'Extra-Tree' algorithms to reduce the bias and the variance. For both models, a transformation of the data was carried out before the regression using the *StackingEstimator* function.

Once we had the algorithms, the test dataset, with 63 948 SSNVs and 2506 indels (Supplementary Table S5), was used to select the best cutoffs for both pipelines. With this purpose, the result from RFcaller was filtered to get the 'QUAL' field for those mutations that passed all filters (Supplementary Table S6). This parameter is calculated considering the initial quality from bcftools and the regression value for SSNV and indels, and only the regression value for homopolymer indels (polyindels):

$$\mathrm{QUAL}_{\mathrm{SNV}} = \text{bcftools qual} * \text{regression value}^2$$

$$\mathrm{QUAL}_{\mathrm{indel}} = \text{bcftools qual}^{\text{regression value}}$$

$$\mathrm{QUAL}_{\mathrm{polyindel}} = \text{regression value}$$

Then, receiver operating characteristic curves were generated and area under the curve (AUC) metrics were calculated using the R package OptimalCutpoints (27) with the *MaxEfficiency* method. False/true positive/negative ratios were calculated using the formulas described in the R package ROCR (28).

Computational cost

To compare the performance of RFcaller with other state-of-the-art tools, the docker container corresponding to the four callers used by PCAWG for the detection of SSNVs was downloaded (https://dockstore.org/organizations/PCAWG/collections/PCAWG). After minor fixes of broken links in the Sanger and DKFZ tools, all of them were run with the default parameters for one random donor. In case the tools allowed to choose the number of threads and RAM to be used, 20 threads and 200 GB of memory were specified. In addition, because RFcaller allows multiple samples to be run simultaneously, four cases were run in parallel using the default parameters to calculate the computational cost. To improve data interpretation, some axes were broken using the R package ggbreak (29).

PCAWG analysis

To validate that the trained models are applicable for liquid and solid tumors and to compare the results to those obtained by the PCAWG pipeline, RFcaller was run for the CLLE-ES and BRCA-EU studies, with an average tumor coverage of 30× and 50×, respectively, and 30× for normal samples in both. PCAWG BAM files were downloaded from the 'collaboratory' repository using the score-client program (Supplementary Table S7). RFcaller was run with its default parameters for all samples and the obtained

results were combined into a single VCF file for each study. A custom panel of normals was used to obtain a more reliable set of mutations by annotating variants in complex regions, allowing elimination of false positives derived from sequence background. The set of mutations detected by the PCAWG pipeline were extracted from the controlled consensus callsets for SSNV/indel. To analyze coding and noncoding mutations, the Variant Effect Predictor tool (30) was launched for both datasets using the following options: `--offline --format vcf --dir_cache homo_sapiens --symbol --force_overwrite --total_length --numbers --ccds --canonical --biotype --pick --vcf --assembly GRCh37`.

To be able to compare both sets of mutations in the most accurate manner, (i) dinucleotides and trinucleotides from RFcaller were split as this feature is not available for PCAWG, (ii) RFcaller mutations located in alternative chromosomes and PCAWG's variants that appear in our custom dbSNP were removed and (iii) only mutations that passed all filters were studied. For this comparison, a mutation was considered as low-VAF mutation when its variant allele frequency (VAF) was <0.15, in accordance with the sensitivity of Sanger sequencing.

For the purpose of calculating the precision and recall for both pipelines in each study, 1% or at least 50 discordant mutations from each section were manually reviewed by a panel of experts, considering the features listed in Supplementary Table S1. This process was blinded as each person reviewed a random set of mutations and did not know to which pipeline corresponded each mutation. After this first step, uncertain mutations were examined individually by each of the three reviewers to reach an agreement. A total of five blocks were checked: mutations detected only by RFcaller and mutations detected between one and four of the callers used by the PCAWG, as the ratio of false positives may be different between them. The results obtained were then extrapolated to the whole set of mutations in order to calculate the total number of true and false positive private mutations. To extrapolate the data, the percentages of true and false positives were computed for the manually reviewed mutations, and then these scores were multiplied by the total number of variants detected for each of the five blocks mentioned above. Mutations detected by both pipelines were classified as true positives. False positive PCAWG-private mutations would be RFcaller true negatives, whereas PCAWG-private true positives would be RFcaller false negative mutations, and vice versa. Using these values, precision and recall were calculated with the *prediction* and *performance* functions of the R package ROCR (28). These same data were used for the comparison between RFcaller and the individual variant callers used by PCAWG, since this consortium specifies which callers detect each mutation. The procedure is detailed in Supplementary Data scripts.

Finally, deep sequencing data generated by previous studies (31,32) for some CLLE-ES cases (Supplementary Table S8) were used to analyze possible low-VAF mutations in driver genes. In order to compare both results, only mutations in CLL driver genes and donors analyzed by both WGS and deep sequencing were selected. In addition, mutations detected by deep sequencing were removed from the analysis if they were germline or there was insufficient coverage or reads supporting the mutation by WGS (Supplementary Table S9).

To compare the performance of RFcaller against other tools based on machine learning [SNooPer (21), NeoMutate (19), Cerebro (20) and DeepSVR (18)], only DeepSVR could be benchmarked. We used the default configuration and the SSNV positions detected after the initial calling step by RFcaller for CLLE-ES and BRCA-EU studies. No indels were analyzed because it does not identify indels.

Sanger validation

To perform verification of private calls obtained from the analysis of CLLE-ES cases, five and two mutations detected only by RFcaller and PCAWG, respectively, were chosen to be verified by Sanger sequencing. These positions were chosen because they appeared in known driver genes for CLL and because tumor and/or normal DNA was available. The list of primers and melting temperatures are listed in Supplementary Table S10.

Exome analysis

To test the performance of RFcaller on exome sequencing data, we selected five CLLE-ES cases previously analyzed by WGS and for which exome data were available (Supplementary Table S8). Average sequencing depth was 50× (33). RFcaller was run with default parameters and LIKELY_GERMINAL variants were removed. Only mutations within the targeted regions of the exome (Agilent, SureSelect Human All Exons V4) were taken into account. Finally, for those mutations not detected by both methods, total coverage and number of mutated reads were extracted in order to determine the cause for loss.

RESULTS

Development of a workflow for the detection of somatic mutations in tumor samples

An overview of the RFcaller's workflow is provided in Figure 1. The pipeline takes as input the BAM files from the normal–tumor paired samples and starts performing a basic variant calling using bcftools (v1.10.2) with the -P option set to 0.1 to enable calling positions with low VAF. Then, indels are normalized, and common SNPs (dbSNP v153), and variants within 5 bp of an indel, are removed. To increase the speed of the pipeline, low-quality calls are filtered (<15 for SSNVs and <40 for indels). Remaining mutations are divided into three different files to be processed independently: SSNVs, short indels (<7 bp) and long indels (≥7 bp).

SSNVs and indels have a specific pipeline where read-level features are extracted for those mutations that meet basic requirements that can be customized, such as having a minimum coverage (≥7), a maximum number of mutated reads in normal (≤3 for SSNVs and ≤2 for indels) or a minimum number of mutated reads in tumor (≥3 for SSNVs and ≥4 for indels). These filters have been chosen because positions that fail to meet these requirements cannot be confidently classified as *bona fide* mutations from the available

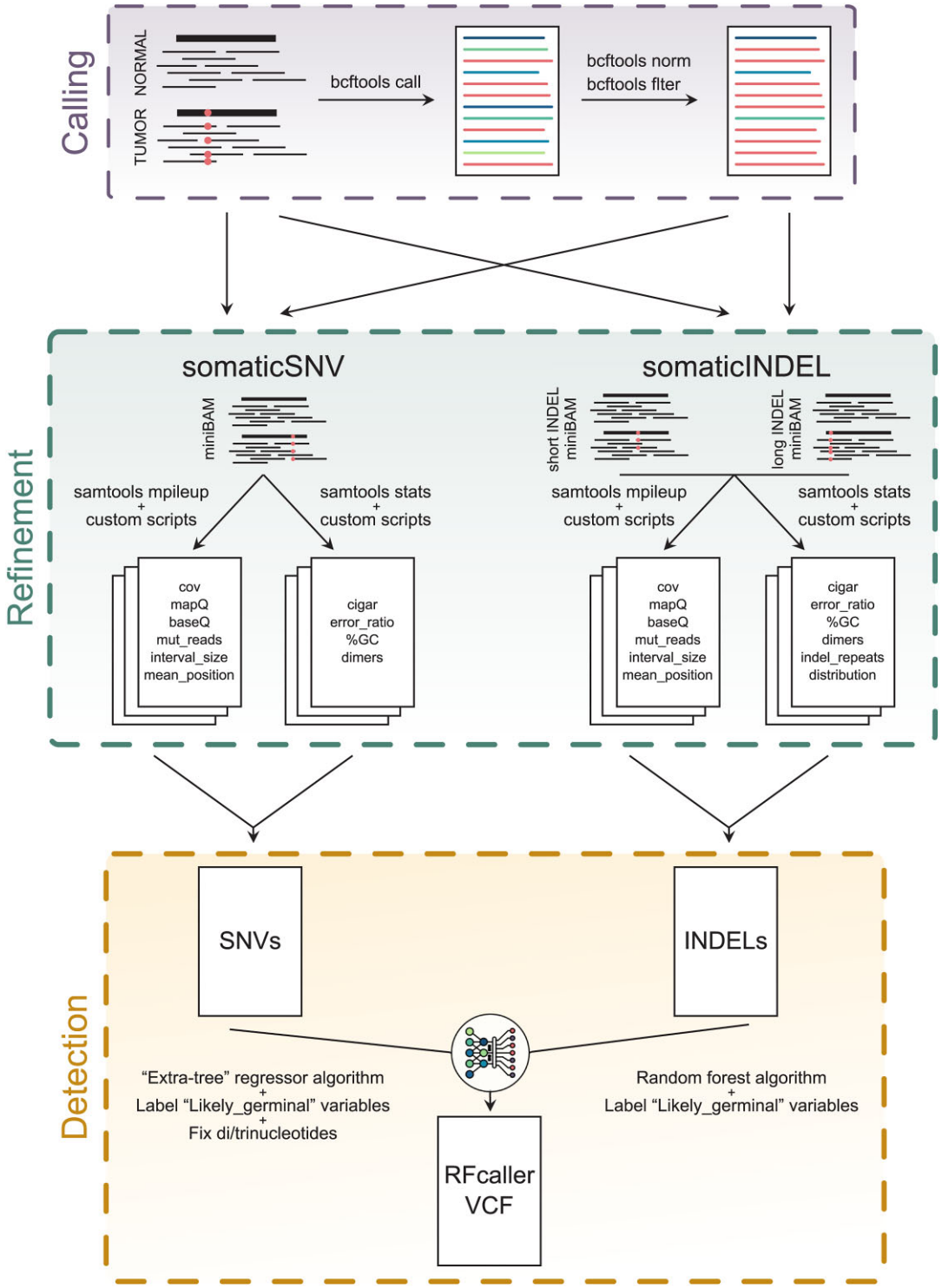

Figure 1. Flowchart of the RFcaller pipeline.

Table 1. Number of total and manually reviewed mutations used for training and testing RFcaller

	Training set				Test set			
	SSNV		Indel		SSNV		Indel	
	TP	TN	TP	TN	TP	TN	TP	TN
Manually reviewed	915	730	321	242	924	959	528	490
Total	8362	57 734	504	427	6909	57 039	696	1810

TP, number of true positive mutations; TN, number of true negative mutations.

data, which would require additional techniques to make a reliable call. These arguments can be modified from command line options.

Some of the selected features are very basic like the coverage or the number of mutated reads, considering in both cases only reliable reads (mapping quality $\geq$30). Other features are the average mapping quality when reads are extracted without mapping quality filters, for both normal and mutated reads independently. This is useful to check the quality of the region or if only mutated reads have lower mapping quality. The percentage of mismatched nucleotides around the mutations is also extracted together with CIGAR information. In this regard, a region with a higher mismatch value or lower number of matched bases is less confident than a region with a lower mismatch ratio and all their bases aligned properly. Another feature is the number of repeated dinucleotides that can be found around the mutations, which is used to know whether it is a microsatellite. Finally, the last two features are the distance between the leftmost and rightmost mutations in the reading and mean position of the mutation along the mutated reads. Manual review of discordant mutations during the first rounds of training allowed us to discover many false positive mutations located at one end of the read, which could be removed after adding these features. In addition, 11 more characteristics were extracted for indels to determine their complexity (Supplementary Table S4).

Once all features have been extracted, a CSV file is generated to be used by the algorithm. The result is a VCF file with the mutations that have passed the threshold for the 'QUAL' field.

To classify mutations that might be germinal but have passed the previous filters, a 95% confidence interval is applied to calculate the expected number of mutant reads in normal, considering the VAF of the mutation in tumor sample, the expected contamination of tumor in normal sample defined by the user and the normal coverage:

$$\text{expectedNormal}_{\text{VAF}} = \text{Tumor}_{\text{VAF}} * \text{Contamination}$$

$$z = 1.96 * \sqrt{\frac{\text{expectedNormal}_{\text{VAF}} * (1 - \text{expectedNormal}_{\text{VAF}})}{\text{Normal}_{\text{coverage}}}}$$

$$\text{maxNormal}_{\text{mut\_reads}} = (\text{expectedNormal}_{\text{VAF}} + \text{cte}) * \text{Normal}_{\text{coverage}}$$

Thus, if the number of mutated reads in normal is greater than the expected upper boundary of 95% confidence interval ($\text{maxNormal}_{\text{mut\_reads}}$), the position is labeled as 'LIKELY_GERMINAL'.

Finally, the RFcaller pipeline for SSNVs searches for dinucleotide or trinucleotide mutations within the results. With this step, if two mutations are found in *cis*, they are merged into a single mutation to be accurate when predicting its functional effect, a step that is usually missed by most commonly used somatic callers.

RFcaller training and computational cost comparison

For the initial training step, previous results from the genomic analysis of two mantle cell lymphomas (25) were used to annotate the set of mutations, and RFcaller was trained with this initial dataset. The obtained results were compared with those used for training, and all discordant positions were manually reviewed to improve the accuracy of the dataset. These steps were repeated until all discrepancies were classified by an expert panel. After that, 2208 and 2901 calls were reviewed for training and testing, respectively, resulting in a high-quality set of mutations to train and test the final versions of the algorithm (Table 1).

In order to select the best cutoff for the pipeline, SSNVs, indels and homopolymer indels were considered independently as they represent mutations whose detection is influenced by different features. The separation between both types of indels (isolated or within a homopolymer trait) was introduced due to the bias of the initial calling performed by bcftools against indels within homopolymeric tracts, giving very low scores to mutations that otherwise appear to be real. Furthermore, different formulas were considered to calculate the 'QUAL' threshold used by RFcaller (Supplementary Table S11).

Although the RFcaller score provided high accuracy, we observed that by combining the regression obtained by RFcaller with the score given by bcftools, the accuracy was improved over each one independently, suggesting both scores complement each other. We did not observe major differences between formulas for SSNVs and indels according to the AUC metric, so we selected the formulas with the highest *F*1 score. Thus, the cutoffs were 10.726 for SSNVs, 32.1418 for indels and 0.7723 for homopolymer indels (Supplementary Figure S1), which achieved 1.3%, 7.18% and 8% of false positive mutations, respectively. We observed that many of the false positives belonged to complex regions such as microsatellites or GC-rich sites, appearing also in normal samples from other donors. Therefore, we used a panel of normals to filter these calls and improve accuracy.

In terms of the number of variables selected, only 16 and 27 read-level features were considered for SSNVs and indels, respectively (Supplementary Table S4). They focused on specific properties such as coverage with certain base

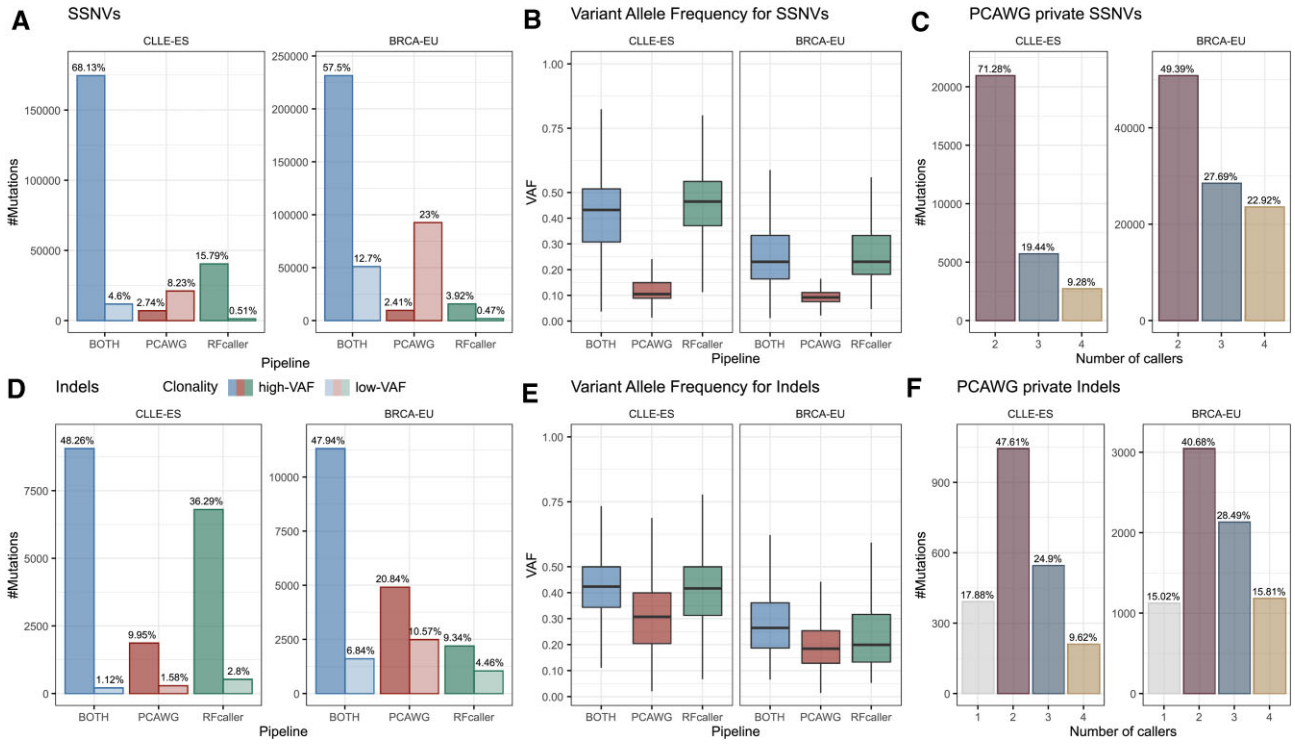

Figure 2. Summary of mutations detected by PCAWG and/or RFcaller pipelines for SSNVs and indels. (**A**, **D**) Classification of mutations according to the pipeline that can detect them. Mutations are divided in high-VAF (VAF $\geq$ 0.15) and low-VAF (VAF $<$ 0.15) mutations. (**B**, **E**) Distribution of the VAF of the mutations identified by both pipelines, or specifically by RFcaller or PCAWG pipeline. (**C**, **F**) Number of callers detecting each of the PCAWG-private mutations.

quality, CIGAR, mapping qualities or error ratio in tumor and normal samples, therefore avoiding overlapping features that can be counterproductive and lead to overfitting. In this regard, although most of the features selected by RFcaller were also included in other ML models, some of them also use many more features, such as NeuSomatic (23) or DeepSVR (18).

Another important aspect we considered during the selection of these features was the difficulty by which they can be extracted, resulting in a fast pipeline for medium-sized servers. Thus, the analysis of four WGS tumor–normal paired samples using 20 threads consumes only ∼5 GiB of RAM and takes ∼3 h per case, while using only 10 processors the analysis is extended up to ∼4.5 h per case (Supplementary Figure S2D).

When RFcaller was compared with the callers used by PCAWG for the detection of SSNVs, only the MuSE variant caller (∼2.5 h) was faster than RFcaller (∼4.8 h) (Supplementary Figure S2), while Sanger variant caller was the slowest, taking >70 h for a single case. In terms of memory consumption (RSS), Mutect variant caller is the most demanding, consuming between 100 and 250 GiB during half the time it is running (∼5 h). In this case, RFcaller and MuSE variant caller consume the least memory with an average of 5 GiB. It is important to note that although we have used the SSNV-specific callers, all of them, except MuSE, also detect indels, which would imply that RFcaller is the fastest and least resource-consuming tool for the simultaneous calling of SSNVs and indels.

Validation of RFcaller pipeline: PCAWG analysis

To test RFcaller against a validated set of cancer WGS cases, we used data from the PCAWG study. The pipeline employed by this consortium incorporates five of the best-performing and extensively tested variant callers, which also gave us the possibility to compare RFcaller with each individual tool. Specifically, two different projects (CLLE-ES and BRCA-EU), representative of liquid and solid tumors, with a total of 89 and 75 cases, respectively, were analyzed (Supplementary Table S8). RFcaller results were compared to those mutations labeled as 'PASS' by the PCAWG mutation calling pipeline. Due to the inherent differences between SSNVs and indels, we performed each analysis independently.

Statistics for SSNVs. After merging RFcaller and PCAWG 'PASS' mutations, we observed that ∼70% of SSNVs were detected by both pipelines in both studies. However, and even though the number of shared mutations was almost the same, for samples from the CLLE-ES project 11% of mutations were detected only by the PCAWG pipeline versus 16.3% mutations specifically detected by RFcaller. For BRCA-EU-derived mutations, only 4.4% mutations were RFcaller-specific, versus 25.4% for the PCAWG pipeline (Figure 2A). A detailed analysis of those differentially called mutations revealed that the mean VAF for SSNVs detected by both pipelines was 0.41 and 0.27 for CLLE-ES and BRCA-EU, respectively (Figure 2B). However, those detected by the PCAWG

Table 2. Distribution of SSNVs detected by each pipeline and extrapolated after manual revision

		SSNVs	TP	FP	TN	FN	Precision	Recall
CLLE-ES								
Private mutations	2 callers	20 956	16 097	4859			76.81%	
	3 callers	5715	5001	714			87.50%	
	4 callers	2729	2519	210			92.31%	
	RFcaller	41 772	41 153	619			98.52%	
	Both	186 361	186 361	0				
Total	PCAWG	257 533	209 977	5784	619	41 153	97.32%	83.61%
	RFcaller	257 533	227 514	619	5784	23 616	99.73%	90.60%
BRCA-EU								
Private mutations	2 callers	50 823	42 804	8019			84.22%	
	3 callers	28 490	26 469	2021			92.91%	
	4 callers	23 580	23 080	500			97.88%	
	RFcaller	17 699	12 613	5086			71.26%	
	Both	282 547	282 547	0				
Total	PCAWG	403 139	374 900	10 540	5086	12 613	97.27%	96.75%
	RFcaller	403 139	295 160	5086	10 540	92 353	98.31%	76.17%

TP, number of true positive SSNVs; FP, number of false positive SSNVs; TN, number of true negative SSNVs; FN, number of false negative SSNVs.

pipeline but not RFcaller had a mean VAF of 0.16 and 0.10 for CLLE-ES and BRCA-EU, respectively (Figure 2B), suggesting that they constitute low-VAF mutations. In fact, only 29% and 50% of them could be detected by more than two callers in the PCAWG pipeline for CLLE-ES and BRCA-EU, respectively (Figure 2C). Furthermore, those SSNVs detected by RFcaller but not the PCAWG pipeline had a mean VAF of 0.46 for CLLE-ES and 0.28 for BRCA-EU, similar to those detected by both pipelines, suggesting that they constitute high-VAF mutations detected by RFcaller. Some of them showed minor tumor in normal contamination (one to three mutant reads), common in hematological tumors, resulting in most callers missing these true positive somatic mutations, while RFcaller is able to retain most of them.

To explore the set of discordant mutations between both pipelines, we randomly selected 1–2% of the pipeline-private calls ($n = 776$ for CLLE-ES and $n = 1233$ for BRCA-EU) to be manually reviewed by a panel of experts (Supplementary Table S12 and Supplementary Figure S3). As expected, PCAWG-specific variants detected by four callers are more precise than those identified by two tools (Table 2). Surprisingly, the difference in precision for RFcaller-private mutations between studies was very high, 98.5% for CLLE-ES and 71.3% for BRCA-EU, probably reflecting the fact that RFcaller was trained using a hematological tumor. However, despite the apparently higher number of false positives, RFcaller-private calls only represent 18.3% and 5.9% of the total number of SSNVs detected in the CLLE-ES and BRCA-EU projects, respectively. Considering the observed number of false positives within these sets, the real precision of RFcaller for SSNVs is 99.7% and 98.3% for CLLE-ES and BRCA-EU, respectively, while the precision of the PCAWG pipeline is 97.3% for both studies (Table 2 and Figure 3). RFcaller performance is similar for tumors with different mutation burden.

The comparison between RFcaller and each of the individual tools used by PCAWG showed that these variant callers have slightly lower recall than the ensemble approach, as it was expected (Supplementary Table S13). In fact, in BRCA-EU where RFcaller recall was lower due to tumor purity, DKFZ and Sanger tools had similar recall to RFcaller for SSNVs (∼75%), MuSE and Mutect2 pipelines being the ones that identify most of the mutations detected by PCAWG (∼95%). In CLLE-ES, all the individual variant callers had lower recall than RFcaller for SSNVs (70–80% versus 90%). On the other hand, precision was very similar among all the callers for both studies and type of mutation (SSNVs or indels) (Supplementary Table S13 and Supplementary Figure S4).

After testing DeepSVR for CLLE-ES and BRCA-EU studies, using the same SSNV dataset as for RFcaller, it was observed that DeepSVR is not suitable for the refinement of basic variant callings, since its precision falls to 50% for both studies (Supplementary Table S14 and Supplementary Figure S5A). Moreover, when DeepSVR-private mutations were removed from the analysis, considering a reliable set of mutations previously detected by PCAWG and RFcaller, the precision of the tool was the same as for PCAWG and RFcaller (Supplementary Table S14 and Supplementary Figure S5B).

Statistics for small indels. The analysis of small indels revealed that there were more differences between pipelines than those seen for SSNVs. In this regard, only ∼50% of indels were detected by both RFcaller and PCAWG pipelines; however, for CLLE-ES, RFcaller-private calls represented 39.1% of the total number of indels, whereas only 11.5% of them were PCAWG-specific. In contrast, in BRCA-EU, RFcaller and PCAWG-private mutations accounted for 13.8% and 31.4%, respectively (Figure 2D). Moreover, among them, <45% of PCAWG-private indels were detected by more than two callers (Figure 2F), reflecting the difficulty to identify somatic indels in tumor samples.

To further explore pipeline-private indels, we selected at least 50 indels from each group for expert review ($n = 283$ for CLLE-ES and $n = 429$ for BRCA-EU) (Supplementary Table S12 and Supplementary Figure S3). We observed that the precision within PCAWG-private indels was very high, varying between 70% and 99% depending on the number of

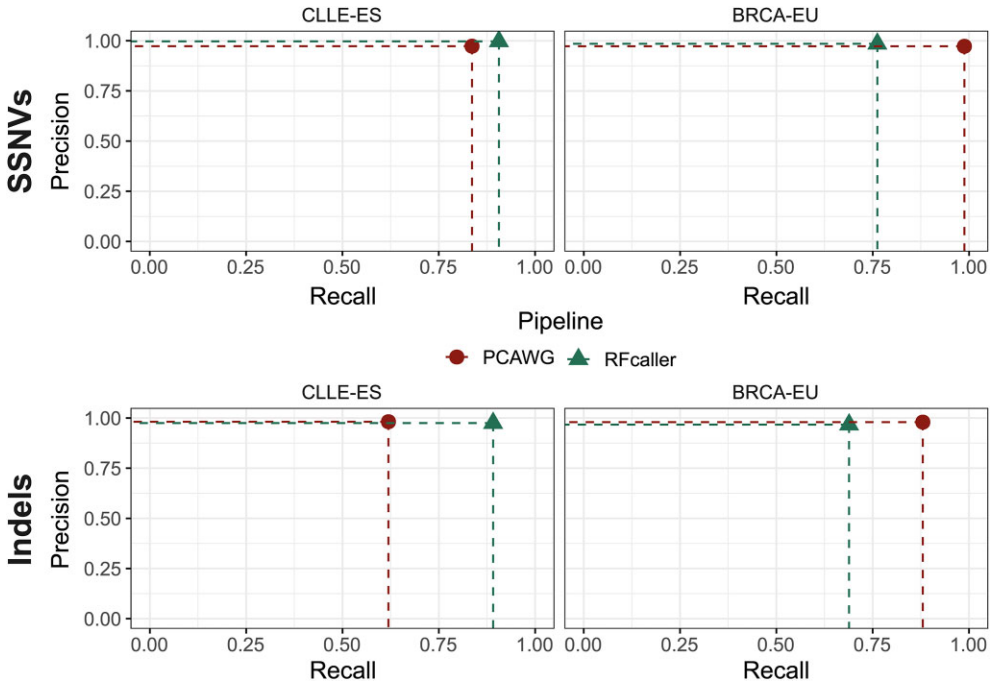

Figure 3. Accuracy of RFcaller and PCAWG pipelines for SSNVs and indels against CLLE-ES and BRCA-EU datasets. RFcaller shows a higher recall in both SSNVs and indels for CLLE-ES, whereas in BRCA-EU the PCAWG manages to detect a higher number of mutations. The precision of the two pipelines is similar in all conditions.

individual callers supporting the call (Table 3). In contrast, the precision observed for RFcaller was 89%, despite the fact that the total number of indels detected by this pipeline was much higher. Similar to SSNVs, the observed VAF was significantly higher in CLLE-ES compared to BRCA-EU (0.42 versus 0.29), probably reflecting higher tumor purity. Nonetheless, we did not observe differences in VAF between pipeline-private indels (Figure 2E), suggesting that pipeline-specific mutations were not due to their allele frequency, as they were for SSNVs, but due to other factors such as alignment issues, size of the indel, the presence of microsatellites or if they were within homopolymer tracks. Despite the higher precision obtained by the PCAWG pipeline for indel calling, this might be at the expense of a larger number of false negative calls in otherwise high-VAF and *bona fide* somatic indels, as shown by the number of true positive calls detected by RFcaller (Table 3 and Figure 3).

The performance of RFcaller against individual callers for indels was very similar to that of SSNVs. The MuSE pipeline was excluded from this analysis as it cannot detect indels. In CLLE-ES, the three callers (Mutect2, Sanger and DKFZ) showed lower recall than RFcaller (∼50% versus 89%), whereas in BRCA-EU only the Sanger pipeline (76%) had a better recall than RFcaller (69%), which was similar to the other two tools (64%) (Supplementary Table S13 and Supplementary Figure S4).

RFcaller performance on exome-derived data

RFcaller was trained with WGS data, but as the features used for the prediction are at read level, this pipeline could also be used for exome analysis. In order to test the ability of RFcaller to detect mutations by whole exome sequencing (WES), exomes from five cases previously analyzed by WGS were run with default parameters. Results were compared with those obtained by RFcaller and PCAWG in the WGS analysis after filtering for mutations within target regions in WES. Thus, 63% ($n = 110$) of mutations detected by WES were also detected by WGS. Additionally, we were able to identify 47 novel mutations for which there was neither coverage nor any mutated read in WGS (Figure 4A). When we made the comparison in the opposite direction, we found that 55% ($n = 136$) of the mutations detected by WGS did not appear by WES. However, 93% ($n = 126$) of these missing mutations had no coverage or any mutated read in the exome or were clearly germinal (Figure 4B). Only 10 mutations detected by WGS had enough coverage in WES and were not detected, constituting false negatives (RFcaller exome recall = 94%). Similarly, considering the 17 mutations that were labeled as germinal by WGS but detected by WES as false positives, RFcaller achieves a precision of 90% (RFcaller_exome_analysis.R, Supplementary Data).

Detection and verification of mutations in driver genes

From the above data we can conclude that RFcaller has a similar accuracy to detect SSNVs, and an increased sensitivity to detect indels at the cost of a slightly lower specificity. To explore whether these differences might allow the detection of previously missed mutations with potential clinical impact, we analyzed somatic mutations on the set of driver genes previously described in these two tumor types (34,35) (Supplementary Table S15). This analysis resulted in the identification of 155 coding mutations in driver genes in the CLLE-ES project and 162 in the BRCA-EU study. Out of those calls, 83% of them were shared between both

Table 3. Distribution of indels detected by each pipeline and extrapolated after manual revision

		Indels	TP	FP	TN	FN	Precision	Recall
CLLE-ES								
Private mutations	1 caller	392	276	116			70.37%	
	2 callers	1044	1002	42			96.00%	
	3 callers	546	495	51			90.74%	
	4 callers	211	207	4			98.18%	
	RFcaller	7335	6916	419			94.29%	
	Both	9267	9267	0				
Total	PCAWG	18 795	11 248	212	419	6916	98.15%	61.92%
	RFcaller	18 795	16 183	419	212	1981	97.48%	89.10%
BRCA-EU								
Private mutations	1 caller	1125	1008	117			89.56%	
	2 callers	3046	2792	254			91.67%	
	3 callers	2133	2097	36			98.31%	
	4 callers	1184	1174	10			99.14%	
	RFcaller	3257	2707	550			83.11%	
	Both	12 925	12 925	0				
Total	PCAWG	23 670	19 995	418	550	2707	97.95%	88.08%
	RFcaller	23 670	15 632	550	418	7070	96.60%	68.86%

TP, number of true positive indels; FP, number of false positive indels; TN, number of true negative indels; FN, number of false negative indels.

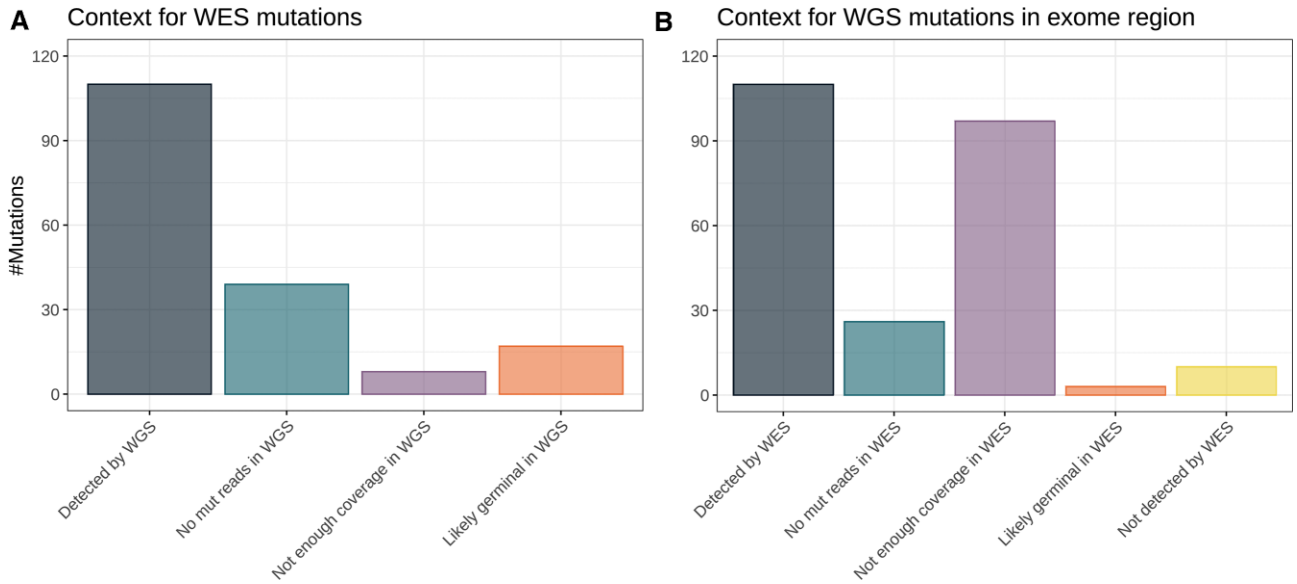

Figure 4. Comparison of mutations detected by analysis of WGS and WES in selected donors. Comparison is limited to exomic regions. (**A**) Mutations detected by WES and analysis of their status in WGS. (**B**) Mutations detected by WGS and analysis of their status in WES samples.

pipelines, while 53 (17%) in 35 driver genes were pipeline-specific (Supplementary Table S16).

Those pipeline-specific mutations were manually reviewed, resulting in the identification of 19 high-VAF mutations detected by RFcaller (12 SSNVs and 7 indels) versus four high-VAF SSNVs detected by the PCAWG pipeline in CLLE-ES. For the BRCA-EU project, eight high-VAF mutations were detected by RFcaller (five SSNVs and three indels) versus four high-VAF detected by the PCAWG pipeline (three SSNVs and one indel).

For seven private calls detected in the CLLE-ES study (five by RFcaller and two by the PCAWG pipeline), tumor and normal DNA was available for verification by Sanger sequencing, except two cases in which only tumor DNA was available (Supplementary Table S8). This analysis resulted in the verification of all RFcaller-private calls (Supplementary Figure S6), as well as one of the PCAWG-private SSNVs. The last call could not be verified because it was a very low-VAF mutation (8.7%), which falls below the detection limit of this technique.

To further perform an orthogonal validation of these pipelines, we took advantage of a previous study in which 26 CLL driver genes had been analyzed by deep sequencing in some of the CLL cases used by PCAWG (31,32). A total of 77 mutations, excluding germline calls, were detected in 28 cases, for which enough coverage was available in WGS to make a call (Supplementary Table S9). Due to the high depth of sequencing, VAF was very variable (range 0.0029–0.9665); therefore, mutations were classified as high-VAF mutations if VAF $\geq$ 0.15 (n = 44, median 0.43) and

low-VAF mutations if VAF < 0.15 ($n = 33$, median 0.03). As expected, most low-VAF mutations could not be detected from WGS data, as each pipeline was only able to detect 6/33 low-VAF mutations (18%). In contrast, most high-VAF mutations detected by deep sequencing could also be identified by RFcaller (39/44, 89%), while the performance of the PCAWG pipeline was slightly lower (31/44, 70%). The mutations specifically detected by RFcaller affected *NOTCH1* (3), *ATM* (2), *TP53*, *RPS15*, *MGA* and *DDX3X* (Supplementary Figure S3), some of which have been associated with poor prognosis and whose presence might impact clinical decisions. The PCAWG pipeline was able to identify a mutation in *ATM* (Supplementary Figure S3) that was not detected by RFcaller due to a very low VAF (0.065). Together, these results support the utility of RFcaller to identify novel high-VAF driver mutations of potential clinical value.

DISCUSSION

The application of NGS techniques for clinical diagnosis in tumor samples requires procedures that provide enough sensitivity and specificity, while at the same time do not require large computing resources to achieve the analysis in a reasonable amount of time. To increase accuracy, a final step of manual review through visual inspection is usually carried out for mutations that might be clinically informative. This manual revision increases the specificity, but at the cost of a labor-intensive process. Recent advances in machine learning approaches are suitable to incorporate features that experts consider when distinguishing between *bona fide* mutations and false positives. However, most available programs that use machine learning approaches for somatic mutation calling have been trained with high depth of coverage WES using *in silico* (19,20) or orthogonal validated mutations (21). Therefore, these programs cannot be used directly for the analysis of whole genome sequences, as they are not prepared for complex intronic or intergenic regions, nor modest coverage where their sensitivity is very low.

In this work, we have taken advantage of a manually curated dataset of real mutations with features that an expert curator might consider when manually reviewing a mutation in a research or clinical context. Thus, we have achieved a very high sensitivity to detect SSNVs and small indels, while at the same time maintaining a low footprint, with low CPU and RAM consumption, being able to analyze a whole genome in <5 h. Moreover, although it has been trained with WGS data, it has shown a good performance in exome samples.

On the other hand, even though our selected features are often used by similar programs, these tools usually process SSNVs and indels in the same manner, when clearly the two types of mutations have different characteristics. Additionally, instead of integrating as many features as possible to train the algorithms like other variant callers based on machine learning techniques (18,20), we have carefully selected the features used to train the algorithms, which simplify the models and improve their training and performance. We analyzed SSNVs and indels separately, which allowed us to detect indels with higher accuracy without affecting the ability to detect SSNVs. Indeed, we have shown that RFcaller performance is similar to that of a combination of complex pipelines used in the PCAWG project to detect high-VAF mutations, with the ability to detect new ones, some of them in driver genes, which might contribute to improve the detection of actionable mutations (36). Furthermore, we showed that RFcaller is able to detect mutations even in the presence of some tumor contamination in the normal sample, a common problem in some hematological tumors that might lead to false negatives with other pipelines. Finally, we have demonstrated that most RFcaller false negatives were very low-VAF mutations mainly associated with lower tumor purity. This suggests that RFcaller performance is better in high-purity tumor samples such as those derived from hematological tumors, while the overall performance in low-purity solid tumors is similar to the individual variant callers employed by PCAWG, which only achieves a better recall than RFcaller due to their ensemble pipeline. In this regard, RFcaller represents a fast and accurate tool for the detection of most mutations in tumor samples, allowing the detection of mutations in most driver genes, and could be complemented with additional tools focused on the detection of very low-VAF mutations that could be relevant in treatment resistance.

In conclusion, we have developed a pipeline called RFcaller that is provided under a docker system, which allows its easy and fast installation without version incompatibilities. This tool allows the identification of clonal mutations with the same efficiency as state-of-the-art pipelines, but with a smaller footprint in computing resources.

DATA AVAILABILITY

RFcaller and the scripts used to train the algorithms are available at the GitHub repository (https://github.com/xa-lab/RFcaller), and a docker with all the requirements and necessary files to run the pipeline has been built to improve reproducibility and facilitate the use of the program (https://hub.docker.com/repository/docker/labxa/rfcaller). Additionally, the scripts with the files we have used to obtain the results shown above can be found in the Supplementary Data.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGMENTS

E.C. is an Academia Researcher of the 'Institució Catalana de Recerca i Estudis Avançats' (ICREA) of the Generalitat de Catalunya. IUOPA is funded by the Asturian Government and Fundación Cajastur. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.
Authors' contributions: A.D.-N. developed the software, performed the bioinformatical work, interpreted data, designed the figures and wrote the manuscript. P.B.-M., F.N. and S.L.-T. contributed to data analysis and performed experimental work. S.B. and E.C. collected samples and interpreted data. X.S.P. designed the study, interpreted data,

wrote the manuscript and directed the research. All authors read, commented on and approved the manuscript.

FUNDING

Ministerio de Ciencia e Innovación [SAF2017-87811-R and PID2020-117185RB-I00 to X.S.P.]; Fundación Científica Asociación Española Contra el Cáncer (AECC) [to X.S.P. and S.B.]; Centro de Investigación Biomédica en Red de Cáncer (CIBERONC) [to X.S.P. and E.C.]; Instituto de Salud Carlos III and co-funded by European Union (ERDF/ESF, 'Investing in your future') [PMP15/00007 to E.C., PI17/01061 to S.B.]; 'La Caixa' Foundation CLLEvolution [HR17-00221 to E.C.]; Ministerio de Economía y Competitividad (MINECO) [RTI2018-094274-B-I00 to E.C.]; Generalitat de Catalunya AGAUR [2021-SGR-01293 to S.B., 2017-SGR-1142 to E.C.]; Department of Education of the Basque Government [PRE_2017_1_0100 to A.D.-N.]; Asturian Government [to S.L.-T.]; 2021 AACR-Amgen Fellowship in Clinical/Translational Cancer Research [21-40-11-NADE to F.N.]; European Hematology Association (EHA) Junior Research Grant 2021 [RG-202012-00245 to F.N.]; Lady Tata Memorial Trust (International Award for Research in Leukaemia 2021–2022) [LADY_TATA_21_3223 to F.N.].

Conflict of interest statement. X.S.P. is co-founder and equity holder of DREAMgenics. F.N. has received honoraria from Janssen, AbbVie and SOPHiA GENETICS for speaking at educational activities. E.C. has been a consultant for Takeda, AbbVie, Genmab, and Illumina; has received research support from AstraZeneca, and honoraria from Takeda, Bristol Meyier Squib, Janssen, and EUSPharma for speaking at educational activities; and is an inventor on a Lymphoma and Leukemia Molecular Profiling Project patent 'Method for subtyping lymphoma subtypes by means of expression profiling' licensed to NanoString Technologies.

REFERENCES

1. Mardis,E.R. (2019) The impact of next-generation sequencing on cancer genomics: from discovery to clinic. *Cold Spring Harb. Perspect. Med.*, **9**, a036269.
2. The International Cancer Genome Consortium (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
3. Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
4. Alexandrov,L.B., Kim,J., Haradhvala,N.J., Huang,M.N., Tian Ng,A.W., Wu,Y., Boot,A., Covington,K.R., Gordenin,D.A., Bergstrom,E.N. *et al.* (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94–101.
5. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
6. Li,Y., Roberts,N.D., Wala,J.A., Shapira,O., Schumacher,S.E., Kumar,K., Khurana,E., Waszak,S., Korbel,J.O., Haber,J.E. *et al.* (2020) Patterns of somatic structural variation in human cancer genomes. *Nature*, **578**, 112–121.
7. Rheinbay,E., Nielsen,M.M., Abascal,F., Wala,J.A., Shapira,O., Tiao,G., Hornshøj,H., Hess,J.M., Juul,R.I., Lin,Z. *et al.* (2020) Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, **578**, 102–111.
8. Jones,D., Raine,K.M., Davies,H., Tarpey,P.S., Butler,A.P., Teague,J.W., Nik-Zainal,S. and Campbell,P.J. (2016) cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics*, **56**, 15.10.1–15.10.18.
9. Benjamin,D., Sato,T., Cibulskis,K., Getz,G., Stewart,C. and Lichtenstein,L. (2019) Calling somatic SNVs and indels with Mutect2. bioRxiv doi: https://doi.org/10.1101/861054, 02 December 2019, preprint: not peer reviewed.
10. Fan,Y., Xi,L., Hughes,D.S.T., Zhang,J., Zhang,J., Futreal,P.A., Wheeler,D.A. and Wang,W. (2016) MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, **17**, 178.
11. Kim,S., Scheffler,K., Halpern,A.L., Bekritsky,M.A., Noh,E., Källberg,M., Chen,X., Kim,Y., Beyter,D., Krusche,P. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
12. Ye,K., Schulz,M.H., Long,Q., Apweiler,R. and Ning,Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
13. Moncunill,V., Gonzalez,S., Beà,S., Andrieux,L.O., Salaverria,I., Royo,C., Martinez,L., Puiggròs,M., Segura-Wang,M., Stütz,A.M. *et al.* (2014) Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.*, **32**, 1106–1112.
14. Ellrott,K., Bailey,M.H., Saksena,G., Covington,K.R., Kandoth,C., Stewart,C., Hess,J., Ma,S., Chiotti,K.E., McLellan,M. *et al.* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, **6**, 271–281.
15. Patel,L., Shukla,T., Huang,X., Ussery,D.W. and Wang,S. (2020) Machine learning methods in drug discovery. *Molecules*, **25**, E5277.
16. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
17. Fang,L.T., Afshar,P.T., Chhibber,A., Mohiyuddin,M., Fan,Y., Mu,J.C., Gibeling,G., Barr,S., Asadi,N.B., Gerstein,M.B. *et al.* (2015) An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.*, **16**, 197.
18. Ainscough,B.J., Barnell,E.K., Ronning,P., Campbell,K.M., Wagner,A.H., Fehniger,T.A., Dunn,G.P., Uppaluri,R., Govindan,R., Rohan,T.E. *et al.* (2018) A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.*, **50**, 1735–1743.
19. Anzar,I., Sverchkova,A., Stratford,R. and Clancy,T. (2019) NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. Genomics*, **12**, 63.
20. Wood,D.E., White,J.R., Georgiadis,A., Van Emburgh,B., Parpart-Li,S., Mitchell,J., Anagnostou,V., Niknafs,N., Karchin,R., Papp,E. *et al.* (2018) A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.*, **10**, eaar7939.
21. Spinella,J.-F., Mehanna,P., Vidal,R., Saillour,V., Cassart,P., Richer,C., Ouimet,M., Healy,J. and Sinnett,D. (2016) SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics*, **17**, 912.
22. Poplin,R., Chang,P.-C., Alexander,D., Schwartz,S., Colthurst,T., Ku,A., Newburger,D., Dijamco,J., Nguyen,N., Afshar,P.T. *et al.* (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, **36**, 983–987.
23. Sahraeian,S.M.E., Liu,R., Lau,B., Podesta,K., Mohiyuddin,M. and Lam,H.Y.K. (2019) Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.*, **10**, 1041.
24. Danecek,P., Schiffels,S. and Durbin,R. (2016) Multiallelic calling model in bcftools (-m). https://samtools.github.io/bcftools/call-m.pdf, (10 March 2023, date last accessed).
25. Nadeu,F., Martin-Garcia,D., Clot,G., Díaz-Navarro,A., Duran-Ferrer,M., Navarro,A., Vilarrasa-Blasi,R., Kulis,M., Royo,R., Gutiérrez-Abril,J. *et al.* (2020) Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. *Blood*, **136**, 1419–1432.
26. Le,T.T., Fu,W. and Moore,J.H. (2020) Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, **36**, 250–256.

27. López-Ratón,M., Rodríguez-Álvarez,M.X., Suárez,C.C. and Sampedro,F.G. (2014) OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *J. Stat. Softw.*, **61**, 1–36.
28. Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.
29. Xu,S., Chen,M., Feng,T., Zhan,L., Zhou,L. and Yu,G. (2021) Use ggbreak to effectively utilize plotting space to deal with large datasets and outliers. *Front. Genet.*, **12**, 774846.
30. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
31. Nadeu,F., Delgado,J., Royo,C., Baumann,T., Stankovic,T., Pinyol,M., Jares,P., Navarro,A., Martín-García,D., Beà,S. *et al.* (2016) Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood*, **127**, 2122–2130.
32. 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**, 645–653.
33. Puente,X.S., Beà,S., Valdés-Mas,R., Villamor,N., Gutiérrez-Abril,J., Martín-Subero,J.I., Munar,M., Rubio-Pérez,C., Jares,P., Aymerich,M. *et al.* (2015) Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **526**, 519–524.
34. Nik-Zainal,S., Davies,H., Staaf,J., Ramakrishna,M., Glodzik,D., Zou,X., Martincorena,I., Alexandrov,L.B., Martin,S., Wedge,D.C. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47–54.
35. Knisbacher,B.A., Lin,Z., Hahn,C.K., Nadeu,F., Duran-Ferrer,M., Stevenson,K.E., Tausch,E., Delgado,J., Barbera-Mourelle,A., Taylor-Weiner,A. *et al.* (2022) Molecular map of chronic lymphocytic leukemia and its impact on outcome. *Nat. Genet.*, **54**, 1664–1674.
36. Garcia-Prieto,C.A., Martínez-Jiménez,F., Valencia,A. and Porta-Pardo,E. (2022) Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. *Bioinformatics*, **38**, 3181–3191.