nature portfolio

Peer Review File

Comprehensive benchmarking of methods for mutation calling in circulating tumor DNA

Corresponding Author: Dr Anders Jacobsen Skanderup

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this study, the authors present a novel approach to build a cfDNA benchmarking dataset, and benchmark a set of variant callers, including cfDNA specific tools, on the dataset. Additionally, the authors develop a Random Forest model for filtering variants in cfDNA NGS data. This work is important as there is a need for well characterised cfDNA NGS data for benchmarking new tools. Furthermore, as cfDNA is increasingly investigated in cancer research, guidance for variant calling tools in cfDNA data is valuable. The approach adopted is thorough and mainly well explained and justified. While the manuscript is well written and clearly presented, we do have some comments for the authors to address and respond to.

1) The pseudo ground truth set.
The pseudo ground truth was derived from consensus calls of 5/8 (4/7 for INDELs) variant callers. This definition is circular when used to then determine performance, and indeed in figure S1, while there is significant overlap with tumour mutations, there is also a large fraction which don't overlap.
Determining a ground truth/gold standard is hard, experimentally validating all calls is generally infeasible. However, more justification and exploration of different methods of obtaining a truth set should be given, as well as a discussion of these limitations.
In particular, how did the authors arrive at the minimal number of callers (5/8 and 4/7)?
How does comparing to databases like COSMIC and TCGA, or considering variants commonly detected across patients within this dataset, improve the truth set and comparisons between the truth set and the tumour mutations in S1?
Can you discuss the limitations of your chosen approach.

2) Synthetic dataset
The dataset was generated in-silico, and is therefore a form of synthetic data. Synthetic datasets do have inherent disadvantages in benchmarking studies. While the approach used generates a useful cfDNA benchmarking resource, and its certainly more realistic than entirely synthetic data, it is important to discuss that the data is ultimately synthetic and comment on the drawbacks of this.

3) Machine Learning section
A random forest model was used, there's no justification given for this choice. Were any other models considered, or is there an objective reason to choose random forests?

Additionally more detail is required for the methods used to train the model. What was the train/test split? Were there any issues regarding imbalance in classes of training data? What were all the features used? A brief description of each feature and acronym is necessary - without this, Figure 5 is difficult to interpret as in contains lots of undefined acronyms.

4) Supplementary Table 4
The authors should consider putting this table, a compressed version, or a more detailed summary, in the main text. This table contains informative data for a benchmarking exercise for readers - compute time and resources in particular would be of value for bioinformaticians considering which tools to use.

5) Some minor typos
L380 - Our analysis [let led] us to derive practical guidelines for users to optimize
L375 - Indeed, 91% of verified mutations [happened overlapped] annotated germline mutations in dbSNP [50].

(Remarks on code availability)
The code is well documented with a detailed README file. We were able to install the tool, although no further testing was performed.

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
Remarks made by Co-reviewer

Reviewer #3

(Remarks to the Author)
The study by Carre et al. introduces a novel framework for benchmarking somatic variant calling algorithms in cell-free DNA (cfDNA) from cancer patients,addressing the challenge of low ctDNA levels and DNA degradation. Using longitudinal samples from colorectal and breast cancer patients, the authors generated patient-matched dilution series that reflect true cfDNA characteristics. They benchmarked nine variant callers using deep WGS and ultra-deep WES, creating a reference set of ~95,000 variants. Performance was assessed across varying ctDNA levels and sequencing depths, with additional machine learning analyses identifying features to improve caller accuracy. Below are some comments to be addressed:

Major points:
1. When establishing the ground truth using a sample with a matched tissue sample, the author report only the p-value. It's important to report also the actual numbers. That is, the number of point mutations and\or indels in ctDNA, and in the tissue, as well as their overlap. A venn diagram can be informative for presenting this data. Also, can the authors speculate about variants not in the overlap? In other words, what is the source of noise for variants detected only in ctDNA, and why are some variants found only in tissue? Is the latter only a matter of low frequency?
2. In continuation to point #1, when evaluating the different methods and reporting the AUPRC, it would be informative for the reader to have the actual number of true mutations that the methods are aiming for.
3. More generally, can the authors speculate on the different features implemented in the different methods that result with the observed differences?
4. It is common in variant calling to apply a Panel of Normal filtering. While one of the methods (ABEMUS) had this requirement built-in, it is worth applying it to the other methods as well following the initial calling to make it more comparable.
5. A section in the method explaining the entire machine learning (i.e., method-tuned) is missing.
6. In relation to the previous point, it is worth applying better or more advanced models such as XGBoost or LightGBM to truly appreciate the potential of this framework.

Minor points:
1. Can the authors explain the sentence on lines 173-175? How come only one read is sufficient for mutation detection? Most methods requires at least 3. This seems like a misunderstanding on my side and I'd appreciate clarification.
2. Lines 214-215 - Mutect2 and VarDict *were* able to outperform.
3. In the PDF, the equations appear as blank squares.

(Remarks on code availability)
The code is still not available.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
All comments have been thoroughly addressed including new analysis which supports findings.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
Code previously reviewed

Reviewer #3

(Remarks to the Author)
The authors have adequately addressed my comments and I suggest accepting the manuscript in its current form.

(Remarks on code availability)

General response to reviewers

We thank the reviewers for their thoughtful and constructive feedback, which has substantially improved the manuscript. In response, we expanded the dataset and added new analyses, and we revised the text and figures for clarity and fairness. Key improvements include:

- **Ground truth evaluation.** We expanded and clarified the ground-truth section, providing a new in-depth analysis of thresholds for constructing the consensus ground truth mutation sets (Supplementary Fig. 2).
- **Model fine-tuning potential.** We added justification for the choice of the Random Forest model for evaluating feature importance and method fine tuning potential (Fig. 5d).
- **Panel-of-Normals (PoN) filtering.** Following Reviewer #3's idea, we implemented a caller-agnostic PoN post-filter and evaluated its impact when applied uniformly across callers.

Below, we provide a point-by-point response to each comment, with specific references to the revised text, figures, and supplementary materials.

Point-by-point response to reviewers

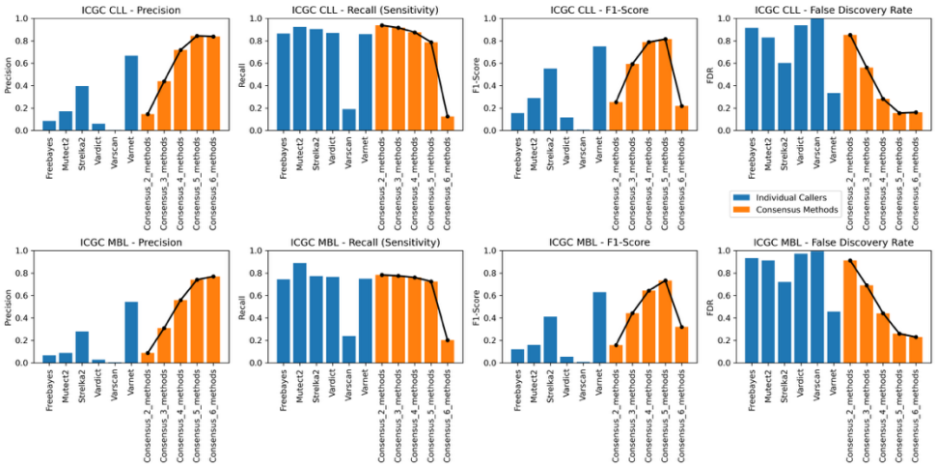
Reviewer 1 comments

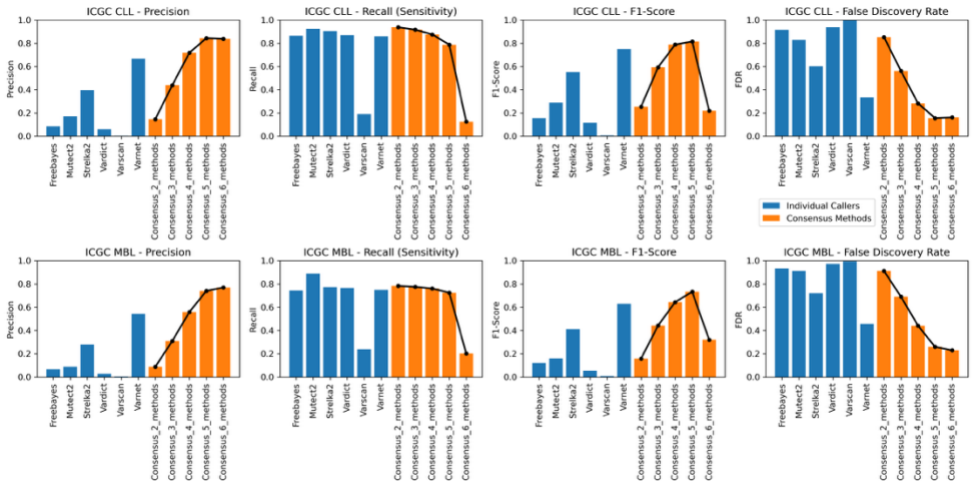
Reviewer 1, General Comments

Reviewer Comment	In this study, the authors present a novel approach to build a cfDNA benchmarking dataset, and benchmark a set of variant callers, including cfDNA specific tools, on the dataset. Additionally, the authors develop a Random Forest model for filtering variants in cfDNA NGS data. This work is important as there is a need for well characterised cfDNA NGS data for benchmarking new tools. Furthermore, as cfDNA is increasingly investigated in cancer research, guidance for variant calling tools in cfDNA data is valuable. The approach adopted is thorough and mainly well explained and justified. While the manuscript is well written and clearly presented, we do have some comments for the authors to address and respond to.
Author Response	We appreciate the reviewer's detailed feedback and excellent suggestions for improvements to the manuscript. Below, we provide a point-by-point response to the comments.

Reviewer 1, Comment 1: The pseudo ground truth set

Reviewer Comment	The pseudo ground truth was derived from consensus calls of 5/8 (4/7 for INDELs) variant callers. This definition is circular when used to then determine performance, and indeed in Supplementary Figure 2, while there is significant overlap with tumour mutations, there is also a large fraction which don't overlap. Determining a ground truth/gold standard is hard, experimentally validating all calls is generally infeasible. However, more justification and exploration of different methods of obtaining a truth set should be given, as well as a discussion of these limitations. In particular, how did the authors arrive at the minimal number of callers (5/8 and 4/7)? How does comparing to databases like COSMIC and TCGA, or considering variants commonly detected across patients within this dataset, improve the truth set and comparisons between the truth set and the tumour mutations in S1? Can you discuss the limitations of your chosen approach?

Author Response	We thank the reviewer for raising this important point. Our approach is based on majority voting: selecting mutations that are predicted by a majority of the methods (5/8 for SNVs, and 4/7 for indels). We reasoned that this would be an unbiased and conservative approach to define the ground truth sets under the expectation that the different methods often yield independent errors. However, we agree with the reviewer that it is worthwhile to explore how other selection approaches could have affected the accuracy of this consensus ground truth set. In a new analysis we evaluated how different ensemble voting schemes ('k of N') could affect the accuracy of somatic mutation call-sets. We based this analysis on ground-truth mutation call-sets from the ICGC (Alioto et al. 2015). This dataset comprises curated ground truth mutations in two tumor/normal sample pairs (MBL and CLL) with high-depth (>100x) whole genome sequencing. By applying our 6 tissue-based variant calling methods on this dataset, we found that the highest accuracy (F1) was achieved with k=4/6 or k=5/6 (Fig. S1a). Notable, at k=6 (intersection of all methods) there was a substantial loss in both sensitivity and accuracy. Mapping proportionally to our 8-caller cfDNA panel yields 5/8 (and 4/7 for indels). These results support our initial hypothesis that a majority-voting approach could be a rational and unbiased approach to define ground truth mutations in our cell-free DNA datasets.  **Supplementary Figure 1: Evaluation of k-of-N consensus calling in ICGC benchmark datasets.** Precision/Recall/F1/FDR for individual callers (blue) and consensus approaches (orange, k-of-N); elbow at 4/6 and 5/6 (F1 and sensitivity) supports a majority voting scheme.
Changes to manuscript	[**line 159-166, excerpt in *Generation of ground truths labels using consensus calling***]

	"We used the high-TF sample from each patient to generate ground truth variant calls using a majority-voting scheme (Fig. 1d, Methods). We reasoned that a majority-voting approach would be an unbiased and conservative approach to define the ground truth sets under the expectation that the different methods often yield independent errors. We explored this assumption in tumor tissue mutation benchmark datasets from ICGC (Alioto et al. 2015). We conducted an evaluation of distinct ensemble voting schemes ('k of N') in this dataset and found that a majority-voting approach yielded the best balance of accuracy and sensitivity (Supplementary Fig. 1)." **[line 913-916, excerpt in *Supplementary Information*]**  ***Supplementary Figure 1: Evaluation of k-of-N consensus calling in ICGC benchmark datasets.*** *Precision/Recall/F1/FDR for individual callers (blue) and consensus approaches (orange, k-of-N); elbow at 4/6 and 5/6 (F1 and sensitivity) supports a majority voting scheme.*

Reviewer 1, Comment 2: Synthetic dataset

Reviewer Comment	The dataset was generated in-silico, and is therefore a form of synthetic data. Synthetic datasets do have inherent disadvantages in benchmarking studies. While the approach used generates a useful cfDNA benchmarking resource, and its certainly more realistic than entirely synthetic data, it is important to discuss that the data is ultimately synthetic and comment on the drawbacks of this.
Author Response	We thank the reviewer for raising this point. We agree that's important to highlight advantages and disadvantages of these approaches. Although

	our benchmark approach can be classified as synthetic — created by in-silico mixing of patient-matched cfDNA samples — it originates entirely from real plasma and thus preserves authentic fragmentation patterns and sequencing error profiles. This is in contrast to the existing SEQC2 benchmark where cell-line DNA enzymatically sheared to mimic/approximate cfDNA, with 91 % of verified sites overlapping dbSNP germline variants (and thus generated via non-cancer mutation processes). By contrast, our design maintains bona-fide fragmentation patterns and somatic mutations while enabling precise, noise-controlled dilutions that are critical for fair comparison of variant-calling algorithms. We have further highlighted the limitations in the discussion of the revised manuscript.
Changes to manuscript	**[line 428-432, excerpt in *Discussion*]** *"Nevertheless, our in-silico high/low-TF dilution design also has key limitations. For example, the low-TF sample imposes a lower limit for the testable range of mutation VAFs. Additionally, our design does not allow for modelling of pre-analytical factors (e.g. extraction or storage-induced artefacts) that could influence cfDNA analyses."*

Reviewer 1, Comment 3: Machine Learning section

Reviewer Comment	A random forest model was used, there's no justification given for this choice. Were any other models considered, or is there an objective reason to choose random forests? Additionally more detail is required for the methods used to train the model. What was the train/test split? Were there any issues regarding imbalance in classes of training data? What were all the features used? A brief description of each feature and acronym is necessary - without this, Figure 5 is difficult to interpret as in contains lots of undefined acronyms.
Author Response	We thank the reviewer for this feedback. Our intent with the model fine-tuning section was two-fold: i) to evaluate the extent that tissue-based methods can be further improved (fine-tuned) for cfDNA variant calling; ii) to determine which features contribute towards this improvement. We chose to evaluate these questions with a Random Forest model as it is a common and robust first choice when modelling tabular datasets.

	Moreover, there are established methods to evaluate feature importance for this model type. To further address the reviewers question, we explored the performance of other machine learning models (logistic regression, support vector machines, and decision trees) in the revised manuscript. This analysis showed that Random Forests provided superior out-of-the-box performance (see new Supplementary Table 4 for benchmarking results). Random Forests were also preferred because they provide internal feature importance scores, which facilitated interpretation of predictive signals (see Methods on 'feature importance analysis'). While single decision trees would allow more granular interpretability, their performance was insufficient for this task. In addition, we have expanded the Methods to describe the training procedure, including the train/test split, how class imbalance was addressed (see 'Model training and evaluation' section). Figure 5 has been updated with a color-coded classification of acronyms by feature category (e.g., mapping quality, strand bias), with full feature names and descriptions in Supplementary Table 5.
Changes to manuscript	- *Figure 5 now uses color-coding to indicate each feature's category.* - *Supplementary Table 4 and 5 were added.* - *Methods sections entitled 'Model training and evaluation' and 'Features and feature importance analysis' were added.* **[line 827-844, excerpt in *Methods*]**

	Model training and evaluation. We evaluated several supervised classification algorithms using the scikit-learn v0.21.3 Python library. The models tested included decision trees (using Gini impurity), random forests (100 estimators), gradient boosting (100 estimators), logistic regression with L1 (Lasso), L2 (Ridge), and ElasticNet penalties (liblinear or saga solvers), and a multilayer perceptron (MLP) with one hidden layer of 100 neurons, ReLU activation, a learning rate of 0.001, batch size of 200, and a maximum of 100 iterations. Models were trained on the combined 150× WGS dataset from three colorectal cancer patients. We applied a leave-one-subject-out cross-validation strategy for the train/test split. We used stratified 10-fold cross-validation to preserve class ratios and report average precision averaged across folds. Performance results are presented in Supplementary Table 4. **Feature importance analysis**. We used features other than standard read counts available in each variant caller's VCF file to train the models. Feature counts per caller were: FreeBayes (37), Mutect2 (21), Strelka2 (19), VarDict (9), VarScan (5), and SMuRF (30). To interpret the random forest classifier, we computed feature importance using the Mean Decrease in Impurity (MDI). This metric quantifies the reduction in Gini impurity associated with each feature across all splits and trees in the ensemble. While informative, MDI is known to favor high-cardinality features and primarily reflects performance on the training set, rather than generalizability. Most important input features and their definitions are listed in Supplementary Table 5 and summarized in Fig. 5.

Reviewer 1, Comment 4: Supplementary Table 4

Reviewer Comment	The authors should consider putting this table, a compressed version, or a more detailed summary, in the main text. This table contains informative data for a benchmarking exercise for readers - compute time and resources in particular would be of value for bioinformaticians considering which tools to use.
Author Response	We thank the reviewer for this excellent idea. We have included a compact version of this table in the main text, which retains the benchmarking information of practical value to bioinformaticians.
Changes to manuscript	[**line 252-258, excerpt in *Variant calling accuracy at 150x sequencing depth***]

	Former Supplementary Table 4 moved to Table 1 in main manuscript

Reviewer 1, Comment 5: Some minor typos

Reviewer Comment	L380 - Our analysis [let led] us to derive practical guidelines for users to optimize L375 - Indeed, 91% of verified mutations [happened overlapped] annotated germline mutations in dbSNP [50].
Author Response	We thank the reviewer for their careful reading. These two typos have been corrected in the revised manuscript (now line 434 and 425).

Reviewer 1, Comment on Code availability

Reviewer Comment	The code is well documented with a detailed README file. We were able to install the tool, although no further testing was performed.
Author Response	We thank the reviewer for taking the time to test the installation and for acknowledging the clarity of the documentation. We appreciate the positive feedback on the README file and are glad to hear that installation was successful.

Reviewer 2 comments

Reviewer 2, General Comments

Reviewer Comment	I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.
Author Response	We thank the reviewer for sharing this information and fully support the Nature Communications initiative to recognize and train Early Career

	Researchers through co-reviewing. We appreciate the collaborative effort that contributed to this thorough review process.

Reviewer 2, Comment on Code availability

Reviewer Comment	Remarks made by Co-reviewer
Author Response	Please refer to the co-reviewer's remarks on code availability above, where we provide our response regarding code documentation and accessibility.

Reviewer 3 comments

Reviewer 3, General Comments

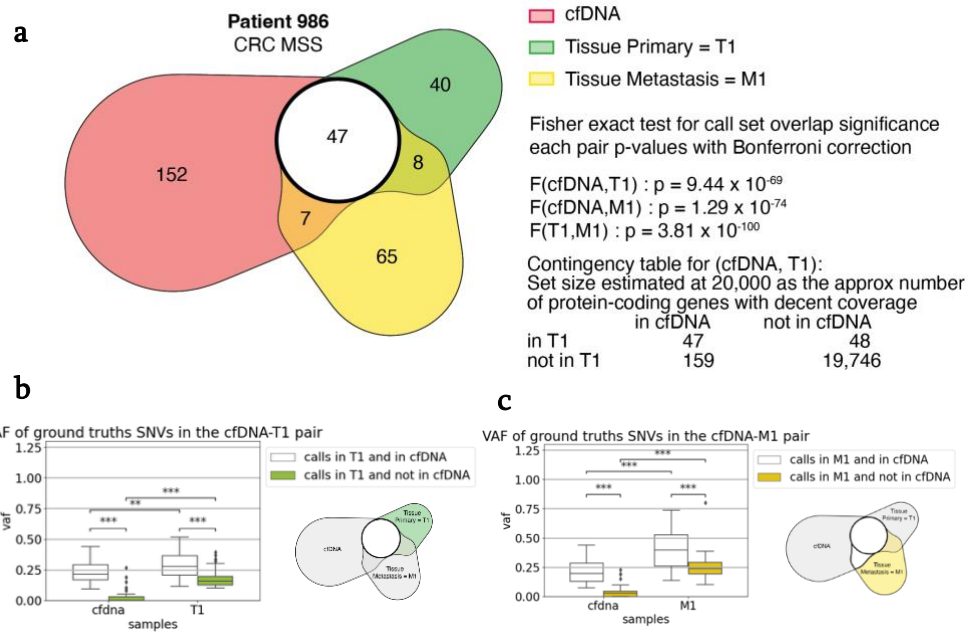
Reviewer Comment	The study by Carrie et al. introduces a novel framework for benchmarking somatic variant calling algorithms in cell-free DNA (cfDNA) from cancer patients,addressing the challenge of low ctDNA levels and DNA degradation. Using longitudinal samples from colorectal and breast cancer patients, the authors generated patient-matched dilution series that reflect true cfDNA characteristics. They benchmarked nine variant callers using deep WGS and ultra-deep WES, creating a reference set of ~95,000 variants. Performance was assessed across varying ctDNA levels and sequencing depths, with additional machine learning analyses identifying features to improve caller accuracy. Below are some comments to be addressed.
Author Response	We thank the reviewer for the thoughtful summary of our work and for recognizing the novelty and relevance of our benchmarking framework using longitudinal cfDNA samples. We appreciate the clear articulation of the study's contributions, please see below for our detailed responses to the specific comments.

Reviewer 3, Major Comment 1

Reviewer Comment	When establishing the ground truth using a sample with a matched tissue sample, the author report only the p-value. It's important to report also the actual numbers. That is, the number of point mutations and\or indels in ctDNA, and in the tissue, as well as their overlap. A venn diagram can be informative for presenting this data. Also, can the authors speculate about variants not in the overlap? In other words, what is the source of noise for variants detected only in ctDNA, and why are some variants found only in tissue? Is the latter only a matter of low frequency?
Author Response	Thank you for this important suggestion. We agree that absolute counts and overlaps should be made explicit. **Counts and presentation.** We now report the exact numbers alongside the p-values. For the patient with matched tissue (CRC_986), the high-ctDNA plasma consensus set contains 206 SNVs; deep WGS of the primary (T1) and metastatic (M1) tissues contains 95 and 120 SNVs, respectively. The overlaps with plasma are 47 (T1∩plasma) and 54 (M1∩plasma), both highly significant after Bonferroni correction (Fisher's exact $p < 10^{-68}$ and $p < 10^{-74}$). We retain a Venn diagram with counts in Supplementary Fig. 2a and now also state these numbers in the Results for visibility. **Interpretation of non-overlap.** To clarify why some variants are tissue-only or ctDNA-only, we added Supplementary Fig. 2b–c, showing VAF distributions for tissue-specific variants in T1 and M1. Variants absent from plasma have markedly lower tissue VAFs and ultra-low/undetectable VAFs in plasma, consistent with sub-clonality and cfDNA detection limits. Additional potential contributors include spatial/temporal tumor heterogeneity (unsampled lesions, evolution between biopsies and blood draw), variable shedding and copy-number context, and technical factors (effective depth, capture/GC biases, caller thresholds). We updated the figure legend and Results accordingly.
Changes to manuscript	**[line 167-178, excerpt in *Generation of ground truths labels using consensus calling*]** *"When a matched tumor tissue was not available, we hypothesized that the high-ctDNA (high-TF) plasma sample could serve as a proxy to derive high-confidence truth labels. Accordingly, for each patient we constructed ground-truth variant sets from the high-TF plasma by consensus across methods (Fig. 1d; Methods). We validated this assumption in the patient with matched tissue (patient 986): the high-TF plasma contained 206 SNVs, while deep WGS of the primary (T1) and metastatic (M1) tissues contained 95 and 120 SNVs, respectively. The overlaps with plasma were*

47 (T1∩plasma) and 54 (M1∩plasma), both highly significant after Bonferroni correction (Fisher's exact test $p<10^{-68}$ and $p<10^{-74}$; Supplementary Fig. 2a), supporting the reliability of the plasma-derived ground truth. Tissue-specific (T1- or M1-only) calls typically had low VAFs, consistent with subclonality, and were rarely detectable in plasma (Supplementary Fig. 2b-c)."

[line 917-932, excerpt in ***Supplementary Information***]
Panels b and c added to Supplementary Figure 2.



Supplementary Figure 2: Pseudo ground truth concordance between cell-free DNA (100x, WGS) and tumor tissue (100x, WGS), restricted to the exome.
(a) *Venn diagram showing the overlap in consensus exomic calls using the high ctDNA sample (K = 5 callers out of 8) vs. the tumor - primary and/or metastasis - tissue samples (K = 3 callers out of 6) for colorectal cancer patient 986. SMuRF is excluded from ground truth generation as it is an ensemble caller. For tissue samples, cfDNA- specific methods ABEMUS and SiNVICT are not applied. Fisher exact test based on contingency tables is applied on each pair of call sets to quantify overlap significance.* ***(b-c)*** *VAF distributions for tissue-specific ground-truth SNVs in T1 (b) and M1 (c). White boxplots denote variants also detected in high-ctDNA plasma; colored boxplots denote variants absent from plasma. Tissue-only variants show lower VAFs in tissue - consistent with subclonality - and ultra-low/zero VAFs in plasma, explaining non-detection. ctDNA: circulating tumor DNA, VAF: Variant Allele Frequency.*

Reviewer 3, Major Comment 2

Reviewer Comment	In continuation to point #1, when evaluating the different methods and reporting the AUPRC, it would be informative for the reader to have the actual number of true mutations that the methods are aiming for.
Author Response	Thank you for the helpful suggestion. To make the benchmark population size explicit, we now report the number of ground-truth somatic mutations (n) for each patient directly in Figure 3 for SNVs and in Supplementary Figure 4 for INDELs. We also specify the pooled number of ground truths to the reader.
Changes to manuscript	**[line 261-274, excerpt in *Variant calling accuracy at 150x sequencing depth*]** ***Addition to Figure 3**: Box with the number of true SNVs per patient.* ***Addition to the legend of Fig. 3**: "SNV ground-truth counts (n) per patient are shown in the small box: CRC_986 n=206, CRC_1014 n=361, CRC_123 n=453, BRA_412 n=285 (pooled N=1,305). Here, n denotes the number of true somatic mutations in the benchmark for that patient."* **[line 962-977, excerpt in *Supplementary Information*]** ***Addition to Supplementary Figure 5**: Box with the number of true INDELs per patient.* ***Addition to the legend of Supplementary Fig. 5**: "INDEL ground-truth counts (n) per patient are shown in the small box: CRC_986 n= 96, CRC_1014 n=277, CRC_123 n=357, BRA_412 n= 79 (pooled N= 809). Here, n denotes the number of true somatic INDELs in the benchmark for that patient."*

Reviewer 3, Major Comment 3

Reviewer Comment	More generally, can the authors speculate on the different features implemented in the different methods that result with the observed differences?

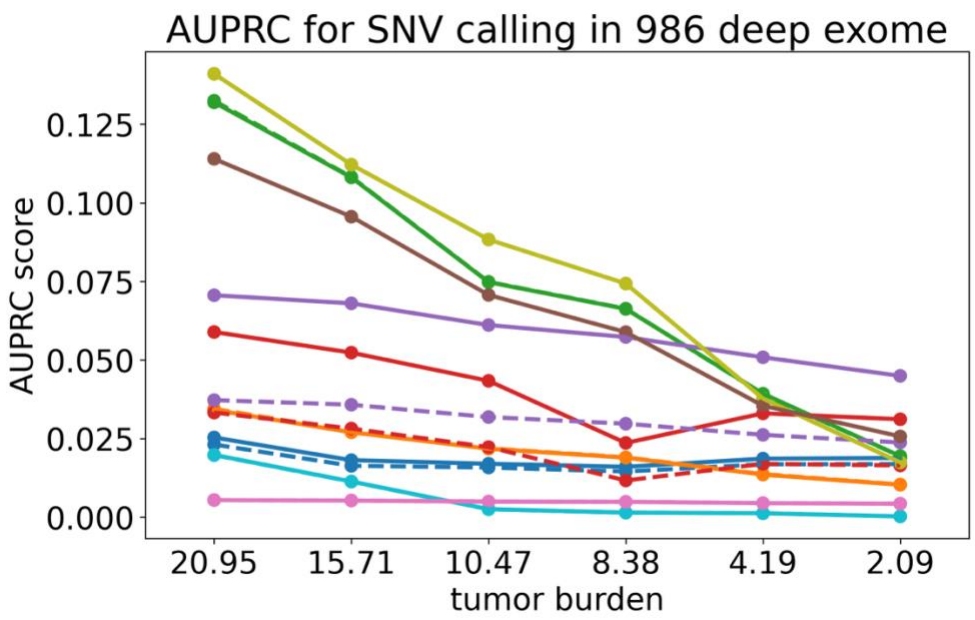
Author Response	We thank the reviewer for raising this point. In our view, performance differences stem from (i) whether evidence is modeled by local assembly or pile-up; (ii) the presence and granularity of background-error priors (matched normal, PoN, sequence-context/read-position/orientation-bias models); and (iii) the use and training domain of machine-learned calibration. Assembly-plus-priors or learned scoring (Strelka2, Mutect2 with artifact filters, SMuRF) appears to preserve precision as ctDNA levels drop (higher AUPRC in discovery), whereas pile-up-centric approaches (VarScan, VarDict) trade precision for sensitivity (higher recall at fixed 3% precision in clinical genotyping). cfDNA-specific callers calibrated for UMI-based, ultra-deep targeted assays (ABEMUS, SiNVICT) underperform at the intermediate-depth WES/WGS used here, consistent with domain mismatch. We have incorporated these observations in the revised manuscript.
Changes to manuscript	**[line 235-251, excerpt in the *Variant calling accuracy at 150x sequencing depth*]** *"We hypothesize that the observed ranking could reflects three design choices: (i) evidence representation (local haplotype assembly vs. pile-up counting); (ii) background-error priors (use of matched normal, panel-of-normals, and sequence-context/read-position/orientation-bias models); and (iii) training/calibration, including whether scoring is rule-based or machine-learned and the domain on which it was trained. Methods that couple assembly with rich priors or learned scoring (e.g., Strelka2, Mutect2 with artifact filters, SMuRF) better preserve precision as ctDNA levels decline, yielding higher AUPRCs in the discovery analysis. In contrast, pile-up–centric callers (VarScan, VarDict) prioritize recall via simpler thresholds and statistical tests, which benefits the fixed-precision clinical genotyping readout but produces an earlier precision drop in discovery. cfDNA-specific tools tuned for UMI-enabled, ultra-deep targeted panels (ABEMUS, SiNVICT) may be disadvantaged on the intermediate-depth exome/genome data used here, consistent with calibration to a different library design. Indel accuracy tracks the availability of local realignment/assembly and realignment-aware features, potentially explaining VarScan and SMuRF's stronger recall and Strelka2's more conservative performance as signal decreases."*

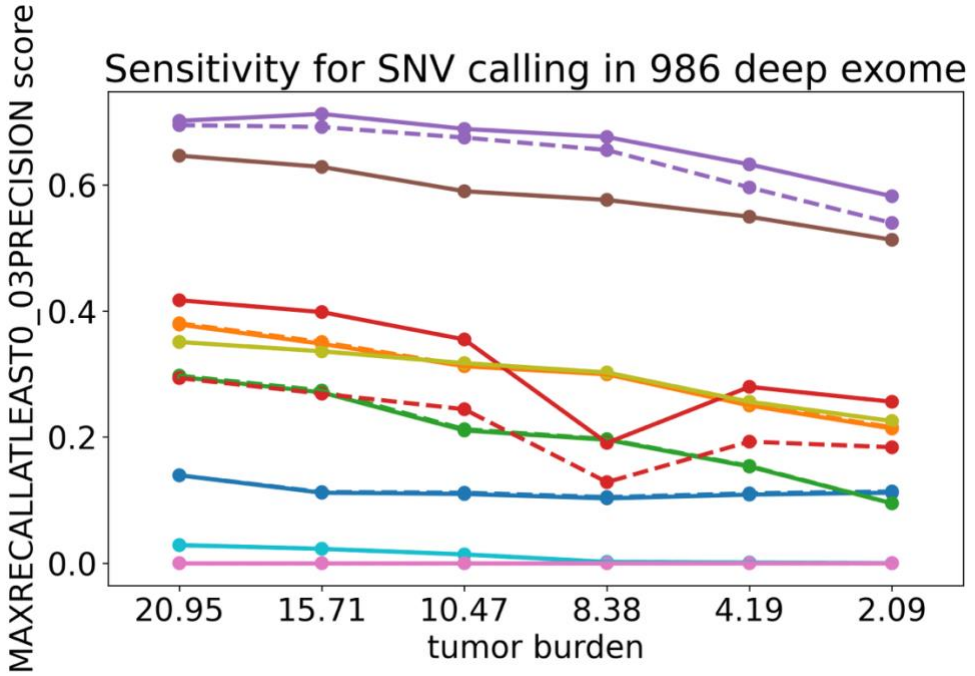
Reviewer 3, Major Comment 4

Reviewer Comment	It is common in variant calling to apply a Panel of Normal filtering. While one of the methods (ABEMUS) had this requirement built-in, it is worth

	applying it to the other methods as well following the initial calling to make it more comparable.
Author Response	We thank the reviewer for this helpful suggestion. We agree that PoN filtering is an interesting approach. We therefore explored the use of a uniform PoN across all methods to improve comparability. To remove any potential advantage from ABEMUS's built-in PoN, we implemented a caller-agnostic PoN post-filter and applied it uniformly to all call sets. PoN post-filter model construction: From the 8 deeply sequenced buffycoat WES samples sequenced with the same protocol, we derived (i) a site-level recurrence score (frequency of artefactual events across normals) and (ii) a context-aware per-base error model (mismatch-based error rates stratified by trinucleotide context, strand, and read position). Model application: We applied this PoN as a post-hoc filter to VCFs produced by all callers on the deep WES series—the dataset where ABEMUS previously exhibited strong performance—without re-calling variants, to ensure identical upstream evidence across methods. Results: As shown in Supplementary Figure 10, a unified PoN-based filtration notably benefits VarScan and VarDict, has a marginal effect on FreeBayes, and has no measurable effect on Mutect2, Strelka2, SMuRF, VarNet, or SiNVICT. The relative ordering of methods was mostly unchanged, except for VarScan, whose AUPRC increased appreciably and remained stably high across low-VAF bins; only the very lowest VAF bin showed a slight attenuation. Importantly, these adjustments do not change our overall comparative conclusions for deep-WES: ABEMUS's strong performance is not attributable solely to PoN usage, as it persists under a uniform PoN applied to all outputs. At the same time, these results highlight the practical value of adding a PoN in cfDNA pipelines to improve mutation-calling performance for certain callers. We added a new subsection 'Uniform Panel-of-Normals post-filtering' in Methods detailing the PoN construction and application, and we added the relevant figures with PoN-adjusted results in Supplementary Figure 10. We also note this change in the main text where performance is summarized.
Changes to manuscript	[line 294-302, excerpt in the ***Variant calling accuracy at 2,000x sequencing depth***] *"Orthogonally, to test whether ABEMUS's gains could reflect PoN usage rather than caller logic, we applied a unified, caller-agnostic Panel-of-Normals post-filter to the deep-WES dataset (Supplementary Fig. 10). PoN filtering improved VarScan (ΔAUPRC +0.033; Δsensitivity at 3% precision*

	+0.043) and VarDict (ΔAUPRC +0.026; Δsensitivity at 3% precision +0.129). In contrast, it had only a marginal effect on FreeBayes (ΔAUPRC +0.002) and produced no measurable changes for Mutect2, Strelka2, SMuRF, VarNet, or SiNVICT. While the relative performance of VarScan was notably improved at low ctDNA levels, the relative ranking of the remaining methods was largely unchanged." [line 809-825, excerpt in the ***Methods***] *"**Panel-of-Normals Post-filtering.** To compare callers fairly, we took the PoN logic from ABEMUS and turned it into a small, standalone post-filter that can be applied to any VCF (Mutect2, Strelka2, VarDict, VarScan, FreeBayes, etc). Using unmatched normal WGS BAMs processed like our cohort, we summarize each covered site across normals to learn:* - *a position- and allele-specific background error rate (how often spurious signal appears at that exact base), and* - *a coverage-aware allelic-fraction (AF) threshold (the percentile of AF seen in normals within depth bins, so low-depth sites have stricter cutoffs than high-depth ones).* *How filtering works (post-hoc on VCFs):* 1. *Basic quality: require minimum total depth and a minimum number of alternate reads.* 2. *AF threshold: require AF ≥ the bin-specific cutoff learned from normals.* 3. *Per-base error test: keep the variant only if its observed alternate reads are unlikely under the site's background error (FDR-controlled).* *If per-allele counts or PoN coverage are missing, we apply steps 1–2 and skip step 3. Multi-allelic sites are checked per allele.* *As a result, we obtain an caller-agnostic PoN screen that removes recurrent artefacts and caller-specific biases, enabling a fair, apples-to-apples comparison."* [line 1,042-1,069, excerpt in the ***Supplementary Information***]





Caller	Max Δ AUPRC	Max Δ Sensitivity at 3% Precision	Interpretation of PoN effect
VarScan	+ 0.0334	+ 0.0425	significant performance improvement
VarDict	+ 0.0256	+ 0.1294	significant performance improvement
FreeBayes	+ 0.0022	0	very slight performance improvement
Mutect2	0	0	No effect
Strelka2	0	0	No effect
SMuRF	0	0	No effect
VarNet	0	0	No effect
SiNVICT	0	0	No effect
ABEMUS	NA	NA	only with PoN result available

	Supplementary Figure 10: Effect of a unified Panel-of-Normals (PoN) post-filter on SNV calling (patient 986 deep-WES). ***(a)*** *Area under the precision–recall curve (AUPRC) versus tumor burden (% TF).* ***(b)*** *Sensitivity at a fixed 3% precision versus tumor burden. Solid lines show results with the unified PoN; dashed lines show results without PoN. Colors indicate callers (FreeBayes, Mutect2, Strelka2, VarDict, VarScan, SMuRF, VarNet, ABEMUS, SiNVICT). The PoN was trained on unmatched normals from the same capture/pipeline and applied post hoc and identically to all VCFs. (c) Summary of maximum changes (post-PoN – pre-PoN) across burdens: VarScan ΔAUPRC +0.0334, ΔSensitivity +0.0425; VarDict ΔAUPRC +0.0256, ΔSensitivity +0.1294; FreeBayes ΔAUPRC +0.0022, ΔSensitivity 0; Mutect2, Strelka2, SMuRF, VarNet, SiNVICT: no measurable effect. ABEMUS: only PoN-enabled outputs available (N/A deltas).*

Reviewer 3, Major Comment 5

Reviewer Comment	A section in the method explaining the entire machine learning (i.e., method-tuned) is missing.
Author Response	We thank the reviewer for this important observation. In the revised Methods, we added a dedicated subsection detailing the complete training and evaluation setup for all machine-learning models. We also describe the Random Forest feature-importance measure (mean decrease in impurity) to aid interpretability. These additions clarify the machine-learning component of the study.
Changes to manuscript	**[line 827-844, excerpt in the *Methods*]**

	Model training and evaluation. We evaluated several supervised classification algorithms using the scikit-learn v0.21.3 Python library. The models tested included decision trees (using Gini impurity), random forests (100 estimators), gradient boosting (100 estimators), logistic regression with L1 (Lasso), L2 (Ridge), and ElasticNet penalties (liblinear or saga solvers), and a multilayer perceptron (MLP) with one hidden layer of 100 neurons, ReLU activation, a learning rate of 0.001, batch size of 200, and a maximum of 100 iterations. Models were trained on the combined 150× WGS dataset from three colorectal cancer patients. We applied a leave-one-subject-out cross-validation strategy for the train/test split. We used stratified 10-fold cross-validation to preserve class ratios and report average precision averaged across folds. Performance results are presented in Supplementary Table 4. **Feature importance analysis**. We used features other than standard read counts available in each variant caller's VCF file to train the models. Feature counts per caller were: FreeBayes (37), Mutect2 (21), Strelka2 (19), VarDict (9), VarScan (5), and SMuRF (30). To interpret the random forest classifier, we computed feature importance using the Mean Decrease in Impurity (MDI). This metric quantifies the reduction in Gini impurity associated with each feature across all splits and trees in the ensemble. While informative, MDI is known to favor high-cardinality features and primarily reflects performance on the training set, rather than generalizability. Most important input features and their definitions are listed in Supplementary Table 5 and summarized in Fig. 5.

Reviewer 3, Major Comment 6

Reviewer Comment	In relation to the previous point, it is worth applying better or more advanced models such as XGBoost or LightGBM to truly appreciate the potential of this framework.
Author Response	We thank the reviewer for the helpful suggestion. Indeed, we evaluated several alternative models before selecting the Random Forest, including Gradient Boosting, Decision Trees, Logistic Regression with L1/L2/Elastic Net regularisation, and a Multi-Layer Perceptron. Performance of all models is reported in Supplementary Fig. 4. Across callers in the revised manuscript. Random Forest was consistently among the top three models for both AUPRC and sensitivity, supporting our choice based on predictive performance. This selection is further supported by its widespread use in

	related studies (e.g., [1, 2]) and the availability of an interpretable feature-importance measure (mean decrease in impurity). By contrast, the MLP offers limited interpretability, while single decision trees and logistic regression, although fully interpretable, underperformed on our data. 1. Huang W, Guo YA, Muthukumar K, et al (2019) SMuRF: portable and accurate ensemble prediction of somatic mutations. Bioinformatics 35:3157–3159. https://doi.org/10.1093/bioinformatics/btz018 2. Christensen MH, Drue SO, Rasmussen MH, et al (2023) DREAMS: deep read-level error model for sequencing data applied to low-frequency variant calling and circulating tumor DNA detection. Genome Biology 24:1–25. https://doi.org/10.1186/s13059-023-02920-1
Changes to manuscript	**[line 1,097-1,106, excerpt in the *Supplementary Information*]** *Supplementary Table 4 has been added to detail the performance of other models tested before selecting Random Forest.*

Reviewer 3, Minor Comment 1

Reviewer Comment	Can the authors explain the sentence on lines 173-175? How come only one read is sufficient for mutation detection? Most methods requires at least 3. This seems like a misunderstanding on my side and I'd appreciate clarification.
Author Response	We thank the reviewer for the opportunity to clarify this point. The sentence refers to how we define which ground-truth variants are considered for benchmarking. Specifically, for a variant to be included in the evaluation, we require that it is actually sampled in the diluted dataset—i.e., it must be covered by at least one variant read. This avoids penalizing the variant caller for failing to detect variants that were not present in the sequencing data due to low allelic fraction and sampling limitations. We do not suggest that one read is sufficient to call a variant; rather, we use this criterion to define which variants are "eligible" to be counted as true positives or false negatives in the evaluation.

Reviewer 3, Minor Comment 2

Reviewer Comment	Lines 214-215 - Mutect2 and VarDict *were* able to outperform.
Author Response	We thank the reviewer for the careful reading. The typo has been corrected in the revised manuscript. (see line 226-227).

Reviewer 3, Minor Comment 3

Reviewer Comment	In the PDF, the equations appear as blank squares.
Author Response	We thank the reviewer for catching this display issue. The problem with equations appearing as blank squares in the PDF has been corrected in the revised version (see lines 784-785 and 795).

Reviewer 3, Comment on Code availability

Reviewer Comment	The code is still not available.
Author Response	The code was made publicly available at the end of March, shortly after reviewer assignments. It is possible that this comment reflects an early version of the manuscript that preceded the code release. Indeed, we note that Reviewer 1 was able to access and evaluate the code. We hereby confirm that the code is fully accessible and remains documented as described.