DREAMS: Deep Read-level Error Model for Sequencing data applied to low-frequency variant calling and circulating tumor DNA detection

Mikkel H. Christensen [1,4] *, Simon Drue [1] *, Mads H. Rasmussen [1,4] *, Amanda Frydendahl [1,4] *, Iben Lyskjær [1,4], Christina Demuth [1], Jesper Nors [1,4], Kåre A. Gotschalck [2,4], Lene H. Iversen [3,4], Claus L. Andersen [1,4]# & Jakob Skou Pedersen [1,4]#

*Shared first author

#Shared senior authors / corresponding authors

[1]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark.

[2]Department of Surgery, Horsens Regional Hospital, Horsens, Denmark.

[3]Department of Surgery, Aarhus University Hospital, Aarhus, Denmark.

[4]Department of Clinical Medicine, Faculty of Health, Aarhus University, Aarhus, Denmark

Keywords:

Circulating tumor DNA, Next-generation sequencing, cancer research, colorectal cancer, Machine learning

Abstract

Circulating tumor DNA detection using Next-Generation Sequencing (NGS) data of plasma DNA is promising for cancer identification and characterization. However, the tumor signal in the blood is often low and difficult to distinguish from errors. We present DREAMS (**D**eep **Rea**d-level **M**odelling of **S**equencing-errors) for estimating error rates of individual read positions. Using DREAMS, we developed statistical methods for variant calling (DREAMS-*vc*) and cancer detection (DREAMS-*cc*).

For evaluation, we generated deep targeted NGS data of matching tumor and plasma DNA from 85 colorectal cancer patients. The DREAMS approach performed better than state-of-the-art methods for variant calling and cancer detection.

Background

Degraded DNA fragments are released into the blood through apoptosis, necrosis and active secretion from a range of cell types and can be detected as circulating free DNA (cfDNA)[1]. Solid tumors also shed DNA into the bloodstream and cfDNA of cancer origin is called circulating tumor DNA (ctDNA)[2]. The ctDNA level in blood is reported to be positively associated with tumor burden[3, 4]. As the half-life of cfDNA is less than an hour, ctDNA measurements can be considered real-time assessments of tumor burden and studies have shown that ctDNA can be more sensitive than radiological imaging[5-7]. This makes ctDNA measurements a promising approach for detecting relapse in patients who have undergone curative surgery[6-10]. Other proposed applications include diagnosis and intervention planning, tracking therapeutic response, monitoring the development of treatment resistance, and ultimately early detection of cancer in screening programs[8, 11]. Since obtaining liquid biopsies, such as plasma from blood samples, is both cost-effective and minimally invasive, techniques for efficient ctDNA detection holds great promise for targeted treatment in precision medicine.

In clinical contexts with low tumor burden, e.g. detection of minimal residual disease after curative-intended surgery and early detection of recurrence, the ctDNA constitute only a minor fraction of the cfDNA, often less than 0.1%. Hence, the error rate of current sequencing methods is in the same order of magnitude as the tumor signal[12], making it challenging to accurately distinguish errors from true mutations in ctDNA applications. Errors can arise in several steps between the initial shedding of cfDNA and the final generation of next-generation sequencing (NGS) reads (**Figure 1**). DNA fragments may be damaged e.g. by deamination or oxidation[13, 14], during PCR amplification of the sequencing library[13], and during sequencing from PCR amplification and/or sequencing artefacts.{{Ma, 2019 #25}} For deep sequencing, some of the PCR and sequencing errors can be rectified using unique molecular identifiers (UMIs). With the use of UMIs, each DNA fragment is labeled with a unique "barcode" prior to PCR amplification, such that replicates of the same fragment can be grouped together. Errors can then be eliminated by comparing the replicates within a group, as errors from PCR amplification and sequencing are likely to be present in only a minority of reads. However, some errors, such as DNA damage introduced prior to UMI labeling remains and continue to challenge the discrimination of true low frequency mutational signal from these errors.

Several methods for detecting low frequency variants using NGS data have been developed. Most of these establish a model for the expected frequency of errors and then assess the mutational signal with a statistical test. They differ greatly in the required data prerequisites, how the errors are modelled and handled, and the final assessment of the mutational signal.

Mutect2[15] and Shearwater[16] are examples of general somatic variant callers applicable for most NGS data. Mutect2 realigns reads in regions with mutational signal and then calculates a log-odds for the existence of the alternative allele using a statistical model in which the error rates are derived from the PHRED scores. Shearwater is developed specifically for low-frequency somatic variant detection for sub-clonal tumor mutations. It builds a position-specific error model based on the observed rate of read alignment mismatches across a set of training samples. A mutation is called if the observed signal exceeds what is expected from the error model. Additionally, this method can incorporate prior knowledge about the probability of the mutations of interest.

Other methods, including MRDetect[17], INVAR [18]and iDES[12], have been specifically tailored to detect ctDNA in NGS data. These methods build on the idea of aggregating the signal across multiple mutations to classify a sample as ctDNA positive or negative, as opposed to calling each individual mutation. For this purpose, a patient specific catalogue of mutations is generated from a matched tumor sample. However, the enhanced performance of these methods come at the expense of general applicability as they assume the presence of curated data from known ctDNA fragments or specialized lab protocols.

Here we develop a generally applicable ctDNA detection method based on a detailed background error model of individual read positions. This approach aims to capture general read-level error behavior and thus be applicable even for genomic regions where training data is not available. Data from reads known to come from ctDNA is not needed, and all data outside known mutated positions, or from independent normal samples can be used as training data. However, training data that was obtained similarly to the test data will provide the most precise model. Thus, severe changes in laboratory protocols should optimally be accompanied by re-training of the model. Some features such as the read position[19], proximity to fragment ends[14], UMI group size[12], GC-content[20] and trinucleotide context[21] have been shown to affect the probability of errors at individual read positions. By modelling their effect, the error rate of individual read positions may be predicted. Thereby, a read alignment mismatch, i.e. a non-reference base, with a low predicted error rate can provide more mutational evidence than a mismatch with a high error rate. This allows for improved cfDNA error modelling, which is key to develop accurate ctDNA applications.

In the following, we demonstrate how cfDNA errors can be modelled accurately using a neural network, by combining read level features with information about the sequencing context. For this we developed DREAMS (**D**eep **Rea**d-level **M**odelling of **S**equencing-errors) that incorporates both read-level and local sequence-context features for positional error rate estimation. Based on DREAMS, we developed a method for variant calling (DREAMS-*vc*) to accurately call individual cancer mutations in cfDNA data. The method was generalized for cancer calling in DREAMS-*cc* that aggregates the signal across a catalogue of mutations for accurate estimation of the tumor fraction and sensitive determination of the overall cancer status. To evaluate the performance of DREAMS, we performed deep-targeted sequencing of pre- and post-operative cfDNA samples from 85 stage I-II colorectal cancer (CRC) patients and compared to state-of-the art methods Mutect2[15] and Shearwater[16].

Results

Plasma cfDNA was extracted from pre-operative (Pre-OP) and post-operative (Post-OP) blood draws of 85 stage I-II CRC patients (**Table 1)** undergoing curative surgery. In addition, two stage III CRC patients were used in the model training. A biopsy from the resected tumor and paired peripheral blood cells was sequenced to generate a patient-specific mutational catalogue. Post-OP samples were collected 2-4 weeks after surgical removal of the primary tumor (**Figure 2**). Each cfDNA sample was sequenced using a custom hybrid-capture panel, designed to capture 41 exonic regions, spanning 15.413 bp, frequently mutated in CRC (**Supplementary section 1** and **Supplementary table 1)**. After UMI collapse the median of the average depths with corresponding interquartile range (IQR) of samples were for Pre-OP; 3307 (IQR: 3560), Post-OP; 7143 (IQR: 8844), buffycoat; 1850 (IQR: 1468), and tumor samples; 2132 (IQR: 2145), no samples had an average read depth below 100. All samples have been mapped and processed through the same pipeline (**Supplementary section 1).**

Table 1: Clinical characteristics

Characteristic	**Count or Median** (percent or range)
Patients	85 (100%)
Gender	
Male	53 (62%)
Female	32 (38%)
Age [years]	71 (49-87)
Tumor location	
Right colon	23 (27%)
Left colon	26 (31%)
Rectum	36 (42%)
Pathological T-stage	
pT1	15 (18%)
pT2	25 (29%)
pT3	41 (48%)
pT4	4 (4.7%)
UICC stage	
I	40 (47%)
II	45 (53%)

We first identified features that are known or expected to affect the error rate (**Figure 3a**). In general, they can be split into two types: local sequence-context features and read-level features. The local sequence-context features capture the genomic sequence context, including the trinucleotide context, information about the sequence complexity (Shannon entropy of nucleotide frequency), and GC contents in an 11 bp window around the position of interest (**Methods**).

The read-level features capture the structural composition of the read, UMI characteristics and sequencing information. The structural composition includes the strand a read aligns to (forward or reverse), the number of insertions and deletions in the read, and the total size of the underlying fragment. In the read pre-processing, UMIs were used to generate consensus reads with lowered error rates (**Supplementary section 2**). For each consensus read, we extracted the UMI-group size, the number of reads disagreeing with the consensus at the position, and the overall number of mismatches outside the position of interest. As sequencing related features, we included the base position in the read (read position) and whether the read is the first to be sequenced from the read-pair. The read quality (PHRED score) was not included, as it had the same high value for all positions in the UMI-collapsed consensus reads.

We evaluated the individual features association with the error rate by analyzing the total set of read alignment mismatches (n=707,562) across all Post-OP samples (**Figure 3b-d**), after excluding mutations and variants found in matching tumor and germline samples. The mismatches were compared to an equal number of randomly sampled matches, to estimate the error rate for each feature across its values (**Supplementary section 3**).

Since fragment lengths of cfDNA are influenced by nucleosome binding patterns, the fragment length distribution have peaks at around 162 bp (mono-nucleosomal) and 340 bp (di-nucleosomal)[22]. The error rate tended to be minimized in fragments of these lengths (**Figure 3b**). As expected, we observed a lower error rate in consensus reads formed by larger UMI groups[12] (**Figure 3c**).

The error distribution for the read position showed an increased error rate in the beginning of the reads (**Figure 3d**). We also observed a clear difference in error distribution along the read between the first and second read of the pair. The 12 different nucleotide alterations showed widely different error rates (**Figure 3e**), which is expected as error-induced mismatches are not equally likely, and the rate further differed between the two strands. However, strand symmetric alterations were generally similar, apart from the mismatches C→T/G→A and C→A/G→T.

Overall, we saw variation in the error rate for all the presented features (the remaining are shown in **Supplementary section 3**). Thus, for a given genomic position, different reads may have different error rates due to differences in read-level features. In the following, we present how this variation can be captured and used to potentially improve detection of ctDNA.

Neural network model and feature selection

To predict the error rate at a given read position, we used a neural network model with the input features described above (**Methods**). The predictive ability of individual features was evaluated using a "leave-one-covariate-out" (LOCO) scheme[23] (**Supplementary section 4)**. In short, we evaluated the performance of a full model containing all features (baseline) and then the relative performances of restricted models where each feature had been left out one by one. We used the latter to measure and rank the importance of each feature (**Figure 4a**). When leaving out the trinucleotide context, the reference base was provided instead to assess only the importance of the two neighboring nucleotides.

We found the most informative feature for modelling the error rate to be the strand (**Figure 4a)**. The second and thirds most informative features were whether the read is the first in a pair and the read position. The trinucleotide context was fourth, indicating that there is a difference in error rate for different contexts, as found by others[18]. The fragment length and the UMI group size also contribute significantly to the model. The remaining features showed little to no effect on the model performance.

An optimal subset of informative features was chosen using a stepwise procedure where features were excluded in order of importance (**Methods**). The set of features chosen was the smallest model that did not perform significantly worse than the full model (**Supplementary section 4)**. The four least important features could be removed without any significant negative effect on the performance (**Figure 4b**). Of the remaining ten features, eight were read-level features, namely the features describing the UMI group, the number of errors in the UMI group, the number of deletions in the read, the number of other errors in the read, the fragment length, read position, strand, and if the read was first in pair. This showed that read-level features do contribute to accurate modelling of the error rate.

The numerical and categorical variables are processed differently in the neural network prior to the hidden layers (**Figure 4c**). The numerical features are batch normalized, the categorical features are one-hot encoded, and the tri-nucleotide context is embedded in three dimensions to handle the large number of possible contexts (**Methods**).

Predictive performance in clinical data

To validate the utilization of the DREAMS error model, we applied it in calling tumor variants (DREAMS-*vc*) and cancer (DREAMS-*cc*) (**Methods)**. We assessed the performance using five repeats of 2-fold cross-validation (5x2 CV) (**Figure 5a**). The model was trained on the Post-OP samples, and Pre-OP samples were used for method validation. The split was done on patient level to ensure that a model is not trained and tested on data from the same patient. This analysis was repeated with five different randomized splits to control for split induced variation.

The performance of calling tumor mutations in the plasma samples was assessed by looking at the area under Receiver Operating Characteristic curves (AUC). The performance of DREAMS-*vc* was compared to state-of-the-art algorithms Mutect2 and Shearwater. Only positions with at least one observed mismatch were included in the performance calculations (**Figure 5b**). Positions without signal was called negative by any method, making them redundant for performance comparisons.

Using DREAMS-*vc*, we aimed to call the tumor mutations of each patient from their respective mutation catalogue. As negative controls, we attempted to call cross-patient mutations, by searching for the mutations found in other patients. Additionally, a validation set of 500 randomly generated alterations within the covered sequencing panel was used as negative controls. Evaluating across the combined negative set of both cross-patient mutations and validation alterations and cancer stages, DREAMS-*vc* performs significantly better than both Shearwater and Mutect2 (**Figure 5b**). Additionally, the performance was assessed separately for stage I and stage II CRC patients. This showed that superior performance of DREAMS-*vc* is predominantly due to the stage II CRC patients (**Figure 5b**). As expected, all models perform better on later stage patient samples as these are expected to have a higher mutational signal in the cfDNA due to a higher tumor burden.

All methods perform similarly on stage I patients, however DREAMS-*vc* has marginally better performance. Performance evaluations for each of the separate negative sets showed that DREAMS performs better than Mutect2 with the cross-patient negative set and better than Shearwater with the validation set as the negative set. The variation in performance of DREAMS-*vc* across splits and folds is lower than for Mutect2 and Shearwater, which indicates that its variant calling is more stable across patients and mutation types.

By maintaining the false positive rate at 5% for the alterations with signal in the validation set for each model, we get comparable thresholds for the three confidence measures: p-values, Bayes factor and TLOD for DREAMS-*vc*, Shearwater, and Mutect2, respectively. This allows for a comparison of the sensitivity of the models at a pre-determined specificity of 95%. The model could then be assessed across an alteration catalogue of 191 true positive mutations from the mutation catalogue and 1290 cross-patient negative calls based on the mutation catalogue of the other patients. Out of the alteration catalogue, 88 true mutations and 1100 cross-patient negative calls had a signal for the alteration.

Using this threshold DREAMS-*vc* called 83% of the tumor mutations with signal, while Shearwater and Mutect2 called 75% and 72.7%, respectively (Table 2). F1 and G-mean scores were calculated to assess the performance of the models by using the cross-patient mutations as negative controls. G-mean is the geometric mean of sensitivity and specificity, and F1 is the harmonic mean of precision and sensitivity. For G-mean, DREAMS-*vc* performed better than Shearwater and Mutect2, however the F1 score of Shearwater was very similar to DREAMS-*vc*, due to lower false-positive rate of shearwater (Table 2). Considering all mutations observed in the tumors, including those without signal in plasma, we found that about 38.2% could be recalled in Pre-OP liquid biopsy samples.

Table 2	**Full alteration catalog**[a]		**Catalogue alterations with signal**[b]			
	Sensitivity	Specificity	Sensitivity	Specificity	F1	G-mean
DREAMS-*vc*	0.382	0.998	0.830	0.957	0.702	**0.891**
Shearwater	0.346	0.998	0.750	0.971	**0.710**	0.853
Mutect2	0.336	0.997	0.727	0.933	0.566	0.831

[a] Full alteration catalogue consisting of n=191 true positive mutations, and n=1290 potential cross-patient negative calls.

[b] Catalogue of alterations with signal consisting of n=88 true positive mutations, and n=1100 potential cross-patient negative calls.

By setting the threshold based on a 5% false positive rate in the cross-patient mutation set, the validation mutation set can be used as negative controls. The true positives are still the same 191 mutations of which 88 has a signal for the alteration. The negatives are the 500 validation positions multiplied with the 87 tested samples, giving a total of 43,500 possible alterations of which 1,350 had a signal. With this set we obtained an 83% true positive rate, compared to 77.3% for Shearwater and 68.2% for Mutect2 (Table 3). DREAMS-*vc* scored highest in both F1 and G-mean scores. Here, DREAMS-*vc* performed distinctly better than Shearwater, while Mutect2 had a more comparable F1 score.

Table 3	**Full alteration catalog**[a]	**Catalogue alterations with signal**[b]

	Sensitivity	Specificity	Sensitivity	Specificity	F1	G-mean
DREAMS-*vc*	0.382	0.998	0.830	0.944	**0.616**	**0.885**
Shearwater	0.356	0.997	0.773	0.911	0.493	0.839
Mutect2	0.314	0.999	0.682	0.962	0.603	0.810

[a] Whole catalogue consisting of n=191 true positive mutations, and n=43500 potential validation set calls.

[b] Catalogue of positions with signal consisting of n=88 true positive mutations, and n=1350 potential validation calls.

A common measure used to predict the presence of ctDNA is the estimated tumor fraction in plasma. DREAMS-*cc* combines the mutational evidence across the mutation catalogue, to estimate the tumor fraction with an accompanying p-value for the presence of cancer (**Methods**). We aimed to detect cancer in the Pre-OP samples, since cancer is present and should, in theory, be detectable given enough ctDNA is present in the blood. As a negative control, we attempted to detect cancer in each Pre-OP sample (Tested Sample) with the mutation catalogue from all other patients (Candidate patient) (**Figure 6a**). In case of shared mutations between the mutation catalogues, these were eliminated to prevent false positives. As a benchmark, we constructed a cancer call score using the product of the individual Bayes factors across the mutation catalogue from Shearwater, resulting in a similar tendency (**Figure 6b**). The performance of calling cancer can be assessed by treating the cross-patient mutation catalogues as expected negatives and calculate an AUC score. Performance was compared using the 5x2 cross validation setup as above (**Figure 5a).** The AUC was very similar between DREAMS-*cc* and Shearwater with respect to calling cancer, however DREAMS-*cc* showed a slightly increased performance (p = 0.0343, one tailed t-test). As for variant calling, we only included the samples with mutational signal to showcase and compare the performance of the different methods in discriminating tumor from error signal.

For the patients with stage I and II CRC, we found tumor supporting reads in 47.5% (19/40) and 73% (33/45) of the Pre-OP samples, respectively. We called cancer in 34% of the stage I CRC patients, corresponding to 74% (14/19) of the patients with a mutational signal. We called cancer in 73% of the stage II CRC patients, corresponding to 94% (31/33) of the patients with signal. These results were obtained whilst still limiting the false positive rate to 5 % in cross-patient cancer calls with a non-zero mutational signal.

Detailed analysis of the false positive cancer calls reveals that most are due to a specific KRAS G12V variant: chr12:25245350 C>A. This variant is common in colon cancer, and it is therefore not surprising to find in the patients [24]. However, the mutation was not found in the patient's corresponding tumor or buffycoat samples. A possible explanation for this is that the mutation is not detected in the tumor biopsy due to sub-clonality [25] or that there is an underlying germline signal that was not caught in the buffycoat.

Discussion

We have developed DREAMS, as a new approach for modelling the error rates in sequencing data that incorporates information from both the local sequence context and read-level information. DREAMS is intended for settings that rely on accurate error identification and quantification. We applied the error model for low-frequency ctDNA variant calling (DREAMS-*vc*) and cancer detection (DREAMS-*cc*).

The error rate was found to vary depending on several of the proposed read-level features. Surprisingly, fragment size was found to be correlated with the error rate, with the smallest error-rates being observed for fragment sizes corresponding to the mono-nucleosomal and di-nucleosomal lengths (**Figure 3b**). Fragments that deviate from these in length may have been degraded in the blood for a longer time and thereby accumulated more errors. Fragments of ctDNA are generally shorter and error rates are generally highest in short fragments, which shows the importance of accurate error modelling[26, 27]. The error rate was also found to vary with the strand, and symmetric mismatches occurred at different rates (**Figure 3e**). The G>T/C>A asymmetry can be explained by the hybridization capture protocol only targeting one strand and thus only capturing oxidative damage of that strand[14]. A similar mechanism might explain the C>T/G>A asymmetry in the case of cytosine deamination. The error rate varied with the position in the read and was especially increased in the beginning of reads (**Figure 3d**). This may be because ends of fragments are prone to damage[14] and in thermodynamic equilibrium with being single stranded. The error rate also varied depending on whether the read was the first or second in the pair (**Figure 3d**). Besides being intermitted by a PCR amplification step, the reads differ in composition and length of adapters sequenced prior to the insert, which might cause this difference.

Training a background error model using DREAMS does not require known mutation sites in reads, as it only models the errors found in aligned reads (BAM-files). These can originate from normal samples or mutation filtered cancer samples, as in this study. Since error patterns are highly dependent on laboratory procedures, the same protocol should be used for training samples and subsequent testing samples. Training across multiple samples gathered over time, is expected to learn the error patterns that are general across samples and batches. Conversely, if the amount of data in a single sample is large, the error model can be trained on the sample itself, which potentially yields a highly specific model that accounts for sample specific error patterns. The model is built to be position agnostic and can therefore be used to predict error rates for positions for which no training data is available. Furthermore, it is fit for both deep sequencing of panels and shallow sequencing of whole genomes.

The error model has been implemented using a neural network, allowing the feature set to be tailored to capture the relevant information of a specific setting. Analysis of the feature importance revealed that several of the proposed read-level features are useful in predicting the error rate in sequencing data (**Figure 4a**). Most features presented in this paper are general to NGS data, however not all sequencing protocols use UMI based error correction, rendering UMI related features redundant. In particular, UMI cannot be exploited for shallow whole-genome sequencing as read groups cannot be formed. In such cases error rates would be increased, making accurate error modelling as performed by DREAMS even more important.

Compared to simpler methods, the presented approach is more computationally demanding, due to training of the neural network model and the use of complex data extracted from BAM-files. A neural network is a simple and flexible approach for bridging the gap between a complex set of contexts and read level features and the error rate of a given read position but might not be the most efficient solution. The model can be trained on a regular laptop within a few hours, which should only be done once, when the training dataset is defined. Using the trained model and the statistical modules adds no significant computation time for calling mutations and cancer in the current setting. However, very large mutation catalogues are expected to increase the computation time for DREAMS-*cc*.

DREAMS was built to exploit read-level features under the assumption that these affect the error rate in sequencing data. Thus, the power of this approach increases with the variability in the error rate explained by read level features. Thereby, less emphasize is put on mismatches that are likely errors, and more confidence in the potential tumor signal from other mismatches. Conversely, if read level features are not improving error prediction, the performance is expected to be similar to methods working with simpler summary data. Although DREAMS use information about the local sequence-context, strong regional effects on the error rate are not expected to be captured by the model.

In all performance comparisons DREAMS-*vc* performed better or equal to the other methods in calling tumor mutations. This indicates that read-position level features can improve performance in separating error from mutational signal. Similarly for cancer detection, DREAMS-*cc* performed equal to calls based on Shearwater. Cancer was detected in most (73%) of stage II CRC cancer patients and a third (34%) of stage I patients.

There are false positive cancer and mutation calls, some of which could potentially be explained by clonal hematopoiesis of indeterminate potential (CHIP) or an unexpected error signal. To reduce the signal from CHIP, we have excluded positions with significant presence of non-reference nucleotides, found in the germline samples, however, a low signal might still be present. Remaining false positive calls might be due to regional effects or sample specific artifacts. Many of the false positive mutation calls in the Pre-OP samples were found to be a mutation leading to the KRAS G12V variant, and it could therefore potentially be explained by a sub-clonal variant that was not identified in the tumor sample or a germline signal of clonal hematopoiesis of indeterminate potential (CHIP) that was not identified in the buffycoat samples.

Sensitive variant calling in liquid biopsies can provide non-invasive insight into tumor genetics, which can potentially enable personalized treatment of patients and be a cost-effective approach for cancer screening. DREAMS-*cc* integrates evidence across a mutation catalogue to increase sensitivity in cancer detection. Cancer detection is expected to get more sensitive as the number of mutations in the catalogue rises. A potential application of DREAMS-*cc* could be tumor agnostic cancer detection based on a catalogue of commonly known tumor variants.

The approach presented in this paper does not utilize tumor specific signals such as the fragment size distribution, fragmentation patterns, mutational signatures, expression information, etc. However, together with the error characterizing properties of DREAMS-*cc*, this could potentially refine the cancer calls. Addition of regional properties and positional information could potentially further increase sensitivity. In this paper, we focus on the single nucleotide variants in the tumor, but the model could be extended to be able to look for indels. The underlying ideas in DREAMS are not restricted to variant calling and could be used in other tasks of sequencing data analysis such as advanced error filtering.

Conclusion

We have presented the DREAMS error rate model and demonstrated the importance of using read-level features for modelling the errors in NGS data. The model was validated in a tumor informed setting, using DREAMS-*vc* for variant calling and DREAMS-*cc* for cancer detection in patients with CRC. DREAMS-*vc* allowed accurate detection of mutation signal in plasma samples extracted prior to curative intended surgery with an improved performance compared to state-of-the-art methods. This highlights the importance of including read-level information in modelling the background error rate. Furthermore, DREAMS-*cc* demonstrated the ability to combine signal from multiple mutations known from the tumor biopsy for improved cancer detection. DREAMS-*cc* was able to call cancer in 73 % of Pre-OP samples from CRC stage II patients, and 34 % of CRC stage I patients. Potential future applications of DREAMS include analysis of WGS data and tumor agnostic cancer detection. The approach presented with DREAMS is generally applicable across NGS applications that need accurate handling and quantifications of errors, and the presented algorithms (DREAMS-*vc* and DREAMS-*cc*) are only examples of how to exploit this. The specific application presented in this paper is implemented as a user-friendly R package [**https://github.com/JakobSkouPedersenLab/dreams**].

Declarations

Ethics approval and consent to participate

The Committees on Biomedical Research Ethics in the Central Region of Denmark have approved the study (J. No. 1-10-72-3-18). The study was performed in accordance with the Declaration of Helsinki and all participants provided written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

Sharing of sensitive patient specific clinical information and raw sequencing data is currently not possible due to ethical and GDPR regulations.

Competing interests

The authors declare that they have no competing interests

Funding:

MHR, AF, LI, JN, and CLA were funded by Aarhus University, Lundbeck Foundation (R180-2014-3998), Dansk Kræftforsknings Fond (FID1839672), Innovationfund Denmark (9068-00046B), Danish Cancer Society (R133-A8520-00-S41, R146-A9466-16-S2, R231-A13845, R257-A14700), NEYE foundation, Frimodt-Heinke Foundation, and Novo Nordisk Foundation (NNF17OC0025052). SD, MHC, and JSP were funded by Aarhus University, the Independent Research Fund Denmark | Medical Sciences (8021-00419B), the Danish Cancer Society (R307-A17932), and Aarhus University Research Foundation (AUFF-E-2020-6-14).

Author contributions:

MHC, SD, CLA, and JSP conceived and designed the study. MHC and SD developed the statistical methods and the software under supervision by JSP with input from MHR and CLA. MHR, AF, IL, CD, JN, KAG, and LHI acquired patient samples and generated patient data, including NGS data. MHC, SD, MHR, AF, CLA, and JSP analyzed and interpreted the patient data. SD and MHC wrote the article under supervision of CLA and JSP with revisions and suggestions from the other authors. All authors read and approved the final manuscript.

Acknowledgements:

We thank the participating CRC patients, and the Danish Cancer Biobank for contributing clinical material. We also thank the IMPROVE study group for patient inclusion: Kåre Andersson Gotschalck (Horsens Hospital), Lene Hjerrild Iversen (Aarhus University Hospital), Uffe Schou Løve (Viborg hospital), Anders Husted Madsen (Herning Hospital), Ole Thorlacius-Ussing (Aalborg University Hospital), Ismail Gögenur (Køge Hospital), Per Vadgaard Andersen (Odense University Hospital), Jakob Lykke (Herlev Hospital), Peter Bondeven (Randers Hospital), and Nis Hallundbæk Schlesinger (Bispebjerg Hospital).

Methods

Error rate prediction using read level information

In this study we present a method called DREAMS (**D**eep **Rea**d-level **M**odelling of **S**equencing-errors) for estimating the error rate at each read position using features of the individual read and the genomic context of the position. In practice, this is achieved by predicting the probability of observing each allele given the describing features of a position in a read and considering the probabilities of observing the alternative alleles as the error rates. The read specific features can include information such as the read position, the strand of the mapped read, the length of the fragment, and UMI-group size. The read position refers to the cycle number at which the position was sequenced starting with the first nucleotide of the fragment, thus disregarding cycles used for reading primers, adapters, unaligned ends etc. Context specific features contain information about the genomic sequence surrounding the position, including the neighboring bases (tri-nucleotide context), the complexity, and GC-content. The local complexity is calculated as the Shannon entropy for both single nucleotides and pairs. Similarly, the local GC content is calculated as the fraction of C and G nucleotides. In principle, any feature that can be thought to affect the error rate of a read position can be added to improve the error rate prediction. Another possible feature would be the positional read quality score given by the sequencing machine. However, the estimated quality for the collapsed consensus reads were all capped at the same high value and thus excluded as they do not include any information for further modelling.

Data

Data for a read position can be extracted from a read mapping (BAM-file) with sequencing data from a next generation sequencing experiment. The training data for the model consists of a set of read positions from multiple samples, for which the observed allele is denoted together with the relevant features. This means that the training data includes both matches, where read positions where the observed allele is equal to the reference allele and mismatches where the observed and reference allele differ. Mismatches that correspond to known single nucleotide polymorphisms found in the germline samples are excluded from the training. Assuming that the training samples are non-cancerous means that all remaining mismatches in the dataset can be assumed to be errors that have occurred on a molecular level in the body or lab, or during sequencing of the sample.

The mismatches are extracted from the BAM-file using the mismatched positions annotated in the MD-tag. The equivalent genomic position is found, and the 11- and 3-mer context is extracted from the reference genome and used for calculation of local sequence-context features. The UMI errors and UMI count are extracted from the cE and cD tags generated by the CallMolecularConsensusReads from fgbio used for calling UMI consensus reads. Information about the insertions and deletions is extracted from the cigar tag. The fragment size is the insert-size (isize), and the read position is the position in the read sequence from the 5'-end of the read. Strand and first in pair are extracted from BAM flag where this information is encoded in a bitwise fashion.

The model assumes that the input data for both training and testing is based on readings of unique fragments, so each position in a fragment is only represented in one read. This can be assured using unique molecular identifiers (UMIs) and by trimming overlapping read positions in the read pairs.

As training on every single read position in every single read is very demanding and inefficient, we employ a methodology akin to importance sampling where we extract all the mismatches from the data and randomly sample a subset of the non-mismatches. To account for this skew induced by down-sampling one category of the training data a rescaling scheme inspired by[28] is used on the predicted error rates. The method outlined in **Supplementary section 5**.

Neural network model

Structure of the neural network

To predict the error rate at a given read position we use a multilayer perceptron (MLP) which is a simple neural network setup with multiple fully connected layers. The neural network allows us to use the features without prior knowledge of how they interact amongst each other or how they affect the error rate. The neural network is trained using a set of read positions where the features describing the read positions are used as inputs and the observed allele as output.

For a given read position the possible observed outcomes are the alleles A, T, C or G. Interpreting this as a random event, the observed allele can be seen as an outcome from a four-dimensional multinomial distribution with one trial. Let X_{ij} represents the observed allele in read j at position i and D_{ij} be the set of observed features for that read position. For a non-mutated, homozygote position the observed allele should predominantly be the reference allele, and any observations of non-reference alleles, would be considered errors. In this situation $P(X_{ij} = A|D_{ij})$ would be close to 1 if A was the reference allele for read position (i,j), and $P(X_{ij} = x|D_{ij})$, $x \in \{T, C, G\}$ would be the error rates for the remaining three alleles. Given a set of observations $\{(x_{ij}, D_{ij})\}_{i=1}^{N}$ it is then possible to write the log-likelihood function for the observed data:

$$
\begin{aligned}
&l\left(\{(x_{ij}, D_{ij})\}_{i,j}\right) \\
&= \sum_{i,j} \log\left(P(X_{ij} = x_{ij}|D_{ij})\right) \\
&= \sum_{i,j:x_{ij}=A} \log\left(P(X_{ij} = A|D_{ij})\right) + \sum_{i,j:x_{ij}=T} \log\left(P(X_{ij} = T|D_{ij})\right) + \\
&\sum_{i,j:x_{ij}=C} \log\left(P(X_{ij} = C|D_{ij})\right) + \sum_{i,j:x_{ij}=G} \log\left(P(X_{ij} = G|D_{ij})\right)
\end{aligned}
$$

The problem now becomes how to estimate the distribution $P(X_{ij}|D_{ij})$ above. To do this, start by defining the probability functions via the SoftMax function:

$$
P(X_{ij} = a|D_{ij}) = \frac{e^{f_a(D_{ij})}}{\sum_{a' \in \{A,T,C,G\}} e^{f_{a'}(D_{ij})}}
$$

, where $f_a(D_{ij})$ is a predictor function for the allele a using the observed information D_{ij}. As an example, for classic multinomial logistic regression a linear predictor function is chosen such that

$f_a(X_i) = \beta_a \cdot X_i$, where β_a is a vector of feature specific weights that can be found by maximizing the log-likelihood function. To get a more flexible model, a neural network is chosen, since this can approximate any arbitrary predictor function well including arbitrary interactions between input features. To do this $P\left(X_{ij} = a \middle| D_{ij}\right)$ can be interpreted as the output from a neural network model where SoftMax is used as the last activation function and $f_a\left(D_{ij}\right)$ is the output from the last hidden layer. To train such a model inspiration is drawn from likelihood theory and the negative log-likelihood function is chosen as the loss function to minimize.

Architecture

The neural network model allows for high flexibility in the choice of features and requires very limited prior knowledge about the effect of the features on the error rate. The neural network was selected to be a MLP with an input layer, three hidden layers and an output layer. The dimension of the input layer depends on the selected input features, the hidden layers have a configuration of 128, 64, and 32 nodes with a ReLu activation function, and the output layer contains 4 nodes with SoftMax activation, as explained above, corresponding to probability of observing each of the 4 alleles. The configuration of hidden layers can be varied, depending on the input data and the available computational resources. The models were training using the Keras library (2.3.0) in R, which is an interface that builds in Tensorflow (2.6.0) [29].

Feature handling / embedding

The features are split into numeric, categorical, and embedded variables and handled accordingly. Categorical features are one-hot encoded, and the numeric features are batch normalized. The trinucleotide context can be seen as the three distinct features: reference allele and the two neighboring bases. These can be handled as categorical features with individual one-hot encoded 4-dimensional inputs using 12 (3x4) input nodes in total. Alternatively, a 64-dimensional (4x4x4) one-hot encoded input of the entire trinucleotide context (TNC) can be used. We will employ another alternative that takes the 64-dimensional feature in the input layer and embeds it into a continuous

3-dimensional vector before including it in the model alongside the remaining input features. Thereby, the model can learn the relationship between the contexts, and cluster contexts that have a similar effect on the error rate close together and vice versa.

Assessing cancer status across a catalogue of multiple mutation candidates

Based on the neural network error model developed above, it can now be assumed that the individual error rates for a given position in each read is known. In this section the error rates will be exploited to develop a statistical framework for estimating the tumor fraction in a sample based on a catalogue of candidate mutations. This framework can ignore some mutation candidates if these are not found in the sample, for example due to relatively low allelic frequency due to sub-clonality in the tumor or due to little tumor in the circulation. Reduction in the candidate mutations allows for a comprehensive mutation catalogue to be used, where mutation candidates with limited evidence may be excluded. The subset of candidate mutations is selected statistically by finding mutations with a consistently high mutational signal, and the tumor fraction is estimated based on these candidates. This subset of mutations is then used in a statistical procedure for testing if the observed mutational signal exceeds what we would expect if no mutated DNA were present, making it possible to determine the cancer status of a patient based on the sample.

The statistical model

Start by introducing Z_i as a variable that controls the presence of a given mutation on the site i, such that $Z_i = 1$ represent the case where the site is mutated, and $Z_i = 0$ when it is not. Furthermore let:

$$Z_i \sim Bernulli(r)$$

Thus, given a catalogue of possible mutations, r is the probability that each of them is present in the sample. For site i let R be the germline reference allele and M the alternative allele of interest. Furthermore, it is assumed that the germline site is homozygote, such that any signal from non-reference alleles must be due to errors or mutational signal from a tumor. To model the molecular composition of the fragments covering site i let $Y_{ij} \in \{R, M\}$ be the true error-free nucleotide of the j'th fragment. If the i'th mutation is not present in the sample ($Z_i = 0$), we are sure that the true nucleotide of the fragment is the reference and thus the following distribution holds:

$$P(Y_{ij} = R | Z_i = 0) = 1, \qquad P(Y_{ij} = M | Z_i = 0) = 0$$

To model the mutational DNA present in the sample let $f > 0$ denote the tumor fraction. This is the fraction of the DNA in the blood that originates from tumor cells. Assuming that the mutation of interest is (sufficiently) clonal in the tumor, i.e. half of the DNA in the tumor has this mutation, the probability of a given fragment having the mutation is $f/2$. Using this the following distribution for Y_{ij} can be assumed when the mutation is present in the sample ($Z_i = 1$):

$$P(Y_{ij} = R | Z_i = 1) = 1 - \frac{f}{2}, \qquad P(Y_{ij} = M | Z_i = 1) = \frac{f}{2}$$

To model the errors that occur in NGS data let X_{ij} be the observed nucleotide in fragment j at position i. Assume that the distribution of X_{ij} depends only on the corresponding true nucleotide Y_{ij}, in the sense that the event $X_{ij} \neq Y_{ij}$ corresponds to the observation being an error. This distribution is exactly what the neural network model described above aims to approximate using the observed features D_{ij}. To simplify notation the dependence of X_{ij} on D_{ij} will be omitted from notation in the following. Note that observations X_{ij} outside $\{R, M\}$ will have little information about the true nucleotide Y_{ij}. Furthermore, since the error rates generally are low, the difference between including interactions between all four possible alleles and only the two allele of interest is negligible. Thus, to simplify the following calculations, we assume that $X_{ij} \in \{R, M\}$. In practice this means that all fragments, j', for which $x_{ij'} \notin \{R, M\}$ are eliminated from the analysis. Using this assumption, we define the probability of observing the alternative allele in a reference allele position as the following error rate:

$$e_{ij}^{R \to M} = P(X_{ij} = M | Y_{ij} = R, X_{ij} \in \{R, M\}) = \frac{P(X_{ij} = M | Y_{ij} = R)}{P(X_{ij} = R | Y_{ij} = R) + P(X_{ij} = M | Y_{ij} = R)}$$

Conversely, for a fragment that stems from a tumor cell and contains the mutated allele we define:

$$e_{ij}^{M\to R} = \frac{P\left(X_{ij} = R | Y_{ij} = M\right)}{P\left(X_{ij} = R | Y_{ij} = M\right) + P\left(X_{ij} = M | Y_{ij} = M\right)}$$

Estimating the tumor fraction and mutation presence

In this section we will develop a procedure for estimating the tumor fraction (f) and mutation presence probability (r). For this, let $i \in \{1, \dots, K\}$ be the index of a catalogue of K candidate mutations, N_i the corresponding number of covering reads and $\left\{\left(x_{ij}\right)_{j\in\{1,\dots,N\}}\right\}_{i\in\{1,\dots,K\}}$ all the observed alleles. First, we write the likelihood function for f and r:

$$L\left(f, r\middle|\left\{\left(x_{ij}\right)\right\}_{i\in\{1,\dots,K\}, j\in\{1,\dots,N\}}\right)$$

$$= \prod_{i=1}^{K} P(Z_i = 0) \cdot$$

$$\prod_{j:x_{ij}=R} \left[P\left(X_{ij} = R|Y_{ij} = R\right)P\left(Y_{ij} = R|Z_{ij} = 0\right) + P\left(X_{ij} = R|Y_{ij} = M\right)P\left(Y_{ij} = M|Z_{ij} = 0\right)\right] \cdot$$

$$\prod_{j:x_{ij}=M} \left[P\left(X_{ij} = M|Y_{ij} = R\right)P\left(Y_{ij} = R|Z_{ij} = 0\right) + P\left(X_{ij} = M|Y_{ij} = M\right)P\left(Y_{ij} = M|Z_{ij} = 0\right)\right] +$$

$$P(Z_i = 1) \cdot$$

$$\prod_{j:x_{ij}=R} \left[P\left(X_{ij} = R|Y_{ij} = R\right)P\left(Y_{ij} = R|Z_{ij} = 1\right) + P\left(X_{ij} = R|Y_{ij} = M\right)P\left(Y_{ij} = M|Z_{ij} = 1\right)\right] \cdot$$

$$\prod_{j:x_{ij}=M} \left[P\left(X_{ij} = M|Y_{ij} = R\right)P\left(Y_{ij} = R|Z_{ij} = 1\right) + P\left(X_{ij} = M|Y_{ij} = M\right)P\left(Y_{ij} = M|Z_{ij} = 1\right)\right]$$

$$= \prod_{i=1}^{K} (1 - r) \cdot \prod_{j:x_{ij}=R} \left(1 - e_{ij}^{R\to M}\right) \cdot \prod_{j:x_{ij}=M} e_{ij}^{R\to M} +$$

$$r \cdot \prod_{j:x_{ij}=R} \left[\left(1 - e_{ij}^{R\to M}\right) \cdot \left(1 - \frac{f}{2}\right) + e_{ij}^{M\to R} \cdot \frac{f}{2}\right] \cdot \prod_{j:x_{ij}=M} \left[e_{ij}^{R\to M} \cdot \left(1 - \frac{f}{2}\right) + \left(1 - e_{ij}^{M\to R}\right) \cdot \frac{f}{2}\right]$$

Getting a maximum likelihood estimate (MLE) of f and r by optimizing this expression analytically is not tractable. However, by seeing Y_{ij} and Z_i as latent variables, estimates can be found by employing an EM-algorithm, which will be developed in a **Supplementary section 6**. For now, assume that $\hat{f}$ and $\hat{r}$ are a MLEs of f and r respectively.

To test if a sample has a significant content of mutational DNA, we focus on the parameter in the model. By representing the hypothesis of a negative sample as a tumor fraction of 0 and no mutations present ($H_0: f, r = 0$) and a positive sample as a positive tumor fraction and some mutations present $\left(H_A: f > 0, r \geq \frac{1}{K}\right)$, a likelihood ratio test can be used to test for significance. Note that $r \geq \frac{1}{K}$ in H_A corresponds to at least one mutation being present in the sample. The LR-test statistic for this test is:

$$Q = -2\log\frac{L\left(0{,}0\middle|\{(x_{ij})\}_{j=1}^{N}\right)}{L\left(\hat{f},\hat{r}\middle|\{(x_{ij})\}_{j=1}^{N}\right)}$$

Since there are 2 free parameters in the model, it can be assumed that Q is approximately $\chi^2(2)$-distributed, and a p-value can be obtained as follows:

$$p_{val} = 1 - F_{\chi^2(2)}(Q)$$

Using this statistical model for cancer calling on top of the error rate predictions from DREAMS we refer to as the DREAMS-cc.

Calling individual mutations

In the special case where the number of mutations in the catalogue is $K = 1$, the algorithm outlined above can be thought of as a regular variant caller. In this case the concept of some mutations not being present in the sample is unnecessary, as the presence of the single mutations of interest can be governed solely by the tumor fraction f. The algorithm above is easily modified to handle this by assuming that $r = 1$, and using one degree of freedom for the χ^2-distribution in the significance test. The equations in the EM-algorithm can also be simplified by making this assumption. We refer to the variant caller will be referred to as DREAMS-*vc.*

Figure legends

Figure 1:

Error generation in Next Generation Sequencing data. Normal cells (grey) and tumor cells (blue) shed DNA into the bloodstream. The tumor DNA (blue) contains a tumor mutation (yellow). The circulating free DNA in the blood becomes damaged both *in vivo* and *in vitro* (green triangle). Errors can be introduced at each PCR duplication during amplification (red circle). Further errors are accumulated during sequencing and mapping (purple square). The final data contains mapped reads, where some mismatches are errors, and others are mutation from tumor cells.

Figure 2:

The data collection setup for tumor-informed relapse detection in colon cancer patients. After the patient is diagnosed with colorectal cancer a liquid biopsy is extracted prior to curative surgery (Pre-OP). A biopsy is taken from the tumor. Following surgery liquid biopsies (Post-OP) can be collected to monitor relapse. All collected samples are sequenced using Next-Generation Sequencing.

Figure 3:

a) Examples of local sequence-context features and read-level features extracted from a read for a single position of interest in a read mapping. Centered at the position of interest, the trinucleotide context is extracted, and the surrounding 11 bp region is used for calculating the regional features, including GC content and K-mer complexity. The read pairs contain a forward and reverse read that are enumerated as either the first or second of the pair according to the order of sequencing. Two read pairs are used for illustration of the read-centric features in the panels on the right. The UMI groups are shown to indicate the variation in the number of reads used for the consensus reads. The read position and fragment size are shown for the consensus reads. b-e) Variation in observed error rate for selected features based on their observed distribution: b) Fragment size, c) UMI group size, d) Read position and the variation between the first and second read in a pair. e) Error type for each strand (forward and reverse). For each feature the 95% confidence interval is indicated by the shaded areas or error bars. See **Supplementary section 3** for how the error rates and confidence intervals are calculated and similar plots of the remaining features.

Figure 4:

a) Features are individually removed one-by-one from the full model containing all features to measure the decrease in validation error. The most important feature is then defined as the one that decreases the validation error the most, and vice versa. The grey points show the mean decrease in validation error for each fold of a 5-fold cross validation. The average of these is used to rank the features by importance, indicated by the black points. b) Based on the importance ranking, the features are cumulatively removed one-by-one to from a full model. If the decrease in validation error compared to the full model is significant, the feature should not be removed from the model. A feature is only kept if removing it worsen the performance in all folds of the 5-fold cross validation. c) Structure of the neural network. The neural network uses three different types of input features: numeric, categorical, and embedded. The input features are processed differently in each group. The input features are then parsed through three hidden layers of decreasing width. The output contains 4 nodes representing the probability of observing each of the four based (A, T, C, G) at the given read position.

Figure 5:

a) Illustration of 5x2-cross-validation procedure for the estimation of performance. The patients are first split into two approximately equally sized folds. The neural network model is trained on the Post-OP data of fold 1 and validated by testing the models on the Pre-OP samples of the other fold (Test B). This is then repeated by swapping the data in fold 1 and 2. The whole process is repeated 5 times. b) Performance of variant calling using DREAMS-*vc* compared to state-of-the-art tools Shearwater and Mutect2. The AUC is estimated based on the different negative sets: The cross-patient calls, 500 random validation alterations and these sets combined (All). The AUC is also estimated for the full group of patients (All), and the patients with stage I and stage II CRC, individually (ns: p≥0.05, *: p<0.05, **: p<0.01, ***: p<0.001, ****: p<0.0001).

Figure 6:

Prediction of cancer using DREAMS-cc (a) and Shearwater (b). For each patient's LB-sample (y-axis) the mutation catalogue (x-axis) for every candidate patient is used for calling cancer. The patients are stratified into patients with stage I and stage II CRC, respectively. The diagonal is showing the result of using a patient's own mutation catalogue for cancer calling and constitutes the expected positives. The off diagonal is the cross-patient results, for which the mutation catalogue is filtered with the patient's tumour and germline variants prior to cancer calling, and thus these are expected to be negative. The colour scheme is chosen based on the matched quantiles from the p-value and combined Bayes factors from DREAMS-cc (a) and Shearwater, respectively. The cancer predictions show the results from one split in the 5x2 CV. c) AUC performance of DREAMS-cc and shearwater with respect to calling cancer.

References

1. Hu Z, Chen H, Long Y, Li P, Gu Y: **The main sources of circulating cell-free DNA: Apoptosis, necrosis and active secretion.** *Critical Reviews in Oncology/Hematology* 2021, **157:**103166.
2. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, et al: **Detection of circulating tumor DNA in early- and late-stage human malignancies.** *Science translational medicine* 2014, **6:**224ra224-224ra224.
3. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, Anagnostou V, Fiksel J, Cristiano S, Papp E, et al: **Direct detection of early-stage cancers using circulating tumor DNA.** *Science Translational Medicine* 2017, **9:**eaan2415.
4. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, Le Quesne J, Moore DA, Veeriah S, Rosenthal R, et al: **Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution.** *Nature* 2017, **545:**446-451.

5. Coakley M, Garcia-Murillas I, Turner NC: **Molecular Residual Disease and Adjuvant Trial Design in Solid Tumors.** *Clinical Cancer Research* 2019, **25:**6026-6034.

6. Henriksen TV, Tarazona N, Frydendahl A, Reinert T, Gimeno-Valiente F, Carbonell-Asins JA, Sharma S, Renner D, Hafez D, Roda D, et al: **Circulating Tumor DNA in Stage III Colorectal Cancer, beyond Minimal Residual Disease Detection, toward Assessment of Adjuvant Therapy Efficacy and Clinical Behavior of Recurrences.** *Clinical Cancer Research* 2022, **28:**507-517.

7. Øgaard N, Reinert T, Henriksen TV, Frydendahl A, Aagaard E, Ørntoft M-BW, Larsen MØ, Knudsen AR, Mortensen FV, Andersen CL: **Tumour-agnostic circulating tumour DNA analysis for improved recurrence surveillance after resection of colorectal liver metastases: A prospective cohort study.** *European Journal of Cancer* 2022, **163:**163-176.

8. Cescon DW, Bratman SV, Chan SM, Siu LL: **Circulating tumor DNA and liquid biopsy in oncology.** *Nature Cancer* 2020, **1:**276-290.

9. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, Thornton K, Agrawal N, Sokoll L, Szabo SA, et al: **Circulating mutant DNA to assess tumor dynamics.** *Nature Medicine* 2008, **14:**985-990.

10. Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, Cheang M, Osin P, Nerurkar A, Kozarewa I, et al: **Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer.** *Science Translational Medicine* 2015, **7:**302ra133-302ra133.

11. Corcoran RB, Chabner BA: **Application of Cell-free DNA Analysis to Cancer Treatment.** *New England Journal of Medicine* 2018, **379:**1754-1765.

12. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, Stehr H, Liu CL, Bratman SV, Say C, et al: **Integrated digital error suppression for improved detection of circulating tumor DNA.** *Nature Biotechnology* 2016, **34:**547-555.

13. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J, et al: **Analysis of error profiles in deep next-generation sequencing data.** *Genome Biology* 2019, **20**.

14. Chen L, Liu P, Evans Thomas C, Ettwiller Laurence M: **DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification.** *Science* 2017, **355:**752-756.

15. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L: **Calling Somatic SNVs and Indels with Mutect2.** Cold Spring Harbor Laboratory; 2019.

16. Gerstung M, Papaemmanuil E, Campbell PJ: **Subclonal variant calling with multiple samples and prior knowledge.** *Bioinformatics* 2014, **30:**1198-1204.

17. Zviran A, Schulman RC, Shah M, Hill STK, Deochand S, Khamnei CC, Maloney D, Patel K, Liao W, Widman AJ, et al: **Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring.** *Nature Medicine* 2020, **26:**1114-1124.

18. Wan JCM, Heider K, Gale D, Murphy S, Fisher E, Mouliere F, Ruiz-Valdepenas A, Santonja A, Morris J, Chandrananda D, et al: **ctDNA monitoring using patient-specific sequencing and integration of variant reads.** *Science Translational Medicine* 2020, **12:**eaaz8084.

19. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G: **Systematic evaluation of error rates and causes in short samples in next-generation sequencing.** *Scientific Reports* 2018, **8:**10950.

20. Huptas C, Scherer S, Wenning M: **Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly.** *BMC Research Notes* 2016, **9:**269.

21. Meacham F, Boffelli D, Dhahbi J, Martin DIK, Singer M, Pachter L: **Identification and correction of systematic error in high-throughput sequence data.** *BMC Bioinformatics* 2011, **12:**451.

22. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR: **Analysis of the Size Distributions of Fetal and Maternal Cell-Free DNA by Paired-End Sequencing.** *Clinical Chemistry* 2010, **56:**1279-1286.

23. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L: **Distribution-Free Predictive Inference for Regression.** *Journal of the American Statistical Association* 2018, **113:**1094-1111.

24. Hayama T, Hashiguchi Y, Okamoto K, Okada Y, Ono K, Shimada R, Ozawa T, Toyoda T, Tsuchiya T, Iinuma H, et al: **G12V and G12C mutations in the gene KRAS are associated with a poorer prognosis in primary colorectal cancer.** *International Journal of Colorectal Disease* 2019, **34:**1491-1496.

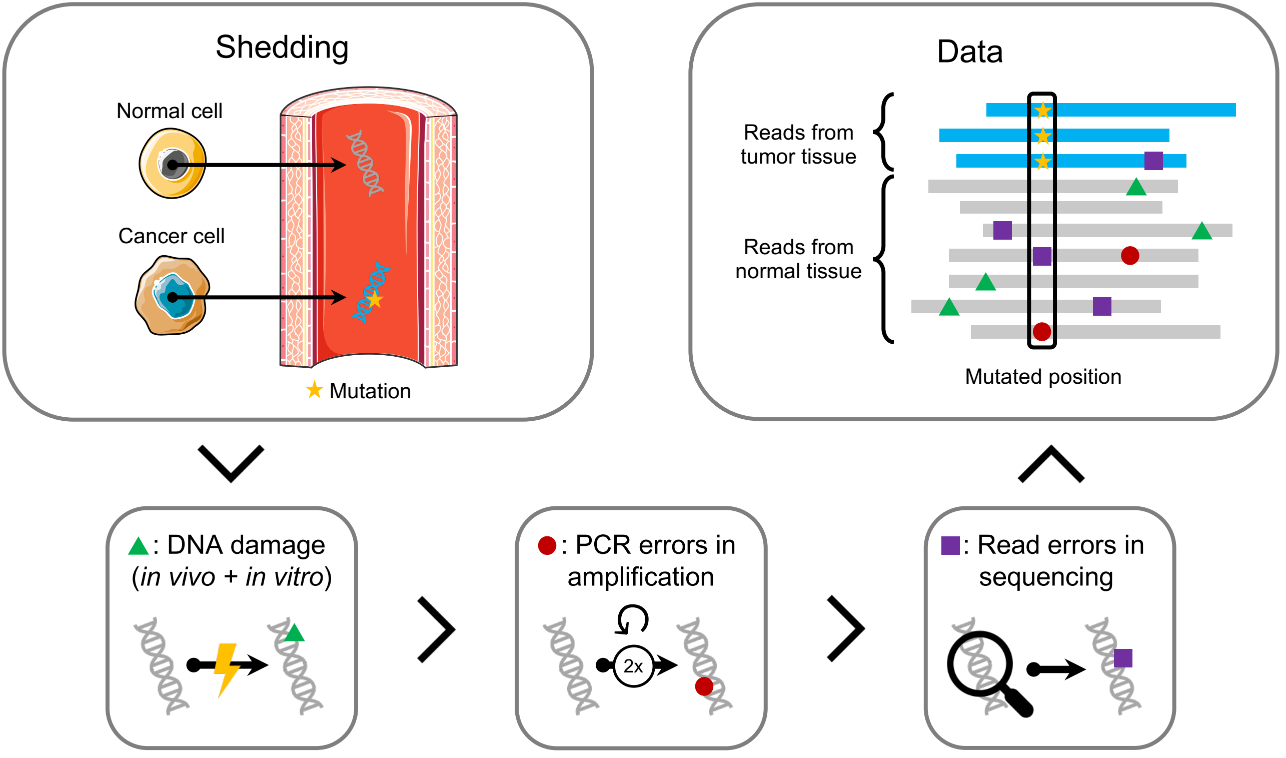
25. Parikh A, Goyal L, Hazar-Rethinam M, Siravegna G, Blaszkowsky L, Russo M, Van Seventer E, Nadres B, Shahzade H, Clark J, et al: **Systematic liquid biopsy identifies novel and heterogeneous mechanisms of acquired resistance in gastrointestinal (GI) cancer patients.** *Annals of Oncology* 2017, **28:**iii137.

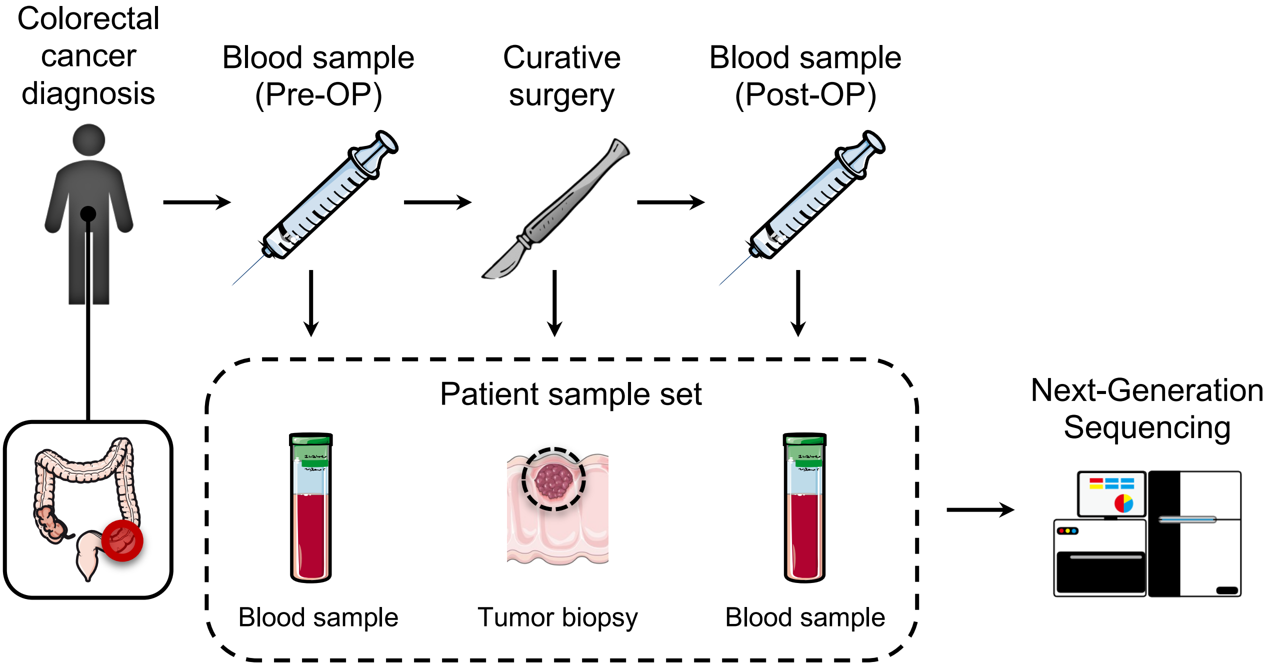
26. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, Jensen S, Medina JE, Hruban C, White JR, et al: **Genome-wide cell-free DNA fragmentation in patients with cancer.** *Nature* 2019, **570:**385-389.

27. Mouliere F, Rosenfeld N: **Circulating tumor-derived DNA is shorter than somatic DNA in plasma.** *Proceedings of the National Academy of Sciences* 2015, **112:**3178-3179.

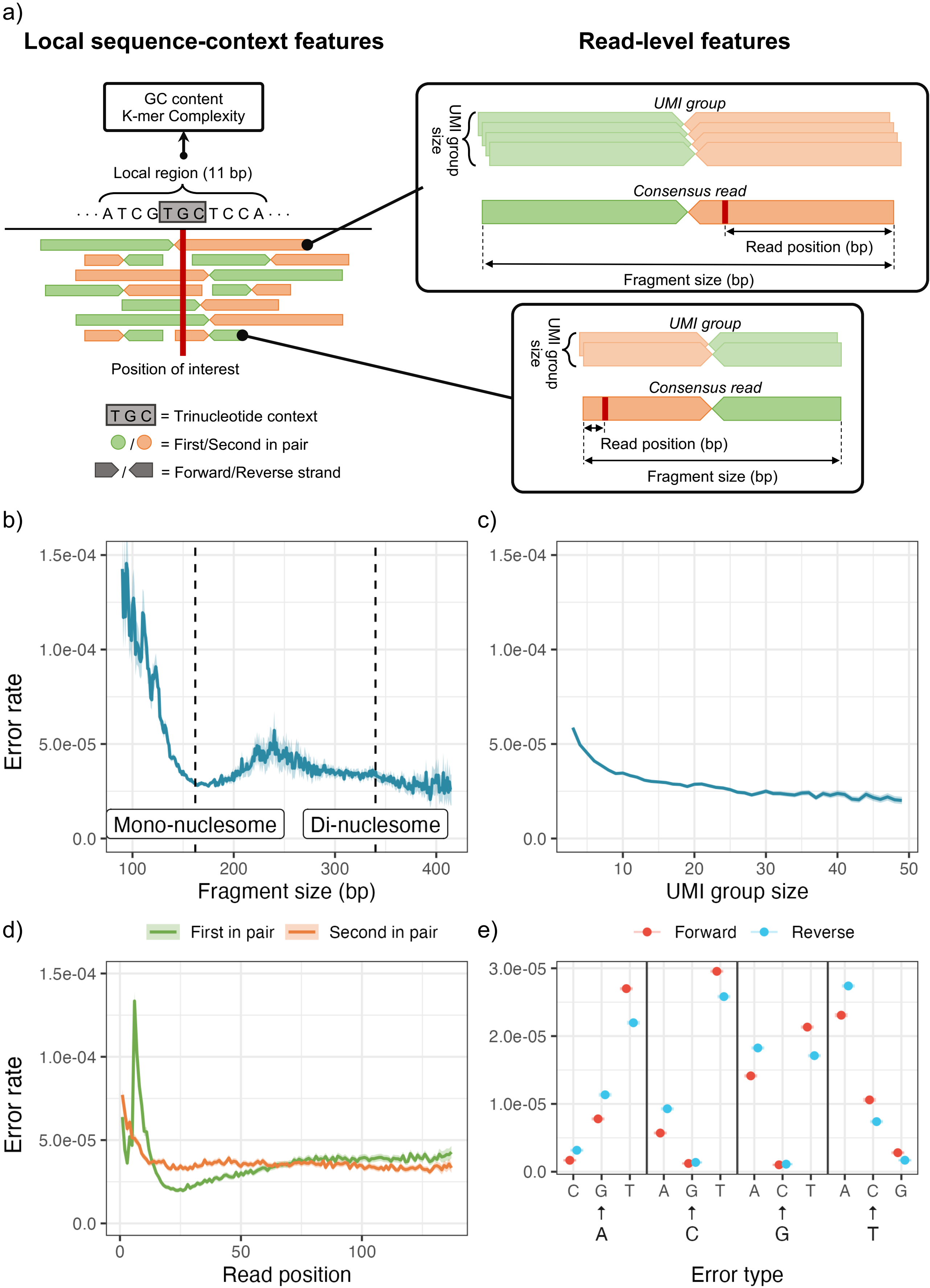
28. Pozzolo AD, Caelen O, Johnson RA, Bontempi G: **Calibrating Probability with Undersampling for Unbalanced Classification.** In; *2015*. IEEE;

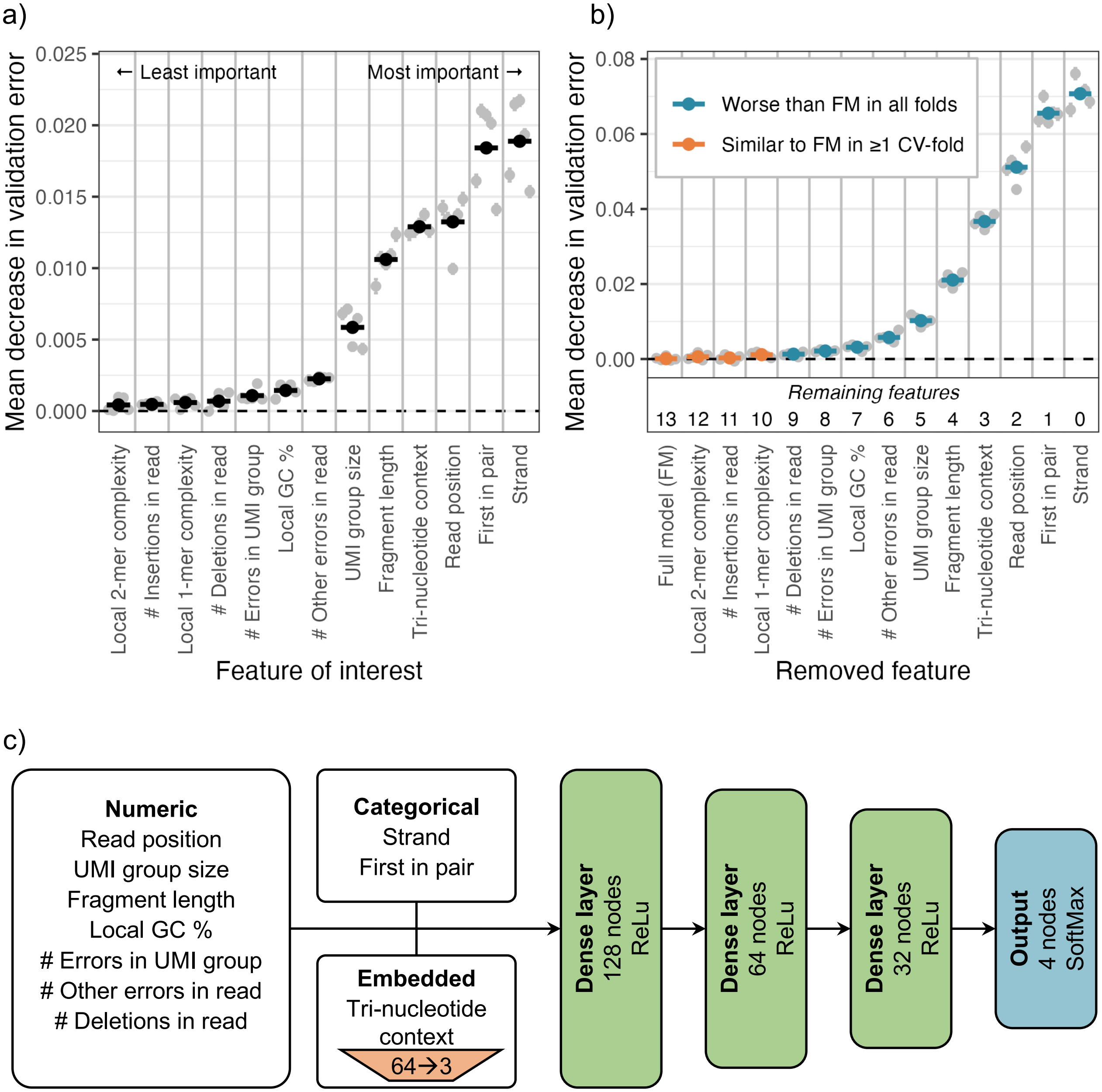
29. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M: **Tensorflow: Large-scale machine learning on heterogeneous distributed systems.** *arXiv preprint arXiv:160304467* 2016.

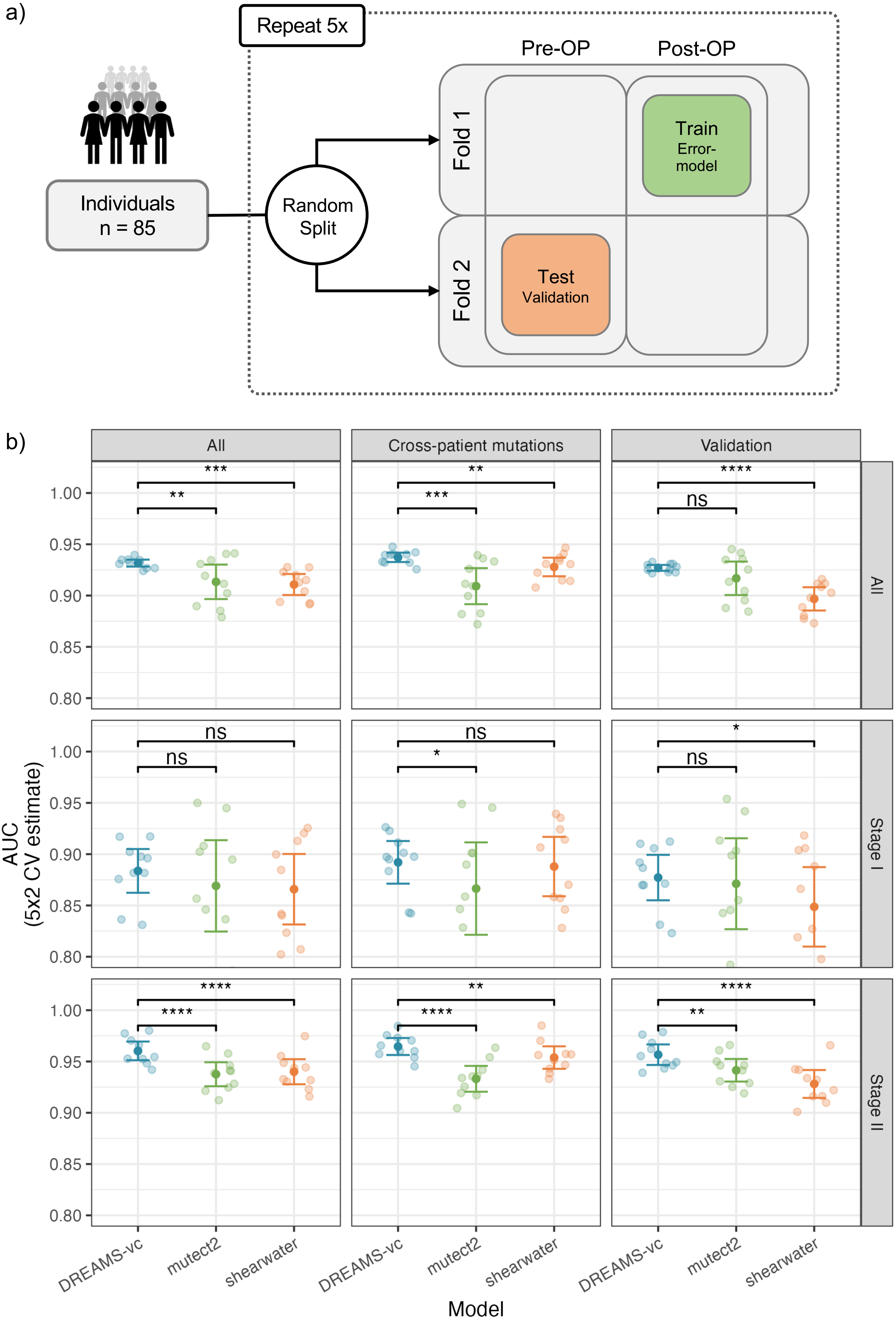

Shedding
Normal cell
Cancer cell
★ Mutation
Data
Reads from tumor tissue
Reads from normal tissue
Mutated position
▲: DNA damage (in vivo + in vitro)
●: PCR errors in amplification
2x
■: Read errors in sequencing

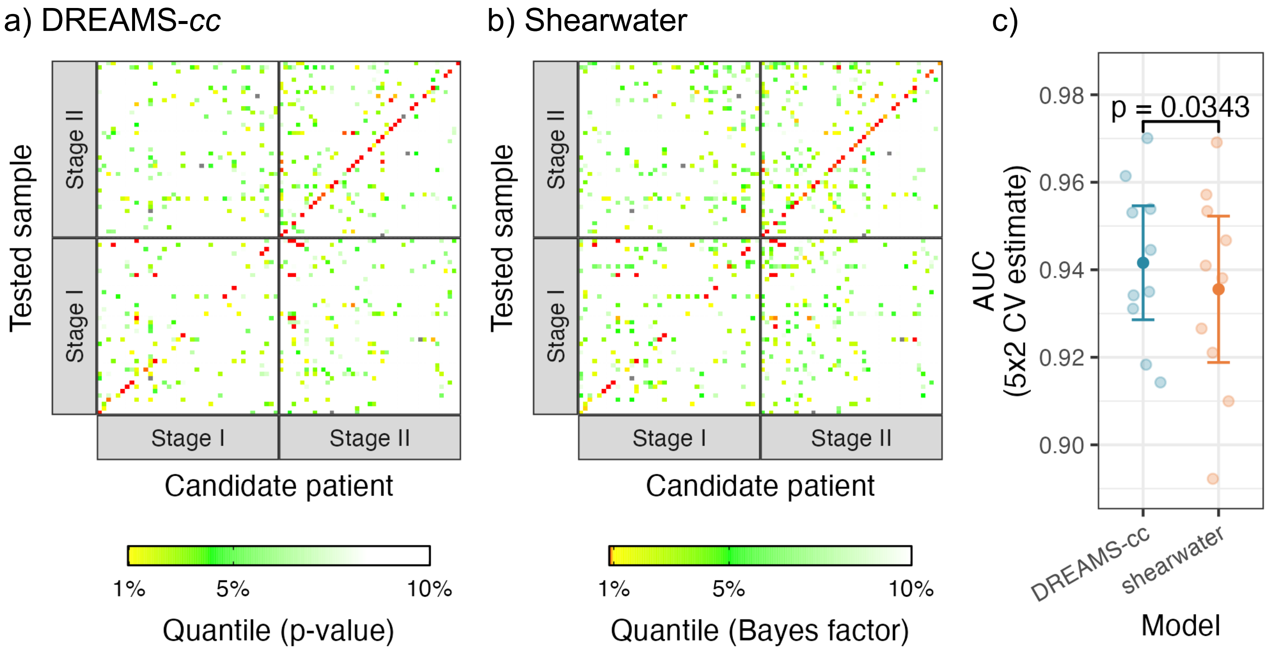


Colorectal cancer diagnosis
Blood sample (Pre-OP)
Curative surgery
Blood sample (Post-OP)
Patient sample set
Blood sample
Tumor biopsy
Blood sample
Next-Generation Sequencing


a)
Local sequence-context features
Read-level features
GC content
K-mer Complexity
Local region (11 bp)
· · · A T C G T G C T C C A · · ·
Position of interest
T G C = Trinucleotide context
● / ● = First/Second in pair
▶ / ◀ = Forward/Reverse strand
UMI group
UMI group size
Consensus read
Read position (bp)
Fragment size (bp)
UMI group
UMI group size
Consensus read
Read position (bp)
Fragment size (bp)
b)
Error rate
1.5e-04
1.0e-04
5.0e-05
0.0
Mono-nucleosome
Di-nucleosome
100
200
300
400
Fragment size (bp)
c)
1.5e-04
1.0e-04
5.0e-05
0.0
10
20
30
40
50
UMI group size
d)
First in pair
Second in pair
Error rate
1.5e-04
1.0e-04
5.0e-05
0.0
0
50
100
Read position
e)
Forward
Reverse
3.0e-05
2.0e-05
1.0e-05
0.0
C G T
A
A G T
C
A C T
G
A C G
T
Error type


a)
Mean decrease in validation error
0.025
0.020
0.015
0.010
0.005
0.000
← Least important
Most important →
Local 2-mer complexity
# Insertions in read
Local 1-mer complexity
# Deletions in read
# Errors in UMI group
Local GC %
# Other errors in read
UMI group size
Fragment length
Tri-nucleotide context
Read position
First in pair
Strand
Feature of interest
b)
Mean decrease in validation error
0.08
0.06
0.04
0.02
0.00
Worse than FM in all folds
Similar to FM in ≥1 CV-fold
Remaining features
13 12 11 10 9 8 7 6 5 4 3 2 1 0
Full model (FM)
Local 2-mer complexity
# Insertions in read
Local 1-mer complexity
# Deletions in read
# Errors in UMI group
Local GC %
# Other errors in read
UMI group size
Fragment length
Tri-nucleotide context
Read position
First in pair
Strand
Removed feature
c)
Numeric
Read position
UMI group size
Fragment length
Local GC %
# Errors in UMI group
# Other errors in read
# Deletions in read
Categorical
Strand
First in pair
Embedded
Tri-nucleotide context
64→3
Dense layer
128 nodes
ReLu
Dense layer
64 nodes
ReLu
Dense layer
32 nodes
ReLu
Output
4 nodes
SoftMax


a)
Repeat 5x
Individuals
n = 85
Random
Split
Pre-OP
Post-OP
Fold 1
Fold 2
Train
Error-
model
Test
Validation
b)
All
Cross-patient mutations
Validation
All
Stage I
Stage II
AUC
(5x2 CV estimate)
1.00
0.95
0.90
0.85
0.80
***
**
****
ns
*
DREAMS-vc
mutect2
shearwater
Model


a) DREAMS-*cc*
b) Shearwater
c)
Tested sample
Stage II
Stage I
Candidate patient
Stage I
Stage II
1%
5%
10%
Quantile (p-value)
Quantile (Bayes factor)
p = 0.0343
AUC
(5x2 CV estimate)
0.98
0.96
0.94
0.92
0.90
DREAMS-cc
shearwater
Model