Error Characterization and Statistical Modeling Improves Circulating Tumor DNA Detection by Droplet Digital PCR

Tenna V. Henriksen,a,b,† Simon O. Drue,a,b,† Amanda Frydendahl,a,b Christina Demuth,a,b Mads H. Rasmussen,a,b Thomas Reinert,a,b Jakob S. Pedersen,a,b,† and Claus L. Andersen*a,b,*,†

BACKGROUND: Droplet digital PCR (ddPCR) is a widely used and sensitive application for circulating tumor DNA (ctDNA) detection. As ctDNA is often found in low abundance, methods to separate low-signal readouts from noise are necessary. We aimed to characterize the ddPCR-generated noise and, informed by this, create a sensitive and specific ctDNA caller.

METHODS: We built 2 novel complimentary ctDNA calling methods: dynamic limit of blank and concentration and assay-specific tumor load estimator (CASTLE). Both methods are informed by empirically established assay-specific noise profiles. Here, we characterized noise for 70 mutation-detecting ddPCR assays by applying each assay to 95 nonmutated samples. Using these profiles, the performance of the 2 new methods was assessed in a total of 9447 negative/positive reference samples and in 1311 real-life plasma samples from colorectal cancer patients. Lastly, performances were compared to 7 literature-established calling methods.

RESULTS: For many assays, noise increased proportionally with the DNA input amount. Assays targeting transition base changes were more error-prone than transversion-targeting assays. Both our calling methods successfully accounted for the additional noise in transition assays and showed consistently high performance regardless of DNA input amount. Calling methods that were not noise-informed performed less well than noise-informed methods. CASTLE was the only calling method providing a statistical estimate of the noise-corrected mutation level and call certainty.

CONCLUSIONS: Accurate error modeling is necessary for sensitive and specific ctDNA detection by ddPCR. Accounting for DNA input amounts ensures specific detection regardless of the sample-specific DNA concentration. Our results demonstrate CASTLE as a powerful tool for ctDNA calling using ddPCR.

Introduction

Circulating tumor DNA (ctDNA) investigation is a rapidly expanding field with escalating requirements for sensitivity and specificity (1–3). A large proportion of ctDNA studies utilizes droplet digital PCR (ddPCR) technology for rare event detection (4–15), as it is simple and inexpensive compared to most sequencing-based alternatives. Additionally, the technology is highly sensitive, allowing for detection of allelic frequencies down to approximately 0.01% (16, 17).

In ddPCR, a DNA sample is partitioned into tens of thousands of droplets prior to PCR amplification. Thereby the DNA fragments become randomly distributed, ideally with most droplets containing either zero or 1 target, but occasionally with multiple targets in 1 droplet (18). The target is subsequently PCR-amplified in each droplet. Fluorescent probes are used to indicate whether a droplet contained the targeted sequence. In mutation-based ctDNA detection, often a single PCR reaction is carried out amplifying both the wild-type and mutated templates, which are distinguished using two differently labeled probes.
The ddPCR technology is widely used both for tracking known resistance variants (4–6) and for personalized ctDNA assessments aimed at minimal residual disease detection (7–13). Relative to the total cell-free DNA (cfDNA), ctDNA is often found in extremely low abundance. Hence, it is important to distinguish between low-signal readouts and experimental noise. Commonly, a signal limit is defined, which needs to be exceeded for a sample to be called positive. In some studies, these limits are set seemingly arbitrarily (5–7, 12, 14), while in others, they are based on empirical assessment of noise in nonmutated samples (9, 10, 13, 15). These noise-informed limits have been implemented either as a static limit for all samples (5–8, 10–13), or as dynamic limits, adapting to the amount of DNA investigated (19, 20). Most often, these approaches result in a TRUE/FALSE call, leaving little room for nuance and interpretation.

For this study, we have developed 2 new approaches for ddPCR-based ctDNA calling, 1 of which also provides a statistical estimate and a CI of the true, noise-corrected concentration of target fragments. We compare our 2 calling approaches against 7 different calling methods previously used for ctDNA detection in the literature. We assess the performance on 9447 laboratory reference samples, as well as 1311 clinical plasma samples from colorectal cancer (CRC) patients, examining data from 70 ddPCR assays.

**Materials and Methods**

**PARTICIPANT INCLUSION**

For model training data, blood samples were obtained from 200 healthy blood donors aged ≥40 from the Blood Bank at Aarhus University Hospital. For clinical performance tests, CRC patients were enrolled prospectively from 10 surgical centers in Denmark. Patient enrollment was performed in accordance with the Declaration of Helsinki, and all participants provided written informed consent. Blood samples were collected before surgery (n = 847) and during follow-up after surgery (n = 464). For details on sample collection and DNA extraction, see the online Supplemental Information 1.1 in the online Data Supplement.

**NEGATIVE REFERENCE SAMPLES**

Peripheral blood leukocyte DNA was sonicated for 140 s using the E220 evolution focused-ultrasonicator (Covaris). Fragmentation was confirmed on TapeStation 4200 (Agilent). For each donor, a 5-point dilution series was made with 150 ng, 60 ng, 24 ng, 9.6 ng, and 3.84 ng DNA. Noise was assessed by applying ddPCR assays (n = 70) to dilution series from 19 donors (19 × 5 = 95 ddPCR-reactions/assay). Noise assessment was made for every new assay batch targeting the same mutation. In total, 103 batches were assessed. These measurements (n = 9168) were also used as negative reference samples.

**POSITIVE REFERENCE SAMPLES**

Tumor DNA harboring the targeted mutation (see online Supplemental Information 1.2–1.3) was diluted stepwise in ATE buffer (Qiagen). DNA from healthy individuals was added for a uniform DNA concentration of 3030 genome equivalents (GE) per µL, resulting in mutant allele frequencies of 1% (A), 0.3% (B), 0.1% (C), and 0.03% (D). All 70 ddPCR assays were tested, yielding 280 measurements, where 1 was excluded due to low droplet counts. These samples (n = 279) were used as positive reference samples.

**CLINICAL SAMPLES**

For each patient, a single clonal mutation, identified through whole exome sequencing (online Supplemental Information 1.2–1.3), was targeted by ddPCR. The cfDNA from 8 mL plasma was investigated with a patient-specific ddPCR assay. On average, our ddPCR setup produced 14 000 (SD 3000) droplets per 20 µL reaction volume. To enable precise quantifications, we aimed for a maximum of 5000 GE cfDNA per reaction. Standardly, cfDNA from 8 mL plasma was divided into 6 separate ddPCR reactions. For samples with high cfDNA levels (more than 6 × 5000 GE), additional reactions were set up, to avoid exceeding this limit.

**DIGITAL DROPLET PCR**

The assays consisted of a single primer set amplifying the target region and 2 probes reporting either the mutation or corresponding wild-type sequence. Primer/probe sequences and PCR protocols are provided in online Supplemental Table 1.

In brief, ddPCR was carried out in 20 µL reactions (online Supplemental Information 1.4–1.6). A water sample was run as a no-template control and a tumor DNA sample carrying the targeted mutation was used as positive control. PCR mixture was made according to manufacturer’s instructions (Bio-Rad). Droplets were formed on the Automated Droplet Generator (Bio-Rad) and analyzed on the QX200™ Droplet Reader (Bio-Rad). Droplet counts and no-template control results are included in online Supplemental Table 2.

**STATISTICAL ANALYSES**

False positive rates (FPRs) were compared by a 2-sided Wilcoxon rank-sum test. P values were adjusted for multiple testing using the Benjamini–Hochberg method, and adjusted P values < 0.05 were considered statistically significant.
Specificity was compared on the negative reference samples used for model fitting through a 10-fold cross validation. All samples from 1 individual were tested simultaneously, to ensure uniform distribution of every concentration in the training data.

Overall performance was assessed using the area under the curve (AUC) of ROC curves, precision-recall curves with F1-values annotated, and the geometric mean (G-mean) of the true-positive rate (TPR) and true negative rate (TNR). A 95% CI was calculated as a Wilson score interval for FPRs, and a bias-corrected and accelerated bootstrap interval (using 2000 replicates) for G-means.

All statistical calculations were made in R (v. 4.0.2).

Results

NOISE CHARACTERIZATION
We analyzed 70 ddPCR assays targeting different mutations. Each assay consisted of a set of primers amplifying the target region and 2 probes, allowing simultaneous quantification of mutated and wild-type fragments (Fig. 1, A). Errors in ddPCR will manifest as mutational signals in nonmutated samples. While their source is not well understood, Fig. 1, B outlines potential causes.

For each assay, a noise profile was established by applying the assay to a set of 95 negative reference samples, covering the expected input span when analyzing plasma cfDNA. In these nonmutated samples, some assays produced an erroneous mutational signal, and for some assays, the number of false-positive droplets increased with increasing DNA input (Fig. 1, C and D). Other assays produced no or few errors (Fig. 1, E). Noise-prone assays were often characterized by a large number of droplets with both mutational and wild-type signals.

Multiple different strategies have previously been employed to separate noise from signal. These strategies fall into 3 groups: (1) observing more mutational droplets than a given cutoff (6, 7, 11–15); (2) calculating a limit of blank (LOB) to be exceeded by the sample readout (9, 19, 20, 22); and (3) statistically testing if the signal observed is greater than expected (20, 22) (Fig. 1, F). The Droplet Cutoff can be used with varying cutoffs. Here, we explored cutoffs of 2 (M2), 3 (M3), and 5 (M5) mutational droplets, as well as 2 mutational droplets including at least 1 single positive for the mutation (M2S1). In the following sections, we propose 2 new calling strategies and compare these with preexisting methods. The requirements and features of all calling methods are compared in online Supplemental Fig. 2.

THE DYNAMIC LOB
We devised a calling strategy, to account for noise increasing with DNA input (Fig. 2, A). An LOB was calculated for each dilution in the negative reference data, and a power regression was used to interpolate the sample-specific LOB given the input. A sample was considered positive if the allele frequency was higher than the sample-specific LOB and more than 2 mutational droplets were observed.

THE CASTLE ALGORITHM
We built the concentration and assay-specific tumor load estimator (CASTLE) algorithm (Fig. 2, B) to estimate the mutant DNA concentration in a sample and quantify the significance of the mutational signal. Using an assay-specific error profile, CASTLE predicts the total droplet counts by modeling the expected signal from errors as well as wild-type and mutational DNA fragments in each droplet.

To estimate an error profile, 3 parameters representing different error sources (Fig. 1, B) were incorporated. Although similar errors would occur for the wild-type signal, these are omitted from the model, as their frequency is negligible compared to the true wild-type signal. First, the α-parameter models arbitrary errors, independent of the DNA concentration. Second, the β-parameter models errors from a wild-type fragment giving off a mutational signal. Since this can happen for each wild-type fragment, β errors would scale with the number of wild-type DNA fragments in a droplet and only give rise to double-positive droplets. Lastly, the γ-parameter models errors of nontargeted probe-DNA interactions, which expectedly will increase with the DNA concentration in the droplet. The combined error rate is

\[
\text{error rate} = \alpha + \beta \cdot w + \gamma \cdot \lambda,
\]

where β is multiplied by the number of wild-type DNA fragments in that droplet (w) and γ by the DNA concentration in the droplet (λ). We then assume the number of wild-type (NW) and mutational (NM) fragments in a droplet to follow a Poisson distribution:

\[
NW \sim \text{Pois}(\lambda),
\]

\[
NM(NW = w) \sim \text{Pois}(\alpha + \beta \cdot w + \gamma \cdot \lambda + \rho),
\]

where ρ is the true mutational DNA concentration in the droplet.

The 4 observable outcomes for a droplet can be described as double negative (NW = 0, NM = 0), wild-type single positive (NW ≥ 1, NM = 0), mutational single positive (NW = 0, NM ≥ 1), and double positive (NW ≥ 1, NM ≥ 1). The sample result is the droplet counts in each of these categories.

To estimate the error rates, a set of nonmutated samples are analyzed. Assuming fixed error rates for an assay and no mutational DNA present (ρ = 0), α, β, and γ can
Fig. 1. Droplet Digital PCR generated noise. (A) Schematic overview of the principle in the used ddPCR assays. One probe (green) targets the wild-type sequence and the other probe (blue) targets the mutated sequence. Fluorescent amplitudes reflect the presence or absence of the targeted DNA fragment. (B) Schematic overview of possible errors in ddPCR. Errors giving rise to droplets with a wild-type-only signal are not shown, as the contribution to the overall wild-type signal is considered negligible. (C) Observed noise when measuring negative reference samples (grey, n = 2561). Example of assay noise and log-linear regression shown in blue for KRAS_c.34G>A_p.G12S. Only samples with minimum 1 observed droplet are depicted. (D) Example of 2D amplitude observed for all negative reference samples run with the KRAS_c.34G>A_p.G12S assay. (E) Example of 2D amplitude observed for all negative reference samples run with the KRAS_c.34G>T_p.G12C assay. (F) Simplified schematic overview of 3 different ddPCR calling strategies: droplet cutoff, LOB calculation, and a statistical estimate. Examples of models have been included static LOB (20) and ALPACA (19) as limit of blank calculations and the Poisson model (20) as a statistical estimate. FP = false-positive count, FPR = false-positive rate (PIF corrected), GE = genome equivalents, LOB = limit of blank, Mut = mutation, PIF = polymerase-induced false event, RFP = rate of false positives (false mutational copies/wild-type copies), WT = wild type.
be estimated (Fig. 2, C) using maximum likelihood theory (online Supplemental Information 2.1–2.3).

Finally, a likelihood-ratio test determines if the observed mutational signal is significantly higher than expected from non-mutated samples.

In this study, CASTLE was trained on 95 nonmutated negative reference samples for every assay but displayed robust performance when trained on only 30 samples per assay, preferably spanning multiple DNA concentrations (online Supplemental Fig. 3). A
complete CASTLE R-package is available at https://simondrue.github.io/castle/.

NOISE LEVELS DEPEND ON TARGETED MUTATION
Assays targeting different mutational base changes at the same genomic position can show different noise profiles (Fig. 1, D and E). To elaborate on this, the $\alpha$, $\beta$, and $\gamma$ error rates were plotted for all assays (Fig. 2, D), grouped by mutational type (transition ($n = 38$), transversion ($n = 29$), or indel ($n = 2$)). In total, 82% (31/38) of transition assays showed higher error rates than the noisiest transversion assay ($P < 0.0001$, Wilcoxon rank-sum test). The noisiest transition assays were $C > T$ and $G > A$, while $A > G$ and $T > C$ showed error rates similar to transversion assays (Fig. 2, E). The dominant error type in the noisiest assays were $\beta$ errors, encompassing PCR errors. In Fig. 1, D, the droplets follow a nonrandom pattern of discrete clouds rising from the wild-type-only signal, consistent with PCR errors in different PCR-cycles. This pattern was visible for 30 of 34 $C > T$ and $G > A$ assays (online Supplemental Fig. 1).

PERFORMANCE IN POSITIVE AND NEGATIVE REFERENCE SAMPLES
We compared the performance of 4 existing calling approaches (droplet cutoff, static LOB, ALPACA, and Poisson) (Fig. 1, F) to the performance of dynamic LOB and CASTLE. The specificity of each method was assessed by 10-fold cross-validation on positive reference samples ($n = 9168$) and the sensitivity on positive reference samples (tumor DNA in a set amount of wild-type DNA; $n = 279$). As samples without mutational droplets always would be classified as negative, we only compared samples containing one or more mutational droplets. This left 2561 negative and 268 positive reference samples for analysis. The CASTLE algorithm had the highest AUC (AUC = 0.97) on ROC curves (Fig. 3, A) and similarly showed the best performance on precision-recall curves ($F1 = 0.89$) (Fig. 3, B).

FPRs were compared between methods as the fraction of negative reference samples called mutation positive. Overall, the CASTLE FPR was significantly lower (1%) than all other FPRs (range 2%–48%, $P < 0.0001$) (Fig. 3, C; online Supplemental Table 3). The same was the case when focusing only on transitions. Conversely, for transversion assays, the droplet cutoff methods reached FPRs (range 1%–8%) comparable to CASTLE (4%), whereas ALPACA had significantly higher FPR (23%, $P < 0.0001$). The static and dynamic LOB approaches had comparable FPRs between transition and transversion assays, while CASTLE, Poisson, and ALPACA had significantly higher FPRs in transversion assays compared to transitions (Fig. 3, C; online Supplemental Table 4). In contrast, all droplet cutoff approaches showed significantly higher FPRs in transitions compared to transversions.

To assess the performance of each model at different mutant concentrations, the geometric mean (G-mean) of the TPR and TNR was compared by determining a TPR for different tumor concentrations in positive reference samples. The TNR was calculated using the negative reference samples. As expected, the G-mean increased with increasing tumor allele frequency (Fig. 3, D). Again, the droplet cutoff methods performed less well than noise-informed methods. Restricting analyses to transition assays revealed worse performance for the Poisson and ALPACA methods in samples with low allele frequency. Here, CASTLE had the best performance of the noise-informed methods. In transversion assays, ALPACA performed less well than all other methods in samples with high allele frequency.

PERFORMANCE ON CLINICAL PLASMA ctDNA SAMPLES
To compare clinical performance, each calling method was used to classify a set of clinical case and control samples. Cases were preoperative plasma samples from stage I to III CRC patients ($n = 847$), and controls were postoperative plasma samples from nonrecurrence CRC patients with at least 1 year of follow-up ($n = 464$). ROC and precision-recall curves were used to compare ctDNA calls stratified by the Union for International Cancer Control (UICC) stages (Fig. 4, A; online Supplemental Table 5). For every method, performance increased with increasing UICC stage, and stage I patients were difficult to classify (AUC range 0.6–0.68). In stage II and III patients, CASTLE, Poisson, and ALPACA had the highest AUCs and best performance on precision-recall curves.

FPRs in control plasma samples were used to compare specificities. Among transitions ($n = 268$), droplet cutoff methods had the highest FPRs (range 22%–60%) (Fig. 4, B). The static LOB method similarly displayed a high FPR (22%), whereas the remaining noise-informed methods had the lowest FPRs (range 3%–13%). Only the Poisson method (3%) had a significantly lower FPR than CASTLE (8%). Methods with high FPRs showed a tendency for positive calls across all investigated DNA inputs (Fig. 4, C). Conversely, methods with low FPRs were either vulnerable to low DNA amounts (ALPACA), vulnerable to high DNA amounts (static LOB, M5), or not affected by changes in DNA input (CASTLE, Poisson, dynamic LOB) (Fig. 4, D). When targeting transversions ($n = 198$) (Fig. 4, D), no significant difference in FPR was observed between CASTLE (3%) and other methods (range 3%–7%) except ALPACA (12%; $P = 0.0015$). The methods showed no difference in tendency for false-positive calls with
Fig. 3. Detection of mutations in laboratory reference samples. ROC (A) and precision-recall (B) curves for 6 calling methods used to detect mutations using ddPCR: CASTLE, Poisson, ALPACA, dynamic LOB, static LOB, and droplet cutoff. Expected negative samples \( n = 2561 \) were negative reference samples (assessed by 10-fold cross validation) with minimum 1 mutational droplet observed. Expected positive samples \( n = 268 \) were positive reference samples with minimum 1 mutational droplet observed. Dots indicate the cutoff for classifying a sample as positive, which was used in the F1 calculation. For CASTLE and Poisson a \( P \) value less than 0.01 was used to classify a sample as positive. For ALPACA, dynamic LOB, and static LOB samples were called as shown in Figs. 1, F and 2, A. For the droplet cutoff method a cutoff of 2 mutational droplets was applied. (C) FPR in negative reference samples with minimum 1 observed mutational droplet assessed by 10-fold cross-validation. FPR in all samples \( n = 2561 \), samples with transition targets \( n = 2473 \) and samples with transversion targets \( n = 121 \) for various calling methods were compared to CASTLE FPR using a Wilcoxon rank-sum test. Pairwise comparisons of all methods can be found in online Supplemental Table 1. Error bars indicate 95% CIs. Significance in difference between CASTLE and other calling methods are indicated as follows: * \( P < 0.05 \); ** \( P < 0.01 \); *** \( P < 0.001 \); **** \( P < 0.0001 \). (D) G-mean of TPR determined in positive reference samples with different allele frequencies and TNR determined in negative reference samples (C). Number of positive reference samples for each dilution point denoted by \( n \) for all assays, transition assays, and transversion assays.
Fig. 4. Detection of mutations in clinical samples. (A) ROC (top) and precision-recall (bottom) curves for the 6 calling methods on plasma samples from CRC patients with UICC stages I to III. Negative control samples (n = 464) were postoperative samples taken after end of definitive therapy, during follow-up in nonrecurrence CRC patients. Expected positive case plasma samples (stage I: n = 222; stage II: n = 416; stage III: n = 209) were preoperative samples from CRC patients. Dots indicate the cutoff (described in Fig. 3) for classifying a sample as positive. (B) FPR in control plasma samples. Rates are grouped according to type of targeted base change. Error bars indicate 95% CIs. Significance in difference between CASTLE and other callers using a Wilcoxon rank sum test are indicated as follows: *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001. (C) Amount of DNA (genome equivalents) investigated for control plasma samples. Gaussian kernel density plot of DNA input in false-positive samples given for each calling method in arbitrary units (a.u.). Plot in the upper-right corner shows distribution for all control plasma samples. (D) G-mean of TPR determined in case plasma samples and TNR determined in control plasma samples (C).
varying input amounts (Fig. 4, C). Every method except ALPACA and Poisson showed a significantly lower FPR for transversions compared to transitions (online Supplemental Table 6).

The G-mean was calculated for each UICC stage, using the TNR calculated on clinical control samples as a constant across every UICC stage (Fig. 4, D). G-mean scores improved with increasing stage. On stage II and III patients, the noise-informed methods (except static LOB) outperformed droplet cutoff methods in transition assays, whereas performance was comparable between all methods (except M5) on transversion assays.

Discussion

For rare-event detection by ddPCR to be clinically relevant, achieving high sensitivity while retaining high specificity is paramount. This requires the ability to accurately measure, model, and control for noise generated by the ddPCR assay.

We tested the clinical performance of 9 calling methods by comparing ctDNA detection in samples from CRC patients collected before surgery (expected positive) and after surgery in nonrecurrence patients (expected negative). Detection of ctDNA in a sample is contingent on the sample containing the targeted DNA sequence. In preoperative samples, this largely depends on the sampling process, which becomes stochastic in low ctDNA concentrations. Previous studies have shown that detection increases with disease stage (24–26), likely reflecting increased shedding and higher ctDNA concentrations. Consistent with this, our results showed that the performance of all calling methods improved with increasing disease stages. Overall, the noise-informed callers performed better compared to un-informed callers. This was also the case in reference samples, with CASTLE showing the lowest FPR and retaining high sensitivity.

For maximal clinical usability, a calling method should be applicable across multiple targets and input DNA concentrations. In plasma samples, methods considering the DNA input (CASTLE, Poisson, ALPACA and dynamic LOB) had the highest overall performance, indicating that utilizing this information is beneficial. Interestingly, among transition mutations, ALPACA was most error-prone in samples with relatively low DNA inputs (approximately 10 000 GE), whereas other methods displayed false-positive calls with high DNA inputs (approximately 50 000 GE). Thus, ALPACA likely accounts differently for the type of noise in transition assays. The static LOB and droplet cutoff methods, which are not informed by the DNA input amount, were vulnerable to changes in input. Of note, when the specificity of the droplet cutoff caller was increased by requiring more positive droplets, there was a shift from false-positive observations being made at all concentrations to false-positive calls being made primarily at high ctDNA concentrations. This highlights the need for dynamic calling methods to adjust for the input DNA amount. Even the most specific of the static calling methods had a 2-fold higher FPR compared to the dynamic algorithms.

In general, model performances were less divergent when applied to data from transversion assays than transition assays, suggesting differing error patterns between mutational types. Our analyses showed that assays targeting transitions were generally more error-prone than assays targeting transversions. The CASTLE-modeled error rates were vastly higher for C > T and G > A transitions compared to T > C and A > G transitions. This pattern fits the known error profile for the Taq-polymerase used in ddPCR reactions (27, 28). As expected, the β error rate estimated by CASTLE to reflect PCR errors (among others) was the biggest contributor to the combined error rate for noise-prone assays. While the Taq-polymerase is known to be more error prone than other high-fidelity polymerases (29, 30), the Taq-polymerase comes standard in the ddPCR reaction mixes provided by several vendors. We hypothesize that replacing the Taq-polymerase with a less error-prone polymerase may reduce the FPR, making ddPCR better suited for rare-event detection. Further studies are needed to confirm this.

Of note, assays with C > T and G > A transitions were the most common in our assay compendium. Therefore, the error estimates in these mutations are likely better generalized than errors occurring with less represented mutations. The compendium is enriched for clonal mutations that occur frequently in CRC. However, when choosing mutations for plasma ddPCR analysis, the mutational base change may be worth keeping in mind to increase the theoretical sensitivity. Conversely, when aiming to detect specific resistance variants using ctDNA, choice of target is limited and accurate error modelling is the only way to ensure sensitive and specific detection.

Which method to choose for ctDNA calling in ddPCR depends on the clinical application. Simply setting a cutoff of mutational droplets may seem attractive, as it does not require any previous noise characterization and requires little computational effort. Although such methods in general performed less well in our testing, performance was comparable to noise-informed methods when targeting transversion mutations. However, better overall performance would be achieved using noise-informed and dynamic methods such as CASTLE, Poisson, ALPACA, or the dynamic LOB. As the performances of these models were comparable, which to choose is a matter of preference. CASTLE and Poisson provide a P value, which informs on call certainty. CASTLE further estimates the number of noise-
corrected mutated fragments in a sample accompanied by a CI, additionally informing calling decisions. Moreover, we have developed CASTLE as a free, easy-to-implement, and well-documented R package (31), which can be used with any ddPCR output data. ALPACA and dynamic LOB offer similar performance, however, provide little to no information on call certainty and require noise samples assessed in predefined bins, which potentially is hard to implement.

In this study, analyses were designed to display differences between calling methods—for example, by only including reference samples with an observed mutational signal. Further, analyses on plasma samples are difficult to interpret, as the “true” mutational status of a sample is unknown. Thus, the performance level observed in this study may not reflect the absolute performance for each calling method.

The calling methods described here do not consider the fluorescent amplitude of each droplet, which may be informative, and user-defined fluorescence thresholds may have limited the data. In the future, ctDNA calling could be improved by instead weighting the amplitude of each droplet.

In conclusion, errors in ddPCR vary in nature. Assays targeting transition base changes were more error-prone than transversion-targeting assays. Accurate error modeling alleviates this problem enabling high sensitivity and specificity, which is central for clinical implementation of ddPCR for rare-event detection. We offer the CASTLE algorithm as an easy-to-implement ctDNA calling tool, which can give a statistical estimate of the true amount of mutational fragments in a sample, while demonstrating high sensitivity and specificity across different mutational types and a wide range of DNA input concentrations.

Supplemental Material

Supplemental material is available at Clinical Chemistry online.

Nonstandard Abbreviations: ctDNA, circulating tumor DNA; ddPCR, droplet digital PCR; cfDNA, cell-free DNA; FPR, false-positive rate; CRC, colorectal cancer; GE, genome equivalents; AUC, area under the curve; G-mean, geometric mean; TPR, true-positive rate; TNR, true-negative rate; LOB, limit of blank; CASTLE, concentration and assay-specific tumor load estimator; UICC, Union for International Cancer Control.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following four requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.

Authors’ Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.

Patents: None declared.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

Acknowledgments: We thank the blood donors, CRC patients, and the Danish Cancer Biobank for contributing clinical material. We also thank the IMPROVE study group for patient inclusion: Kåre Anderson Gotschalck, Lene Hjerrild Iversen, Uffe Schou Løve, Anders Husted Madsen, Ole Thorlacius-Ussing, Ismail Gogener, Per Vadgaard Andersen, Jakob Lykke, Allan Gorm Pedersen, and Nis Hallundbæk Schlesinger.

References

Error Characterization and Modeling in Digital PCR


