Measuring DNA quality by digital PCR using probability calculations

David Gleerup, [£, †, ‡] Yao Chen, [£,¶,‡] Willem Van Snippenberg[‖], Cedric Valcke,[□] Olivier Thas, [¶,‡,@,△] Wim Trypsteen[‖,‡] and Ward De Spiegelaere*[†, ‡]

Author Information

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. £ These authors contributed equally.

†Laboratory of Veterinary Morphology, Ghent University, Belgium.

‡Ghent University Digital PCR Consortium, Ghent University, Belgium.

¶Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

‖HIV Cure Research Center, Department of Internal Medicine and Pediatrics, Ghent University and Ghent University Hospital, Belgium.

@National Institute for Applied Statistics Research Australia (NIASRA),University of Wollongong, Australia

△Data Science Institute, I-BioStat, Hasselt University, Belgium

□ Department of Healthcare, HOGENT University of Applied Sciences and Arts, Belgium

Corresponding Author

Ward De Spiegelaere – Faculty of veterinary medicine, Ghent university, Campus Merelbeke, Salisburylaan 133, Merelbeke 9820, Belgium; https://orcid.org/ 0000-0003-2097-8439; email: Ward.DeSpiegelaere@ugent.be; Phone: +321292647738

Author emails

David Gleerup: Davidgrandjean.gleerup@ugent.be

Yao Chen: yao.chen@ugent.be

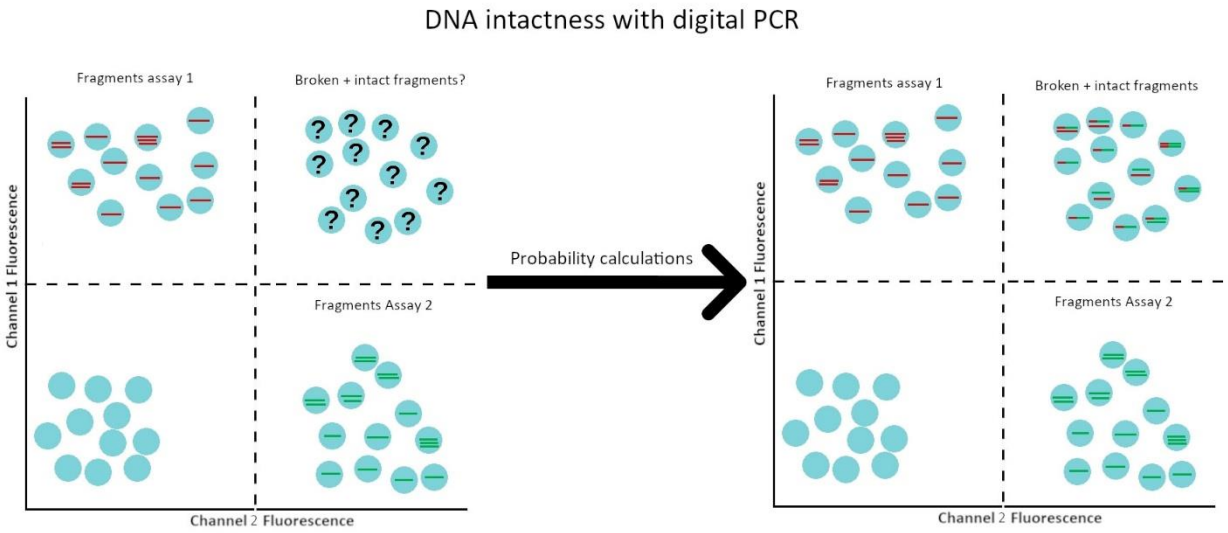
Willem Van Snippenberg: willem.vansnippenberg@ugent.be

Cedric Valcke: Cedric.Valcke@gmail.com

Olivier Thas: olivier.thas@ugent.be

Wim Trypsteen: wim.trypsteen@ugent.be

Graphical Abstract



Abstract

Background

Accurate methods to assess DNA integrity are needed for many biomolecular methods. A multiplex digital PCR (dPCR) method designed for interspaced target sequences can be used to assess sequence integrity of large DNA strands. The ratio of single positive partitions versus double positive partitions is then used to calculate the sheared DNA strands. However, this simple calculation is only valid with low DNA concentration. We here describe a method based on probability calculations which enables DNA quality analysis in a large dynamic range of DNA concentrations.

Results

Known DNA integrity percentages were mimicked using artificial double stranded DNA in low, intermediate and high DNA concentration scenarios, respectively 600, 12500 and 30000 copies of DNA per reaction. At low concentrations both methods were similar. However, at the intermediate concentration (12500 copies per reaction) the ratio based method started producing a larger error than the proposed probability calculation method with a mean relative error of 20.7 and 16.7 for the Bruner and the proposed method respectively. At the high concentration (30000 copies per reaction) only the proposed method provided accurate measurements with a mean relative error of 60.9 and 9.3 for the ratio based and the proposed method respectively. Furthermore, while both methods have a bias, it is constant for the proposed method, while it decreases with the integrity of the DNA for the ratio based method. The probability calculation equation was extended to 4 dimensions and a proof of concept experiment was performed, the data suggested that the 4 dimensional equation is valid.

Significance and novelty

We here validate a method of estimating DNA integrity with dPCR using multiple probe combinations, allowing fast and flexible DNA integrity analysis. Additionally, we extend the method from 2 to 4 plex for more accurate DNA integrity measurements

Keywords

dPCR, HIV, DNA integrity, digital PCR.

1. Introduction

Obtaining good quality DNA is important for many biomolecular methods, such as PCR and sequencing. Especially the increasing interest in long-read DNA sequencing technology depends on obtaining long intact stretches of DNA. Hence, good methods to estimate DNA integrity are needed. Up to date, most techniques to assess DNA quality are based on electrophoresis techniques and are at best semiquantitative. Digital PCR (dPCR) is an interesting method to investigate DNA quality. In dPCR the PCR reaction is divided into thousands of partitions using droplets or microwells. Each partition contains either no, a single or at most a few target molecules which can be accurately calculated using the Poisson distribution [1]. The output of the data is considered binary, hence digital, since a partition is either positive or negative. The single molecule resolution of dPCR allows the analysis of DNA integrity.

One method for assessing DNA integrity was developed by *Didelot et al. (2013)* [2]. This method utilizes multiplexing using Evagreen dye chemistry with 4 amplicons that differ in length, i.e. 78, 159, 197 and 550 bp's. If the DNA is highly fragmented only the shortest amplicons are amplified, resulting in partitions with low levels of fluorescence. Samples with intact DNA generate longer amplicons resulting in partitions with higher peak fluorescence. A similar method was used by Kint et al. (2018) [3] to assess DNA quality after bisulfite conversion. The drawback of these amplicon size based methods is that they are limited to fairly short DNA stretches as too long amplicons would not efficiently amplify.

Another elegant approach was independently discovered by Bruner et al. (2019) [4] and Regan et al. (2015) [5]. methods are based on detect the physical connection or dissociation of two amplicons that are located on the same DNA molecule, but are separated from each other by a large interval (>1000bp). These methods take advantage of the fact that DNA molecules will enter dPCR partitions at random. If the DNA molecule is sheared, the two amplicons will no longer be physically linked and will end up in a different partition resulting in two single positive partitions (Figure 1, A). However, if the two amplicons reside on the same, intact, strand of DNA a double positive partition will be registered. Of note, in dPCR it is not only an intact molecule that can cause a partition to become double positive. Since a single partition can harbour multiple molecules, double positive partitions can also be caused by multiple sheared molecules entering the same partition (Figure 1, B). Bruner solved this issue by including a limiting dilution step. If the concentration is low enough, then the co-occurrence of multiple unlinked target amplicons in a partition becomes so low that the error caused by multiple sheared molecules is negligible. Using this approach they can then estimate integrity by a relatively simple ratio based equation (equation 1), using only the ratio of single to double positive partitions [4]. Regan et al. took their solution a step further by developing a statistical probability based approach which theoretically, allows them to use the full dynamic range of dPCR thereby eliminating the need for a limiting dilution step by compensating for the probability of co-occurrence of target sequences in partitions [5].

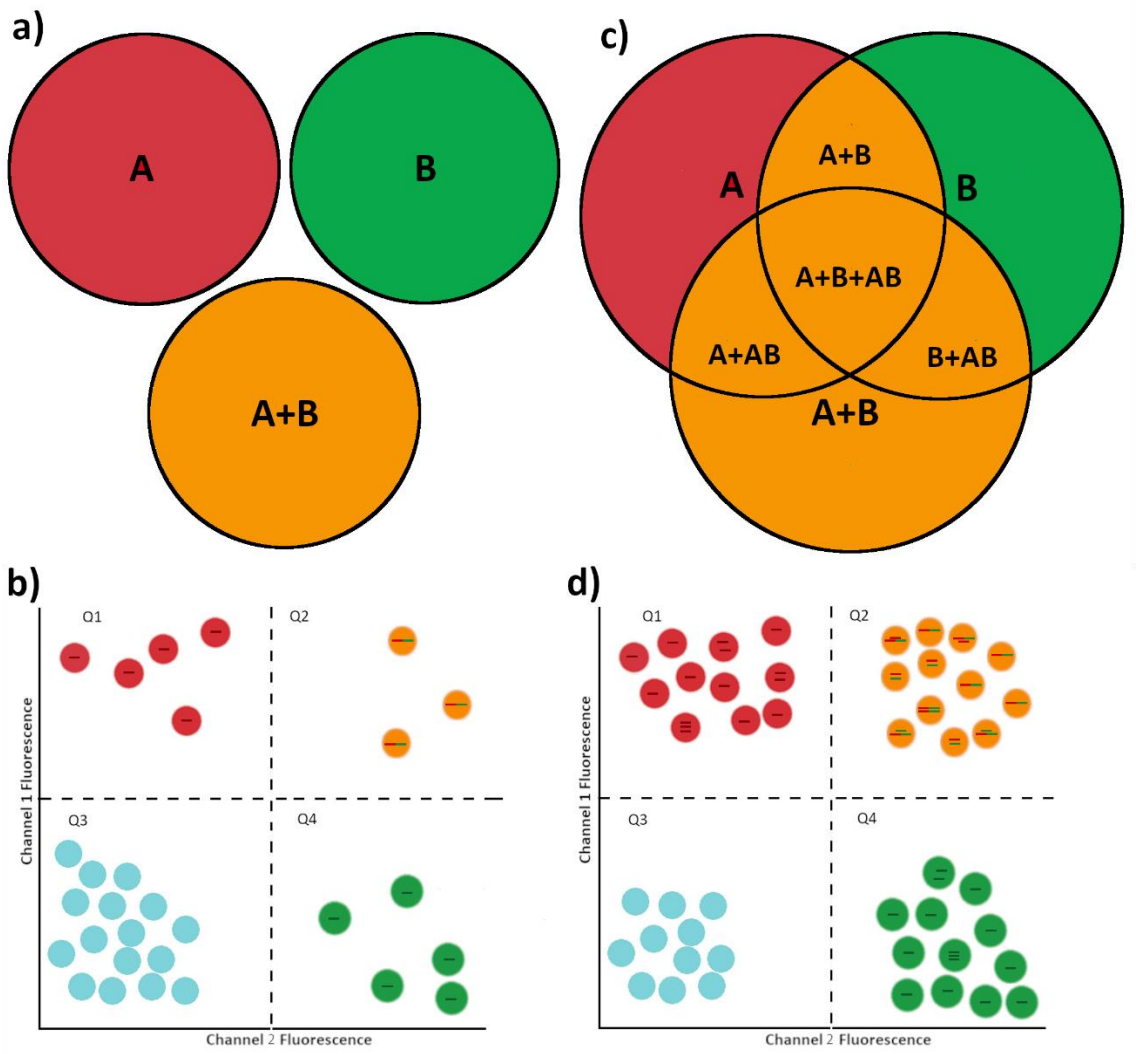


Figure 1, a) a Venn diagram showing a very low DNA copy number scenario of the populations, no overlap between the populations will exist. b) the expected populations from a dPCR experiment in which very low concentrations were used, Q1 and Q4 shows the single positive populations while Q2 shows the double positive, note that only intact DNA molecules exist in Q2. C) Venn diagram depicting a higher DNA copy number scenario, here the populations overlap, resulting in several possible combinations of double positives (orange colour). d) the expected populations for a dPCR experiment with high DNA copy number, notice that the Q2 population contains several combinations of molecules most of which will bias the estimated the DNA integrity.

Here, we validate Regans approach using experimental data. In addition, by using a different approach to derive the probability equation we enable the method to be extended to multiple dimensions. So that more than two amplicons can be used, enabling

a more detailed assessment of DNA integrity [6][7]. Using this method, we extended the equation to 3 and 4 targets and validated these using simulations and a proof of concept experiment with experimentally sheared DNA.

Equation 1

$$DSI_B = \frac{\left(\frac{K_{Q1} + K_{Q4}}{2}\right)}{\left(\frac{K_{Q1} + K_{Q4}}{2} + K_{Q2}\right)}$$

Where K_{Q1}, K_{Q2}, K_{Q3} and K_{Q4} refer to the number of partitions in the 4 quadrants.

2. Materials and methods

2.1 Materials

Primers and probes were used as previously described [4] and [8]. As positive control material, four gBlocks were designed to match the sequences of the primers and probes. Two single target gBlocks [IDT, USA] were based on the ENV or GAG genes from the HIV genome and were used to simulate sheared DNA. Two double-target gBlocks [IDT, USA], i.e., ENV-GAG or GAG-ENV were designed to be used as intact DNA equivalents. All gBlock sequences were initially tested *in silico* for secondary structure formation using mFold [9], complete sequences can be found in supplementary S.3. After initial quantification by dPCR, the gBlocks were diluted to 50,000 cp per µl. From these stock solutions, 3 dilutions of 30,000 12,500 and 600 cp per µl were created of either the double target gBlock or the two single targets. All gBlocks were diluted in HPLC grade water and stored at -20 °C. Aliquots were made and freeze/thaw cycles were kept below 4 as per manufacturer recommendation. Since the double target gBlocks were not 100% intact, an initial integrity percentage of the double target gBlock was calculated by using equation 2 where C_A is the concentration of target A, C_B is the concentration of target B and C_{AB} is the concentration of the intact sequence AB , this integrity value was used as a rough measure of integrity for the subsequent experiments. Finally, the double target and the single target mixes were used to create known integrity standards of 75%, 50% and 25% integrity, resulting in a total of 12 combinations of concentration and integrity percentages. Additionally, we also calculated the results using both the Bruner and the proposed methods, to test that the initial estimation did not affect the conclusion, see table S.1 for results.

Equation 2

$$Integrity = 1 - \frac{\left(\frac{C_A + C_B}{2}\right)}{\left(\frac{C_A + C_B}{2}\right) + C_{AB}}$$

2.2 Sonication.

HIV DNA extracted from the clonal Jurkat cell line 8.4 (J-Lat 8.4) was received from the UZ Ghent HIV cure research center, description of the original cell line can be found in [10]. The sample DNA was diluted 1:30 and aliquoted into tubes containing 200 µl each. To create varying degrees of DNA integrity the samples were then subjected to sonication with a Misonix sonicator ultrasonic processor XL [Bioventus, USA] for either: 0s, 10s, 20s, 30s, 40s, 50s or 60s.

2.3 dPCR

gBlocks

The 12 gBlock dilutions were run in triplicate on the Naica 3 color dPCR [Stilla technologies, France]. With the following PCR conditions: 10 min at 95 °C,40 cycles at 60 °C for 30 seconds and 95 °C for 15 seconds followed by 10 cycles at 55 °C for 30 seconds and 95 °C for 15 seconds to increase separation of the positive partitions as described previously [11]. All ddPCR runs were done using perfecta-multiplex-qPCR-Toughmix (Quanta Biosciences, USA, cat#: 95147-250). 4 Non Template Controls(NTC's) were included in the setup to serve as quality control. Furthermore, A spillover compensation matrix was made using the *Stilla Crystal Miner software version 2.4.0.3* and applied for both assays. Subsequently, the raw data was exported as .csv files for further analysis using a custom R-script (supplemental data). The data were baselined using the median of the negative populations and

a common threshold (5000 RFU for the red channel and 1000 RFU for the green channel) for all samples was manually applied. Due to suboptimal separation in the green channel (figure S.3) a separate threshold had to be applied for the NTC's. All wells with less than 10000 partitions were excluded from the data analysis. To comply with the dMIQE guidelines further technical information regarding the dPCR can be found in section S.3 in supplementary.

Sonicated DNA

Each aliquot of sonicated DNA were run on the Naica 6 color dPCR [Stilla technologies, France] in triplicates. A total of 4 target genes were included: PSI, GAG, POL and RU5. With the PCR conditions as follows: 10 min at 95 °C followed by 40 cycles of 15 sec at 95 °C then 30s at 65 °C. NTC's were included in every run and all wells with less than 10000 partitions were excluded from the data analysis. All runs with sonicated DNA was performed using the Naica multiplex PCR mix 10X [Stilla technologies, France, Ref#: R10104]. Spillover compensation was performed as described above. A total of 2 runs were performed. For the first run, the automatic thresholding from Crystal miner was quite good and was kept as is. For the second run a few thresholds proved problematic and a manual threshold was applied at the following RFU's: Blue=17866, Green=2822, Red=6458 and yellow=3999. To comply with the dMIQE guidelines further technical information regarding the dPCR can be found in section S.3 in supplementary.

2.4 Statistical analysis

In dPCR experiments, the entry of molecules into a partition is random. Therefore, in a duplex reaction, 4 populations will be generated, see figure 1, for target A (K_{Q1}), 1 for target B (K_{Q4}), one negative (K_{Q3}) and one containing both A and B – either as an intact molecule or as 2 singles in one partition (K_{Q2}). Using the probability that a molecule will enter a partition we can estimate the 3 positive populations as shown in equation 3.

Equation 3

$$K_A = \frac{\boldsymbol{N} * K_{Q1}}{(K_{Q1} + K_{Q3})}$$

$$K_B = \frac{\boldsymbol{N} * K_{Q4}}{(K_{Q4} + K_{Q3})}$$

$$\boldsymbol{K_{AB}} = \frac{K_{Q2} * K_{Q3} - K_{Q1} * K_{Q4}}{K_{Q3}}$$

To calculate the integrity we first need to calculate the number partitions that contain only broken A, broken B and intact fragments AB: x_A, x_B and x_{AB} respectively. And N is the total number of partitions. We then apply Poisson estimation to calculate the actual amounts of molecules in those partitions,

Equation 4

$$\lambda_A = -\log\left(1 - \frac{K_{Q1}}{(K_{Q1} + K_{Q3})}\right)$$

$$\lambda_B = -\log\left(1 - \frac{K_{Q4}}{(K_{Q4} + K_{Q3})}\right)$$

$$\lambda_{AB} = -\log\left(1 - \frac{K_{Q2} * K_{Q3} - K_{Q1} * K_{Q4}}{\boldsymbol{N}K_{Q3}}\right)$$

Where λ_A, λ_B and λ_{AB} are the average number of broken A, broken B and intact fragments per partition respectively. Integrity is estimated accordingly,

Equation 5

$$\text{Integrity} = 1 - \left(\frac{\left(\frac{\lambda_A + \lambda_B}{2}\right)}{\left(\frac{\lambda_A + \lambda_B}{2} + \lambda_{AB}\right)}\right)$$

This can be easily extended to higher dimensions using probability calculation, which enables more accurate integrity calculation. We give a 3-dimension example. The proportion of partitions containing broken A, B and C can be estimated as,

$$P(A) = \frac{K_{Q1}}{K_{Q1} + K_{Q3}}$$

$$P(B) = \frac{K_{Q4}}{K_{Q4} + K_{Q3}}$$

$$P(C) = \frac{K_{Q5}}{K_{Q5} + K_{Q3}}$$

where K_{Q1} , K_{Q3}, K_{Q4}, k_{Q5} are the number of partitions that contain only broken A, no molecules, broken B and C molecules respectively. Subsequently, we derive $\boldsymbol{P_{AB}}$, $\boldsymbol{P_{AC}}$, $\boldsymbol{P_{BC}}$, $\boldsymbol{P_{ABC}}$ (for more details, please see the Appendix S.1). We developed an easy R function to automate the calculation.

We also give a 95% confidence interval of the estimates, for more details, please refer to Chen et al [12] and Appendix S.1. The confidence interval takes the variation of the estimates and is the range of values we can expect at a certain level of confidence if the experiment is redone.

The performance of this method was tested in different scenarios with varying concentrations and degrees of integrity. We then compared the results of the proposed method with the method of Bruner to the theoretical integrity percentage using 2 linear models, and the mean relative error. All data analysis was done in R (Version 4.2.0) [13] codes can be found at:

https://github.com/emmachenlingo/dpcr-flexible-methods-for-standard-error-calculation/blob/digital-PCR/dsi_functions.R.

3. Results and discussion

DNA integrity comparison.

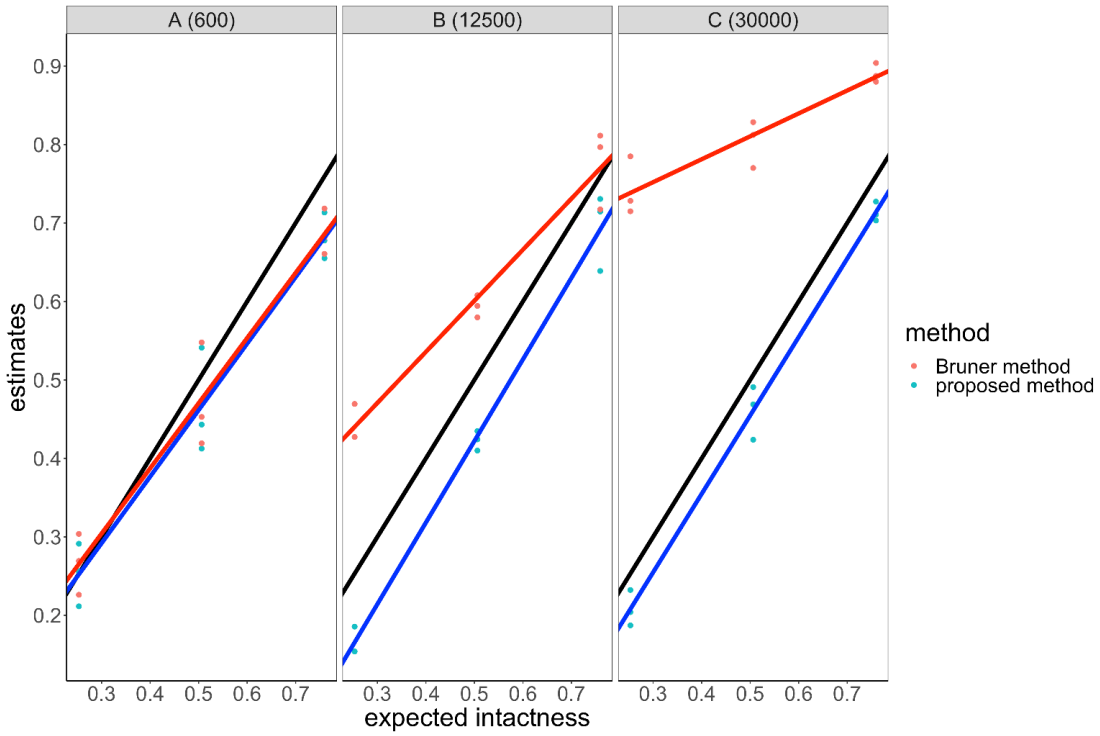


Figure 2, The fitted models on the expected and the estimates by the two methods under A, B and C scenarios. The x-axis represents the expected (theoretical) integrity percentage, and y-axis is the estimates given by the proposed (the blue dots)) and the Bruner method (the red dots). The black line shows the reference values calculated by the simple method, the blue line is the fitted model based on the DNA integrity values measured by the proposed method and the red line is the model based on the Bruner method. In panel A where the total DNA copies is low, the two methods give similar estimates, whereas in panel C, Bruner's method gives much higher DNA integrity estimates than the proposed one. All integrity values are shown in percentage.

The proposed method and the method by *Bruner et. Al.* (2019) was used to calculate the integrity percentage. We used MRE (mean relative error) to measure the deviation from the expected values. For the low copy number scenario (600 cp per reaction), a strong fit was observed between both regression lines and the reference line. Furthermore, the MRE (mean relative error) were almost identical for both methods, 14.4% vs. 13.7% for the proposed and Bruner methods respectively, that means on average 14.4% off the expected value by the proposed method and 13.7% off by Bruner method, as shown in Table 2.

At this low copy number (scenario A), the λ value is sufficiently low (median λ = 0.022), which minimizes the chance of multiple molecules per partition. Hence, utilizing the probabilities of double positive events holds no advantage when compared with the Bruner method. At higher copy numbers (scenario B), the two methods start to differ. As can be seen in Figure 2 the proposed method has a closer fit with the expected line, this becomes more evident when examining the MRE (16.7% vs. 20.7% for the proposed and Bruner methods) as shown in table 2. The data in Table 2 shows that at the copy numbers in scenario B and C, the Bruner method overestimates the expected DNA integrity. At these copy numbers the averaged Poisson probability that a single partition would harbour 2 or more molecules is 0.06086 and 0.272 for scenario B and C respectively. The difference between the two methods is, however, not very prominent in scenario B. For scenario C, the difference between the linear models in figure 2 becomes very clear, and the overestimation of the Bruner method becomes problematic. At these higher copy numbers the λ value (median λ = 1.019) has increased to such an amount that the Bruner method is no longer applicable which is shown by the huge MRE value (60%). The proposed method underestimates as compared to the expected values, but the strength of relying on the probabilities of double positive events is shown, as the MRE are essentially the same as for scenario A (scenario A= 14.4 vs scenario C= 9.29). It should also be noted that the bias of the proposed method is constant, while that of the Bruner method is non-constant – decreasing with the integrity of the DNA. Indicating that the bias on the estimate will be higher in low integrity samples. Of note, when investigating the bias between the estimated integrity and the expected, we should take into account that the expected integrity may be biased on its own. A bias can be introduced by pipetting errors. In addition, since the original gBlocks were not 100% intact, we derived the expected integrity using a simple ratio based calculation (equation 2). This method is prone to the same error as the Bruner method.

Table 1, Table containing the MRE and DNA integrity values of the proposed and Bruner method while using the simple method as estimates for DNA Integrity as expected values. 50% DNA integrity is shown as a representative measurement.

Scenario	A	B	C
Total copies of DNA	600	12500	30000
Integrity Proposed	46.555 %	42.3 %	46.116 %
MRE proposed	14.443	16.743	9.294
Integrity Bruner	47.337 %	59.4 %	80.365 %
MRE Bruner	13.748	20.698	60.877

To also provide the level of uncertainty on the Integrity percentages, we developed a method for calculating the confidence interval of the Integrity estimate based on a bootstrap method[12], (figure 3). The classical method to construct a confidence interval for dPCR data is to use delta method[14]. However, this method makes assumption of binomial distribution of the number of positive partitions and also independence of the ratio variables. For the integrity, the amount of sheared and total DNA molecules are not independent. Also with the existence of experimental errors such as pipette error and misclassification, the binomial assumption may not be valid [7, 10]. The NonPVar method in *Chen et al*[12] makes use of replicates and is robust against such errors because it does rely on the distributional assumption of the number of positive partitions.

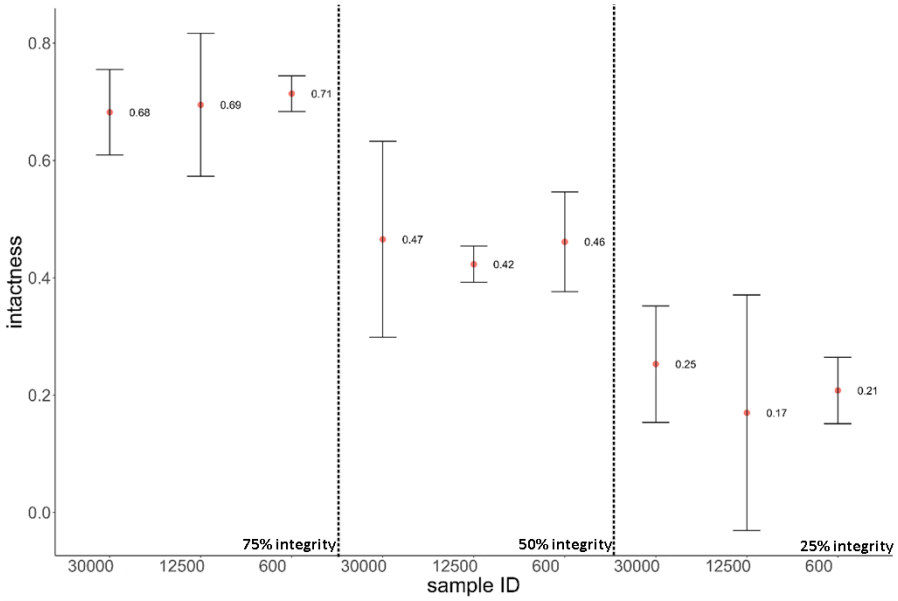


Figure 3, 95% Confidence interval of the estimated DNA integrity by the proposed method. X-axis represents the different combinations of copy number and integrity percentages. The red dot and texts beside them indicate the estimated integrity by the proposed method. For each combination, we have 3 replicates.

As proof of concept for the 4 dimension DNA integrity equation an experiment was performed using sonicated DNA (Figure 4). For the first 3 samples (T = 0 – 20 sec sonication) the measurements behave as expected, at 0 sec the DNA is between 53% and 58% intact these values then drop with higher sonication time in a trend that resembles a logarithmic decline, which would be expected.

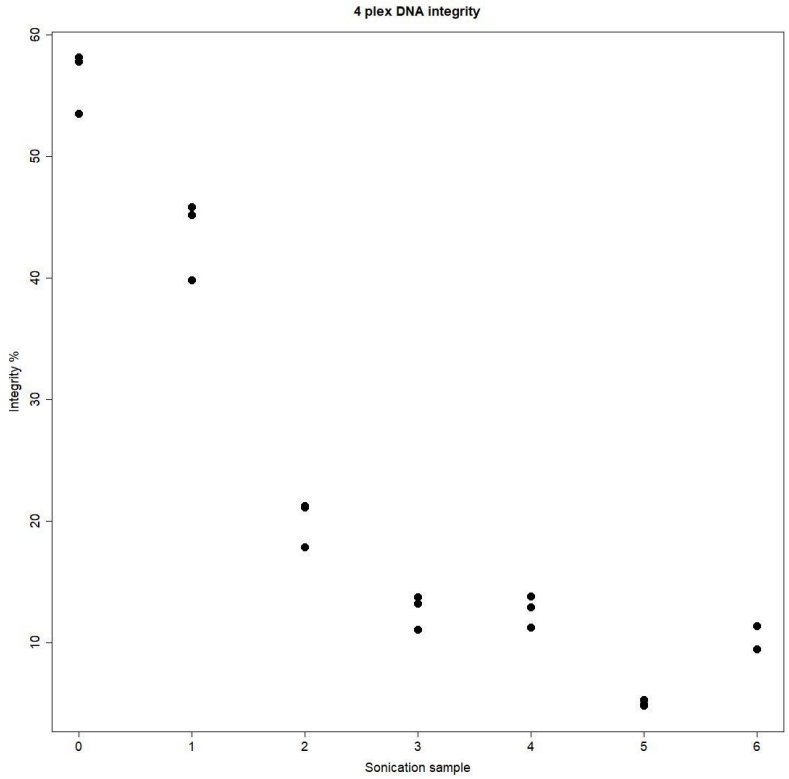


Figure 4, shows the results of the sonication experiment, x-axis contains the various sonication samples (0 = 0 sec, 1 = 10 sec, 2 = 20 sec etc.). Y-axis shows the measured integrity percentage.

4. Conclusion

At low concentrations both methods perform well, but at higher concentrations the Bruner method starts to overestimate the Integrity of the DNA due to the increasing appearance of partitions harbouring more than one molecule. The proposed method maintains an acceptable performance, even at higher DNA copy numbers and although it slightly underestimates the DNA integrity based on the tested reference material, this bias is constant. Furthermore, the sonication experiment shows that the extension of the probability equation into multiple dimensions is possible and might also indicate a lower limit of DNA integrity for the calculations. Additionally, a freely available web-based tool for calculating DNA integrity using both 2, 3 and 4 targets can be found on https://dpcr-ugent.shinyapps.io/intactness_assay_dsi/.

Acknowledgements

This work was financially supported by an interdisciplinary grant of the special research fund of Ghent University (01IO0420).

References

[1] Vogelstein, B.; Kinzler, K. W. Digital PCR. Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 9236.

[2] Didelot, A.; Didelot, S.K.; Kotsopoulos, A.; Lupo, D.; Pekin, X.; Li, I.; Atochin, P.; Srinivasan, Q.; Zhong, J.; Olson, D.R.; Link, P.; Laurent-Puig, P; Blons, H; Hutchison, J.B.; Taly, V.; Multiplex picoliter-droplet digital PCR for quantitative assessment of DNA integrity in clinical samples. Clin. Chem. 2013, 59, 815–823.

[3] Kint, S; De Spiegelaere, W; De Kesel, J; Vandekerckhove, L; Van Criekinge, W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. PloS One. 2018 Jun 14;13(6):e0199091. Doi: 10.1371/journal.pone.0199091.

[4] Bruner, K.M.; Wang, Z.; Simonetti, F:R.; Bender, A.M.; Kwon, K.J.; Sengupta, S.; Fray, E.J.; Beg S.A.; Antar, A.A.R.; Jenike, K.M.; Bertagnolli, L.N.; Capoferri, A.A.; Kufera, J.T.; Timmons, A.; Nobles, C.; Gregg, J.; Wada, N.; Ho, Y.C.; Zhang, H.; Margolick, J.B.; Blankson, J.N.; Deeks, S.G.; Bushman, F.D.; Siliciano, J.D.; Laird, G.M.; Siliciano, R.F. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. Nature 2019, 566, 120.

[5] Regan, JF; Kamitaki, N; Legler, T; Cooper, S; Klitgord, N; Karlin-Neumann, G; Wong, C; Hodges, S; Koehler, R; Tzonev, S; McCarroll, SA. A rapid molecular approach for chromosomal phasing. PloS One. 2015 Mar 4;10(3):e0118270. Doi: 10.1371/journal.pone.0118270.

[6] Gaebler, C; Falcinelli, SD; Stoffel, E; Read, J; Murtagh, R; Oliveira, TY; Ramos, V; Lorenzi, JCC; Kirchherr, J; James, KS; Allard, B; Baker, C; Kuruc, JD; Caskey, M; Archin, NM; Siliciano, RF; Margolis, DM; Nussenzweig, MC. Sequence Evaluation and Comparative Analysis of Novel Assays for Intact Proviral HIV-1 DNA. J Virol. 2021 Feb 24;95(6):e01986-20. Doi: 10.1128/JVI.01986-20.

[7] van Snippenberg, W; Gleerup, D; Rutsaert, S; Vandekerckhove, L; De Spiegelaere, W; Trypsteen, W. Triplex digital PCR assays for the quantification of intact proviral HIV-1 DNA. Methods. 2022 May;201:41-48. Doi: 10.1016/j.ymeth.2021.05.006. Epub 2021 May 13.

[8] Palmer, S; Wiegand, AP; Maldarelli, F; Bazmi, H; Mican, JM; Polis, M; Dewar, RL; Planta, A; Liu, S; Metcalf, JA; Mellors, JW; Coffin, JM. New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. J Clin Microbiol. 2003 Oct;41(10):4531-6. Doi: 10.1128/JCM.41.10.4531-4536.2003.

[9] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 2003, 31, 3406.

[10] Jordan, A; Bisgrove, D and Verdin, A. "HIV Reproducibly Establishes a Latent Infection after Acute Infection of T Cells *in Vitro*." EMBO J. 22 (2003): 1868-1877.

[11] De Spiegelaere, W; Malatinkova, E; Kiselinova, M; Bonczkowski, P; Verhofstede, C; Vogelaers, D; Vandekerckhove, L; Touchdown digital polymerase chain reaction for quantification of highly conserved sequences in the HIV-1 genome. Anal. Biochem. 2013, 439, 201–203.

[12] Chen, Y; De Spiegelaere, W; Trypsteen, W; Gleerup, D; Vandesompele, J; Thas, O; Flexible Methods for Standard Error Calculation in digital PCR Experiments. Submitted.

[13] R Core Team (2022) R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. URL https://www.R-Project.org/.

[14] Whale, A.;, Huggett, J.; Cowen, S.; Speirs, V.; Shaw, J.; Ellison, S.; Scott, D. Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. Nucleic acids research. 2012, 40(11), 82.

[15] Trypsteen W, Vynck M, De Neve J, Bonczkowski P, Kiselinova M, Malatinkova E, Vervisch K, Thas O, Vandekerckhove L, De Spiegelaere W. ddpcRquant: threshold determination for single channel droplet digital PCR experiments. Anal Bioanal Chem. 2015 Jul;407(19):5827-34.

Supporting information

S.1

The probability of $k_{(Q1)}$ can be seen as with single positive partitions of broken A, no positives of broken B and no positives of intact AB,

$$\frac{k_{Q1}}{N} = P(A) * \big(1 - P(\mathrm{B})\big) * \big(1 - P(\mathrm{AB})\big) \qquad \text{(Equation S1)}$$

The probability of k_{Q3} can be seen as with no single positive partitions of broken A, no positives of broken B and no positives of intact AB,

$$\frac{k_{Q3}}{N} = (1 - P(A)) * \big(1 - P(\mathrm{B})\big) * \big(1 - P(\mathrm{AB})\big) \qquad \text{(Equation S2)}$$

The probability of k_{Q4} can be seen as with no single positive partitions of broken A, with positives of broken B and no positives of intact AB,

$$\frac{k_{Q4}}{N} = (1 - P(A)) * P(\mathrm{B}) * \big(1 - P(\mathrm{AB})\big) \qquad \text{(Equation S3)}$$

k_{Q1}, k_{Q2}, k_{Q3} and k_{Q4} are all known to us, $P(A)$, $P(B)$ and $P(AB)$ are the proportion of partitions containing broken A (including those in double positives k_{Q2}), broken B and intact fragments AB respectively. We derive $P(A)$ from dividing Equation S1 by Equation S2.

$$\frac{k_{Q1}}{k_{Q3}} = \frac{P(A)}{1 - P(A)}$$

$$P(A) = \frac{k_{Q1}}{(k_{Q1} + k_{Q3})}$$

We can derive P(A) and P(B) the same way from dividing Equation S3 by Equation S2,

$$P(A) = \frac{k_{Q1}}{(k_{Q1}+k_{Q3})}$$

$$P(B) = \frac{K_{Q4}}{(K_{Q4}+K_{Q3})}$$

Since we have $P(A)$ and $P(B)$, we can estimate $P(AB)$ in Equation S1,

$$\frac{K_{Q1}}{N}=\frac{K_{Q1}}{(K_{Q1}+K_{Q3})}*(1-\frac{K_{Q4}}{(K_{Q4}+K_{Q3})})(1-P(AB))$$

$$\frac{K_{Q1}}{N}=\frac{K_{Q1}}{(K_{Q1}+K_{Q3})}*\frac{K_{Q3}}{(K_{Q4}+K_{Q3})}(1-P(AB))$$

$$\frac{(K_{Q1}+K_{Q3})*(K_{Q4}+K_{Q3})}{NK_{Q3}}=1-P(AB)$$

$$P(AB)=1-\frac{(K_{Q1}+K_{Q3})*(K_{Q4}+K_{Q3})}{NK_{Q3}}$$

$$P(AB)=\frac{NK_{Q3}-(K_{Q1}+K_{Q3})*(K_{Q4}+K_{Q3})}{NK_{Q3}}$$

$$P(AB)=\frac{(K_{Q1}+K_{Q2}+K_{Q3}+K_{Q4})*K_{Q3}-(K_{Q1}+K_{Q3})*(K_{Q4}+K_{Q3})}{QQ3}$$

$$P(AB)=\frac{K_{Q2}*K_{Q3}-K_{Q1}*K_{Q4}}{NK_{Q3}}$$

With estimated proportions of positive partitions for broken A, broken B and intact fragments, we further proceed with Poisson estimation.

This can be easily extended to multiple dimensions. We will first start with a 3-dimension example.

$\frac{K_{Q1}}{N} = P(A) * (1-P(B)) * (1-P(C)) * (1-P(AB))*(1-P(BC)) * (1-P(ABC))$ (Equation S4)

$\frac{K_{Q2}}{N} = (P(A) * P(B) + P(AB) - P(A) * P(B) * P(AB)) * (1-P(C))*(1-P(BC)) * (1-P(ABC))$ (Equation S7)

$\frac{K_{Q3}}{N} = (1-P(A)) * (1-P(B)) * (1-P(C)) * (1-P(AB))*(1-P(BC)) * (1-P(ABC))$ (Equation S6)

$\frac{K_{Q4}}{N} = (1-P(A)) * P(B) * (1-P(C)) * (1-P(AB))*(1-P(BC)) * (1-P(ABC))$ (Equation S5)

$\frac{K_{Q5}}{N} = (1-P(A)) * (1-P(B)) * P(C) * (1-P(AB))*(1-P(BC)) * (1-P(ABC))$ (Equation S8)

$\frac{K_{Q6}}{N} = (P(x_A) * P(x_C))*(1-P(AB)) * (1-P(BC)) * (1-P(ABC))$ (Equation S9)

$\frac{K_{Q7}}{N} = (P(B) * P(C) + P(BC) - P(BC) * P(B) * P(C))*(1-P(AB)) * (1-P(ABC))$ (Equation S10)

where k_{Q1}, k_{Q2}, k_{Q3}, k_{Q4}, k_{Q5}, k_{Q6}, k_{Q7} are the number of partitions that contain only broken A, B, no molecules, AB, C and BC (AC does not exist). We can infer $P(A)$ from dividing Equation S4 by Equation S6. The same applies to $P(B)$ and $P(C)$.

$$P(A) = \frac{K_{Q1}}{K_{Q1}+K_{Q3}}$$

$$P(B) = \frac{K_{Q4}}{K_{Q4}+K_{Q3}}$$

$$P(C) = \frac{K_{Q5}}{K_{Q5}+K_{Q3}}$$

Then we derive $P(AB)$ as,

$$P(AB) = \frac{k_{Q2}S2 - K_{Q3}S1}{K_{Q3}+K_{Q2}S2-K_{Q3}S1}$$

where $S1 = P(A) * P(B)$, and $S2 = \big(1 - P(A)\big) * \big(1 - P(B)\big)$.

The same applies to $P(B)$.

As the combinations increase considerably with the number of dimensions, we wrote a R function to do the calculation automatically, which is also embedded in our R shiny app.

S.2

We also did a simulation study to evaluate the accuracy of the estimation. The simulation set-ups are varied by the total number of molecules from 100 to 10000, and the integrity percentage from 25% to 75%. Altogether, we have 9 scenarios. The total number of partitions is fixed at 20000. The result is shown in Fig. S.1, S.2, S.3, S.4. The estimates given by the proposed methods are quite close to the true integrity in all scenarios. We also compared the proposed method and Bruner's method in 2-dimension scenarios. The result shows that when the concentration is high (10000 molecules) and the intactness percentage is low, Bruner's method can give a very wrong estimate.

The simulation starts with random assignment of molecules (including broken and intact ones) into partitions. The Qs are found by the intersections, unions and non-overlapping of the partitions. In the simulation study, we know the integrity percentage in advance. The codes for simulation can be found at https://github.com/emmachenlingo/dpcr-flexible-methods-for-standard-error-calculation/blob/digital-PCR/simulation%20study.R .

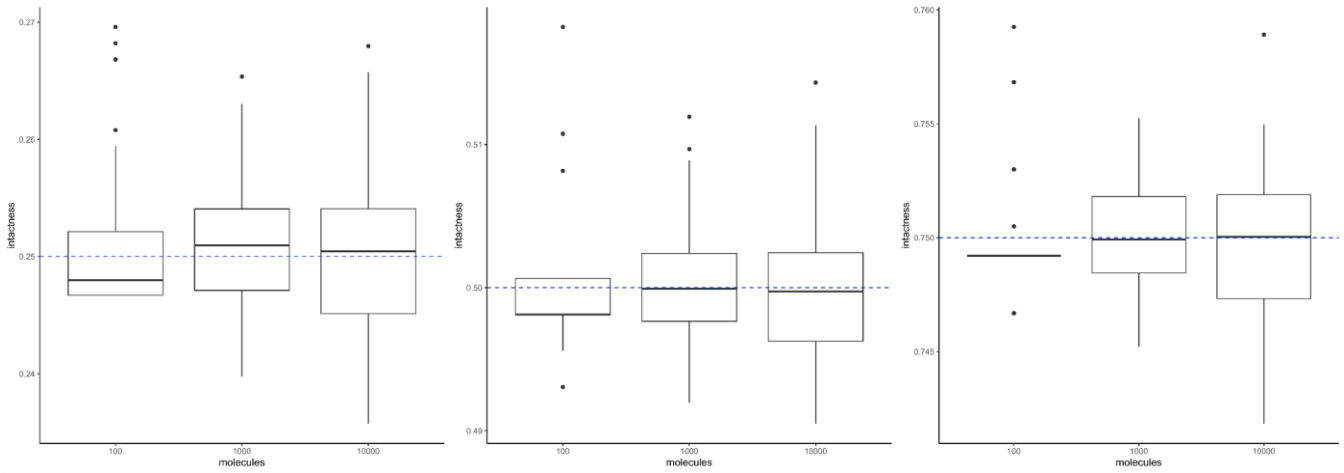


Figure S.1, The estimated integrity at different concentration levels for 2 targets using the proposed method. The dashed blue lines represent the true integrity. The x-axis is the total number of molecules in a sample, and y-axis is the estimated integrity percentage.

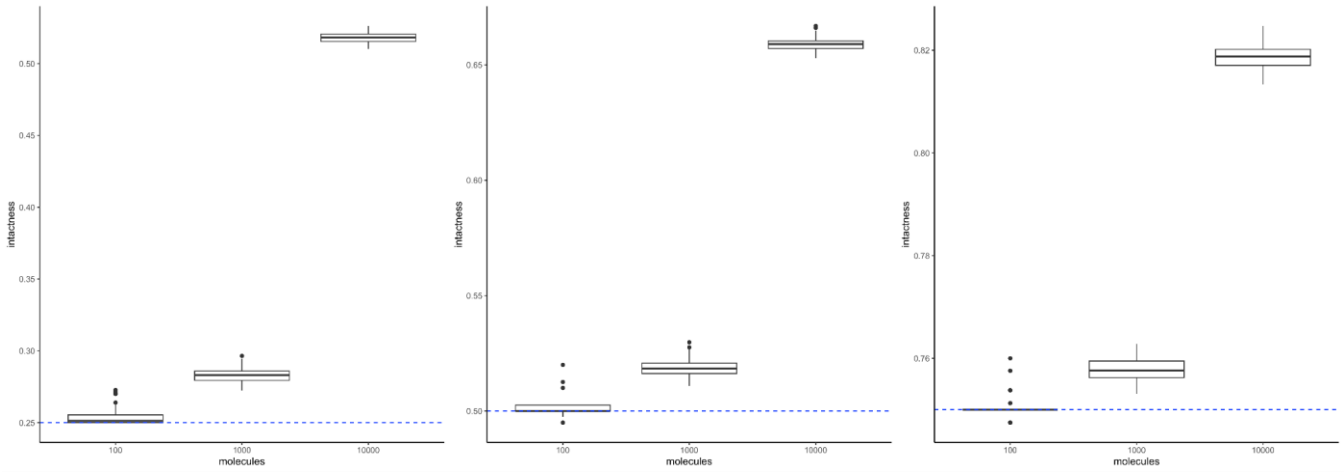


Figure S.2, The estimated integrity at different concentration levels for 2 targets using Bruner's method. The dashed blue lines represent the true integrity. The x-axis is the total number of molecules in a sample, and y-axis is the estimated integrity percentage.

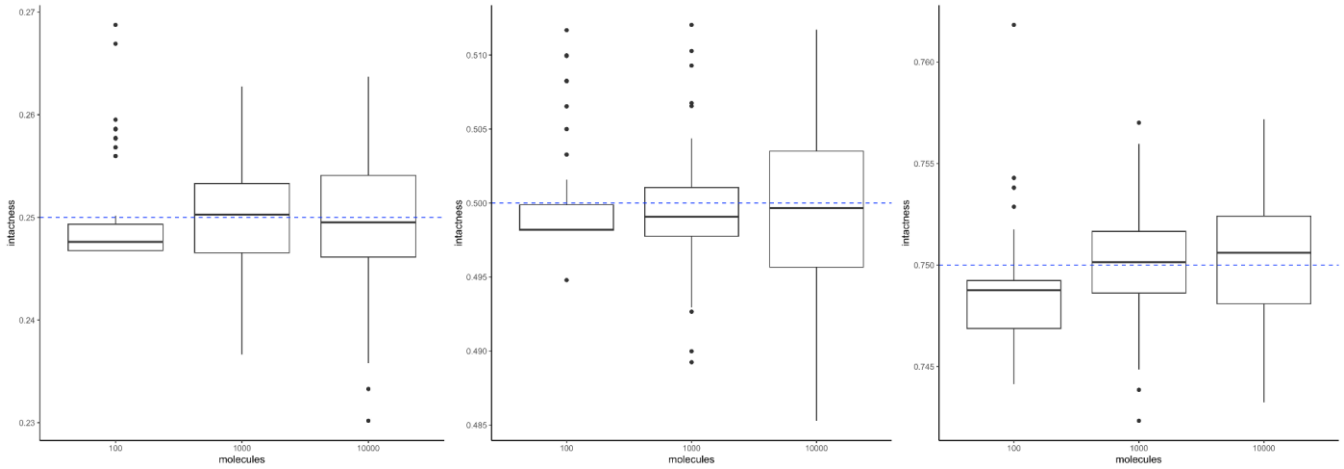


Figure S.3, The estimated integrity at different concentration levels for 3 targets using the proposed method. The dashed blue lines represent the true integrity. The x-axis is the total number of molecules in a sample, and y-axis is the estimated integrity percentage.

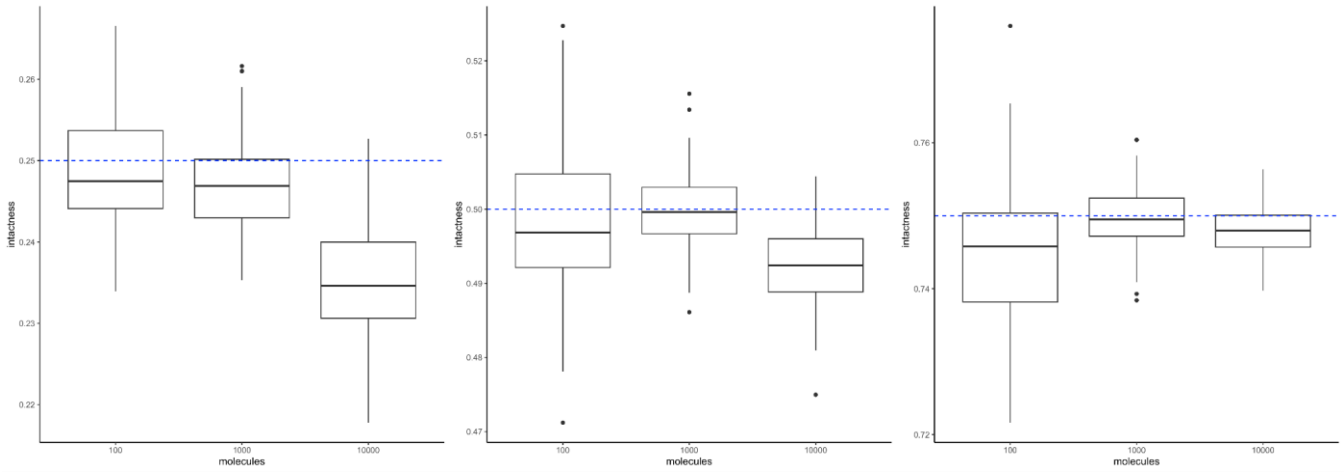


Figure S.4, The estimated integrity at different concentration levels for 4 targets using the proposed method. The dashed blue lines represent the true integrity. The x-axis is the total number of molecules in a sample, and y-axis is the estimated integrity percentage.

S.2 MRE values for the different starting scenarios

Table S.1, Table containing the MRE values of the proposed and Bruner method based on three different initial calculations of integrity.

Scenario	Total copies of DNA	MRE proposed	MRE Bruner
A Simple	600	14.443	13.748
A Bruner	600	66.928	66.969
A Proposed	600	64.115	64.153
B Simple	12,500	16.743	20.698
B Bruner	12,500	24.323	18.653
B Proposed	12,500	19.191	20.022
C Simple	30000	9.294	60.877
C Bruner	30000	15.755	58.455
C Proposed	30000	11.383	60.094

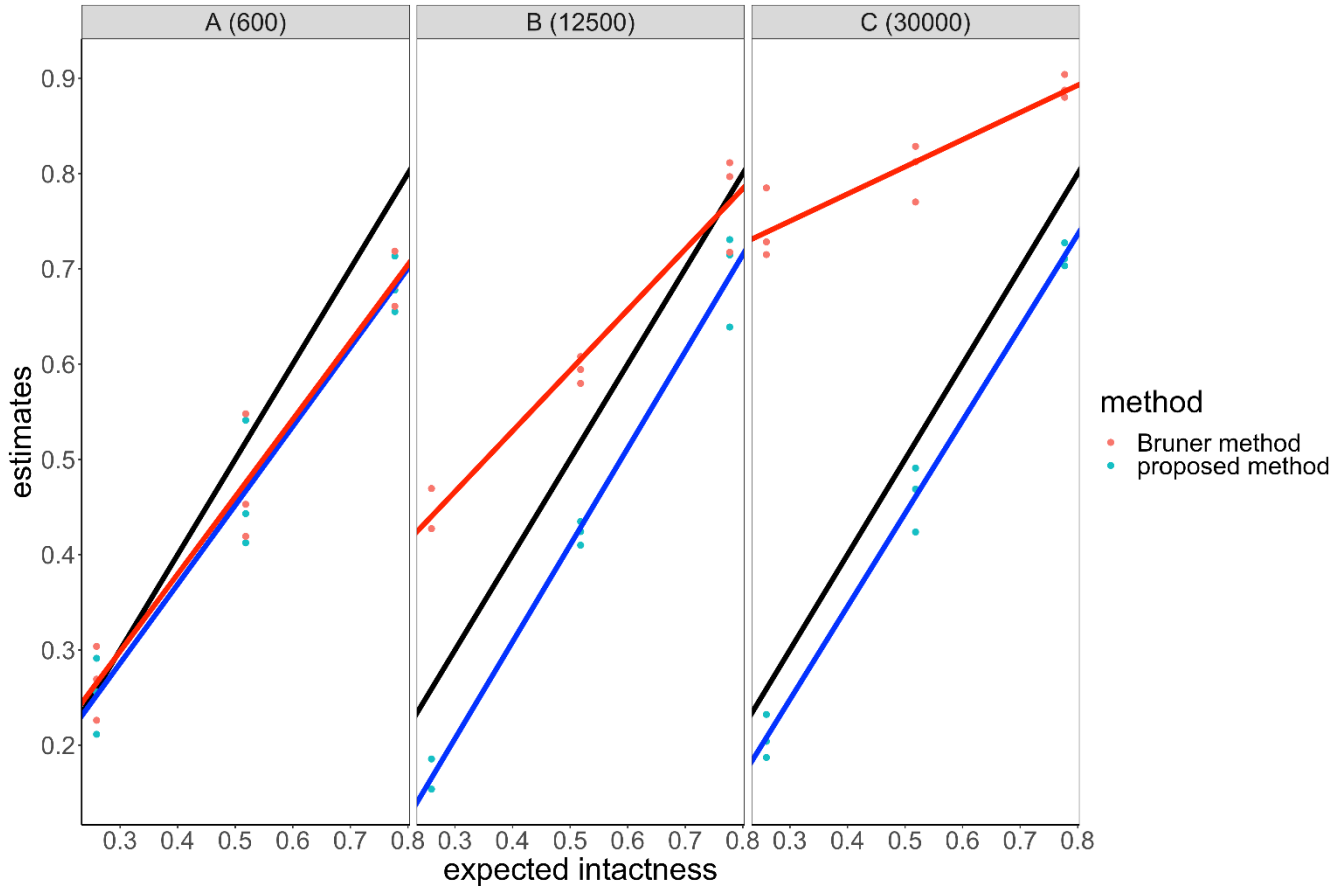


Figure S.5, regression lines if Bruner method was used for initial calculations

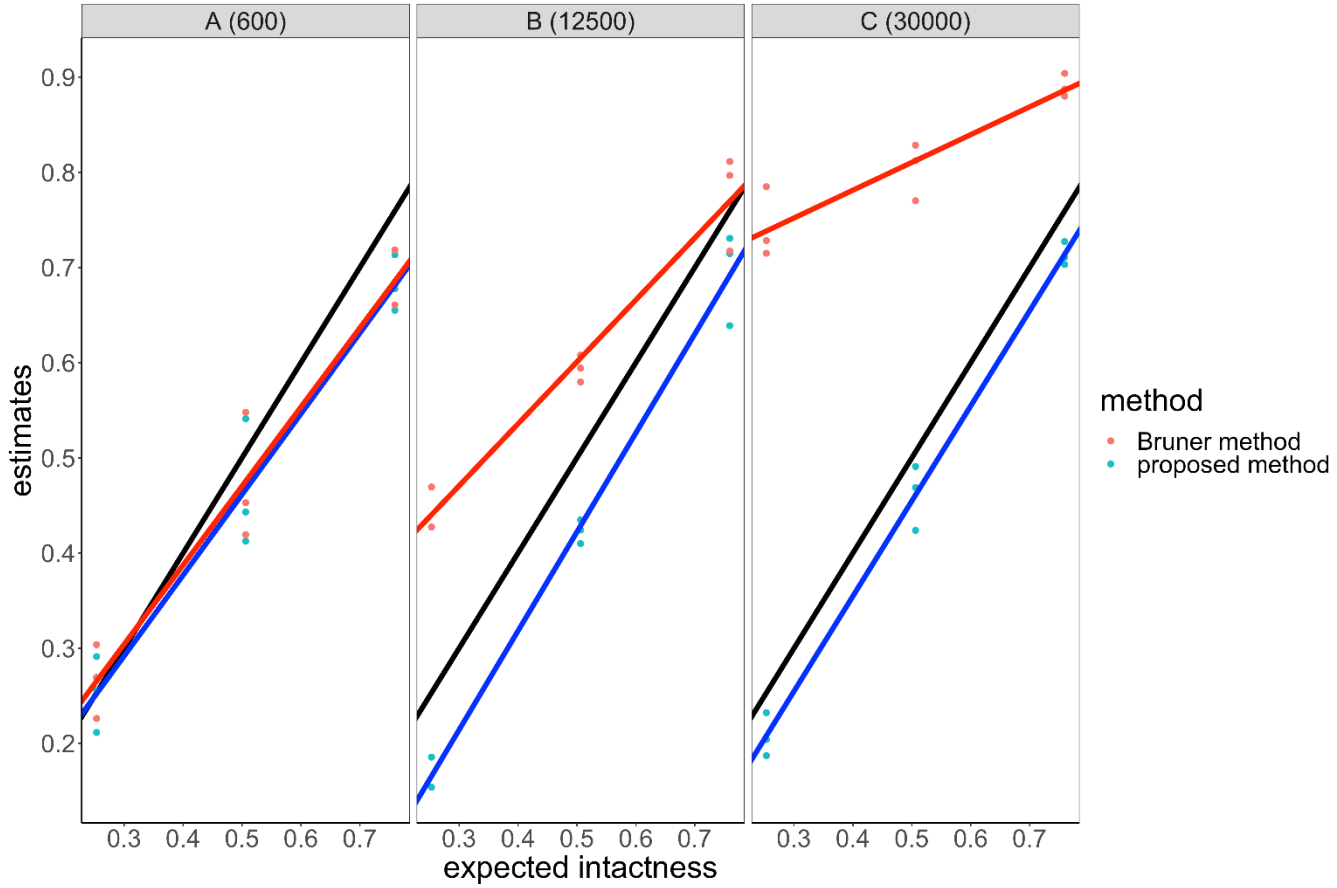


Figure S.6 Regression lines if Proposed method was used for initial calculations

S.3 DMIQE GUIDELINES

ddPCR thresholding details

In order to process the data and apply the *integrity* calculations method the data was first baselined. Two different methods were attempted, the ddpcRquant method [15], in which the half sample mode of the negative population is used for baselining. However, the lack of variation (extreme values) in the red channel meant that this method resulted in the automated threshold being placed inside the negative population. The second method used the manual threshold and the median of the negative population rather than the half sample mode for baselining, this was deemed acceptable. The NTC's were then merged using R and thresholding was applied as described in methods. The resulting plot is shown in figure 7. For the green channel 2 positive partitions can be seen. However, since both positive partitions originated in 1 well, and the red channel have 0 positives these were judged as being false positives rather than contamination. Furthermore, 2 false positives would likely have a negligible effect on the results of the DNA shearing index.

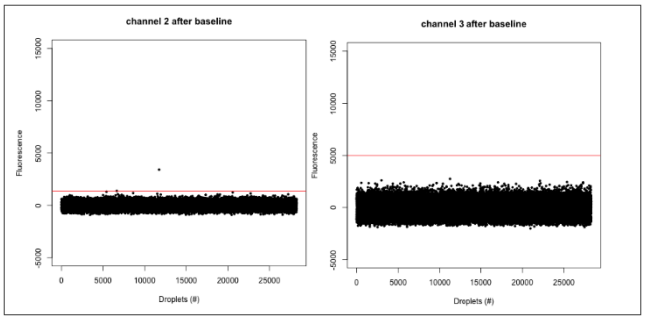


Figure S.7, The merged NTC's after baselining was performed. Channel 2 shows the ENV assay and channel 3 shows the GAG assay. Two positive droplets can be seen for channel 2, due to the absence of positives in channel 3 these were deemed to be false positives rather than contamination.

DPCR TECHNICAL INFORMATION

To be in accordance with the dMIQE guidelines the average partition count along with the IQR was calculated. The median partition count across all reactions was 13781 with an IQR of 1376.75. Furthermore, the following information is included: The average partition volume is 0.20 nL according to the manufacturer. Figure S.8 shows an example of positive results. Since gBlocks was used as the target DNA no negative results was obtained (besides the NTC's). The λ values for the three scenarios can be seen in table S.2

Table S.2, The average lambdas with standard deviation for the 3 scenarios

Scenario	λ	STD
A (600)	0.0202	0.00291
B (12500)	0.406	0.0483
C (30000)	1.020	0.0591

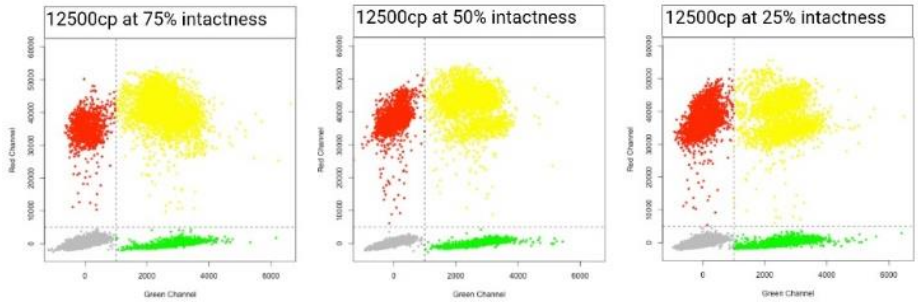


Figure S.8, example of dPCR data results, showing the scatter plots for 12500 copies at 75%, 50% and 25% integrity.

DILUTION OF GBLOCKS.

The gBlocks were initially resuspended to 10 ng per µl in HPLC grade water as per manufacturers (IDT, USA) recommendations. This concentration were then converted to copies per reaction (also described by manufacturer).

All gBlocks were diluted following the same procedure, first diluted using 1:100 steps (1µL gBlock: 99 µL water) until a concentration close to 30000 copies but still far enough above it that corrections could be made before the final steps. For GAG-ENV this was 42420 copies per µl. At this step an aliquot of the dilution was measured using the Naica dPCR (Stilla, France). The measured concentration was 35350 copies per µl. This was then diluted to 30000 copies per reaction (highest concentration used for the experiment) by adding 14.144 µL of gBlock to 5.856 µL of water. For the single gBlocks the target concentration was 60000 copies per reaction such that a 50/50 mix could be made to reach 30000 copies per reaction of 100% shearing.

dPCR mixes

Table S.3, Shows the information relevant for the dPCR runs with gBlocks. HPLC grade water was added to the mix until the final PCR volume was reached

perfecta-multiplex-qPCR-Toughmix 5X	1X
ENV forward	0,8 µM
ENV reverse	0,8 µM
ENV probe	0.4 µM
GAG forward	0,8 µM
GAG reverse	0,8 µM
GAG probe	0.4 µM
sample volume	1 µl
Final PCR volume	7 µl
Amplicon length ENV	116 bp
Amplicon length GAG	81 bp

Table S.4, shows information relevant to the dPCR run with sonicated DNA, HPLC grade water was added to the mix until the final PCR volume was reached. Note that the Naica master mix used consists of 2 buffers (A and B).

Buffer A	1x
Buffer B	2%
Forward primer (PSI, GAG, POL and Ru5)	0,8 µM
Reverse primer (PSI, GAG, POL and Ru5)	0,8 µM
Probe ((GAG, POL and Ru5)	0,4 µM
Probe (PSI)	0.7 µM
Sample	1 µl
Final PCR volume	25 µl

GBlock sequences

Intact gBlock sequence (GAG-ENV)

AGAAGGAAGTGATACCCATGTTTTCAGCATTATCAGAAGGAGCCACCCCACAAGATTTAAACACCATGCTAAACGCTACAGTGGGGGGACATCA
AGCAGCCATGCTCGAGAGGAAGTGGTGCAGAGAGAAAAAAGAGCAGTGGGAATAGGAGCTTTGTTCCTTGGGTTCTTGGGAGCAGCAGGAA
GCACTATGGGCGCAGCCTCAATGACGCTGACGGTACAGGCCAGAC

Single 1 (ENV)

GAGTGGTGCAGAGAGAAAAAAGAGCAGTGGGAATAGGAGCTTTGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACTATGGGCGCAGCCTCA
ATGACGCTGACGGTACAGGCCAGACAATTATTGT

Single 2 (GAG)

TAGAAGGAAGTGATACCCATGTTTTCAGCATTATCAGAAGGAGCCACCCCACAAGATTTAAACACCATGCTAAACGCTACAGTGGGGGGACATC
AAGCAGCCATGGCCATGCAAATGTTAAAAGAT

189

190