Computational Biology and Chemistry 68 (2017) 1–5

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Algorithm Note

Simultaneous estimation of detection sensitivity and absolute copy number from digital PCR serial dilution

Xutao Deng[a,b,*], Brian S. Custer[a,b], Michael P. Busch[a,b], Sonia Bakkour[a,b], Tzong-Hae Lee[a,b]

[a] Blood Systems Research Institute, San Francisco, CA 94118, USA
[b] Department of Laboratory Medicine, University of California at San Francisco, San Francisco, CA 94107, USA

ARTICLE INFO

Article history:
Received 23 November 2016
Accepted 30 January 2017
Available online 1 February 2017

Keywords:
Digital PCR
Copy number estimation
Serial dilution
Poisson distribution
Virus
Maximum likelihood

ABSTRACT

Digital polymerase chain reaction (dPCR) is a refinement of the conventional PCR approach to nucleic acid detection and absolute quantification. Digital PCR works by partitioning a sample of DNA or cDNA into many individual, parallel PCR reactions. Current quantification methods rely on the assumption that the PCR reactions are always able to detect single target molecules. When the assumption does not hold, the copy numbers will be severely underestimated. We developed a novel dPCR quantification method which determines whether the single copy assumption is violated or not by simultaneously estimating the assay sensitivity and the copy numbers using serial dilution data sets. The implemented method is available as an R package "digitalPCR".

1. Introduction

Digital polymerase chain reaction (dPCR) is a refinement of conventional PCR methods that can be used to directly quantify nucleic acids (Sykes et al., 1992; Kalinina et al., 1997; Vogelstein and Kinzler, 1999). In dPCR, each sample is divided into many aliquots or partitions and the reaction is carried out in each partition individually. Each partition will report negative or positive reactions depending on whether target molecules are present in the partition. This dPCR working principle eliminates the reliance on the uncertain number of exponential amplification cycles and effectively increases the number of data points and thus statistical power. Therefore, dPCR allows a more accurate quantification of the absolute copy numbers than the conventional end-point PCR and reverse transcription-qPCR (RT-qPCR) (Heid et al., 1996).

Current dPCR copy number quantification methods (Hindson et al., 2011) are based on two assumptions: 1) that the number of molecules in a partition is randomly distributed following the Poisson distribution (Ross, 2010); and 2) the PCR reaction need only one target molecule in the reaction mixture to give rise to a positive reading. At the limit of dilution, where some partitions are positive and some are negative, the Poisson distribution can usually provide an accurate estimate of copy number of the initial sample (Sykes et al., 1992). The second assumption of single-copy sensitivity, however, holds true only in ideal PCR conditions. Partially degraded target molecules, DNA hairpins, DNA methylations, non-specific priming, polymerase errors and many other factors all contribute to lower sensitivity of PCR reactions (Lindahl and Nyberg, 1972; Eckert and Kunkel, 1991; Pavlov et al., 2002; Kiselev et al., 2014). Assuming single-copy PCR sensitivity when it is not true will result in severely underestimated copy numbers. It is also important to be able to generate warning messages from the statistical analysis when the data indicate that the single-copy assumption is violated.

To address this important issue, we propose a computational method that estimates dPCR copy number without assuming single-copy PCR sensitivity. Instead, the copy number and assay sensitivity are simultaneously estimated using dPCR data from serial dilutions. A serial dilution is the stepwise dilution of a substance in solution. Usually the dilution factor at each step is constant, resulting in a geometric progression of the concentration in a logarithmic fashion, for example, a 4-step 1-2-4-8 fold serial dilution. The assay sensitivity is an integer, indicating the minimum threshold for the target molecules that the dPCR reports a positive reading. For example, single-copy sensitivity indicates

Abbreviations: dPCR, digital polymerase chain reaction; qPCR, quantitative real-time PCR; RT-qPCR, reverse transcription-qPCR; MLE, maximum-likelihood estimation; SE, standard error.

* Corresponding author at: 270 Masonic Ave. San Francisco, CA 94118, USA.
E-mail addresses: xdeng@bloodsystems.org (X. Deng), bcuster@bloodsystems.org (B.S. Custer), mbusch@bloodsystems.org (M.P. Busch), sbakkour@bloodsystems.org (S. Bakkour), tlee@bloodsystems.org (T.-H. Lee).

http://dx.doi.org/10.1016/j.compbiolchem.2017.01.015
1476-9271/

one copy of target molecules is sufficient to generate positive PCR readings. Five-copy sensitivity indicates 5 or more copies of target molecules are needed to generate positive PCR readings. The copy number reported is the number of target molecules in each partition in the initial dilution. Our method is based on maximum-likelihood estimation (MLE) for copy number and assay sensitivity estimation. We also developed a bootstrapping method to assess the accuracy of the estimates.

2. Methods

2.1. Algorithm for estimating copy number and sensitivity

For a partition with expected λ copies of molecule, its actual copy number S is a Poisson random variable with mean λ. The probability that S equals a specific copy number i given λ is expressed as:

$$\Pr(S = i;\lambda) = \frac{e^{-\lambda}\lambda^{i}}{i!}, \quad i = 0, 1, 2 \ldots$$

A PCR reaction has a sensitivity threshold k when it produces a positive reading if and only if k or more copies of molecules are present in the partition. For example, single-copy sensitive PCR has a threshold k of 1. The probability of negative and positive readings for the partition can be calculated as:

$$p_{neg}(\lambda, k) = \Pr(S < k;\lambda) = \sum_{i=0}^{k-1}\frac{e^{-\lambda}\lambda^{i}}{i!}$$

$$p_{pos}(\lambda, k) = 1 - p_{neg}(\lambda, k)$$

For a dPCR experiment with x positive partitions and y negative partitions, the probability of observing x and y given λ and k is governed by the binomial distribution expressed as:

$$\Pr(x, y;\lambda, k) = \binom{x+y}{x} \cdot \left(p_{pos}(\lambda, k)\right)^{x} \cdot \left(p_{neg}(\lambda, k)\right)^{y}$$

The log-likelihood of observing (x, y) given unknown parameters λ and k is therefore:

$$LL(\lambda, k; x, y) = C + x \cdot \log(p_{pos}(\lambda, k)) + y \cdot \log(p_{neg}(\lambda, k))$$

, where C is a constant with regard to parameters λ and k. For simplicity, we can set C = 0 without affecting our estimates.

When the sample is serially diluted, for example, 1-fold, 2-fold, 4-fold, and 8-fold, and the observed data are (x_1,y_1), (x_2,y_2), (x_3,y_3), and (x_4,y_4) respectively, the total likelihood is defined as:

$$TL(\lambda, k; D) = LL(\lambda, k; x_1, y_1) + LL(\lambda/2, k; x_2, y_2) + LL(\lambda/4, k; x_{3,}, y_3) + LL(\lambda/8, k; x_4, y_4)$$

, where D is the collective observed data (x_1,y_1), (x_2,y_2), (x_3,y_3), and (x_4,y_4). The total likelihood for other serial dilution scenarios can be similarly defined. Note that the log likelihood and total likelihood are functions of λ and k given observed data.

According to the method of maximum likelihood estimation (MLE), we can search the parameter values of λ and k that maximize total likelihood (TL) as the point estimates for λ and k.

$$\left(\hat{\lambda}, \hat{k}\right) = \underset{\lambda, k}{\operatorname{argmax}} TL(\lambda, k; D)$$

We use the following iterative search algorithm to find the MLE estimates in the parameter space $(0, \lambda_{max})$ and $(1, k_{max})$ where λ_{max} and k_{max} are maximum parameter values to be searched:

2.2. MLE Algorithm for point estimate of λ and k

Input: D = (x_1,y_1), (x_2,y_2), . . . (x_n,y_n), λ_{max} k_{max}
Output: $\left(\hat{\lambda}, \hat{k}\right)$
Function: MLE(D)

```
{
    max =−∞
    For λ = 0.1, 0.2, ..., λmax {
        For k = 1,2,...,kmax {
            If max < TL(λ,k; D) {
                max= TL(λ,k; D)
                (λ̂, k̂)= (λ,k)
            }
        }
    }
    return (λ̂, k̂)
}
```

Table 1
Estimated copy numbers with standard error and thresholds based on simulated datasets (100 bootstraps).

Simulation Paraeters		Bootstrap Estimates								
		#partitions=50			#partitions=500			#partitions=5000		
Thres	Copy	Estimated Copy	SE Copy	Estimated Threshold	Estimated Copy	SE Copy	Estimated Threshold	Estimated Copy	SE Copy	Estimated Threshold
1	1	1.11	0.5	1.09	0.99	0.04	1	1	0	1
1	2	1.94	0.2	1	1.99	0.06	1	2	0.01	1
1	4	4.13	1.1	1.05	3.98	0.13	1	3.99	0.05	1
2	2	2.38	0.79	2.32	2	0.06	2	2	0.01	2
2	4	4.49	1.31	2.25	4.01	0.11	2	4	0.03	2
2	8	7.57	2.13	1.98	7.97	0.25	2	7.99	0.06	2
5	5	6.18	2.25	6.23	5.24	0.61	5.21	5	0.03	5
5	10	11.15	2.76	5.65	10.14	0.95	5.1	9.97	0.05	5
5	20	16.19	6.09	4.36	19.64	1.67	4.96	20	0	5

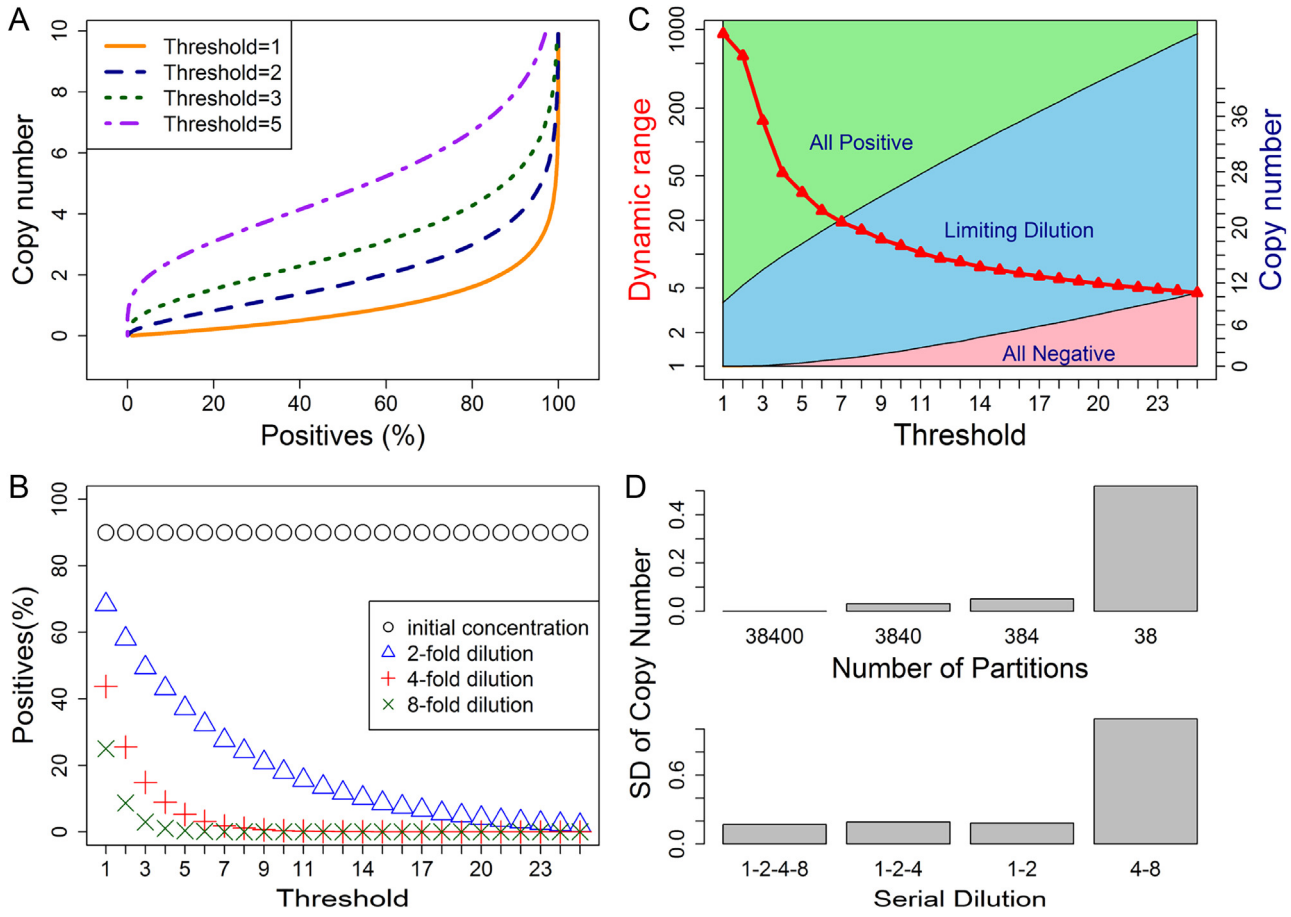

Fig. 1. The factors that affect dPCR copy number estimation: sensitivity thresholds, limiting dilution, and number of partitions. A. The sensitivity thresholds significantly impact copy number estimates from dPCR positive readings. Under the same positive readings, higher sensitivity assays yield lower copy number estimates (lower curves) B. Pattern of serial dilution readings in different thresholds when reading is 90% positive in initial concentration. Different thresholds would produce different reading patterns under the same serial dilution scheme. This fact allows us to infer the thresholds from reading patterns yielded from the serial dilution. C. The copy number and threshold define three distinct regions in dPCR: Saturated region where all readings are positive (green); depleted region where all readings are negative (pink); and limiting dilution region where the readings are informative (blue). The majority of the serial dilutions should fall in this region for better estimation accuracy. Higher thresholds lead to lower assay dynamic range which is defined as the ratio of upper and lower copy numbers bound in the limiting dilution region (blue). D. The number of partitions and serial dilutions affect accuracy of copy number estimates. The y-axis is the standard deviation of bootstrapping copy number estimation. Higher number of partitions would produce higher copy number accuracy (lower standard deviation). Similarly, having dilutions that are in the limiting dilution region is important to produce more accurate estimates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Datasets and copy number (standard error) and threshold estimates (1000 bootstraps).

Target Molecule	Serial Dilutions	dPCR positive readings (# partitions = 384)	Mean Copy Number (SE)	Threshold (frequency)
Dengue type 1	1-2-4-8	352,211,120,60	3.64(1.01)	1(229), 2(771)
Dengue type 1 (rep)	1-2-4-8 3-9-27	221,127,70,2897,39,14	0.83(0.05)	1(1000)
Dengue type 2	1-2-4-8 3-9-27	227,153,76,48102,42,11	0.94(0.05)	1(1000)
CHIKV Plasmid	1-2-4-8	127,80,32,25	0.42(0.04)	1(1000)
CHIKV type 1	1-2-4-8	79,40,19,9	0.24(0.15)	1(961), 2(39)
SFTSV	1-2-4-8 3-9-27	137,81,35,15 64,22,8	0.47(0.05)	1(1000)

In order to estimate the confidence interval of $(\hat{\lambda}, \hat{k})$, we developed the following bootstrapping algorithm. From initial sample of x positives and y negatives, a bootstrap sample (x', y') is constructed by randomly taking x + y data points from the initial sample with replacement. The above MLE is applied on (x', y') and the solution $(\hat{\lambda}, \hat{k})'$ is obtained. This process is repeated for a large number of times (typically 1000 or 10,000 times), to generate a histogram (distribution) of bootstraps $(\hat{\lambda}, \hat{k})'$. This provides an interval estimate of the $(\hat{\lambda}, \hat{k})$.

2.3. Bootstrapping algorithm for interval estimate of λ and k

Repeat 1000 times the following:
D' = bootstrap(D)
$(\hat{\lambda}, \hat{k})'$ = MLE(D')

Table 3
Zika virus copy number estimates (standard error) comparing with experimentally determined true copy numbers.

Target Molecule	True Copy Number	Serial Dilutions fold	dPCR (# partitions)	Positive readings	Negative readings	Mean Copy Number (SE)	Threshold (frequency)
Zika Virus	69.0 (determined experimentally)	1	20	20	0	71.32 (9.75)	1(1000)
		3	20	20	0		
		9	40	40	0		
		27	40	36	4		
		81	40	26	14		
		243	40	12	28		
		729	20	1	19		
		2187	20	0	20		

Compute standard deviation from the bootstrapped distribution $\left(\hat{\lambda}, \hat{k}\right)$' as the interval estimates of λ and k

2.4. dPCR experiments to demonstrate the proposed method

The target molecules used were Zika Virus (ZIKV), Dengue type 1, Dengue type 2, Chikungunya virus (CHIKV) type 1, CHIKV plasmid, and Severe Fever with Thrombocytopenia Syndrome Virus (SFTSV). The ZIKV were French Polynesia isolate stock virus, made of an inactivated culture supernatant of strain H/PF/2013 provided by the European Virus Archive. The Dengue viruses and SFTSV were virions from culture supernatant provided by the CDC; the CHIKV was cultured at the Blood Systems Research Institute. The starting concentration (copy number) of these viruses was unknown. RNA was extracted using the Qiagen Viral RNA kit per manufacturer's instructions, and reverse transcribed into cDNA followed by real-time PCR amplification.

Real-time PCR amplification was performed using 384-well plates in the Roche Lightcycler 480 with cycle conditions of 1 min at 95 °C followed by 45 cycles of 30 s at 95 °C and 60 s at 60 °C. The following sequence-specific primers and TaqMan probes were used to amplify the cDNA: ZIKV (1086/1162c/1107-FAM), Dengue subtype 1 (DEN-1F/1C and DEN-1P), Dengue subtype 2 (DEN-2F/2C and DEN-2P), CHIKV (CHIK-F5/R5 and CHIK-P5), and SFTSV (S-F-3/S-R-3 and S-P-3). An amplification was interpreted to be positive if two criteria were met: the cycle threshold (Ct) must be less than 45 cycles and the melting-curve analysis confirmed that the melting temperature of the PCR product matched that of the positive control well, which indicates that the PCR product is specific.

3. Results and Discussions

3.1. Simulation results

We simulated a number of sensitivity threshold and copy number combinations and tested the accuracy and precision of the proposed method with the simulated data. Table 1 shows simulation parameters with sensitivity thresholds 1, 2, 5, and initial dilution copy number 1 to 20 with serial dilution 1-2-4-8-16 and 50, 500, or 5000 partitions. We applied our method to these simulated data, and the estimated copy number and sensitivity threshold are shown in Table 1. It shows that with as little as 50-partition digital PCR, our method obtained reasonably good estimates for copy number and assay sensitivity threshold. The estimated copy numbers are mostly within 10% of the true numbers except when threshold is high (low sensitivity), where copy number can be off by 20%. These estimates should be sufficient for many quantitative applications. We observed (in Fig. 1) that low sensitivity assays have low dynamic range which limits the usefulness of the serial dilution data. When number of partitions is high, such as 500 and 5000, the estimates are increasingly better, usually within a few percent of the true numbers. The simulation results demonstrate that our algorithm is able to accurately estimate dPCR copy numbers with unknown assay sensitivity, when there are a reasonable number of partitions.

3.2. Copy number estimation for a number of target molecules

Table 2 shows the estimated copy numbers and thresholds of 5 different target virus molecules. In many targets, the estimated thresholds are indeed 1, that is, single copy sensitive and the estimated copy number is highly accurate with small confidence intervals. However for Dengue type 1, the estimated copy number has a high standard error. The 1000 bootstrap threshold yield is 2 for 771 times and 1 for 229 times, suggesting that the single copy assumption is violated. In this case, we repeated dPCR experiments for Dengue type 1 and obtained more definitive results, where single copy sensitivity is established This example demonstrated that our method is able to generate a warning when single-copy sensitivity is violated, which helps the users identify experimental issues early. For CHIKV type 1, the estimated copy number had a high level of variance, which resulted from the low initial concentration.

Table 3 shows the estimated copy number of the ZIKV sample whose true copy numbers were determined experimentally by the European Virus Archive. The serial dilutions involve 7 different concentrations, with a 3-fold dilution performed at each concentration. The number of partitions at each concentration varies from 20 to 40. Based on 1000 bootstraps, the estimated copy number for the initial concentration is 71.32 copies per partition which is close to the true copy number at 69 copies per partition.

3.3. The factors affecting dPCR copy number estimation

The relationships between assay threshold, copy number, serial dilution and dPCR readings are illustrated in Fig. 1. Fig. 1A shows that lower thresholds lead to higher copy number estimates under the same dPCR positive readings. When the single-copy assumption (lowest curve) does not hold, the copy number based on the false assumption will result in severe underestimation, because the true copy number is actually at a higher curve. Our method, however, attempts to fit data to different threshold curves for the best match. This is demonstrated in Fig. 1B, which demonstrates the serial dilution patterns of different thresholds while holding the initial positive reading at 90%. For the 2, 4, and 8-fold dilutions, higher thresholds will result in lower numbers of positive readings. Therefore the pattern of serial dilution readings is an indicator of thresholds, which forms the basis of our method. Fig. 1C shows the 3 dPCR reading regions defined by copy numbers and sensitivity thresholds. The upper "All Positive" region is where copy number is too high such that all partitions are positive (probability of positive >0.9999). The lower "All Negative" region is when copy number is too low and all partitions are negative (probability of positive <0.0001). The middle "Limiting Dilution" region, which is most informative for copy number estimation, reflects a situation where

both negative and positive partitions are present. We define the assay dynamic range as the ratio of copy number upper and lower bounds in the limiting region. Fig. 1C demonstrated that although the dynamic range of dPCR can approach 1000 fold for single-copy dPCR, it rapidly declines when threshold increases. Based on the data of Dengue type 2, Fig. 1D shows that the number of partitions and serial dilutions greatly affect the accuracy (standard deviation) of copy number estimates. The upper panel shows that the required number of partitions should ideally be in the hundreds to accurately estimate copy numbers. The lower panel shows that it is more important to have initial concentration that is in the informative limiting dilution region than to have more dilutions for accurate estimation.

4. Conclusions

In summary, we developed an accurate method for copy number estimation from serial dilution dPCR data without the assumption that the assay is single copy sensitive. For serial dilution data with dPCR readings from multiple concentrations, we use MLE to combine the Poisson estimators for all dilutions and generate a copy number estimate that optimally fits the overall serial dilution data. These estimates were numerically obtained by searching a grid of parameter space of sensitivity thresholds and copy numbers. In order to obtain the variance of the estimated copy number and assay sensitivity, a bootstrapping procedure was developed which repeatedly and randomly re-sampled the serial dilution data many times with replacement. We also demonstrated the conditions when such estimation is favorable or not. Higher number of partitions, higher sensitivity and more dilutions in the informative limiting dilution region lead to more definitive and accurate estimates.

Competing Interests

The author(s) declare(s) that they have no competing interests.

Ethics Statement

The author(s) declare(s) that no human subjects or human data is used in this study.

Author's Contribution

THL oversaw all dPCR experiments and conceived of the study. XD designed the algorithm, performed computational work and wrote the first draft. BSC and MPB participated in the design of the study and manuscript writing. All authors read and approved the final manuscript.

Availability of Data and Materials

The software is freely available as the R digitalPCR package at https://cran.r-project.org/web/packages/digitalPCR/index.html.

Acknowledgement

The authors would like to thank Mr. Li Wen for carrying out digital PCR experiments.

References

Sykes, P.J., Neoh, S.H., Brisco, M.J., Hughes, E., Condon, J., Morley, A.A., 1992. Quantitation of targets for PCR by use of limiting dilution. BioTechniques 13, 444–449.
Kalinina, O., Lebedeva, I., Brown, J., Silver, J., 1997. Nanoliter scale PCR with TaqMan detection. Nucleic Acids Res. 25, 1999–2004.
Vogelstein, B., Kinzler, K.W., 1999. Digital PCR. Proc. Nat. Acad. Sci. U. S. A. 96, 9236–9241.
Heid, C.A., Stevens, J., Livak, K.J., Williams, P.M., 1996. Real time quantitative PCR. Genome Res. 6, 986–994.
Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A. J., Bright, I.J., Lucero, M.Y., Hiddessen, A.L., Legler, T.C., 2011. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal. Chem. 83, 8604–8610.
Ross, S.M., 2010. Introduction to Probability Models, 10th ed. Academic Press, Amsterdam; Boston.
Lindahl, T., Nyberg, B., 1972. Rate of depurination of native deoxyribonucleic acid. Biochemistry 11, 3610–3618.
Eckert, K.A., Kunkel, T.A., 1991. DNA polymerase fidelity and the polymerase chain reaction. PCR Methods Appl. 1, 17–24.
Pavlov, A.R., Belova, G.I., Kozyavkin, S.A., Slesarev, A.I., 2002. Helix-hairpin-helix motifs confer salt resistance and processivity on chimeric DNA polymerases. Proc. Nat. Acad. Sci. U. S. A. 99, 13510–13515.
Kiselev, K.V., Dubrovina, A.S., Tyunin, A.P., 2014. The methylation status of plant genomic DNA influences PCR efficiency. J. Plant Physiol. 175C, 59–67.