Nucleic Acids Research Advance Access published August 7, 2014

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku603

Direct elicitation of template concentration from quantification cycle (C_q) distributions in digital PCR

Mitra Mojtahedi[1,†], Aymeric Fouquier d'Hérouël[2,3,*,†] and Sui Huang[1,2,*]

[1]Department of Biological Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada, [2]Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109, USA and [3]Luxembourg Centre for Systems Biomedicine, L-4362 Esch-sur-Alzette, Luxembourg

Received April 9, 2014; Revised June 17, 2014; Accepted June 23, 2014

ABSTRACT

Digital PCR (dPCR) exploits limiting dilution of a template into an array of PCR reactions. From this array the number of reactions that contain at least one (as opposed to zero) initial template is determined, allowing inferring the original template concentration. Here we present a novel protocol to efficiently infer the concentration of a sample and its optimal dilution for dPCR from few targeted qPCR assays. By taking advantage of the real-time amplification feature of qPCR as opposed to relying on endpoint PCR assessment as in standard dPCR prior knowledge of template concentration is not necessary. This eliminates the need for serial dilutions in a separate titration and reduces the number of necessary reactions. We describe the theory underlying our approach and discuss experimental moments that contribute to uncertainty. We present data from a controlled experiment where the initial template concentration is known as proof of principle and apply our method on directly monitoring transcript level change during cell differentiation as well as gauging amplicon numbers in cDNA samples after pre-amplification.

INTRODUCTION

Digital polymerase chain reaction (dPCR) is the most accurate and sensitive method to measure the abundance of specific nucleic acids. Conceptually simple, it is based on distributing a diluted sample over many replica reactions and direct counting of positive and negative reactions. The number of positive reactions N^+ following 'end-point' amplification directly yields a reliable estimate of the total initial number of templates using standard Poisson statistics, provided the sample is dilute enough to minimize the probability of reactions that contain more than one template. With increasing N^+, however, dPCR becomes prone to saturation and inevitably fails when all reactions are positive. The method we present in the following overcomes this caveat by analyzing quantitative information from each positive reaction in an assay— information which is discarded in conventional dPCR.

Conceived in 1992 (1), dPCR was first implemented in 1999 to quantify disease-associated mutations in cancer patients (2) and is now available as a standard repertoire of many laboratories. Advances in microfluidics and robotic micromanipulation have led to the development of automated and cost-effective dPCR systems that can run hundreds of parallel replica reactions in the nanoliter to picoliter range. Such integrated systems have been used as tools in clinical studies, such as viral detection (3), biomarker analysis (4) and prognostic monitoring (5,6) as well as in research studies for bacteriophage-host interactions (7), DNA copy number variation (8) and transcription factor profiling (9). An essential requirement for dPCR is the proper concentration of the sample, since the method becomes uninformative when all replica reactions are positive, i.e. when each reaction contains (at least) one template molecule. The accuracy of dPCR depends on the dilution range: a situation with ~20% negative reactions yields the most informative results (10,11).

For appropriately dilute samples, dPCR achieves higher precision and sensitivity for absolute quantification than classical real-time quantitative PCR (qPCR), owing to its independence from a standard curve and its digital, directly countable output (12). Conversely, if the analyzed sample is not in the proper dilution range ('high' concentration), accuracy is impeded and eventually the dPCR method may fail entirely ('too high' concentration). An estimate of sample concentration range is therefore crucial for accurate measurements in dPCR. Routinely, this is obtained by performing an initial dPCR titration experiment, which can be a limiting factor, particularly in case of rare samples. Moreover, this has to be repeated for each nucleic acid sequence to be analyzed in the same sample. While complications of dPCR for samples with low copy numbers of template

*To whom correspondence should be addressed. Tel: +1 206 732 1208; Fax: +1 206 732 1299; Email: sui.huang@systemsbiology.org
Correspondence may also be addressed to Aymeric Fouquier d'Hérouël. Tel: +352 466 644 6723; Fax: +352 466 644 3 6723; Email: aymeric.dherouel@uni.lu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

have been discussed elsewhere (12), the case of high template concentrations remains unaddressed to the best of our knowledge. To alleviate the need for prior titration experiments, we devised a method that allows us to recover absolute quantification in low dilution (high concentration) regimes or when the number of replica reactions is limited by the instrument. Our method, which we term 'retroflex' PCR, takes advantage of quantification cycle (C_q) distributions measured across the replica reactions (Figure 1a–c). By comparing this empirical distribution (Figure 1d) to a theoretical model of expected C_q-values (Figure 1e), we are able to infer the amount of template molecules in the initial sample (Figure 2), in the absence of a standard curve or relative controls. More importantly, our approach permits users to infer gene expression from entirely positive sets of replica reactions where Poisson statistics fails to determine the number of template copies. We validate the method on dilution series of plasmids carrying murine GATA1 and PU1 genes and demonstrate its applicability for quantifying the expression of these genes in mouse progenitor EML cells along with differentiated erythroid and myeloid cells.

MATERIALS AND METHODS

Plasmid templates

GATA1-pSPORT1 and Sfpi1-pCMV-pSPORT6 plasmids (clone IDs 30039400 and 3600260) containing Mus musculus GATA1 and PU1 cloned target sequences were obtained from Open BioSystems (Huntsville, AL, USA). Plasmids were propagated in DH5α *Escherichia coli* strain (Invitrogen) and plasmid DNA was isolated using QIAprep spin miniprep kit (Qiagen). DNA concentration c_{DNA} was measured using NanoDrop 1000 spectrophotometer (Thermo Scientific) and converted (13) to plasmid copy number n_c via

$$n_c = \frac{N_A c_{DNA}}{N_{bp} m_{bp}}, \quad (1)$$

where N_A and N_{bp} represent Avogadro's number and the base pairs of the plasmid, respectively, and average weight of a single base pair is considered to be $m_{bp} \approx 660$Da. To compare amplification efficiency of circular and linearized plasmids, GATA1-pSPORT1 and Sfpi1-pCMV-pSPORT6 were treated with Not1 restriction enzyme (Roche) at 37°C for 1 h, using 1 unit of enzyme per 1 μg of plasmid in reaction volumes of 25 μl. The integrity and the conformation of plasmid samples were confirmed using 1% agarose gel electrophoresis (Supplementary Figure S1). The concentration of the linearized plasmids was measured on a Nanodrop 1000 (Thermo Scientific). Aliquots of both circular and linearized plasmids were subsequently stored at –20°C to avoid repeated freeze-thaw cycles.

Serial dilutions of the plasmid samples were prepared to 950 and 900, 475 and 450, 238 and 225, 119 and 113, 59 and 56, 30 and 28, 15 and 14 copies/μl for GATA1-pSPORT1 and Sfpi1-pCMV-pSPORT6, respectively (Supplementary Tables S1 and S2). Each subarray of 64 reactions was loaded with 2 μl of sample. To minimize subsampling errors, final dilution steps were prepared with a single excess volume. Low-binding tips and tubes (Applied Biosystems) were used throughout this study since a significant influence of plastics on PCR results has been observed (Supplementary Figure S2).

Culture and differentiation of EML cells

Differentiated EML cells of erythroid (ERY) and myeloid (MYL) lineage were used to evaluate the performance of dPCR.

EML-C1 cells (ATCC CRL-11691) were maintained in Iscove's modified Dulbecco's medium (IMDM, Invitrogen), supplemented with 20% (vol/vol) donor horse serum (DHS, Invitrogen) and 12% (vol/vol) BHK/MKL conditioned medium. BHK/MKL cell line was established by transfecting BHK cells with an expression vector containing a cDNA encoding the secretory form of mouse SCF. Conditioned medium from BHK/MKL cells was used for maintaining EML-C1 cells as the source of mouse SCF (14).

To obtain myeloid cells (MYL), EML cells were suspended in growth medium containing 10 ng/ml IL-3 (Sigma) and 10 μM ATRA (Sigma) at an initial density of 2×10^5 cells/ml. After incubation for 3 days at 37°C in 5% CO_2, cells were resuspended in IMDM + 20% DHS + 2% BHK/MKL conditioned medium + 1% glutamine + 10 ng/ml IL-3 + 10μM ATRA + 10 ng/ml mGM-CSF (Peprotech). Cells were further incubated at 37°C for 2 days before additional 10 ng/ml mGM-CSF (Peprotech) was added (14).

To obtain erythroid cells (ERY), EML cells were suspended in growth medium containing 10 ng/ml erythropoietin (EPO, Sigma) at an initial density of 2.5×10^5 cells/ml. Cells were incubated for 3 days at 37°C in 5% CO_2 and then are resuspended in IMDM + 20% DHS + 2% BHK/MKL conditioned medium + 1% glutamine + 10 ng/ml EPO. Cells were further incubated at 37°C for 2 days, after which additional 10 ng/ml EPO was added (14).

Analysis of EML and differentiating cells

Populations of 2000 EML, ERY and MYL cells were sorted on a FACSAria III instrument (BD Biosciences) into lysis solution and cDNA was synthesized directly from cell lysate without prior RNA isolation (cf. **Direct cDNA synthesis**). These samples were used to address the condition of high template concentration (saturation of reactions) in dPCR quantification. Reverse transcription (RT) solution of 1000 EML sorted cells, and 250 ERY and MYL cell equivalents were used to prepare a 4-fold, five-point serial dilution (Supplementary Table S3). Higher equivalents were used for EML cells to anticipate the lower transcript numbers of GATA1 and PU1 relative to ERY and MYL cells, respectively (15). Four subarrays (256 replica reactions) were analyzed for each dilution step. For each subarray, 2 μl of cDNA was used as template for dPCR. No-template controls were included on each array. The number of positive amplifications (counts) was used for analysis of GATA1 and PU1 in these samples (cf. **dPCR analysis**).

To count positive reactions in dPCR, the sample needs to be diluted sufficiently such that ideally each reaction is either empty or contains a single template molecule (4). If a specific transcript is expressed at higher copy numbers, a

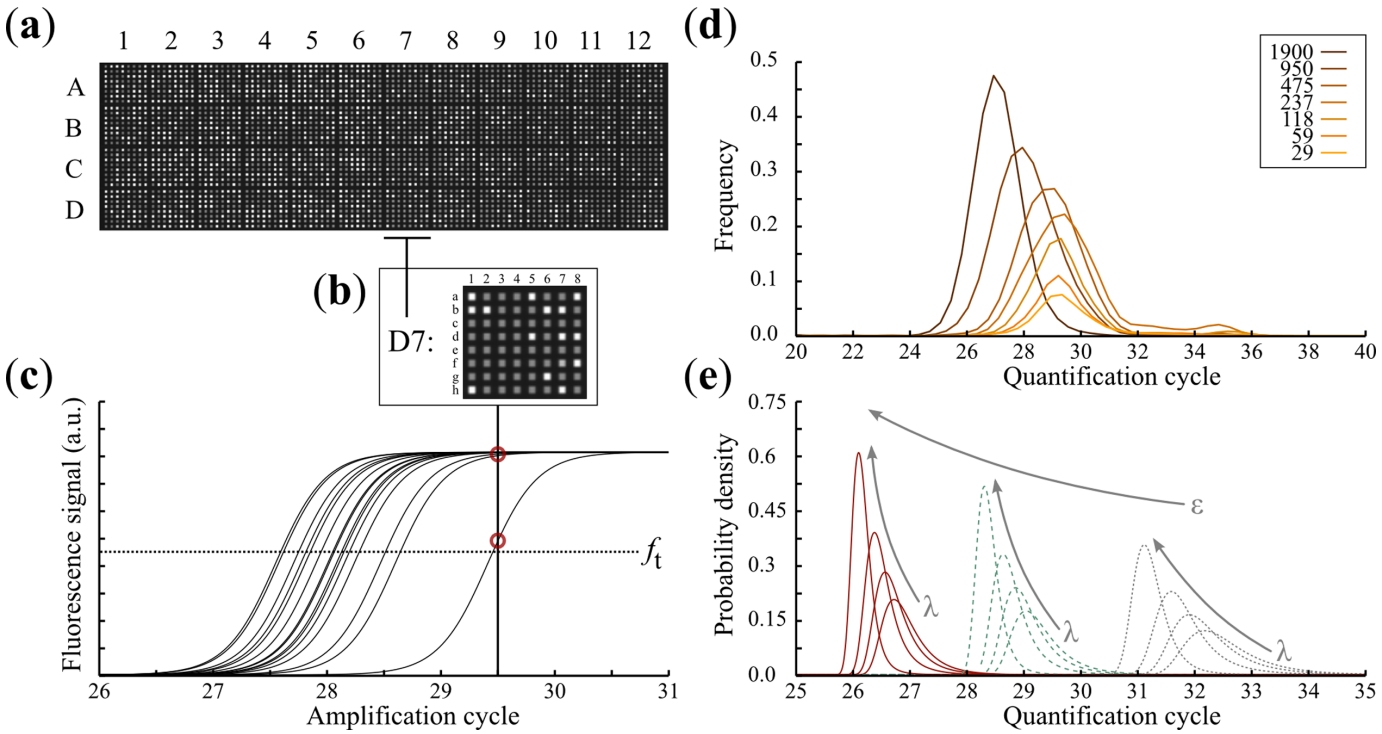

Figure 1. Elements of retroflex method. Layout and typical experimental readout of a dPCR array (**a**), enlargement of subarray D7 (**b**) showing positive (light gray) and negative (dark gray) reactions, illustration of corresponding amplification curves (**c**) with marked acquisition endpoints (red circles) and quantification threshold f_t, empirical C_q distributions (**d**) for different template dilutions (expected copies per subarray) indicated in the legend box, analytical model of C_q distributions (**e**) illustrating the impact of modifying parameters λ and ε (arrows denote increase), curves within each bundle (solid, dashed, dotted) share the same $\varepsilon \in \{2, 1.8, 1.5\}$ and are plotted for $\lambda \in \{5, 2, 1, 0.5\}$ (from left to right).

lower starting concentration should be used in dPCR reactions to maximize accurate quantification.

Analysis of gene-specific pre-amplified cDNA

For the study of pre-amplified samples, single ERY and MYL cells were sorted on a FACSAria III (BD Biosciences) with respect to high and low surface expression of the lymphocyte antigen Sca-1 on the third day of differentiation. Two cells of each type and Sca-1 level were collected for analysis of GATA1 and PU1 expression. RNA from the resulting eight samples was reverse-transcribed (cf. **Direct cDNA synthesis**) and resulting cDNA was pre-amplified. Pre-amplfication reactions were performed in volumes of a 35.75 µl, using 16.5 µl cDNA sample (entire reaction volume from single cell sample), 17.5 µl TaqMan PreAmp master mix and 1.75 µl pooled primer/assay mix. Final primer concentration in pre-amplification reactions was 45 nM. Pre-amplification was performed on an iQ5 thermocycler (BioRad) with the following program: denaturation at 95°C for 10 min and 18 cycles of amplification (15 s at 95°C, 4 min at 60°C). Unincorporated primers were digested by adding a 6 µl solution containing 4 units Exonuclease I (New England BioLabs M0293L) in 1× Exonuclease I Reaction Buffer (New England BioLabs B0293S) and using the following thermal protocol: 30 min at 37°C, 15 min at 80°C, hold at 4°C. The pre-amplified products were then diluted 10-fold with Tris-ethylenediaminetetraacetic acid (EDTA) buffer. Ten-fold diluted pre-amplified samples were analyzed on the dPCR platform (cf. **dPCR analysis**) and reported C_q-values were used for the inference of transcript numbers.

Direct cDNA synthesis

Cell samples were sorted into 5.0 µl CellsDirect (Invitrogen) lysis buffer containing 4.25 µl Resuspension Buffer, 0.25 µl Lysis Enhancer and 0.5 µl RNase out. Note that (i) all heating steps were performed in an iQ5 PCR thermocycler (Bio-Rad); (ii) after sorting, each tube was immediately sealed and heated to 75°C for 10 min and either immediately processed or frozen at –80°C; (iii) processed samples were incubated at room temperature for 10 min with 2.5 µl DNase I (Invitrogen) and 0.8 µl DNase I buffer (Invitrogen) to remove genomic DNA. After the DNase step, 0.6 µl of 25 mM EDTA was added to the sample, and each tube was heated to 70°C for 5 min to inactivate the enzyme. cDNA was reverse transcribed using a mix of random primers and oligo(dT)20 (1 µl oligo(dT)20 at 50 µM, 0.6 µl random primers at 75 ng/µl, 0.5 µl 10 mM dNTPs) and each tube was heated to 70°C for 5 min. After priming, RT was initiated by adding 3 µl RT buffer, 0.5 µl RNaseOut, 1.0 µl SuperScript III RT and 1.0 µl DTT to each tube (25°C for 10 min, 50°C for 60 min, 85°C for 5 min). Samples were subsequently stored at –20°C. RNase-free solutions as well as sterile, disposable lab ware were used for all RNA processing steps.

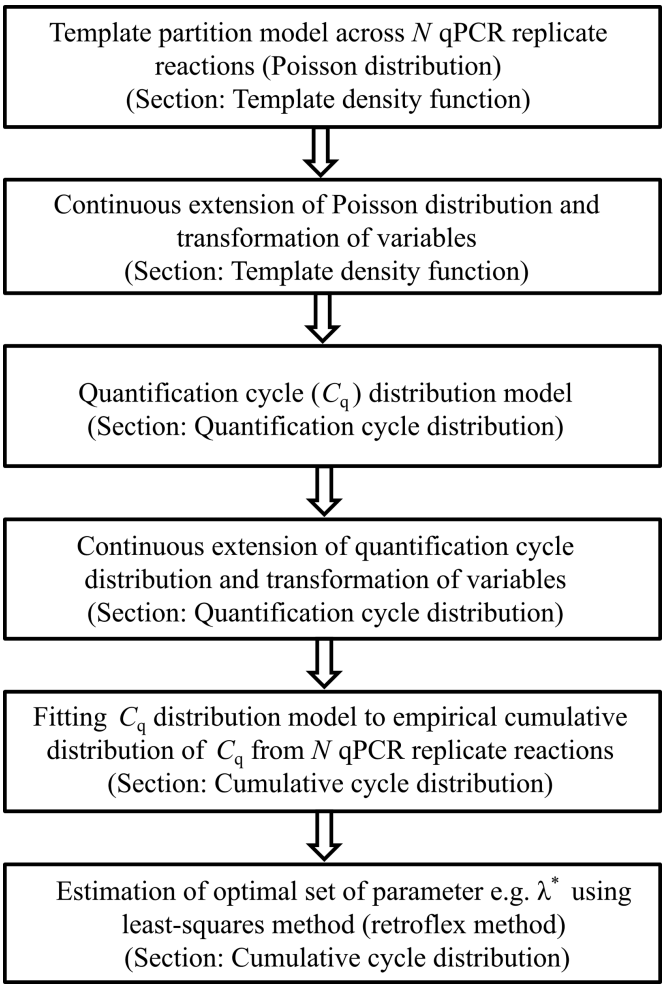

Figure 2. Flow diagram of retroflex procedure. Template molecules are partitioned within the reaction wells according to the Poisson distribution. A continuous extension of the Poisson distribution is used in the approach and variables are transformed accordingly. A model of quantification cycles is established from the expected distribution of templates in replica reactions. This model is fitted to empirical distribution of quantification cycles and a set of optimal parameters is estimated.

Gene-specific qPCR

Inventoried TaqMan assays (Applied Biosystems) contained 900 nM of forward and reverse primers and 250 nM of a FAM-labeled minor groove-binding hydrolysis probe. The GATA1 assay (Mm01352636_m1) targeted exon junction 4–5, producing a 73 bp amplicon. The Sfpi1 assay (Mm00488140_m1) targeted exon junction 1–2, producing an 82 bp amplicon. The manufacturer does not provide primer and probe sequences. All assays were optimized and assessed initially by qPCR on an iQ5 real-time PCR system (Bio-Rad) by altering primer concentration and estimating assay efficiency using standard curves. Reactions of 8 µl consisting of 5 µl TaqMan Universal PCR Master Mix (Applied Biosystems), 0.5 µl sequence-specific gene assay (GATA1 or PU1, primers at 9 µM and probe at 2 µM), 2.5 µl molecular grade water (Gibco) and 2 µL of DNA at various concentrations (Supplementary Table S3). Samples were run in quadruplicate on 96-well plates (Bio-Rad). PCR efficiency was estimated using six-point, 10-fold, serially diluted standard curves (triplicate measurements performed at each point). The qPCR reaction was performed using the following parameters: 50°C for 2 min, 95°C for 10 min, 40 cycles of 95°C for 15 s and 60°C for 1 min. No-template controls were included to rule out any possible contamination.

dPCR analysis

dPCR was performed using the OpenArray system (Applied Biosystems). Each dPCR plate consists of 12 × 4 subarrays and each subarray contains 8 × 8 reaction chambers. Each dilution was analyzed across 11 subarrays (704 reactions) (Figure 1a), reserving the 12th for no-template and water controls.

The instrument software generates positive/negative amplification calls for each of the 3072 reaction chambers (64 × 48) (Figure 1b). Following amplification, digital raw data were processed by the OpenArray Digital PCR Analysis software (version I10R3) to measure average copies/reaction. Amplification curves (Figure 1c) and quantification cycles (C_q) were extracted using the OpenArray Real-Time qPCR Analysis software (version 1.0.4) with a quantification threshold of 100(5), yielding C_q-values between 20 and 40 for the reaction chambers with positive amplification (cf. Supplementary Figure S3 for an example with negative and positive amplification curves). For each subarray, 2 µl of target sample was loaded into a well of a 384-well plate (Applied Biosystems); subsequently 3 µl master mix reactions consisting of TaqMan OpenArray Digital Master Mix (Applied Biosystems) and TaqMan sequence-specific gene assay (Applied Biosystems) was added. Target and master mix were mixed, centrifuged and the 384-well plate was processed in the OpenArray AccuFill system (Applied Biosystems), where 2.1 µl of reaction solution is transferred automatically from each well of the 384-well plate into the corresponding subarrays of a dPCR plate. The reaction solution enters into the reaction wells due to differential hydrophilic–hydrophobic coating between wells and surface of the dPCR array (24). Reactions were performed using following thermocycling conditions: 50°C for 2 min, 95°C for 10 min, 40 cycles of 95°C for 15 s and 60°C for 1 min.

Template density function

When a volume V containing n template molecules is partitioned across $N \gg 1$ replica reactions of identical volume $v = V/N$, then the expected number of molecules in each replicate is $\lambda = n/N$ and the probability of finding k_0 molecules in any replicate is given by the Poisson distribution

$$p(k_0; \lambda) = e^{-\lambda} \frac{\lambda^{k_0}}{k_0!} \quad (2)$$

Equation (2) is well defined for integer-valued k_0. In this work, a continuous extension of the discrete Poisson distribution is used to facilitate the analytical treatment (cf. **Discussion**). Replacing k_0 by a 'continuous' template number x and introducing the associated random variable X, the

probability density function of the partitioning problem is given by the derivative

$$\rho_X(x;\lambda) = \frac{\partial}{\partial x}\frac{\Gamma_\lambda(x)}{\Gamma(x)} = \frac{\partial}{\partial x}Q_\lambda(x), \tag{3}$$

where $\Gamma_\lambda(x) = \int_\lambda^\infty \mathrm{d}t e^{-t}t^{x-1}$ and $\Gamma \equiv \Gamma_0$ denote the 'upper incomplete' and 'complete' gamma functions, respectively. The fraction in Equation (3) is commonly referred to as the 'regularized' gamma function Q_λ. Although the resulting density is similar to the well-known Erlang distribution (16), it is noteworthy that here the usual notion of parameters and variables is inverted.

Quantification cycle distribution

A statistical model of quantification cycles (Figure 1e) is established from the expected distribution of templates in replica reactions and the equation describing PCR. The amplification of templates by PCR is well described by an exponential function $k(C) = k_0\varepsilon^C$ of the amplification cycle C (17), with k_0 initial templates and an efficiency of $\varepsilon \in (1, 2]$. dPCR calls a reaction positive when a given threshold signal f_t is exceeded (Figure 1c). This signal is accordingly proportional to a threshold template number $k_t = k(C_q)$ with associated 'quantification cycle' C_q. This specifies a parametric relationship between the threshold k_t and the initial template number k_0. As with the template number, the quantification cycle is relaxed from being discrete and replaced by a continuous variable $y \in (0, \infty)$. The relationship $x(y) = k_t\varepsilon^{-y}$ then yields a model for the quantification cycle distribution of the random variable Y in terms of the density

$$\rho_Y(y;\lambda,k_t,\varepsilon) = \log\varepsilon\frac{k_t\varepsilon^{-y}}{\Gamma(k_t\varepsilon^{-y})^2}\int_\lambda^\infty \mathrm{d}r\int_0^\lambda \mathrm{d}s\, \mathrm{e}^{-(r+s)}(rs)^{k_t\varepsilon^{-y}-1}\log\frac{r}{s}. \tag{4}$$

The compact notation of the integral part is adapted from (18) and the pre-factor results from the transformation of variables (see the Supplementary information for a derivation). This model can then be fitted to the empirical distribution $\tilde{\rho}\left(C_q^{(i)}\right)$ of observed threshold cycles $C_q^{(1)}, \ldots, C_q^{(N)}$ from N replica reactions (Figure 1d). To find the optimal set of parameters, $\{\lambda^*, k_t^*, \varepsilon^*\}$, we adopt the least-squares method, thus

$$\{\lambda^*, k_t^*, \varepsilon^*\} = \underset{\{\lambda,k_t,\varepsilon\}}{\operatorname{argmin}} \sum_{i=1}^{N}\left(\tilde{\rho}\left(C_q^{(i)}\right) - \rho_Y\left(C_q^{(i)};\lambda,k_t,\varepsilon\right)\right)^2. \tag{5}$$

When quantification threshold k_t and amplification efficiency ε are known from control experiments, these parameters are fixed and excluded from the optimization step (see Table 1 for a summary of all parameters and variables in our model).

Table 1. Parameters and variable of the model

Symbol	Quantity	Role in model
λ	Expected molecules per reaction	Central quantity to be inferred; shape parameter of theoretical distribution
ε	Amplification efficiency	Shape parameter of the model; estimated or from experiment
k_t	Quantification threshold number	Shape parameter of the model; estimated from optimization
x, X^*	Amount of template	Auxiliary random variable; used to set up the model
y, Y^*	Quantification cycle	Central random variable; observable in experiment

*Upper case denotes random variables and lower case realizations of these.

Cumulative cycle distribution

To find the optimal set of parameters that describe a set of replicates in a computationally more efficient way, one may employ the cumulative distribution function

$$P_Y(y;\lambda,k_t,\varepsilon) = 1 - Q_\lambda\left(k_t\varepsilon^{-y}\right) \tag{6}$$

instead of the density. With the empirical cumulative distribution $\tilde{P}\left(C_q^{(i)}\right)$ from measurements, the minimization reads as in Equation (5), but replacing ρ by P. The procedure of finding λ^* by fitting the model function P_Y to an empirical cumulative distribution $\tilde{P}$ is henceforth termed *retroflex* method (see Figure 2 for an outline of steps involved).

Retroflex estimation of template number

Empirical C_q distributions are established exclusively from the N^+ positive reactions of a set of replica reactions. With an inferred template concentration λ_r^*, the retroflex estimator for the initial number of template molecule is given as

$$n_r^* = N^+\lambda_r^*. \tag{7}$$

Poisson correction of standard dPCR

In a currently employed method, the observed number of positive replica reactions N^+ among the total number N in dPCR is used to infer the initial amount of templates across all replicates applying the following argument. The probability of placing more than one template molecule in any replica reaction is $P(k_0 \geq 1;\lambda) = 1 - \mathrm{e}^{-\lambda}$. An estimator for the true number of templates, n_p^*, is self-sufficiently defined by the expectation $\mathrm{E}[N^+] = NP\left(k_0 \geq 1;\lambda_p^*\right)$. With $\mathrm{E}[N^+] \cong N^+$ and $\lambda_p^* = n_p^*/N$, the estimator reads

$$n_p^* \cong N\log\frac{N}{N-N^+}. \tag{8}$$

Intuitively, this estimator reports the number of positive replica reactions as the amount of templates when $N^+ \approx 1$. However, it diverges when $N^+ = N$, the situation when all replica reactions are positive and dPCR ultimately becomes uninformative.

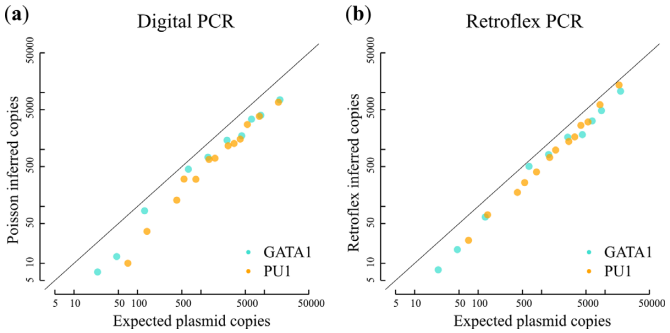

Figure 3. dPCR and retroflex method on plasmid dilution series. Inferred copies for GATA1 and PU1 plasmid dilutions are pooled and compared to expected concentrations. Each dot corresponds to one dilution step. The offset from the diagonal (perfect concurrence between inference and expectation) indicates the loss of material in the measurement process.

RESULTS

Retroflex method concurs with dPCR for plasmid dilutions

GATA1 and PU1 carrying plasmids were prepared in serial dilutions and measured in replica reactions on the OpenArray platform. Results were pooled to compare the outcomes of the reference method of (**Poisson**) corrected dPCR (Figure 3a) and our newly developed (**Retroflex**) approach (Figure 3b) on large data sets. Comparison of expected and inferred copy numbers clearly shows the concurrence between dPCR and the retroflex method.

The quantification threshold k_t, a machine-dependent quantity, is a free parameter of the model and can be estimated in presence of sufficient observations. It is not expected to vary between different experiments. Therefore, k_t was estimated using the dilution factor for which most data (reported C_q-values) was available. Consistently for both GATA1 and PU1 plasmids at highest concentration, it was found to be $k_t = 8.03 \cdot 10^8$ from an optimization of the model, while ε was taken from efficiency measurements (Supplementary Figure S4).

Adsorption reduces effective template number in serial dilutions

Template concentrations inferred from both Poisson corrected dPCR and retroflex PCR were lower than expected for an unbiased dilution (Figure 3). The nanoliter fluidic digital array used in this study enables partitioning of a parent sample across a subarray comprising 64 individual reactions of ~33 nl volume each (19). The total volume of each subarray is hence 2.1 µl, while the loading volume for each subarray inlet is 5 µl. The resulting dead volume of 58% is taken into account in the calculation of expected plasmid copies (Supplementary Table S4).

Figure 3 illustrates the reduction in plasmid copies, reflected as off-diagonal shift in both Poisson corrected dPCR and the retroflex method. These shifts correspond to average deviations by factors of 2.63 ± 1.17 (Figure 3a) and 1.94 ± 0.48 (Figure 3b) from expectation using the former and latter method, respectively. We ascribe this deviation partially to residual adsorption of DNA by pipetting tips and plastic tubes. This adsorption leads to an effective loss of material since overall yield was in fact affected by the choice of plastic (Supplementary Figure S2). Other possible causes such as electrostatic effects in the array filling robot or adsorption by the reaction wells remain to be investigated.

Template circularity marginally impairs amplification

Purity of the extracted DNA plasmid was found to be within an acceptable range (260/280 absorption ratio of 1.78). Circularity of plasmids has been reported by others to have negative effect on PCR efficiency; therefore, it has been recommended to use linearized molecules as standards for absolute quantification by qPCR (20). To test whether template circularity indeed had an impact on our results, serial 10-fold dilutions ranging from 10^7 to 10^2 copies/µl were generated for each type of plasmid (Supplementary Table S3). The efficiency of amplification was evaluated for circular and linearized starting templates. Standard DNA was freshly prepared before the experiment to avoid degradation that may occur during storage. Interestingly, the PCR efficiency was high for both plasmids (90%, $\varepsilon = 1.90$, for GATA1-pSPORT1 and 94%, $\varepsilon = 1.94$, for Sfpi1-pCMV-pSPORT6); furthermore, similar efficiency was found for both plasmids in both circular and linearized conformations (Supplementary Figure S4). To put the reported efficiencies into perspective the reported efficiencies for qPCR primers used on (shorter and less complex) cDNA templates were higher, with 100% ($\varepsilon = 2.00$) using GATA1 primers, and 96% ($\varepsilon = 1.96$) for PU1 (Supplementary Figure S5). Reported C_q-values for circular plasmids were consistently larger than those of the linearized molecules at every dilution step, although the shift was not significant for either GATA1-pSPORT1 (P-value 0.63) or Sfpi1-pCMV-pSPORT6 (P-value 0.69) according to the t-test.

Retroflex method exceeds the dynamic range of dPCR

The applicability of our approach was tested on the quantification of gene expression from cell lysates. Gene expression was measured in different cell numbers from the mouse multipotent hematopoietic progenitor line EML with developmental potential similar to that of the human common myeloid progenitor cells (CMP). These cells are able to differentiate to several downstream lineages (14), such as erythroid (ERY) and myeloid (MYL) cells that we studied here. Two thousand cells of each type, EML, ERY and MYL, were sorted into lysis buffer, reverse transcribed and first diluted or directly analyzed for their expression of GATA1 and PU1 (Supplementary Figure S6). These genes are of specific interest due to their central role in fate-determination of the progenitor cells (21). Serial dilutions of the lysate were examined to further compare the performance of the retroflex method to that of classical dPCR (Figure 4). Retroflex and Poisson inferred GATA1 levels are in good agreement for the EML and ERY samples (Figure 4a), whereas perfect agreement is evident across all samples for PU1 (Figure 4b). We speculate that the lower concurrence for GATA1 is due to a higher tendency of its amplicons to be adsorbed by plastic surfaces (Supplementary Figure S2).

In the next step, expression levels were measured in undiluted samples. Here, all replica reactions were positive, a

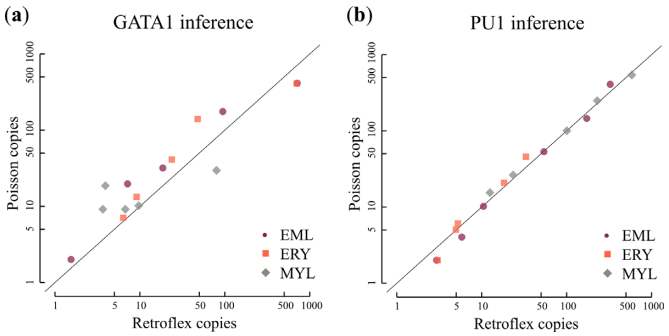

Figure 4. Inferred transcript numbers in directly analyzed cells. Absolute expression levels of GATA1 and PU1 in undifferentiated EML as well as ERY and MYL cells are compared using Poisson corrected dPCR (n_p^*, Equation (8)) and retroflex (n_r^*, Equation (7)), showing concurrence of both methods in dilution ranges that are optimal for dPCR.

Table 2. Inferred transcript numbers from saturated reactions

	Average transcripts per cell	
Cell type	*GATA1*	*PU1*
EML	01.02	03.58
ERY	19.70	–*
MYL	–*	52.31

*Not assessed.

situation where our approach yields results, while classical dPCR fails. Inferred average transcript numbers per cell (Table 2) agree well with corresponding single-cell analysis results (in preparation) and a report on PU1 levels in murine CMP cells (9).

To further illustrate the applicability of the retroflex method, levels of GATA1 and PU1 were analyzed in pre-amplified cDNA from single ERY and MYL cells of high and low expression of Sca-1 (Supplementary Figure S7) by distributing the samples from both cell types over 192 and 320 replica reactions, corresponding to 3 and 5 subarrays, respectively (Supplementary Figures S8). All reactions were positive, rendering classical dPCR incompatible with this assay. Distributions of measured C_q-values (Supplementary Figure S9) allowed us to find the amplicon numbers per subarray reported in Figure 5. With an estimated dead volume of 58%, primer efficiencies on cDNA as reported above of 100% for GATA1 and 96% for PU1, and 18 cycles of amplification, the results yield molecule numbers in the pre-amplified samples as reported in Table 3 together with corresponding extrapolations of transcript numbers at the single cell level. The reported numbers concur well with the previous results.

Retroflex method reduces the number of necessary replica reactions

We addressed the question on the reaction numbers required for the retroflex method to outperform classical dPCR by asking at which significance level it can distinguish between a template concentration λ and $\lambda' = \lambda + \Delta\lambda$ when observing N^+ positive reactions. To this aim, we sampled sets of C_q-values from $\rho_Y(y; \lambda, k_t, \varepsilon)$ and assessed the

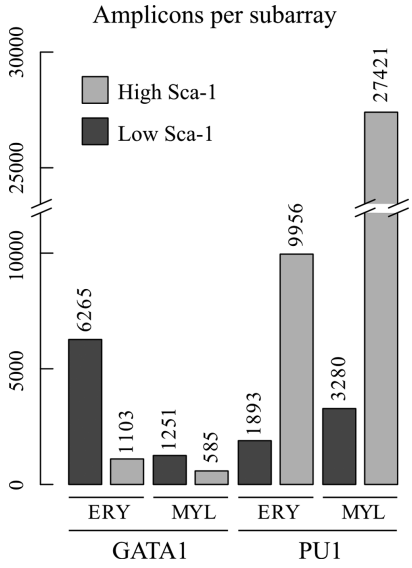

Figure 5. Inferred amplicon numbers in pre-amplified samples. Bars indicate the inferred GATA1 and PU1 amplicon numbers per subarray in pre-amplified single cell cDNA. Individual ERY and MYL cells were sorted on day 3 of differentiation and selected for high or low expression of Sca-1. Each subarray probed 2 µl of 10-fold diluted pre-amplified cDNA sample.

Table 3. Inferred molecule numbers from pre-amplified cDNA.

			Molecule number	
Cell type	Sca-1†	Sample*	*GATA1*	*PU1*
ERY	H	A	$0.46{\cdot}10^6$	$4.15{\cdot}10^6$
		C	4	45
	L	A	$2.61{\cdot}10^6$	$0.79{\cdot}10^6$
		C	20	8
MYL	H	A	$0.24{\cdot}10^6$	$11.43{\cdot}10^6$
		C	2	123
	L	A	$0.52{\cdot}10^6$	$1.37{\cdot}10^6$
		C	4	15

†High (H) or low (L) surface expression level of Sca-1 of single cells measured in flow cytometry.
*Number equivalents in pre-amplified samples (A) or corresponding single cells (C).

probability of these sets to stem from $\rho_Y(y; \lambda + \Delta\lambda, k_t, \varepsilon)$ using the Anderson–Darling test. Repeating the assessment for 100 trials yielded empirical p-value distributions for different N^+ (Figure 6). To compare these results to classical dPCR, we asked how many positive reactions we would need to observe to distinguish λ from λ' at 95% confidence, as described by others (22). Using the formalism derived in (22), we find the necessary positive observations as

$$N^+_{\mathrm{dPCR}} = \mathrm{e}^{\lambda}\left(\frac{z_{\mathrm{conf}}}{1 - e^{-\Delta\lambda}}\right)^2, \qquad (9)$$

with $z_{\mathrm{conf}} = 1.96$ for 95% confidence. For the analyzed values of $\lambda = 0.5, 1, 2, 4,$ and 8, this translates to required numbers of positives $N^+_{\mathrm{dPCR}} \approx$ 41, 67, 183, 1355, and 73968, respectively, as given by

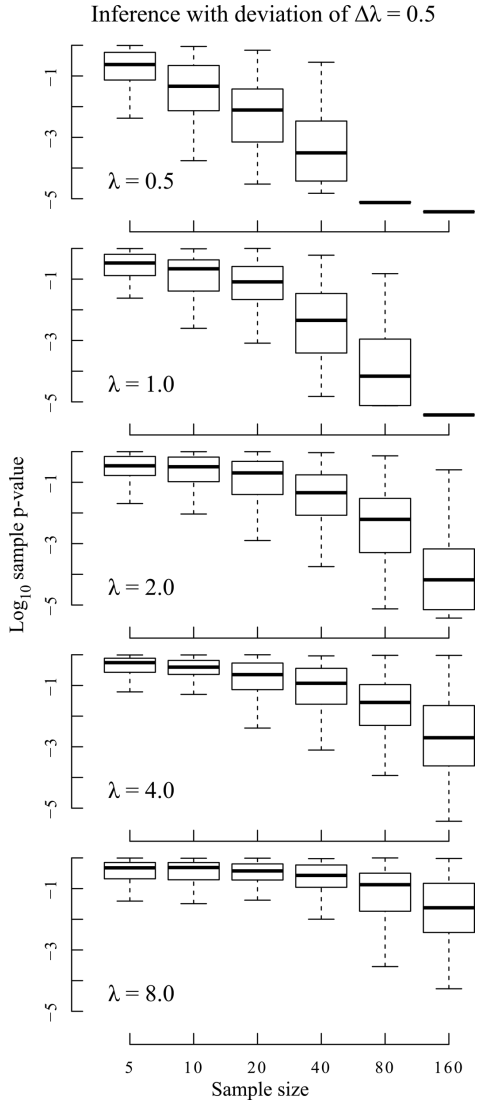

Figure 6. Error probability in estimation of concentration parameter. *P*-value distribution box-plots from Anderson–Darling test assessing the probability of inferring $\lambda' = \lambda + \Delta\lambda$ from samples parametrized by λ. For the shown distributions, we assumed optimal efficiency ($\varepsilon = 2$) and a fixed threshold of $k_t = 8 \cdot 10^8$.

Equation (9). In contrast, retroflex expectably discerns distributions that differ by $\Delta\lambda = 0.5$ at reasonable confidence (p-value < 0.005), e.g. for $\lambda = 4$ at 160 positive observations in 61 out of 100 trials (Figure 6).

DISCUSSION

We developed the retroflex method as a novel estimator for absolute template numbers in massive-parallel dPCR-like investigations. Comparison with the traditional naïve Poisson estimator, which is based on counting positive and negative reactions, has demonstrated the advantage of retroflex, specifically when all reactions are positive. A naïve estimator similar to Equation (8) has been introduced in (11) for the discussion of dPCR error models, and an alternative method based on a binomial distribution of template molecules across replica reactions, which yields congruent results and equivalently breaks down as all reactions become positive, has been discussed in (23). It is worth underlining the difference between retroflex and classic dPCR: while we use a mathematical model to infer the concentration of template molecules independently of the actual dilution, the classic approach is optimal only at a fixed concentration.

In addition to the binary readout of dPCR, our new method exploits information on the relative abundance of templates provided by the quantification cycle (C_q) values of individual reactions. The approach thus essentially bridges the gap between classical dPCR and qPCR. It is based on a probabilistic description of C_q distributions in a dPCR-like setting with replicate amplification reactions of a DNA template. Here, λ, the expected number of template molecules in individual reactions parameterizes the theoretical model of C_q and the fit to a measured distribution yields a reliable estimate for the initial number of templates, which we have demonstrated this in dilution series experiments. Our continuous model affords a straightforward analytical treatment. However, as the number of expected molecules approaches zero ($\lambda \to 0$), the probability density approaches a uniform distribution, spanning the complete range of allowed C_q-values. A widening of the distribution (Figure 1e) with decreasing λ is also visible in the data, reflecting larger uncertainties in terms of expected quantification cycles due to small variations in highly diluted samples. In this situation, a fit of the empirical distribution from few observed positive events is futile, and straightforward counting of positive reactions is preferable in such a 'low concentration' regime (12). This scenario, however, is readily identified.

We also assessed the impact of linearizing plasmids on dPCR results, a matter that is currently debated in the literature with conflicting reports (references hereto and a brief discussion in the supplement). Indeed, we found only a marginal effect of plasmid linearity on amplification efficiency and dPCR results.

Classical dPCR requires either high dilution factors, with each dilution step increasing the variance of the actual templates (Supplementary Figure S10) and therewith the uncertainty of the measured results, or a very large numbers of reactions. Ideally, samples should be partitioned to have an average of 1.59 molecules per reaction (10), amounting to a case where 79.6% of the prepared wells yield positive reactions. Conversely, for a number of n template molecules, $N_{opt} = 0.63 \cdot n$ is the ideal amount of reactions across which the sample should be distributed. It is clearly a limiting factor to require the same order of reactions as there are template molecules, especially for population analysis. In contrast, we have shown that the retroflex method yields statistically significant results for smaller numbers of reactions.

Overall, by taking the C_q-values into account in a dPCR setting, the retroflex method achieves more accurate quantification, has a larger dynamic range, requires a smaller number of parallel reactions and obviates the need for tedious calibration PCRs to determine the initial dilution. However, the quantification results of retroflex, albeit improved compared to current qPCR, still fail to meet the theoretical expectation (Figure 3) due to non-optimal efficiency of amplification. This may be addressed by future improvements in liquid handling.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Zhiming Gu, Dr. Elena Grigorenko, Dr. Mark Gray, Dr. Jeff Johnston and Dr. Elen Ortenberg from the BioTrove team (now part of Life Technologies) for logistical support, Dr. Alexander Skupin for critical discussions and useful comments on the manuscript and Dr. Payman Pirzadeh for careful proofreading.

FUNDING

Canadian Institutes for Health Research (CIHR) [to S.H.]; Institute for Systems Biology, Seattle [to A.F.d'H. and S.H.]; NIGMS/NIH Center for Systems Biology [2P50 GM076547 to S.H.]; Alberta Innovates Technology Futures (AITF) Scholar Award [to S.H.]. Source of open access funding: CIHR and AITF.
Conflict of interest statement. None declared.

REFERENCES

1. Sykes,P.J., Neoh,S.H., Brisco,M.J., Hughes,E., Condon,J. and Morley,A.A. (1992) Quantitation of targets for PCR by use of limiting dilution. *BioTechniques*, **13**, 444–449.
2. Vogelstein,B. and Kinzler,K.W. (1999) Digital PCR. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 9236–9241.
3. Shen,F., Sun,B., Kreutz,J.E., Davydova,E.K., Du,W., Reddy,P.L., Joseph,L.J. and Ismagilov,R.F. (2011) Multiplexed quantification of nucleic acids with large dynamic range using multivolume digital RT-PCR on a rotational SlipChip tested with HIV and hepatitis C viral load. *J. Am. Chem. Soc.*, **133**, 17705–17712.
4. Day,E., Dear,P.H. and McCaughan,F. (2013) Digital PCR strategies in the development and analysis of molecular biomarkers for personalized medicine. *Methods*, **59**, 101–107.
5. Belgrader,P., Tanner,S.C., Regan,J.F., Koehler,R., Hindson,B.J. and Brown,A.S. (2013) Droplet digital PCR measurement of HER2 copy number alteration in formalin-fixed paraffin-embedded breast carcinoma tissue. *Clil. Chem.*, **59**, 991–994.
6. Wang,J., Ramakrishnan,R., Tang,Z., Fan,W., Kluge,A., Dowlati,A., Jones,R.C. and Ma,P.C. (2010) Quantifying EGFR alterations in the lung cancer genome with nanofluidic digital PCR arrays. *Clin. Chem.*, **56**, 623–632.
7. Tadmor,A.D., Ottesen,E.A., Leadbetter,J.R. and Phillips,R. (2011) Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science*, **333**, 58–62.
8. 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. *et al.* (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.*, **83**, 8604–8610.
9. Warren,L., Bryder,D., Weissman,I.L. and Quake,S.R. (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 17807–17812.
10. Fazekas de St Groth,S. (1982) The evaluation of limiting dilution assays. *J. Immunol. Methods*, **49**, R11–R23.
11. Weaver,S., Dube,S., Mir,A., Qin,J., Sun,G., Ramakrishnan,R., Jones,R.C. and Livak,K.J. (2010) Taking qPCR to a higher level: analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods*, **50**, 271–276.
12. Sanders,R., Huggett,J.F., Bushell,C.A., Cowen,S., Scott,D.J. and Foy,C.A. (2011) Evaluation of digital PCR for absolute DNA quantification. *Anal. Chem.*, **83**, 6474–6484.
13. Dhanasekaran,S., Doherty,T.M., Kenneth,J. and Group,T.T.S. (2010) Comparison of different standards for real-time PCR-based absolute quantification. *J. Immunol. Methods*, **354**, 34–39.
14. Tsai,S., Bartelmez,S., Sitnicka,E. and Collins,S. (1994) Lymphohematopoietic progenitors immortalized by a retroviral vector harboring a dominant-negative retinoic acid receptor can recapitulate lymphoid, myeloid, and erythroid development. *Genes Dev.*, **8**, 2831–2841.
15. Enver,T., Heyworth,C.M. and Dexter,T.M. (1998) Do stem cells play dice? *Blood*, **92**, 348–352.
16. Forbes,C., Evans,M., Hastings,N. and Peacock,B. (2011) *Statistical Distributions*. Wiley, Hoboken, NJ.
17. Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, **25**, 402–408.
18. Ilienko,A. (2013) Continuous counterparts of Poisson and binomial distributions and their properties. *Annales Univ. Sci. Budapest*, *Sect. Comp.*, **39**, 137–147.
19. Roberts,D.G., Morrison,T.B., Liu-Cordero,S.N., Cho,J., Garcia,J., Kanigan,T.S., Munnelly,K. and Brenan,C.J.H. (2009) A nanoliter fluidic platform for large-scale single nucleotide polymorphism genotyping. *BioTechniques*, **46**, ix–xiii.
20. Lin,C.-H., Chen,Y.-C. and Pan,T.-M. (2011) Quantification bias caused by plasmid DNA conformation in quantitative real-time PCR assay. *PLoS ONE*, **6**, e29101.
21. Zhang,P., Behre,G., Pan,J., Iwama,A., Wara-Aswapati,N., Radomska,H.S., Auron,P.E., Tenen,D.G. and Sun,Z. (1999) Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 8705–8710.
22. Dube,S., Qin,J. and Ramakrishnan,R. (2008) Mathematical analysis of copy number variation in a DNA sample using digital PCR on a nanofluidic device. *PLoS ONE*, **3**, e2876.
23. Warren,L., Weinstein,J. and Quake,S. (2007) The digital array response curve. *In Thesis*: *Luigi Warren*, Stanford University, Stanford, CA.
24. Morrison,T., Hurley,J., Garcia,J., Yoder,K., Katz,A., Roberts,D., Cho,J., Kanigan,T., Ilyin,S.E., Horowitz,D., Dixon,J.M. and Brenan,C.J.H. (2006) Nanoliter high throughput quantitative PCR. *Nucleic Acids Res.*, **34**, e123.