Methods 59 (2013) 32–46

Contents lists available at SciVerse ScienceDirect

Methods

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

# Evaluation of qPCR curve analysis methods for reliable biomarker discovery: Bias, resolution, precision, and implications

Jan M. Ruijter [a,*], Michael W. Pfaffl [b], Sheng Zhao [c], Andrej N. Spiess [d], Gregory Boggy [e], Jochen Blom [f], Robert G. Rutledge [g], Davide Sisti [h], Antoon Lievens [i], Katleen De Preter [j], Stefaan Derveaux [j,1], Jan Hellemans [k], Jo Vandesompele [k]

[a] Department of Anatomy, Embryology & Physiology, Academic Medical Center, Meibergdreef 15, 1100AZ Amsterdam, The Netherlands
[b] Physiology Weihenstephan, Center of Life and Food Sciences Weihenstephan, Technical University of Munich, Germany
[c] Department of Psychology and Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720-1650, USA
[d] Department of Andrology, University Hospital Hamburg-Eppendorf, Germany
[e] DNA Software Inc., Ann Arbor, MI 48104, USA
[f] Bioinformatics Resource Facility, Center for Biotechnology, Bielefeld University, Germany
[g] Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, 1055 du P.E.P.S., Quebec, Canada G1V 4C7
[h] Department of Biomolecular Science, Piazza Rinascimento 6, University of Urbino, 61029 Urbino (PU), Italy
[i] Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium
[j] Center for Medical Genetics, Ghent University, Ghent, Belgium
[k] Biogazelle, Zwijnaarde, Belgium, Center for Medical Genetics, Ghent University, Ghent, Belgium

ARTICLE INFO

*Article history:*
Available online 3 September 2012

Communicated by Kenneth Adolph

*Keywords:*
qPCR curve analysis
Transcriptional biomarker
Bias
Precision
Resolution
Benchmark

## ABSTRACT

RNA transcripts such as mRNA or microRNA are frequently used as biomarkers to determine disease state or response to therapy. Reverse transcription (RT) in combination with quantitative PCR (qPCR) has become the method of choice to quantify small amounts of such RNA molecules. In parallel with the democratization of RT-qPCR and its increasing use in biomedical research or biomarker discovery, we witnessed a growth in the number of gene expression data analysis methods. Most of these methods are based on the principle that the position of the amplification curve with respect to the cycle-axis is a measure for the initial target quantity: the later the curve, the lower the target quantity. However, most methods differ in the mathematical algorithms used to determine this position, as well as in the way the efficiency of the PCR reaction (the fold increase of product per cycle) is determined and applied in the calculations. Moreover, there is dispute about whether the PCR efficiency is constant or continuously decreasing. Together this has lead to the development of different methods to analyze amplification curves. In published comparisons of these methods, available algorithms were typically applied in a restricted or outdated way, which does not do them justice. Therefore, we aimed at development of a framework for robust and unbiased assessment of curve analysis performance whereby various publicly available curve analysis methods were thoroughly compared using a previously published large clinical data set (Vermeulen et al., 2009) [11]. The original developers of these methods applied their algorithms and are co-author on this study. We assessed the curve analysis methods' impact on transcriptional biomarker identification in terms of expression level, statistical significance, and patient-classification accuracy. The concentration series per gene, together with data sets from unpublished technical performance experiments, were analyzed in order to assess the algorithms' precision, bias, and resolution. While large differences exist between methods when considering the technical performance experiments, most methods perform relatively well on the biomarker data. The data and the analysis results per method are made available to serve as benchmark for further development and evaluation of qPCR curve analysis methods (http://qPCRDataMethods.hfrc.nl).

© 2012 Elsevier Inc. All rights reserved.

[*] Corresponding author. Fax: +31 20 6876177.

*E-mail addresses:* j.m.ruijter@amc.uva.nl (J.M. Ruijter), michael.pfaffl@wzw.tum.de (M.W. Pfaffl), windupzs@gmail.com (S. Zhao), a.spiess@uke.de (A.N. Spiess), greg@dnasoftware.com (G. Boggy), jblom@cebitec.uni-bielefeld.de (J. Blom), bob.rutledge@nrcan.gc.ca (R.G. Rutledge), davide.sisti@uniurb.it (D. Sisti), antoon.lievens@ugent.be (A. Lievens), katleen.depreter@ugent.be (K. De Preter), jan.hellemans@biogazelle.com (J. Hellemans), joke.vandesompele@ugent.be (J. Vandesompele).

[1] Current address: Wafergen, Fremont, CA, USA.

1046-2023/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2012.08.011

## 1. Introduction

### 1.1. Aim of the current study

Fluorescent labeling of DNA enables real-time monitoring of the accumulation of the amount of reaction product in a PCR reaction [2]. The observed amplification data can thus be used to determine the initial target quantity which makes quantitative PCR (qPCR) currently the method of choice to determine concentrations of low amounts of DNA [3,4]. The combination of qPCR with reverse transcription (RT) enables the quantification of minute amounts of RNA species and makes RT-qPCR a suitable method for discovery and validation of expressed biomarkers.

An ideal biomarker should be sensitive and specific, should differentiate between disease states, and should be easy and accurately detectable [5]. The recent advent of miRNA-based biomarker studies (e.g. [6–8] shows that this RNA species fulfills these criteria. Also mRNAs have been proposed as biomarkers for diseases ranging from depression [9] to cancer such as prostate carcinoma [10] or neuroblastoma [11]; RT-qPCR was also used to address soil contamination [12] and viral infections in cattle [13]. Currently, most RT-qPCR based transcriptional biomarker research employs the comparative-$C_q$ method [14] to determine gene expression ratios without considering the actual amplification efficiency and without relying on multiple reference genes for accurate normalization.

The evolution of methods for analysis of qPCR data, that over the last decade has paralleled the evolution of RT-qPCR in the laboratory, thus seems to be neglected. However, when PCR efficiency correction was employed, the observed discriminative genes, as well as their fold-change in expression, were shown to differ considerably [10]. This illustrates the need to address and compare the performance of the available qPCR curve analysis methods in terms of precision, bias and resolution.

In published comparisons the different analysis methods are often applied on a limited number of samples or assays and in a restricted or outdated way that either does not do them justice or that may result in over-fitting of the data. We therefore embarked upon a combined effort to compare published data analysis methods that have a publicly available user interface. The original developers of these algorithms or interfaces are co-author on this paper and each method was applied in the currently proposed and implemented way to a large and published clinical data set, used for improved outcome prediction of children with neuroblastoma [11].

### 1.2. Background

Basic PCR kinetics are described by $N_n = N_0E^n$, in which $N_0$ and $N_n$, are the starting concentration of the target DNA and the concentration after $n$ cycles, respectively. $N_0$ and $N_n$ are linearly related to $F_n$, the fluorescent signal after n cycles and $F_0$, the fluorescence that would be associated with the starting amount of the target DNA [15]. Note that the linear relation between target concentration and fluorescence may differ between targets and may not hold in the plateau phase of the PCR reaction; most efficiency-based qPCR data analysis assume this relation to be proportional in the exponential phase of the PCR reaction. In the basic equation, the parameter $E$, the base of the exponential amplification function, is the PCR efficiency defined as the fold increase of the amount of DNA per cycle. Thus defined, PCR efficiency ranges from 1 (no amplification) to 2 (complete doubling). Note that in the MIQE guidelines, the PCR efficiency is defined as the increase of product per cycle as fraction of the amount present at the start of the cycle and then ranges from 0 to 1 [1]. In this paper we will use the former definition of $E$ and thus deviate from MIQE.

The more copies of input DNA in the PCR reaction, the fewer cycles of amplification are needed to reach a specific amount of product [16]. This principle forms the basis of the original qPCR data analysis, which still forms the starting point of most current day qPCR analysis methods. The approach is to set a quantification fluorescence threshold ($F_q$) and to determine the number of cycles required to reach that threshold (the quantification cycle, $C_q$).

A series of samples with decreasing but known amounts of the target-of-interest can be used to construct a (mainly linear) calibration line (by plotting the observed $C_q$ against the logarithm of the known nucleic acid inputs) and the target quantity of unknown samples can be derived from this calibration plot [15,17]. This so-called 'absolute quantification' method requires such a calibration plot to be constructed for every target that is measured [18]. Moreover, it assumes the PCR efficiency to be equal in the calibration samples and the biological samples. To correct for differences in sample composition and the yield of the reverse transcriptase reaction [19] a 'relative quantification' approach is required, using, preferably multiple, stably expressed reference genes [14,20,21].

When an estimate of the PCR efficiency is available, the set $F_q$ threshold and observed $C_q$ values can be used to calculate $F_0$, the fluorescence associated with the target quantity, using the equation $F_0 = F_q/E^{C_q}$. In the first reported qPCR quantification model, the PCR efficiency was assumed to be 2 and identical for all measured targets. However, it was shown that this assumption may lead to bias in the results [20,22,23]. The recommendation to implement PCR efficiency values per target resulted in 'efficiency-corrected relative quantification' [20,24].

Depending on the adopted model, amplification efficiency is considered to be constant or continuously decreasing per cycle. In the 'constant efficiency' model, the efficiency only starts to decrease after the exponential phase when competition between primer and amplicon, and/or decreasing concentrations of reagents or fluorochrome, lead to a decrease in reaction efficiency or reduced fluorescence increase per cycle [25,26]. In the 'continuously decreasing efficiency' model, limiting reaction conditions are considered to influence the reaction efficiency from the first cycle onward [27–30].

In the beginning of this millennium several approaches to determine the parameters $F_q$, $C_q$ and $E$, needed for the calculation of (relative) target quantities, were published. Most of these methods are based on a constant efficiency in the exponential phase but differ in the way they determine the fluorescence baseline, exponential phase, efficiency, quantification threshold and $C_q$. To satisfy the notion that the PCR efficiency might differ between targets and reactions [19], several schemes were developed to derive an efficiency estimate from each individual amplification curve. These approaches focused on the exponential phase of amplification and involved using all points [31–33], a set of points [34,35] or 2 points [36,37]. The estimation of the fluorescence baseline is a crucial step in qPCR data analysis and is therefore a recurring theme in qPCR data analysis methods [32,33,38]. All methods compared in this study include a baseline estimation. The calculation of the SDM identifies the cycle with the steepest increase in fluorescence and thus the end of the exponential phase [32,39]. In data analysis based on sigmoidal curves, the $F_q$ value associated with this SDM value determines the observed or modeled $C_q$ [33,40]. To accommodate the fact that PCR amplification curves are not symmetrical, the sigmoidal curve was extended to 5-parameters [37] or changed into a logistic model [32]. To correct the $C_q$ value for differences in amplification efficiency between biological samples and the standard curve the Cy0 value was introduced [41]. Despite this plethora of analysis approaches and the confusion in literature, the similarities between these methods are striking as they are all based on the basic kinetics equation and all calculate a target quantity using an efficiency value and a $C_q$ value (Table 1). A real

**Table 1**
Estimation of parameters in the analysis of qPCR amplification curves. $E$: Efficiency, $C_q$: Quantification cycle, $F_0$: Initial fluorescence at cycle 0, $F_q$: Fluorescence at $C_q$, $F_b$: baseline fluorescence.

| Parameters | | | | | | | |
|---|---|---|---|---|---|---|---|
| Name | Class | Fitting | $F_b$ | $E$ | $C_q$ | $F_q$ | $F_0$ |
| CAmpER: DART [34] | Partial curve | Linear | 3-parameter saturation function fitted to $F_{2–10}$, subtraction from $F_n$ | Calculation of ‘midpoint’ region $F_M$, linear regression of 10-fold region around $F_M$. $E$ is slope of regression. (uses $E$ per reaction) | Calculated from intersection of the linear function derived from the regression and $F_q$ | Multiple of $\sigma(F_{1–10})$, depends on PCR system | $F_0 = F_q/E^{Cq}$ |
| CAmpER: FPLM [32] | Partial curve | Nonlinear, linear | 3-parameter saturation function fitted to $F_{2–10}$, subtraction from $F_n$ | Fitting of three-parameter exponential model to $C_q$ range with $F_b$ as baseline parameter, $E$ is regressed parameter.(uses $E$ per reaction) | Calculated from intersection of three-parameter exponential function and $F_q$ | Multiple of $\sigma(F_{1–10})$, depends on PCR system | $F_0 = F_q/E^{Cq}$ |
| MAK2/MAK2+slope [29] | Partial curve | Nonlinear, recursive | Evaluated as a parameter in the recursive model | | | | Evaluated as a parameter in the recursive model |
| FPK-PCR [30] | Complete curve | Nonlinear, linear | Linear regression function fitted from $F_3$ to $F_n$, where $n$ = cycle number with 5% fluorescence of $F_{max}$ Subtraction from $F_n$ | Cycle-dependent $E_n$ obtained from fitting a six-parameter bilinear model to double-logarithmized $F_n$ (uses $E$ per reaction) | | | Calculated from fitting the cumulative product of all $E_n$ to $F_n$ |
| LinRegPCR [38] | Partial curve | Linear | Iteratively evaluated to deliver a straight line of data points downward from SDM; subtracted from $F_n$ | Linear regression fitted on logarithm of $F$ in a window of 4 cycles, that delivers the least $\sigma(E)$ between reactions per amplicon (uses $E_{mean}$) | One cycle below the top border of the best window of cycles | Fluorescence at $C_q$ | $F_0 = F_q/E^{Cq}$ |
| LRE qPCR [42]<br>LRE-Emax/LRE-E100 | Partial curve | Linear | Average of $F_{3–8}$, subtraction from $F_n$, assumes constant baseline fluorescence | Linear regression of $E_n$ versus $F_n$ fitted to the largest possible window, defined by the difference to averaged $F_0$ values. $E$ is intercept of the regression (Emax). Alternatively, to reduce variance Emax is fixed to 100% | | | $F_0 = \frac{F_{max}}{1+\left(\frac{F_{max}}{F_C}-1\right)(E_{max}+1)^C}$ |
| Cy0 [41] | Complete curve | Nonlinear, linear | | Linear regression of Cy0 values from dilution setup | Intersection of the tangent to the first derivative maximum of a five-parameter logistic model with the abscissa (Cy0) | | $F_0 = 1/E^{Cy0}$ |
| 5PSM [37] | Complete curve | Nonlinear | Intersection of five-parameter logistic model, subtraction from $F_n$ | Efficiency at $C_q$ (uses $E$ per reaction) | Second derivative maximum of the fitted five-parameter logistic model | Fluorescence at $C_q$ | $F_0 = F_q/E^{Cq}$ |
| PCR-Miner [33] | Partial curve | Nonlinear | Weighted fitting of three-parameter exponential model from $C_q$ range with $F_b$ is regressed baseline parameter ($y0$) | Weighted fitting of three-parameter exponential model to $C_q$ range with $E$ as regressed parameter (uses $E_{mean}$) | $C_q$ range: cycle range from $C_{noise}$ to the SDM of the four-parameter logistic model; $C_{noise}$ is obtained as the standard error level of the baseline parameter ($y0$) of the logistic model | The mean of $y0$ and SDM of the logistic model | $F_0 = 1/E^{Cq}$ |

difference in approach lies between those 'constant efficiency' algorithms [32–34,37,38,41] and the methods that are based on continuously decreasing efficiency values [29,30,42] (Table 1 and Supplemental Description of methods).

The aim of the current comparison of qPCR curve analysis methods, carried out by the original authors of the published methods, is to test their precision, bias and resolution, and their reliability in transcriptional biomarker identification. This comparison is based on a large number of raw mRNA expression data sets containing fluorescence readings for each cycle. The data comes from two different real-time PCR instruments and 3 different PCR master mixes (see Section 2). The heterogeneity of the data sets avoids favoring algorithms that may work optimally on a given instrument and reagent combination, but underperform on other combinations. As such this study can function as a first benchmark for future development and evaluation of qPCR curve analysis methods.

## 2. Material and methods

### 2.1. qPCR datasets

#### 2.1.1. Biomarker gene expression profiling in tumor biopsies

Data comes from a previously published study in which a 59-mRNA gene expression signature was developed and validated for improved outcome prediction of children with neuroblastoma [11]. In short, 59 biomarkers and 5 reference genes were measured in 8 μl reactions in a 384-well plate using the LightCycler480 SYBR Green Master (Roche) in a sample maximization experiment design [24]. The 59 genes were carefully selected as being previously reported as prognostic genes in neuroblastoma in atleast 2 independent studies. Each plate contained 366 cDNA samples ($n = 1$) from primary tumor biopsies, a 5-point 10-fold serial dilution series based on an external oligonucleotide standard ($n = 3$, from 150,000 to 15 copies), and a no template control (NTC, $n = 3$). Raw (baseline uncorrected) fluorescent data and $C_q$ values were exported from the LightCycler480 instrument software (using the maximum of the second derivative algorithm). This data set will be referred to as 'biomarker dataset' in this paper.

#### 2.1.2. Four-point 10-fold dilution series

An external oligonucleotide standard was synthesized for the human MYCN gene. The sequence of each standard consists of the forward primer sequence of that particular gene, a stuffer sequence (sequence consisting of an ACTG repeat) in the middle and the reverse complement sequence of the reverse primer of that gene at the end (total length of 55 nucleotides; forward primer GCGAGCTGATCCTCAAACG; reverse primer CGCCTCGCTCTTTAT CTTCTTC; template GCGAGCTGATCCTCAAACGactgactgactgacGAA-GAAGATAAAGAGCGAGGCG). No secondary structures in the template sequence were found upon UNAFold analysis [43]. The external oligonucleotide standard was PAGE purified and blocked at its 3′-end with a phosphate group to avoid participation in the PCR amplification process (Biolegio, the Netherlands). The manufacturer's supplied concentration was confirmed using the Nanodrop 1000 Spectrophotometer (Thermo Scientific). A dilution series consisting of four 10-fold serial dilution points, starting from 15,000 molecules down to 15 molecules was created using 10 ng/μl yeast tRNA as carrier (Roche). The same dilution of the carrier was used to create the NTC sample. qPCR was done on a CFX 384 instrument (Bio-Rad). A 384-well qPCR plate was prepared using a 96-well head pipetting robot (Tecan Freedom Evo 150). qPCR amplifications were performed in 8 μl containing 4 μl iQ SYBR Green Supermix (Bio-Rad), 0.4 μl forward and 0.4 μl reverse primer (5 μM each), 0.2 μl nuclease-free water and 3 μl of standard oligonucleotide. A total of 94 replicated reactions were dispensed for each of 4 dilution points. In addition, the NTC reaction was analyzed in 8 replicates. All reactions were performed in 384-well plates (Hard-Shell 384-well microplates and Microseal B clear using adhesive seals (Bio-Rad). The cycling conditions were comprised of 3 min polymerase activation at 95 °C and 45 cycles of 15 s at 95 °C and 30 s at 60 °C followed by a dissociation curve analysis from 60 to 95 °C. This dataset will be referred to as '94-replicates-4-dilutions set'.

#### 2.1.3. Replicates for assessment of precision

A dilution consisting of 15,000 molecules of the MYCN oligonucleotide was created in 10 ng/μl yeast tRNA carrier. qPCR amplifications were performed in 380 replicated 8 μl reactions and quadruplicated reactions of the NTC sample were performed on the same 384-well plate. qPCR reaction set-up and thermal conditions were identical as mentioned above. This dataset will be referred to as '380-replicates set'.

#### 2.1.4. Competimer primers for PCR efficiency modulation

Forward and reverse primers were designed to amplify human AluSx repeats (forward TGGTGAAACCCCGTCTCTACTAA, reverse CCTCAGCCTCCCGAGTAGCT). Competitive primers were synthesized on basis of identical sequence and blocked by amination at the 3′ end to allow annealing, but avoid elongation during the PCR process. A six-point 4-fold serial dilution series made from reference human genomic DNA (Roche), starting from 64 ng/μl down to 0.0625 ng/μl, was created in 10 ng/μl yeast tRNA as carrier. The same dilution of the carrier was used to create a NTC sample. qPCR amplifications were performed in 7.5 μl total reaction volume containing 3.75 μl 2× custom made qPCR SYBR green I Mastermix (Eurogentec), 0.375 μl forward primer (5 μM), 0.375 μl reverse primer (5 μM), 1 μl of a mixture of nuclease-free water and equal amounts of both forward and reverse competitive (aminated) primers, and 2 μl diluted standard. A total of 7 'competitive' mixes were prepared for each dilution point, containing 0%, 5%, 10%, 20%, 30%, 40%, and 50% (of the total amount of primer) competitive (aminated) forward and reverse primers. Each reaction was run in triplicate. The qPCR cycling was performed on the LightCycler480 (Roche) using white LightCycler480 384-multiwell plates with LightCycler480 sealing foils (Roche). The cycling conditions were comprised of 10 min polymerase activation at 95 °C, and 45 cycles of 15 s at 95 °C, 30 s at 60 °C, followed by a dissociation curve analysis from 60 to 95 °C. This dataset will be referred to as 'competimer set'.

### 2.2. qPCR curve analysis methods

Detailed descriptions of the included curve analysis methods can be found in the Supplemental Description of methods. In the text these methods will be referred to with their preferred abbreviations FPLM, 5PSM, DART, PCR-Miner, LinRegPCR, Cy0, MAK2, FPK-PCR, and LRE-qPCR. FPLM [32] and DART [34] are included with their implementation in CAmpER. For LRE-qPCR two implementations were compared, LRE-Emax and LRE-E100 with an estimated maximum efficiency (Emax) and an Emax set to 100%, respectively. In the descriptions of the curve analysis algorithms the mathematical symbols in the equations are defined per analysis method. Therefore, the same parameter can be represented by a different symbol in different methods and vice versa. The original symbols were preserved to enable the reader to easily refer to the original papers for each method. The way in which the main analysis parameters are estimated by the different curve analysis methods is summarized in Table 1.

The results of all curve analysis methods were compared to the results obtained with the original $C_q$ values, exported from the LightCycler480 software, and the PCR efficiency value derived from the standard curve based on these $C_q$ values (method: Standard-$C_q$).

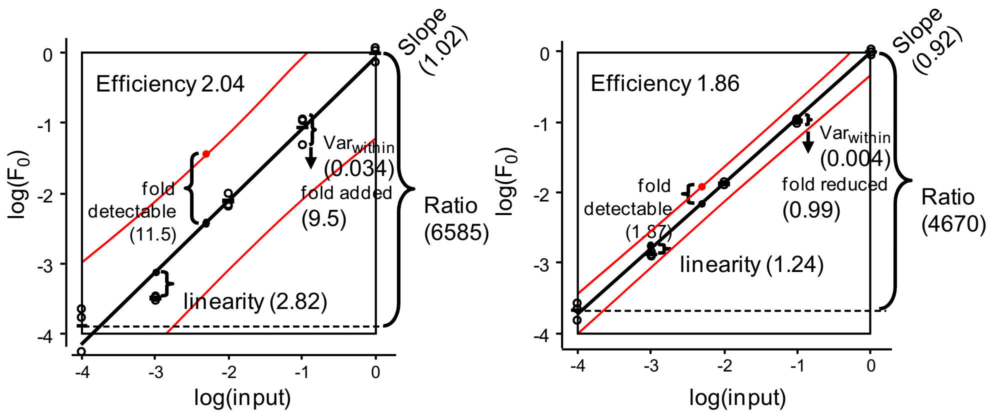


**Fig. 1.** Performance indicators. The calculation of performance indicators based on the analysis of a concentration series in illustrated. The graphs show a representative gene (PRKACB) analyzed with FPK-PCR (left) and LinRegPCR (right). Input data ($x$-axis) and observed $F_0$ values ($y$-axis) are both scaled to set the mean of the highest input and output to 1 and are log-transformed (base10).

### *2.3. Comparison of curve analysis results*

#### *2.3.1. Biomarker data analysis*

In a first step, the number of missing data points resulting from the different curve analysis methods was counted. This analysis was done on the entire dataset including the 59 genes of interest as well as the 5 reference genes.

In a second step, the target quantities determined by the 9 methods (excluding LRE-E100) were log (base 10) transformed and normalized by subtraction of the arithmetic mean of the log-transformed expression values of 4 reference genes (HPRT1, HMBS, SDHA and UBC; AluSq values could not be determined by some methods and were therefore not used), previously determined to be stably expressed in neuroblastoma [21]. As a reference, the original $C_q$ values (generated by the LightCycler480 software) were included in the study. After linearization of these $C_q$ values ($2^{-Cq}$), the same log-transformation and normalization procedure was applied to this reference set. For classification performance assessment, missing values were imputed by the lowest expression value of the gene across all samples minus 1 log-unit.

As third step, we investigated the impact of the curve analysis method on the significance of differential expression of marker genes between two risk groups of cancer patients (high-risk versus non-high-risk, according to previously established criteria [44]. The high-risk subgroup comprised neuroblastoma patients older than 12 months at diagnosis with International Neuroblastoma Staging System (INSS) stage 4 tumors (irrespective of MYCN status) or with INSS stages 2 and 3 tumors with MYCN amplification and patients younger than 12 months with INSS stages 2–4 tumors with MYCN amplification. The non-high-risk subgroup comprised of all other patients. To this purpose, Mann–Whitney tests were run on the 10 datasets for the 59 normalized biomarker genes. Both the fold-change values as the $-\log_{10}$ transformed $p$-values were compared among the methods.

Finally, we studied the effect of the curve analysis method on patient classification performance. For this purpose, a 59-gene expression signature was built using 20 training samples (10 high-risk patients that died of disease and 10 low-risk patients with atleast 36 months event-free follow-up) using the Prediction Analysis of Microarrays (PAM) method as previously described [11]. The performance of the classifiers generated in the 10 different (imputed) datasets was evaluated using receiver operating characteristic (ROC) area under the curve (AUC) analyses using only the patients with an event (death or relapse/progression) or with at least 36 months of follow-up (281 patients in total).

Data-processing and statistical analysis was performed using the R program v2.14.1 with the packages MCRestimate and ROCR (http://bioconductor.org/).

#### *2.3.2. Performance analysis*

For each of the 63 genes in the biomarker dataset (excluding AluSq) a concentration series from 150,000 to 15 DNA copies was measured in triplicate. The raw fluorescence data were analyzed by each of the curve analysis methods. The observed target quantities were used to compare the performance indicators such as precision, bias, resolution and linearity.

All analysis steps were also performed on target quantities that were calculated with the original $C_q$ values and the PCR efficiency value derived from the standard curve based on these $C_q$ values (method: Standard-$C_q$).

The performance analysis was carried out per analysis method. The calculation of the different performance indicators is illustrated in Fig. 1.

1. The data were scaled to set the highest input and its mean output both to 1. This step removed the different measurement scales used by the analysis methods.
2. The fold difference between the mean observed target quantity of the highest and that of the lowest input was calculated. The expected value for this ratio is 10,000; any deviation indicates bias.
3. The scaled input and $F_0$ data were then log-transformed (base 10) which ensures that every concentration will have the same weight in the following calculations. This is required because qPCR results should be valid for the whole range of inputs.
4. The within-triplicate variance was calculated and summed for the 5 concentrations. The expected value per gene is the same for all analysis methods because they all analyzed the same fluorescence data. The observed variances are thus a measure for the precision of the method; systematic differences in variance indicate differences in reproducibility.
5. The variance calculated in step 4 was compared to the variance that is observed when the target quantities were calculated with a conventional standard curve and $C_q$ analysis (method Standard-$C_q$). The resulting ratio between variances shows whether the analysis method leads to increased or reduced variation compared to the conventional analysis.
6. A linear regression analysis of log(output) on log(input) was performed and the 95% CI around the regression line was constructed. The width of this interval was converted into a fold-deviation from the regression line and the geometric mean for

the 5 inputs was calculated. This average fold-deviation is a measure for the fold-difference that deviates significantly from the regression line and is thus a measure for the resolution of the method (lower is better).

7. After linear regression on the log-transformed data the residual variation around the regression line can be split into the deviation of the triplicate data points around the mean $F_0$ per concentration (within-triplicate variance, step 4) and the deviation of those means from the regression line. The latter variance can be considered a measure for the linearity of the input–output relation.

When the analysis methods perform similarly there should be no preferred order between methods in the ranking of the analyzed performance indicators per gene. The non-parametric Friedman test was used to test this null hypothesis per indicator. In case of rejection of the null hypothesis a multiple comparison between methods was carried out to determine which (groups of) methods performed differently [45].

The same performance indicators were determined, if applicable, for the three technical data sets. Additionally, for the competimer set, the observed PCR efficiency values were compared to the expected efficiency that can be calculated from the competimer percentage and the observed efficiency without competimer: $E_{\text{expected}} = 1 + (E_{\text{observed}} - 1) \times (100 - \%\text{competimer})/100$. To compare the variability in performance of the analysis methods in handling those technical datasets, an *F*-test between variances was applied per indicator and the results were used to determine subsets of methods with similar variance.

## 3. Results

### 3.1. Comparison of the results of the analysis of the biomarker set

The different curve analysis methods generated datasets with different numbers of missing values, as summarized in Fig. 2A. Missing values are due to expression below limit of detection, or

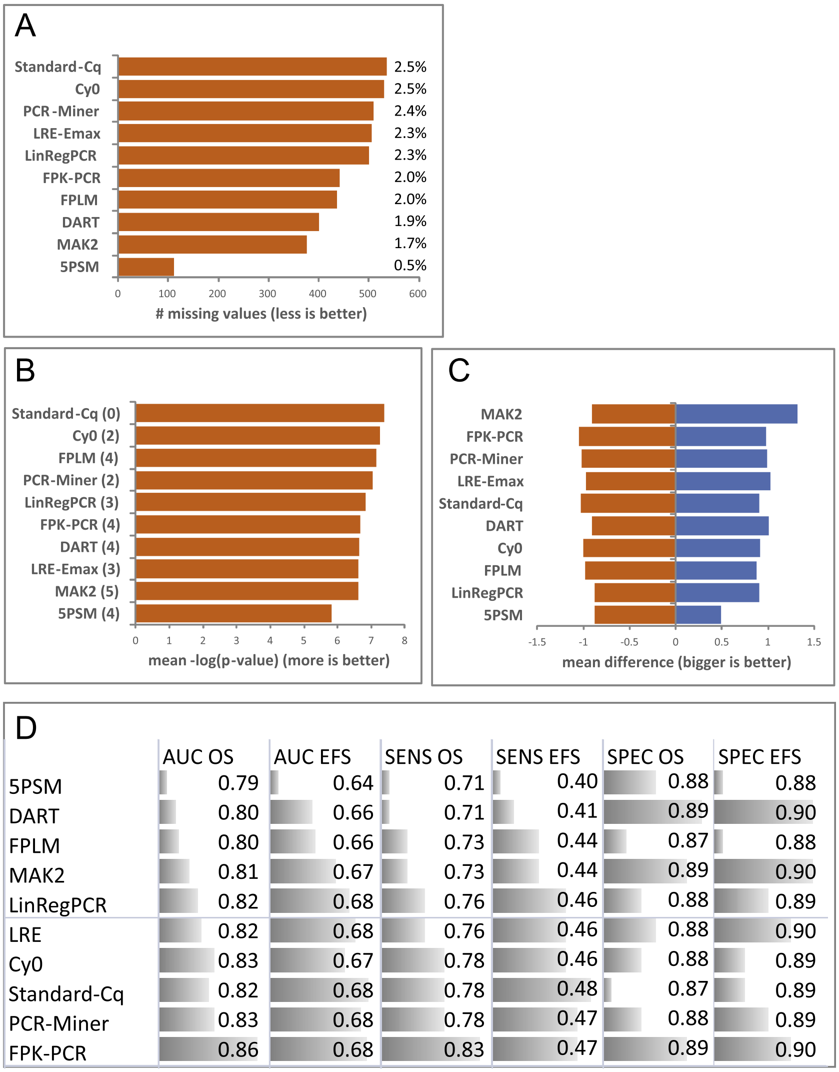


**Fig. 2.** Comparison of the results of the biomarker analysis. This analysis was done on the entire biomarker dataset including the 59 genes of interest as well as 4 reference genes measured in 366 tissue samples. (A) Number of missing values (out of a total of 59 $*$ 366 = 21,594 normalized measurements) generated in the biomarker dataset by the different curve analysis methods (*Note:* exclusion of target AluSq, as LRE-Emax and Cy0 did not have any result for this gene). (B) The mean of the $-\log_{10}(p)$ values of a Mann–Whitney test in which the normalized expression of the 59 biomarker genes are compared in neuroblastoma tumors from high-risk versus non-high-risk patients. Number between brackets indicates number of non-significant genes ($p > 0.05$). (C) The mean differences ($\log_{10}$ scale) of the genes lower expressed (orange) and higher expressed (blue) in high-risk vs. low-risk for the different methods. Methods are ordered by largest average difference. (D) Comparison of the patient classification performance based on the biomarker dataset among the tested curve analysis methods (AUC = area under the curve, SENS = sensitivity, SPEC = specificity, OS = overall survival, EFS = event free survival).

**Table 2**
Analysis of performance parameters per method. For each method, the mean rank (averaged over the 63 genes) is given for each of the performance indicators bias, linearity, precision, and resolution. Between parentheses is the rank of the methods' performance per indicator; the methods are sorted based on the average of these ranks. The Friedman test shows subsets of methods for which there is no evidence that they perform differently when all indicators are considered.

| Method | Bias | Linearity | Precision | Resolution | Mean rank | Friedman test subsets | | | |
|---|---|---|---|---|---|---|---|---|---|
| Cy0 | 1.98 (2) | 2.73 (1) | 3.13 (2) | 2.71 (2) | 1.75 | ■ | | | |
| LinRegPCR | 6.54 (6) | 3.57 (2) | 2.51 (1) | 2.63 (1) | 2.5 | ■ | ■ | | |
| Standard-$C_q$ | 1.92 (1) | 3.84 (3) | 3.71 (3) | 3.59 (3) | 2.5 | ■ | ■ | | |
| PCR-Miner | 5.98 (5) | 4.13 (4) | 4.43 4) | 4.17 (4) | 4.25 | | ■ | ■ | |
| MAK2 | 5.02 (3) | 4.63 (5) | 4.79 (5) | 4.71 (5) | 4.5 | | ■ | ■ | |
| LRE-E100 | 5.27 (4) | 4.79 (6) | 5.10 (6) | 4.95 (6) | 5.5 | | | ■ | |
| 5PSM | 8.97 (11) | 6.40 (7) | 6.27 (7) | 6.51 (7) | 8.0 | | | | ■ |
| DART | 8.43 (10 | 7.98 (8) | 7.95 (8) | 8.17 (8) | 8.5 | | | | ■ |
| FPLM | 7.84 (9) | 8.59 (9) | 8.81 (9) | 8.68 (9) | 9.0 | | | | ■ |
| LRE-Emax | 7.11 (8) | 9.14 (10) | 9.52 (10) | 9.52 (10) | 9.5 | | | | ■ |
| FPK-PCR | 6.94 (7) | 10.19 (11) | 9.78 (11) | 10.33 (11) | 10.0 | | | | ■ |

the inability of the method to properly process the amplification curve. Based on this comparison, the 5PSM method outperforms the other methods, generating the lowest number of missing values in the biomarker dataset.

Next, we investigated the impact of the curve analysis method on the significance of differential expression of the biomarker genes on the basis of the *p*-values. In Fig. 2B, the mean of the $-\log_{10}(p)$ values of a Mann–Whitney test, comparing high-risk versus non-high-risk patients, is visualized. Overall, there is not much difference between the methods, with 5PSM resulting in slightly lower significance values. The variance for all methods is similar (not shown). Also the mean expression differences (Fig. 2C) are very similar among the methods, with MAK2 appearing to have an overestimation of the magnitude of the over-expressed genes in the high-risk group, and 5PSM an underestimation for the same set of genes.

Finally, the effect of the curve analysis method on patient classification performance was investigated (Fig. 2D). The sensitivity refers to the ability of the gene expression classifier to correctly identify those patients that will die (overall survival) or relapse (event-free survival). The specificity refers to the ability of the gene expression classifier to correctly identify those patients that will survive or show no relapse. The area under the (receiver operator) curve (AUC) represents the overall accuracy of the classifier. The classification results are quite similar among the methods. Overall best classification performance was observed for the FPK-PCR method.

## 3.2. Performance indicators based on the biomarker data set

For each of the performance indicators derived from the analysis of the concentration series per gene, line graphs were prepared to show the results per gene and method. Each of these line graphs contains the result of the Friedman test and the subsets of similarly performing methods. Table 2 summarizes the results of the Friedman tests for each of the performance indicators. A low mean rank indicates better performance of that method. Box-and-whisker plots were prepared to illustrate the distribution of each indicator per method (Fig. 3).

### 3.2.1. Efficiency

The range of efficiency values per method shows that efficiency values differ between genes in value as well as in variation (Fig. 4). This variability is the sum of differences in efficiency between genes and differences that result from the estimation method. Therefore differences in variability between methods cannot be interpreted. Apart from DART and FPLM, which have overlapping distributions, all methods result in different median *E* values (Fig. 3A). Some methods have a substantial number of efficiency values that are above 2. For all methods the observed efficiency values (Fig. 3) are significantly different from the values derived from the conventional standard curve (Standard-$C_q$) or the standard curve approach based on Cy0 values (Cy0) which, for the 63 genes, both estimate a median efficiency of 1.95 (IQR: 1.92–1.97).

### 3.2.2. Bias

Bias is defined as the deviation of the observed values from the expected values. The expected value for the ratio between the mean $F_0$ of highest and lowest input is 10,000. Fig. 3B and Supplemental Fig. S1 show that only the standard curve based methods (Standard-$C_q$ and Cy0) reach such observed ratios. However, this does not mean that those methods are unbiased. Those two methods calculate the efficiency value from the slope of the relation between $C_q$ (or Cy0) and the log(input). This efficiency value and the same $C_q$ or Cy0 values are then used to calculate $F_0$. Claiming that these methods are unbiased is thus the result of a circular reasoning.

The other curve analysis methods are consistently positively or negatively biased (Fig. 3B and the cumulative graph in Supplemental Fig. S2). The variation between genes makes that this bias will not easily be canceled out by normalization with reference genes. Methods that show a wide variation in bias between genes will therefore suffer from variation in gene expression ratios.

The regression line fitted to the scaled and log-transformed input and $F_0$ data should have a slope of 1 when the method is unbiased. Indeed, the graph of these slope values per gene and analysis method (Supplemental Fig. S3) is very similar to that of the ratio of the mean observed $F_0$ for the highest and lowest input (Supplemental Fig. S1).

### 3.2.3. Precision

The concentration series was measured in triplicate and the resulting fluorescence data were analyzed. Therefore, the variance within these triplicates should be small, only reflecting random variation in laboratory procedures and fluorescence measurement, and this variance should be the same for every analysis method. The resulting within-triplicate variance can be considered a measure for the precision of the analysis method (Fig. 3C and Supplemental Fig. S4). The within-triplicate variance is strongly method-dependent. The Friedman test shows that the six methods that show low variability (LinRegPCR, Cy0, Standard-$C_q$, PCR-Miner, MAK2, and LRE-E100) form overlapping subsets of two or three methods (Supplemental Fig. S4). The other methods show up to 5 times as much within-triplicate variance.

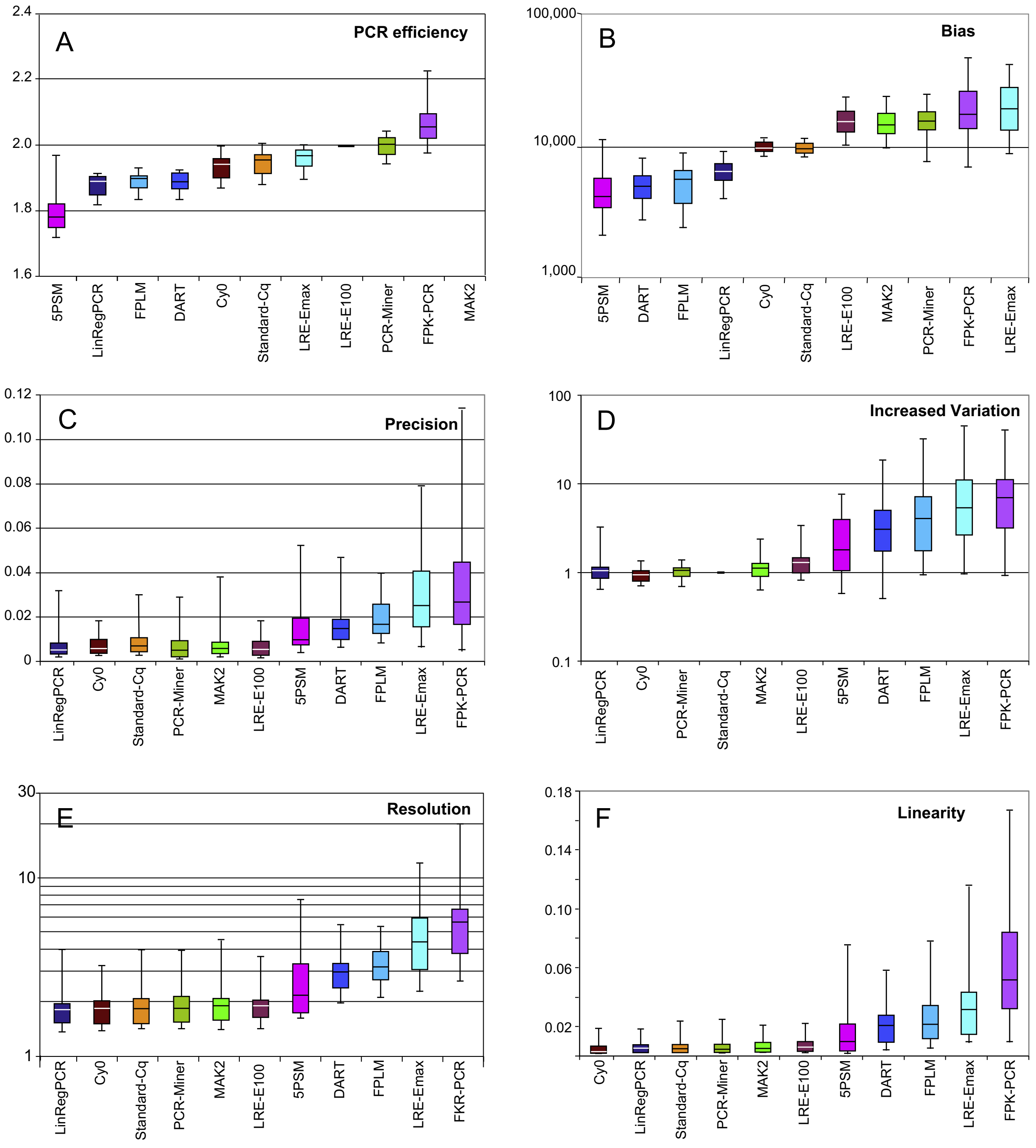


**Fig. 3.** Performance indicators per method. The performance indicator values determined from the concentration series included in the measurement of the 63 genes are summarized in box-and-whisker plots. The boxes range from the 25th to the 75th percentile and are divided by the median; the whiskers are set at the 5th and 95th percentile. (A) PCR efficiency. For LRE-E100 all values are 2; MAK2 does not determine PCR efficiency. No expected value can be defined. (B) Bias is determined as ratio of the mean $F_0$ value observed in the highest and the lowest input. The expected value is 10,000. (C) Precision or reproducibility is determined as the within-triplicate variance and should have the same, low, value in all methods. (D) Increased or reduced variation compared to conventional standard curve–$C_q$ analysis. Values below 1 indicate reduced variance. (E) Resolution defined as the fold-chance that would result in the detection of a difference at a 5% significance level; lower is better. (F) Linearity is defined as the variance due to the deviation of the mean of the observed triplicate $F_0$ values from the regression line; lower is better. See Supplemental Fig. S1–S8 for the performance indicator values per gene, the Friedman test results and subsets of similarly performing methods.

### 3.2.4. Increased of reduced variance

The ratio of the within-triplicate variance of each method and the variance resulting from the conventional standard curve derived efficiency and $C_q$ method (Standard-$C_q$ in Fig. 3C) was calculated to show the increased or reduced variance for each method compared to this conventional approach (Fig. 3D and Supplemental Fig. S5). A ratio of 1 would mean that the method results in the same variance as the conventional method; a value below 1 would indicate that the method results in reduced variance whereas a value above 1 indicates increased variance. The results show that, although all methods can reach the same or better precision for some genes, only 3 methods, Cy0, LinRegPCR and PCR-Miner do so for 50% of the genes; 5PSM, MAK2 and LRE-E100 reduce variance for 20% of the genes (Supplemental Fig. S6). FPK-PCR, FPLM, DART and LRE-Emax display hardly any reduced variance and an up to 30-fold increase instead.

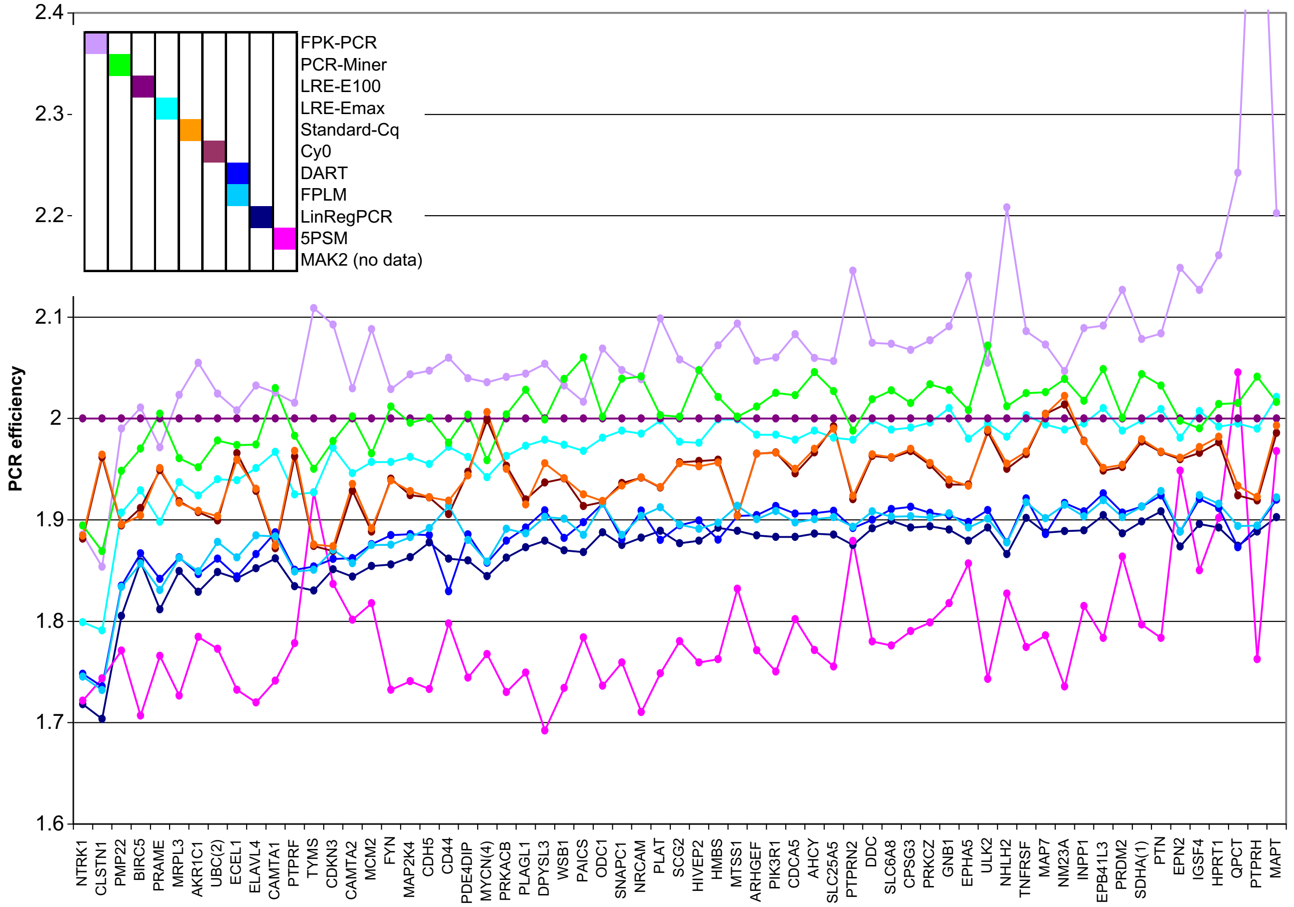


**Fig. 4.** PCR efficiency per gene and curve analysis method. Distribution of amplification efficiency values per method. For methods that determine the efficiency per reaction the mean of those efficiency values per gene is plotted. The legend panel provides the color scheme as well as the results of the Friedman test; when two methods do appear in the same column they do not estimate significantly different efficiency values; only FPLM and DART form such a subset, all other methods differ significantly. Genes are ordered by increasing geometric mean efficiency.

### 3.2.5. Resolution

The total variance around a linear regression line can be used to construct the confidence interval (CI) of this line; data points outside the 95% CI are considered to deviate significantly from this line. According to this reasoning an observation outside the 95% CI of the regression line fitted to the concentration series would be judged to be significantly different from the observed output. By converting the width of the CI into fold-difference from the fitted line, the detectable fold-difference was calculated per gene which can be considered a measure for the resolution of the method.

The resolution (Fig. 3E and Supplemental Fig. S7) varies per method. The cumulative distribution plot (Fig. 5) shows for each method which percentage of the observations reaches a specific detectable difference. LinRegPCR and Cy0 form a first subset with highest resolution followed by Standard-$C_q$, PCR-Miner, MAK2 and LRE-E100 that form overlapping subsets. With these 6 methods a 2-fold observed difference would be significant for about 75% of the genes. For most other methods the resolution lies between the 2 and 3-fold-difference. However, for FPK-PCR and LRE-Emax a substantial fraction of the genes require at least a 4-fold difference.

Note that this analysis only includes the technical variation of the qPCR analysis. In a real biomarker expression study, additional technical variation related to normalization and additional biological variation will be the limiting factor in the resolution of the method [46]. However, because the current comparison of analysis

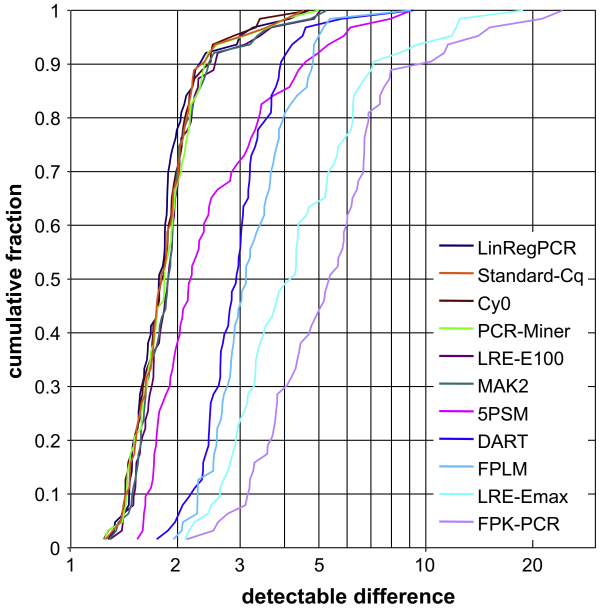


**Fig. 5.** Resolution per method. The cumulative graphs of the difference that would be significant at the 0.05 level are shown per method. For a group of 6 methods (Standard-$C_q$, Cy0, LinRegPCR, MAK2, PCR-Miner and LRE-E100) a difference of 2-fold would be significant for 70% of the genes. See Supplemental Fig. S7 for resolution values per gene.

methods uses the same datasets for each method, the differences shown in Fig. 3E can be used to compare the effect of the analysis method on the this performance indicator.

### 3.2.6. Linearity

The distance between the mean of the observed $F_0$ values per concentration and the expected value on the regression line is considered a measure for the linearity of the input–output relation (Fig. 3F and Supplemental Fig. S8). A high value means that the regression line does not fit the observed $F_0$ values per concentration. This linearity measure is related to the variance of $F_0$ values, but a high within-variance can occur together with a low deviation from linearity when the $F_0$ values are far apart but their mean is on the regression line. The lowest values are observed for Cy0 and LinRegPCR; the latter forms overlapping subsets with Standard-$C_q$, PCR-Miner, MAK2 and LRE-E100 (Supplemental Fig. S8). Of the other methods, only DART, FPLM and LRE-Emax show overlapping subsets.

Note that the within-triplicate variance and the variance due to deviation from regression, together form the total variance around the regression line which is a factor in the calculation of the correlation coefficient. Although the latter is often used to judge linearity, it does in fact only describe the fit of the line in terms of total residual variation. The observed correlation coefficients ($r$) per gene and method are shown in Supplemental Fig. S9.

## 3.3. Performance based on the technical datasets

### 3.3.1. 380-replicates set

The 380-replicates set was included in the comparison of curve analysis methods to determine the precision of the analysis methods. Because all methods analyze the same fluorescence dataset, with its small random technical variation, a difference in the distribution of observed indicators is due to variation introduced by the analysis method. The 380-replicates set could not be analyzed by LRE-qPCR. The Cy0 method did generate a Cy0 value but could not calculate an $F_0$ value because no dilution series was present to determine an efficiency value.

*3.3.1.1. Efficiency.* Individual PCR Efficiency values were determined by 6 methods. The observed efficiency values vary significantly in median value as well as variability (Fig. 6A). DART and FPLM show the same low variation in efficiency values.

$C_q$. To include the Cy0 method into the analysis of the 380-replicates set, a graph was prepared to show the variance in $C_q$ and Cy0 values (Fig. 6B). The Cy0 values show the least variance, followed by a subset including LinRegPCR, 5PSM and FPK-PCR.

*3.3.1.2. Target quantity.* The observed $F_0$ values were scaled to set the median to 1 which enabled the comparison of the $F_0$ values that were determined on different scales (Fig. 6C). The lowest variability in $F_0$ values was observed in PCR-Miner, followed by a subset of LinRegPCR and MAK2; the other methods show significantly more variance.

The analysis of the 380-replicates series shows that, in methods that use efficiency values per reaction, the low variance in $C_q$ values combined with the variance in efficiency values can lead to high variance in $F_0$ values. On the other hand, PCR-Miner and LinRegPCR, with intermediate variance in $C_q$ values, and using a mean efficiency value, reach low variance in $F_0$ values.

### 3.3.2. 94-replicates-per-4-point 10-fold dilution set

The 94-replicates-4-dilutions set was included to compare variability but could also be used to determine other performance indicators. Similar to the analysis of the concentration series in the biomarker datasets, this dilution series with an extended number of replication per concentration ($n = 94$) was used to compare variation in efficiency values, bias, reproducibility, linearity and resolution between methods.

*3.3.2.1. Efficiency.* The distribution of individual efficiency values, determined per reaction, is available for 6 methods (Fig. 6D). Although the methods differ with respect to the median efficiency value, all methods show a narrow distribution. The lowest variance is observed in FPK-PCR and FPLM, with significantly increasing variance in the other methods.

*3.3.2.2. Target quantity: bias, linearity and precision..* $F_0$ values in the 94-replicates-4-dilutions set were determined by 8 analysis methods (Fig. 6E). The variance in $F_0$ values is lowest in LinRegPCR and increases significantly for all other methods; DART and 5PSM show similar high variance. The deviation of the mean $F_0$ from the expected value (Fig. 6E; horizontal lines) shows systematically positive or negative bias per analysis method. The least bias is observed in Cy0, PCR-Miner, LinRegPCR and MAK2 (Fig. 6F); FPK-PCR displays a strong underestimation of $F_0$ values whereas 5PSM shows a strong overestimation.

The methods with the highest precision reach the highest resolution (Fig. 6F). However, despite its high variance, DART displays good linearity; the mean values per dilution are as close to the regression line as those of LinRegPCR. Supplemental Fig. S10 shows the log(input) versus log($F_0$) graph per method which includes all 4 times 94 data points. The graphs clearly illustrate the differences in bias, precision and linearity.

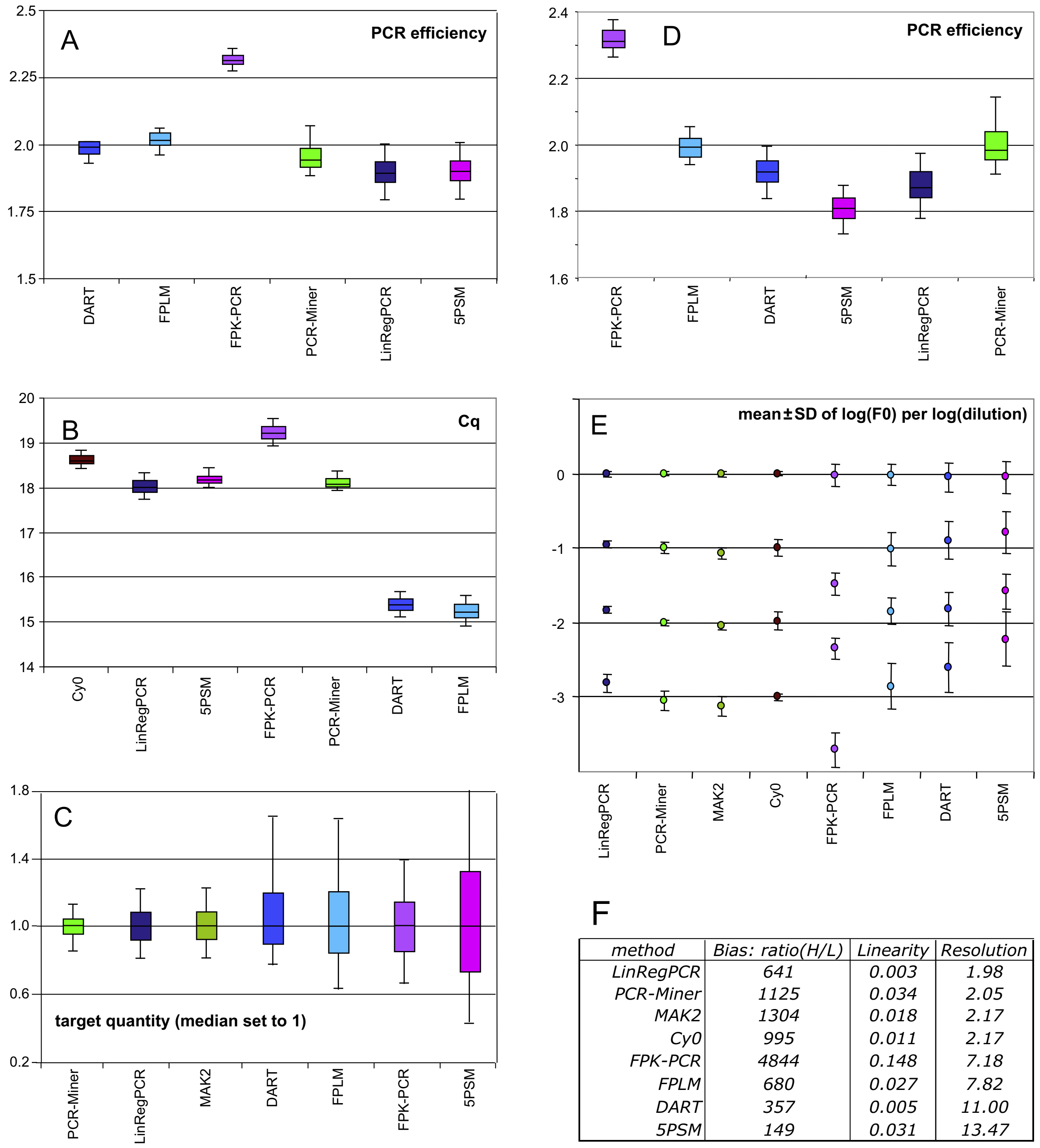
### 3.3.3. Competimer dataset

The competimer dataset was included to compare the degree to which the different methods can handle (known) variations in efficiency values. This approach to manipulate the amplification efficiency is artificial; in field samples inhibitors are more likely to interact with the polymerase. However, competitive primer binding is currently the only way to change PCR efficiency in a predictable way.

The methods differed strongly in the number of resulting missing values (Supplemental Table S1), ranging from 1 (MAK2) to 34 (FPLM and FPK-PCR) (from a total of 126 reactions; 21 observations for each DNA dilution, 18 for each competimer concentration). Of note, missing values occurred in the ‘early’ amplification curves with high DNA input and un-inhibited PCR efficiency. When primer competition lowered the efficiency value most methods were able to handle the displaced amplification curves.

*3.3.3.1. Efficiency.* The variance in the observed efficiency values depended on the analysis method and the competimer percentage (Fig. 7A) with lowest variances in LinRegPCR and DART. Most methods show more variable efficiency values when more competimer is present. The ratio between observed and expected efficiency depends on the analysis method, competimer percentage and DNA input (Fig. 8). No method showed the expected ratio of 1. For all methods, the observed efficiency was too low and this deficit increased with increasing competimer percentage, leading to decreasing ratios for almost every method and cDNA input.

*3.3.3.2. Target quantity.* Because of the missing values in the highest DNA inputs for some methods the analysis of the performance indicators using the observed $F_0$ values is restricted to the DNA inputs of 4 ng/μl and lower. The mean observed $F_0$ value for the DNA input of 4 ng/μl was set to 1 and transformed to log (base 10). To illustrate the variance and bias per method the correlation coefficient and slope of the regression line are shown in Fig. 7B. For these scaled and log-transformed data a slope of close to 1 indicates ‘no bias’ which is the case for most methods except MAK2. The variation in bias increases with decreasing correlation coefficient.



| method | Bias: ratio(H/L) | Linearity | Resolution |
|---|---|---|---|
| LinRegPCR | 641 | 0.003 | 1.98 |
| PCR-Miner | 1125 | 0.034 | 2.05 |
| MAK2 | 1304 | 0.018 | 2.17 |
| Cy0 | 995 | 0.011 | 2.17 |
| FPK-PCR | 4844 | 0.148 | 7.18 |
| FPLM | 680 | 0.027 | 7.82 |
| DART | 357 | 0.005 | 11.00 |
| 5PSM | 149 | 0.031 | 13.47 |

**Fig. 6.** Technical datasets. Performance indicator analysis in the 380-replicates set (A–C) and the 94-replicates-4-dilutions set (D–F) are shown. Methods are ranked according to decreasing performance; in these datasets low variance means better performance. (A) Distributions of observed efficiency values. MAK2 does not estimate an efficiency value. *F*-tests show that only DART and FPLM have similar low variance. (B) Distribution of observed $C_q$ (or Cy0) values. MAK2 does not estimate a $C_q$ value. Variances do not differ between LinRegPCR, 5PSM and FPK-PCR. (C) Variability in observed $F_0$ values. Cy0 could not determine the $F_0$ value because there was no dilution series available. MAK2 and LinRegPCR have a similar intermediate variance. (D) Distribution of observed efficiency values. FPLM and FPK-PCR show similar variance. (E) Mean observed $F_0$ value per input concentration and method (highest input and mean observed $F_0$ both set to 1). The *y*-axis is log (base 10) of the scaled input. Variance increases between all methods, except for DART and FPLM. (F) Bias, linearity and resolution per method. The expected value for bias is 1,000; linearity and resolution should both be as low as possible.

For further specification the variance was calculated per DNA input and competimer percentage (Supplemental Fig. S13). These results show a large variance in FPLM, 5PSM, DART and FPK-PCR and lower variation in LinRegPCR, MAK2 and PCR-Miner. Note, however, that with respect to the mean $F_0$ per cDNA MAK2 shows a large variation in low cDNA inputs because its bias is dependent on the competimer percentage (Supplemental Fig. S12).

The log(input)-log($F_0$) plots (Supplemental Fig. S11) show a clear difference in linearity and reproducibility between methods. The results obtained for different competimer percentages are combined in these graphs. Some methods (LinRegPCR and PCR-Miner) show non-linearity (underestimating the highest input). MAK2 shows separation between competimer percentages resulting in a series of diverging lines.

The log(competimer percentage)-log($F_0$) plot (Supplemental Fig. S12) should show a horizontal line for each input concentration. For some methods this configuration is observed for part of the input concentrations. However, most methods perform worse

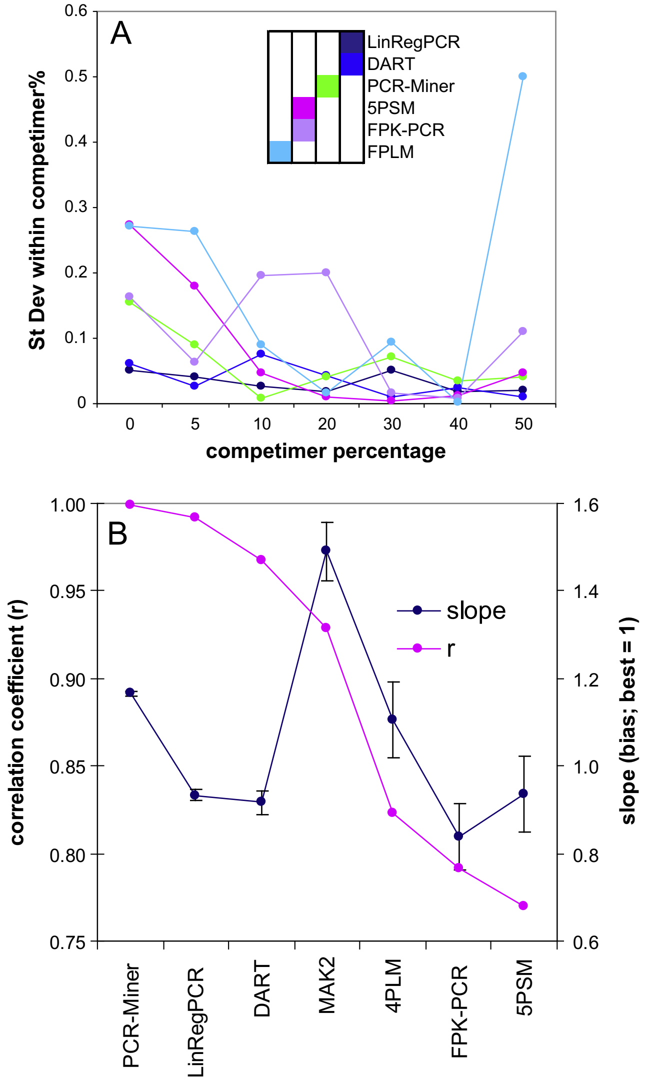


**Fig. 7.** Analysis of the competimer dataset. (A) Variance in observed efficiency value per competimer percentage and analysis method for cDNA input 4 ng/μl. The other cDNA inputs show a similar distribution. The within-replicate variance was pooled to determine total variance and the standard deviation derived from this total variance was displayed. The *F*-tests results show similar variance for LinRegPCR and DART and for 5PSM and FPK-PCR. (B) Correlation coefficient (*r*) and slope of the regression line fitted to the observed $F_0$ values (mean $F_0$ for cDNA input 4 ng/μl set to 1). Large variance in bias, as shown by the displayed standard error of the slope, correlates with a low correlation coefficient.

for the lowest and highest inputs. MAK2 shows a downward trend with increasing competimer concentration for every input.

## 4. General discussion and conclusions

Ten different curve analysis methods were applied to a large biomarker dataset as well as three unpublished technical datasets to determine their performance. All methods were applied by their original developers and with the latest implementation of the algorithms. Performance in the biomarker dataset was compared with respect to the number of missing values, differences in single gene expression level between high and low risk groups and the ability to classify the patient population into known risk groups based on a 59 multi-gene transcript signature. Additionally, the concentrations series of all 63 genes included in this set, as well as the technical datasets were used to determine bias, reproducibility, linearity and resolution.

When it comes to patient classification accuracy and significance and magnitude of differential expression, all tested methods perform relatively similar. This is in great part due to the fact that results are normalized using multiple stably expressed reference genes and that results are either averaged across a large set of 59 biomarkers, or the result of classification using a multi-gene signature of 59 genes. As such, variable or suboptimal results for specific genes may be compensated by normalization or averaging across different genes. This is encouraging news and indicates that the chosen curve analysis method may not have great impact on relative quantification accuracy.

The performance comparison shows that for each indicator there are large differences between genes. This illustrates that a large number of targets or datasets are required to really appreciate the performance of a given curve analysis method. The currently used datasets are made available to allow future tests and comparisons (http://qPCRDataMethods.hfrc.nl).

Overall, methods that use one efficiency value per gene perform better than methods that calculate $F_0$ with an efficiency value per reaction. A notable exception is MAK2, a method that does not depend on calculation of an efficiency value. From a mathematical perspective, this result is expected as variance in efficiency is propagated more severely than variance in $C_q$ values. Hence, similar variance in efficiency and $C_q$ values will result in greater variance of the $F_0$ quantities for the former (Supplemental Fig. S14).

### 4.1. Comments of original developers

#### 4.1.1. CAmpER: DART and FPLM

Given that DART and FPLM are the oldest approaches used in this comparison, both being initially published in 2003, the performance of both methods is surprisingly good, although they are clearly outperformed by more recent approaches. In the analysis of the biomarker dataset both DART and FPLM show results comparable to the other approaches. In the analysis of the concentration series of this data the variance of calculated efficiencies is quite low, but both approaches show an overall too high variance and therefore are prone to add bias. One reason is the use of static $F_q$ values, as these are, to a certain degree, arbitrary and thus add bias to the calculation. A second reason is the use of single sample efficiencies, which seem to be inferior to mean efficiencies for replicates as used in LinRegPCR or PCR Miner. These two issues may be addressed in future CAmpER versions.

#### 4.1.2. Mak2

MAK2. quantification of qPCR data is an a curve analysis method that does not assume constant amplification efficiency per gene, yet performs as well as the methods that make this assumption. While other model-fitting methods employ empirical models (such as a sigmoidal or exponential), used because they appear to have the same shape as qPCR data, the MAK2 model is derived from the molecular mechanism of the low temperature step of the polymerase chain reaction. Only a mechanistic model could be used to accurately infer behavior in early qPCR cycles, where qPCR signal is dominated by noise, from behavior observed when qPCR signal is measurable. It is therefore not surprising that MAK2 quantification outperformed the empirical model-fitting methods evaluated.

Theoretical analysis of the polymerase chain reaction, from which MAK2 was derived, reveals that amplification efficiency is not constant throughout PCR. The amplification efficiency of a given cycle of PCR is determined by competition between the annealing of primer to its target and the reannealing of complementary DNA strands. As target DNA concentration builds up, amplification efficiency declines. The efficiency that can be obtained from a $C_q$ standard curve is an average efficiency that works reasonably well for quantification when PCR is assumed to follow

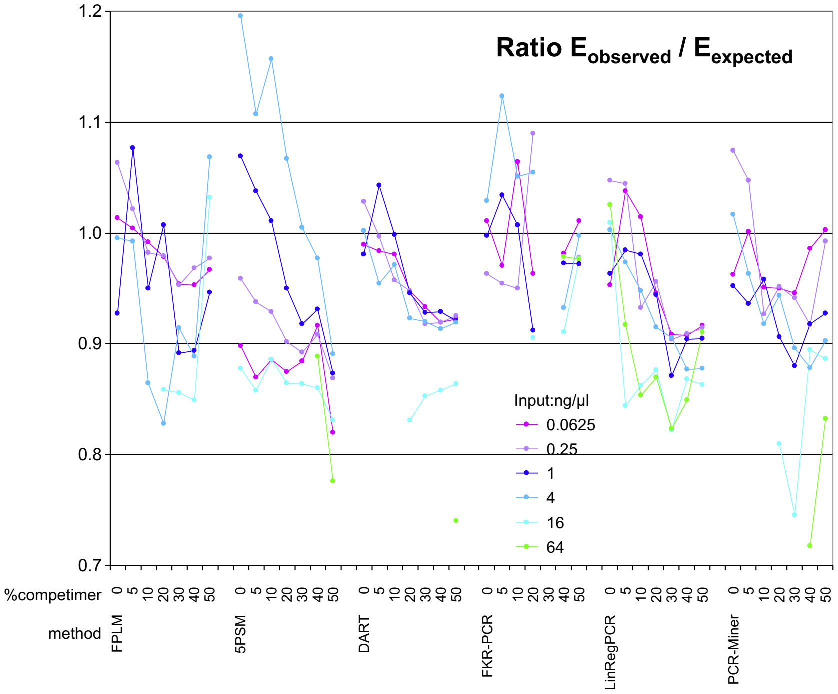


**Fig. 8.** Observed and expected PCR efficiency value. The ratio between observed and expected PCR efficiency was calculated and its mean was calculated per competimer percentage and DNA input. The expected ratio is 1. In general, the observed/expected ratio is low for high competimer percentages and/or high DNA inputs.

exponential behavior. The development of the MAK2 model is significant because MAK2 quantification liberates qPCR users from reliance on amplification efficiencies derived from $C_q$ standard curves.

While MAK2 quantification provides reliable estimates of target DNA concentration in a sample under normal qPCR conditions, MAK2 quantification does not reliably quantify target concentration for qPCR assays with competimers. This is because competimer assays violate one of the assumptions of the MAK2 model: the assumption that primer concentration can be assumed constant, due to the overwhelming abundance of primers, throughout the exponential phase of qPCR. Competimers, which compete for binding sites with primers, cause increasing inhibition with each PCR cycle because primers are consumed during PCR while competimers are not, so that competimers increase in concentration relative to primer concentration. Competimers thus accelerate the decline in amplification efficiency over that which occurs under normal PCR conditions. To accurately quantify competimer data using the MAK2 model, the model would need to be modified in order to take real-time primer and competimer concentrations into account.

### 4.1.3. FPK-PCR

The development of the FPK-PCR approach focused on improved estimation of PCR efficiency as it evolves over the amplification cycles. Results from the competimer dataset indeed confirm that this approach is highly suitable for detecting kinetic outliers (inhibition) and that its elevated efficiency estimates are not precluding their use in comparing reactions. The performance of FPK-PCR in the analysis of the biomarker data further illustrates its potential as a valuable tool for qPCR data analysis. Upon inspection, increased variability of the FPK-PCR results in the other datasets is associated with imprecise estimation of the initial number of target copies, a parameter outside the focus of estimation in our original development. Current implementation relies heavily on the assumption that all changes in reaction fluorescence are due to the amplification process. Any alternative process that adds variation to the final observed fluorescence (i.e. plateau variability) therefore translates into additional variation of these $F_0$ estimates. This explains why the FPK-PCR performs poorly in terms of variance while the bias is on target. An advanced $F_0$ estimation method capable of discounting the extra source of variation is under development as part of a planned update of the FPK-PCR algorithm.

### 4.1.4. LinRegPCR

Because LinRegPCR uses a baseline estimation that is aimed at reconstructing an exponential phase in which the data points are on a straight line, the PCR efficiencies derived from these data points are less variable [38]. Using the mean efficiency per amplicon further reduces the variance. The performance of LinRegPCR shows that, at least till the start of the plateau phase the assumption that the PCR efficiency is constant is not violated by the expected decrease due to limiting reaction conditions. The determination of the start of the plateau phase enables the use of data points in the late exponential phase which avoids ground phase noise. The large number of missing values in the biomarker dataset is partly due to the fact that LinRegPCR assigns a default missing value to reactions that do not show amplification or do not reach the plateau. The interface should be extended with an option to enable the imputation of a large $C_q$ value. The strong non-linear behavior in the competimer dataset, when all inputs are considered, is mostly due to the very early appearance of these amplification curves. Because LinRegPCR does not use the ground phase cycles for baseline estimation it can estimate a baseline for such early curves. However, on the cost of a biased $F_0$ value.

### 4.1.5. LRE qPCR

LRE qPCR originated from the recognition that amplification efficiency progressively decreases during PCR amplification, and that this loss is directly proportional to the mass of amplicon DNA present at the beginning of each cycle performed. In addition to refuting the exponential nature of PCR amplification [27], the application of optical calibration allows absolute quantification to be conducted without the need to construct target-specific standard curves [47,48]. Development of an open source program that automates LRE qPCR [42] further provides the ability to conduct large-scale absolute quantification with little or no user intervention. While absolute quantification was not a consideration in this study, it has substantive implications, particularly for gene expression profiling, in that, among other attributes, it generates universally comparable data, an ability that is difficult to achieve using conventional qPCR methodologies [11]. In addition to the performance evaluation conducted in this study, earlier studies have demonstrated that LRE qPCR has the ability to routinely generate absolute accuracies of ±15–25% and has the ability to maintain accuracy down to a single target molecule [47,48], capabilities that may not be evident in this study due to the inherent limitations of relative quantification.

#### 4.1.6. Cy0

The Cy0 value has been defined as the intersection point between the abscissa axis and the tangent of the inflection point of the Richards curve obtained by the non-linear regression of raw data. Indeed, the shapes of amplification curves in qPCR range from a perfect sigmoidal to a strongly asymmetric shape, for example in presence of inhibitors [49]. More simply, the Cy0 method is a threshold-based method like $C_t$ but with the key difference that the threshold value is dynamic and depends on amplification kinetic and possibly it should compensate for small variations among the samples to be compared. This is a method in which the stability and reliability of a standard curve approach is combined with a fitting procedure to overcome the key problem of PCR efficiency determination in real-time PCR nucleic acid quantification. Cy0 values were calculated using a web interface (http://www.cy0-method.org) specifically developed by the authors for the analysis in the current paper. The data reported herein show that the Cy0 method is a valid alternative to the Standard-$C_q$ method for obtaining reliable and precise nucleic acid quantification even when amplification efficiency differs slightly between reactions.

#### 4.1.7. 5-parameter sigmoidal model (5PSM)

The five-parameter model has an additional parameter that can account for asymmetrical structures in qPCR curves, twisting the fitted curve around the point of inflection and delivering significantly better sigmoidal fits to the curve. This might be the reason that this method delivers the lowest number of missing values (Fig. 2A). However, improving the fit of the curve by adjusting the asymmetry parameter also affects other parameters of the curve such as the second derivative maximum and the slope. This results in relatively low efficiency values (Figs. 3A, 6A and D). Even more, the increased fitting performance (sensitivity) in the exponential region results in a relative high variability in efficiency estimation in this region, which essentially results in increased bias and variability in the estimation of $F_0$ and delivers a relatively poor performance in a dilution experiment setup.

#### 4.1.8. PCR Miner

Similar to the LinRegPCR method, PCR Miner also implemented additional pre-evaluation functions in software to automatically exclude the samples for efficiency estimation with bad exponential phase (not fitting exponential model at all), not reaching plateau (too little input or too less total cycles), or amplification happens too early (too much input). The benefits of this strategy are that it results in more reliable averaged efficiency (per gene) and good $F_0$ estimation since all used individual efficiencies are only from the typical complete PCR curve, although on the cost of larger number of missing values. For those excluded samples, refining the experiment (e.g. DNA inputs, total cycle number) is recommended. For the PCR competimer test, since only a small number (~8) of available points within exponential phase can be used for individual efficiency calculation, variation in efficiencies from only triplicate samples is very likely to result in a considerable effect on $F_0$ because any error in the measured efficiencies will be exponentially magnified. Using more replicates ($n \geqslant 6$) to calculate the mean of the efficiency within each competition group to acquire a comparable $F_0$ is advised.

### 4.2. Overall conclusions

A large scale and systematic analysis of amplification curve analysis methods with respect to their bias, precision, linearity, resolution and impact on transcriptional biomarker identification is currently lacking in the literature. Here we provide an analytical framework that enables assessment of multiple qPCR curve analysis method performances using testable hypotheses.

Large differences are noticeable among methods, targets, and samples. While we describe trends and rank methods with respect to the various performance indicators, we cannot explain all observations. It is clear that methods that use reaction-specific PCR efficiencies generally perform less accurate, due to higher sample specific variability, compared to methods that average the efficiency across all reactions per target gene. However, we have no clear idea what other factors contribute to differential performance. There appears to be a trend that methods that model more parameters are generally more imprecise, but this observation is confounded by the way efficiency is handled.

The strengths of this study are the large number of tested curve analysis methods, the use of both real clinical samples and technical performance datasets, the inclusion of reactions with predictably lower PCR efficiencies, and the use of different instruments and PCR mixes. The limitations of the study are that most assays are of relatively high quality in terms of efficiency, sensitivity, and specificity. This may render the conclusions of the study less representative when data is coming from suboptimal assays having low PCR efficiencies and high variations. Another limitation is that this study has not included samples that are spread across runs, so potential run dependent bias and inter-run variability are not addressed. This is relevant as some methods use run dependent settings, the validity of which was not tested here.

Based on the results and experiences gained from this study, we provide a few recommendations for future studies in this field. Even if this study included 2 different qPCR instruments and 3 different PCR mixes, this is not nearly representative and future studies should include more instances, in order to better assess the robustness and general applicability of the analysis method. Particularly, datasets with lower fluorescence dynamics based on probe detection chemistries should be included as only SYBR Green I datasets were tested here. Another recommendation that we may provide is that novel analysis methods or re-assessments of old methods should indicate in more detail what type of data (instrument, PCR mix, detection chemistry, number of assays) was used to develop and validate the method. This may help assessment of potential robustness of the method and give pointers to users outside the tested ranges. The provided analytical framework in this study should certainly help the evaluation of these future studies and make results comparable with ours.

Finally, it was not the aim of this study to acclaim one particular curve analysis method with best overall performance. Furthermore, choice of one method may depend on the set goals of the quantification study whereby one performance indicator should deserve more weight. Nevertheless, we believe accuracy is generally more important than precision because imprecision can be overcome by running more replicates. We hope the current study may help users to select the ideal method for their studies and developers to modify and improve their methods with testable hypotheses.

## 5. Datasets

The datasets used in this study are made available (http://qPCRDataMethods.hfrc.nl).

An additional compilation of 25 published datasets acquired with different qPCR platforms and chemistries that are useful in testing novel methods and algorithms can be obtained from http://www.dr-spiess.de/qpcR/datasets.html. These datasets have been preformatted into Excel files and a link to the original references is given.

## Acknowledgements

M.W.P. acknowledges Ales Tichopad for the support during the development of FPLM. G.B. would like to thank John SantaLucia, Jr. and Zheng Ouyang for their contributions to the development of *qPCR AutoFit* and Andrej-Nikolai Spiess for development of MAK2-fitting in the R package *qpcR*. J.M.R. thanks Maurice van den Hoff and Antoon Moorman for their support during the development of LinRegPCR. S.Z. thanks Prof. Russell D. Fernald for support when developing Real-time PCR Miner. A.N.S. was supported by Grant Sp721/1-3 of the German Research Foundation (DFG). A.L. would like to thank the GMOlab section of the Institute of Public Health in Brussels. J.V. and A.L. acknowledge the support of Ghent University's Multidisciplinary Research Partnership ''Bioinformatics: from nucleotides to networks''. S.D. acknowledges the Institute for the Promotion of Scientific Research Flanders (IWT). D.S. thanks to Dr. Michele Guescini for the support.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2012.08.011.

## References

[1] S.A. Bustin, V. Benes, J.A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M.W. Pfaffl, G.L. Shipley, J. Vandesompele, C.T. Wittwer, Clin. Chem. 55 (2009) 611–622.
[2] R. Higuchi, C. Fockler, G. Dollinger, R. Watson, Biotechnology (N.Y.) 11 (1993) 1026–1030.
[3] S.A. Bustin, V. Benes, T. Nolan, M.W. Pfaffl, J. Mol. Endocrinol. 34 (2005) 597–601.
[4] H.D. VanGuilder, K.E. Vrana, W.M. Freeman, Biotechniques 44 (2008) 619–626.
[5] E.E. Creemers, A.J. Tijsen, Y.M. Pinto, Circ. Res. 110 (2012) 483–495.
[6] A.J. Tijsen, E.E. Creemers, P.D. Moerland, L.J. De Windt, Circ. Res. 106 (2010) 1035–1039.
[7] A. Fendler, M. Jung, C. Stephan, R.J. Honey, R.J. Stewart, K.T. Pace, A. Erbersdobler, S. Samaan, K. Jung, G.M. Yousef, Int. J. Oncol. 39 (2011) 1183–1192.
[8] K. de Preter, P. Mestdagh, J. Vermeulen, F. Zeka, A. Naranjo, I. Bray, V. Castel, C. Chen, E. Drozynska, A. Eggert, M.D. Hogarty, E. Izycka-Swieszewska, W.B. London, R. Noguera, M. Piqueras, K. Bryan, B. Schowe, R.R. van, J.J. Molenaar, A. Schramm, J.H. Schulte, R.L. Stallings, R. Versteeg, G. Laureys, R.N. Van, F. Speleman, J. Vandesompele, Clin. Cancer Res. 17 (2011) 7684–7692.
[9] R. Belzeaux, C. Formisano-Treziny, A. Loundou, L. Boyer, J. Gabert, J.C. Samuelian, F. Feron, J. Naudin, E.C. Ibrahim, J. Psychiatr. Res. 44 (2010) 1205–1213.
[10] S.E. Larkin, S. Holmes, I.A. Cree, T. Walker, V. Basketter, B. Bickers, S. Harris, S.D. Garbis, P.A. Townsend, C. Aukim-Hastie, Br. J. Cancer 106 (2012) 157–165.
[11] J. Vermeulen, K. De Preter, A. Naranjo, L. Vercruysse, R.N. Van, J. Hellemans, K. Swerts, S. Bravo, P. Scaruffi, G.P. Tonini, B.B. De, R. Noguera, M. Piqueras, A. Canete, V. Castel, I. Janoueix-Lerosey, O. Delattre, G. Schleiermacher, J. Michon, V. Combaret, M. Fischer, A. Oberthuer, P.F. Ambros, K. Beiske, J. Benard, B. Marques, H. Rubie, J. Kohler, U. Potschger, R. Ladenstein, M.D. Hogarty, P. McGrady, W.B. London, G. Laureys, F. Speleman, J. Vandesompele, Lancet Oncol. 10 (2009) 663–671.
[12] M.E. de Boer, S. Berg, M.E. Timmermans, M.E. den Dunnen, N.M. van Straalen, J. Ellers, D. Roelofs, BMC Mol. Biol. 12 (2011) 11.
[13] M. Izzo, P. Kirkland, X. Gu, Y. Lele, A. Gunn, J. House, Aus. Vet. J. 90 (2012) 122–129.
[14] K.J. Livak, T.D. Schmittgen, Methods 25 (2001) 402–408.
[15] R.G. Rutledge, C. Cote, Nucleic Acids Res. 31 (2003) e93.
[16] N.J. Walker, Science 296 (2002) 557–559.
[17] T. Nolan, R.E. Hands, S.A. Bustin, Nat. Protoc. 1 (2006) 1559–1582.
[18] D.G. Ginzinger, Exp. Hematol. 30 (2002) 503–512.
[19] W.M. Freeman, S.J. Walker, K.E. Vrana, Biotechniques 26 (1999) 112–115.
[20] M.W. Pfaffl, M. Hageleit, Biotechnol. Lett. 23 (2001) 275–282.
[21] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, R.N. Van, P.A. De, F. Speleman, Genome Biol. 3 (2002). RESEARCH0034.
[22] S. Fleige, V. Walf, S. Huch, C. Prgomet, J. Sehm, M.W. Pfaffl, Biotechnol. Lett. 28 (2006) 1601–1613.
[23] Y. Karlen, A. McNair, S. Perseguers, C. Mazza, N. Mermod, BMC Bioinformatics 8 (2007) 131.
[24] J. Hellemans, G. Mortier, P.A. De, F. Speleman, J. Vandesompele, Genome Biol. 8 (2007) R19.
[25] G. Stolovitzky, G. Cecchi, Natl. Acad. Sci. USA 93 (1996) 12947–12952.
[26] J. Peccoud, C. Jacob, Biophys. J. 71 (1996) 101–108.
[27] R.G. Rutledge, D. Stewart, Mol. Biol. 9 (2008) 96.
[28] M.J. Alvarez, G.J. Vila-Ortiz, M.C. Salibe, O.L. Podhajcer, F.J. Pitossi, BMC Bioinformatics 8 (2007) 85.
[29] G.J. Boggy, P.J. Woolf, PLoS ONE 5 (2010) e12355.
[30] A. Lievens, S. Van Aelst, M. Van den Bulcke, E. Goetghebeur, Nucleic Acids Res. 40 (2012) e10.
[31] A. Gentle, F. Anastasopoulos, N.A. McBrien, BioTechniques 31 (2001) 502. 504–506, 508.
[32] A. Tichopad, M. Dilger, G. Schwarz, M.W. Pfaffl, Nucleic Acids Res. 31 (2003) e122.
[33] S. Zhao, R.D. Fernald, J. Comput. Biol. 12 (2005) 1047–1064.
[34] S.N. Peirson, J.N. Butler, R.G. Foster, Nucleic Acids Res. 31 (2003) e73.
[35] C. Ramakers, J.M. Ruijter, R.H. Lekanne Deprez, A.F.M. Moorman, Neurosci. Lett. 339 (2003) 62–66.
[36] W. Liu, D.A. Saint, Anal. Biochem. 302 (2002) 52–59.
[37] A.N. Spiess, C. Feig, C. Ritz, BMC Bioinformatics 9 (2008) 221.
[38] J.M. Ruijter, C. Ramakers, W.M. Hoogaars, Y. Karlen, O. Bakker, Nucleic Acids Res. 37 (2009) e45.
[39] W. Liu, D.A. Saint, Biochem. Biophys. Res. Commun. 294 (2002) 347–353.
[40] A. Larionov, A. Krause, W. Miller, BMC Bioinformatics 6 (2005) 62.
[41] M. Guescini, D. Sisti, M.B. Rocchi, L. Stocchi, V. Stocchi, BMC Bioinformatics 9 (2008) 326.
[42] R.G. Rutledge, PLoS ONE 6 (2011) e17636.
[43] N.R. Markham, M. Zuker, Methods Mol. Biol. 453 (2008) 3–31.
[44] J. Vermeulen, K. De Preter, S. Lefever, J. Nuytens, V.F. De, S. Derveaux, J. Hellemans, F. Speleman, J. Vandesompele, Nucleic Acids Res. 39 (2011) e63.
[45] W.J. Conover, Practical Nonparametric Statistics, J Wiley, New York, 1980.
[46] A. Tichopad, R. Kitchen, I. Riedmaier, C. Becker, A. Stahlberg, M. Kubista, Clin. Chem. 55 (2009) 1816–1823.
[47] R.G. Rutledge, D. Stewart, BMC Biotechnol. 8 (2008) 47.
[48] R.G. Rutledge, D. Stewart, PLoS ONE 5 (2010) e9731.
[49] D. Sisti, M. Guescini, M.B. Rocchi, P. Tibollo, M. D'Atri, V. Stocchi, BMC Bioinformatics 11 (2010) 186.

## Glossary

*Raw fluorescence data:* observed fluorochrome dependent fluorescence, corrected for technical background but not for fluorescence baseline.

*Baseline-corrected data:* amplification dependent fluorescence; fluorescence data corrected for fluorescence that is independent of amplification.

*E or PCR efficiency:* amplification efficiency defined as the fold increase per cycle and thus ranging from 1 to 2.

$\boldsymbol{C_q}$ *or quantification cycle:* fractional number of cycles needed to reach the fluorescence threshold. $\boldsymbol{C_q}$ is also known as $\boldsymbol{C_p}$ or $\boldsymbol{C_t}$ but the use of those terms is discouraged by the MIQE guidelines.

$\boldsymbol{F_q}$ *or quantification fluorescence threshold:* threshold set to determine the $\boldsymbol{C_q}$ value.

$\boldsymbol{F_0}$: fluorescence associated with the target quantity or starting concentration of the DNA-of-interest, expressed in arbitrary fluorescence units.

*NTC:* no template control.

*Reaction:* individual reaction unit, corresponds to one well in a PCR run.

*SDM:* cycle at which the second derivative of the fluorescence values reaches its maximum.

*Sample or tissue:* biological material in which target nucleic acid levels have to be determined. A tissue can be measured in different reactions to measure different targets.

*Target or amplicon:* DNA of-interest, product of the PCR reaction specified by a pair of primers.