

Normalization of qPCR data

Daniel Uddenberg

Dept. of Physiological Botany, UU

daniel.uddenberg@ebc.uu.se

(although sitting door-to-door to Alyona)

Some of what we will cover today

Normalization

Absolute quantification

Relative quantification

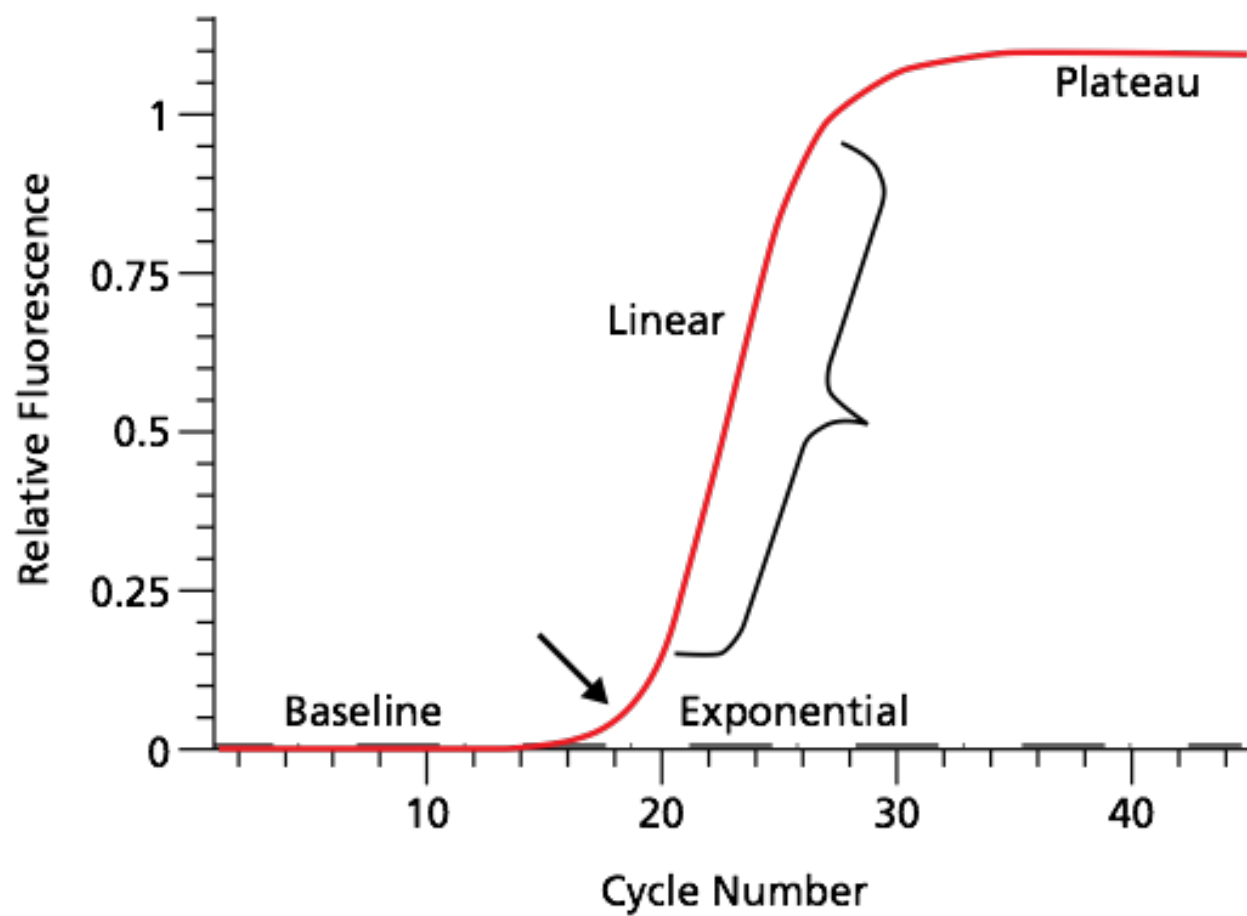
Standard curves



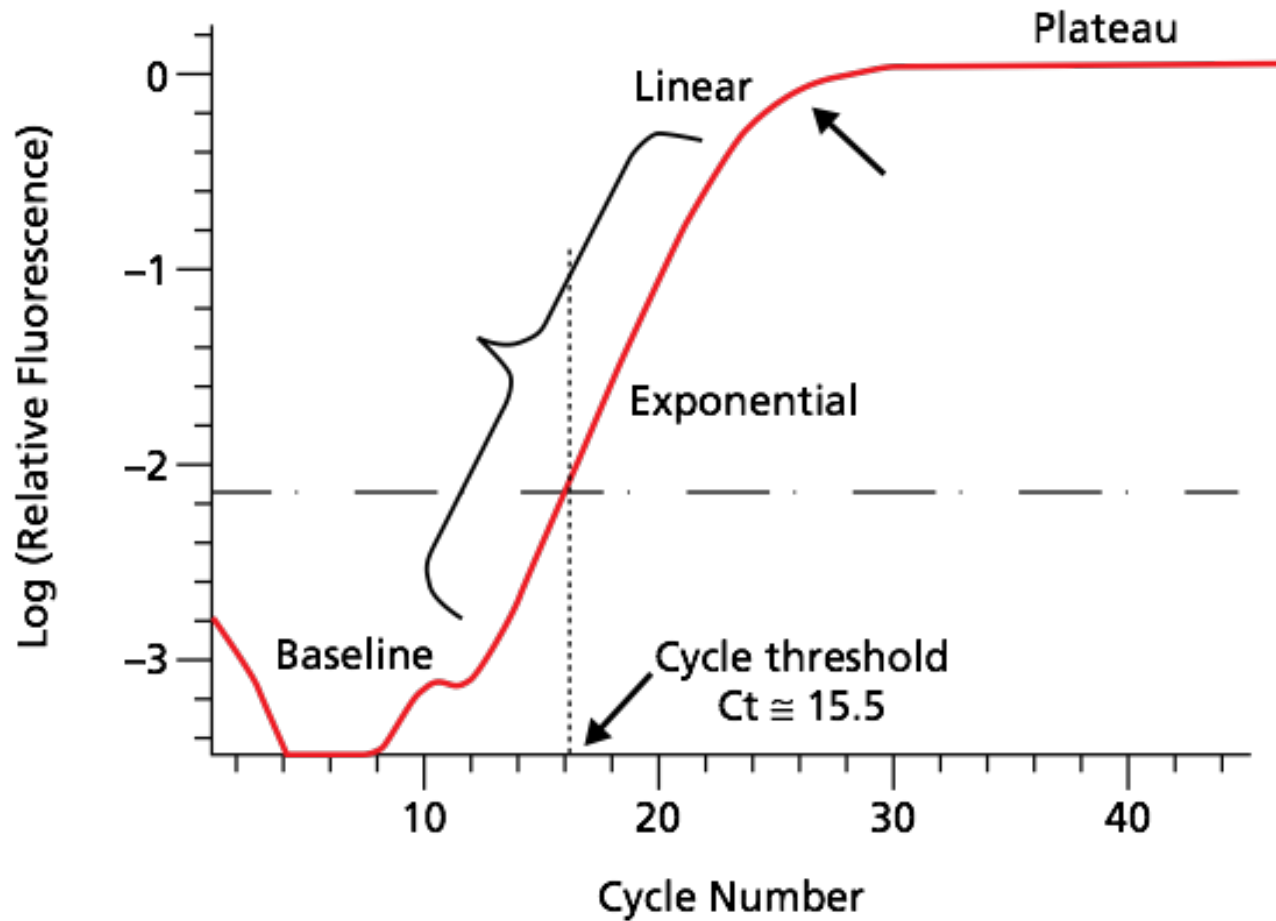
Assay setup

Reference genes

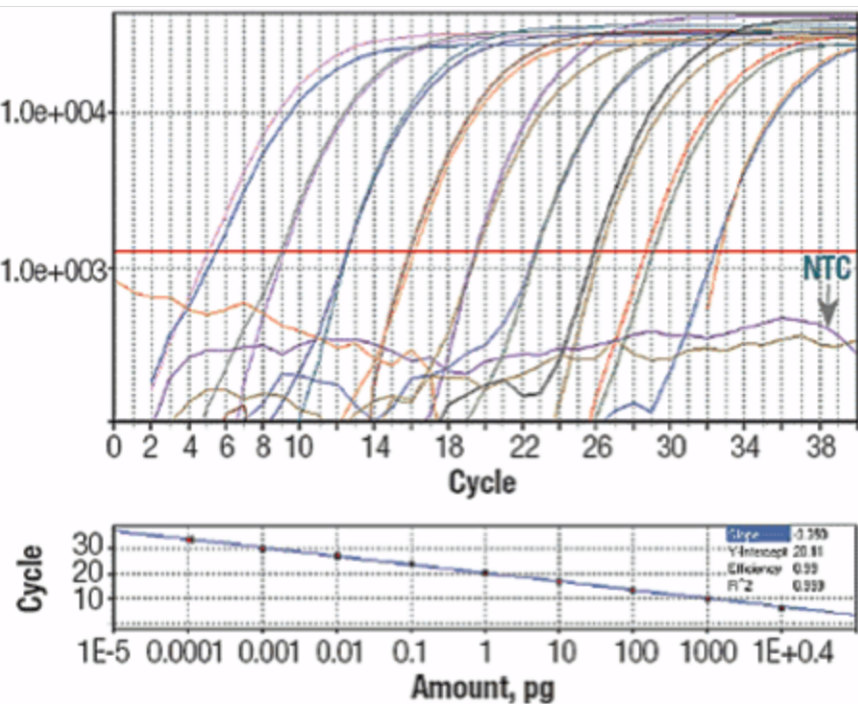
Standard X-Y Plot



Semi-log Plot



$$C_T = C_q = C_p$$



Dilution series

If perfect doubling of amplicons:

$2^n = \text{dilution factor}$ (n = number of cycles between Ct values)

e.g. 10-fold dilution $\rightarrow 2^n = 10$

$\rightarrow n = \log_2(10)$

$\rightarrow n = 3,32$ cycles between Ct values

Why normalizing data?

Two sources of variation in gene expression results

- biological variation
- experimentally induced variation

Examples of experimental variation in qPCR data?

- Input quantity and quality

The purpose of normalization is to reduce experimental variation

Absolute vs Relative quantification

Absolute quantification

- Why: - How many, in relation to “something”?
- Application: - Chromosome or gene copy determination, viral load measurements, etc.
- Result: - **A quantity** – nucleic acid (copy nr, μg) per given amount of sample (per cell, per μg nucleic acid)

Relative quantification

- Why: - To compare levels or changes in gene expression
“What is the fold difference?”
- Application: - Developmental biology, disease research, siRNA, etc.
- Result: - **A ratio** - or fold change between e.g. control and treatment

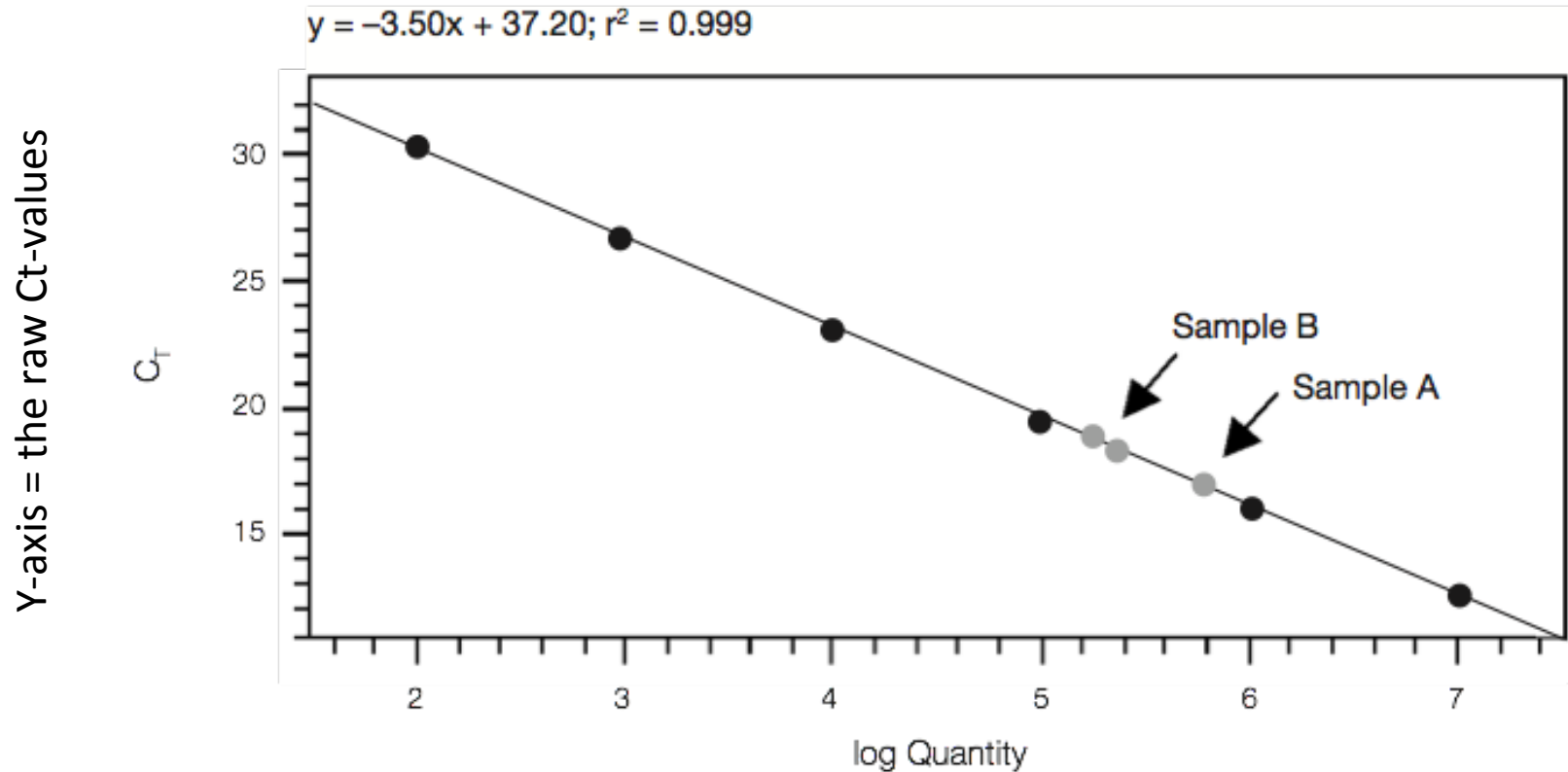
Absolute quantification

Important things to consider

- Standards must be amplified in parallel to your samples every time
- The RNA/DNA for the standard curve must be a single pure species
- Stability of the diluted standards important
- Accurate pipetting over the serial dilution
- The concentration of the standards must span your sample concentration

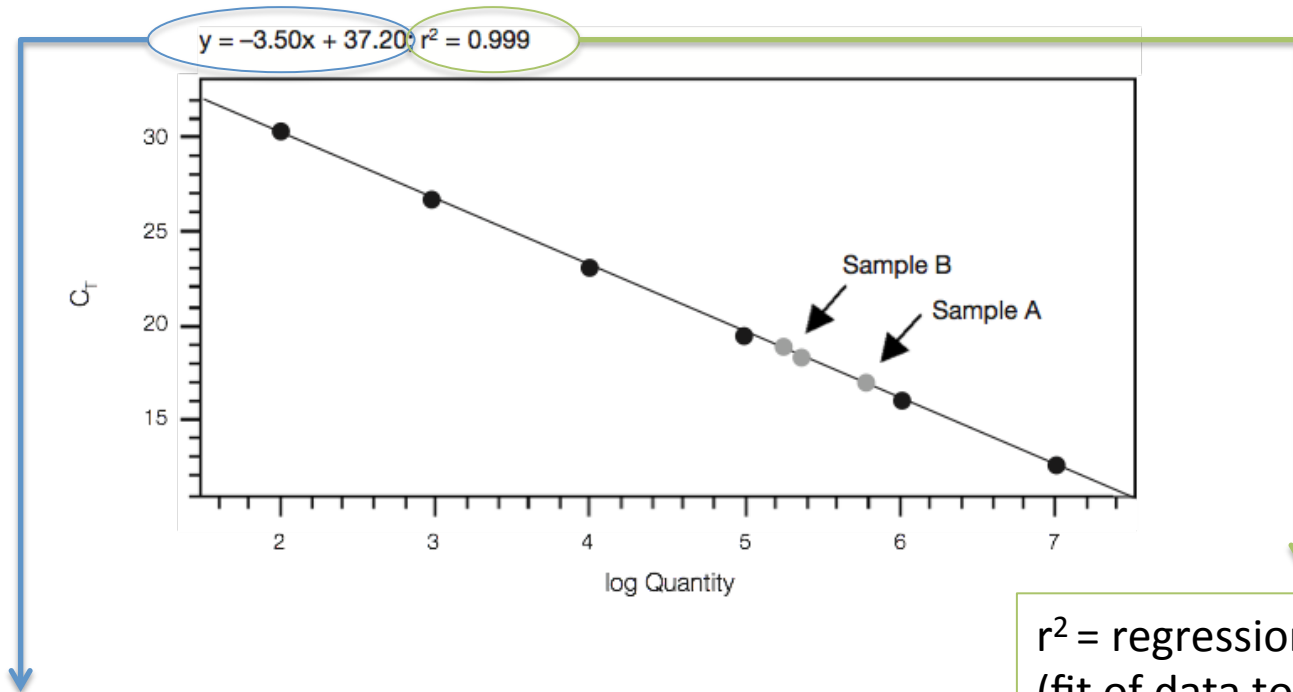
Absolute quantification

Based on comparisons to a standard curve



X-axis = the log quantity of the initial (copy number/unit) of the standards

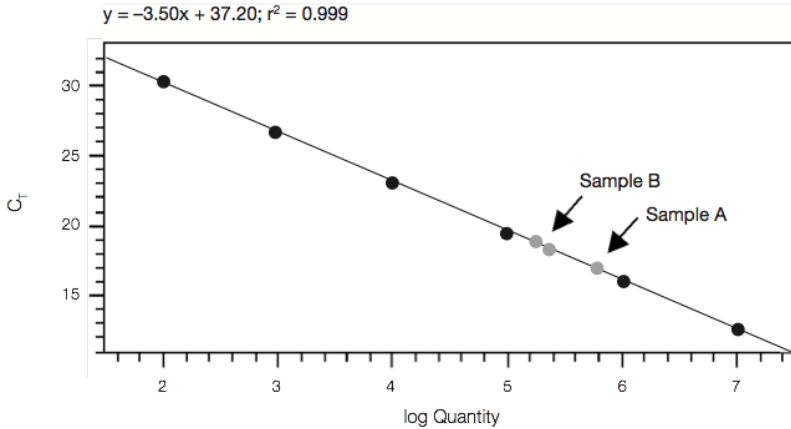
Absolute quantification



The equation of the linear regression: $y = kx + m$

y: Ct value
k: Slope
x: log quantity
m: y-intercept

r^2 = regression coefficient
(fit of data to trendline, 0-1)



y: Ct value
k: Slope
x: log quantity
m: y-intercept

The equation of the linear regression: $y = kx + m$

From this regression equation we can derive the following formula to determine the quantity of our unknown sample:

$$N_n = 10^{((n-m)/k)}, \text{ where } n = Ct$$

N: Quantity
n: unknown sample
k: slope
m: y-intercept

Example 1

Relative quantification

Normalized against unit mass (cell nr. or μg nucleic acid)

	Test (e.g. treatment)	Calibrator (e.g. control)
Target Gene (GOI)	$C_{T(\text{GOI, test})}$	$C_{T(\text{GOI, calibrator})}$

$$2^{\Delta C_T}$$

Where, $\Delta C_T = C_{T(\text{calibrator})} - C_{T(\text{test})}$

Requires accurate quantification of starting material

Relative quantification

Normalized to a reference gene(s)

Circumvents the need for accurate quantification and loading of starting material

Requires available reference genes with constant expression over samples that are non-affected by treatments in your study

Relative quantification

Normalized to a reference gene(s)

	Test (e.g. treatment)	Calibrator (e.g. control)
Target Gene (GOI)	$C_{T(\text{GOI, test})}$	$C_{T(\text{GOI, calibrator})}$
Reference gene (REF)	$C_{T(\text{REF, test})}$	$C_{T(\text{REF, calibrator})}$

First, normalize C_T of GOI to C_T of REF, for both the test sample and the calibrator sample

$$\Delta C_{T(\text{test})} = C_{T(\text{GOI, test})} - C_{T(\text{REF, test})}$$

$$\Delta C_{T(\text{calibrator})} = C_{T(\text{GOI, calibrator})} - C_{T(\text{REF, calibrator})}$$

First, normalize C_T of GOI to C_T of REF, for both the test sample and the calibrator sample

$$\Delta C_{T(\text{test})} = C_{T(\text{GOI, test})} - C_{T(\text{REF, test})}$$

$$\Delta C_{T(\text{calibrator})} = C_{T(\text{GOI, calibrator})} - C_{T(\text{REF, calibrator})}$$

Second, normalize ΔC_T of test to ΔC_T of calibrator

$$\Delta\Delta C_T = \Delta C_{T(\text{test})} - \Delta C_{T(\text{calibrator})}$$

First, normalize C_T of GOI to C_T of REF, for both the test sample and the calibrator sample

$$\Delta C_{T(\text{test})} = C_{T(\text{GOI, test})} - C_{T(\text{REF, test})}$$

$$\Delta C_{T(\text{calibrator})} = C_{T(\text{GOI, calibrator})} - C_{T(\text{REF, calibrator})}$$

Second, normalize ΔC_T of test to ΔC_T of calibrator

$$\Delta\Delta C_T = \Delta C_{T(\text{test})} - \Delta C_{T(\text{calibrator})}$$

Finally, calculate expression ratio

$$2^{-\Delta\Delta C_T} = \text{Normalized expression ratio}$$

Finally, calculate expression ratio

$$2^{-\Delta\Delta C_T} = \text{Normalized expression ratio}$$



$$\text{Ratio} = \frac{(2)^{\Delta C_T, \text{GOI (calibrator, test)}}}{(2)^{\Delta C_T, \text{REF (calibrator, test)}}$$

Assuming ~100% efficient primers



$$\text{Pfaffl ratio} = \frac{(E_{\text{GOI}})^{\Delta C_T, \text{GOI (calibrator, test)}}}{(E_{\text{REF}})^{\Delta C_T, \text{REF (calibrator, test)}}$$

$$E = 10^{-1/\text{slope}}$$

How to select reference genes?

Available primers (published or shared)

- preferentially already tested in my own species/system
- REFs that usually are inter-sample stable across species, e.g. Actin, GADPH...

* There are specific databases for tested qPCR primers:

The screenshot shows the RTPrimerDB website. At the top, there is a navigation bar with 'Home', 'Search', and 'In silico evaluation' links, and a 'Log in' button. Below this is a search bar with fields for 'RTPrimerDB ID', 'Gene', and 'Organism', and a 'Search' button. The main content area is divided into two columns. The left column contains a sidebar with links: 'Home', 'Statistics', 'Links', 'News', 'Citations', 'Faq', 'Comments', and 'Downloads'. The right column contains the 'Introduction' section, which describes the database as a public resource for real-time PCR primers and probes. It mentions that the database is used for various applications like Gene Expression Quantification, DNA Copy Number Quantification, SNP Detection, etc. It also states that the database is currently available for 5762 genes and 8368 real-time PCR assays. Below the introduction is a 'Publications' section with a list of three references, each with a 'Download' link. At the bottom left, there are logos for 'BIO RAD', 'Roche', and 'Eurogentec'.

RT PRIMER DB

Quicksearch - Filter settings

RTPrimerDB ID Gene Organism Search

Exact phrase Substring

Home Search In silico evaluation Log in

Home Statistics Links News Citations Faq Comments Downloads

RTPrimerDB is generously sponsored by

BIO RAD Roche Eurogentec

Introduction

RTPrimerDB is a public database for primer and probe sequences used in real-time PCR assays employing popular chemistries (SYBR Green I, Taqman, Hybridisation Probes, Molecular Beacon) to prevent time-consuming primer design and experimental optimisation, and to introduce a certain level of uniformity and standardisation among different laboratories.

We strongly encourage researchers to submit their validated primer and probe sequence, so that other users can benefit from their expertise. The database can be [queried](#) using the official gene name or symbol, [Entrez](#) or [Ensembl](#) Gene identifier, [SNP](#) identifier, or oligonucleotide sequence.

Different [options](#) make it possible to restrict a query to a particular application (Gene Expression Quantification/Detection, DNA Copy Number Quantification/Detection, SNP Detection, Mutation Analysis, Fusion Gene Quantification/Detection, Chromatin immunoprecipitation (ChIP)), organism (Human, Mouse, Rat, and others) or detection chemistry. Data submission is allowed after [free registration](#) whereby you obtain a login name and password.

Currently, [8368 real-time PCR assays](#) for 5762 genes are available, submitted by 226 people.

Last submission: link

Publications

- PATTYN, F., SPELEMAN, F., DE PAEPE A. & VANDESOMPELE, J. (2003). RTPrimerDB: the Real-Time PCR primer and probe database. *Nucleic Acids Research*, 31(1): 122-123. [Download](#)
- PATTYN, F., ROBBRECHT, P., SPELEMAN, F., DE PAEPE A. & VANDESOMPELE, J. (2006). RTPrimerDB: the Real-Time PCR primer and probe database, major update 2006. *Nucleic Acids Research*, 34(Database issue): D684-688. [Download](#)
- LEFEVER S, VANDESOMPELE J, SPELEMAN F, PATTYN F. (2008). RTPrimerDB: the portal for real-time PCR primers and probes. *Nucleic Acids Research*, Oct 23. [Epub ahead of print] [Download](#)

How to select reference genes?

Global transcript profiling datasets

- Mine for stably expressed genes in microarrays and RNA-seq data

* There are also specific databases for this purpose:



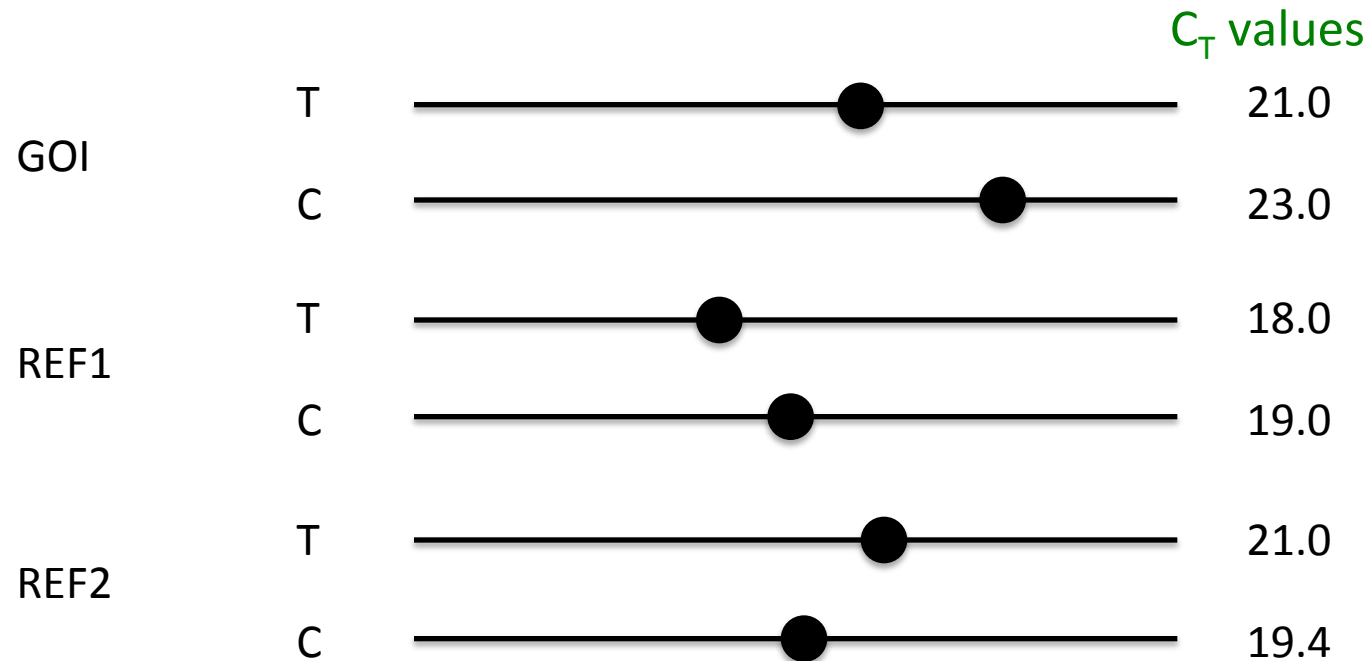
<http://www.refgenes.org/rg/>

Relative quantification

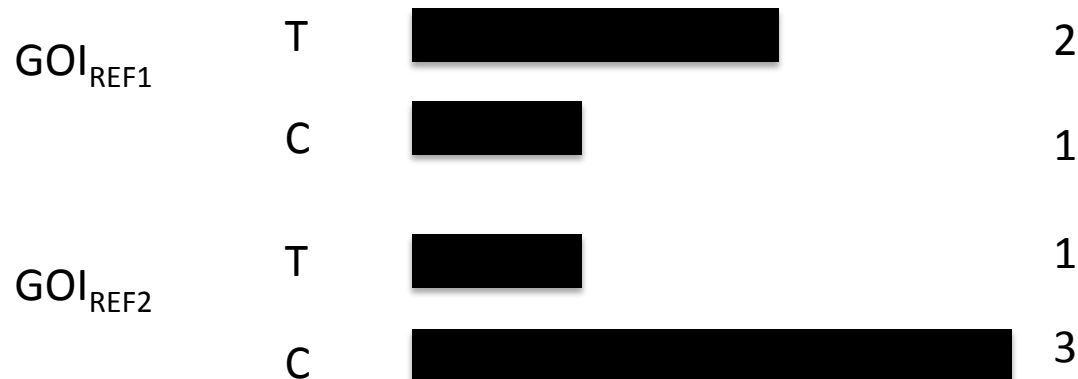
Normalized to a reference gene(s)

Do we need more than one reference gene?

Do we need more than one reference gene?



Normalized relative quantities



6-fold difference

Relative quantification

Normalized to reference genes

In most cases you do!

Quantified errors related to the use of one REF:
> 3-fold in 25% of cases; > 6-fold in 10% of cases

Research

Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes

Jo Vandesompele, Katleen De Preter, Filip Pattyn, Bruce Poppe,
Nadine Van Roy, Anne De Paepe and Frank Speleman

Address: Center for Medical Genetics, Ghent University Hospital 1K5, De Pintelaan 185, B-9000 Ghent, Belgium.

Correspondence: Frank Speleman. E-mail: franki.speleman@rug.ac.be

Published: 18 June 2002

Genome **Biology** 2002, **3**(7):research0034.1-0034.11

Received: 20 December 2001

Revised: 10 April 2002

Accepted: 7 May 2002

Normalization against multiple reference genes

- 2-5 **Validated** stably expressed genes
- **Tested** across all samples that will be used in subsequent experiments
- Enables quality control on their stability

Validating REFs

If you, in a pilot experiment, do this correct you are probably set for the remainder of your lab-life!

General recommendations

- Include all or most of the sample types (tissue type, treatment, time-series...) that you will later assay, ≥ 10
- Test at least 8 different candidate genes

Example of a simple pilot setup

	1	2	3	4	5	6	7	8	9	10	11	12
A	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C
B	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C
C	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C
D	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C
E	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C
F	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C
G	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C
H	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10		C

- ✓ 10 different sample (S1-S10)
- ✓ 8 different candidate reference genes (color coded)
- ✓ C = negative controls (no template control)

How do we Validate REFs

There are programs that can help you with statistics:

- e.g. **geNorm**, Normfinder, BestKeeper etc.



BestKeeper

<http://www.gene-quantification.de/bestkeeper.html>

NormFinder

<http://moma.dk/normfinder-software>

! Old MS excel macros...

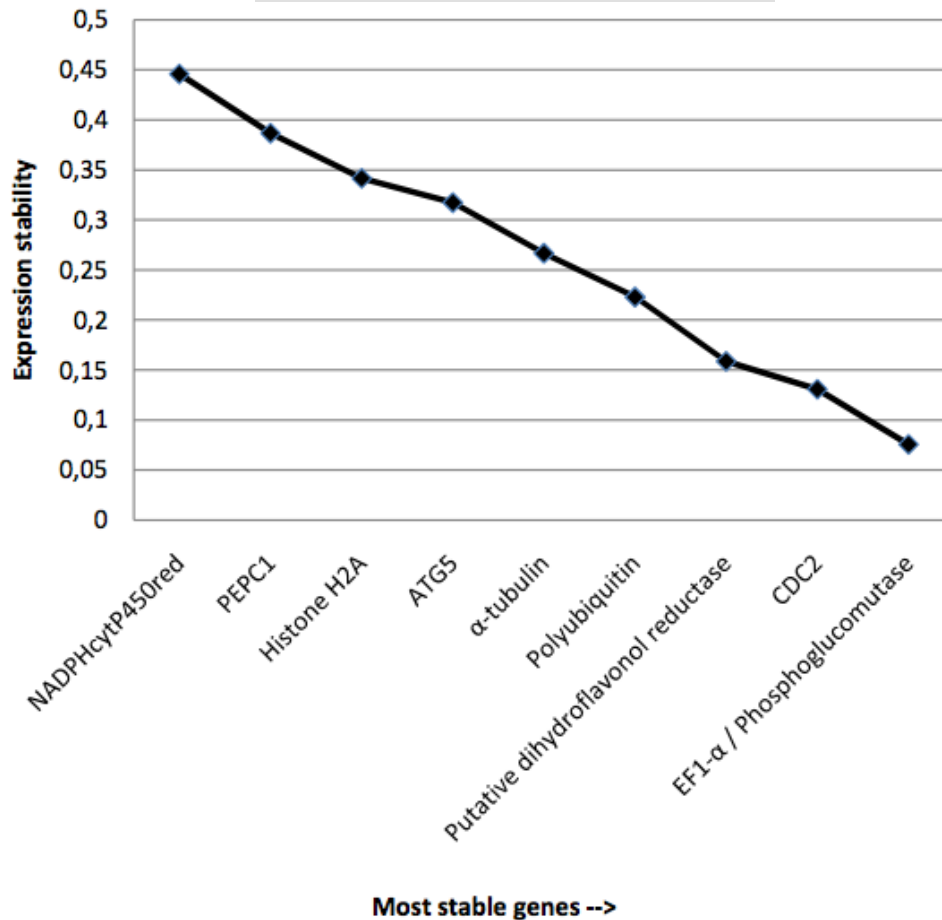
...However, some have versions coded in R

<http://normalisation.gene-quantification.info>

geNorm

(qbase⁺, Biogazelle)

Example from geNorm




Calculates a stability index, M

M-value calculation
possible in **CFX-manager**

Gene stability measure M

- Average pairwise variation of V of a given REF with all other candidate REFs
- Iterative procedure to remove the worst REF followed by recalculation of M-values

	Gene A	Gene B	
Sample 1	A1	B1	$\text{Log2}(A1/B1)$
Sample 2	A2	B2	$\text{Log2}(A2/B2)$
Sample 3	A3	B3	$\text{Log2}(A3/B3)$
...
Sample n	A(n)	B(n)	$\text{Log2}(An/Bn)$



Standard deviation = V

Example 2

However, one thing that the algorithms do not take into account: Systematic variation across samples (co-expression)



Choose genes that are expressed in different pathways!!!

Sample maximization vs. gene maximization



	1	2	3	4	5	6	7	8	9	10	11	12
A	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
B	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
C	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
D	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
E	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
F	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
G	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
H	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C



	1	2	3	4	5	6	7	8	9	10	11	12
A	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1
B	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2
C	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3
D	S4	S4	S4	S4	S4	S4	S4	S4	S4	S4	S4	S4
E	S5	S5	S5	S5	S5	S5	S5	S5	S5	S5	S5	S5
F	S6	S6	S6	S6	S6	S6	S6	S6	S6	S6	S6	S6
G	S7	S7	S7	S7	S7	S7	S7	S7	S7	S7	S7	S7
H	C	C	C	C	C	C	C	C	C	C	C	C

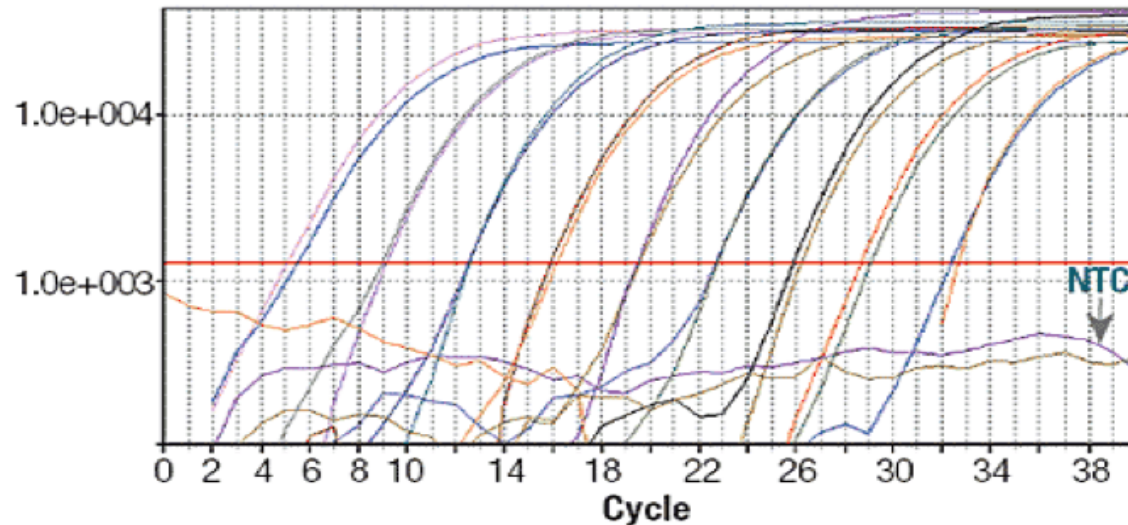
Which method is to prefer?

	1	2	3	4	5	6	7	8	9	10	11	12
A	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
B	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
C	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
D	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	C
E												
F												
G												
H												

	1	2	3	4	5	6	7	8	9	10	11	12
IRC1 A	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1
IRC2 B	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2
IRC3 C	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3
D	S8	S8	S8	S8	S8	S8	S8	S8	S8	S8	S8	S8
E	S9	S9	S9	S9	S9	S9	S9	S9	S9	S9	S9	S9
F	S10	S10	S10	S10	S10	S10	S10	S10	S10	S10	S10	S10
G	S11	S11	S11	S11	S11	S11	S11	S11	S11	S11	S11	S11
H	C	C	C	C	C	C	C	C	C	C	C	C

Gene max. requires linking between plates

- Inter-run calibration is only necessary when the samples of the same gene is run on separate plates
- No need to measure REFs on the same plate as GOIs



The ONLY criterion necessary for REFs is that they are stably expressed.

Importance of RNA quality

Samples from two experiments, using both low quality and high quality RNA

Step*	Degraded RNA (CRS samples)	Intact RNA (CRS samples)	Degraded RNA (NP samples)	Intact RNA (NP samples)
1	HPRT1	GAPD	HPRT1	YWHAZ
2	YWHAZ	YWHAZ	ACTB	B2M
3	B2M	RPL3IA	RPL3IA	RPL3IA
4	TBP	B2M	GAPD	UBC
5	RPL3IA	UBC	TBP	GAPD
6	UBC	HPRT1	YWHAZ	HMBS
7	ACTB	TBP	HMBS	HPRT1
8	GAPD	ACTB	SDHA	SDHA
9	HMBS- SDHA	HMBS- SDHA	B2M- UBc	ACTB- TBP

Least stable



Most stable

How do we calculate the normalization factor for multiple reference genes?

Arithmetic mean “average” = $(a + b + c) / 3$

Geometric mean of REF expression levels

Geometric mean = $(a \times b \times c)^{1/3}$

- Controls for outliers
- Compensates for differences in expression levels between reference genes

Algorithms already included in Real-Time qPCR instruments

Why normalizing data?

To make the data biologically meaningful

To avoid
sample-to-sample variations
and run-to-run variations

The most important factor to get accurate qPCR results

Further reading

<http://www.gene-quantification.info>

A resource covering everything there is to now, and more

How to do successful gene expression analysis using real-time PCR
(2010) Derveaux et al. Methods 50(4):227-230

A good review paper

Real-time PCR – Applications guide (Bio-Rad)

[http://www.bio-rad.com/webroot/web/pdf/
lsr/literature/Bulletin_5279.pdf](http://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_5279.pdf)