Gene expression profiling: accurate normalisation and automated data-analysis

Jo Vandesompele
Center for Medical Genetics
Ghent University Hospital, Belgium

qPCR Satellite Symposium, March 11, 2005
Leipzig, Germany

outline

- pitfalls in qPCR based gene expression analysis
- accurate normalisation of gene expression using multiple references genes
  - geNorm concept
  - other approaches
- automated qPCR data-analysis
  - limitations of current analysis tools
  - qBASE demonstration

pitfalls

- template quality
  - Perez-Novo et al., submitted
- primer-dimer formation using SYBR Green I 1-step RT-PCR
  DNA contamination of RNA preparations
  - Vandesompele et al., Analytical Biochemistry, 2002
- primer design
  - RTPrimerDB: public database of primers and probes
    http://medgen.ugent.be/rtprimerdb/
    Pattyn et al., Nucleic Acids Research, 2003
  - secondary structures amplicons
    Hoebeeck et al., Laboratory Investigation, 2005
- splice isoform quantification
  - Vandenbroucke et al., Nucleic Acids Research, 2001
- normalisation of gene expression levels
  - Vandesompele et al., Genome Biology, 2002
- data-analysis
  - qBASE (Hellemans et al., in preparation)

normalization: what’s the problem ?

- gene-specific (biological) variation
- non-specific (technical) variation
  - RNA quantity & quality
  - RT efficiency
  - PCR efficiency (inhibitors)

normalization: what’s the solution ?

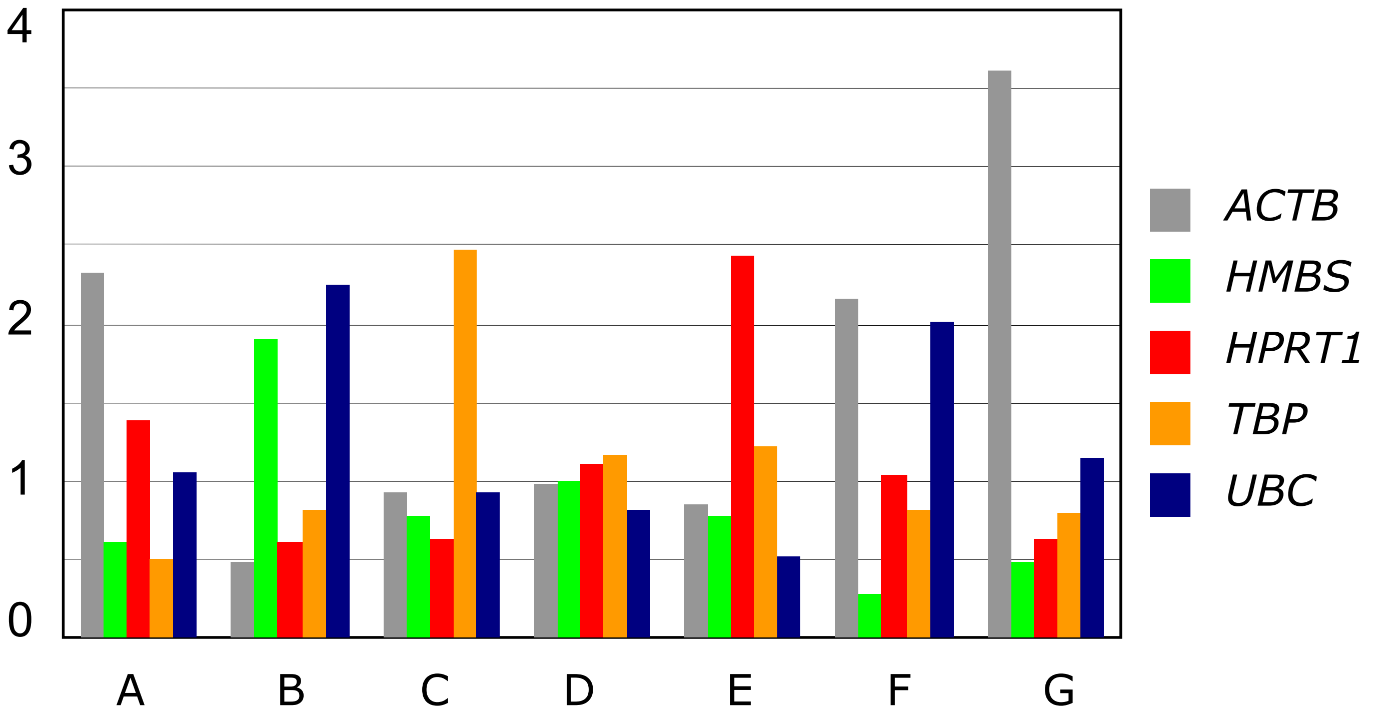
- many different strategies
- reference gene concept
  - most popular
  - captures most variation
- attention!
  - reference genes (might) vary in expression
  - (until recently) non-validated reference genes were used (assuming stable expression)

normalisation: our (geNorm) solution

- framework for qPCR gene expression normalisation using the reference gene concept (Genome Biology, 2002):
  - quantified errors related to the use of a single reference gene (> 3 fold in 25% of the cases; > 6 fold in 10% of the cases)
  - developed a robust algorithm for assessment of expression stability of candidate reference genes
  - proposed the geometric mean of at least 3 reference genes for accurate and reliable normalisation

real-life internal controls

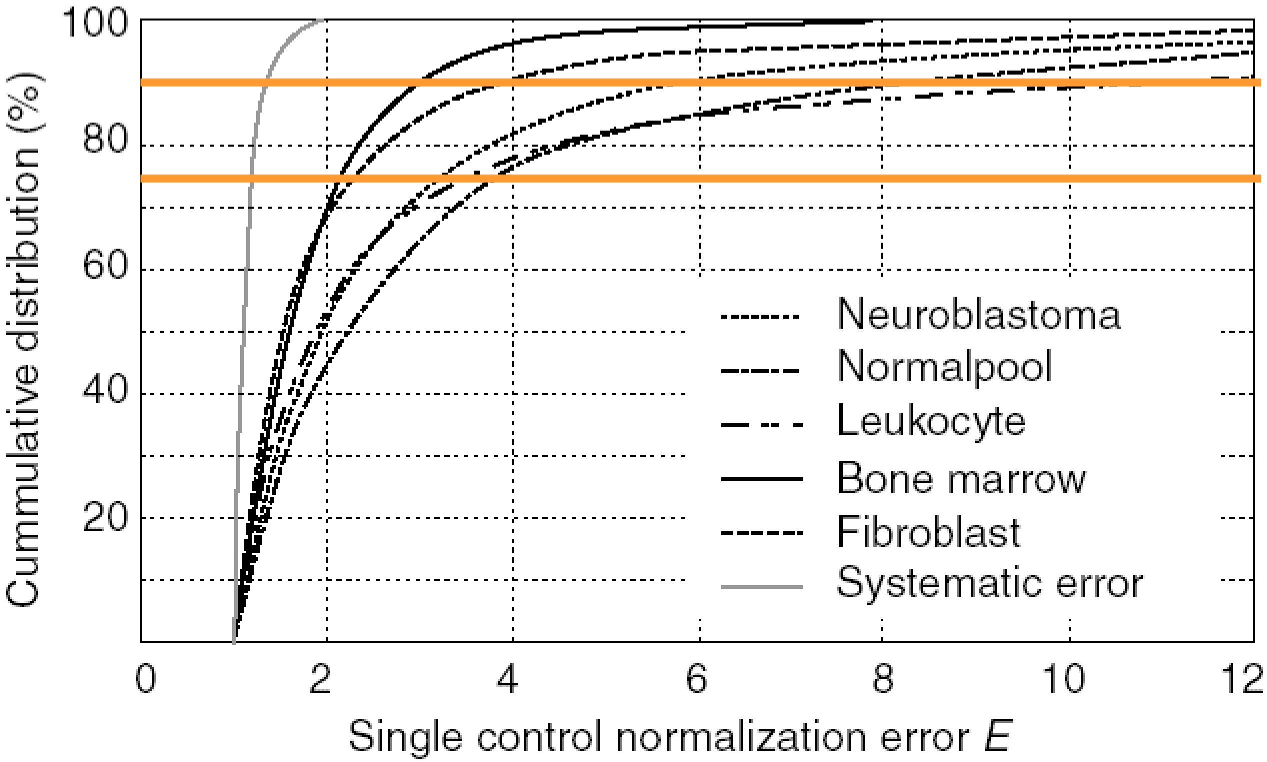
- quantitative RT-PCR analysis of 10 reference genes (belonging to different functional and abundance classes) on 85 samples < 13 different human tissues



15 fold difference between A and B if normalized by only one gene (*ACTB* or *HMBS)*

real-life internal controls

- single reference gene normalization error



  - up to 3 fold in 25% of the cases
  - up to 6.4 fold in 10% of the cases

the need for multiple reference genes

- given the extreme sensitivity, reproducibility and large dynamic range of quantitative RT-PCR
- the observed expression differences between so-called housekeeping genes
- absence of sufficient data to determine the biological significance of 2- to 3-fold expression differences

- we propose the use of multiple reference genes for accurate normalization

- which, how many, how?

robust assumption-free parameter

👎 assess the (standard) variation of the reference gene
> assume equal input of equal quality RNA

👍 compare 2 (or more) reference genes

geNorm expression stability parameter

- pairwise variation V (between 2 genes)

	gene A	gene B	
sample 1	a1	b1	log2(a1/b1)
sample 2	a2	b2	log2(a2/b2)
sample 3	a3	b3	log2(a3/b3)
…	…	…	…
sample n	an	bn	log2(an/bn)

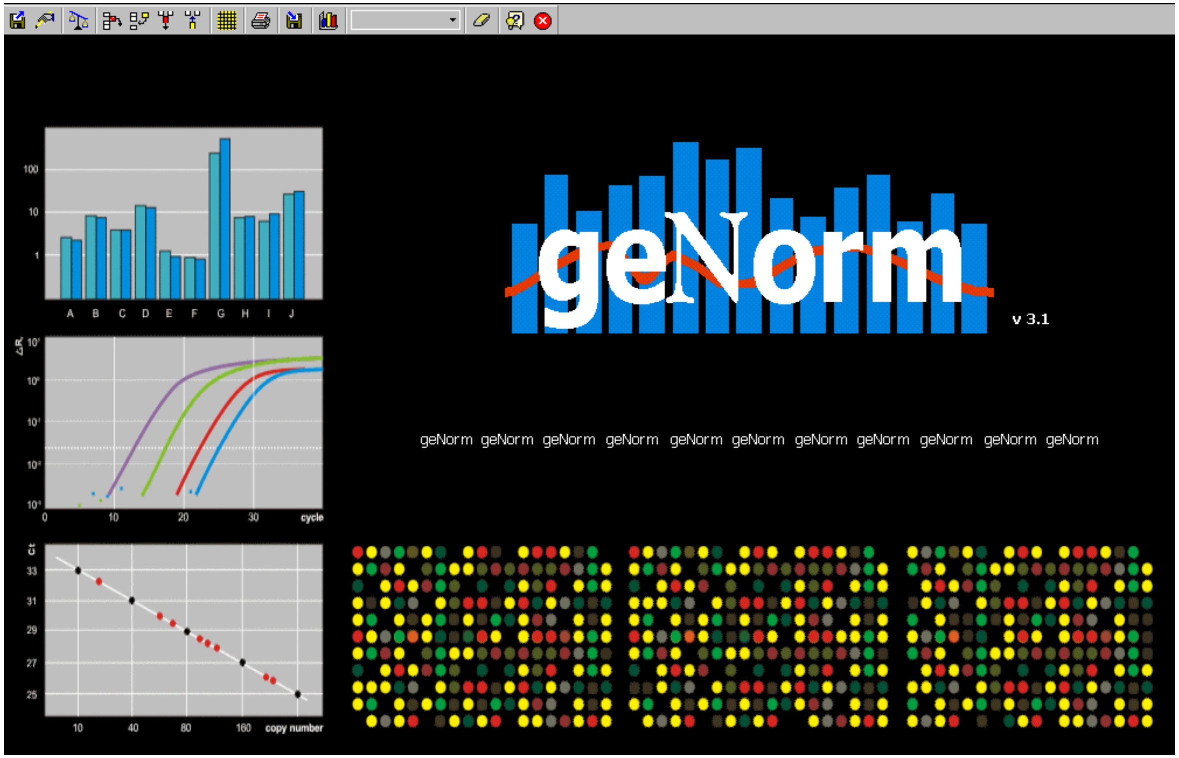
↓

standard deviation = V

- gene stability measure M
  average pairwise variation V of a gene with all other genes

geNorm

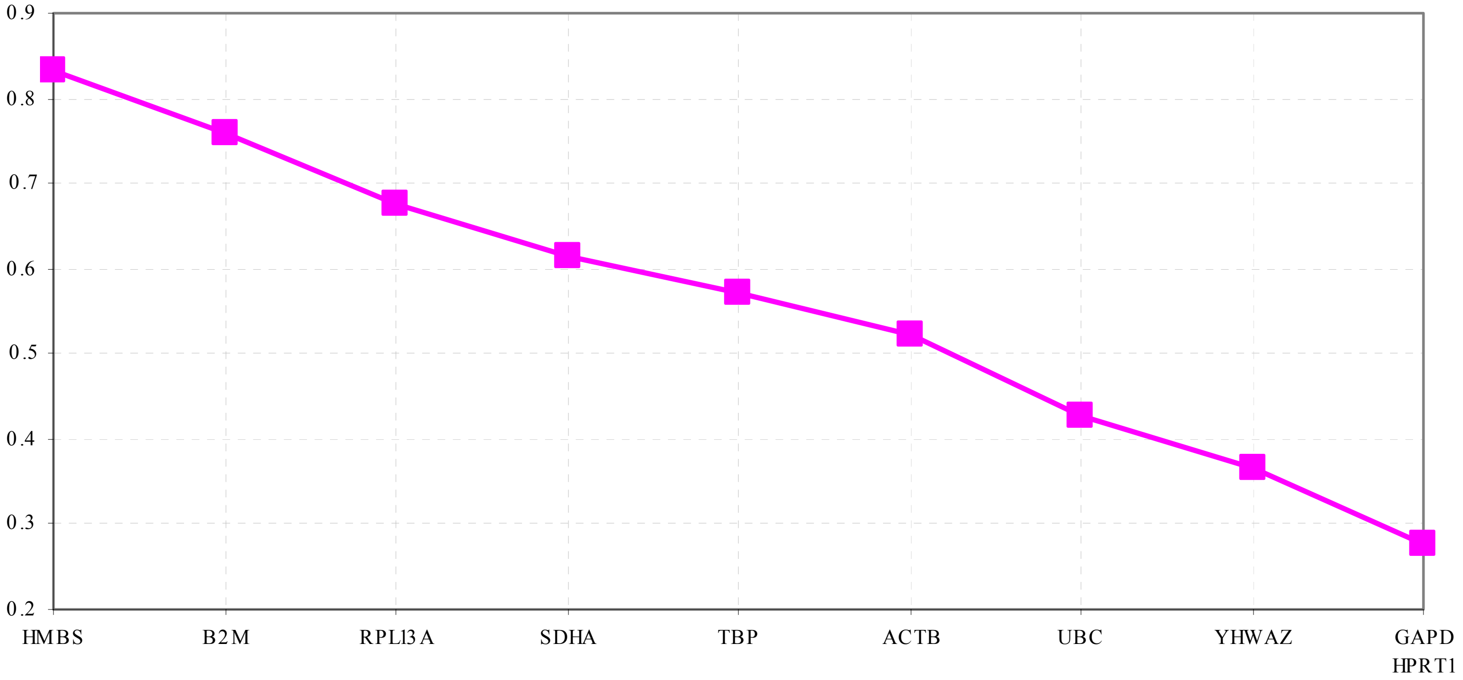
- automated analysis
  - ranking of candidate reference genes according to their stability
  - determination of how many genes are required for reliable normalization



medgen.ugent.be/~jvdesomp/genorm/

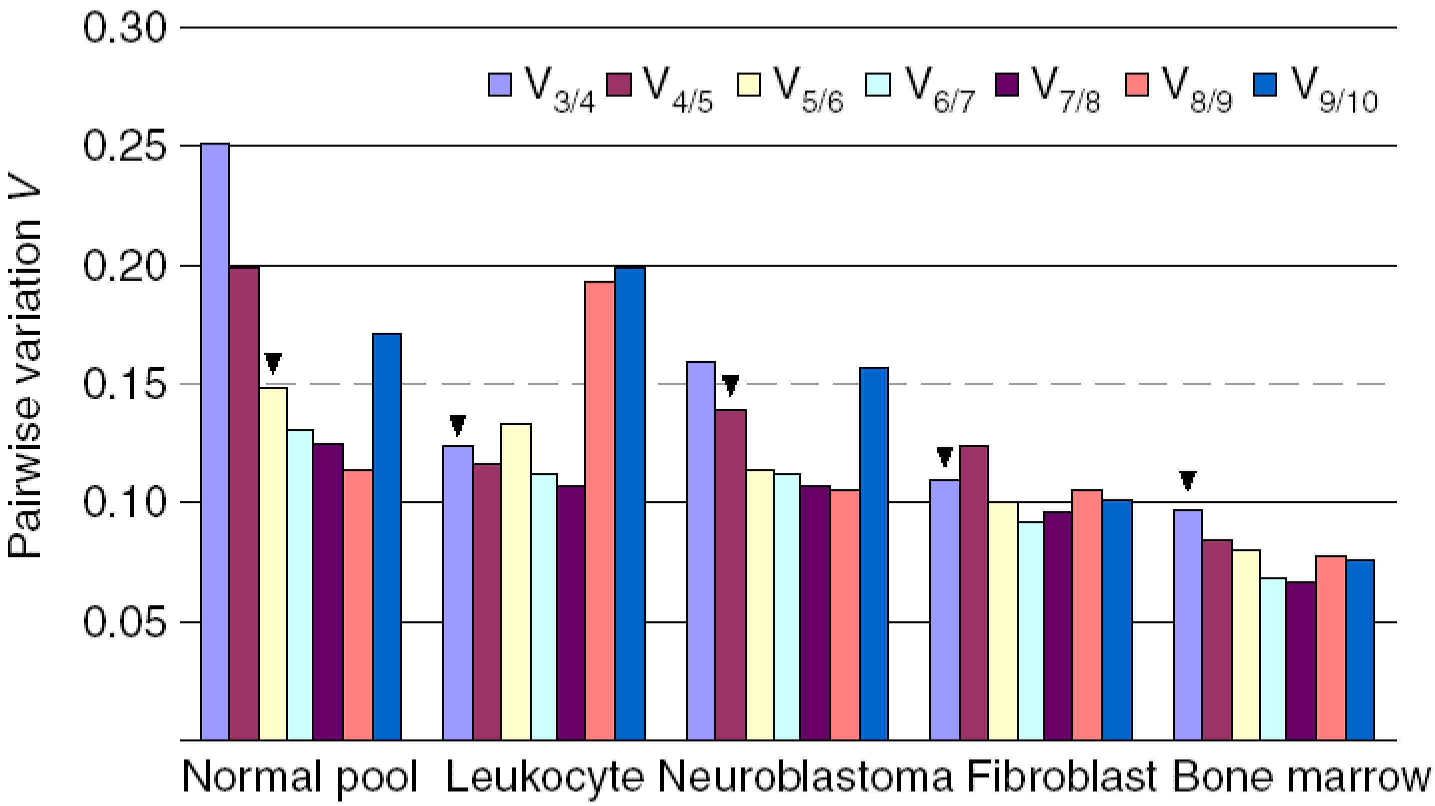
geNorm

- ranking of candidate reference genes according to their stability



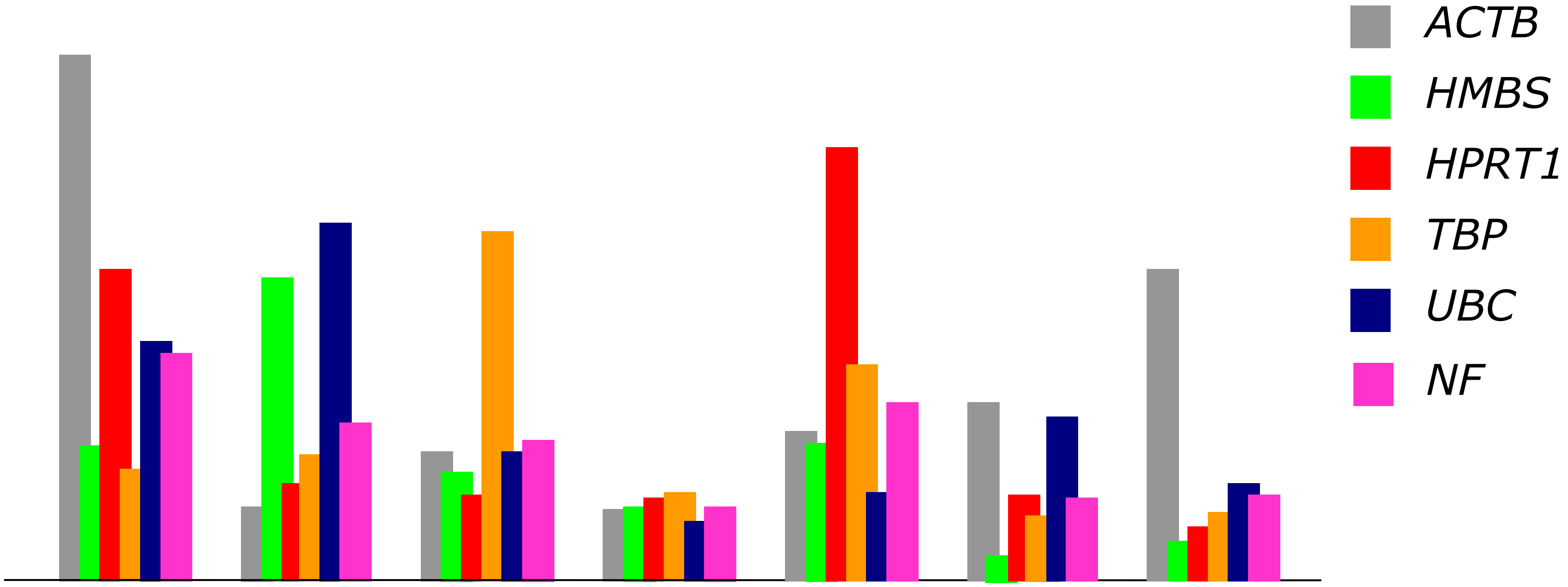
geNorm

- determination of the optimal number of reference genes



geNorm validation

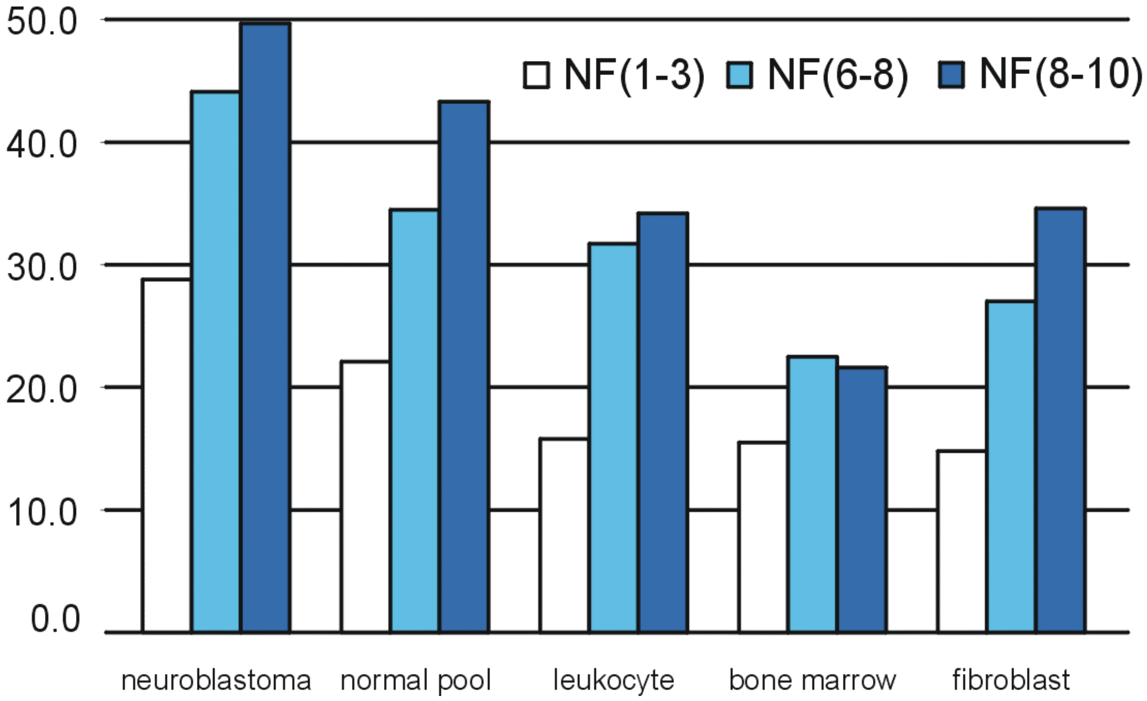
- robust – insensitive to outliers



Center for Medical Genetics

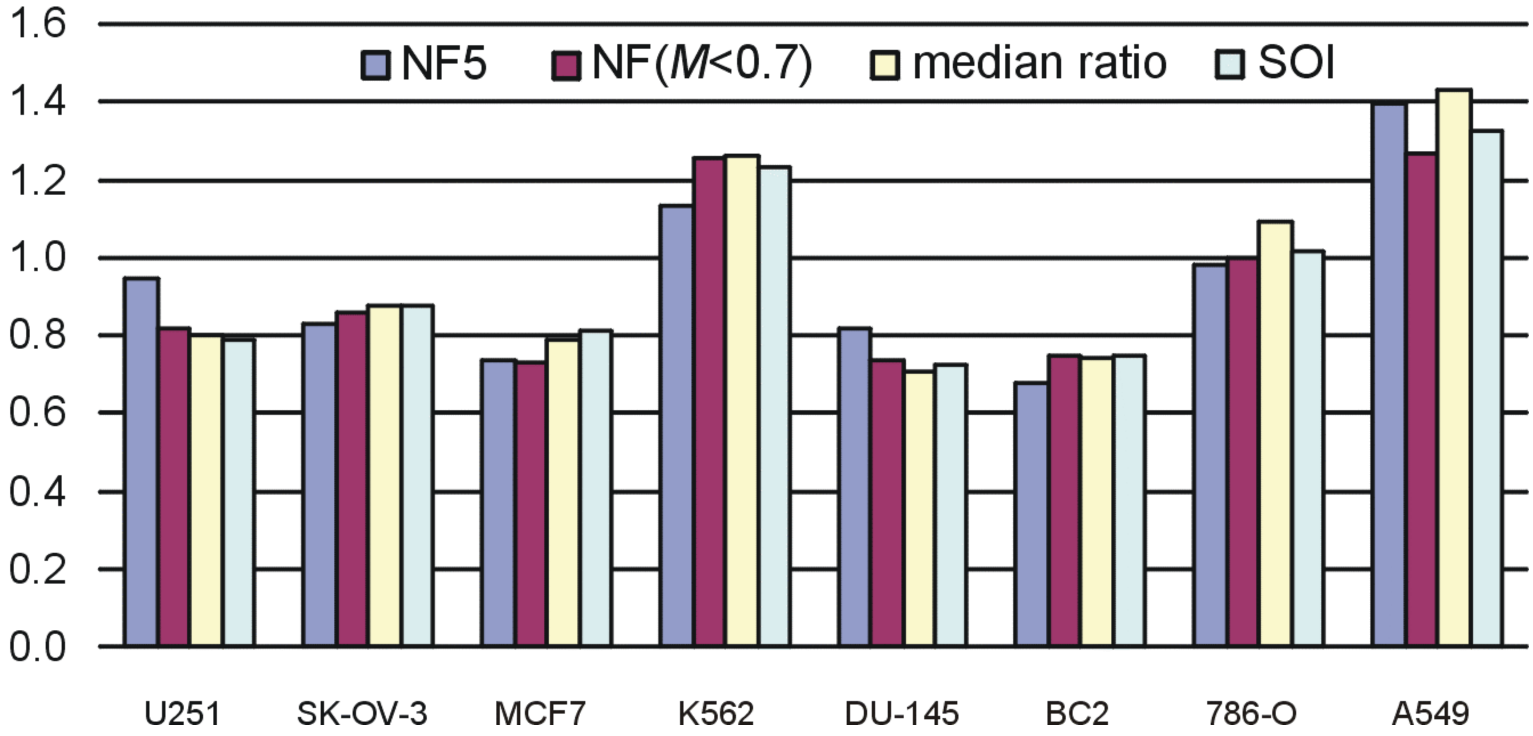
geNorm validation

- purpose of normalization: removal of non-specific variation



geNorm validation

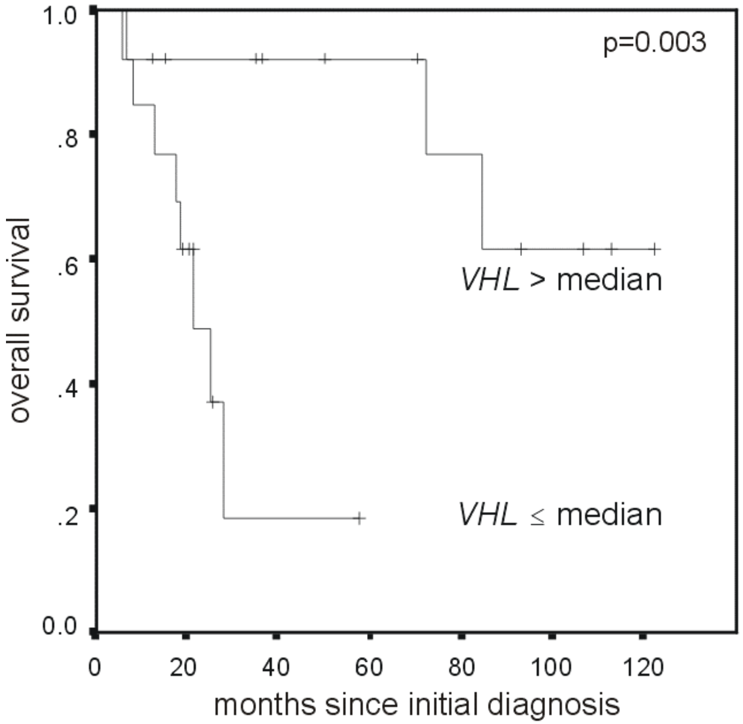
- comparison with microarray normalization factors



Center for Medical Genetics

geNorm validation

- cancer patients survival curve



log rank statistics

NF4

0.003

NF1

0.006
0.021
0.023
0.056

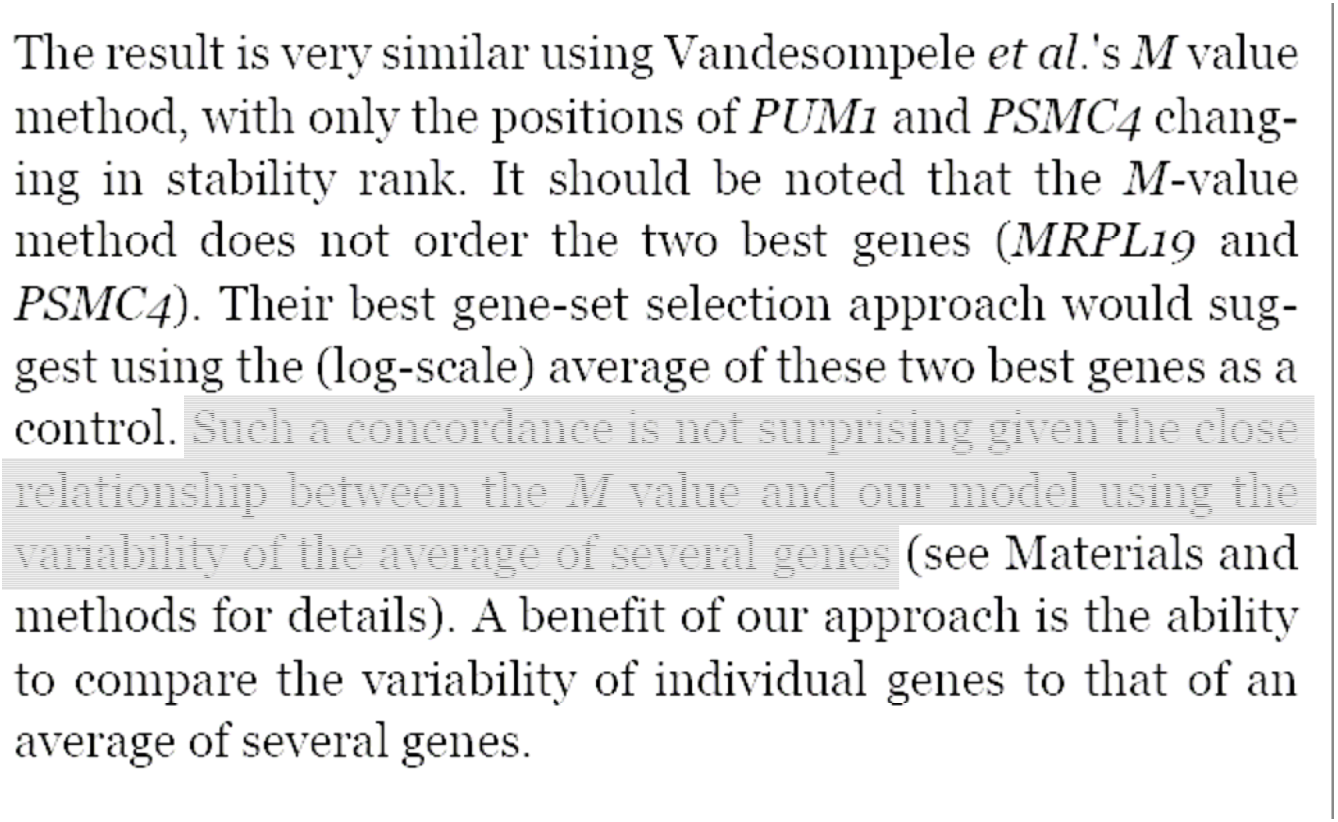
normalisation using multiple stable reference genes

- people really start to pay attention to the problem and are willing to deal with the issue
  - > 130 citations of our Genome Biology (2002) paper
  - > 1300 geNorm downloads in 40 countries

- other approaches
  - Global Pattern Recognition (Akilesh et al., Genome Research, 2003)
  - BestKeeper (Pfaffl et al., Biotechnology Letters, 2004)
  - Normfinder (Andersen et al., Cancer Research, 2004)
  - Szabo et al., Genome Biology, 2004

present mathematical (linear mixed-effects) models to further analyze candidate reference genes

$$\log yij = \mu + Ti + Gj + \varepsilon ij$$



The result is very similar using Vandesompele *et al.*'s *M* value method, with only the positions of *PUM1* and *PSMC4* changing in stability rank. It should be noted that the *M*-value method does not order the two best genes (*MRPL19* and *PSMC4*). Their best gene-set selection approach would suggest using the (log-scale) average of these two best genes as a control. Such a concordance is not surprising given the close relationship between the *M* value and our model using the variability of the average of several genes (see Materials and methods for details). A benefit of our approach is the ability to compare the variability of individual genes to that of an average of several genes.

Vandesompele *et al.*'s *M*-value is the average of relative standard deviations of the log-expression levels. Under Model 1, the *M*-value of the gene is closely related to its variance (under Models 2 and 3 below, the similar relationships can be derived):

$$V_{jk} = SD\left(\left\{\log\left(y_{ij}/y_{ik}\right)\right\}_{i=1}^{n}\right) = SD\left(\left\{\log\left(y_{ij}\right) - \log\left(y_{ik}\right)\right\}_{i=1}^{n}\right) = \sqrt{\sigma_j^2 + \sigma_k^2}$$

$$M_j = \sum_{\substack{k=1,\ldots,g \\ k \neq j}} V_{jk} / (g-1) = \sigma_j^2 \frac{\sum_{k \neq j} \sqrt{1 + \sigma_k^2 / \sigma_j^2}}{g-1}$$

$$\sigma_j^2 \sqrt{1 + 1/R^2} \le M_j \le \sigma_j^2 \sqrt{1 + R^2}, \text{ where } R = \max_{i,k} \sigma_k / \sigma_i$$

automated data-analysis

- limitations current qPCR gene expression analysis tools
  - only one reference gene
  - limited to one run/plate
  - limited number of samples or genes
  - fixed number of replicates
  - dedicated for one instrument
  - lack of data quality controls

    replicate variability

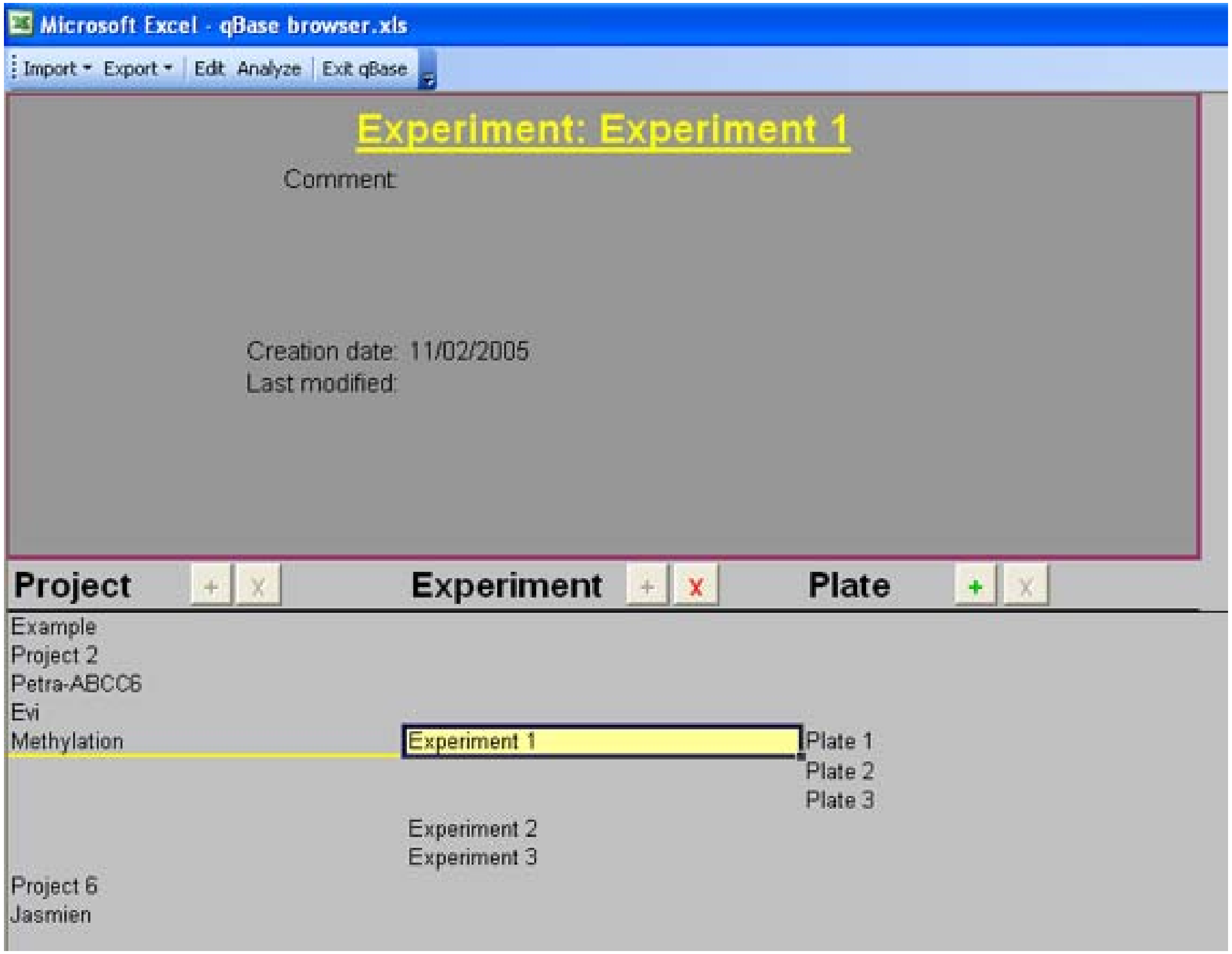
    standard curve

    NTC control
  - cumbersome data import
  - lack of experiment archive
  - inaccurate error propagation / quantification
  - limited visualisation / rescaling
  - closed architecture

automated data-analysis

- qBASE: qPCR data analysis and database-like software for gene expression analysis
  - 96/384/rotor-based formats
  - unlimited number of genes/samples/replicates
  - multiple reference genes for normalization
  - accurate error propagation and quantification
  - easy exchange of data between different users
  - database of raw data (Ct values) from all your qPCR runs organised into projects and experiments
  - data quality controls
  - data-analysis from multiple runs
  - rescaling and re-ordering options for result visualisation
  - free for non-commercial use
  - open source

experiment browser



Microsoft Excel - qBase browser.xls

Experiment: Experiment 1

Comment:

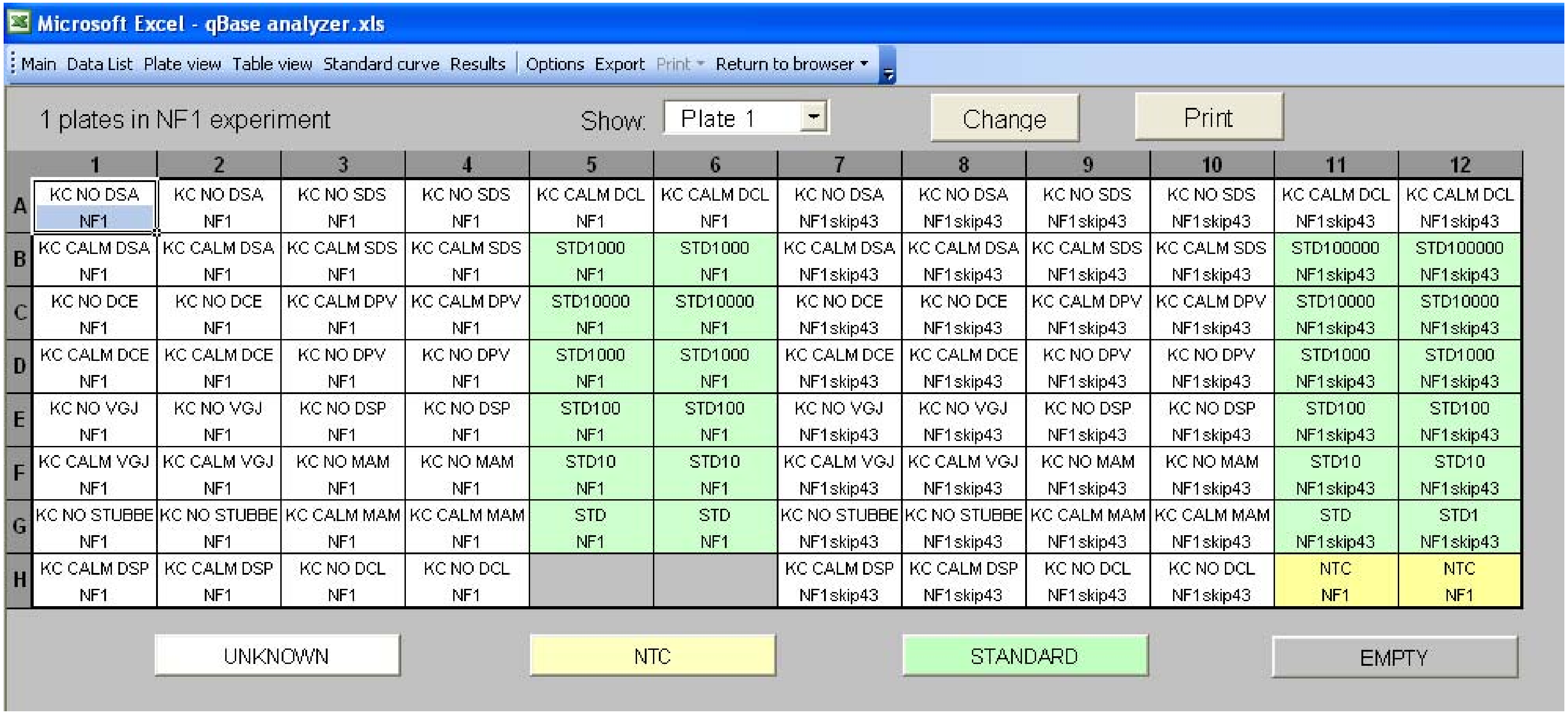
Creation date: 11/02/2005
Last modified:

Project + X Experiment + X Plate + X

Example
Project 2
Petra-ABCC6
Evi
Methylation — Experiment 1 — Plate 1
Plate 2
Plate 3
Experiment 2
Experiment 3
Project 6
Jasmien

- hierarchical organisation: projects > experiments > runs
- database of raw (Ct) data (instrument export files)

plate view & editing


Microsoft Excel - qBase analyzer.xls
Main Data List Plate view Table view Standard curve Results Options Export Print Return to browser
Change
Print

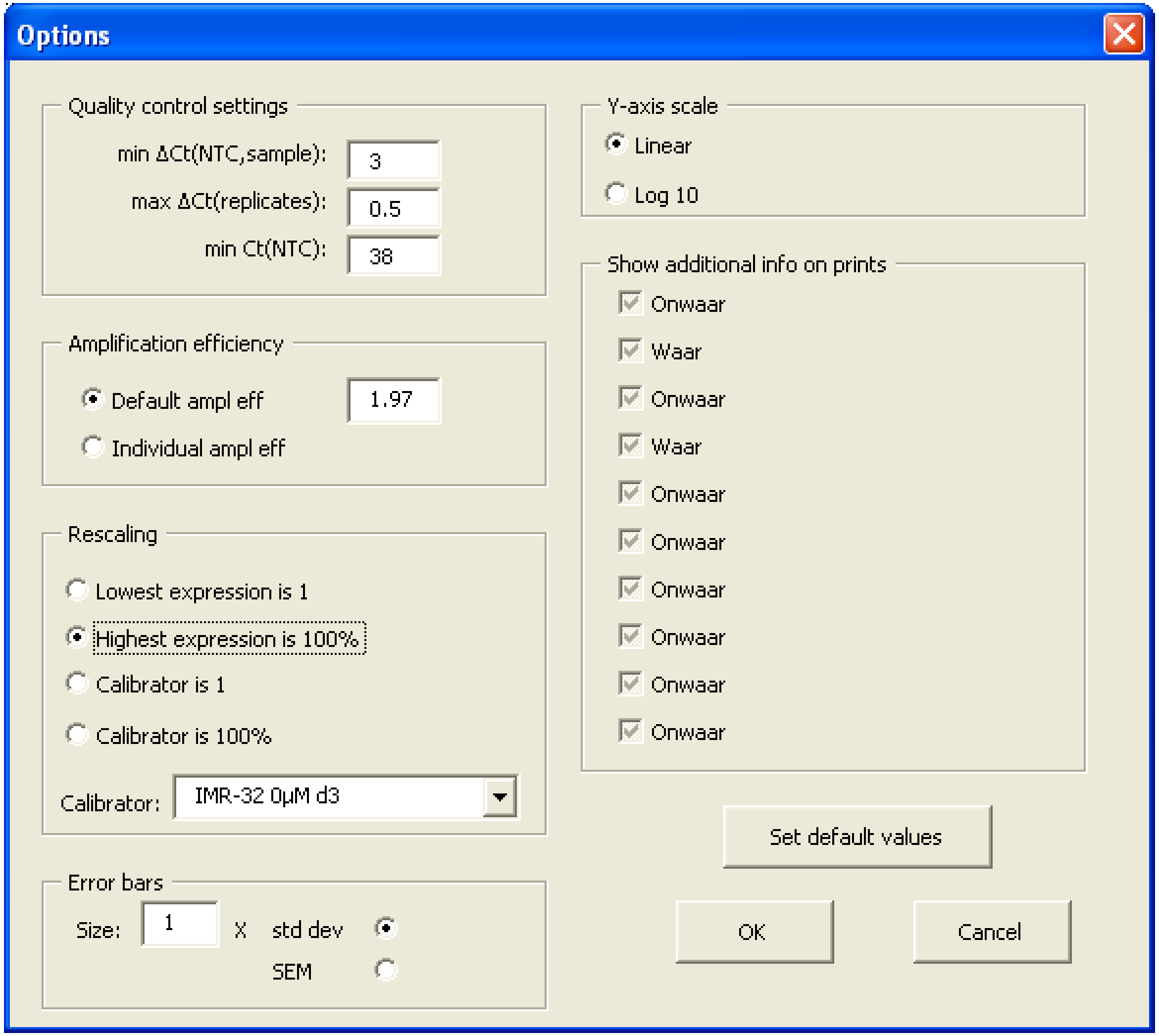
1 plates in NF1 experiment

Show: Plate 1

	1	2	3	4	5	6	7	8	9	10	11	12
A	KC NO DSA NF1	KC NO DSA NF1	KC NO SDS NF1	KC NO SDS NF1	KC CALM DCL NF1	KC CALM DCL NF1	KC NO DSA NF1skip43	KC NO DSA NF1skip43	KC NO SDS NF1skip43	KC NO SDS NF1skip43	KC CALM DCL NF1skip43	KC CALM DCL NF1skip43
B	KC CALM DSA NF1	KC CALM DSA NF1	KC CALM SDS NF1	KC CALM SDS NF1	STD1000 NF1	STD1000 NF1	KC CALM DSA NF1skip43	KC CALM DSA NF1skip43	KC CALM SDS NF1skip43	KC CALM SDS NF1skip43	STD100000 NF1skip43	STD100000 NF1skip43
C	KC NO DCE NF1	KC NO DCE NF1	KC CALM DPV NF1	KC CALM DPV NF1	STD10000 NF1	STD10000 NF1	KC NO DCE NF1skip43	KC NO DCE NF1skip43	KC CALM DPV NF1skip43	KC CALM DPV NF1skip43	STD10000 NF1skip43	STD10000 NF1skip43
D	KC CALM DCE NF1	KC CALM DCE NF1	KC NO DPV NF1	KC NO DPV NF1	STD1000 NF1	STD1000 NF1	KC CALM DCE NF1skip43	KC CALM DCE NF1skip43	KC NO DPV NF1skip43	KC NO DPV NF1skip43	STD1000 NF1skip43	STD1000 NF1skip43
E	KC NO VGJ NF1	KC NO VGJ NF1	KC NO DSP NF1	KC NO DSP NF1	STD100 NF1	STD100 NF1	KC NO VGJ NF1skip43	KC NO VGJ NF1skip43	KC NO DSP NF1skip43	KC NO DSP NF1skip43	STD100 NF1skip43	STD100 NF1skip43
F	KC CALM VGJ NF1	KC CALM VGJ NF1	KC NO MAM NF1	KC NO MAM NF1	STD10 NF1	STD10 NF1	KC CALM VGJ NF1skip43	KC CALM VGJ NF1skip43	KC NO MAM NF1skip43	KC NO MAM NF1skip43	STD10 NF1skip43	STD10 NF1skip43
G	KC NO STUBBE NF1	KC NO STUBBE NF1	KC CALM MAM NF1	KC CALM MAM NF1	STD NF1	STD NF1	KC NO STUBBE NF1skip43	KC NO STUBBE NF1skip43	KC CALM MAM NF1skip43	KC CALM MAM NF1skip43	STD NF1skip43	STD1 NF1skip43
H	KC CALM DSP NF1	KC CALM DSP NF1	KC NO DCL NF1	KC NO DCL NF1			KC CALM DSP NF1skip43	KC CALM DSP NF1skip43	KC NO DCL NF1skip43	KC NO DCL NF1skip43	NTC NF1	NTC NF1

- sample name, gene name, sample type (NTC, UNKN, STD), STD quantity
- easy editing

options



Options

Quality control settings

min ΔCt(NTC,sample): 3

max ΔCt(replicates): 0.5

min Ct(NTC): 38

Amplification efficiency

● Default ampl eff 1.97

○ Individual ampl eff

Rescaling

○ Lowest expression is 1

● Highest expression is 100%

○ Calibrator is 1

○ Calibrator is 100%

Calibrator: IMR-32 0µM d3

Error bars

Size: 1 X std dev ●

SEM ○

Y-axis scale

● Linear

○ Log 10

Show additional info on prints

☑ Onwaar

☑ Waar

☑ Onwaar

☑ Waar

☑ Onwaar

☑ Onwaar

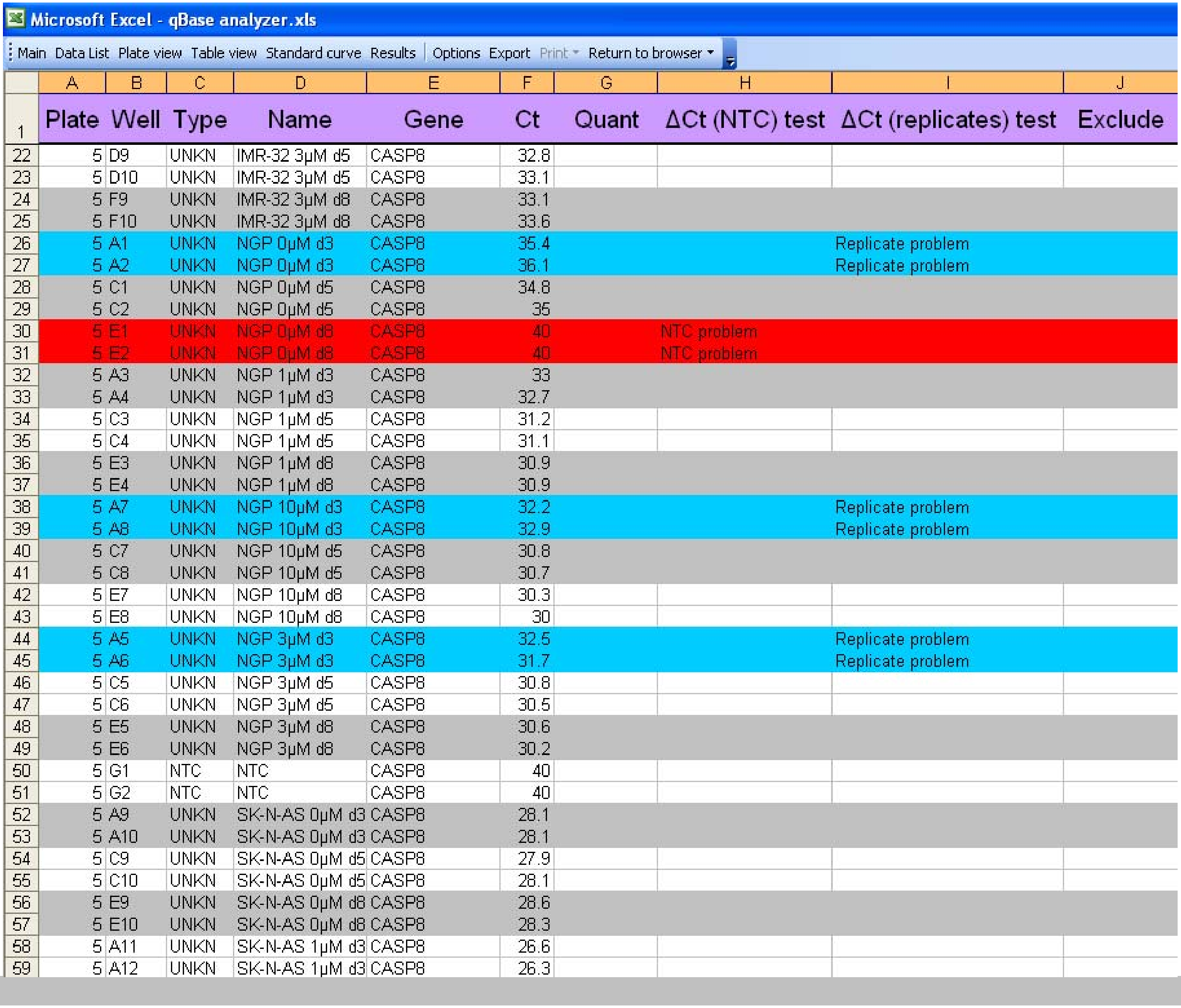
☑ Onwaar

☑ Onwaar

☑ Onwaar

☑ Onwaar

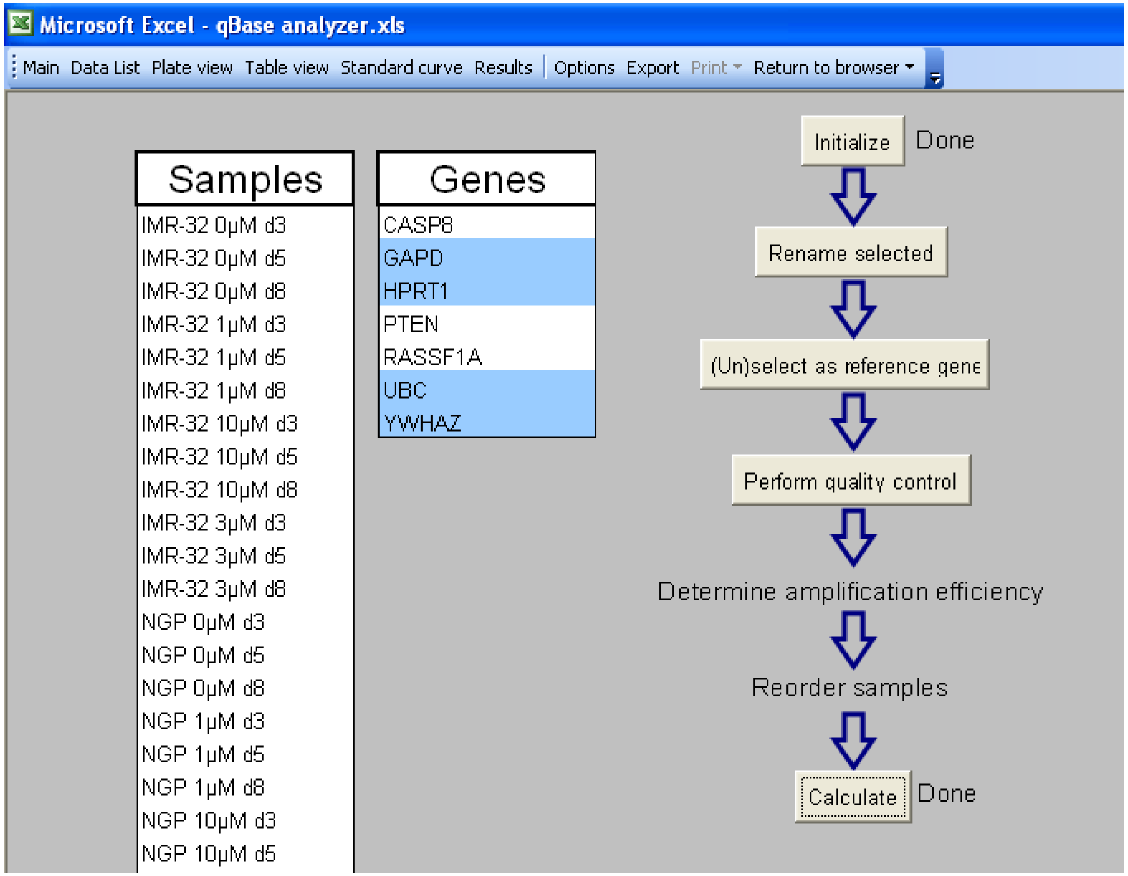
raw data list – quality controlled


Microsoft Excel - qBase analyzer.xls
Main Data List Plate view Table view Standard curve Results Options Export Print Return to browser

	A	B	C	D	E	F	G	H	I	J
1	Plate	Well	Type	Name	Gene	Ct	Quant	ΔCt (NTC) test	ΔCt (replicates) test	Exclude
22	5	D9	UNKN	IMR-32 3µM d5	CASP8	32.8				
23	5	D10	UNKN	IMR-32 3µM d5	CASP8	33.1				
24	5	F9	UNKN	IMR-32 3µM d8	CASP8	33.1				
25	5	F10	UNKN	IMR-32 3µM d8	CASP8	33.6				
26	5	A1	UNKN	NGP 0µM d3	CASP8	35.4			Replicate problem	
27	5	A2	UNKN	NGP 0µM d3	CASP8	36.1			Replicate problem	
28	5	C1	UNKN	NGP 0µM d5	CASP8	34.8				
29	5	C2	UNKN	NGP 0µM d5	CASP8	35				
30	5	E1	UNKN	NGP 0µM d8	CASP8	40		NTC problem		
31	5	E2	UNKN	NGP 0µM d8	CASP8	40		NTC problem		
32	5	A3	UNKN	NGP 1µM d3	CASP8	33				
33	5	A4	UNKN	NGP 1µM d3	CASP8	32.7				
34	5	C3	UNKN	NGP 1µM d5	CASP8	31.2				
35	5	C4	UNKN	NGP 1µM d5	CASP8	31.1				
36	5	E3	UNKN	NGP 1µM d8	CASP8	30.9				
37	5	E4	UNKN	NGP 1µM d8	CASP8	30.9				
38	5	A7	UNKN	NGP 10µM d3	CASP8	32.2			Replicate problem	
39	5	A8	UNKN	NGP 10µM d3	CASP8	32.9			Replicate problem	
40	5	C7	UNKN	NGP 10µM d5	CASP8	30.8				
41	5	C8	UNKN	NGP 10µM d5	CASP8	30.7				
42	5	E7	UNKN	NGP 10µM d8	CASP8	30.3				
43	5	E8	UNKN	NGP 10µM d8	CASP8	30				
44	5	A5	UNKN	NGP 3µM d3	CASP8	32.5			Replicate problem	
45	5	A6	UNKN	NGP 3µM d3	CASP8	31.7			Replicate problem	
46	5	C5	UNKN	NGP 3µM d5	CASP8	30.8				
47	5	C6	UNKN	NGP 3µM d5	CASP8	30.5				
48	5	E5	UNKN	NGP 3µM d8	CASP8	30.6				
49	5	E6	UNKN	NGP 3µM d8	CASP8	30.2				
50	5	G1	NTC	NTC	CASP8	40				
51	5	G2	NTC	NTC	CASP8	40				
52	5	A9	UNKN	SK-N-AS 0µM d3	CASP8	28.1				
53	5	A10	UNKN	SK-N-AS 0µM d3	CASP8	28.1				
54	5	C9	UNKN	SK-N-AS 0µM d5	CASP8	27.9				
55	5	C10	UNKN	SK-N-AS 0µM d5	CASP8	28.1				
56	5	E9	UNKN	SK-N-AS 0µM d8	CASP8	28.6				
57	5	E10	UNKN	SK-N-AS 0µM d8	CASP8	28.3				
58	5	A11	UNKN	SK-N-AS 1µM d3	CASP8	26.6				
59	5	A12	UNKN	SK-N-AS 1µM d3	CASP8	26.3				

Center for
Medical
Genetics

main view



Samples

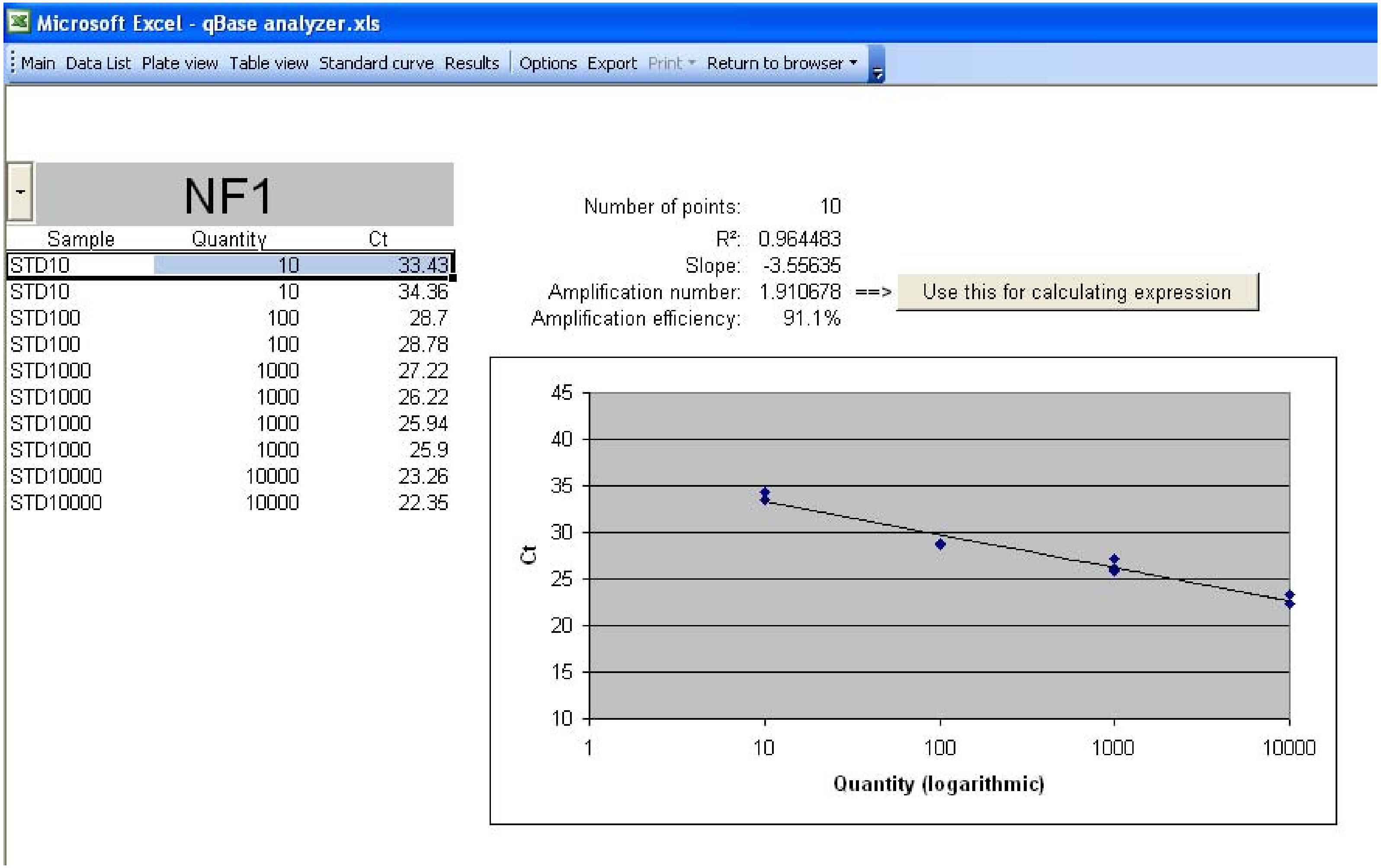
IMR-32 0μM d3
IMR-32 0μM d5
IMR-32 0μM d8
IMR-32 1μM d3
IMR-32 1μM d5
IMR-32 1μM d8
IMR-32 10μM d3
IMR-32 10μM d5
IMR-32 10μM d8
IMR-32 3μM d3
IMR-32 3μM d5
IMR-32 3μM d8
NGP 0μM d3
NGP 0μM d5
NGP 0μM d8
NGP 1μM d3
NGP 1μM d5
NGP 1μM d8
NGP 10μM d3
NGP 10μM d5

Genes

CASP8
GAPD
HPRT1
PTEN
RASSF1A
UBC
YWHAZ

- samples and (reference) genes (from multiple runs belonging to the same experiment)
- data processing workflow

standard curve – efficiency estimation

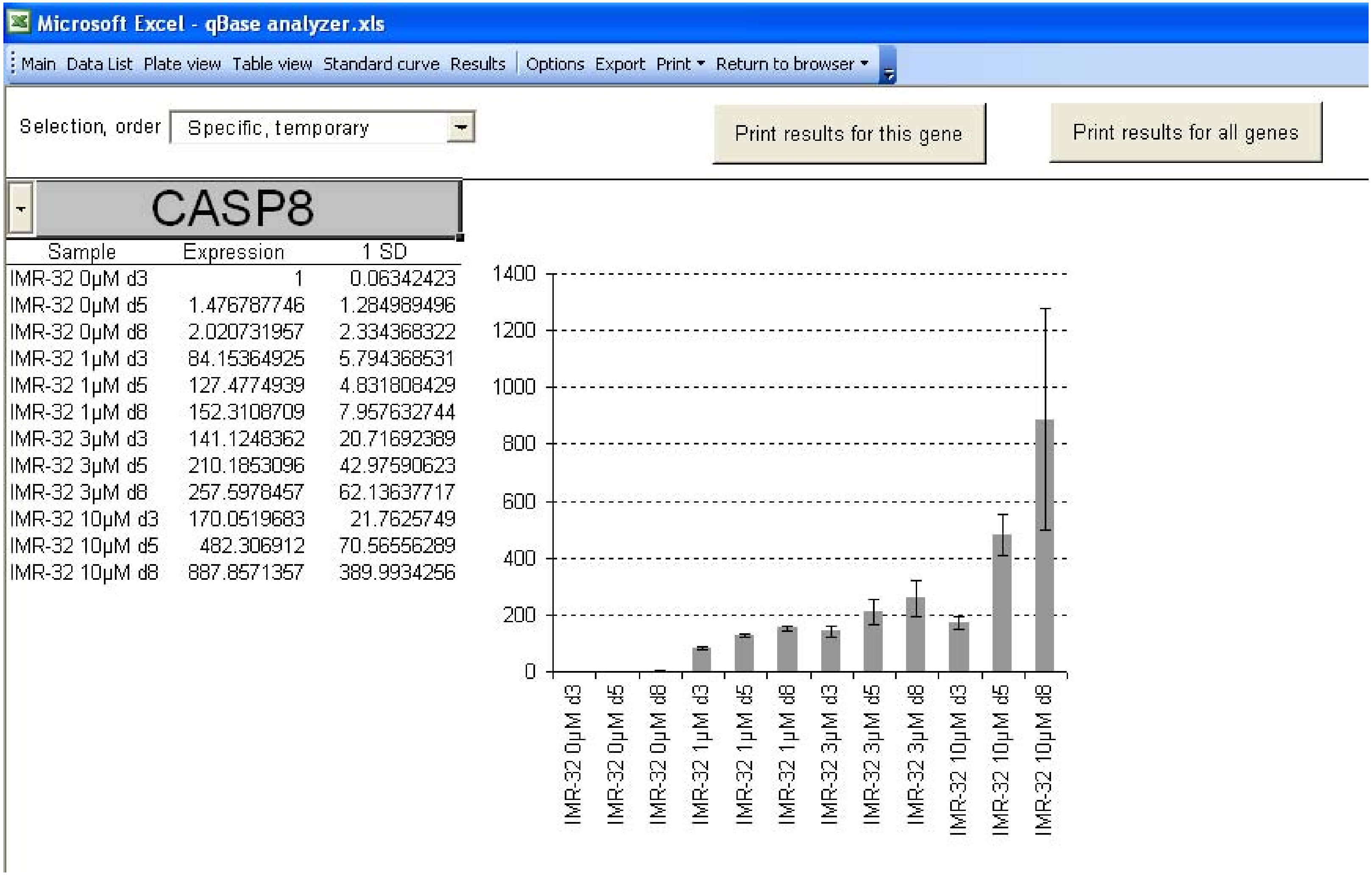


NF1

Sample	Quantity	Ct
STD10	10	33.43
STD10	10	34.36
STD100	100	28.7
STD100	100	28.78
STD1000	1000	27.22
STD1000	1000	26.22
STD1000	1000	25.94
STD1000	1000	25.9
STD10000	10000	23.26
STD10000	10000	22.35

Number of points:	10
R²:	0.964483
Slope:	-3.55635
Amplification number:	1.910678 ==>
Amplification efficiency:	91.1%

result viewer

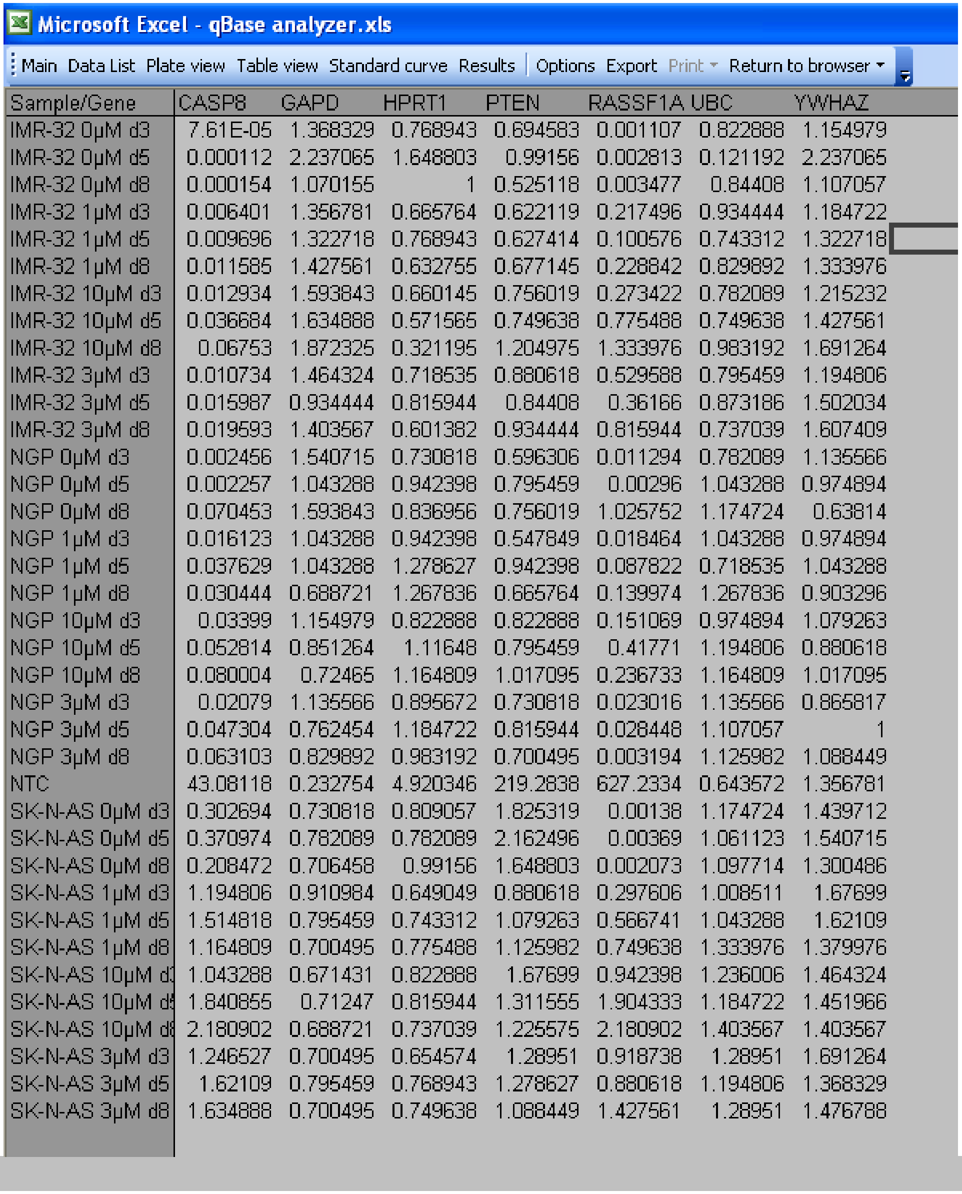


Selection, order: Specific, temporary

CASP8

Sample	Expression	1 SD
IMR-32 0µM d3	1	0.06342423
IMR-32 0µM d5	1.476787746	1.284989496
IMR-32 0µM d8	2.020731957	2.334368322
IMR-32 1µM d3	84.15364925	5.794368531
IMR-32 1µM d5	127.4774939	4.831808429
IMR-32 1µM d8	152.3108709	7.957632744
IMR-32 3µM d3	141.1248362	20.71692389
IMR-32 3µM d5	210.1853096	42.97590623
IMR-32 3µM d8	257.5978457	62.13637717
IMR-32 10µM d3	170.0519683	21.7625749
IMR-32 10µM d5	482.306912	70.56556289
IMR-32 10µM d8	887.8571357	389.9934256

tabulated expression levels



Sample/Gene	CASP8	GAPD	HPRT1	PTEN	RASSF1A	UBC	YWHAZ
IMR-32 0µM d3	7.61E-05	1.368329	0.768943	0.694583	0.001107	0.822888	1.154979
IMR-32 0µM d5	0.000112	2.237065	1.648803	0.99156	0.002813	0.121192	2.237065
IMR-32 0µM d8	0.000154	1.070155	1	0.525118	0.003477	0.84408	1.107057
IMR-32 1µM d3	0.006401	1.356781	0.665764	0.622119	0.217496	0.934444	1.184722
IMR-32 1µM d5	0.009696	1.322718	0.768943	0.627414	0.100576	0.743312	1.322718
IMR-32 1µM d8	0.011585	1.427561	0.632755	0.677145	0.228842	0.829892	1.333976
IMR-32 10µM d3	0.012934	1.593843	0.660145	0.756019	0.273422	0.782089	1.215232
IMR-32 10µM d5	0.036684	1.634888	0.571565	0.749638	0.775488	0.749638	1.427561
IMR-32 10µM d8	0.06753	1.872325	0.321195	1.204975	1.333976	0.983192	1.691264
IMR-32 3µM d3	0.010734	1.464324	0.718535	0.880618	0.529588	0.795459	1.194806
IMR-32 3µM d5	0.015987	0.934444	0.815944	0.84408	0.36166	0.873186	1.502034
IMR-32 3µM d8	0.019593	1.403567	0.601382	0.934444	0.815944	0.737039	1.607409
NGP 0µM d3	0.002456	1.540715	0.730818	0.596306	0.011294	0.782089	1.135566
NGP 0µM d5	0.002257	1.043288	0.942398	0.795459	0.00296	1.043288	0.974894
NGP 0µM d8	0.070453	1.593843	0.836956	0.756019	1.025752	1.174724	0.63814
NGP 1µM d3	0.016123	1.043288	0.942398	0.547849	0.018464	1.043288	0.974894
NGP 1µM d5	0.037629	1.043288	1.278627	0.942398	0.087822	0.718535	1.043288
NGP 1µM d8	0.030444	0.688721	1.267836	0.665764	0.139974	1.267836	0.903296
NGP 10µM d3	0.03399	1.154979	0.822888	0.822888	0.151069	0.974894	1.079263
NGP 10µM d5	0.052814	0.851264	1.11648	0.795459	0.41771	1.194806	0.880618
NGP 10µM d8	0.080004	0.72465	1.164809	1.017095	0.236733	1.164809	1.017095
NGP 3µM d3	0.02079	1.135566	0.895672	0.730818	0.023016	1.135566	0.865817
NGP 3µM d5	0.047304	0.762454	1.184722	0.815944	0.028448	1.107057	1
NGP 3µM d8	0.063103	0.829892	0.983192	0.700495	0.003194	1.125982	1.088449
NTC	43.08118	0.232754	4.920346	219.2838	627.2334	0.643572	1.356781
SK-N-AS 0µM d3	0.302694	0.730818	0.809057	1.825319	0.00138	1.174724	1.439712
SK-N-AS 0µM d5	0.370974	0.782089	0.782089	2.162496	0.00369	1.061123	1.540715
SK-N-AS 0µM d8	0.208472	0.706458	0.99156	1.648803	0.002073	1.097714	1.300486
SK-N-AS 1µM d3	1.194806	0.910984	0.649049	0.880618	0.297606	1.008511	1.67699
SK-N-AS 1µM d5	1.514818	0.795459	0.743312	1.079263	0.566741	1.043288	1.62109
SK-N-AS 1µM d8	1.164809	0.700495	0.775488	1.125982	0.749638	1.333976	1.379976
SK-N-AS 10µM d3	1.043288	0.671431	0.822888	1.67699	0.942398	1.236006	1.464324
SK-N-AS 10µM d5	1.840855	0.71247	0.815944	1.311555	1.904333	1.184722	1.451966
SK-N-AS 10µM d8	2.180902	0.688721	0.737039	1.225575	2.180902	1.403567	1.403567
SK-N-AS 3µM d3	1.246527	0.700495	0.654574	1.28951	0.918738	1.28951	1.691264
SK-N-AS 3µM d5	1.62109	0.795459	0.768943	1.278627	0.880618	1.194806	1.368329
SK-N-AS 3µM d8	1.634888	0.700495	0.749638	1.088449	1.427561	1.28951	1.476788

- ready for export to other applications (e.g. dedicated statistical software)

acknowledgments

- Jan Hellemans
- Katleen De Preter
- Filip Pattyn
- Els De Smet
- Nurten Yigit
- Anne De Paepe
- Geert Mortier
- Frank Speleman

qBase

geNorm

Joke.Vandesompele@UGent.be
http://medgen.ugent.be

Center for
Medical
Genetics