# Normalization of gene expression: state of the art and preview on a new strategy using expressed Alu repeats

Jo Vandesompele
Center for Medical Genetics
Ghent University Hospital, Belgium

2nd International qPCR Symposium
September 6, 2005
Freising-Weihenstephan, Germany

# outline

- our software tools along the qPCR workflow
- accurate normalisation of gene expression using multiple references genes
  - geNorm concept
  - other approaches
- EAR normalisation
  - expressed Alu repeats as references
  - gene expression normalisation
  - gene copy number (DNA) quantification

# qPCR tools from our center

- RTPrimerDB (Pattyn et al., Nucleic Acids Research, 2003)
  http://medgen.ugent.be/rtprimerdb/
  - primer and probe database of experimentally verified assays
  - in silico assay evaluation
  - gene expression assay viewer
  - Wednesday 7th, Bioinformatics Session (11:40)

- geNorm (Vandesompele et al., Genome Biology, 2002)
  http://medgen.ugent.be/genorm/
  - reference gene validation and normalisation
  - this session

- qBase (Hellemans et al., in preparation)
  http://medgen.ugent.be/qbase/
  - relative quantification software for management and automated data analysis
  - Wednesday 7th, Bioinformatics Session (9:40)

# normalisation: what's the problem ?

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

# normalisation: what's the solution (part I)?

- Huggett et al., Genes and Immunity, 2005
  *Real-time RT-PCR normalisation; strategies and considerations*

- sampling size (number of cells, volume or mass of the sample)
  - reproducible extraction
  - not always possible (e.g. microdissected tissue)
- total RNA amount
  - not always possible (e.g. embryo)
  - quality (inhibitors)
  - cDNA synthesis efficiency is not taken into account
  - total RNA (rRNA) is not always representative of the mRNA fraction
- spiking (alien RNA)
  - corrects for enzymatic efficiency differences
  - not assumption-free (equal input template)

- Ståhlberg et al., Clinical Chemistry, 2004

# normalisation: what's the solution (part II)?

- reference genes
  - most popular
  - captures most variation
- attention!
  - reference genes (might) vary in expression
  - until recently, non-validated reference genes were used (assuming stable expression)
- normalisation against 3 or more validated reference genes is considered as the most appropriate and universally applicable method
  - 3rd London qPCR Symposium (April 2005)
  - which genes?
  - how to do the calculations?

# normalisation: our (geNorm) solution

- framework for qPCR gene expression normalisation using the reference gene concept:
  - 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
  - Vandesompele et al., Genome Biology, 2002


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

# geNorm expression stability parameter

- pairwise variation V (between 2 genes)

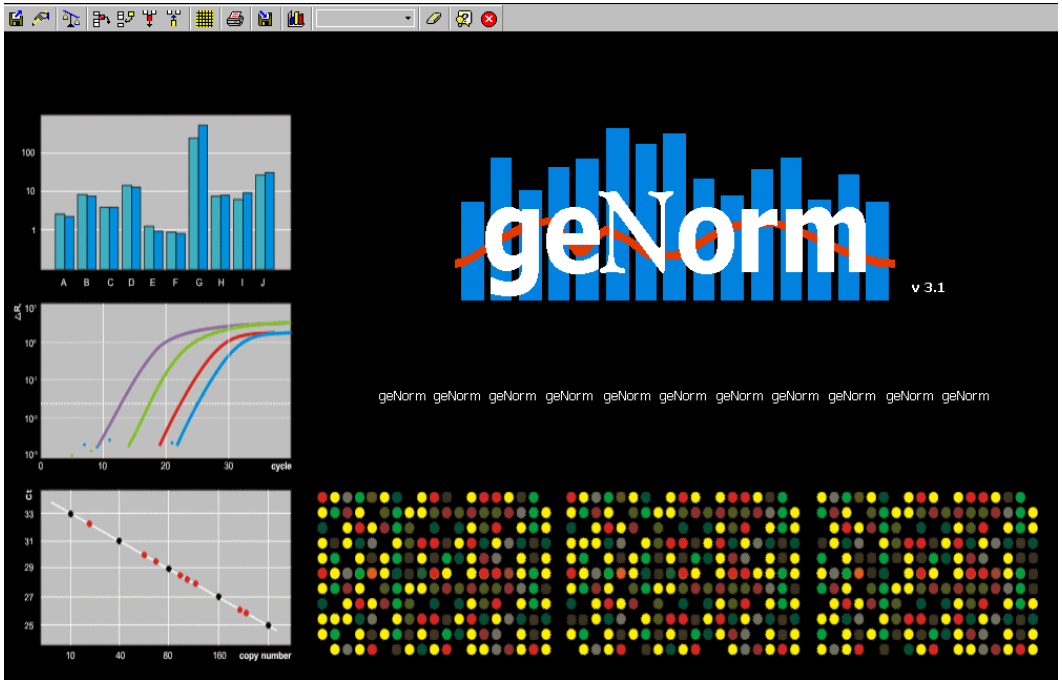
| | gene A | gene B | |
|---|---|---|---|
| sample 1 | a1 | b1 | log2(a1/b1) |
| sample 2 | a2 | b2 | log2(a2/b2) |
| sample 3 | a3 | b3 | log2(a3/b3) |
| … | … | … | … |
| sample n | an | bn | log2(an/bn) |

standard deviation = V

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

# geNorm

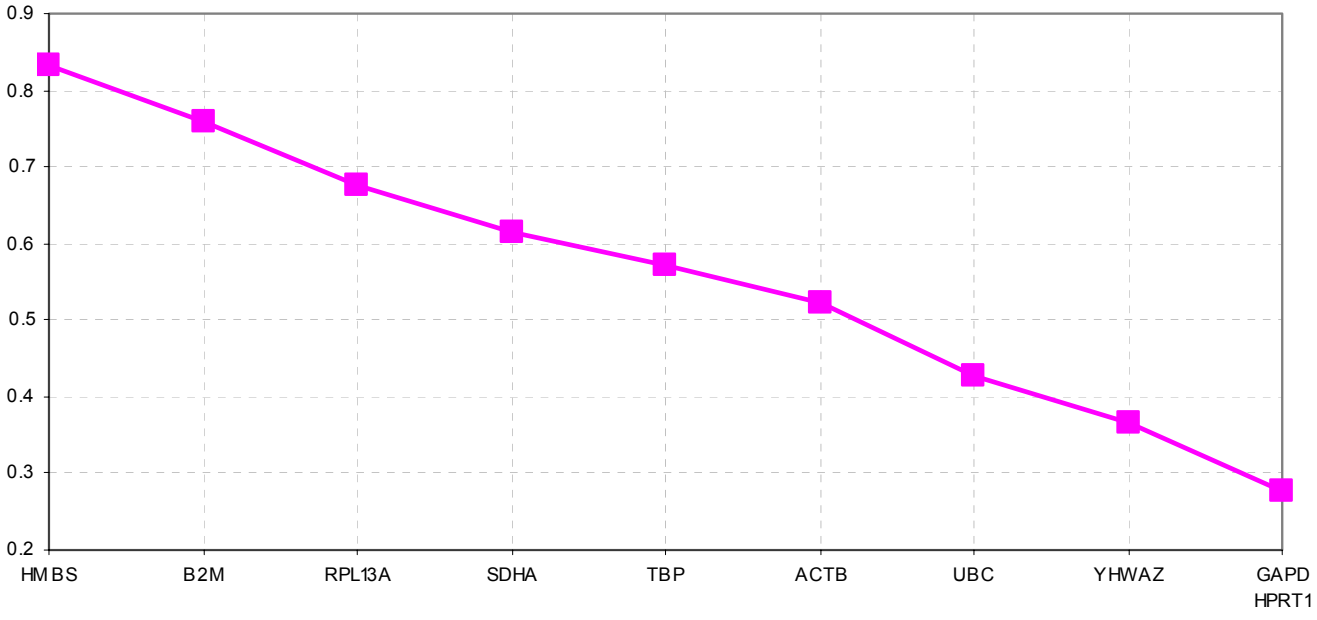
- automated analysis
  - ranking of candidate reference genes according to their stability
  - determination of how many genes are required for reliable normalization




http://medgen.ugent.be/genorm/

# geNorm

- ranking of candidate reference genes according to their stability



# calculation of the normalization factor

- geometric mean of 3 reference gene expression levels

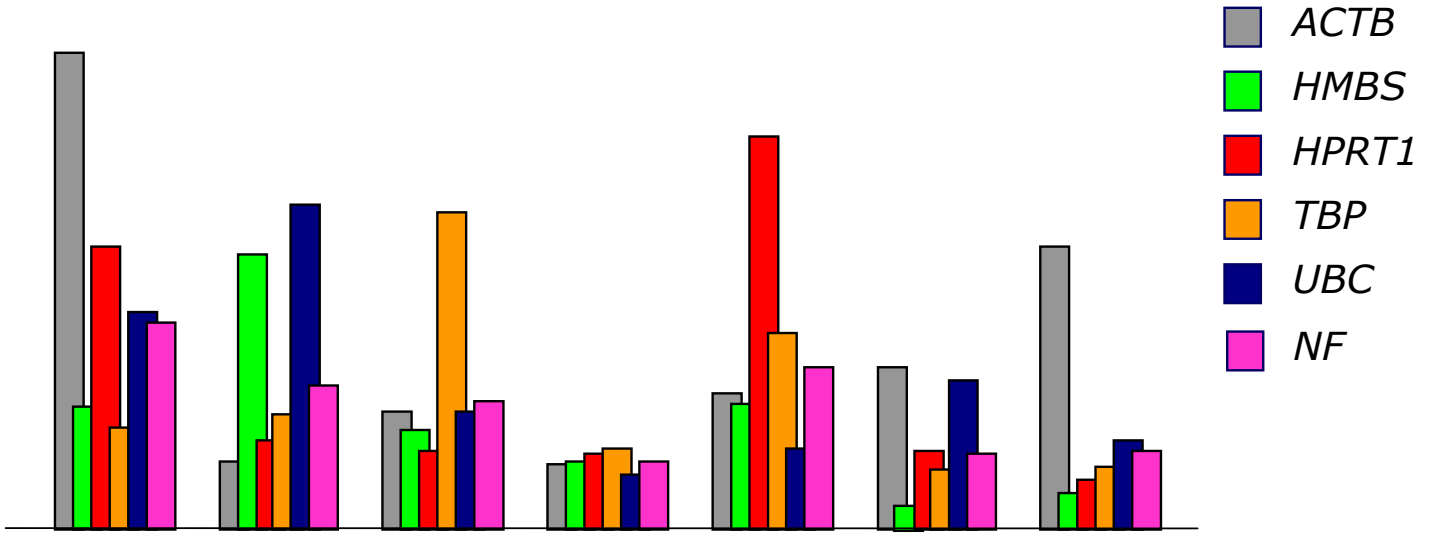
$$\text{geometric mean} = (a \times b \times c)^{1/3}$$

$$\text{arithmetic mean} = \frac{a + b + c}{3}$$

- controls for outliers
- compensates for differences in expression level between the reference genes

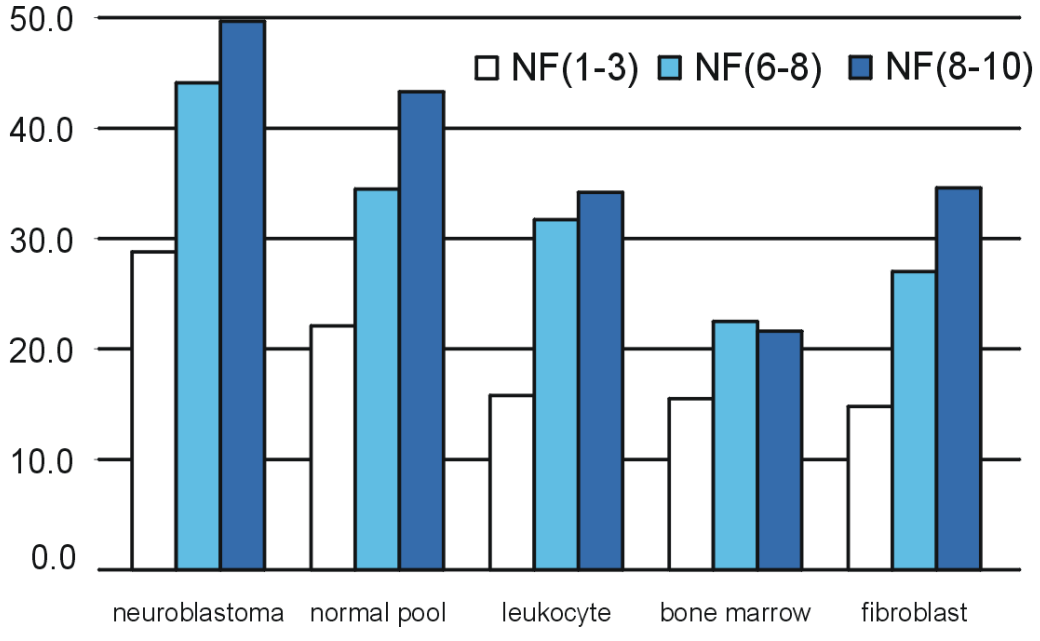
# geNorm validation (I)

- robust – insensitive to outliers



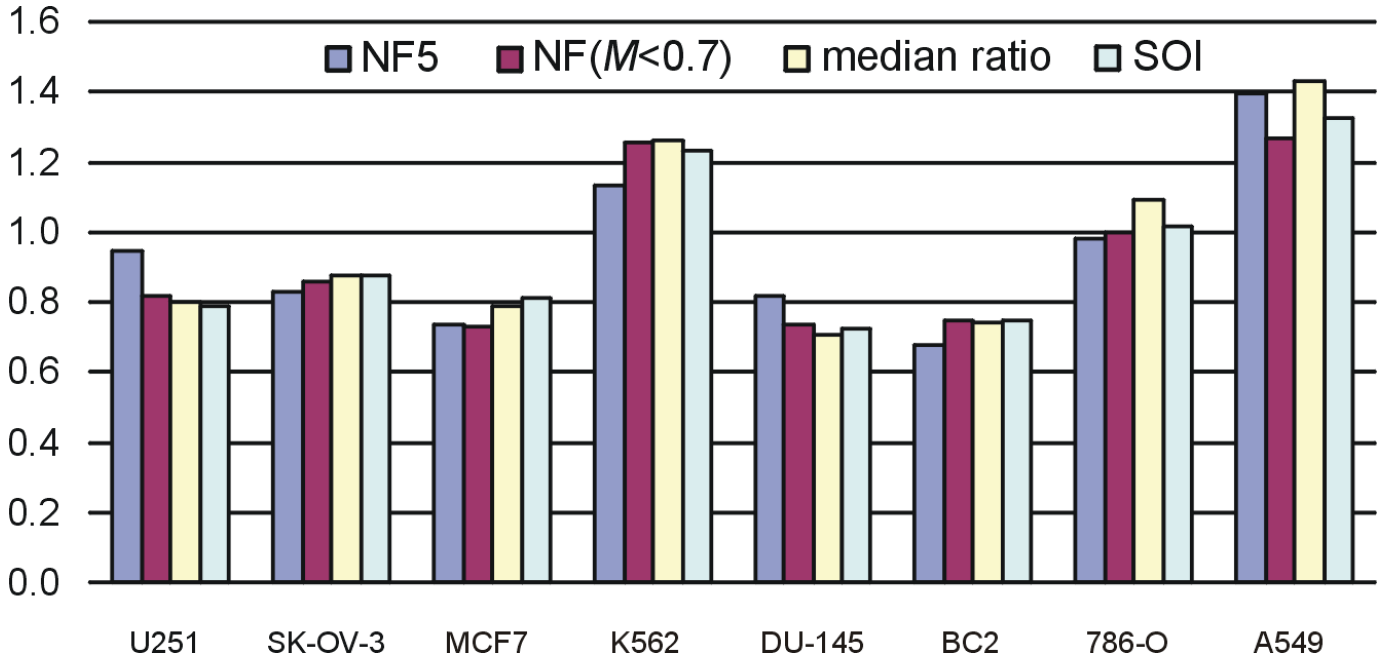
# geNorm validation (II)

- purpose of normalization: removal of non-specific variation



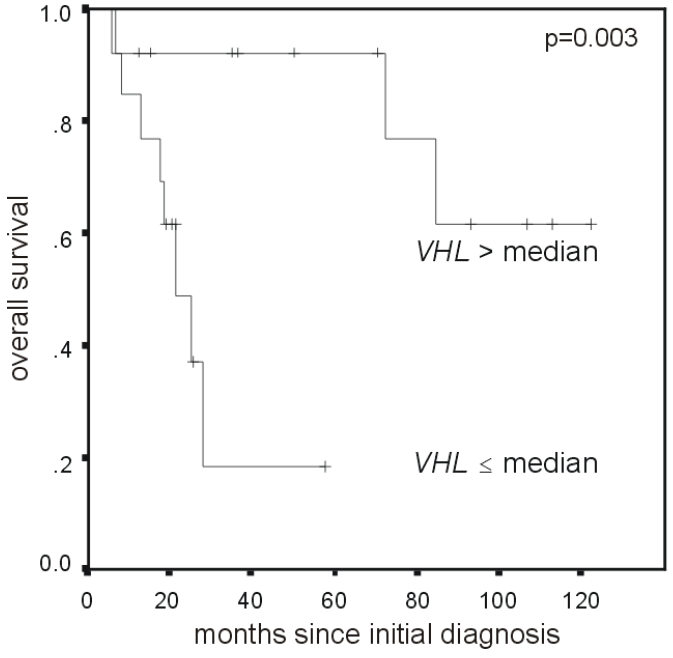
# geNorm validation (III)

- comparison with microarray normalization factors



# geNorm validation (IV)

- cancer patients survival curve




log rank statistics

NF4

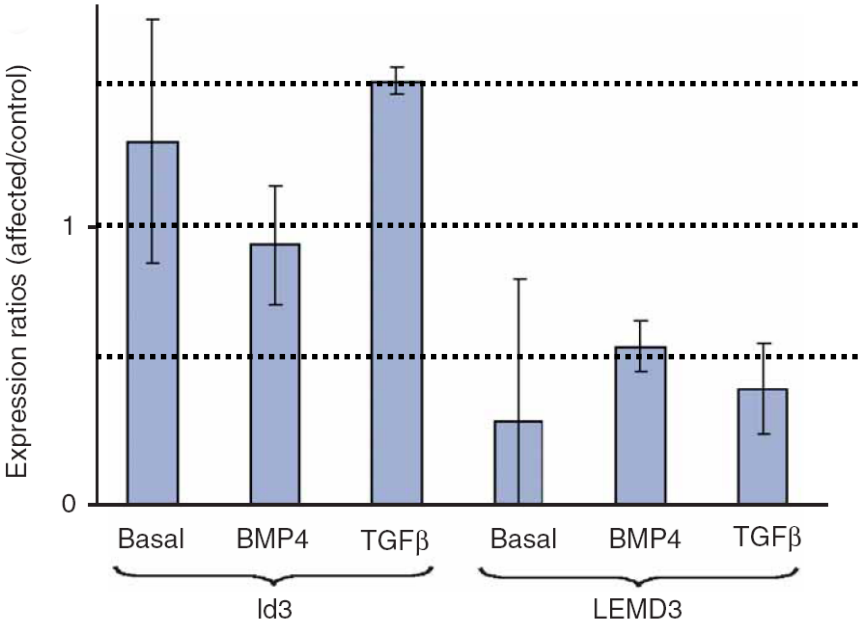
0.003

NF1

0.006
0.021
0.023
0.056

# geNorm validation (V)

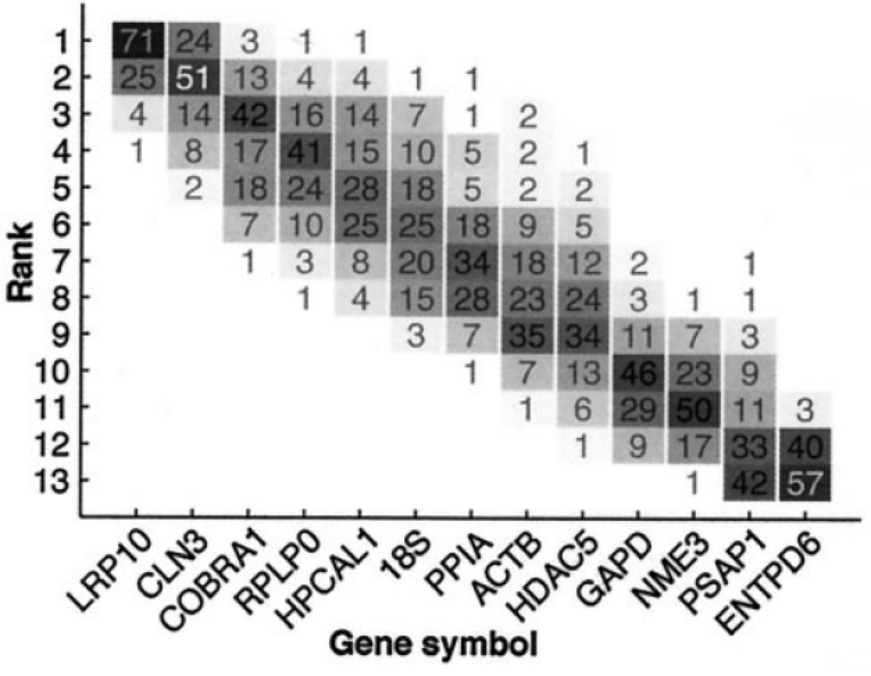
- mRNA haploinsufficiency measurements




Hellemans et al., Nature Genetics, 2004

# geNorm validation (VI)

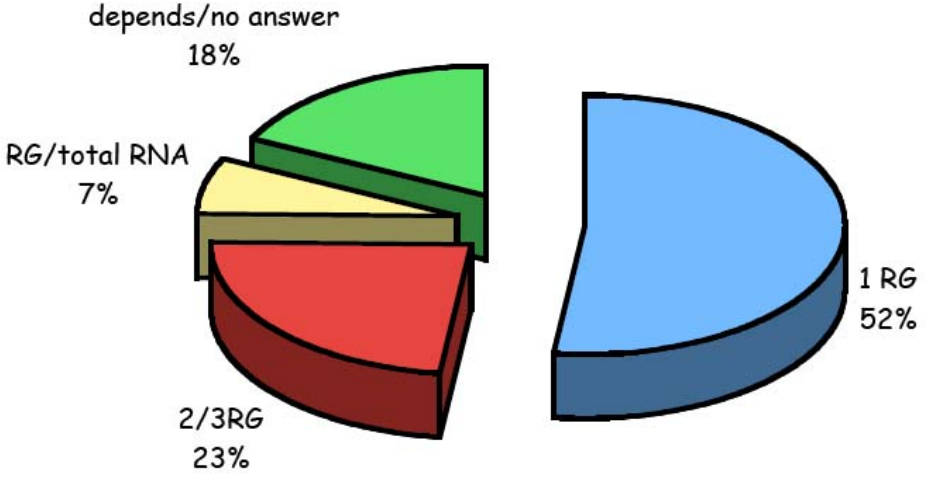
- bootstrapped version of geNorm
  - leaving out samples
  - leaving out outliers log ratios




Gabrielsson et al., Obesity Research, 2005

# normalisation using multiple stable reference genes

- people really start to pay attention to the problem and are willing to deal with the issue
  - > 200 citations of our Genome Biology (2002) paper
  - > 1800 geNorm downloads in 50 countries
  - 3rd London qPCR Symposium survey / EMBO 2005 qPCR course



# selection of stable reference genes

- other approaches
  - Global Pattern Recognition (Akilesh et al., Genome Research, 2003)
  - BestKeeper (Pfaffl et al., Biotechnology Letters, 2004)
  - Equivalence test (Haller et al., Analytical Biochemistry, 2004)
  - ANOVA test (Brunner et al., BMC Plant Biology, 2004)
  - Normfinder (Andersen et al., Cancer Research, 2004)
  - Szabo et al., Genome Biology, 2004
  - Abruzzo et al., Biotechniques, 2005

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

  $$\log yij = \mu + \mathrm{Ti} + \mathrm{Gj} + \varepsilon\mathrm{ij}$$

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):

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

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

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

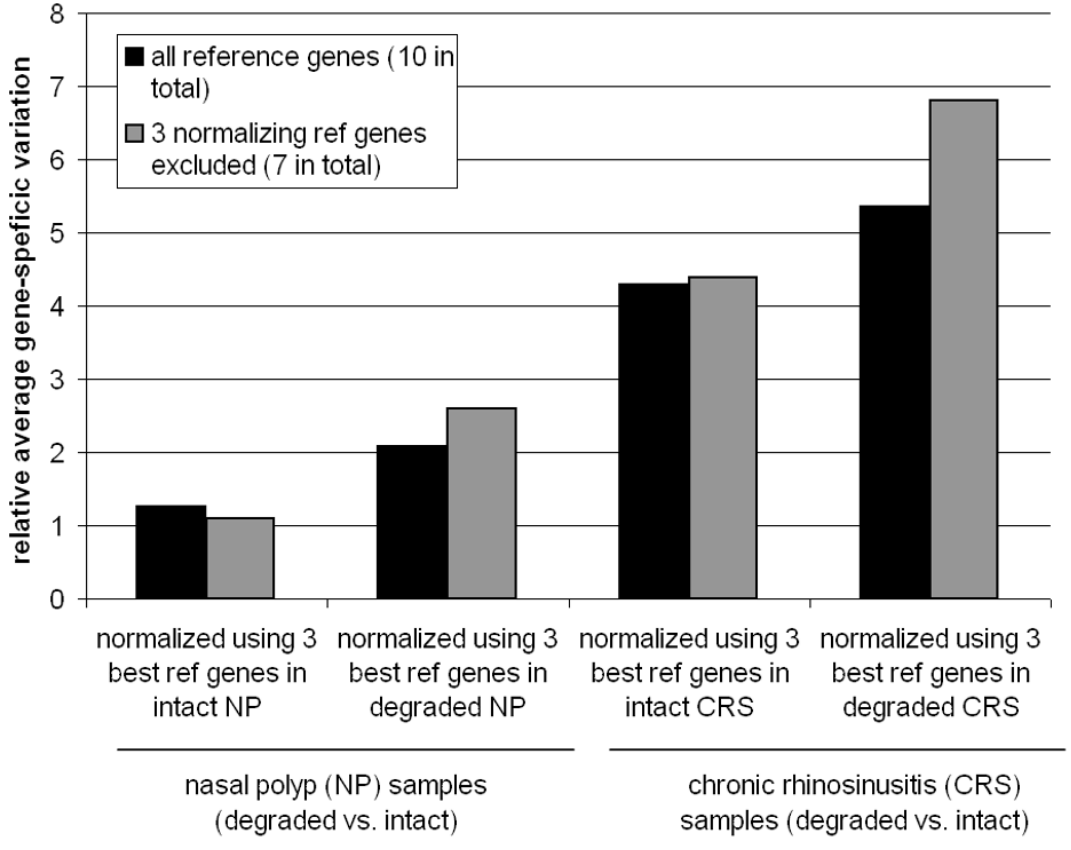
# impact of RNA quality on expression stability

- differences in reference gene ranking
  (Perez-Novo et al., Biotechniques, 2005)

| 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 |

# impact of RNA quality on expression stability

- higher variation in degraded samples
  (Perez-Novo et al., Biotechniques, 2005)




Center for Medical Genetics

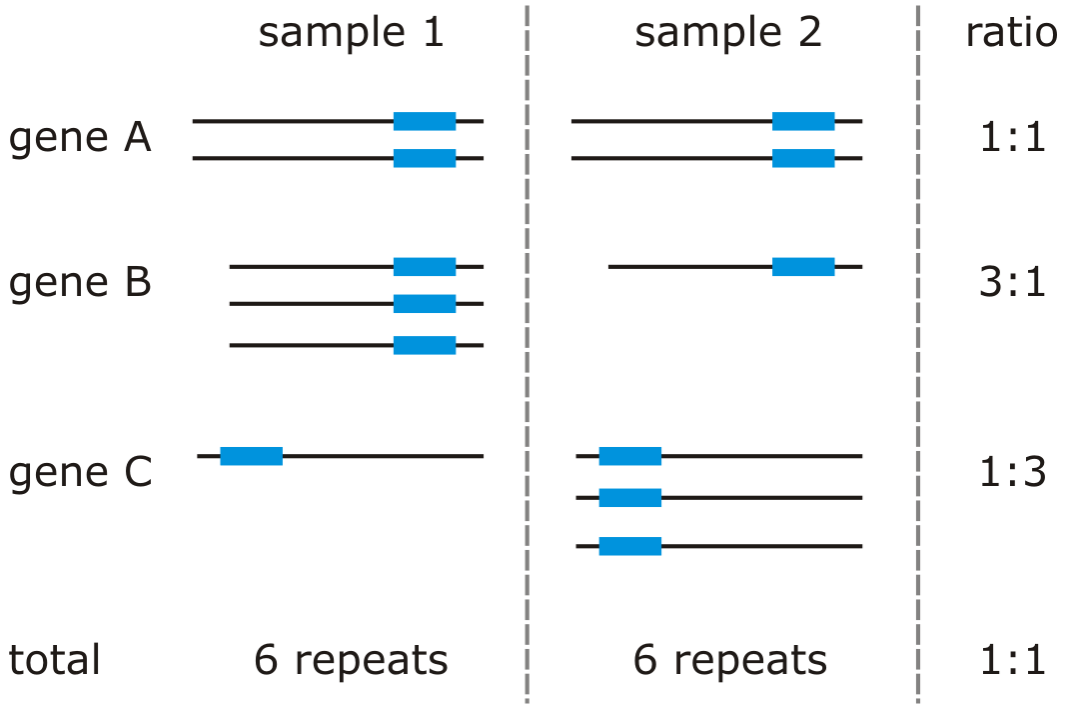
# new strategies for normalisation

- need for something new
  - reference gene validation requires (extensive) experimental work
  - sometimes not possible (lack of sample material, funding, time or devotion)

- there must be something better
  - total DNA content (Li, Bustin et al., 2005)
  - EAR normalisation (Expressed Alu Repeat)

    "using a repetitive sequence in the human transcriptome as a measure for the mRNA fraction"

# EAR normalisation - principle

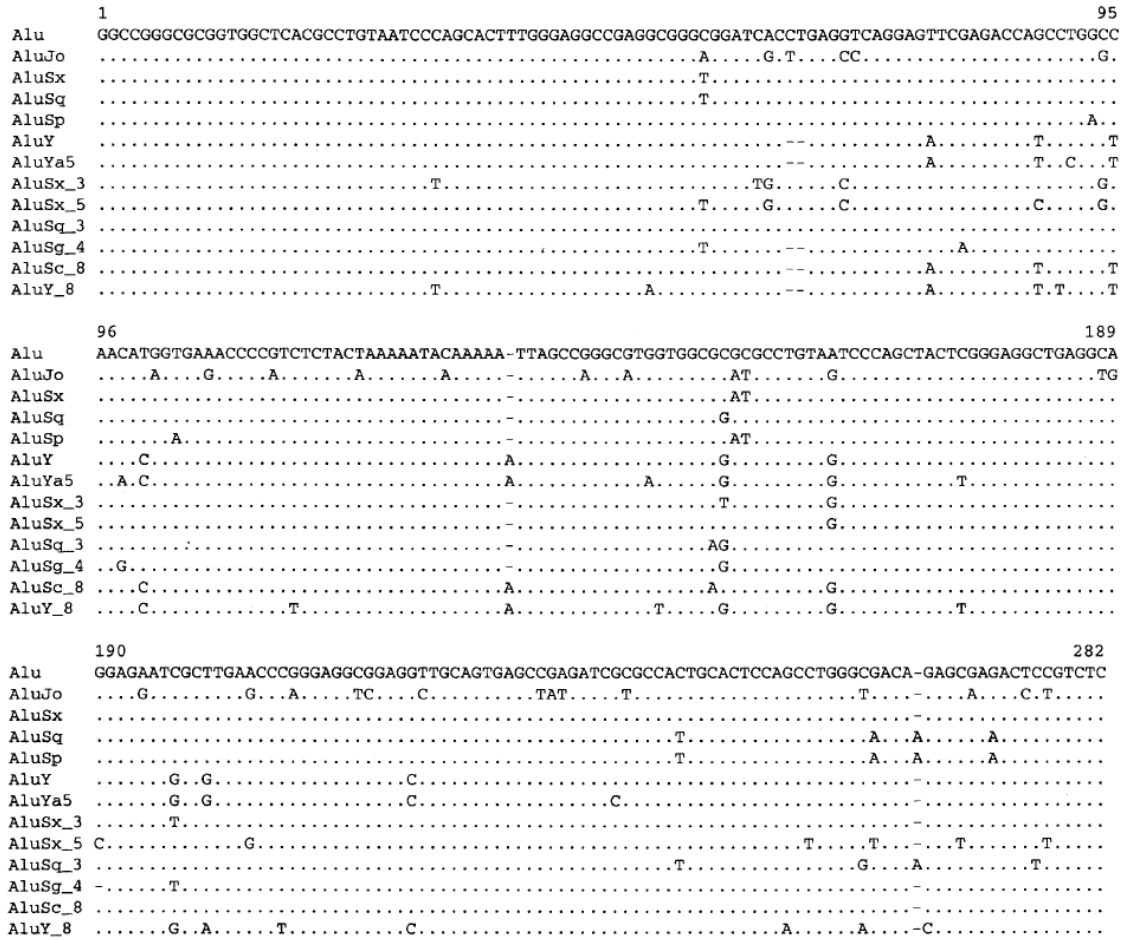



**rationale**: repeat sequences are present in the UTR of many genes, and the differential expression of a small number of genes won't influence the overall repeat abundance in the transcriptome

# Alu repeat elements

- by far the most abundant repeats in the human genome
- 1 million copies (10% of the genome), 31 subfamilies (well conserved)
- short interspersed elements (SINE) replicating via retrotransposition
- ~280 bp long, followed by a variable poly-A tail
- no known biological function
- implicated in human disease (unequal recombination)

# Alu repeat element sequence conservation




Center for Medical Genetics

# in silico transcriptome analysis

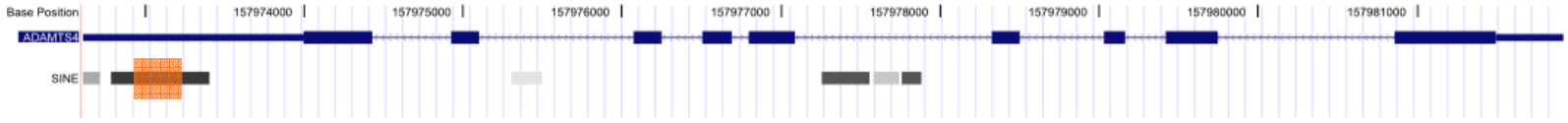
- extraction of all Alu repeat elements in the human genome
  - UCSC genome browser table function
- database with repeat element info and gene structure information for all human genes -> 'expressed Alu repeats'
  - MySQL
- Alu subfamily sequence alignment
  - PHP script 'Alu FASTA generator'
  - wEMBOSS clustalW alignment
- primer design

- roughly 1500 human genes contain one or more Alu repeats

| | |
|---|---|
| AluSx | 532 |
| AluJo | 250 |
| AluJb | 236 |
| AluSq | 178 |
| AluY | 169 |
| AluSg | 161 |
| FLAM_C | 102 |

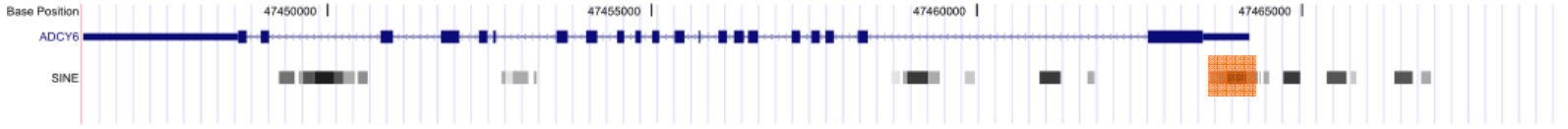
# examples Alu containing genes

*ADAMTS4* (1q23.3)




AluSq

*ADCY6* (12q13.12)




AluSg

Center for Medical Genetics

# Alu subfamily sequence alignment

>AluJo chr1:1517856-1518155 - 83858
GGCCAGGCTTGGTGGCTCATCCTGTAGAGGCGGGAGGATCTCCTGAGGCCAAGAGTTGGAGACCAGCCTGGGCATTATAGCAAGACTCGGTCTCTACAACAAATTTTTAGCATTAGCGGGCGTGGTG
GTGCACGCCTCTGGGGCTAATTAGGGCCTAGTAGGGTATGGCTACTGGACAGGCTGAGGCAGGAGGATCACTTGAGCCTGGGAGGTCGAGGCTGCAGTGAGCTATAATCACACCACTGCACTCCAGC
CTGGGCTCAGAGTAAGAACCCATCTCTAAAAGGAAAAAGAGAGAAA

>AluJo chr1:1699225-1699371 + 985
AGCCGGACACGGCGGCTGATGGCTGTAATCCCAGCACTTTAGGAGGCCGAGGCAGGAGGATCACTTGAGATAAGGAGTTCAGGACCAGCATGGGCAACACAGCGAGACCCCATCTCTATAGAAAACA
CAAAAATGAGGCTGGGGGTG

>AluJo chr1:1699225-1699371 + 9906
AGCCGGACACGGCGGCTGATGGCTGTAATCCCAGCACTTTAGGAGGCCGAGGCAGGAGGATCACTTGAGATAAGGAGTTCAGGACCAGCATGGGCAACACAGCGAGACCCCATCTCTATAGAAAACA
CAAAAATGAGGCTGGGGGTG

>AluJo chr1:6589073-6589395 + 9903
AGCTGGGCGTGGTGGTTCACGCCTATAATCCTAGCACTTTGGGAGGCTGAGACAGGAGGGTCACTTGAGGCCAGGAGTGTGAGACCAGCCTGGGTAACACAGCAAGACCCTGTCTCTATAAAAATTT
TAAGAATATCCTTTTGTATCTACATATGAGAAAAAAAATTACTGGGTGTGGTGGCACACGCCTGTAGTCTCAGCTATTCAGGAAGGTGCGGTGGGAGGACTGCGAGTCAAGGAGGTGGAGGCTGCAG
TGAGCCATGATTGTACCACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGTCTCAAAAATAAAAAA

>AluJo chr1:9121232-9121503 - 80045
GGCCCAGTGTGGTGGCTTGTACCCGTAATCCCAGCACTTTGGGAGGCCGAGGCAGGAGCATCGCTTGAGCCCAGGAGTTCAACTGGCCAACACAGTGAGGCTTTGTCTCTACTAAAAATTTAAAAAT
TAGCTGGGCATGGTGGCGCATGCCTGTAGTTCCAGCTACTTGTGAGGCTGAGGTGGGAGGATCGCTTGAGCCTGGGAGGTCGAGGCTGTAGTGAGCTATGATTGCAGACAGAGCAACACCCTCTCTC
AAAAAAAAAAGAAAAAAA

>AluJo chr1:19288492-19288789 - 23065
GGCTGGGCGTGATGGCTCATGCCTATAATCCCAGCACTTTGGGAGGCCATAGTGGAGGATCTCTTGAGCCCAGGAGTTCCAGACCAGCCTGGGCAACATAGCAAGACTCTATCTACAAAAAATAAAA
AAAAATTAGCCGGGCCTGGTGGCATGTGCCTGTGGTCCTAGCTACTCAGGAGGCTGAGGTAGGAGGATCACTTGAGCCTGGGAGGTCGAGGCTGCAGTGAGCCATGAACATGCTACTGCATTCCAGC
CTGGGCAACAGAGTGAGACCCTGGCTCAAAAACAAAAACAAAAA

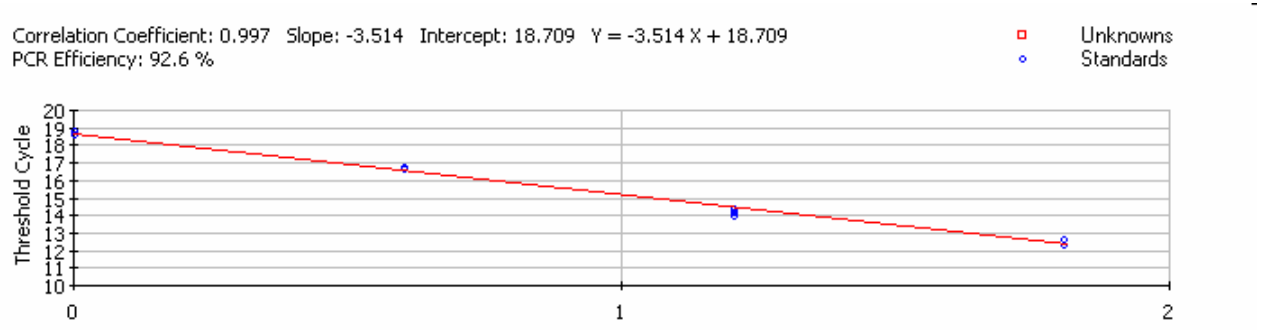
>AluJo chr1:19338130-19338402 - 246181
GTGGCTGACACCTGTAATCTCAGCACTTTGAGAGGCCAAGGCAGTAGGACTGATTGAAGACAGGAGTTCCAGACCAGTCTGGGAAACAAAGCGAGACCCTGTCTCCACTAAACATAAAAACAAAATT
ACTGGGGCCCCATGGCACACACCTGTAGTCCCAGCTGCTCGGGAAGCTGAGATGGGCGGATTGCTTGAGCCCAGGTATTCAAGTCTGGAGTGAGCTATGACTGTGCCACTGCACTCCAGCCTGGGCG
ACAGAGCAAACCCTGTTTC

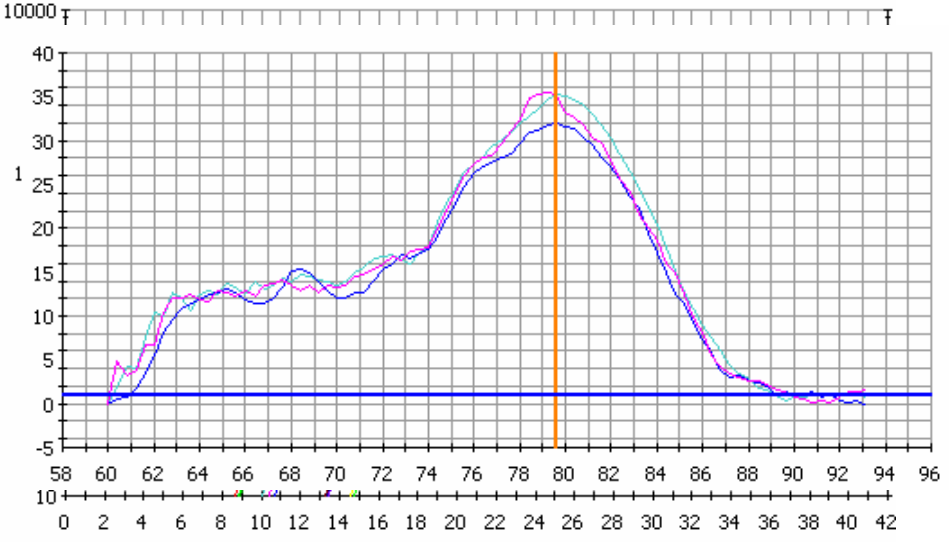
>AluJo chr1:23547487-23547778 + 55616
GGCCAGGCACAAGTGGCTCATGCTTGTAACCCTAACATTTTGGGAGGCCAAGGCAAGAGGATCACTTAAGCCCAAGAGTTTGAAACCAGCCTAGGCAACACAGCGAGACCTCATCTTTACCAAGAGA
AGAAACAAATTAGCATGGATGTGGTGGCATGTGCCTGTGGAAGCACGAGGATTGCTTAAGCCCAGGAATTCAAGGCTGTGGTGAGCTATGACTGCACCACTGCATTCCAGCCTGGGCAAAAGAAGGA
GACCCTGTCTCTAAAAAACAAAAAAATTAAAAAAAAAAA

>AluJo chr1:24543088-24543394 + 57185
AGCTGGGCACAGTGGCTCATGCCTGCAATCCCAGCACTTTGGGAGGCCAAGGCTGGAGGATCACTTGAGCCCAGGAGTTTTATATCAGCCTGGGCAACATAGCAAGACCTCATCTCTGCTAAAAATT
AAAAATAAATAAATTAGCTGGGTGTGGTGGTGCATGCCTGTGATCCTAGCTACTCAGGAGGCTGAGGTGGGAGGATCGCTAGAGCCCAGAGAGCCAAGGCTACAGTGAGCCATGATCATGCTACTGC
ACTCCAGCCTGGGTGACAATGAGACCATGGCTCAAAAAAAAAAAAAAAAAAAA

# Alu repeat assay evaluation

- AluSx assay (AluSq | AluJ)



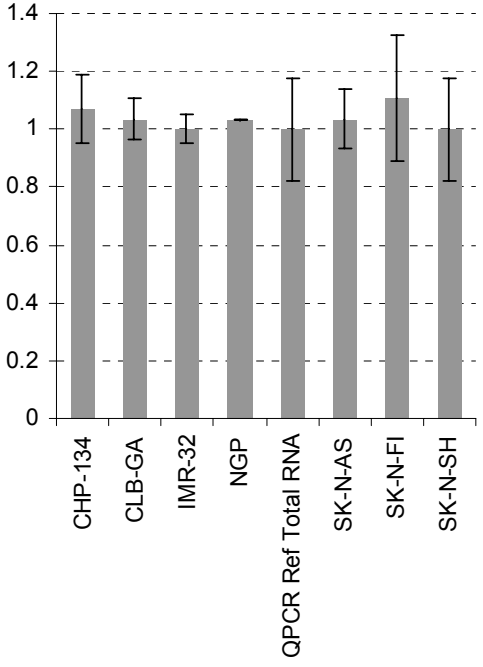




64, 16, 4 and 1 ng
QPCR Reference Total RNA
(Stratagene)

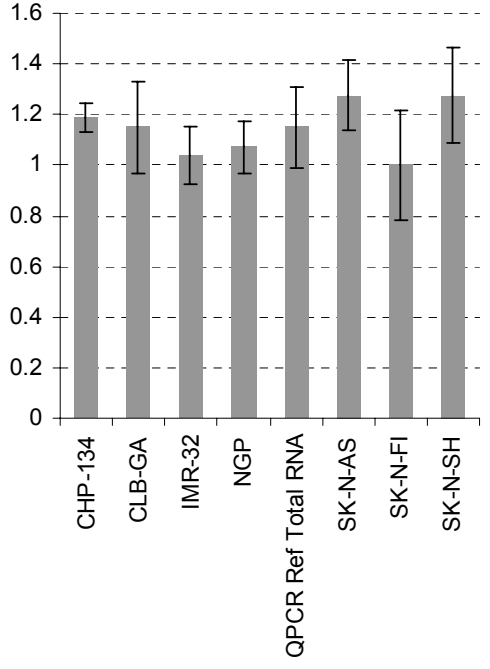
Center for Medical Genetics

# AluSq normalisation

- AluJ



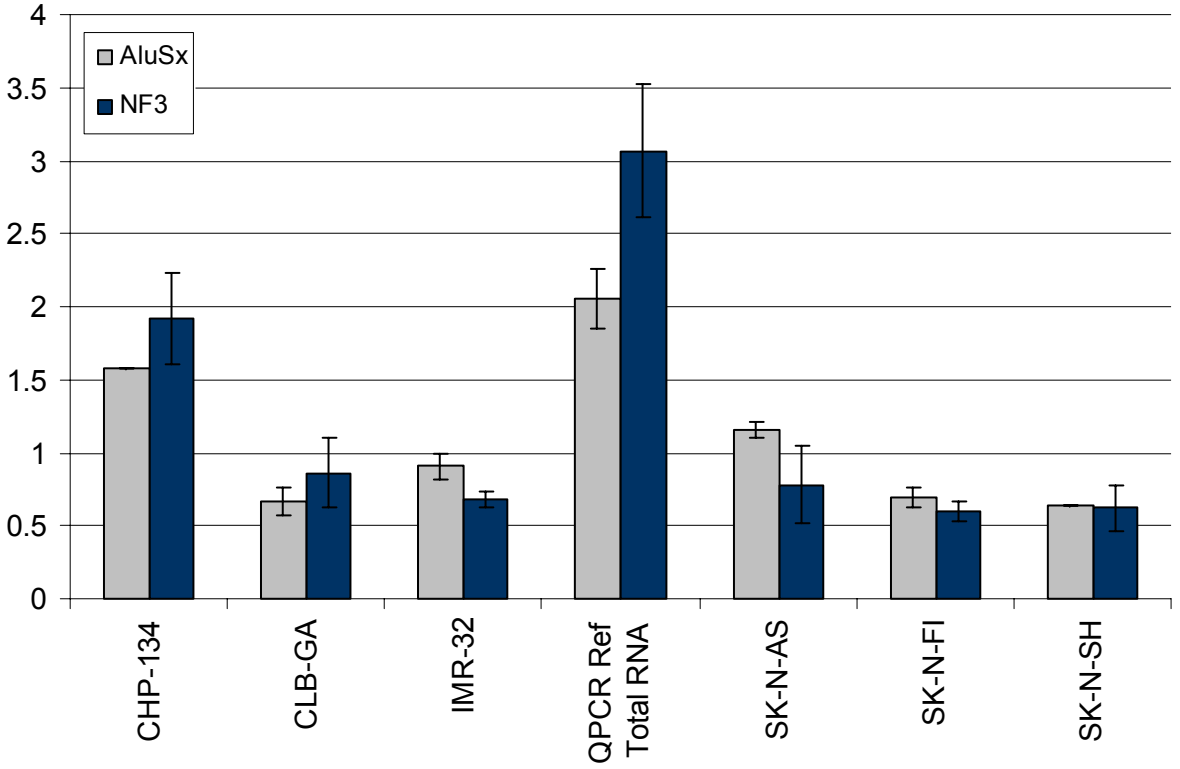

- AluSx




Center for Medical Genetics

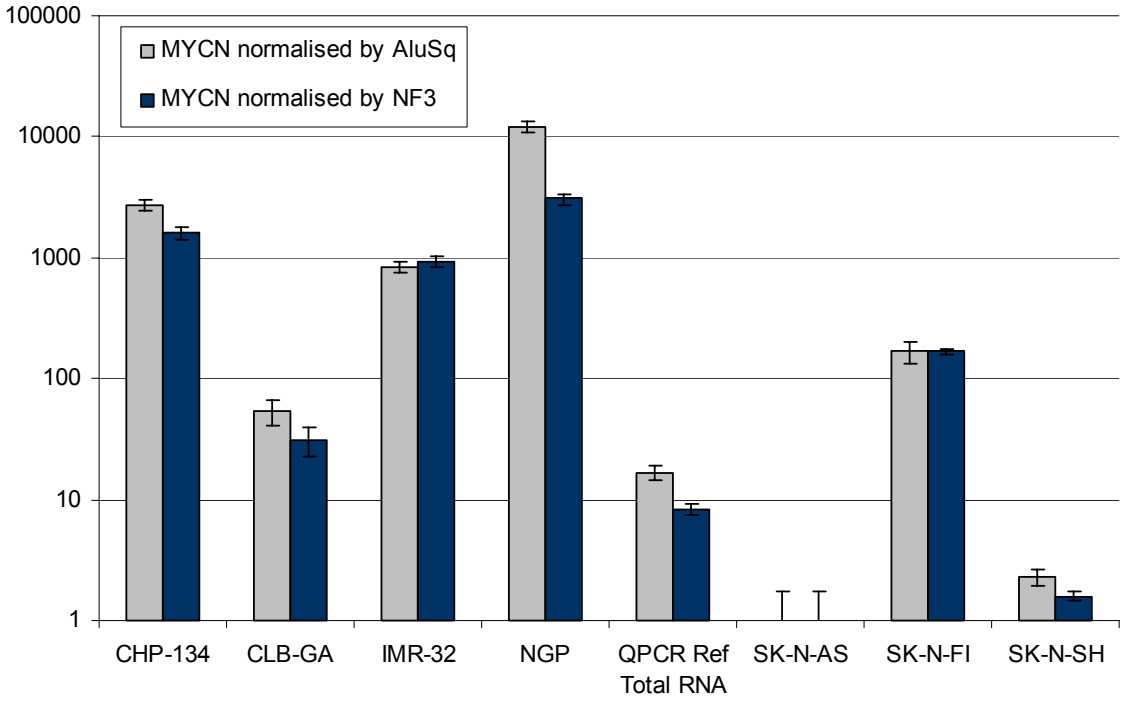
# EAR normalisation

- comparison of AluSq levels and NF based on 3 best reference genes
  Pearsons correlation 0.943 (p=0.0014)



# EAR normalisation

- MYCN expression levels normalised by AluSq or NF3

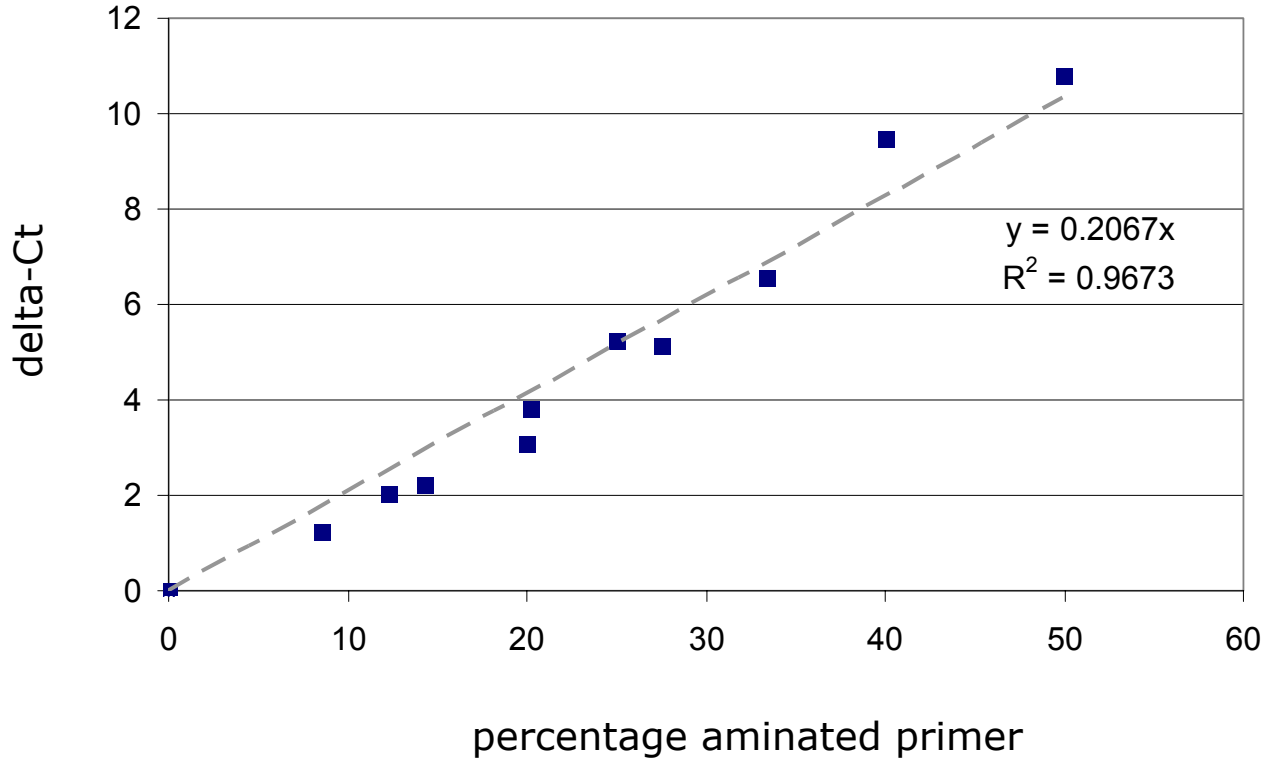


# Alu based genomic DNA assay normalization

- about 1 million Alu repeats in the human genome
  - AluSx 336,949 elements
  - AluSq 94,824 elements
  - …
- 10 ng of DNA as input: Ct value of ~8
  - use less DNA (adsorption / Poisson effects)
  - add competitive non-functional primers

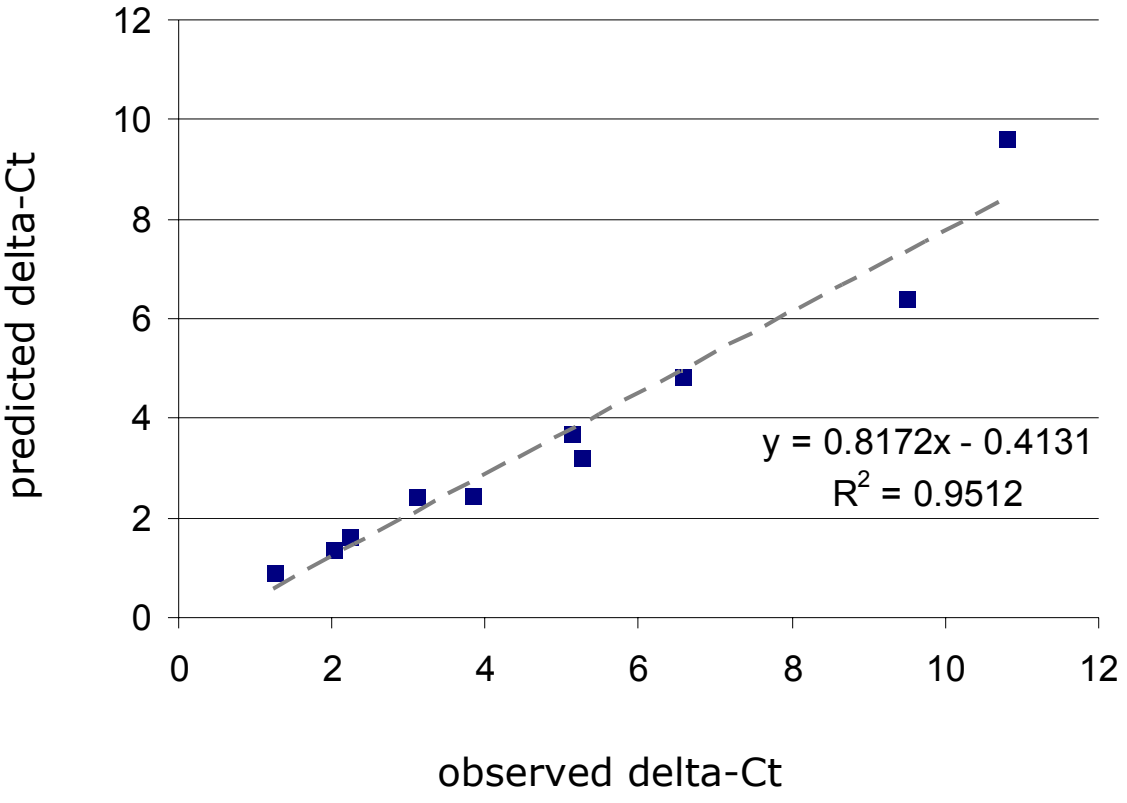
# Alu based genomic DNA assay normalization

- Alu competimer evaluation



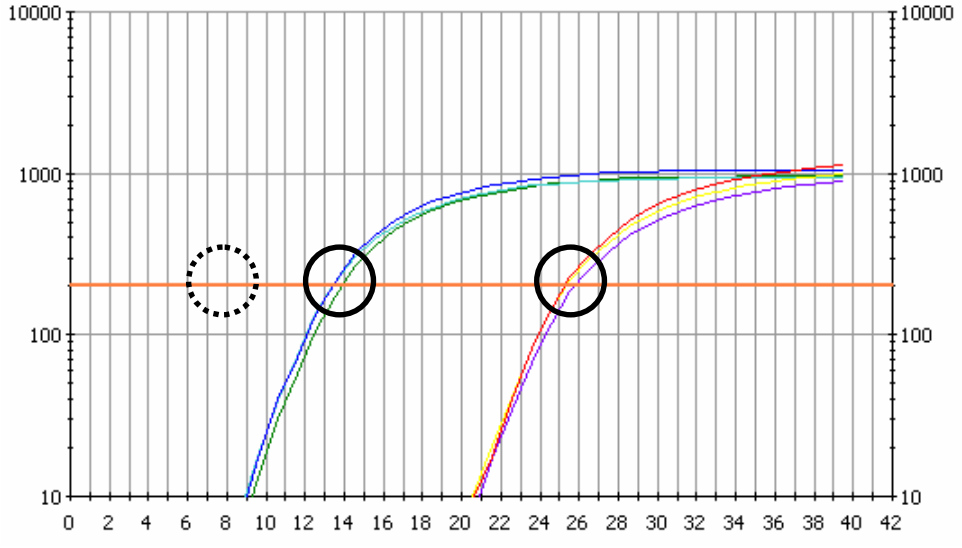
# Alu based genomic DNA assay normalization

- Alu competimer evaluation



# Alu based genomic DNA assay normalization

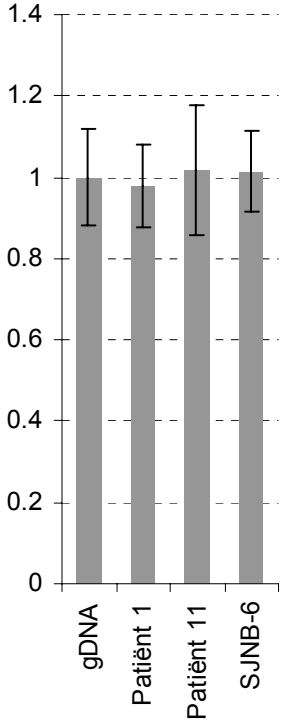
- 1 ng of DNA + 20 % competimers: 6 cycles shift to the right



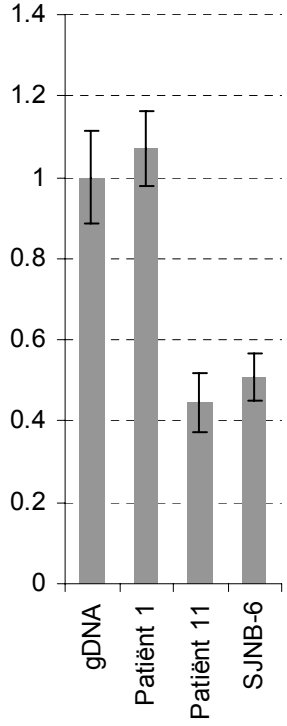

AluSx single copy gene

# Alu based genomic DNA assay normalization

- *GPR15* (unaffected gene)




- *VHL* exon 2 (deleted)



# conclusions Alu repeat normalisation

- preliminary data suggests it works for
  - gene expression normalisation (cDNA) (EAR normalisation)
  - gene copy number quantification (DNA)
- no (extensive) experimental validation required
- only limited sample amount required
- strategy could be expanded to other expressed repeats

# acknowledgments


- Katleen De Preter
- Filip Pattyn
- Jan Hellemans
- Jasmien Hoebeeck
- Els De Smet
- Nurten Yigit
- Pieter Mestdag
- Anne De Paepe
- Frank Speleman

- Rob Powell – PrimerDesign Ltd., Southampton, UK




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