



**tatabiocenter**

Design of real-time PCR experiments and  
statistical analysis of the measured data

Mikael Kubista

**NEWS**

2007-09-20

Fisher Scientific in collaboration with TATAA Biocenter arranges a two-day theoretical course in PCR and real time quantitative PCR (qPCR).

[Read more...](#)
[Visit External Link](#)

2007-09-18

Free Webinar in US on October 8

[Read more...](#)
[Visit External Link](#)

2007-09-18

Free Webinar in Asia on October 5

[Read more...](#)
[Visit External Link](#)
[View all news](#)
**TESTIMONIALS**

"The quality of the course is very good. They are easy-talking persons, and easy to follow. The course was very important for my work. I



## TATAA BIOCENTER

TATAA Biocenter is a commercial research provider that offers commissioned research and training within molecular diagnostics and gene expression analysis using real-time PCR and other molecular techniques to quantify nucleic acids.

Our competence is based on knowledge and experience accumulated through years of research at leading European Universities. Our services comprise the entire field of real-time PCR services including commissioned research, hands-on training, and custom design of real-time PCR assays.

### AMONG OUR ACTIVITIES ARE

- Open and tailor-made hands-on training courses
- Commissioned research and development
- Custom design of QPCR assays and assay validation
- Product development


**PRODUCTS**

**SERVICES**

**COURSES**

**RESEARCH**


APPLY TO OUR **COURSES**  
[CLICK HERE](#)

### UPCOMING COURSES

24:th-28:th of September 2007

Freising - Germany  
[Register](#)

1:st-3:rd of October 2007  
 Comet Assay course - Oslo - Norway

[Register](#)

8:th-12:th of October 2007

Göteborg - Sweden  
[Register](#)

**View all courses.**

Tataa courses are supported by



# Requirements

- Experimental design should:
  - be defined before the experiment starts
  - address a hypothesis
  - be as simple as possible
  - contain minimum number of factors that are not under control
  - be technically and economically feasible
  - result in a practicable statistical test

# Variance in data

The total variance of the data has three contributions:

## Confounding variance

### Processing variance

**Due to sample handling:**

- Sampling
- RT
- PCR

**To decrease:**

- Use replicates
- Normalize internal standard

### Individual biological variance

**Each subject is different.**

- Different baseline expression
- Different response to treatment

**To decrease:**

- Use repeated measurements
- Normalize to control group

## Studied variance

### Treatment variance

The difference between groups.

**To increase:**

- Sample randomly
- Take large sample



## qPCR experiment

- Good experimental design should minimize the confounding variance and maximize the treatment variance
- Consider whether the experiment can be based on repeated measurements.
  - If control samples and samples after treatment can be taken from the same subject confounding biological variance is reduced
- Most experiments can be designed as a comparison of two group.
  - Even if more groups are studied, analyses typically ends up by using post-tests for pair-wise comparisons after initial analysis by eg. ANOVA.

# Hypothesis

- Hypothesis is the suggested explanation for some phenomenon
- Is the explanation correct or not?
- Set the probability level *alpha*
- One-tailed or two-tailed?
  
- Examples:
  1. The expression of *alpha synuclein* in Cerebellum is different between healthy and Parkinson patients.
  2. The expression of *alpha synuclein* in Cerebellum is lower in healthy patients than in Parkinson patients.
  3. The expression of *alpha synuclein* differs between regions of brain.
  4. The expression of *alpha synuclein* increases with age.

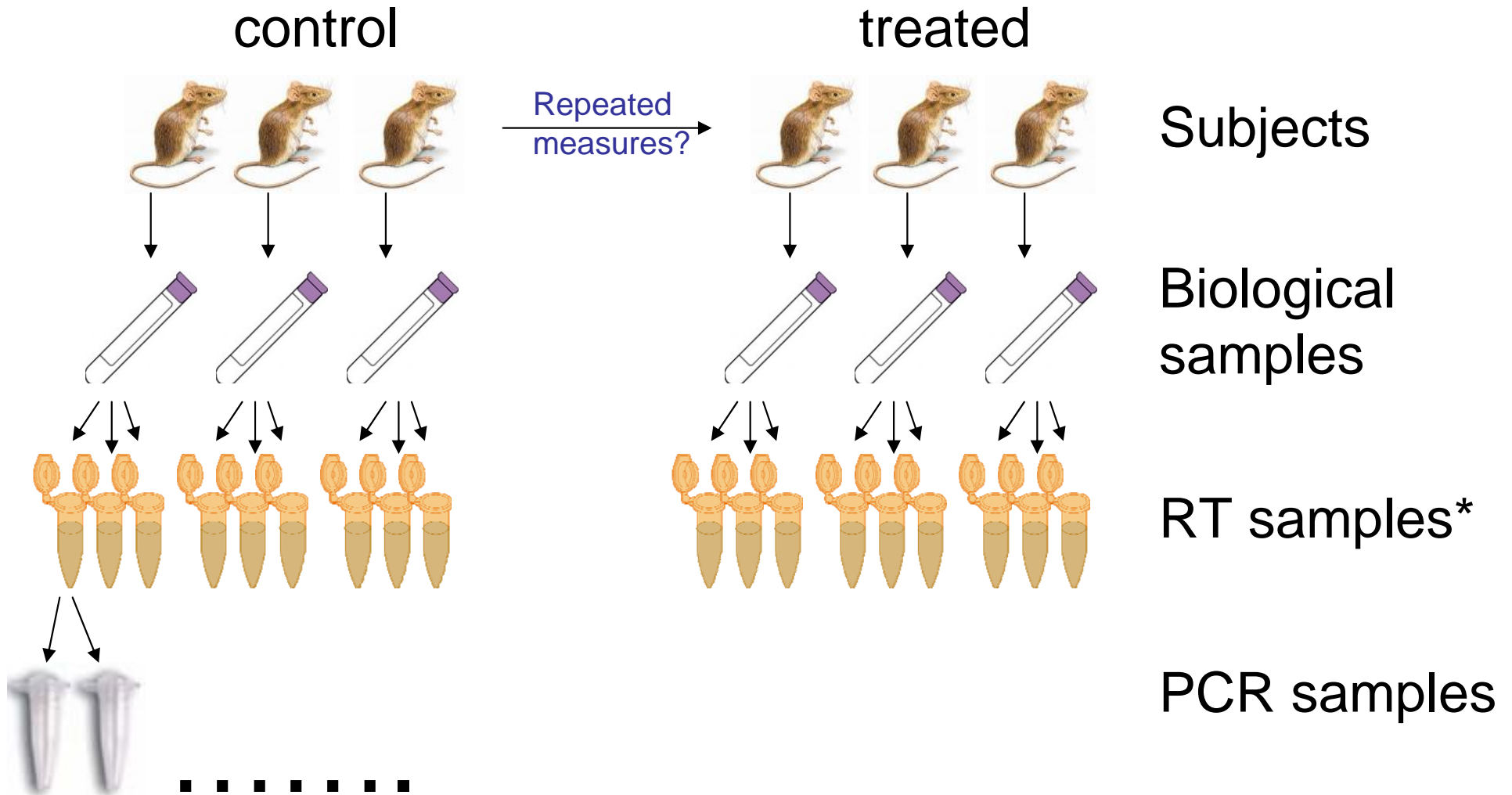
## Define variables

- What variables will be explained?
  - Select the candidate gene before the experiment
- What variables will be explanatory?
  - Nominal variables: Treatment type, Method of extraction ....
  - Metric variables: Dose, Age, Time point ....
- Examples:
  - The explained variable is the expression of alpha synuclein transcript normalized to the expression of GAPDH and the explanatory variable is the disease status (healthy / ill)
  - The explained variable is the expression of alpha synuclein transcript normalized to the expression of GAPDH and the explanatory variable is the brain region (Cerebellum, Medula oblongata....)
  - The explained variable is the expression of alpha synuclein transcript normalized to the expression of GAPDH and the explanatory variable is the age

# Design experiment

- Define population
- Define statistical sample (a subset of the population that will be investigated)
  - The sample should represent the population
- Classify subjects in the sample (e.g. healthy vs. ill patients, time points of sampling)
  - Subjects should be selected or allocated randomly.
- Consider the source of confounding variance and design appropriate replicated observations for:
  - Subjects
  - Treatment
  - Sampling
  - RT
  - PCR
- Consider pairing of subjects and repeated measures to decrease confounding variance
- Always minimize uncontrolled effects = minimize confounding variability

# Experimental Design



\*RT is the main source of variance. Therefore three replicates are built for each biological sample to assure more precise mean estimate.

## Data pre-treatment

1. Perform quality control of the data (outlier detection)
2. Correct for efficiency variations
3. Compensate for variations between runs (inter-plate calibration)
4. Normalize to the same amount of sample
5. Average QPCR technical repeats
6. Normalize with reference genes
7. Average RT technical repeats
8. Calculate relative quantities
9. Calculate fold-changes

Some operations are optional.

Most can be performed in reverse order, but not all

## Efficiency correction

E depends on both the gene/assay and the degree of inhibition in the sample.

For a series of similar samples one can assume that the PCR efficiency only depends on the assay and test this assumption using, for example, KOD.

$$CT_{E=100\%} = CT_E \frac{\log(1+E)}{\log(2)}$$

E is estimated from an appropriate standard curve, response curve, in situ calibration, or assumed to be the same as in previous studies.

## Interplate calibration

- If data were run in multiple plates, run-to-run variation should be corrected for using interplate calibrators.
  - Interplate calibrators: standard samples present in all plates that are analyzed for all genes (MG + RG).

$$CT_{plate\ norm} = CT_i - CT_i^{IC} + \frac{1}{no.\ runs} \sum_{i=1}^{no.\ runs} CT_i^{IC}$$

## Normalization to the amount of sample material

- Normalize samples to the same amount of material (volume, number of cells, total RNA etc)

$$CT_{conc=1} = CT_{conc} + \log_2(conc)$$

## Normalize QPCR repeats

- If QPCR repeats were performed they should be averaged.

$$CT_{QPCR\_average} = \frac{1}{no. \ repeats} \sum_{i=1}^{no. \ repeats} CT_i^{repeat}$$

## Normalization to reference genes

$$CT_{MG, norm} = CT_{MG} - \frac{1}{no. RG} \sum_{i=1}^{no. RG} CT_i^{RG}$$

- Use only validated reference genes

<http://www.tataa.com/Products/Human-Endogenous-Control-Panel.html>

## Normalize RT repeats

- If RT repeats were performed they should be averaged.

$$CT_{QPCR\_average} = \frac{1}{no. \ repeats} \sum_{i=1}^{no. \ repeats} CT_i^{repeat}$$

## Relative quantities

- Calculate the relative expression in the different samples

$$RQ_{max} = 2^{CT_{min} - CT}$$

$$RQ_{min} = 2^{CT_{max} - CT}$$

$$RQ_{mean} = 2^{CT_{mean} - CT}$$

$$RQ_{sample} = 2^{CT_{sample} - CT}$$

## Fold changes

- Most statistical and multivariate methods assume Normal (Gaussian) distribution. Gene expression is usually Normal distributed in log scale.
- Traditionally log base 2 ( $\log_2$ ) is used

$$FC = \log_2(RQ)$$

Linear	$\log_2$
0.125	-3
0.25	-2
0.5	-1
1	0
2	1
4	2
8	3

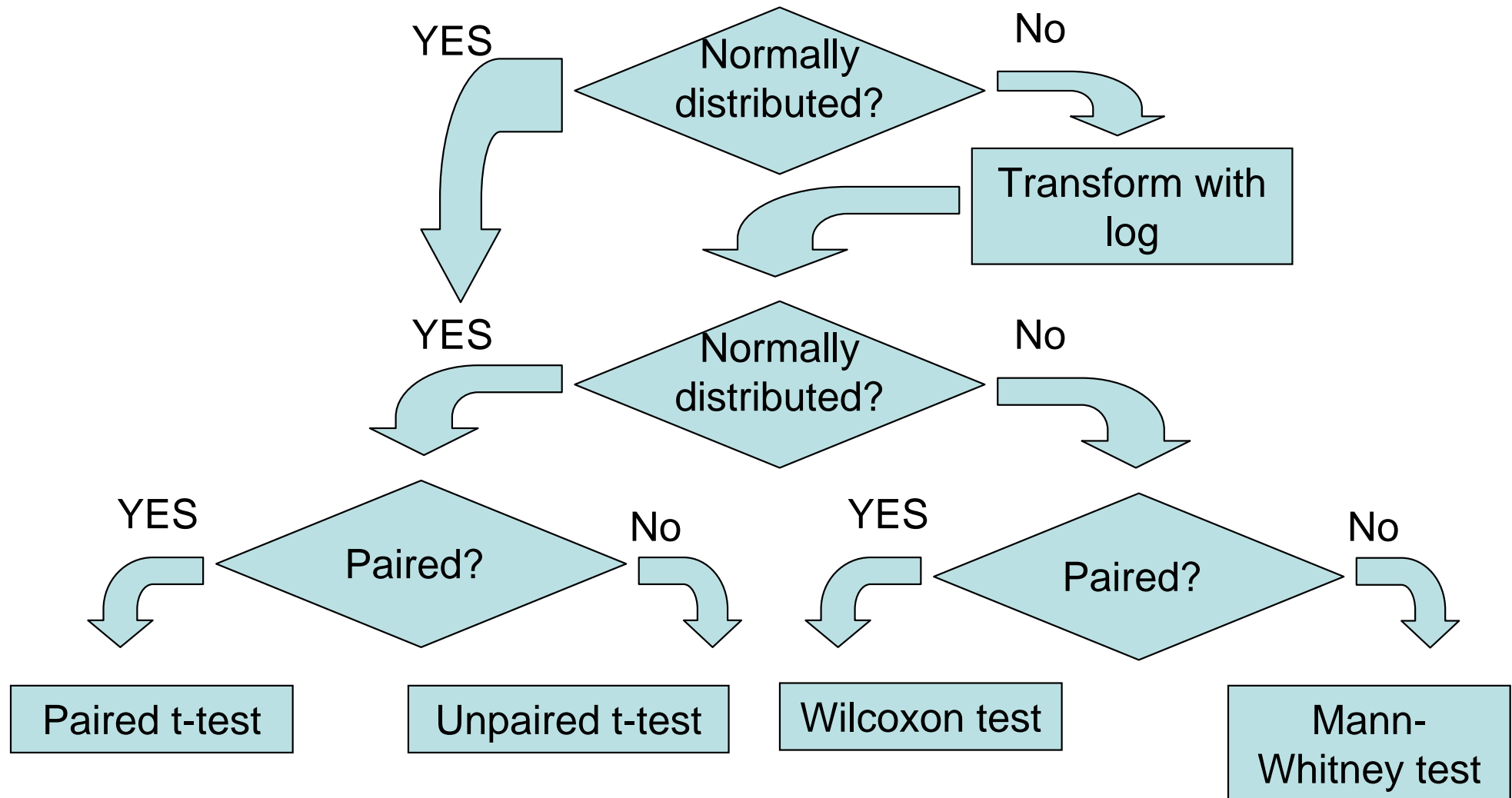
## Perform test

- Consider design
- Consider distribution
- Consider data size
- Select appropriate test
- Set the confidence level (consider multiple comparison)
- Make sure that you have the right software available!

## Parametric vs. nonparametric test

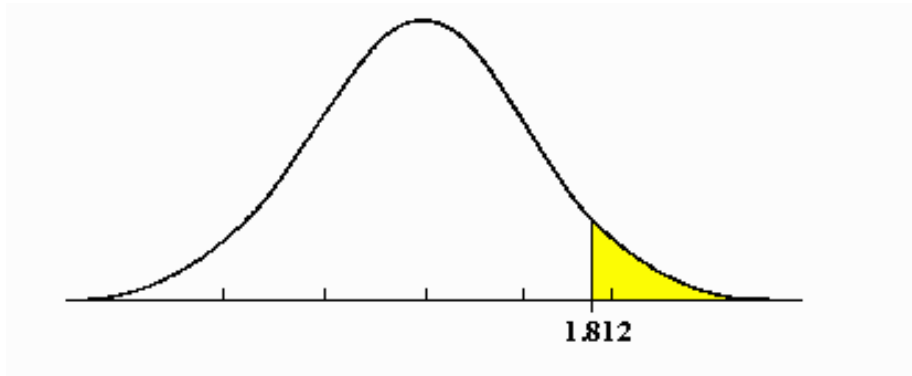
- Use **parametric** tests whenever possible!
  - Even small groups can be compared (e.g.  $N=5$ )
  - Make sure that data are approximately normally distributed
  - If not try log transformation
- Use **non-parametric** test if normal distribution is not obtained even after log-transformation
  - Make sure that you have enough data ( $N>20$ )
  - The power is weak

# Comparison of 2 groups



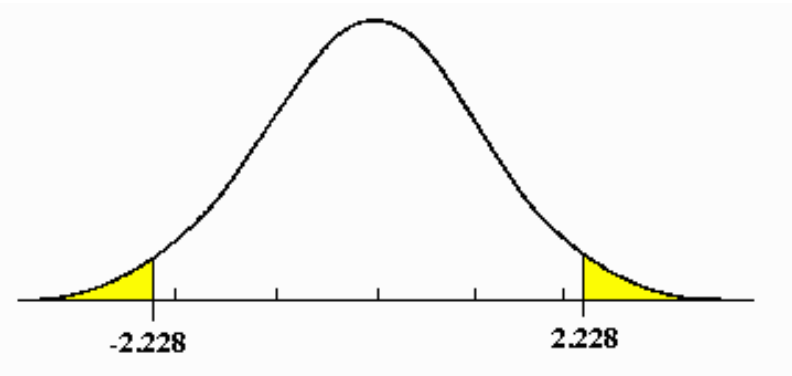
# One sided vs. two sided test

Example of t-distribution for  $df=10$ .



Is there a significant increase?

A one-tailed t-test locates the critical probability into one tail only. The critical range is wider.



Is there a significant difference?

A two-tailed t-test divides the critical probability region in half, placing half in the each tail.

# Relative quantification

Data editor 1

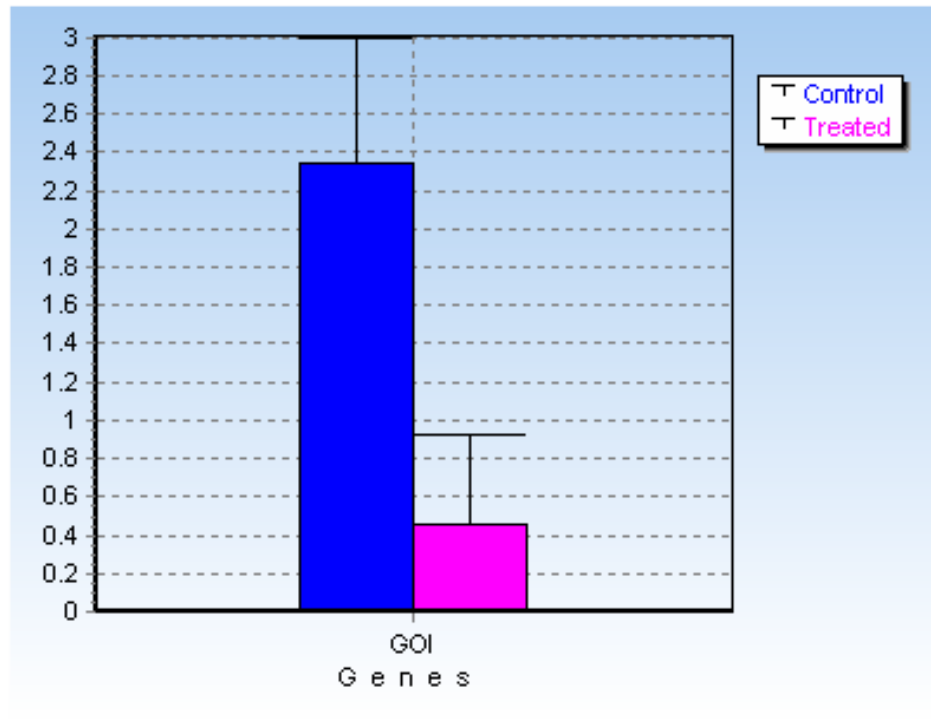
File Edit Grid Pre-processing

	A	B	C	D	E	F	
1		MG	RG	#QPCR	#RT	#Treatment	#RNA
2	S1	24.39	21.87	1	1	1	125
3	S2	24.15	21.87	1	1	1	125
4	S3	24.58	21.86	2	1	1	125
5	S4	24.52	21.72	2	1	1	125
6	S5	24.64	21.97	3	1	1	125
7	S6	24.62	21.9	3	1	1	125
8	S7	24.56	22.7	4	2	1	125
9	S8	24.72	22.69	4	2	1	125
10	S9	24.42	22.53	5	2	1	125
11	S10		22.59	5	2	1	125
12	S11	24.45	22.72	6	2	1	125
13	S12	24.72	22.81	6	2	1	125
14	S13	24.86	21.99	7	3	1	125
15	S14	24.93	22.02	7	3	1	125
16	S15	25.1	21.83	8	3	1	125
17	S16	25.2		8	3	1	125
18	S17	25.53	21.96	9	3	1	125
19	S18	25.64	21.85	9	3	1	125
20	S19	25.86	22.04	10	4	1	125
21	S20	25.88	21.94	10	4	1	125
22	IC 1-36	19.93	21.07	37	13	0	200
23	IC 37-72	19.37	20.79	38	14	0	200

Source: C:\Program Files\Multid\Genex\Data\examples\RT\_QPCR\_Treatment\_repeats.mdf

# Statistical comparison

## Mean and 95% CI



## Unpaired 2-sided t-test

A screenshot of an 'Unpaired t-test' software window. The window displays a table of statistical results for two groups: GOI (Control) and GOI (Treated). The results are summarized in the following table:

Statistic	GOI (Control)	GOI (Treated)
KS	0.119631706142443	0.190967750327373
Norm. Dist.	TRUE	TRUE
KS P-Value	>0.1	>0.1
Count	6	6
Mean	2.34741624543726	0.45502605476186
STDEV	0.619226045757627	0.445475777156404
df		10
SD^2		0.290944781888864
t		6.07687527664882
P (2-tail)		0.000119316

The software window also shows the file path: GIT\_univariate, Set1, [GIT\_processed.mdf], Groups(Control, Treated).

(E(RG) = 0.95, E(MG) = 0.90)

# Acknowledgement

- My colleagues at the Department of Gene Expression at the Institute of Molecular Genetics division for Biotechnology of the Czech Academy of Sciences
  - Radek Sindelka
  - David Svec
  - Vlasta Ctrnacta
- 
- The examples used and free trial version of the GenEx software is available on: [www.multid.se](http://www.multid.se)

Contact info: [mikael.kubista@tataa.com](mailto:mikael.kubista@tataa.com)