# Real-time RT–PCR

for identification of differentially expressed genes.

(with Schizophrenia application)

Rolf Sundberg, Stockholm Univ.

Göteborg, May 2006

# “Real-time PCR for mRNA quantitation”

## Review paper (Wong & Medrano, 2005) ≈ citations:

Real-time PCR and real-time RT–PCR has dramatically changed the field of measuring gene expression.

It is a class of techniques that

- has a large dynamic range,
- boasts tremendous sensitivity,
- requires much less RNA template than other methods,
- can be highly sequence-specific,
- has little or no post-amplification,
- is amenable to increasing sample throughput.
- But therefore requires sound experimental design and
- in-depth understanding of normalization techniques

# Steps in real-time PCR

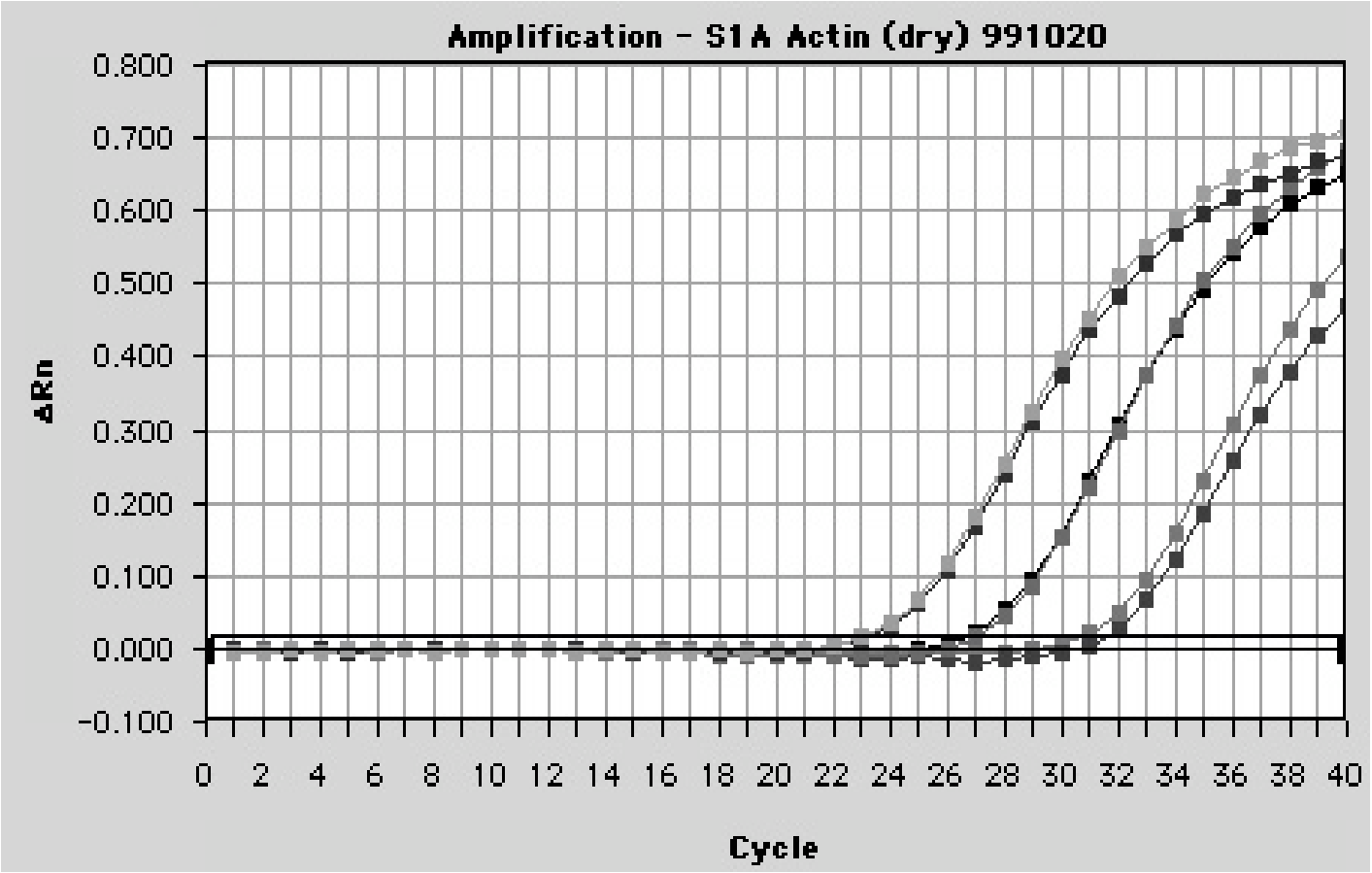
- RNA isolation and characterization
- cDNA synthesis
- Real-time PCR data acquisition (during the process)
  - Incl. adjustment to baseline, setting cycle threshold
- Generation of normalization factors
  - (using house-keeping genes)
- Normalized data
- Data analysis

# “Real-time”,
## with determination of “cycle threshold” $C_t$

$y_0 e^{bC_t} = a$

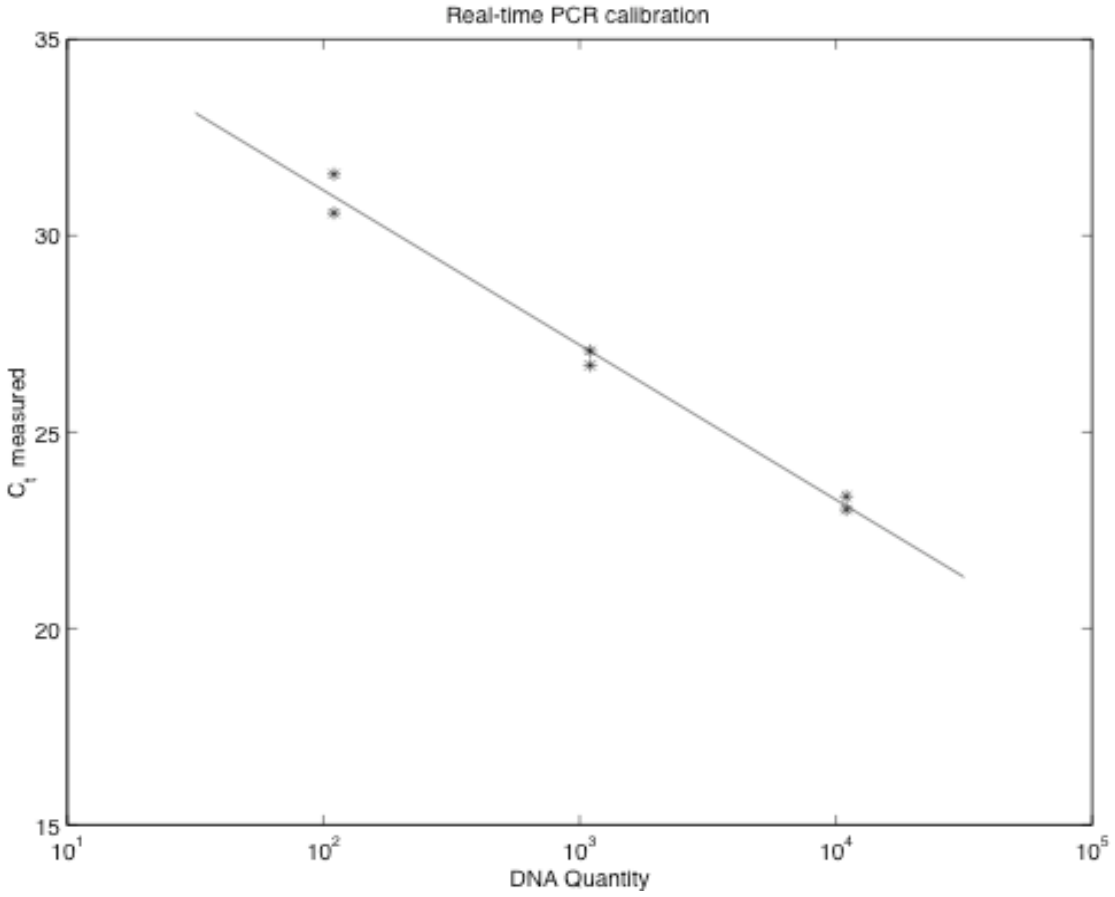
$\Rightarrow$

$\log y_0 = \log a - bC_t$



# Crude calibration by standard curve for genomic DNA

$C_t$ versus
$\log y_0$



# Schizophrenia study

Castensson et al. *Biological Psychiatry* 2003

Sundberg et al. *Biostatistics* 2006

- Patients and controls (55 of each, = 110)
- Several (brain) samples per individual (2)
- Put on plates with < 96 wells per plate
- Fluorescence measurements of mRNA by Real-time RT–PCR combined with TaqMan assay: One *master-plate* => many *replica plates*, one per gene

# Statistical aspects

- Design: Balanced incomplete design on plates
- Basic model: MRANCOVA, i.e.
  multivariate nested random effects
  analysis of covariance model (see below)
- Inference:
  (1) Reference genes for increased precision
  (2) Prediction aspects
  (3) Minor problems: plate effect estimation
  left-censoring for low-expressing genes, outliers,
  non-constant variances, multiple testing, etc.

# Modelling

Basic ≈ MRANCOVA model, for *controls*:

$Y$ = log(*fluoresc.*) vector  (gene <=> comp. $y$)

$$y_{hij} = \mu + \alpha_h + \beta' u_{hi} + \gamma_{k(hij)} + \delta_{hi} + \varepsilon_{hij}$$

h = stratum index (brain bank, sex),

i = individuals within stratum h,

j = samples within individual,

k = plate number allocation,

$u$ = individual covariate (age, time post mortem)

Nested variance components from δ and ε

# Testing and further inference

- Test $H_0$: Absence of disease effect
- Under significance, estimation or prediction?

  Explore effect distribution (interactions? affected subgroup? effects correlated btw genes?)

# Multivariate aspects

- Nested components δ and ε are multivariate, i.e. represented by covariance matrices, dimension = #genes
- Correlations btw components (genes) were high in ε, and even higher in δ.
- Motivates use of unaffected *reference gene(s)*, for statistical efficiency. ("house-keeping" gene)
- Predict candidate gene values from ref-genes, adjusting for other covariates

# For candidate genes

- With $x$ like $y$, but for ref-gene, fit $E(y|x)$,

  $y_{\mathrm{hij}}$ = as before + $\theta\, x_{\mathrm{hij}}$ ,

  or correspondingly for averages $y_{\mathrm{hi.}}$
  Note: parameters have new interpretations, and some are no longer needed in model

# Prediction aspects

- Alternative interpretation of $E(y|x)$:
  Predict candidate gene values from ref–genes, for each individual, adjusting for other factors.
- Predict patient values via model fitted to the unaffected controls, to explore non-constant disease effects
  Varying disease effect => loss of power in standard two-sample tests

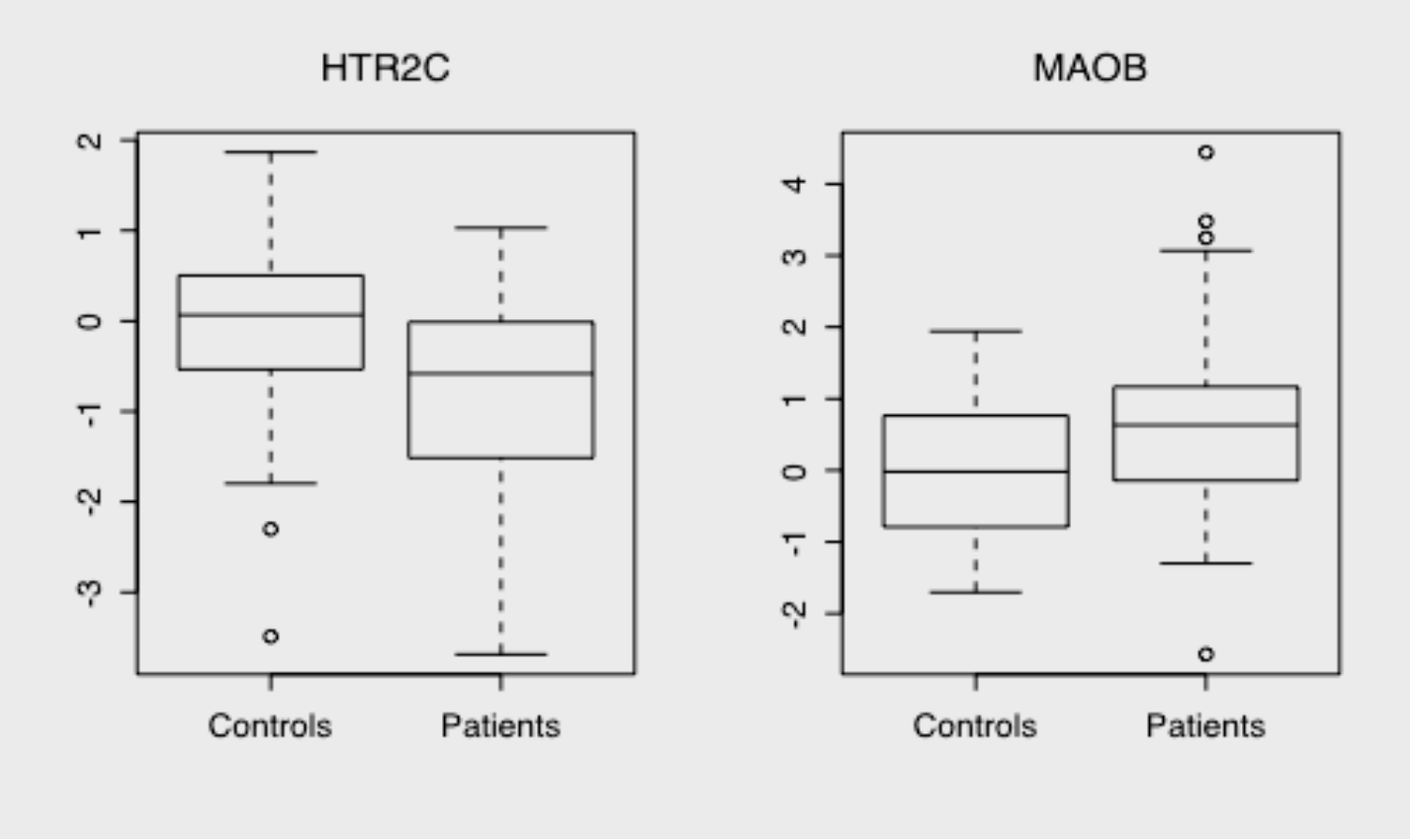
# Plate effects and averaging

- Incomplete design motivates plate effect estimation within individuals, for statistical efficiency
- But regression on *x* ‘within individuals’ will be different from regression on *x* ‘between individuals’
- => sacrifice ‘within’ plate effect estimates, and average over samples from individual

# Results

- Gain from use of reference genes:
  Std error typically reduced by factor 2 – 3,
  crucial for obtaining significant effects.
- 2 out of 16 genes were found significant,
  see box-plots etc
- Their individual effects were correlated,
  see scatter-plot

# Standardized residuals/Prediction errors for controls and patients, two significant genes: HTR2C & MAOB



# Observed against predicted f. two genes

Controls: **o**

Patients: *

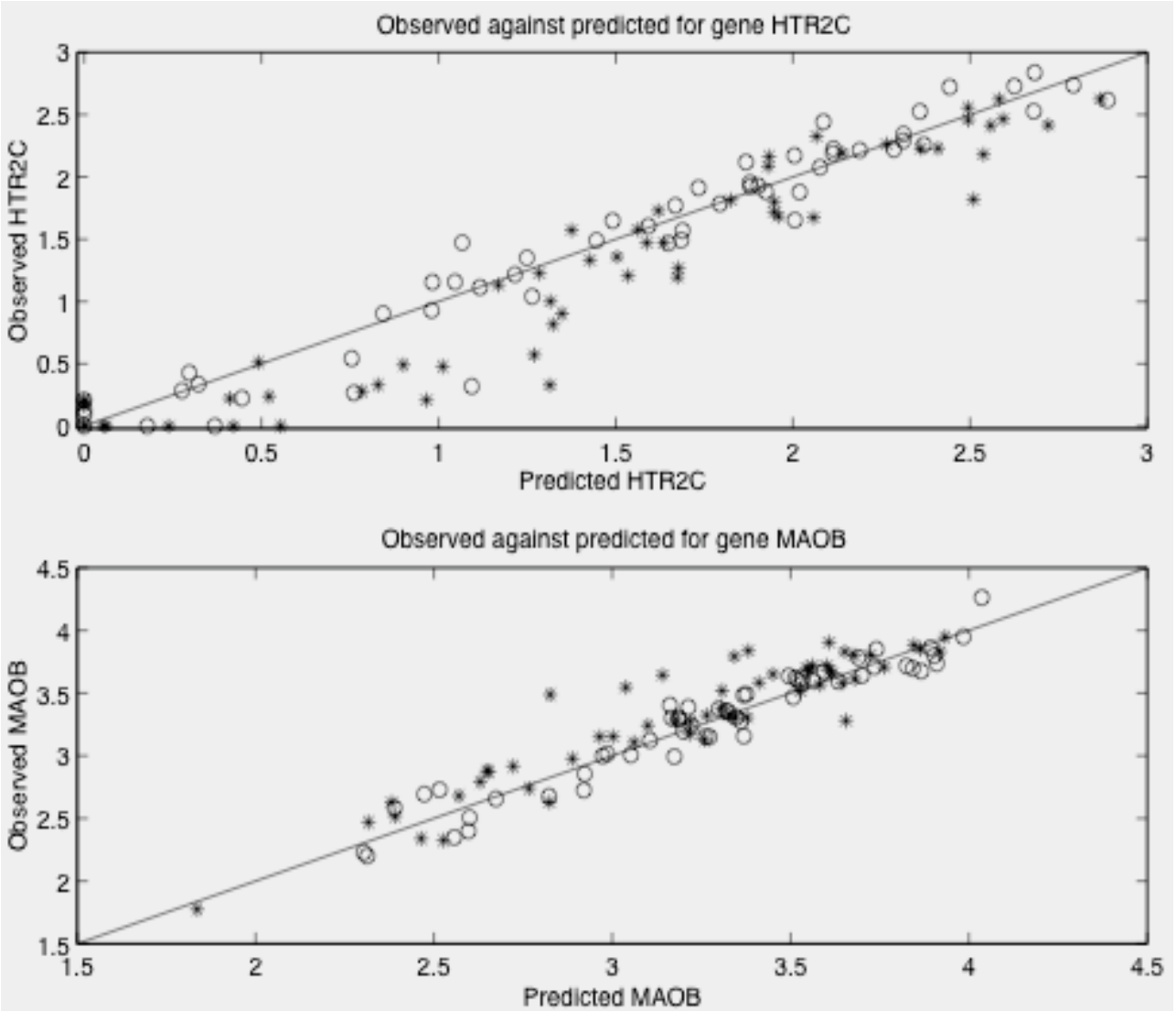
55 of each

Upper:

Gene HTR2C

Lower:

Gene MAOB



# Standardized residuals/predict. errors jointly for the two genes

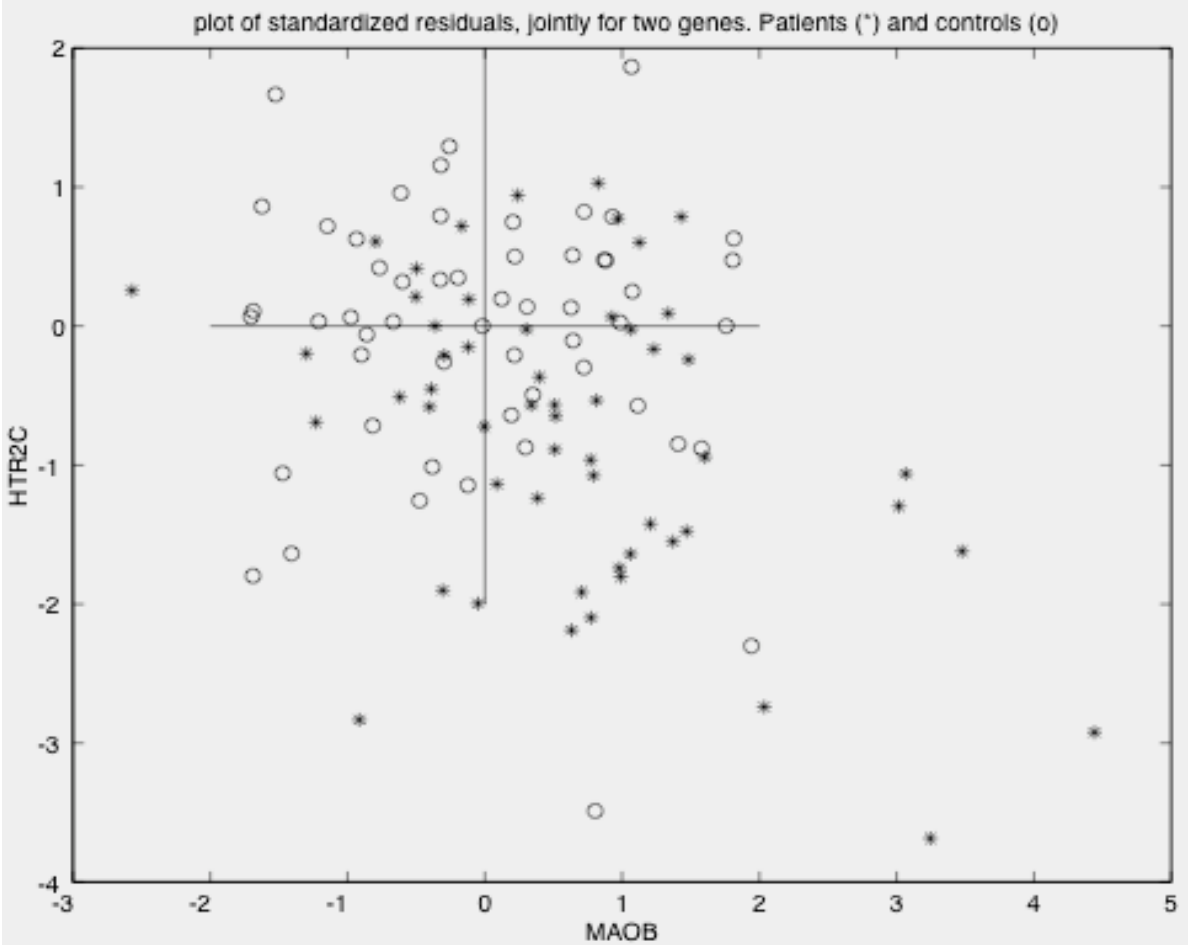
Controls: **o**

$r = -0.02 \approx 0$

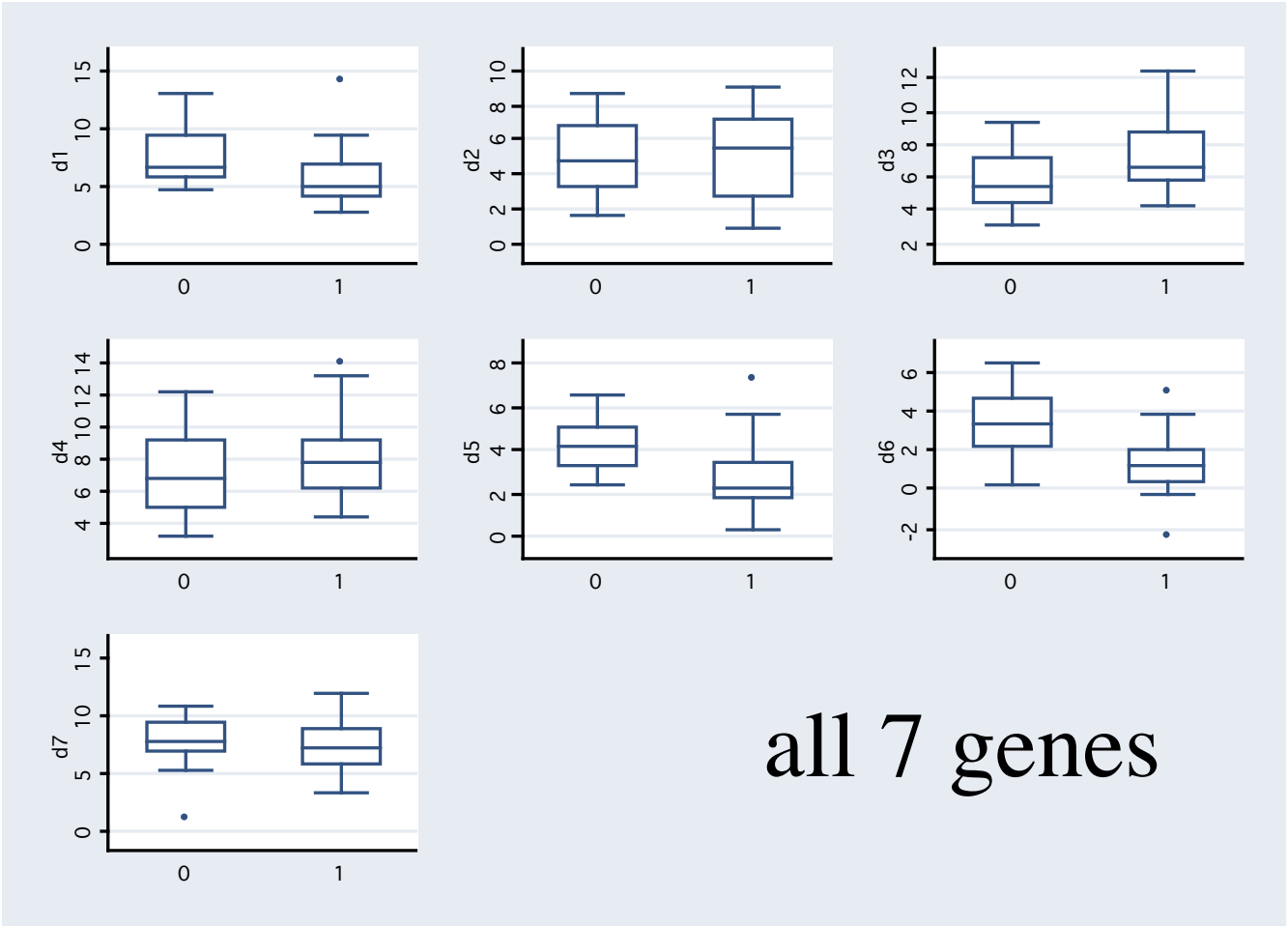
Patients: *

$r = -0.45$

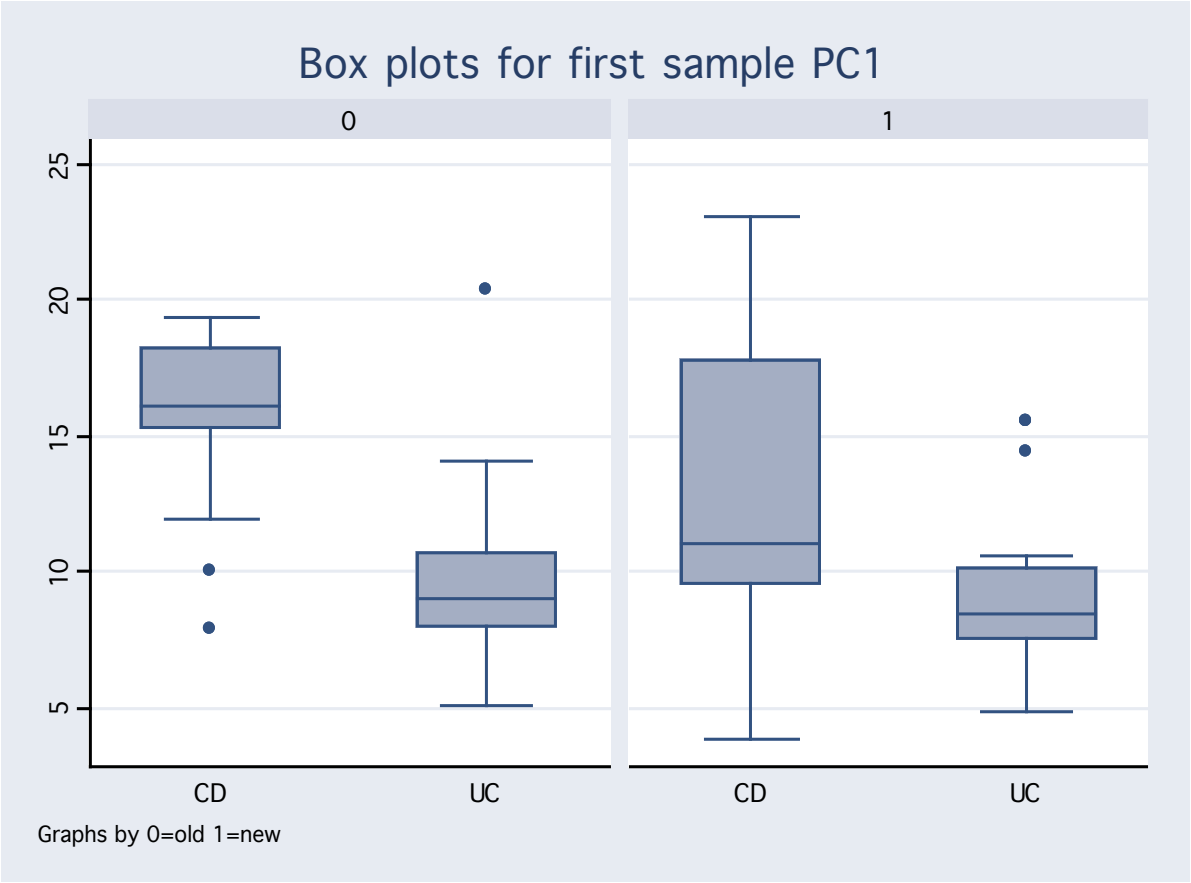
=> *differential coregulation*



# Another project: 7 genes for discrimination. Box plots for old sample and new sample



# Box plots for old sample and new sample, and the two diagnoses CD and UC Variable: PC1 from old sample

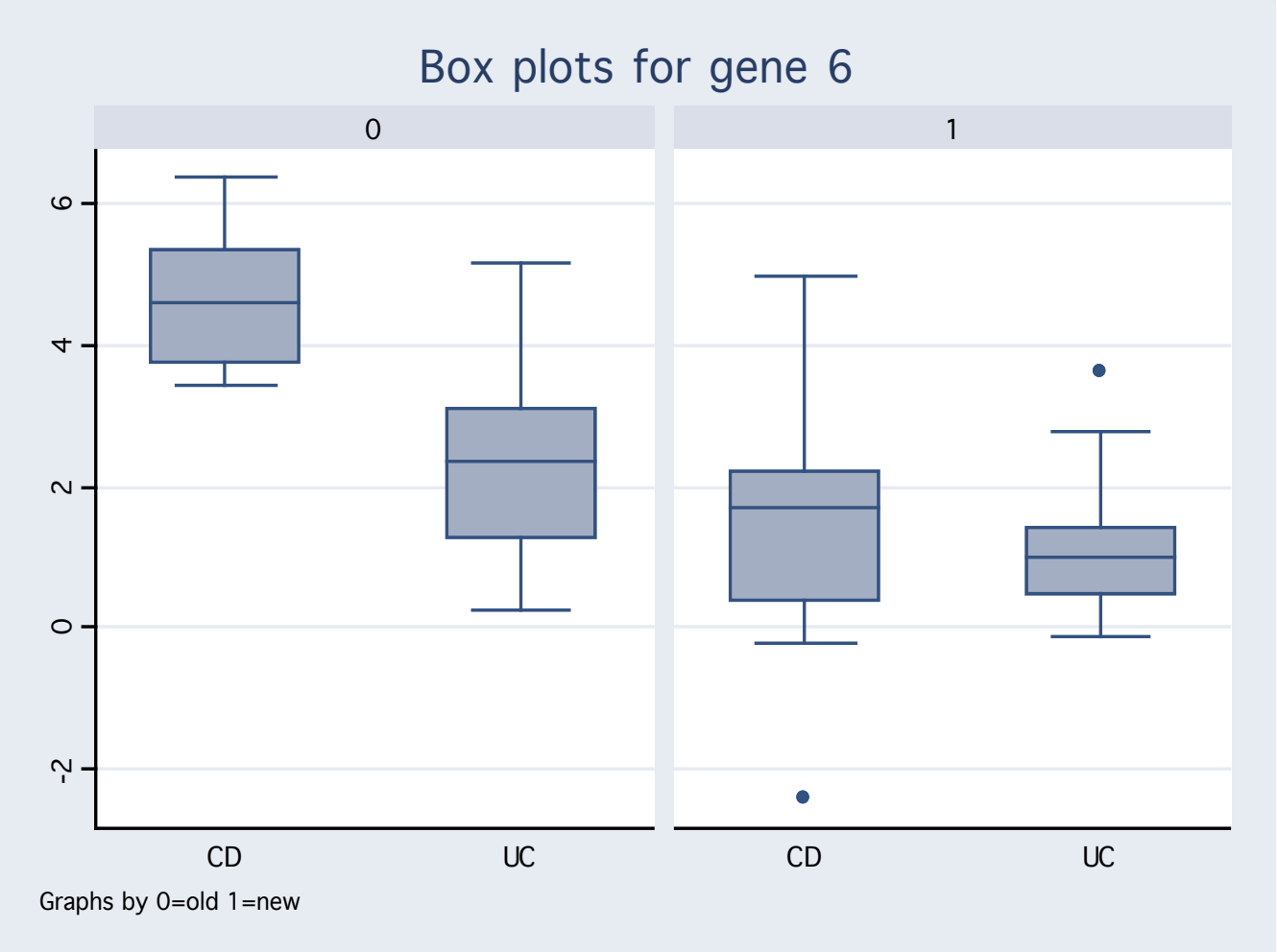


# Box plots for old sample and new sample, and the two diagnoses CD and UC

Gene 6

Old: 0

New: 1



# Kolak: Ref-genes replicates ANOVA

## Failure because too much variation btw runs

| **Source of variation** | **DF** | **MSE(RPLPO)** | **MSE(TBP)** |
|---|---|---|---|
| Treatment | 1 | 0.3 | 0.3 |
| Individuals | 8 | 0.5 | 0.1 |
| **Runs ("time")** | 5 / 2 | **4.4** | **4.1** |
| Pairwise interactions | 53 / 26 | 0.2 | 0.2 |
| Residual | 40 / 16 | 0.4 | 0.1 |

# CONCLUSIONS

Real-time RT–PCR

Can be a powerful technique

But it sometimes fails

Use of reference genes is important