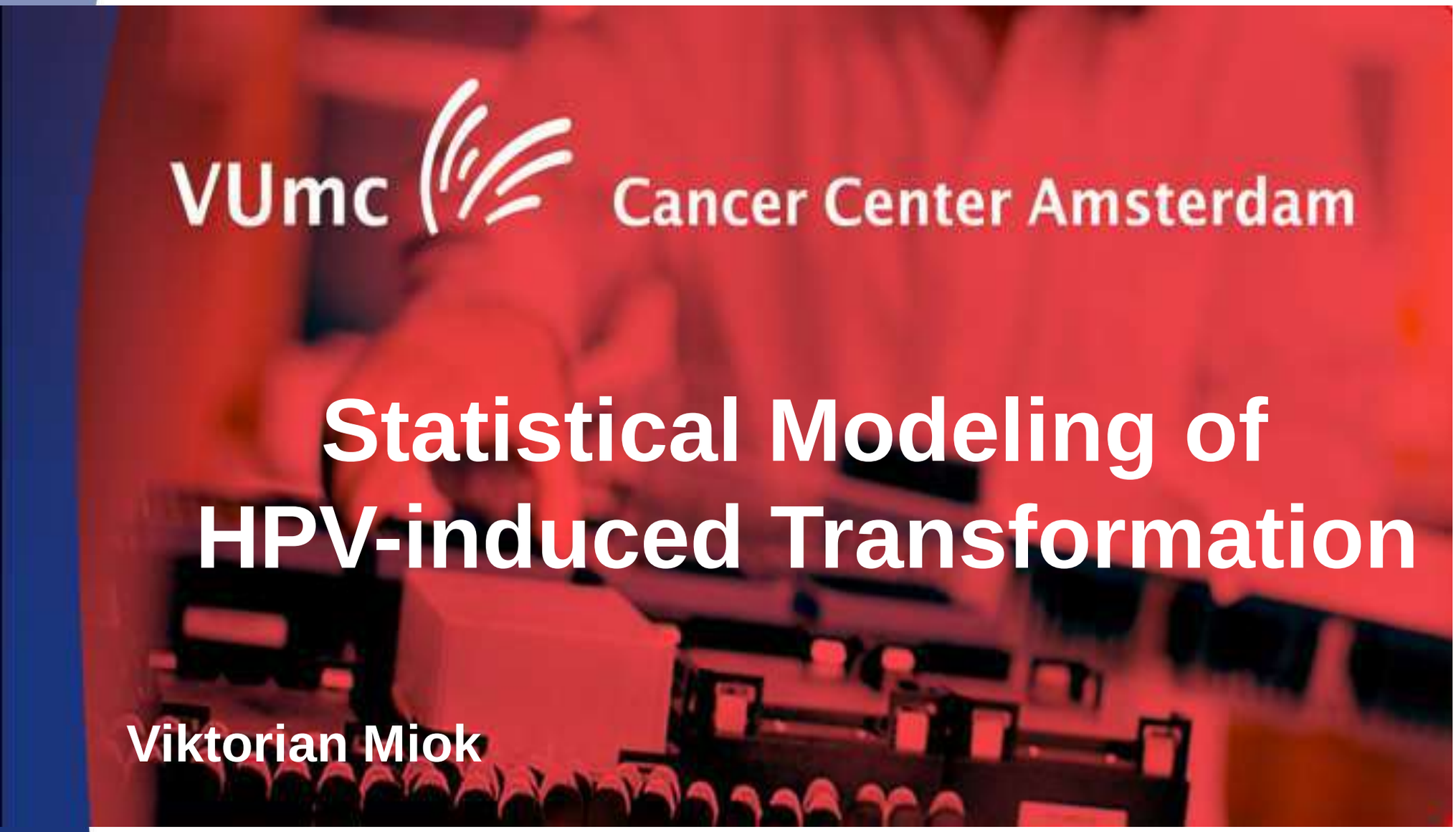


Statistical Modeling of HPV-induced Transformation

Viktorian Miok

Overview

- **Introduction**
- **Method**
- **Comparison**
- **Application**

Introduction

Cervical Cancer

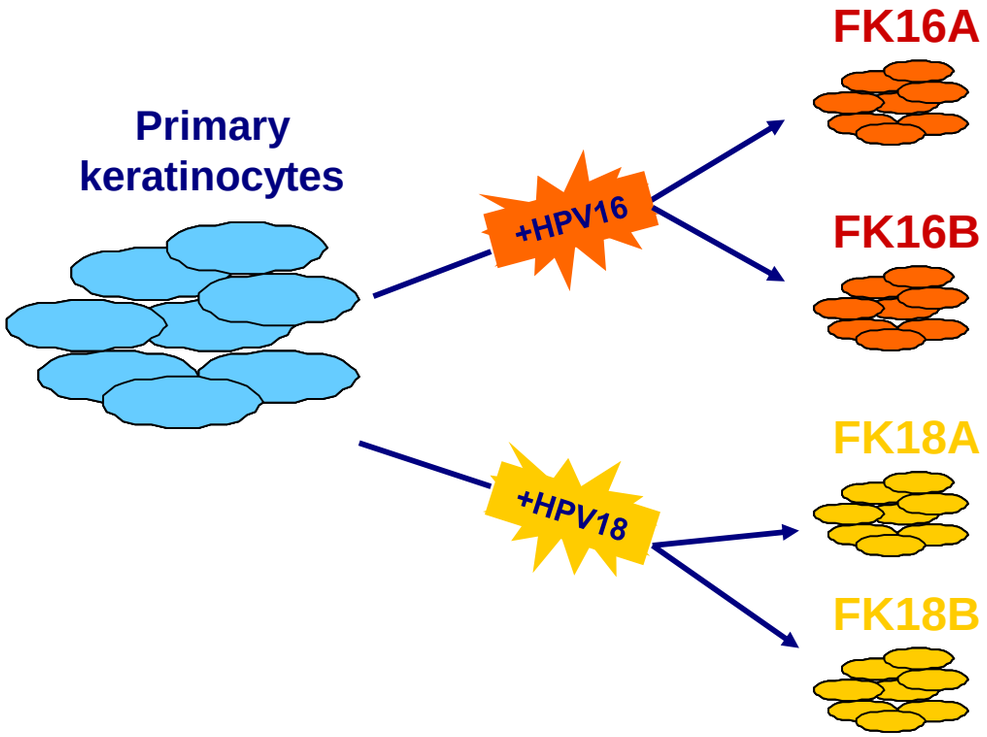
- Second most common cancer in women worldwide.
- The best known example how virus causes cancer.
- Caused by HPV virus, in 80% cases HPV16 and HPV18.
- Genetic disease.

Our approach to investigate HPV-induced transformation

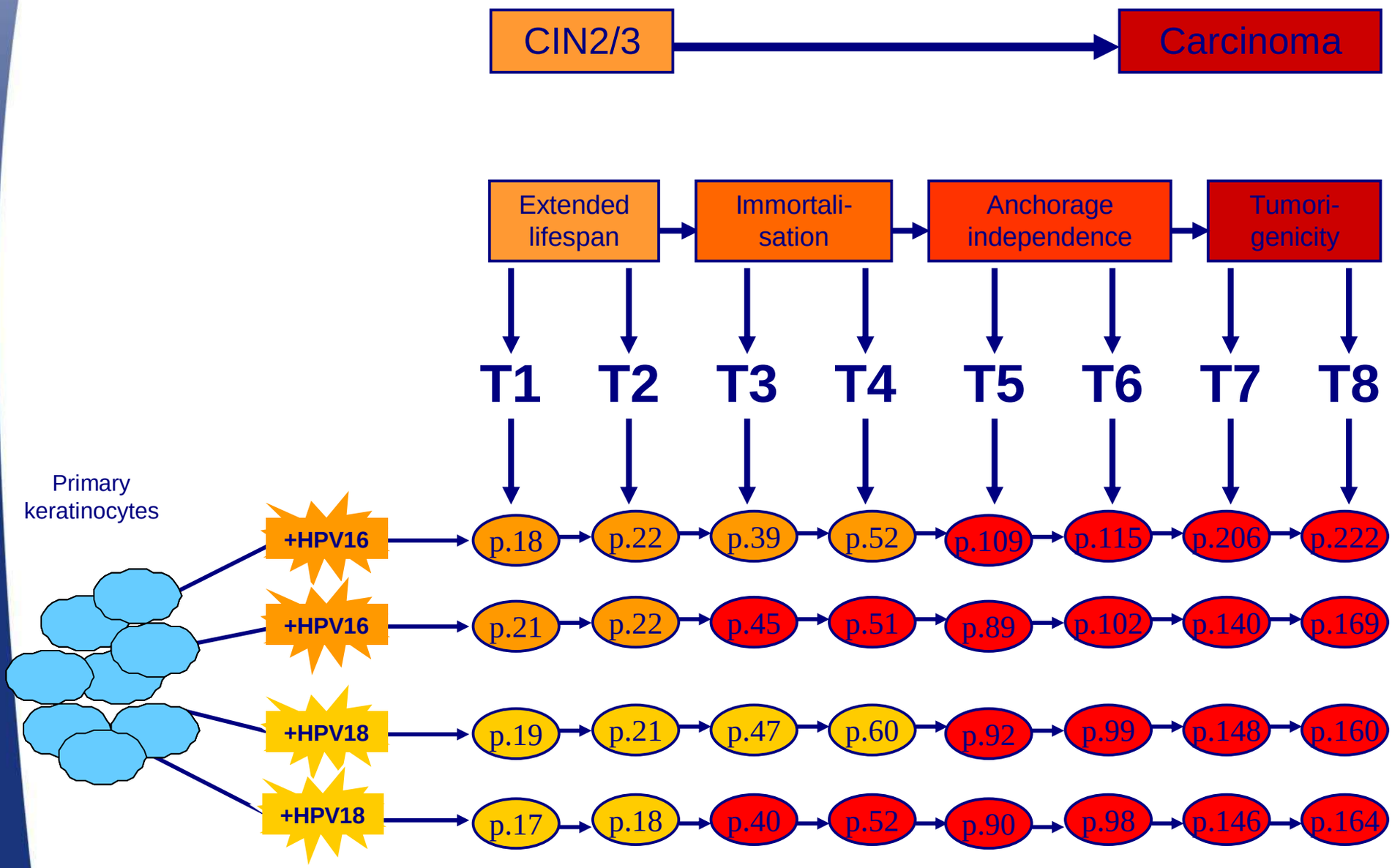
- Cell line model – in vitro model system of HPV-induced transformation
- Time structure – profiling at different moments in time for multiple molecular levels
- Integration – high-throughput multi level molecular data sets
- Aim: Insights in necessary (epi) genetic changes

In vitro model system of HPV-induced transformation

- Suitable model to investigate HPV-induced cancer development.
- Originate from same parental cells.



Time Effect

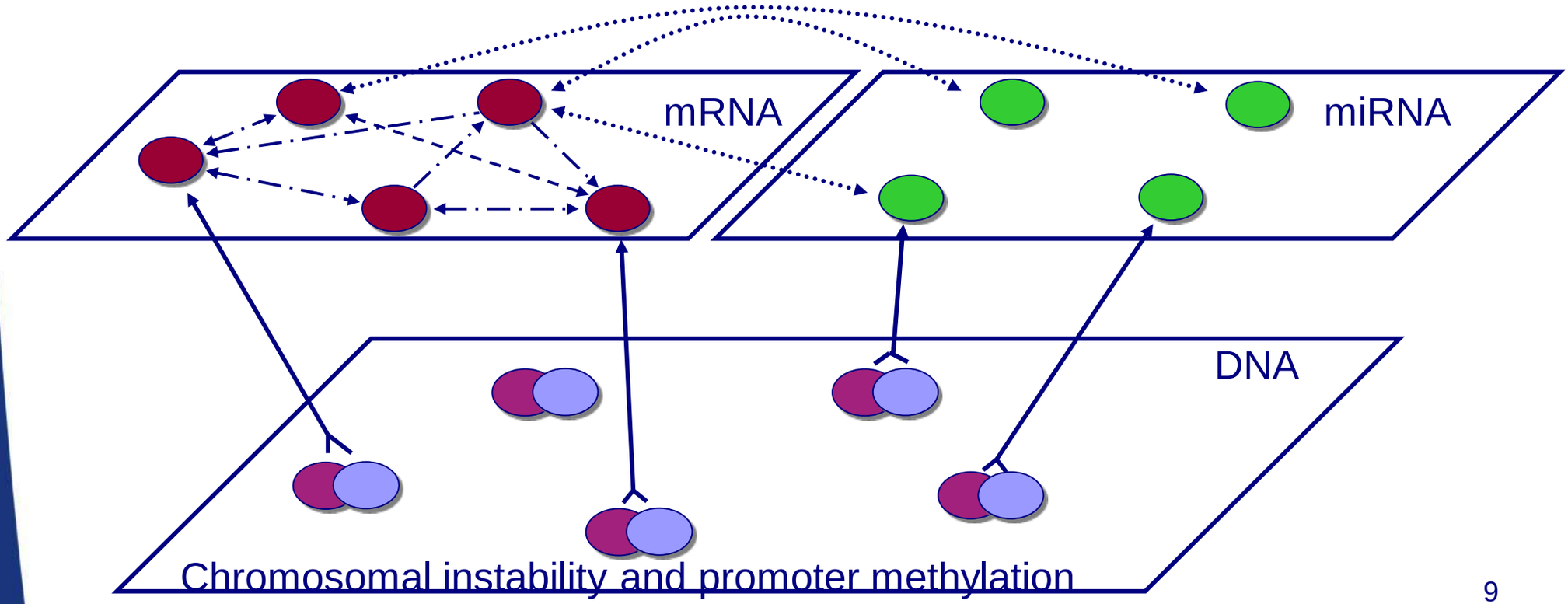

CIN2/3
Carcinoma
Extended lifespan
Immortali-sation
Anchorage independence
Tumori-genicity
T1
T2
T3
T4
T5
T6
T7
T8
Primary keratinocytes
+HPV16
p.18
p.22
p.39
p.52
p.109
p.115
p.206
p.222
+HPV16
p.21
p.22
p.45
p.51
p.89
p.102
p.140
p.169
+HPV18
p.19
p.21
p.47
p.60
p.92
p.99
p.148
p.160
+HPV18
p.17
p.18
p.40
p.52
p.90
p.98
p.146
p.164

DNA Methylation

- Mechanism of epigenetic deregulation – gene silencing.
- Known to be involved in HPV-induced transformation.
- Measured in mRNA and miRNA gene expression experiments
- Comparing mRNA and miRNA gene expression in the samples treated with DAC and in samples without treatment.
- Results are partially validated using Illumina human methylation arrays.

Multilevel Integrative Analysis

- Goal of project is to perform multilevel integrative analysis of longitudinal data sets
- Necessity of novel integrative statistical methods



Method

Integration

GE = Cell Line + Time

↓

GE = Cell Line + CN + Time

↓

GE = Cell Line + CN+ Time + Methylation

Temporal Differential Gene Expression

- Generalized linear mixed model

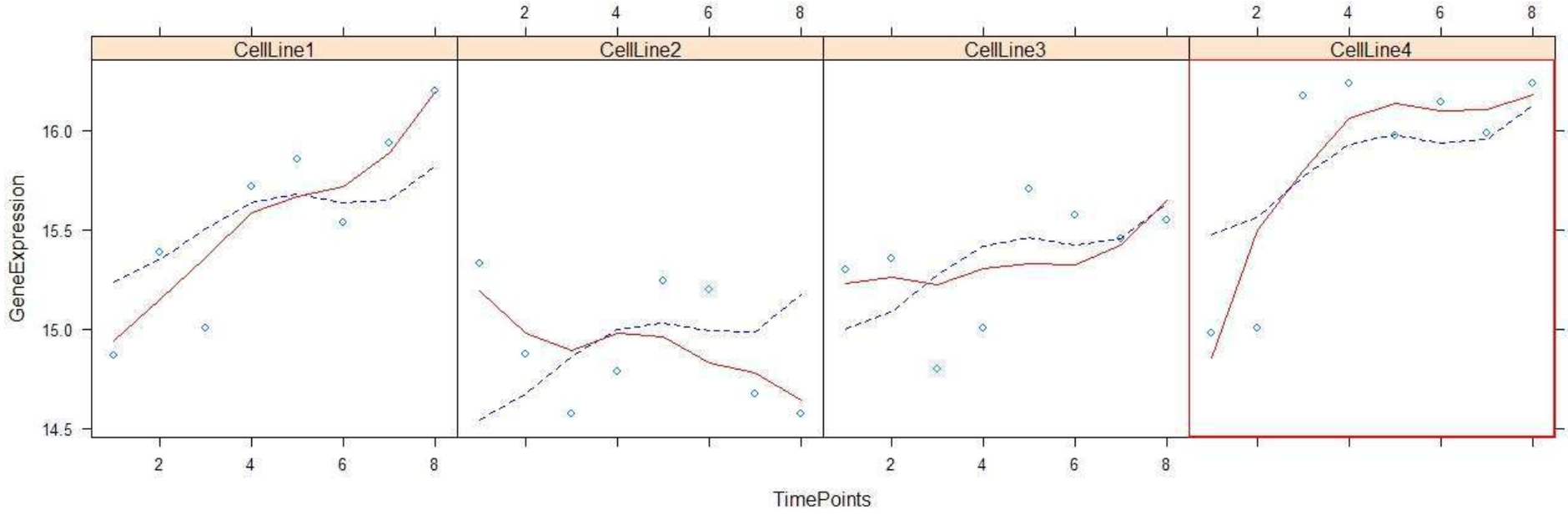
GE = Cell Line + CN + Time

fixed effect random effect

- Penalized splines
- Parameters and hyper-parameters estimated using empirical Bayes procedure employing INLA
- Shrinkage of dispersion-related parameters
- The likelihood ratio test

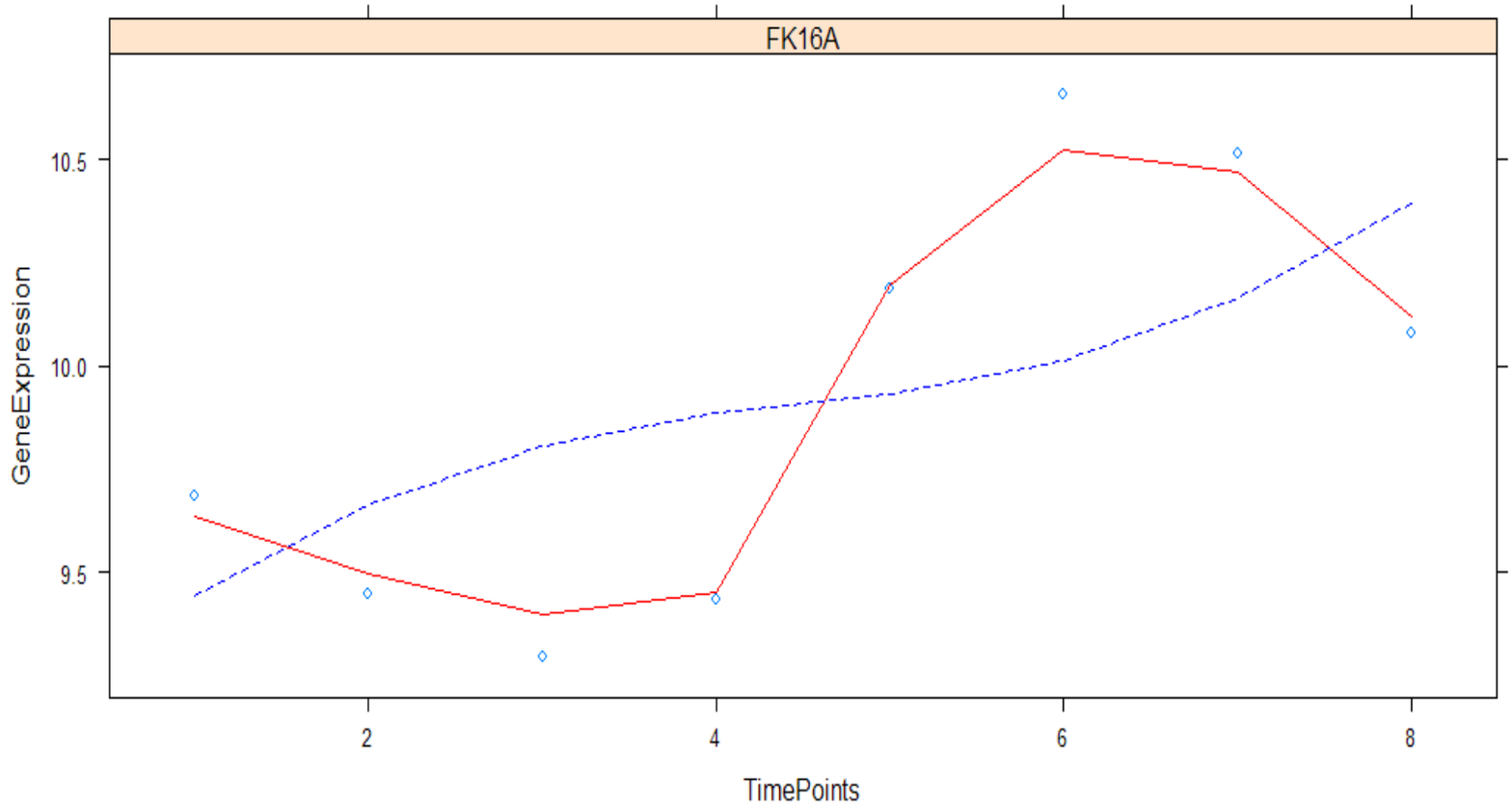
Time effect

- Different spline per cell line
- Same spline for all cell lines



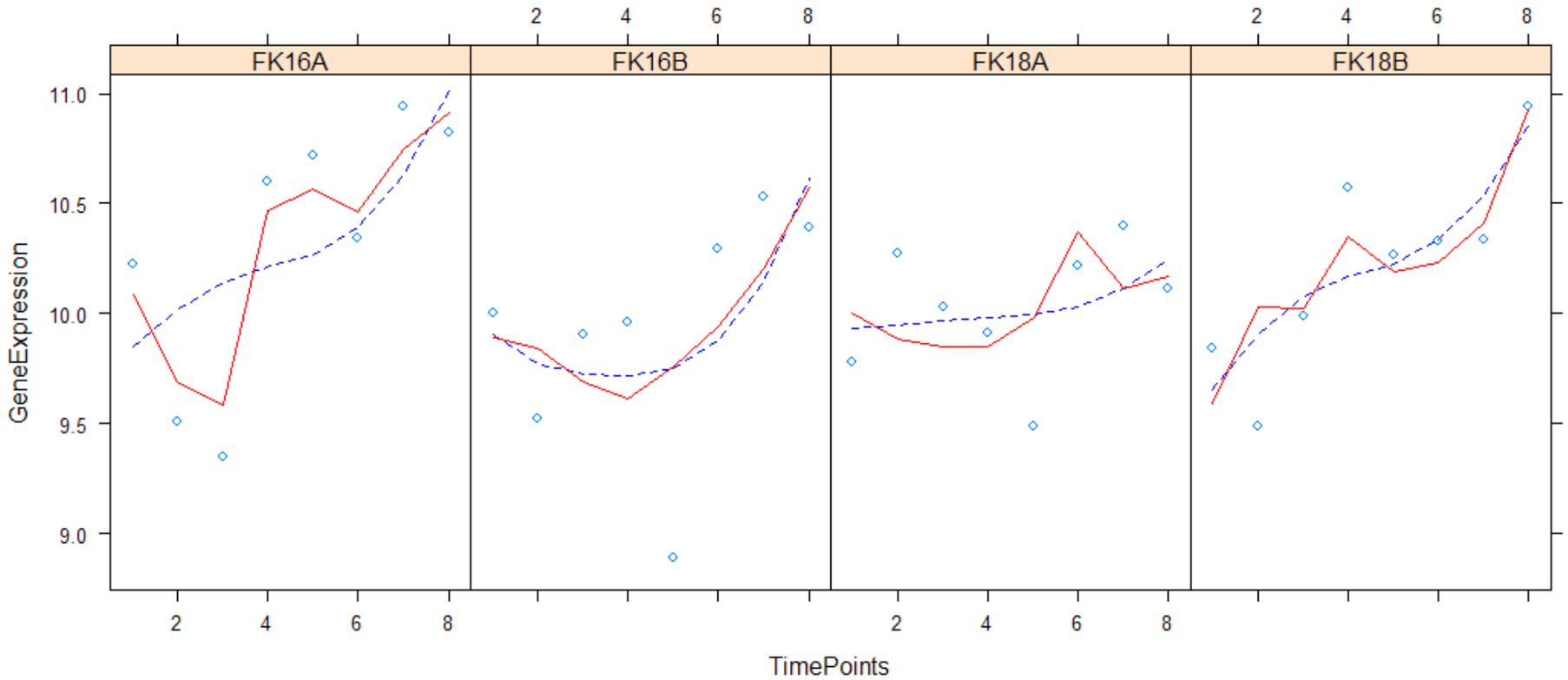
Copy number Effect

- Cell line + CN + Time
- Cell line + Time



Gene - SLC25A36

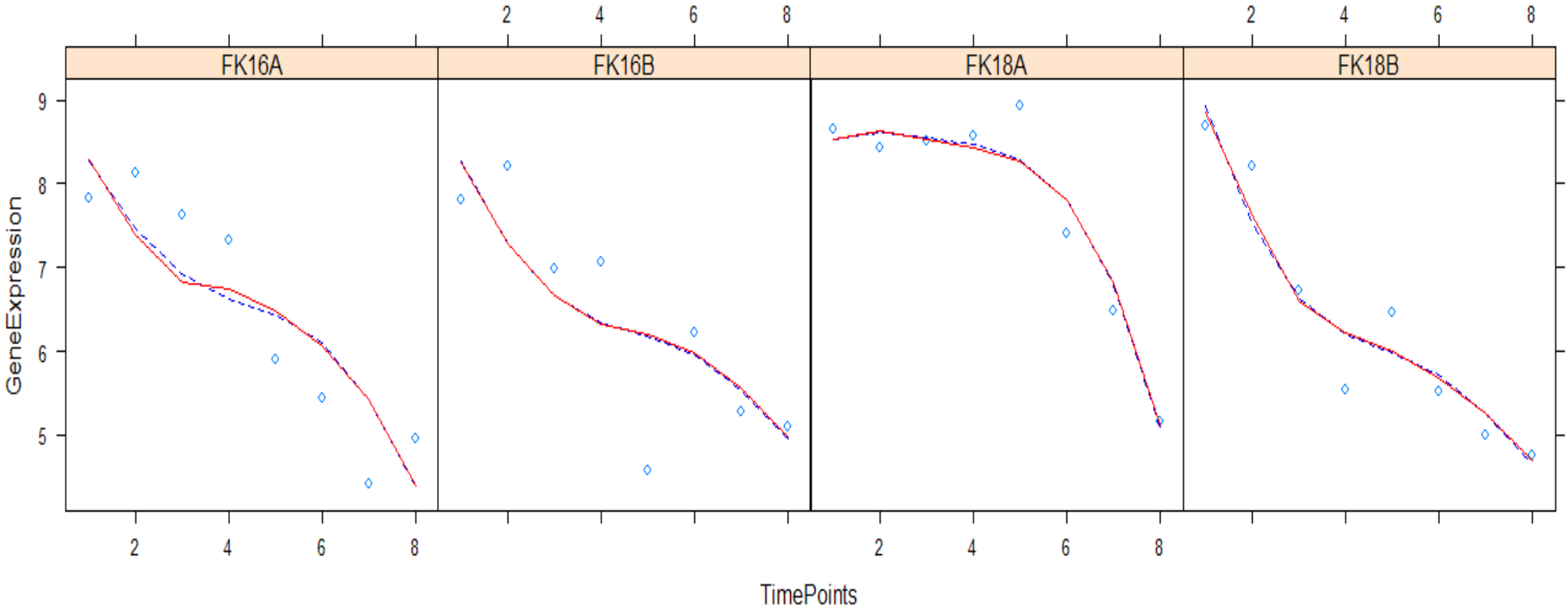
- Cell line + CN + Time
- Cell line + Time



Wilting et al., Genes,Cromosomes&Cancer, 2008

Gene - CADM1

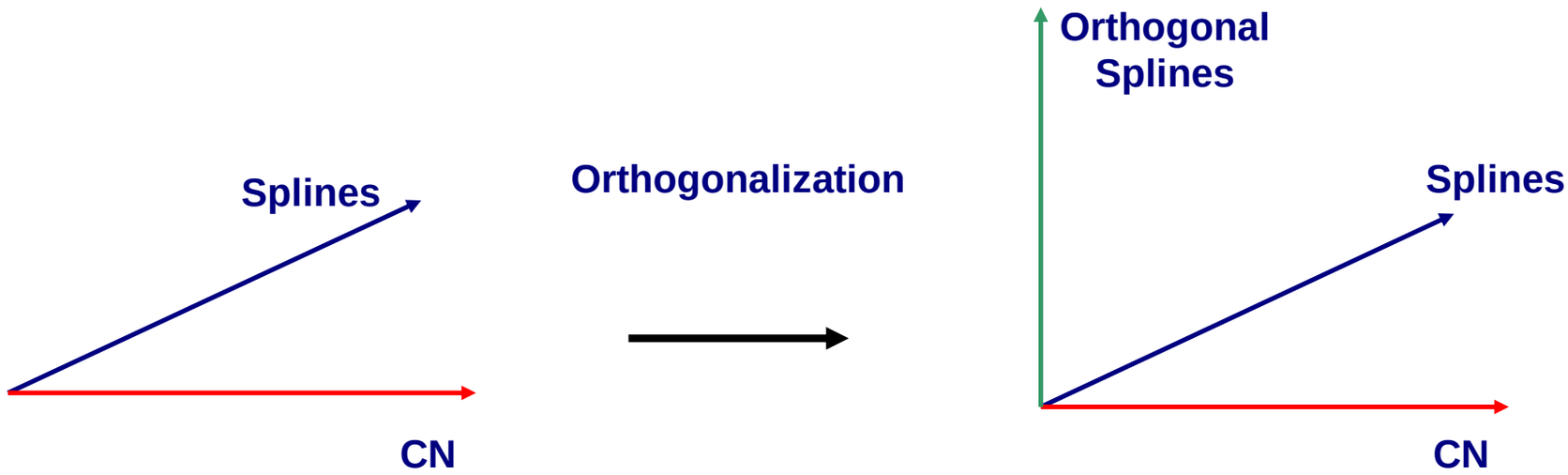
- Cell line + CN + Time
- Cell line + Time



Steenbergen et al.,JNCI, 2004
Overmeer et al., J Pathol., 2008

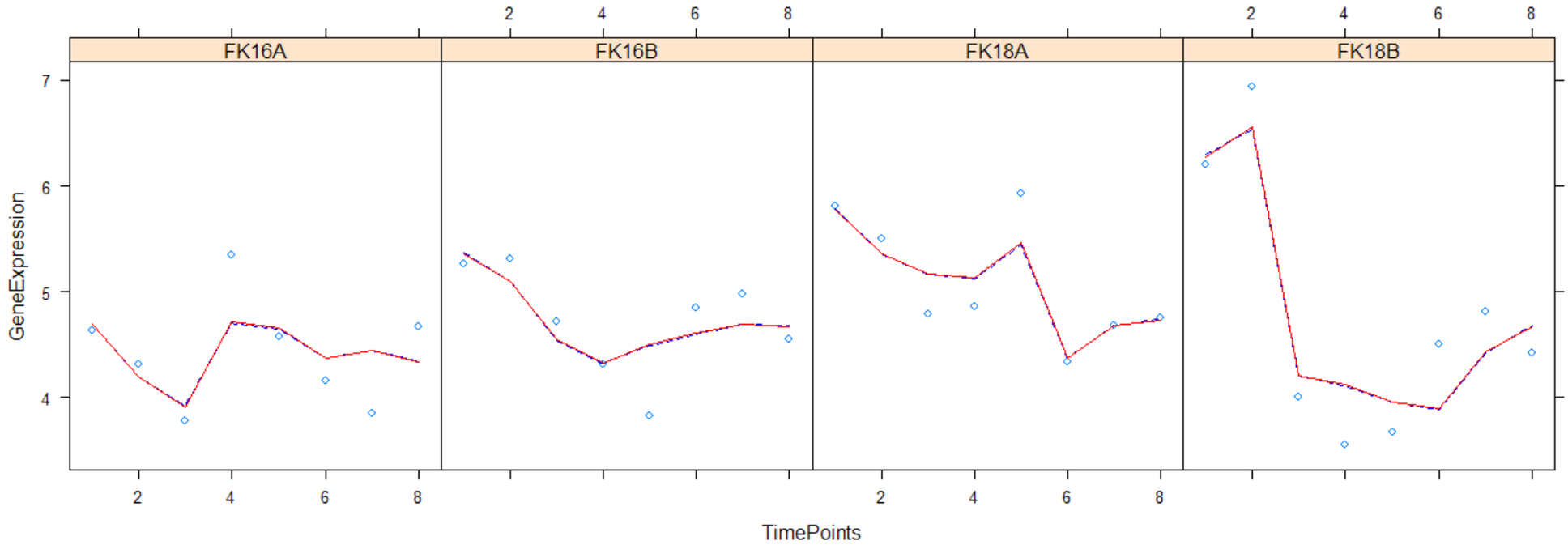
Competition between time and CN effect

- Time effect – flexibility of splines consume effect of copy number
- Potential solution
  - Orthogonalization of splines onto copy number design matrix



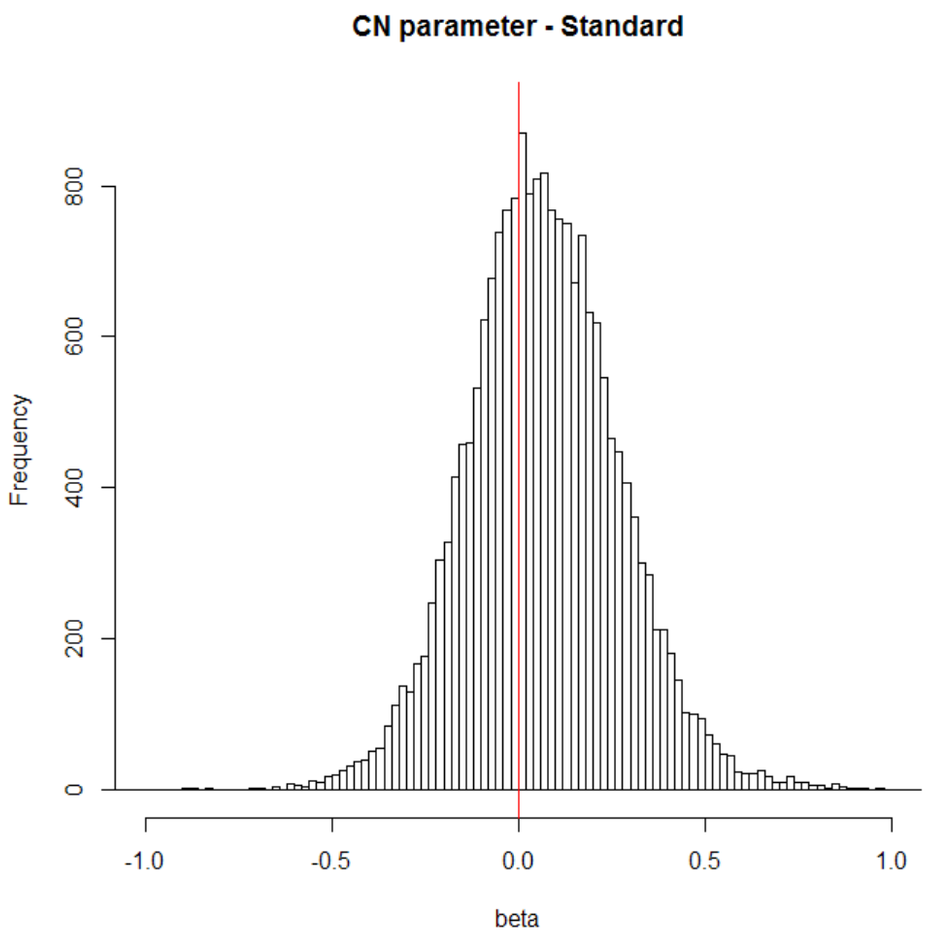
Fitting - orthogonal vs. standard

- Cell Line + CN + orthogonal (Time)
- Cell Line + CN +Time

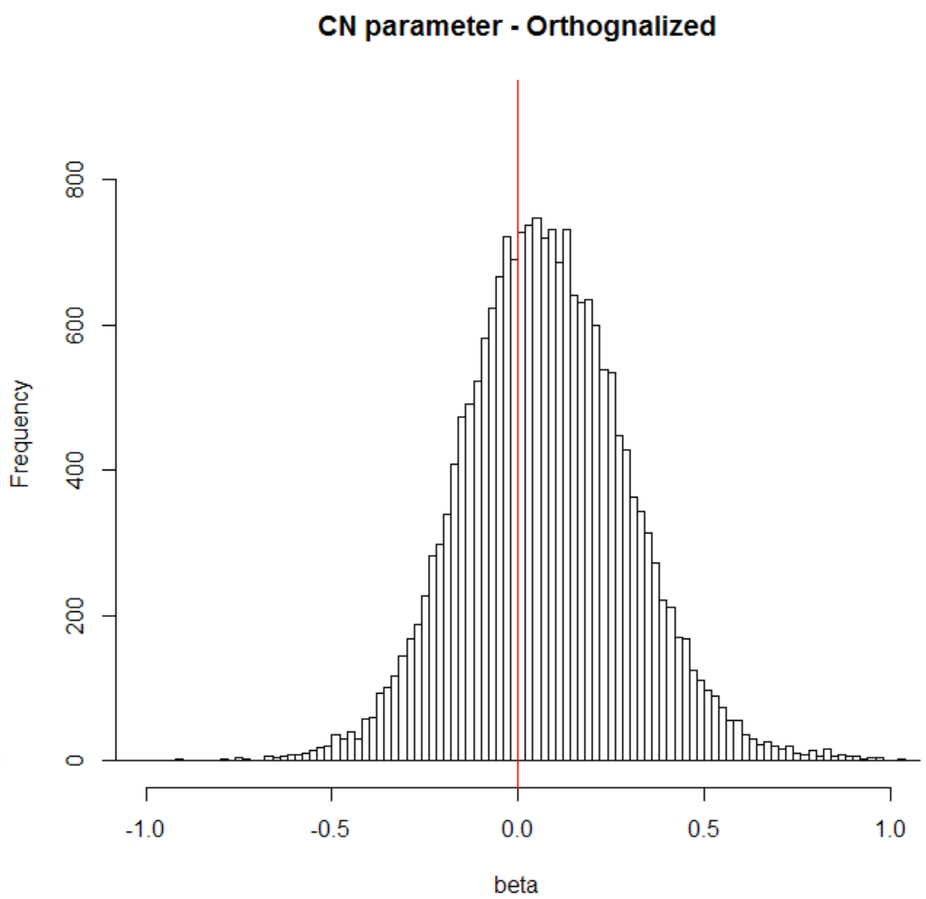


Parameter of CN effect

Standard

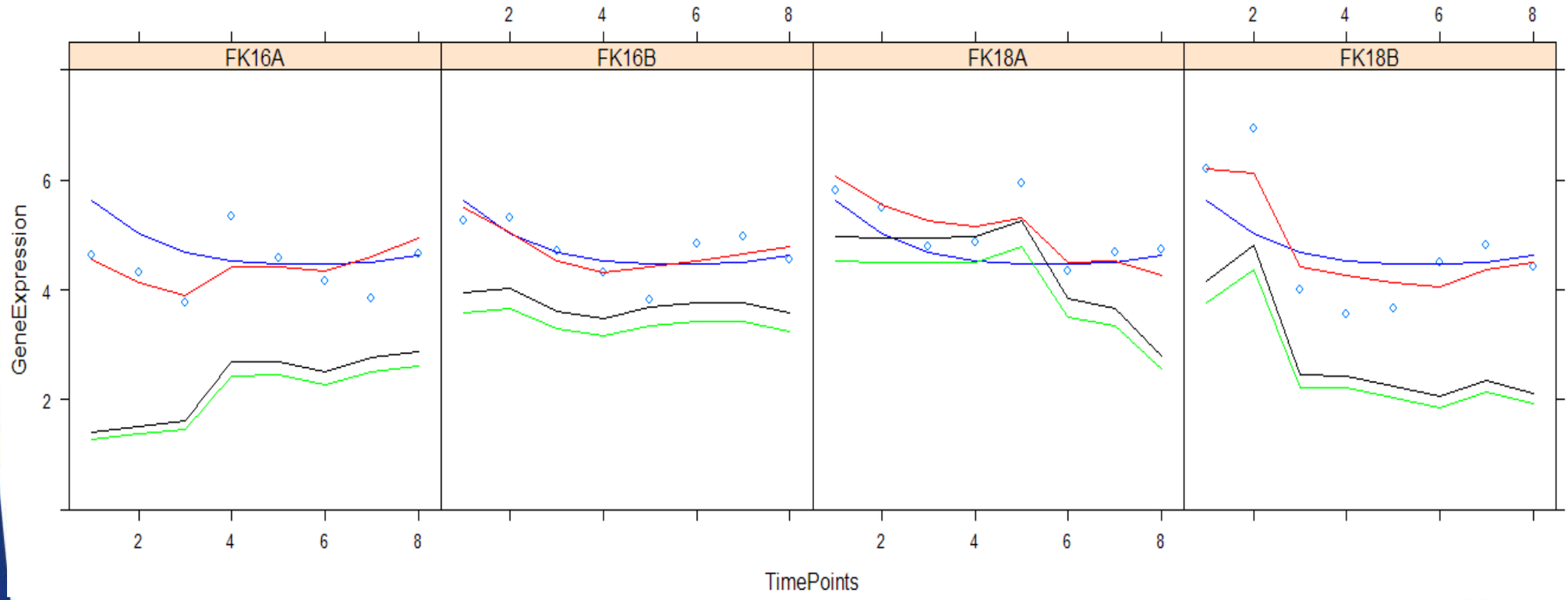


Orthogonalized



Fitting - orthogonal vs. standard design

- Full model
- Time
- Orthogonal (CN)
- CN



Comparison

Comparison

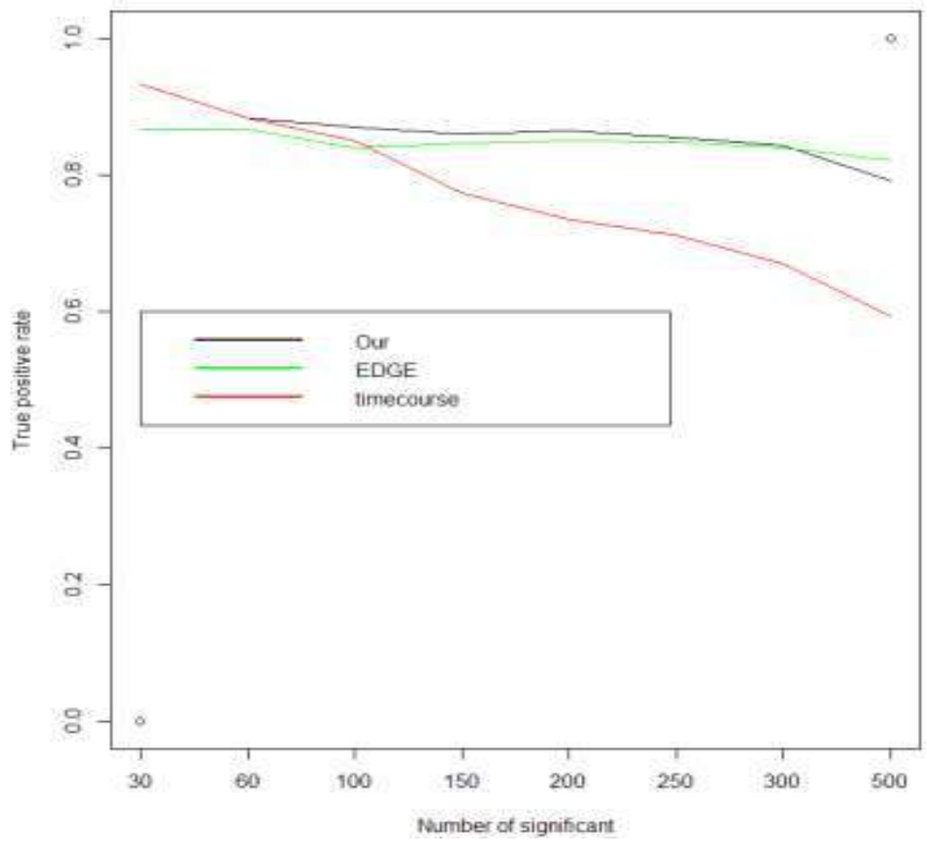
- Comparison of our method with
  - **timecourse** – Tai and Speed, Annals of Statistics, 2006.
  - **EDGE** – Storey et al., PNAS, 2005.
- Fair comparison – model without copy number
- Method is applied on two data sets
  - Data from our experiment
  - Data from Storey et al., PNAS, 2005.

Comparison with **timecourse** and **EDGE**

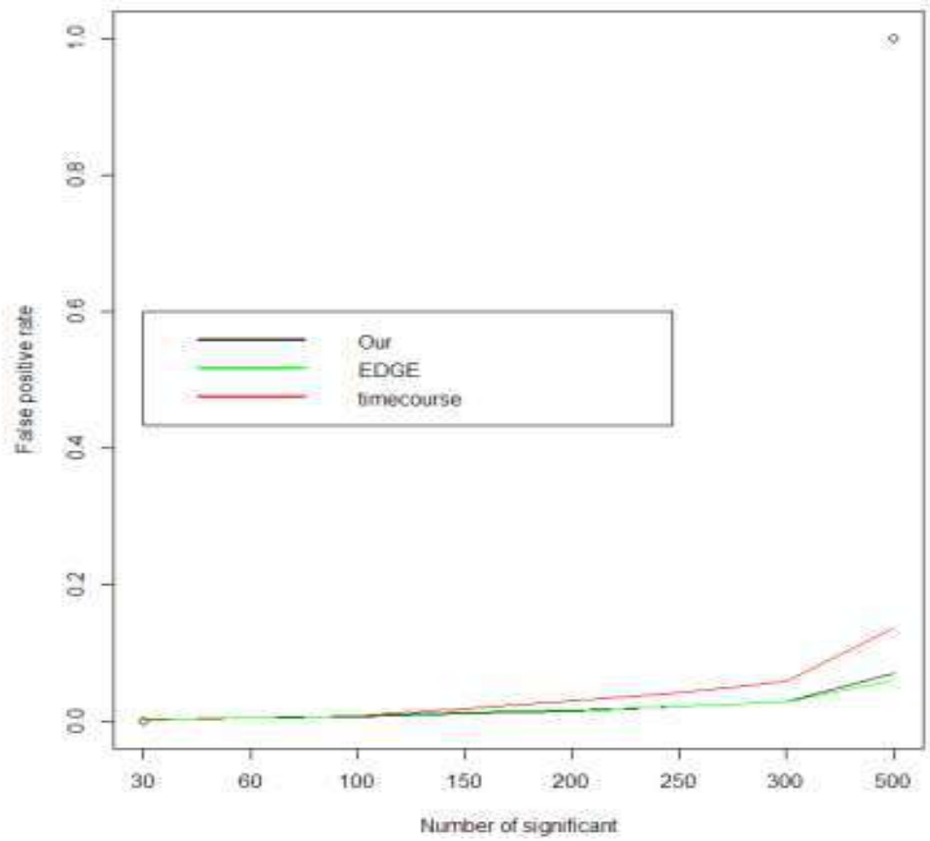
- Sensitivity and specificity
  - Truth – significant genes among methods.
  - For different number of significant genes calculate true positive and false positive rate.
- Reproducibility
  - Equally divided data set in two groups
  - For different number of significant genes calculate number of overlap genes.

Sensitivity and specificity – our data

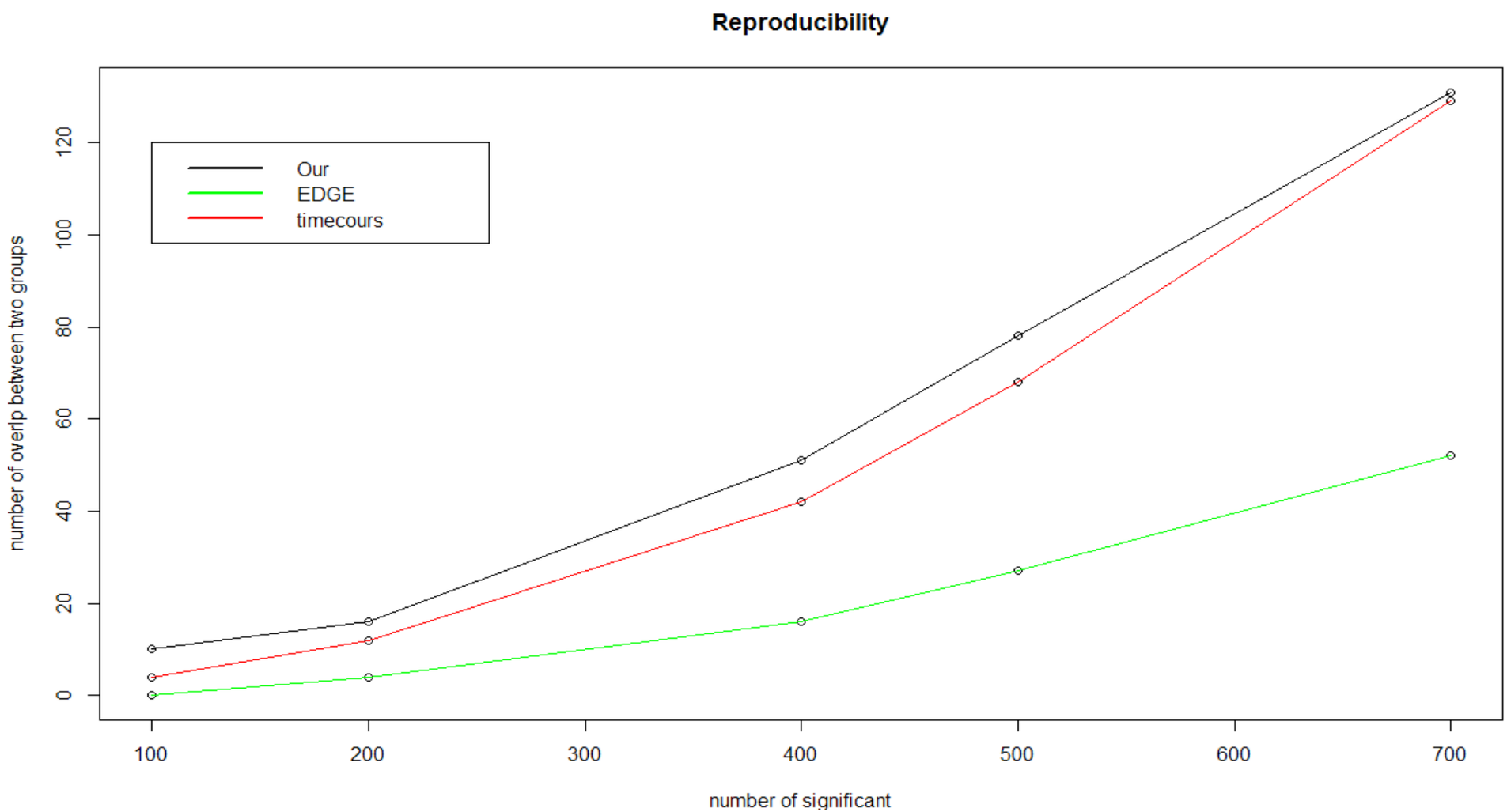
- Sensitivity



- Specificity

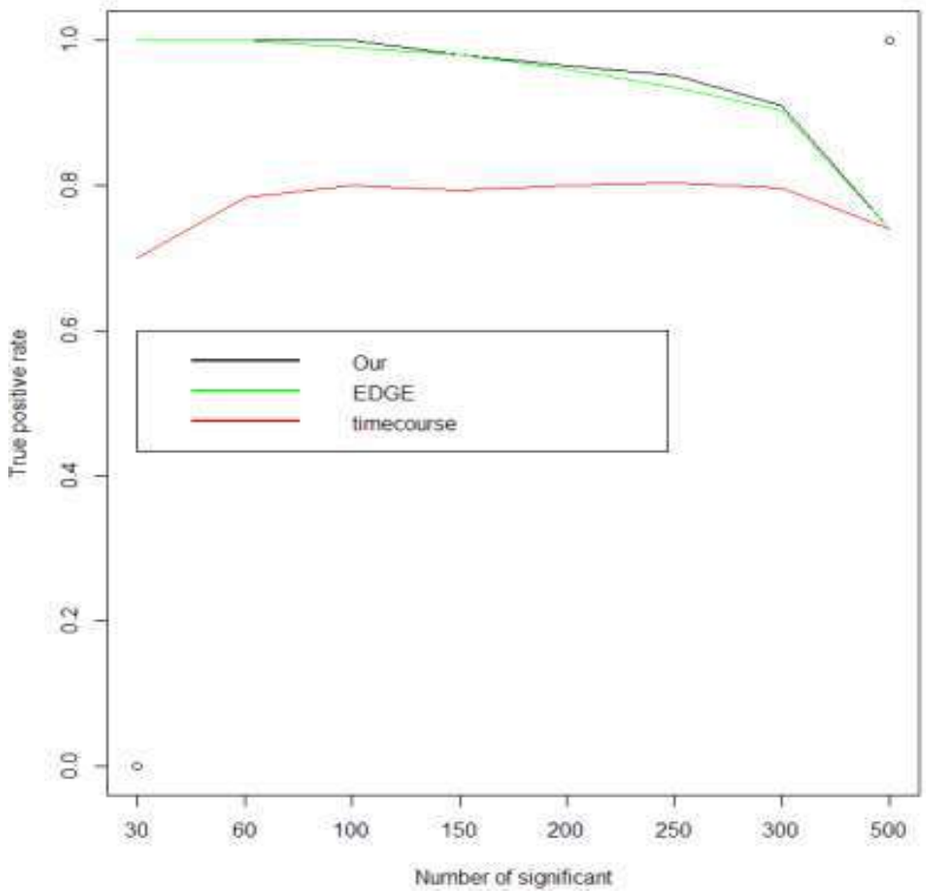


Reproducibility – our data

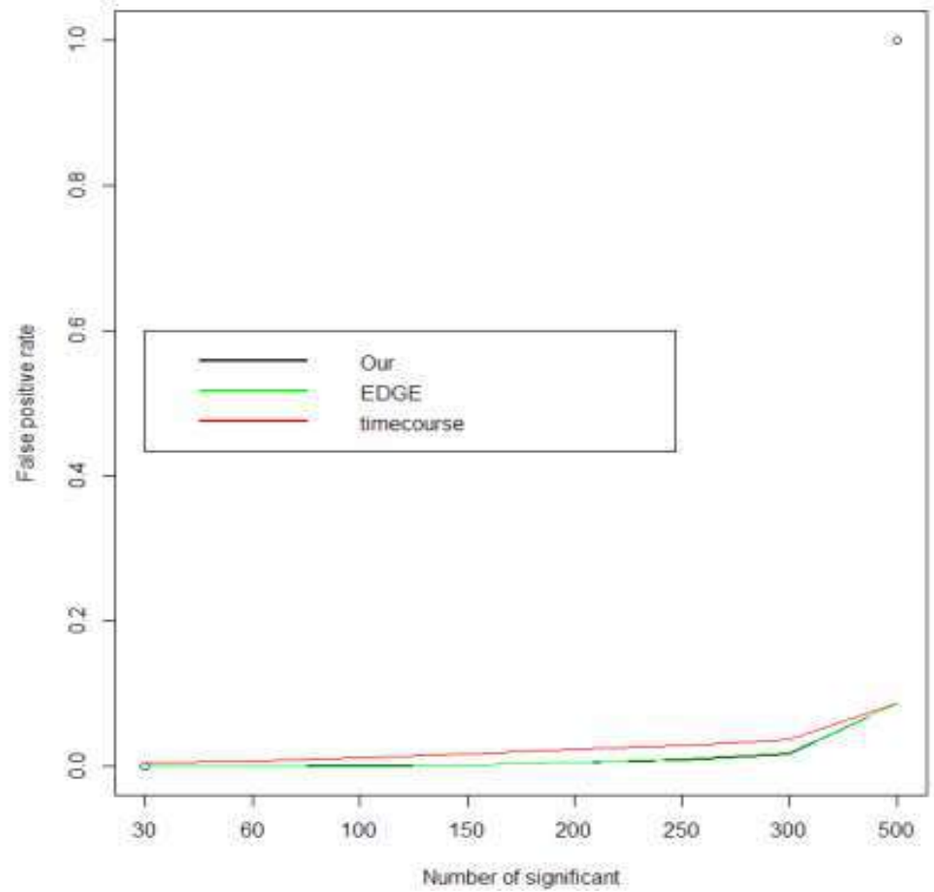


Sensitivity and specificity - EDGE data

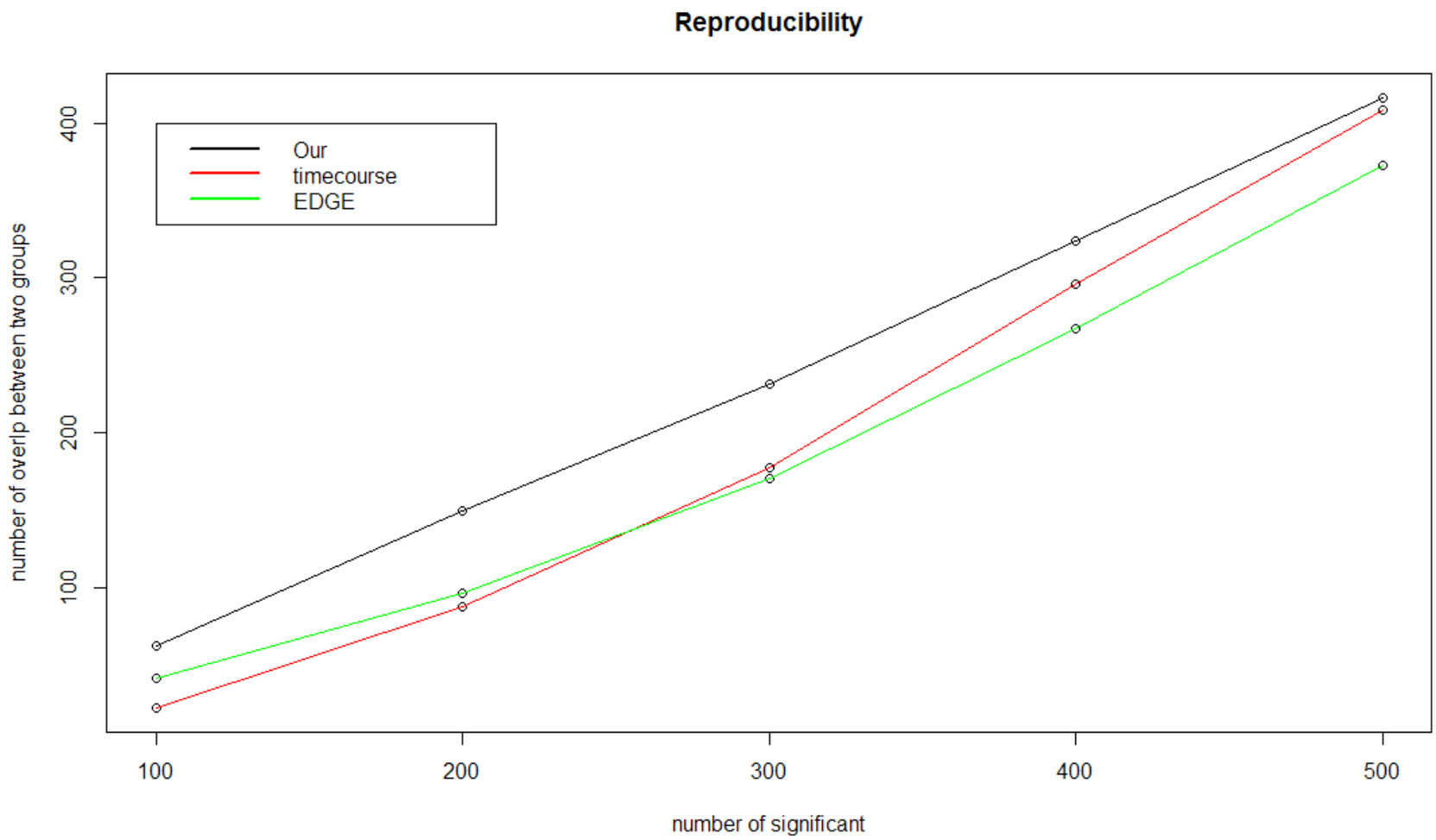
- Sensitivity



- Specificity



Reproducibility – EDGE data



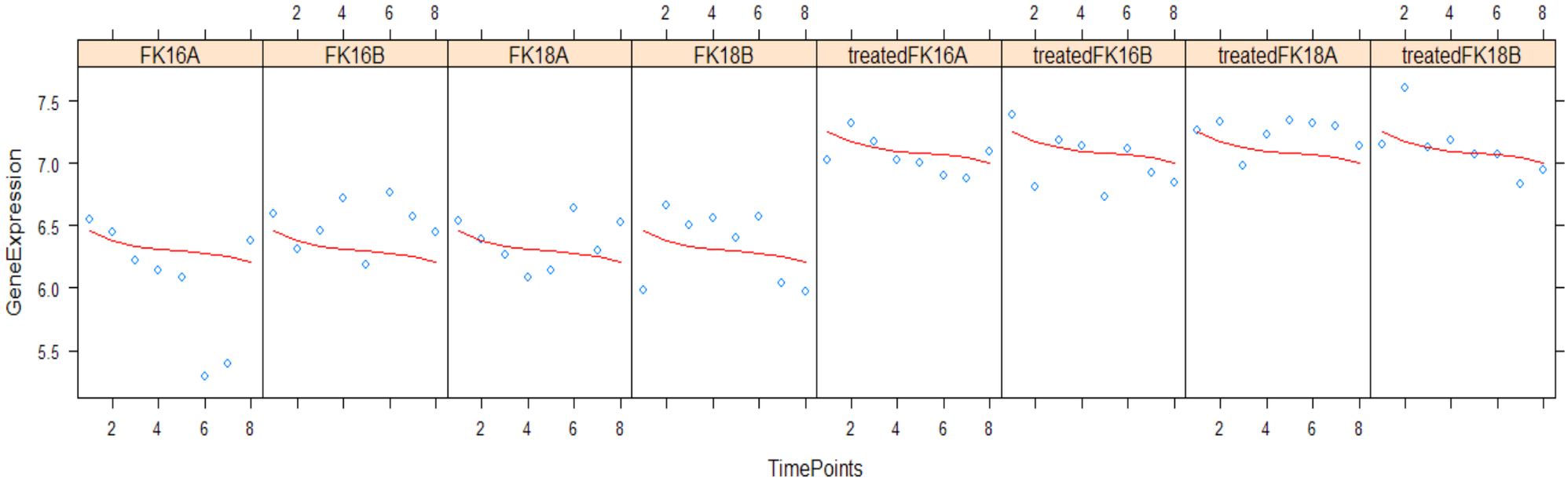
Application

Testing for differential gene expression

Number of significant genes at 5% FDR level	Same Spline		Different Spline	
	Standard	**Orthogonal**	**Standard**	**Orthogonal**
CN Effect	524	544	923	936
Time Effect	4959	5011	10487	10567

Testing for treatment effect

- Treatment effect - 16976 genes significant at 5% FDR level



Summary

- Generated a reliable multilevel molecular dataset.
- Developed generalized linear mixed model for identification of temporal differential gene expression.
- Improvement in identification of variation in genes due to copy number effect
- Method show improvements in sensitivity, specificity and reproducibility comparing with other methods.

Future plans

- Developing methodology to include spatial effect over genome for copy number
- Including DNA methylation in model
- Integrative analysis for miRNA
  - Applying method on miRNA data
  - Pathway analysis

Acknowledgments

- Wessel van Wieringen
- Saskia Wilting
- Annelieke Jaspers
- Mark van de Wiel
- Peter Snijders
- Renske Steenbergen

Thank you for your attention