Integrative statistical modeling of time-series omics data from HPV-induced carcinogenesis

Viktorian Miok

Epidemiology & Biostatistics
VUmc

VU VRIJE UNIVERSITEIT AMSTERDAM

Contributors

Biostatistics department

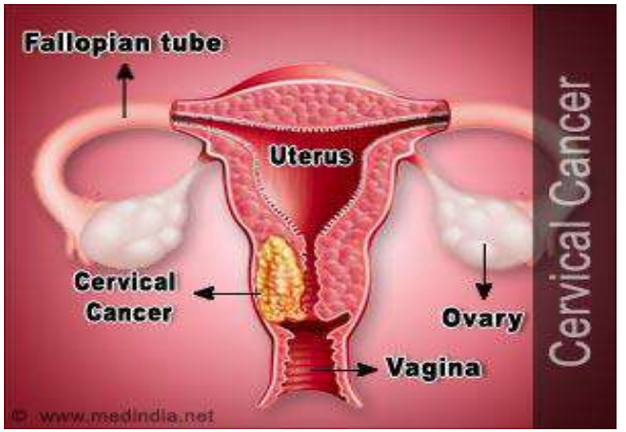
- Viktorian Miok
- Wessel van Wieringen
- Mark van de Wiel

Pathology department

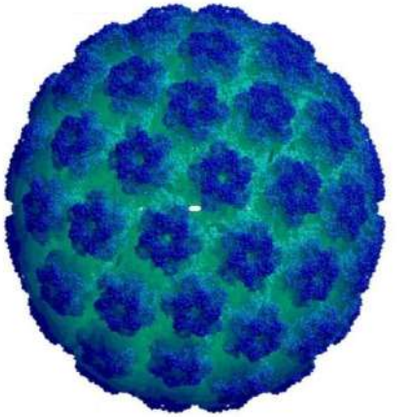
- Saskia Wilting
- Annelieke Jaspers
- Renske Steenbergen
- Peter Snijders

Introduction

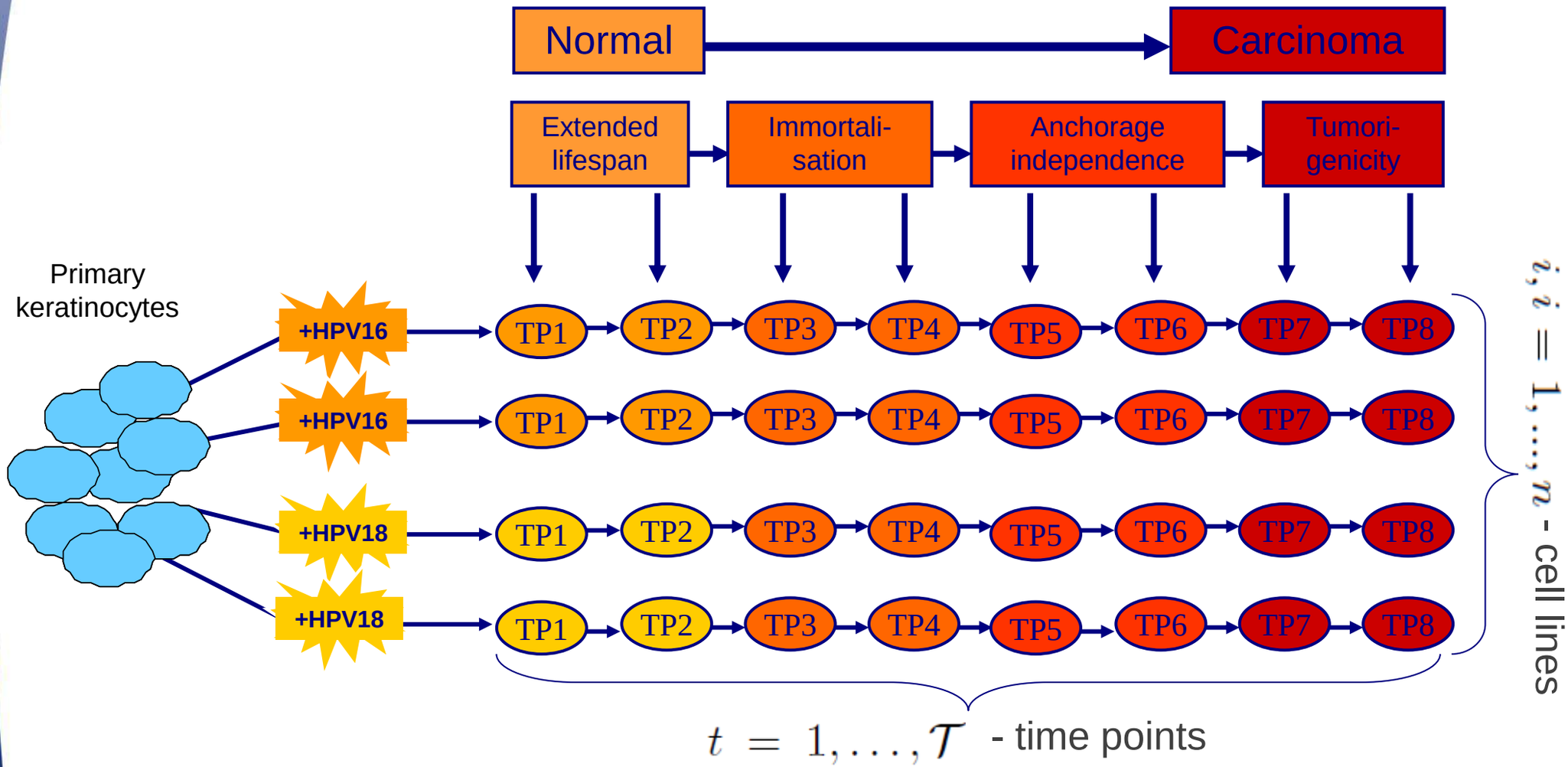
Cervical cancer study



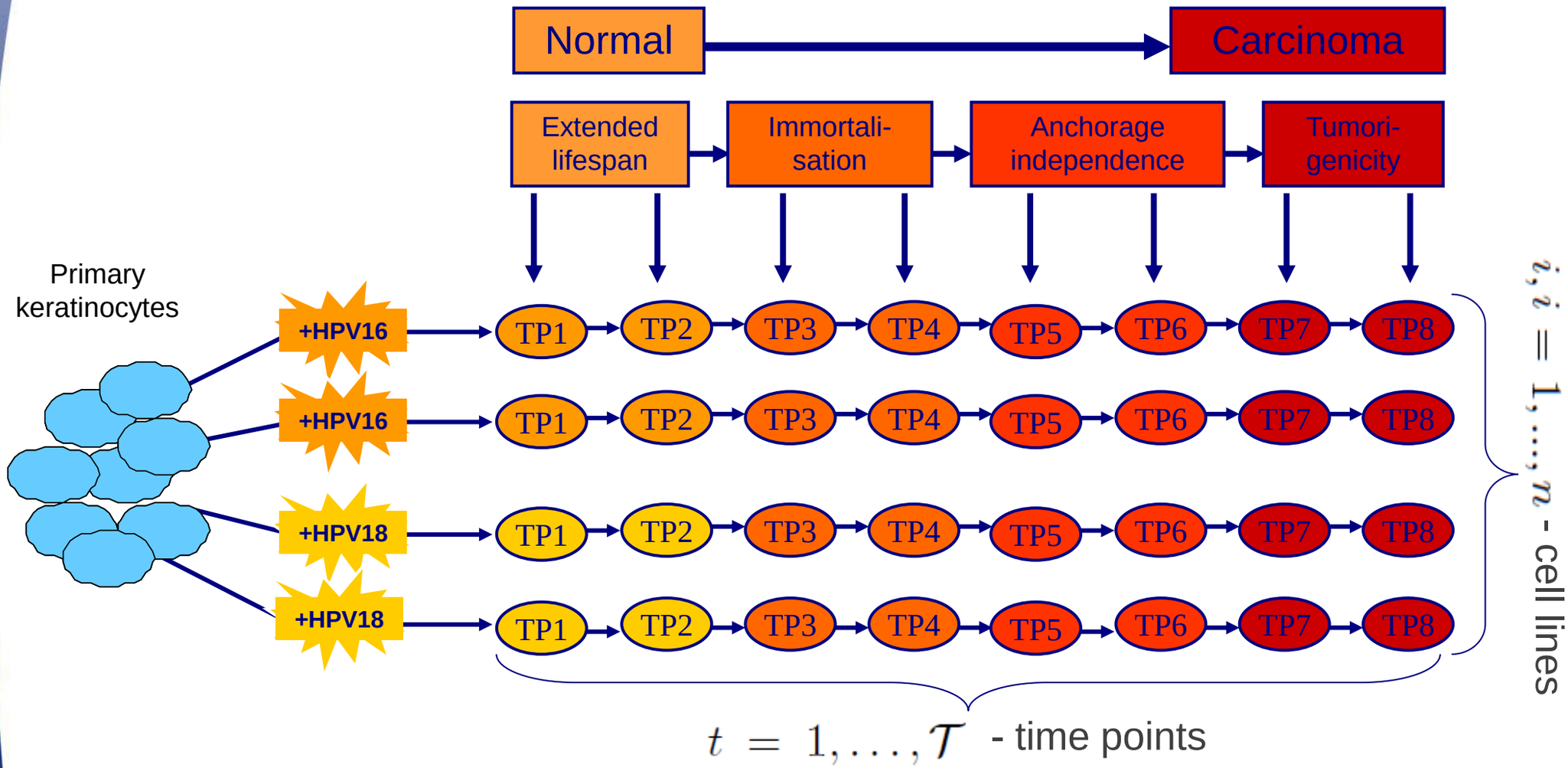
- Second most common cancer in women worldwide.
- Caused by HPV virus (70% cases HPV16 and HPV18) and followed by additional (epi)genetic abnormalities.
- Cell line model – in vitro model system of HPV-induced transformation.
- Integration – high-throughput multi level molecular data sets.
- Understand molecular mechanism driving cervical carcinogenesis



Time-course experiment



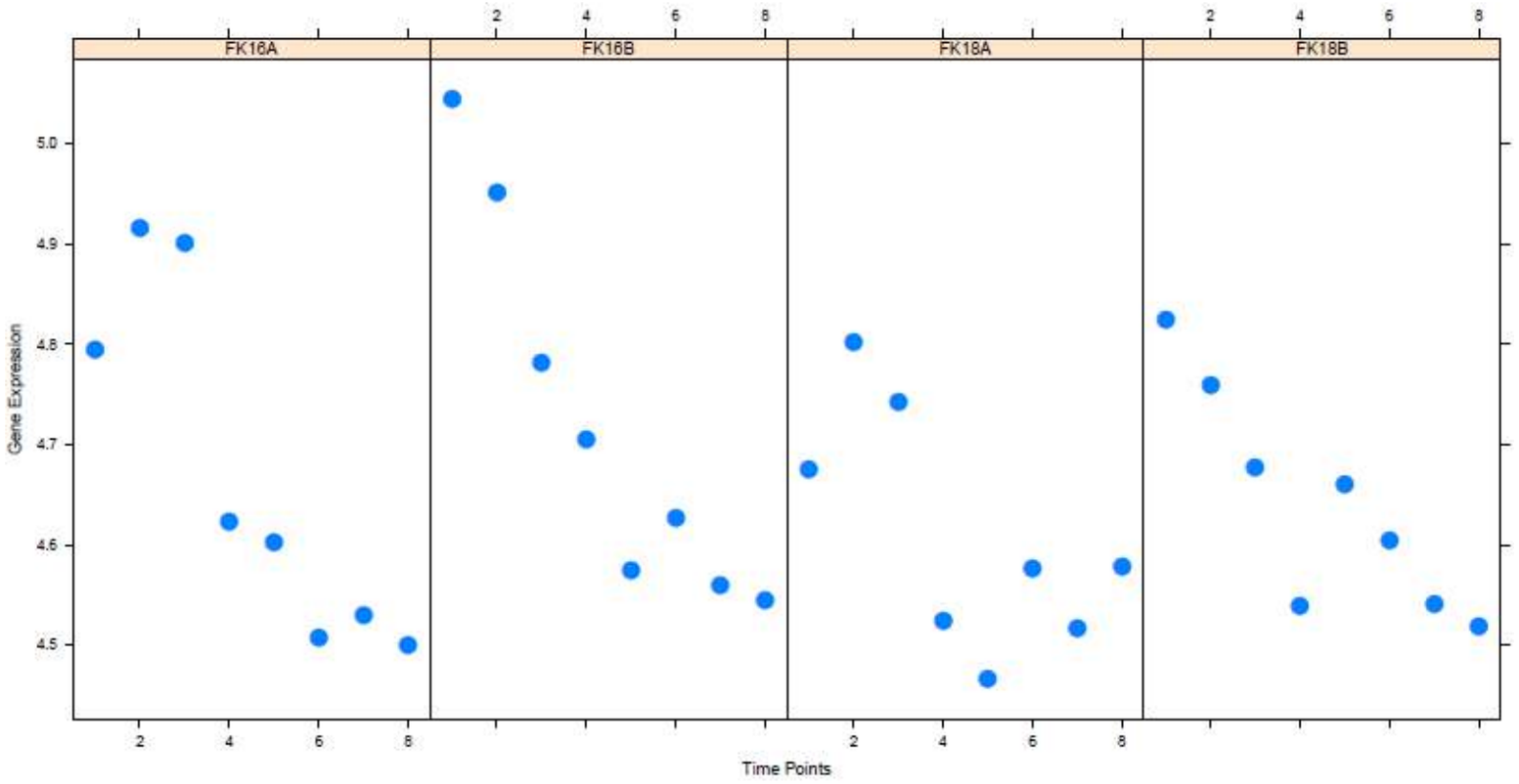
Time-course experiment



	ProbeName	GeneName	X16A_T1.	X16A_T2.	X16A_T3.	X16A_T4.	X16A_T5.	X16A_T6.	X16A_T7.	X16A_T8.
probe1	A_24_P413470	TP73	6.321781	5.988753	6.135580	6.269486	6.281496	6.219492	5.959211	6.674865
probe2	A_32_P73775	KIAA0495	7.225902	7.466130	6.650675	7.079626	7.112270	6.173071	5.477608	6.873047
probe3	A_23_P381478	CCDC27	4.840329	5.249147	7.181347	4.883199	4.773940	7.975527	7.901096	5.269618
probe4	A_23_P96775	LRRC47	10.239232	10.223634	9.958672	10.333902	9.799422	9.213111	9.656384	10.101973
probe5	A_23_P200043	KIAA0562	9.513586	9.292091	8.590563	9.070757	9.020399	8.126005	8.808979	9.636166
probe6	A_23_P44546	DFFB	6.757972	7.046299	6.541672	6.517207	6.675371	7.754047	6.935462	7.113258

Why time-course experiments?

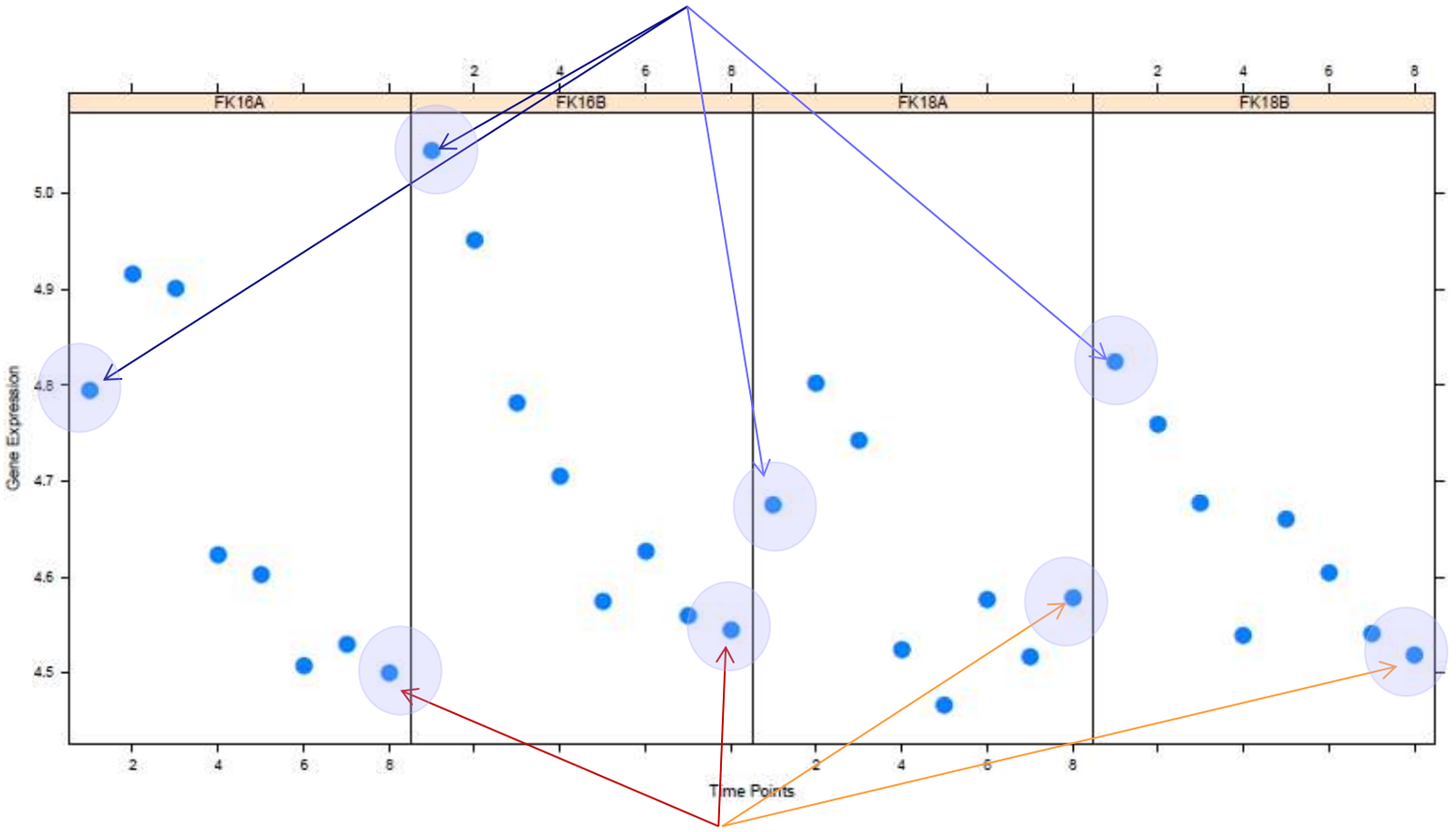
miR-218:



Pick one moment in time

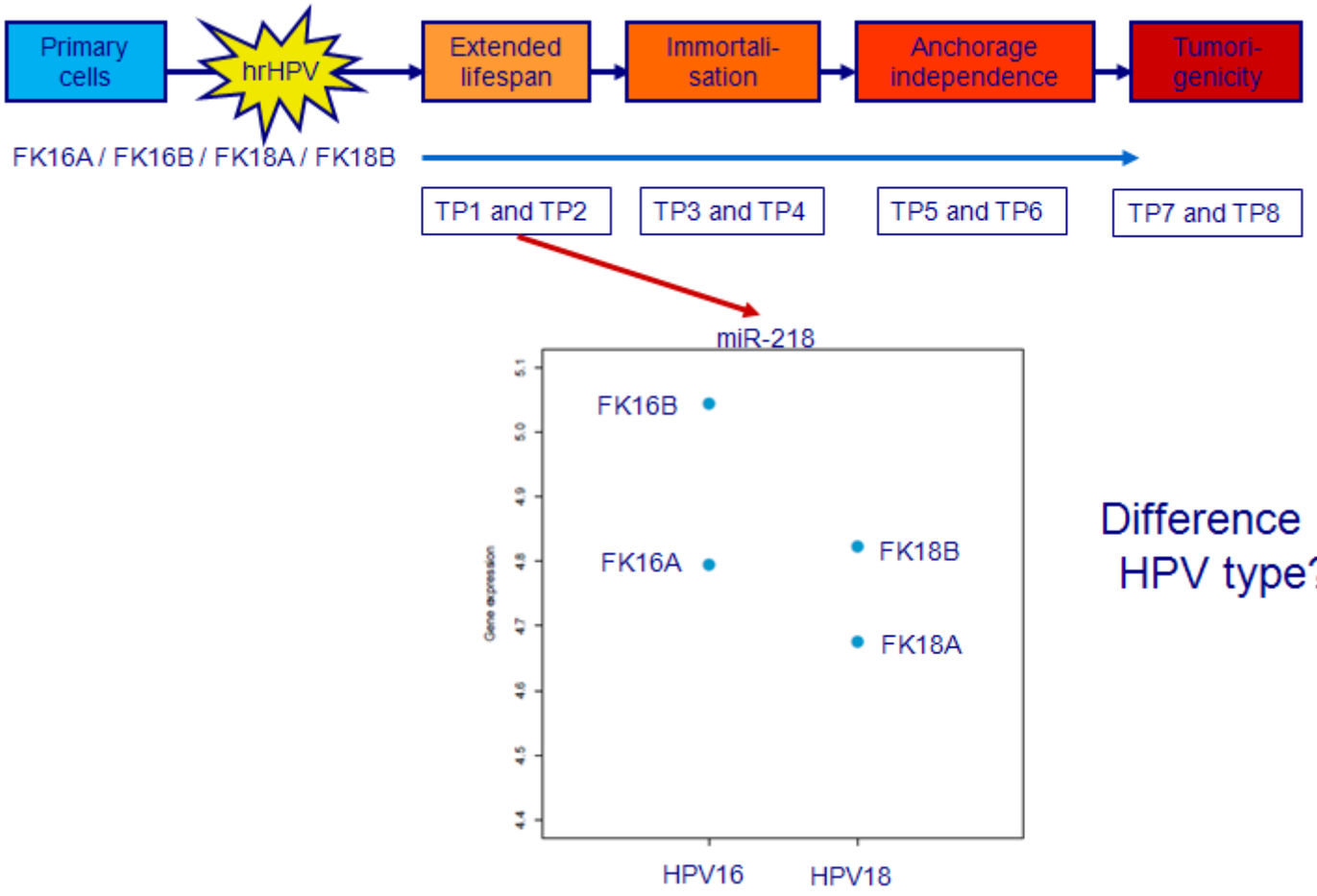
miR-218:

one moment in time: TP1



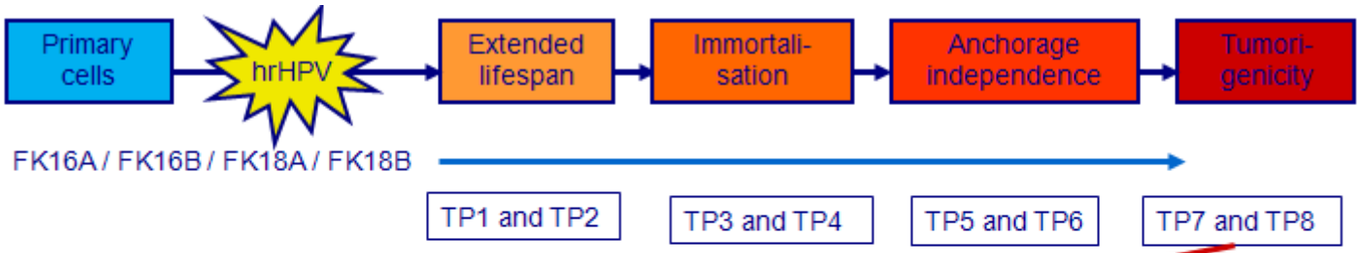
one moment in time: TP8

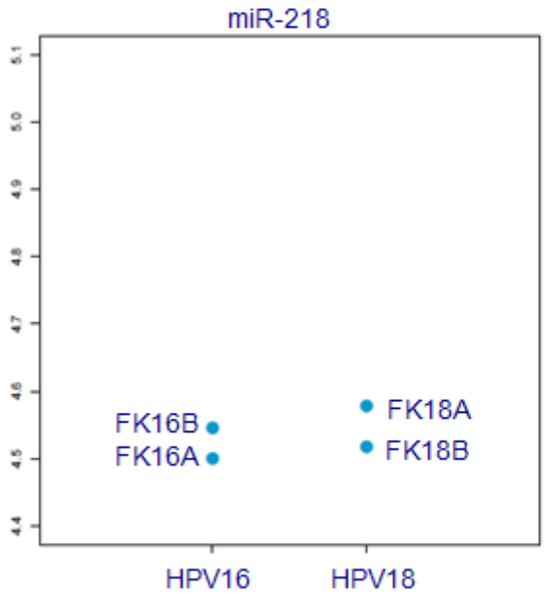
Inference based on TP1



Difference in HPV type?

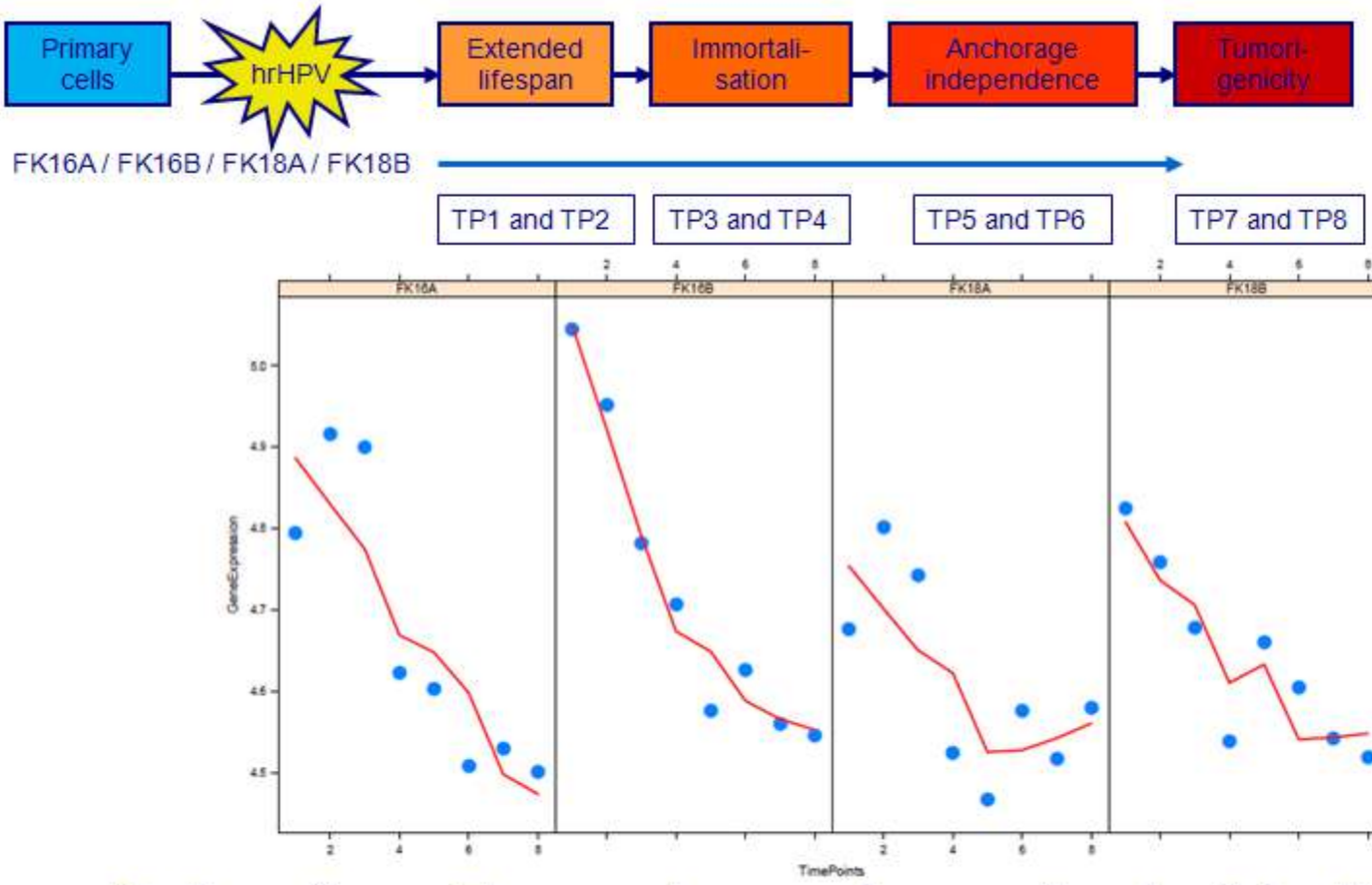
Inference based on TP8





No difference?

Strength of time-course



Similar pattern of decreased expression over time in all 4 cell lines

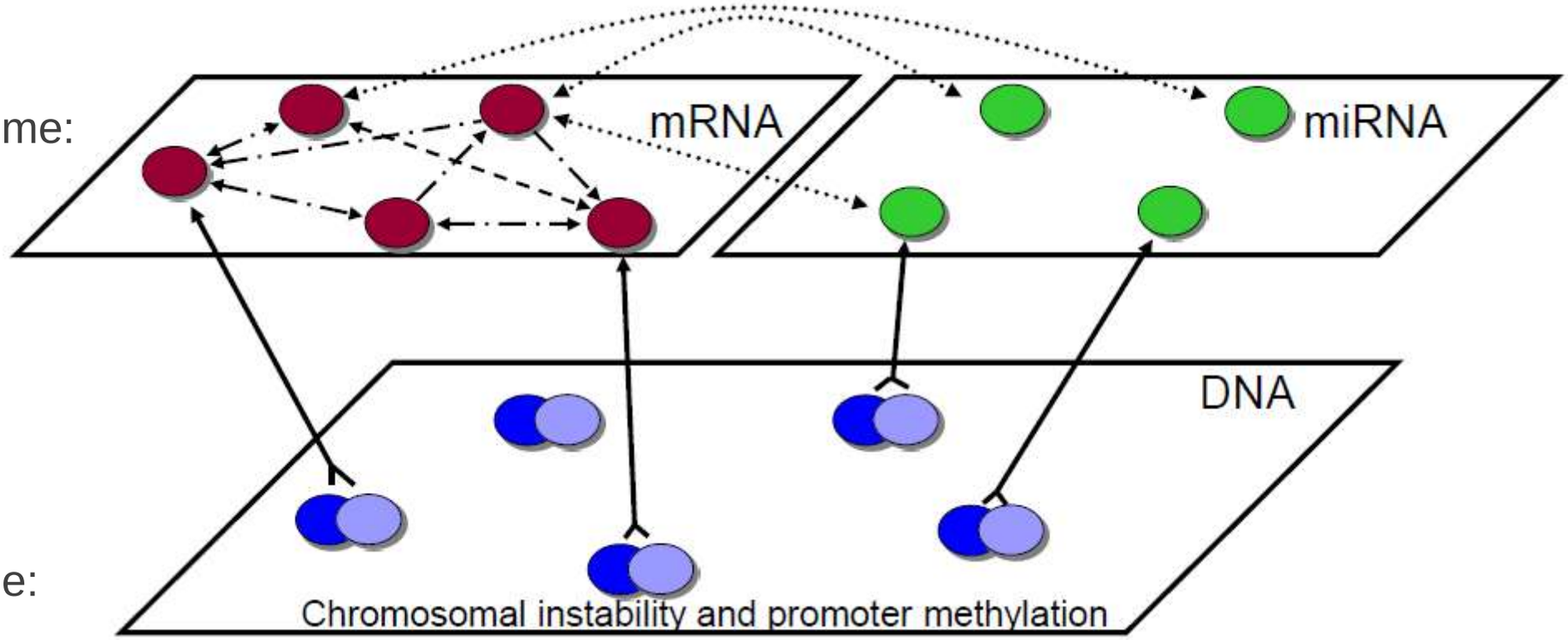
Why integration?

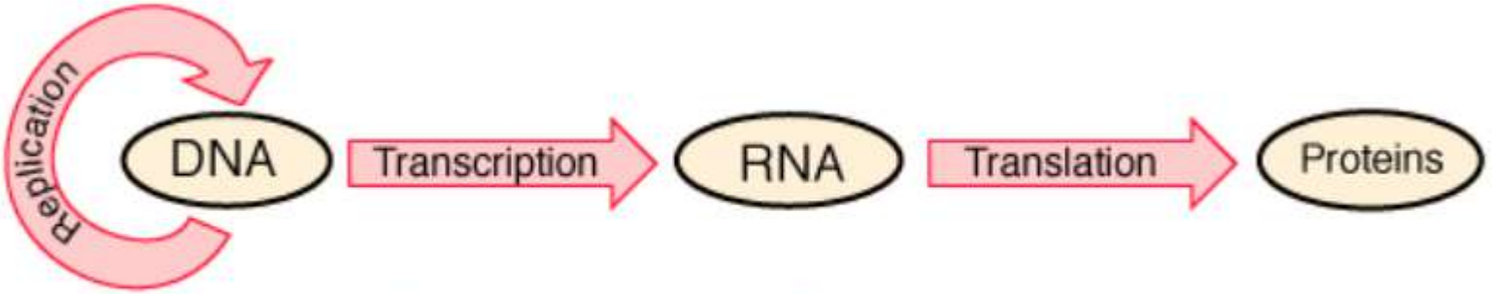
“Blind monks examining an elephant” by Itcho Hanabusa 1888

Multi-omics data integration

Transcriptome:

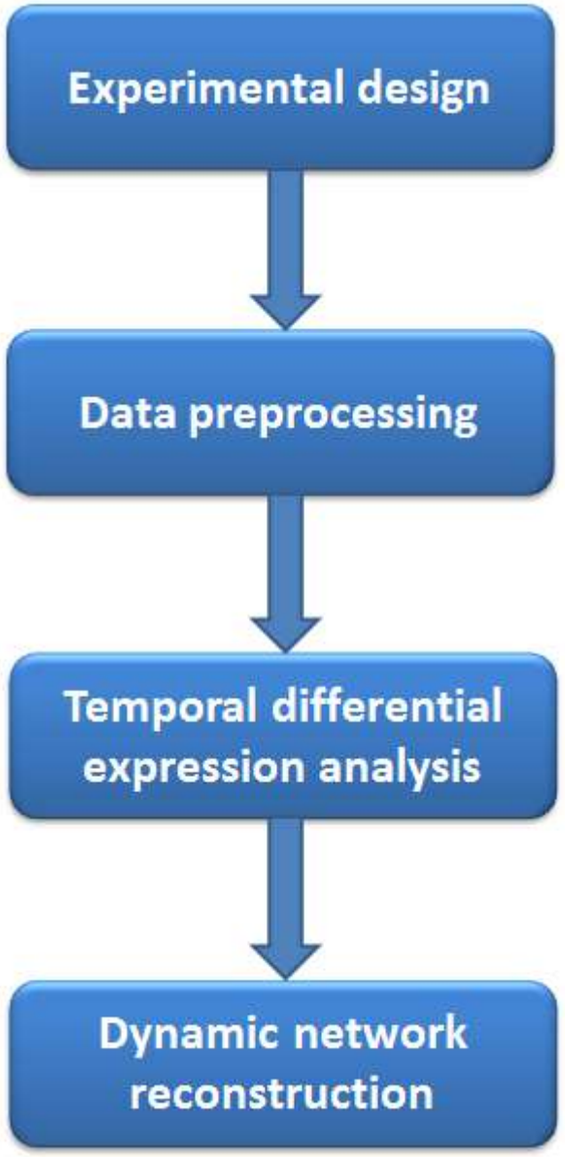
Genome:





Central dogma of molecular biology

What we did?



mRNA: 45K probes arrays
miRNA: 60K probes arrays
CN: 180K probes arrays

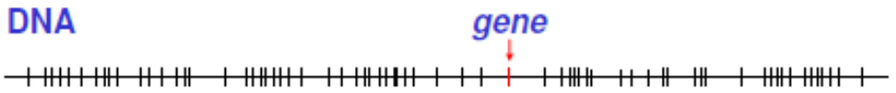
mRNA: 27637 genes
miRNA: 1187 genes
CN: 27637 genes

mRNA: 3642 genes
miRNA: 106 genes

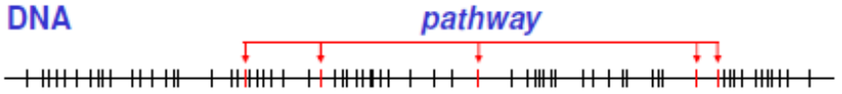
mRNA: 64 genes linked to p53 signaling pathway
miRNA: 106 genes which target mRNA

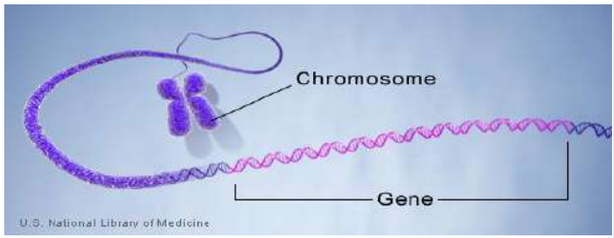
Statistical unit

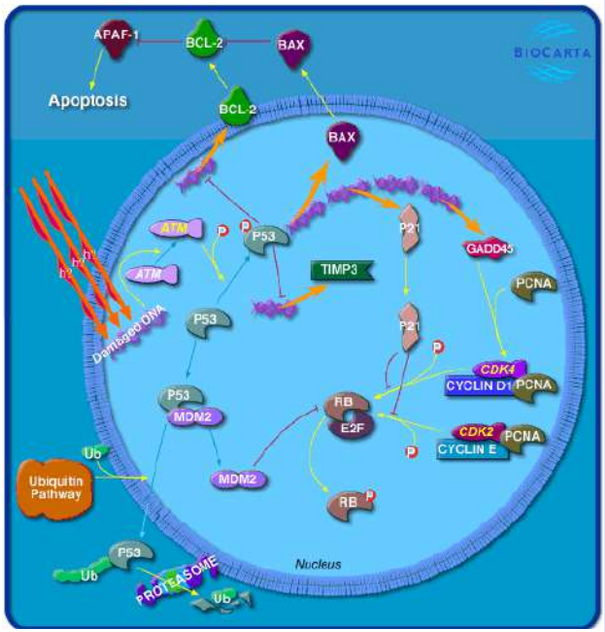
- Gene – measured part of the genome



- Pathway – group of genes which work together

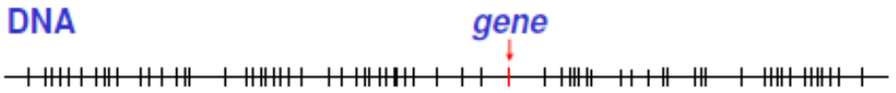


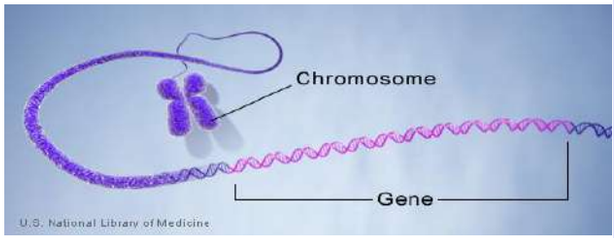




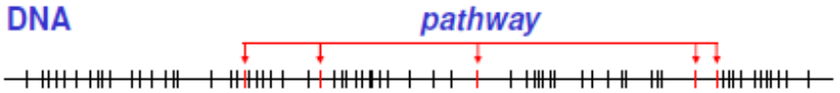
Statistical unit

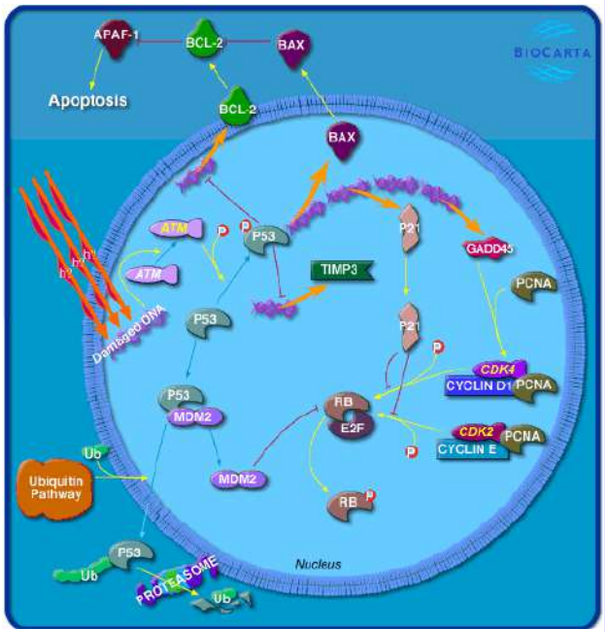
- Gene – measured part of the genome





- Pathway – group of genes which work together





Approach	Statistical Unit
- Restrict dimension model - Test model across genome - Employ familywise error control	Individual features
- Employ regularization - Enabling estimation and inference when $p > n$	Pathways

Temporal differential expression analysis

Model

$j, j = 1, ..., p$ - genes

$\mathbf{Y}_{*,*,t} = (\mathbf{Y}_{1,*,t}, ..., \mathbf{Y}_{n,*,t})$ - mRNA gene expression

Bayesian GLMM: $Y_{i,j,t} \sim \mathcal{N}(\mu_{i,j,t}, \sigma^2_{\varepsilon,j})$

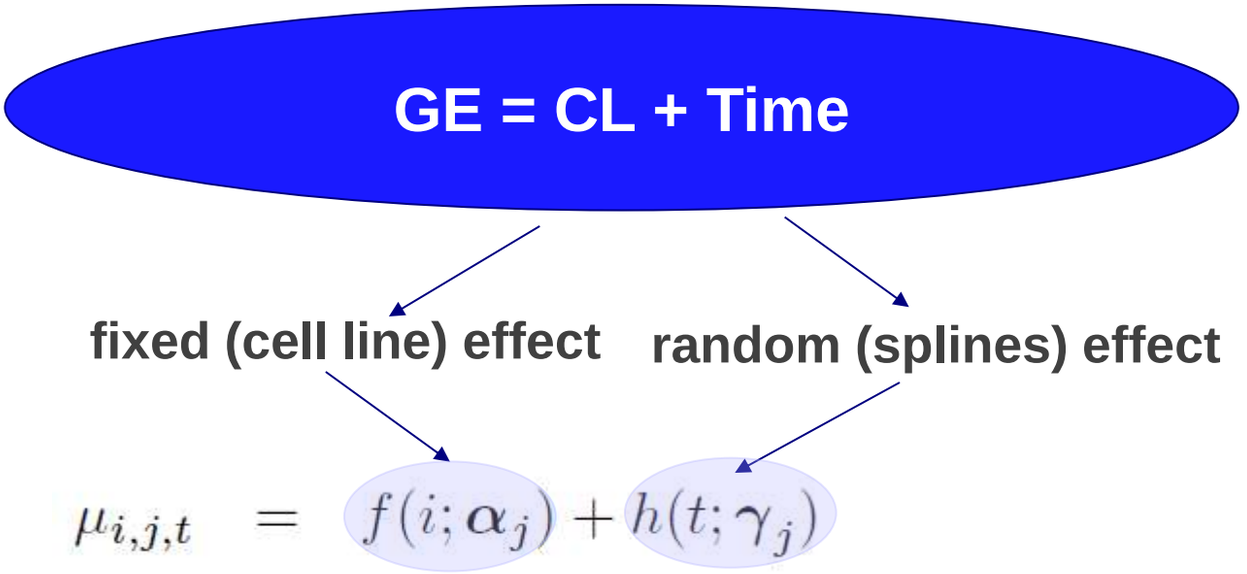
GE = CL + Time

Model

$j, j = 1, ..., p$ - genes

$\mathbf{Y}_{*,*,t} = (\mathbf{Y}_{1,*,t}, ..., \mathbf{Y}_{n,*,t})$ - mRNA gene expression

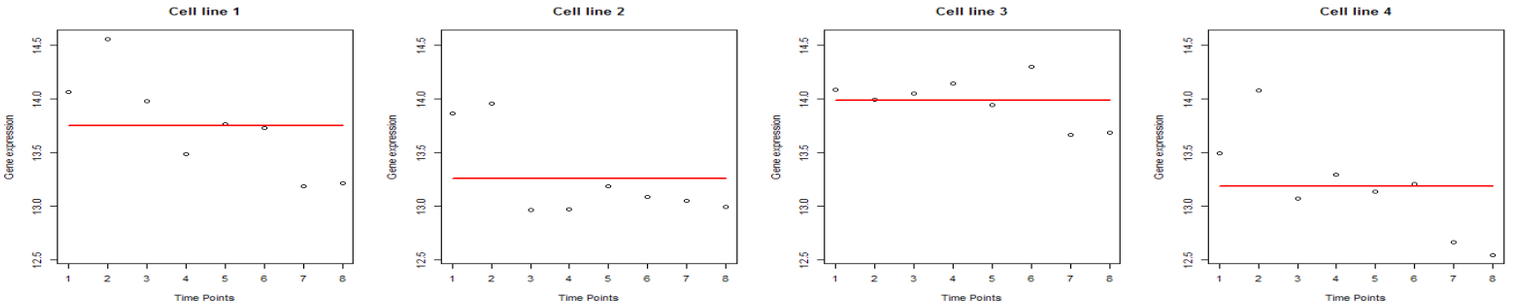
Bayesian GLMM: $Y_{i,j,t} \sim \mathcal{N}(\mu_{i,j,t}, \sigma^2_{\varepsilon,j})$



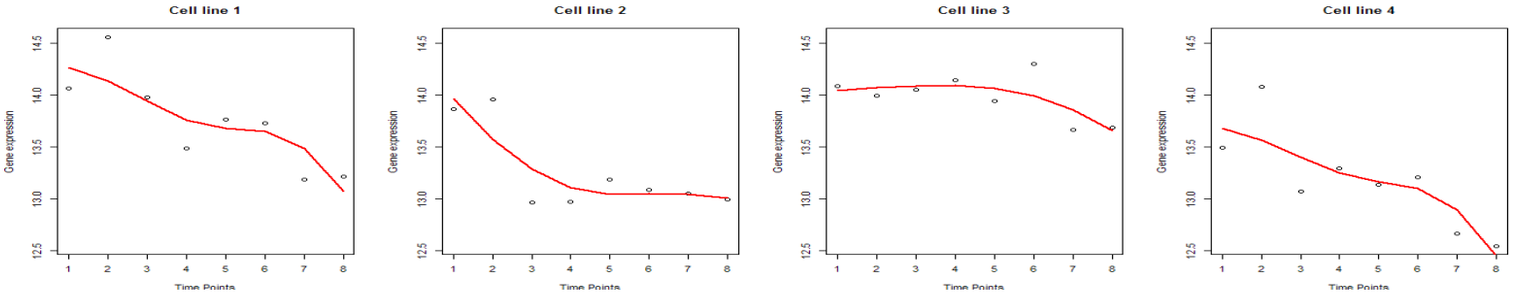
α, γ - Gaussian distribution assumption

Fixed and random effects

Cell line effect:



Cell line + time effect:

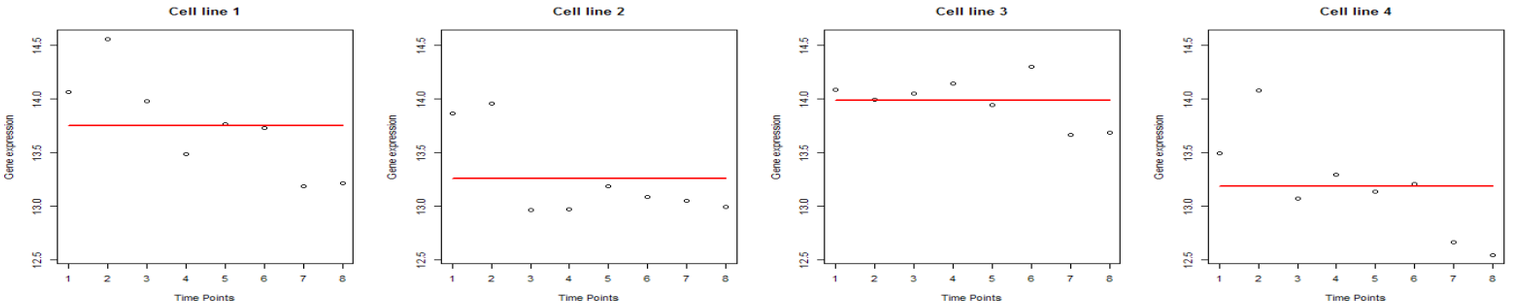


Fixed and random effects

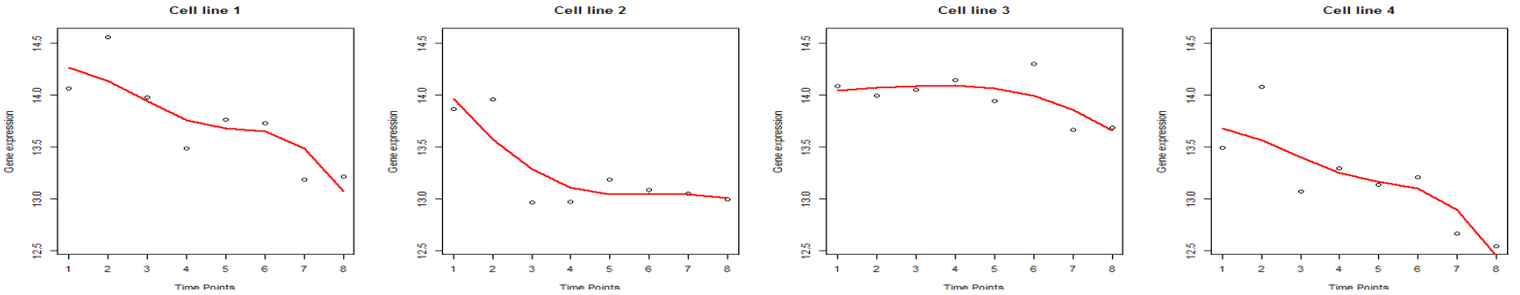
Fixed effect: cell line $f(i;\boldsymbol{\alpha_j}) = \boldsymbol{\alpha}_{i,j}$

Random effect: time $h(t;\boldsymbol{\gamma}_j) = \sum_{k=1}^{K} \gamma_{j,k} |t - \kappa_k|^3$

Cell line effect:



Cell line + time effect:

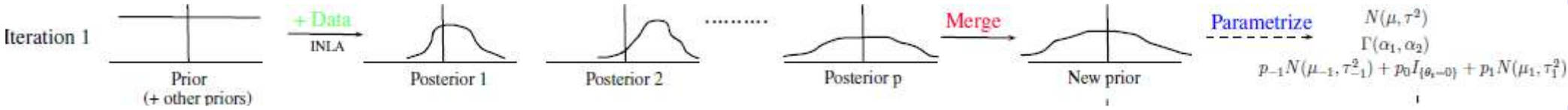


Matrix notation: $Y_{i,j,t} = \alpha_{i,j} + \tilde{\mathbf{Z}}_t \tilde{\boldsymbol{\gamma}}_j + \varepsilon_{i,j,t}$

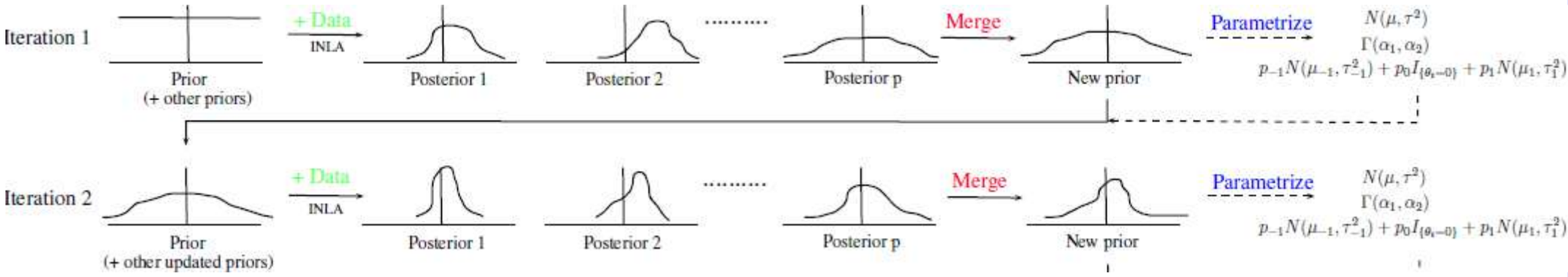
Spline basis: $\mathbf{Z}_t = (|t - \kappa_1|^3, \ldots, |t - \kappa_K|^3)$

Spline coefficients: $\boldsymbol{\gamma}_j = (\gamma_j, \ldots, \gamma_{j,K})^{\mathrm{T}}$

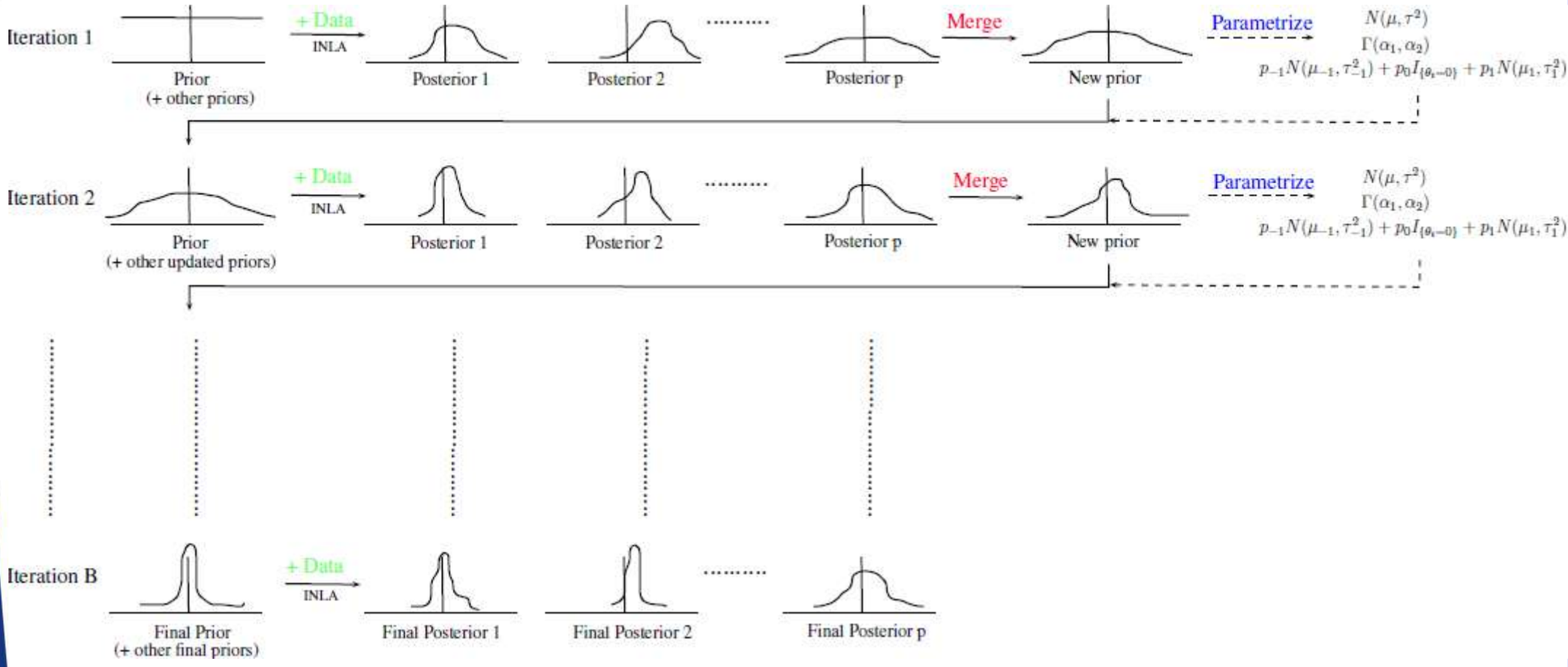
Model parameters estimation



Model parameters estimation



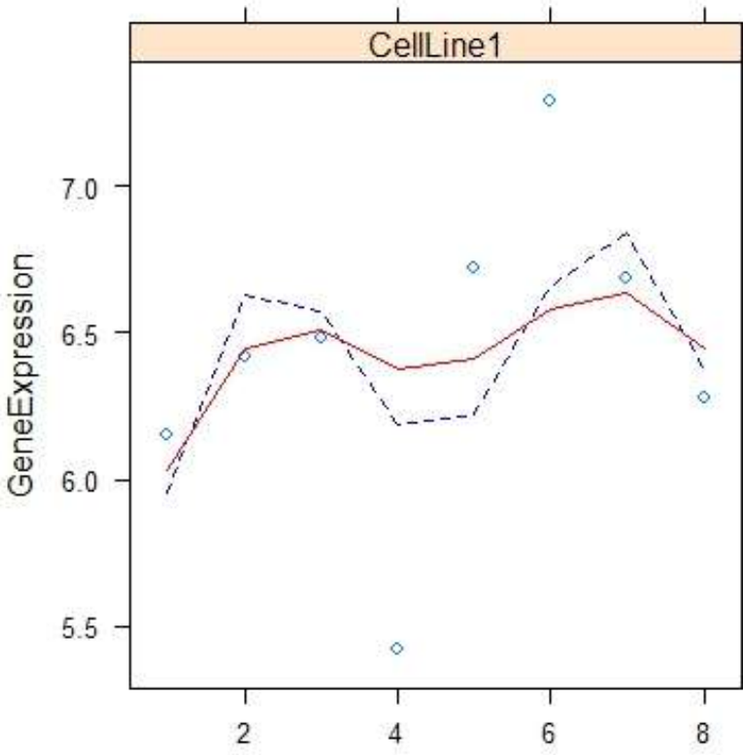
Model parameters estimation



INLA: Marginal posterior are estimated using integrated nested Laplace approximation

Shrinkage

- borrowing information across the genes
- better control of false positives
- improvement of reproducibility
- leads to more stable estimates



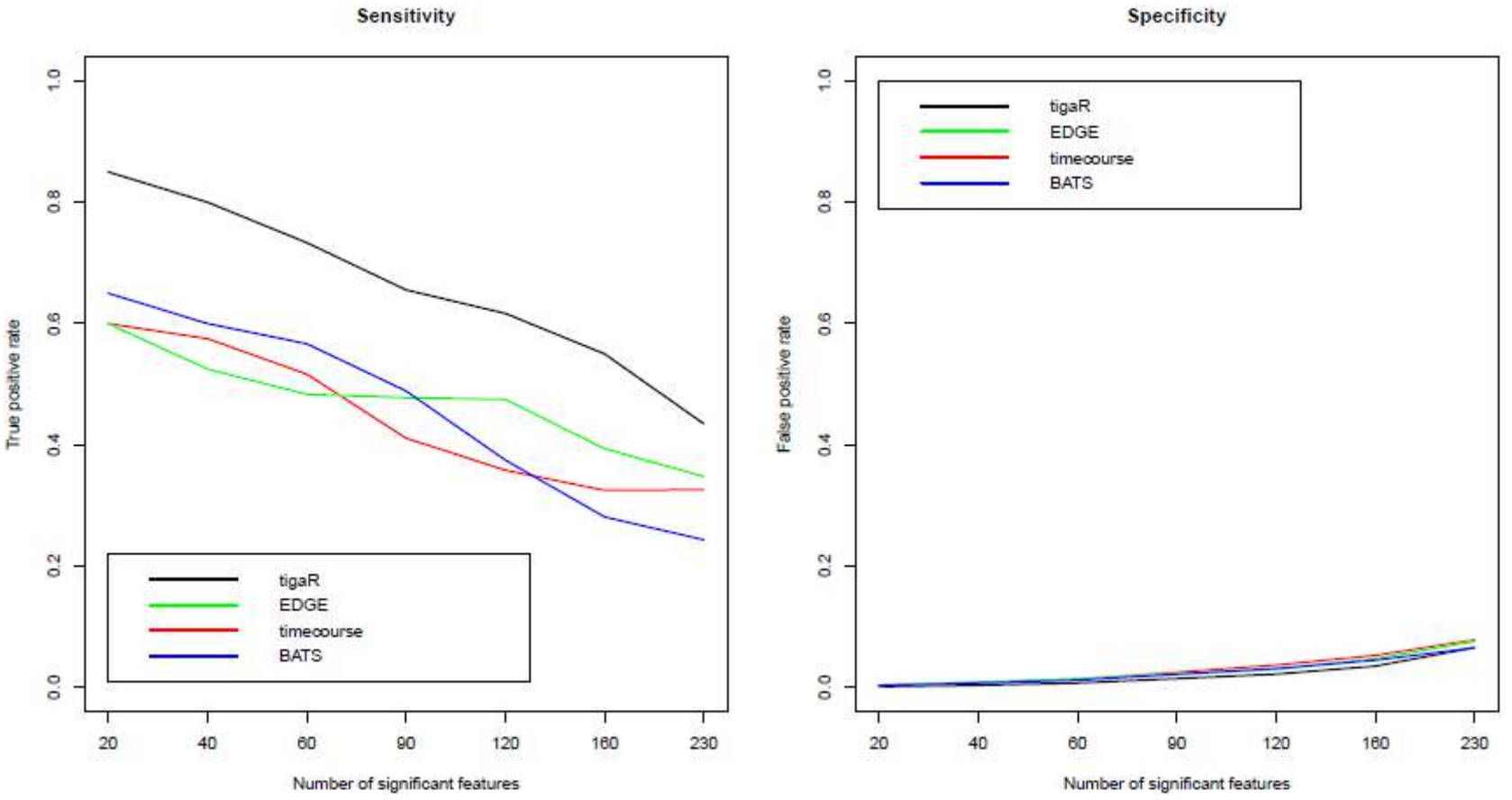
Comparison

- Comparison of following methods:
  - **timecourse** – Tai and Speed, Annals of Statistics, 2006.
  - **EDGE** – Storey et al., PNAS, 2005.
  - **BATS** – Angelini et al., BMC Bioinformatics, 2008.
  - **tigaR** – Miok et al., BMC Bioinformatics, 2014.

- Method is applied on two data sets
  - Data from our experiment (only mRNA data)
  - Data from Storey et al., PNAS, 2005.

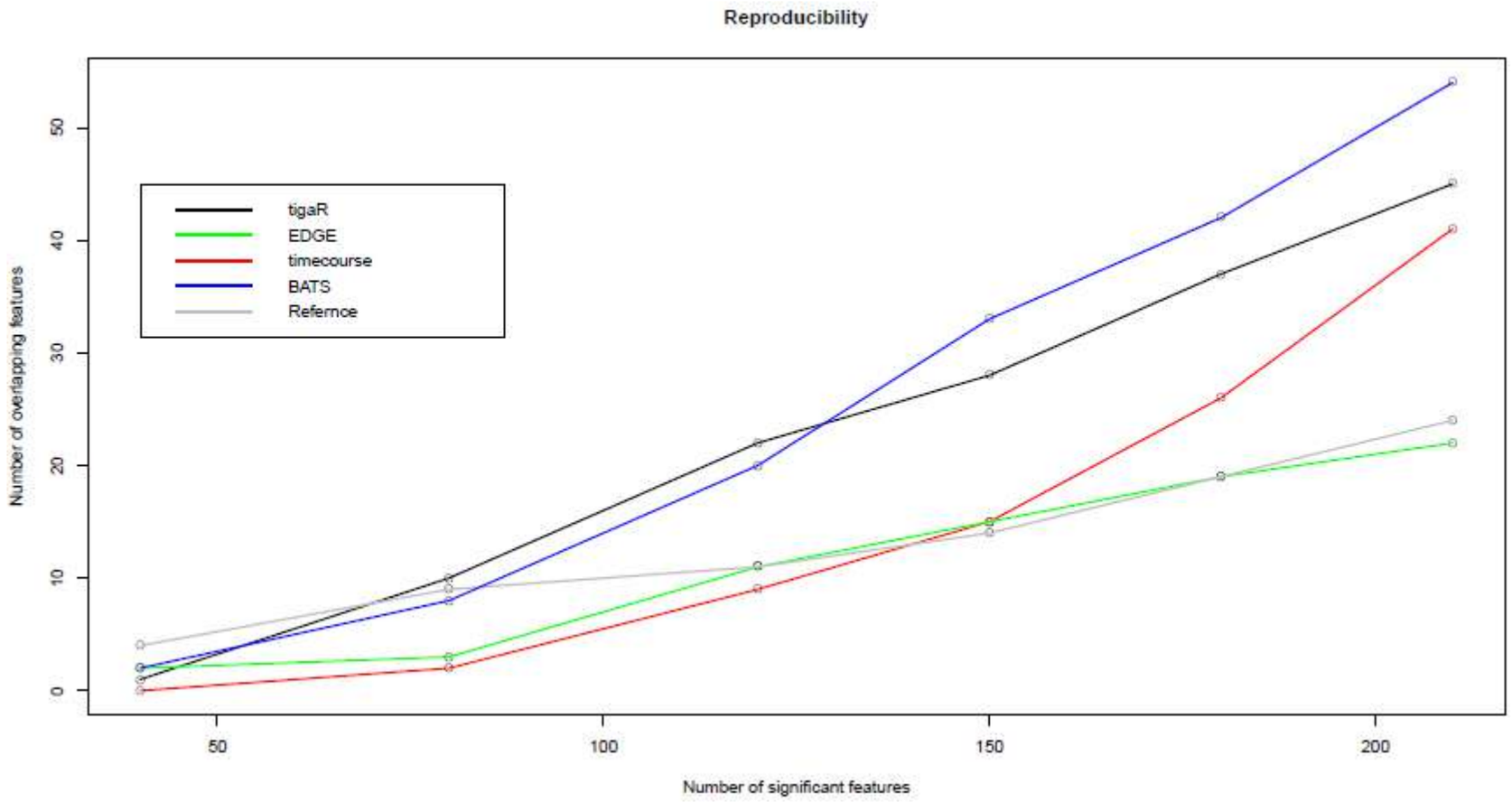
Sensitivity and specificity

- Truth – overlap of significant genes among methods.



Reproducibility

- Equally divided data set in two groups.



DNA copy number (CN)

GE = CL + CN + Time

DNA copy number (CN)

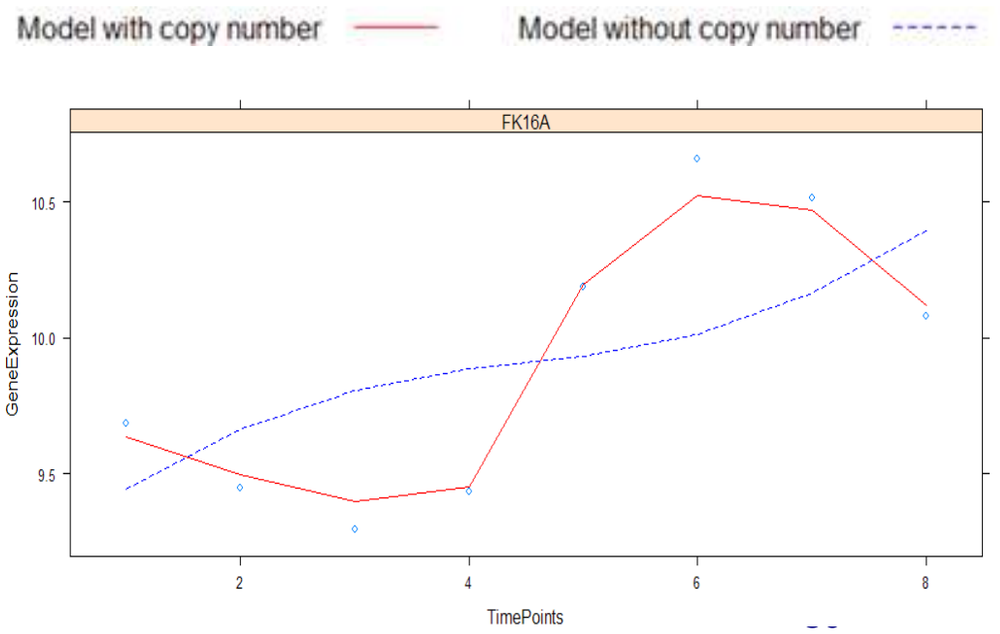
GE = CL + CN + Time

$$\mathbf{X}_{*,*,t} = (\mathbf{X}_{1,*,t}, \ldots, \mathbf{X}_{n,*,t})$$ - CN

Cell line CN Time Error

$$Y_{i,j,t} = \alpha_{i,j} + \beta_j x_{i,j,t} + \tilde{\mathbf{Z}}_t \tilde{\boldsymbol{\gamma}}_j + \varepsilon_{i,j,t}$$

Gene GSTM3:



Hypothesis testing

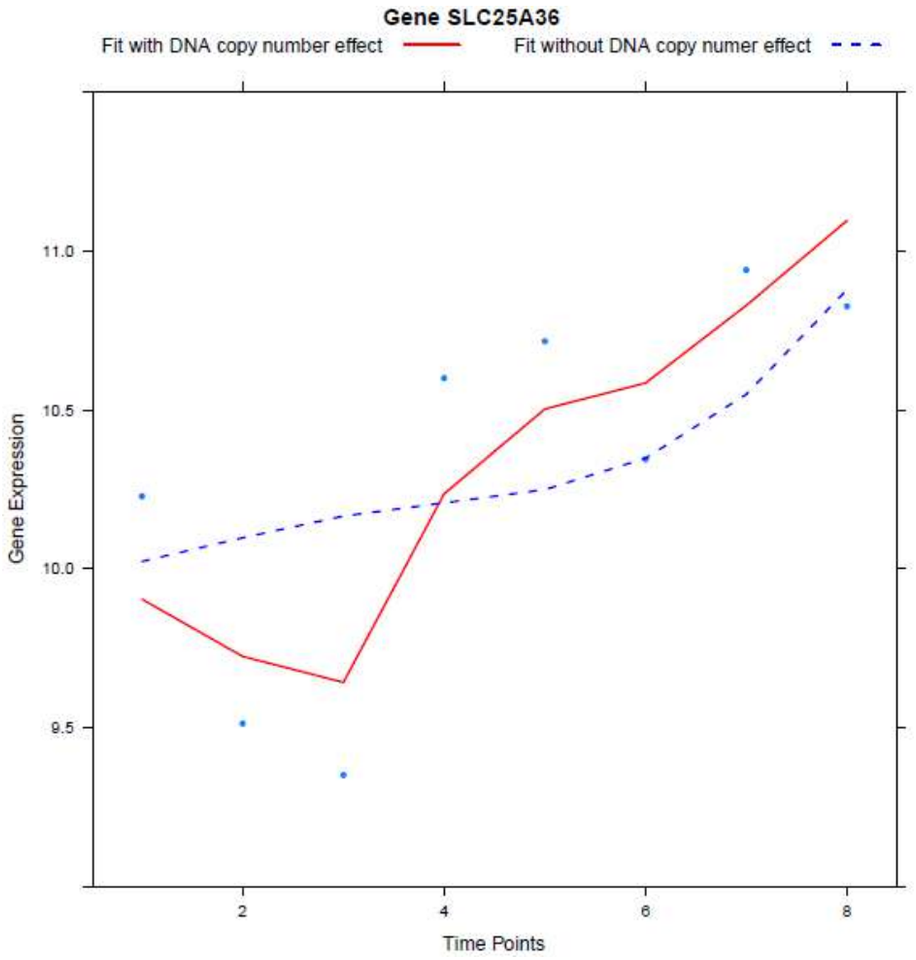
- Questions?
  - Is there differential expression over time?
  - Does DNA copy number drive gene expression?
  - Is there a difference between the cell lines?

- Hypothesis are evaluated by means of the likelihood ratio statistics

$$D_j = \log\left[L\left(\underset{j}{\widehat{\boldsymbol{\alpha}}}^{(H_A)}, \widehat{\beta}_j^{(H_A)}, \widehat{\sigma}^{2,(H_A)}_{\gamma,j}, \widehat{\sigma}^{2,(H_A)}_{\varepsilon,j}\right)\right] - \log\left[L\left(\underset{j}{\widehat{\boldsymbol{\alpha}}}^{(H_0)}, 0, \widehat{\sigma}^{2,(H_0)}_{\gamma,j}, \widehat{\sigma}^{2,(H_0)}_{\varepsilon,j}\right)\right]$$

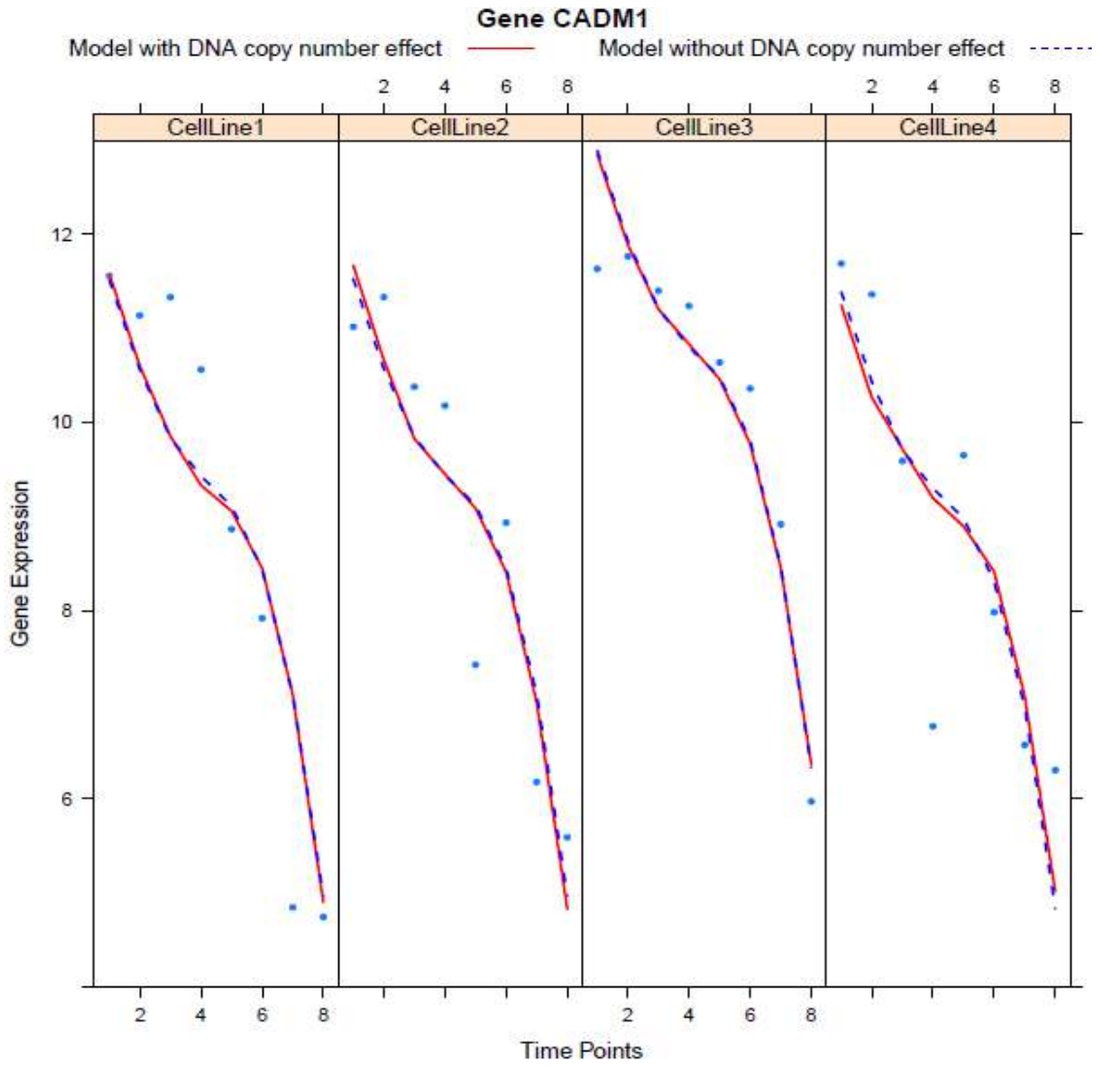
- To account for multiplicity the False Discovery Rate is controlled

SLC25A36 – gene with CN effect



Wilting et al., Genes, Chromosomes and Cancer, 2008.

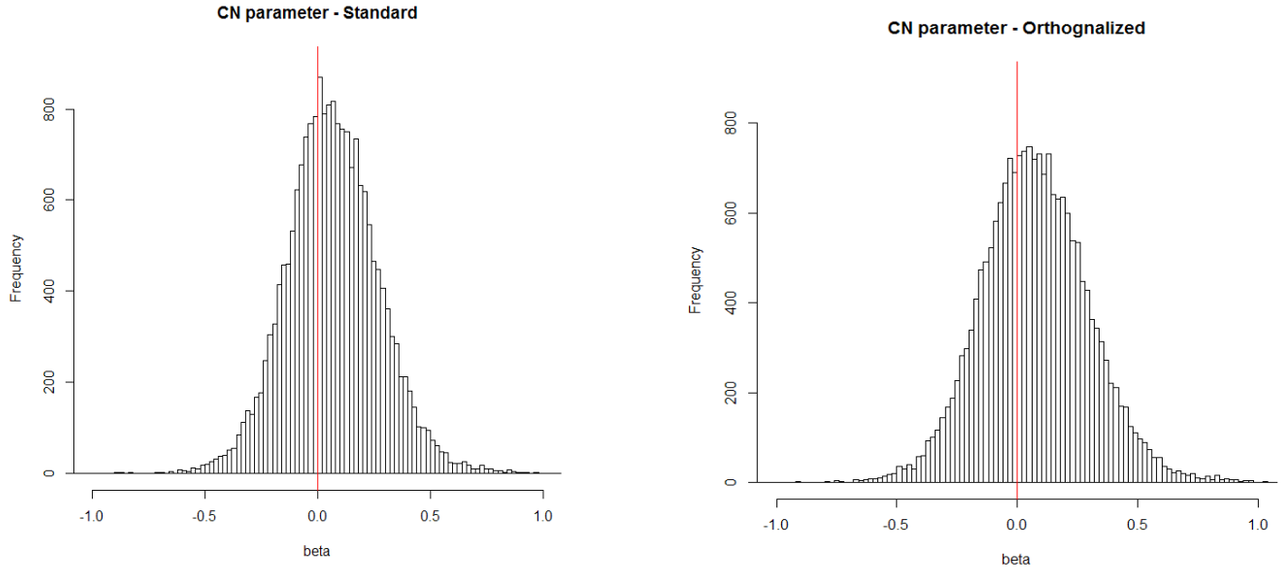
CADM1- gene without CN effect



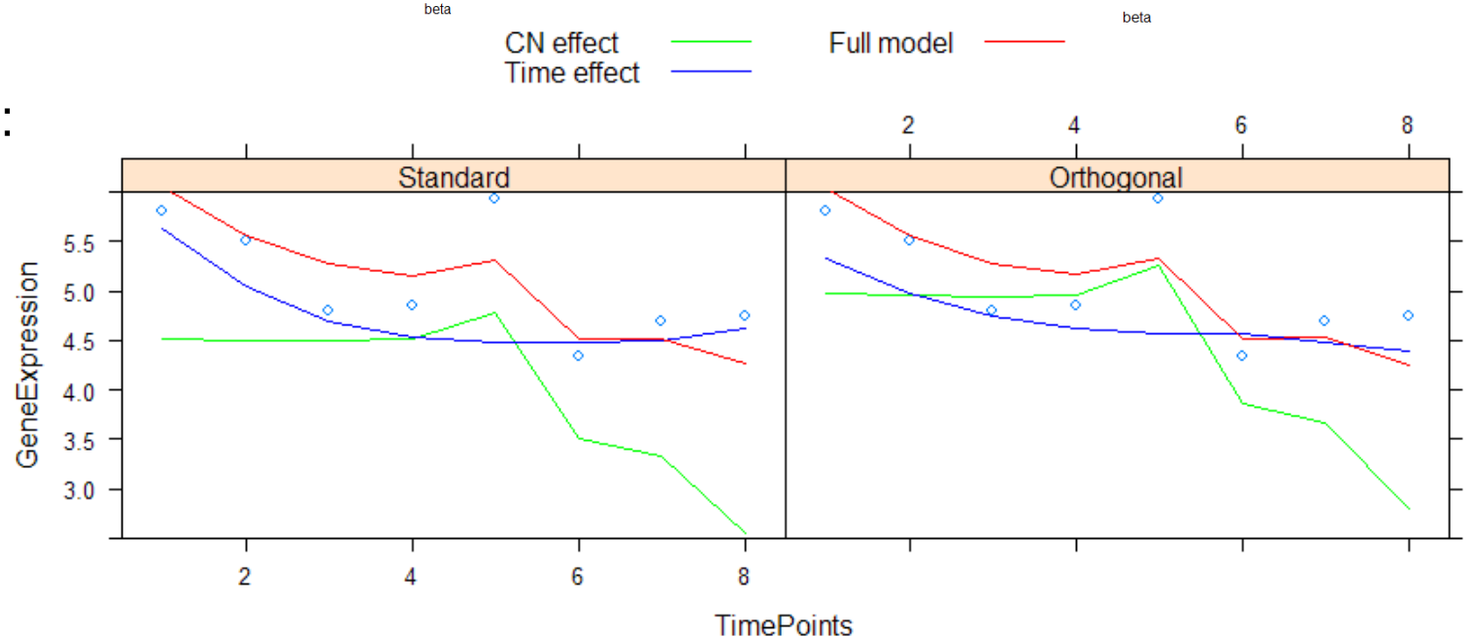
Overmeer et al., J Pathol., 2008.

Orthogonalization splines onto CN

CN parameter standard vs. orthogonal:



Fit of the model standard vs. orthogonal:

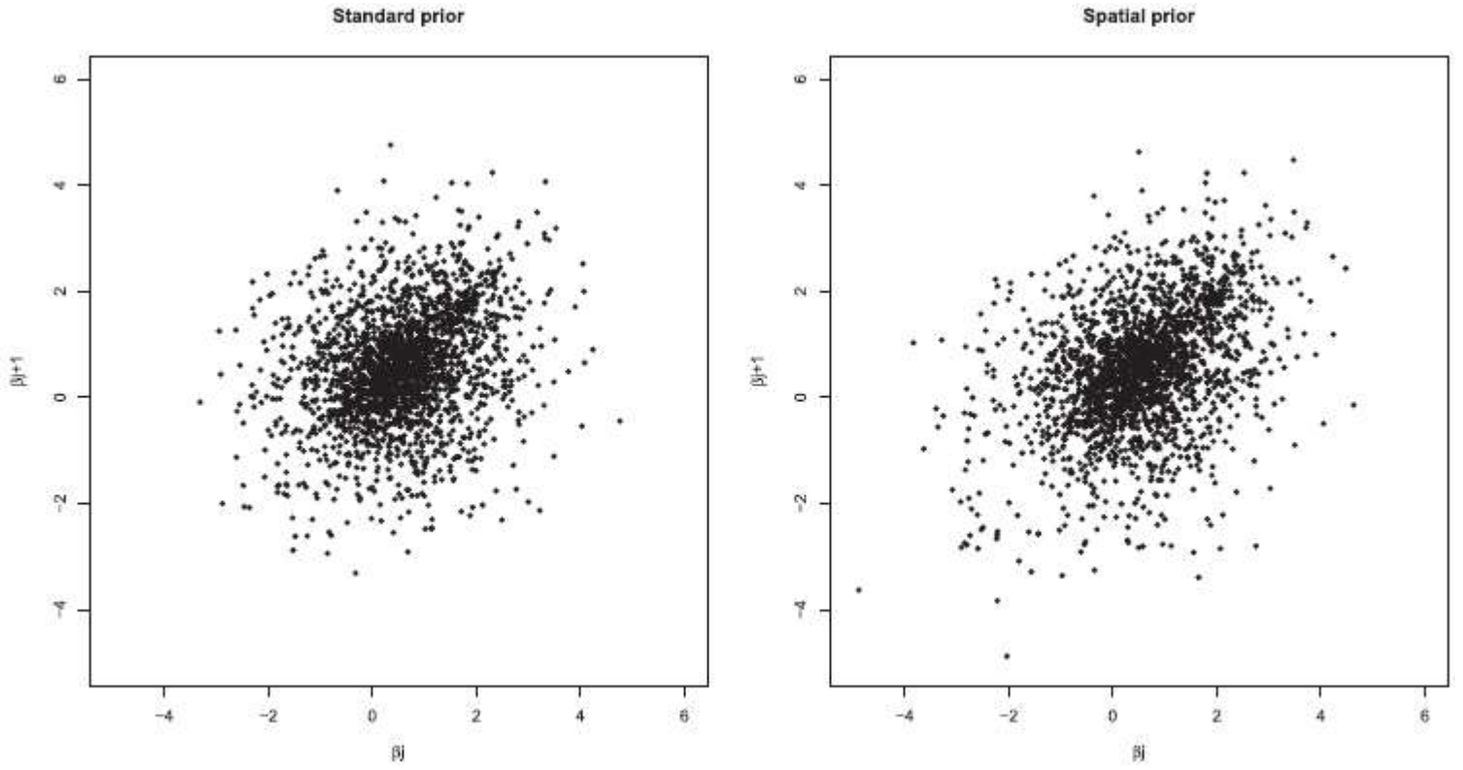


Spatial multivariate prior for CN

Multivariate prior:

$$\begin{pmatrix} \beta_{j-1} \\ \beta_j \\ \beta_{j+1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{j-1}^2 & \sigma_{j-1}\sigma_j\rho & \sigma_{j-1}\sigma_{j+1}\rho^2 \\ \sigma_{j-1}\sigma_j\rho & \sigma_j^2 & \sigma_j\sigma_{j+1}\rho \\ \sigma_{j-1}\sigma_{j+1}\rho^2 & \sigma_j\sigma_{j+1}\rho & \sigma_{j+1}^2 \end{pmatrix} \right)$$

Improvement in partial correlation of CN parameters:



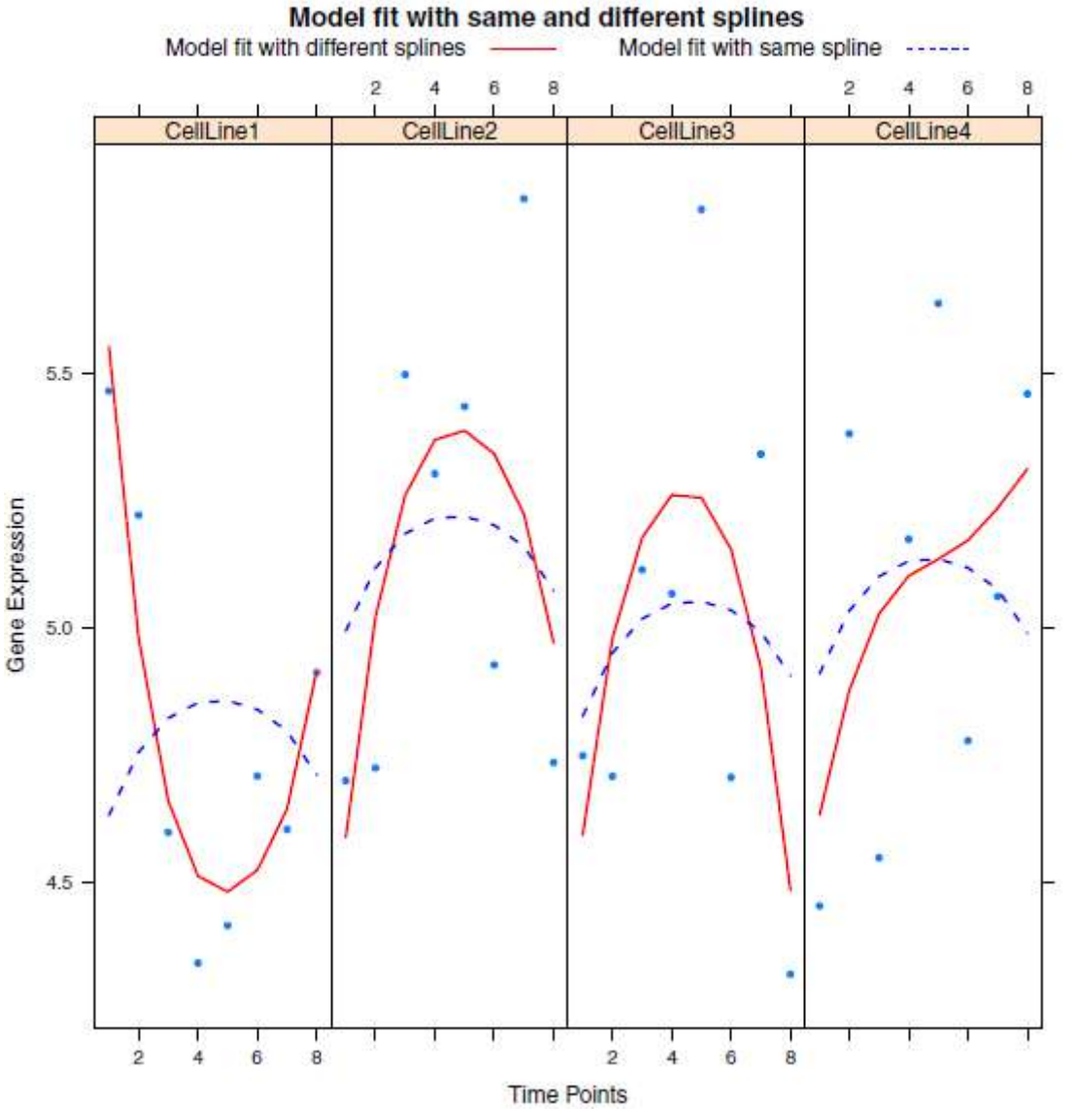
Splines flexibility

Same spline – up/down regulated genes

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{Z}} \otimes \mathbf{1}_{n \times n}$$

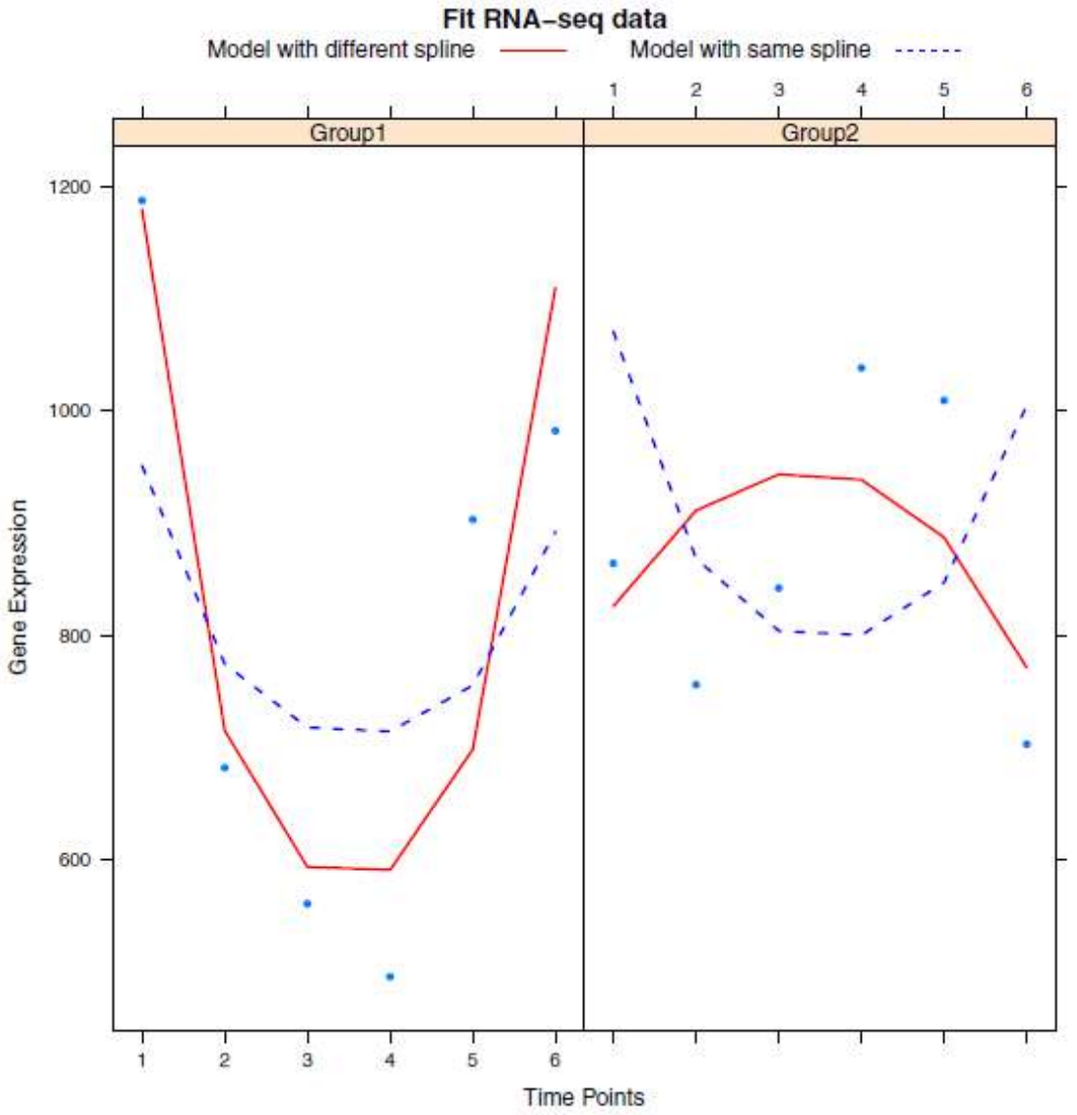
Different spline – allow more flexibility

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{Z}} \otimes \mathbf{I}_{n \times n}$$



RNA-seq data

- Changing link function method can deal with count data.
- Two group time-course RNA-seq data.



Article + R-package

Miok et al. BMC Bioinformatics 2014, 15:327
http://www.biomedcentral.com/1471-2105/15/327

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

tigaR: integrative significance analysis of temporal differential gene expression induced by genomic abnormalities

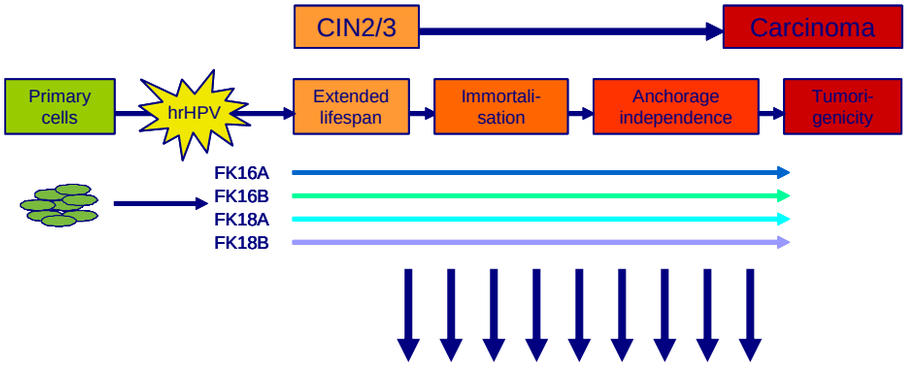
Viktorian Miok[1,2], Saskia M Wilting[2], Mark A van de Wiel[1,3], Annelieke Jaspers[2], Paula I van Noort[4], Ruud H Brakenhoff[5], Peter JF Snijders[2], Renske DM Steenbergen[2] and Wessel N van Wieringen[1,3*]

GitHub

tigaR: temporal integrative genomic analysis in R

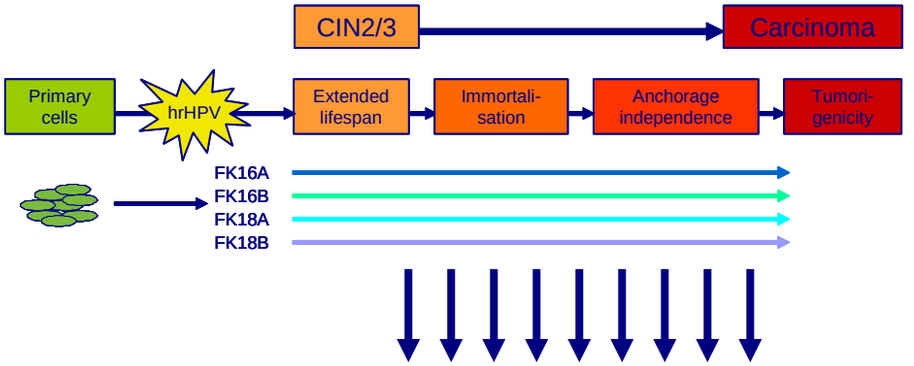
https://github.com/viktormiok/tigaR

tigaR analysis



Measured expression of 1187 miRNAs and 27637 mRNAs

tigaR analysis

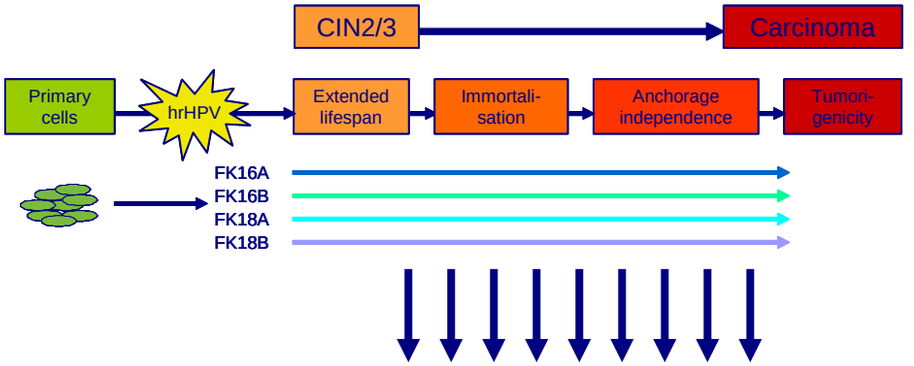


Measured expression of 1187 miRNAs and 27637 mRNAs

106 significant miRNAs and 3642 significant mRNAs

(concordant change in expression in at least 3 cell lines)

tigaR analysis



Measured expression of 1187 miRNAs and 27637 mRNAs

106 significant miRNAs and 3642 significant mRNAs

(concordant change in expression in at least 3 cell lines)

36 miRNAs and 1233 mRNAs linked with CN

(~34% of altered expression in both cases)

Thank you for your attention!