Longitudinal modeling of omics data from HPV-induced carcinogenesis

Viktorian Miok

Epidemiology & Biostatistics
VUmc
VU VRIJE UNIVERSITEIT AMSTERDAM

Contributors

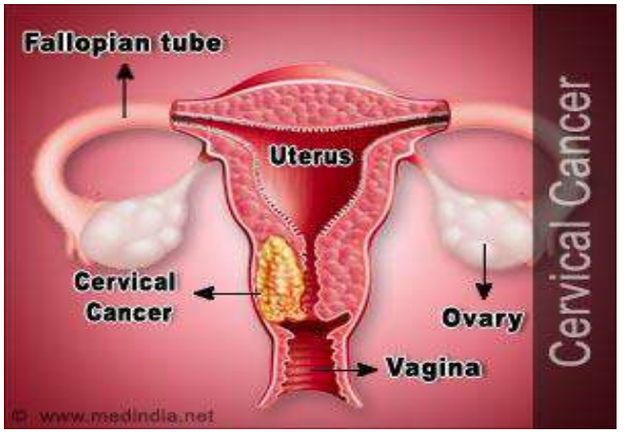
Biostatistics department

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

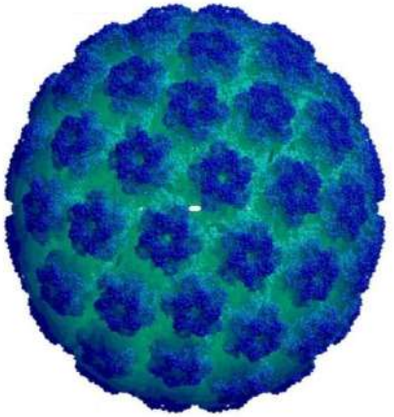
Pathology department

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

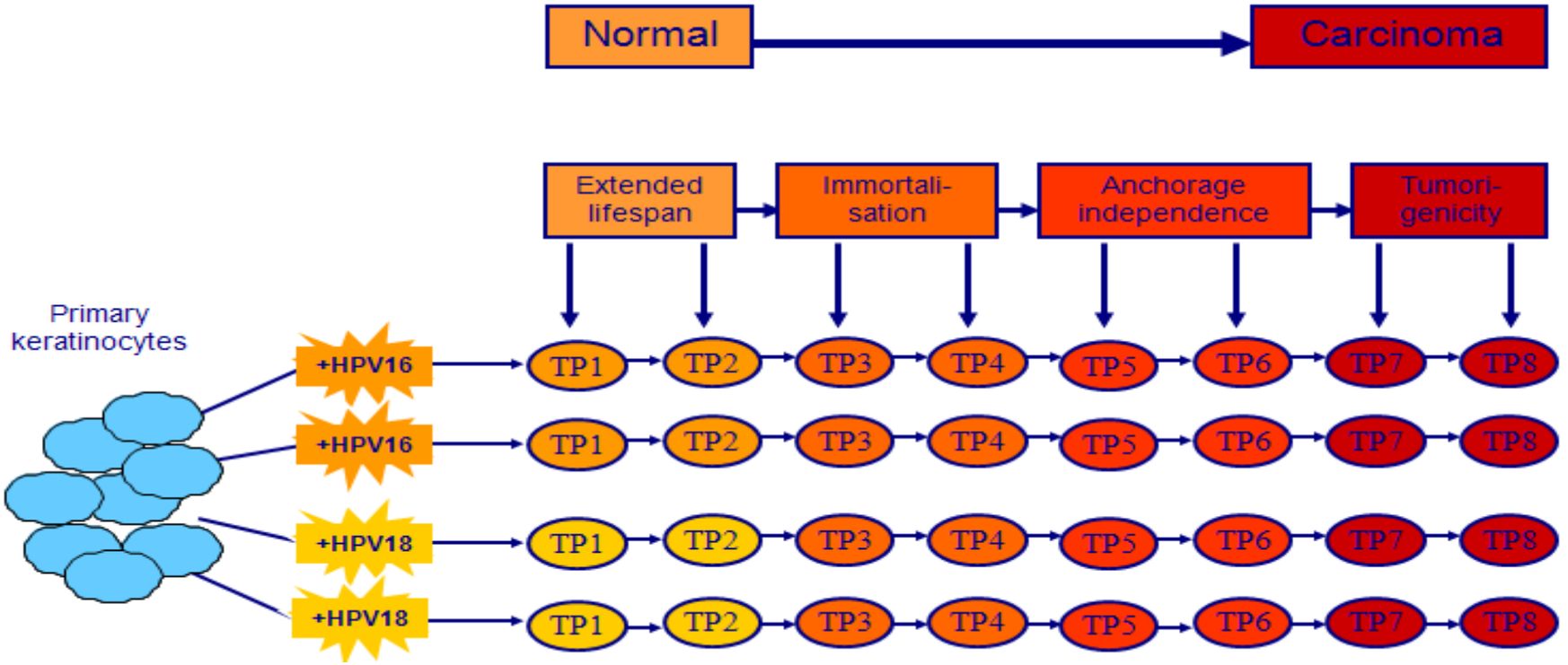
Cervical cancer study



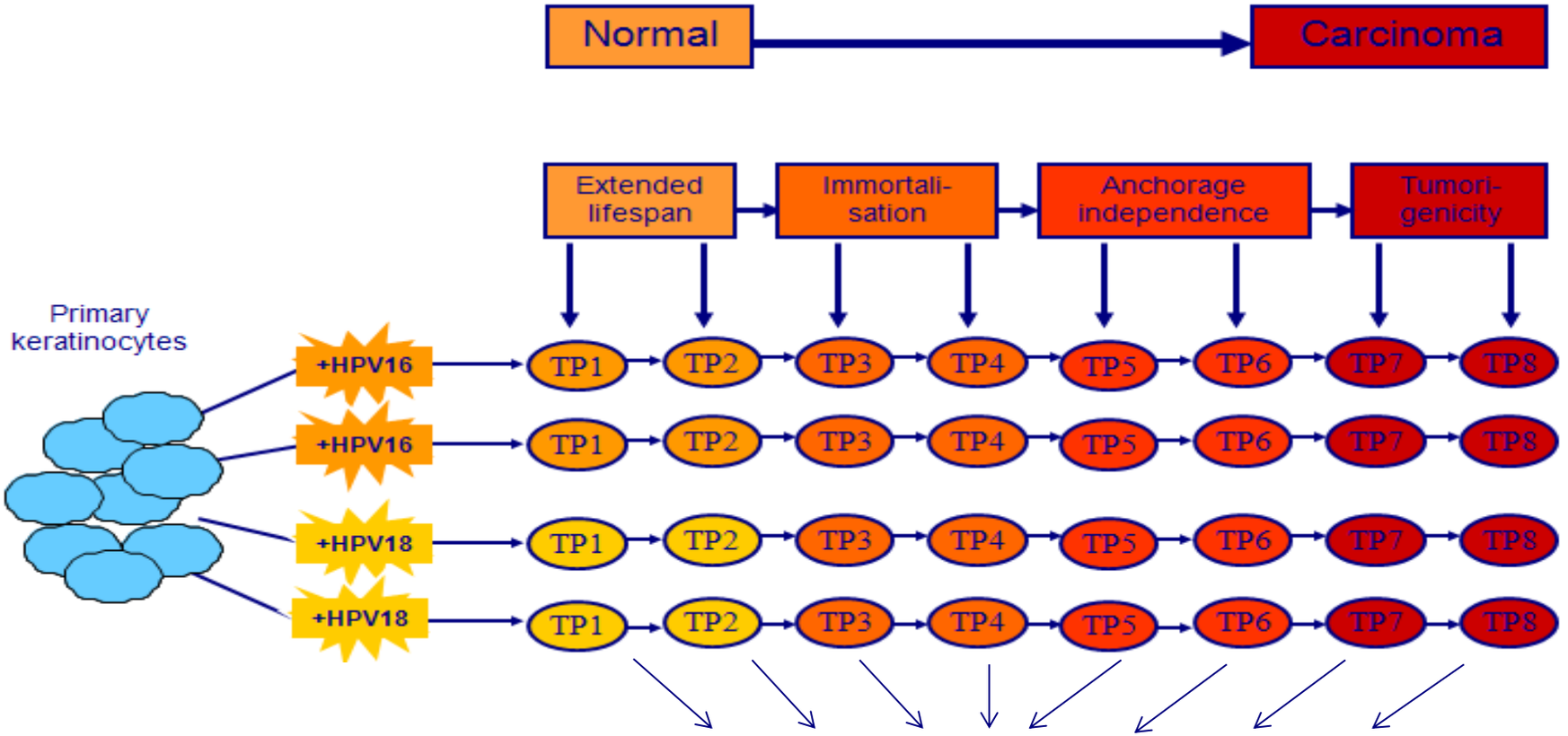
- Third most common cancer in women worldwide.
- Caused by HPV virus (70% cases HPV16 and HPV18) and followed by additional (epi)genetic abnormalities.
- Cell line model – in vitro model system of HPV-induced transformation.
- Integration – high-throughput multi level molecular data sets.
- Understand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



Time-course experiment



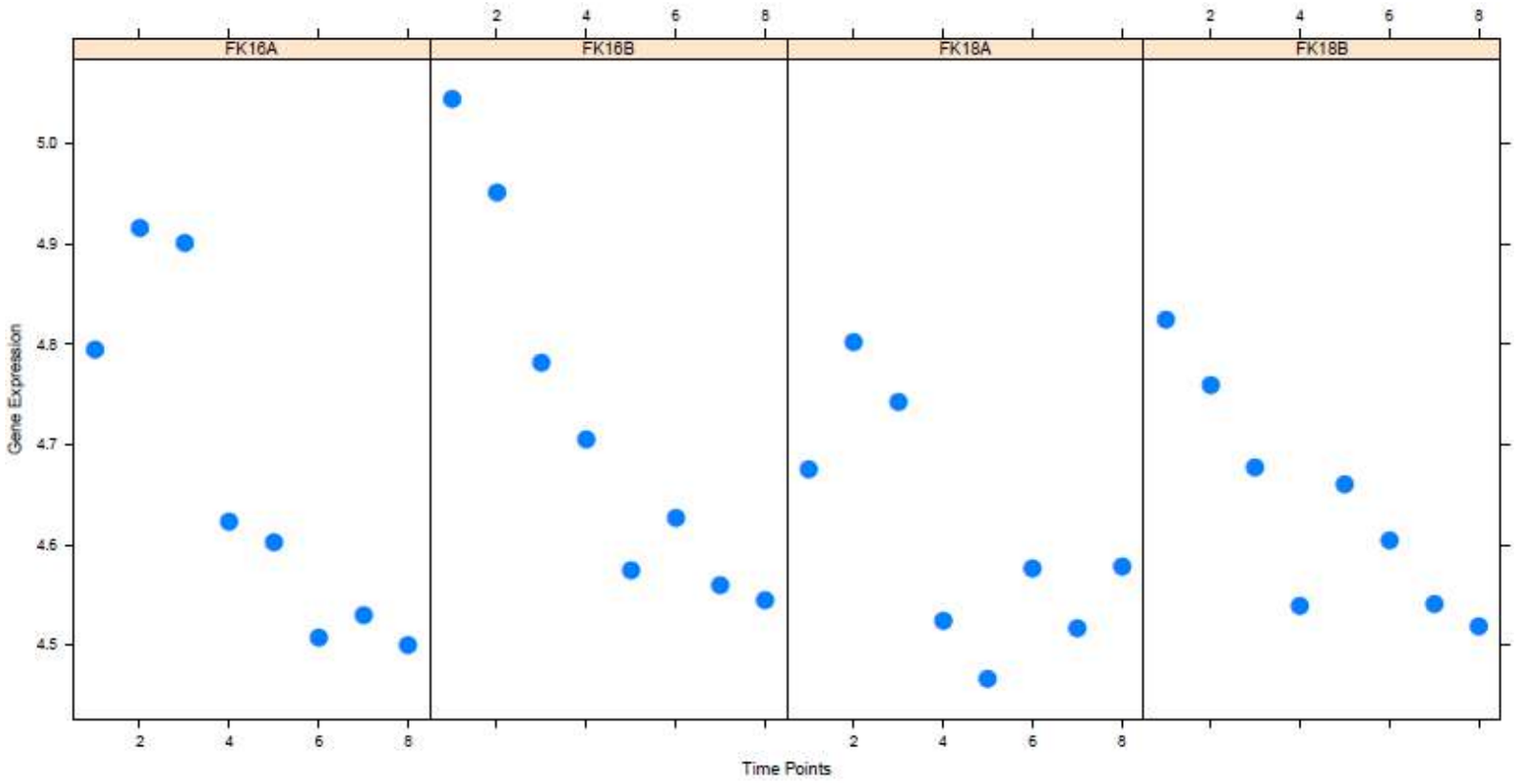
Time-course experiment



	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?

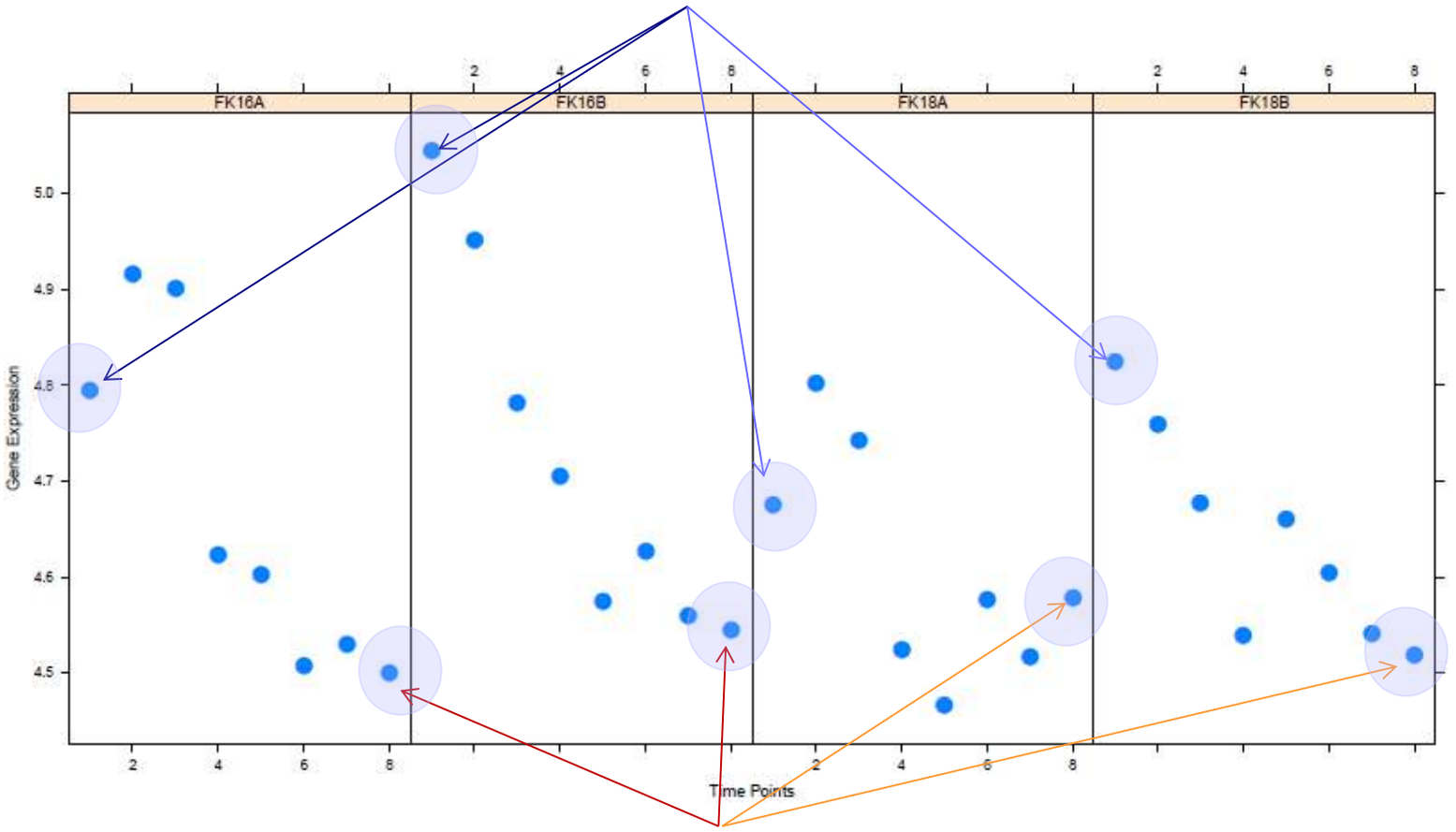
miR-218:



Pick one moment in time

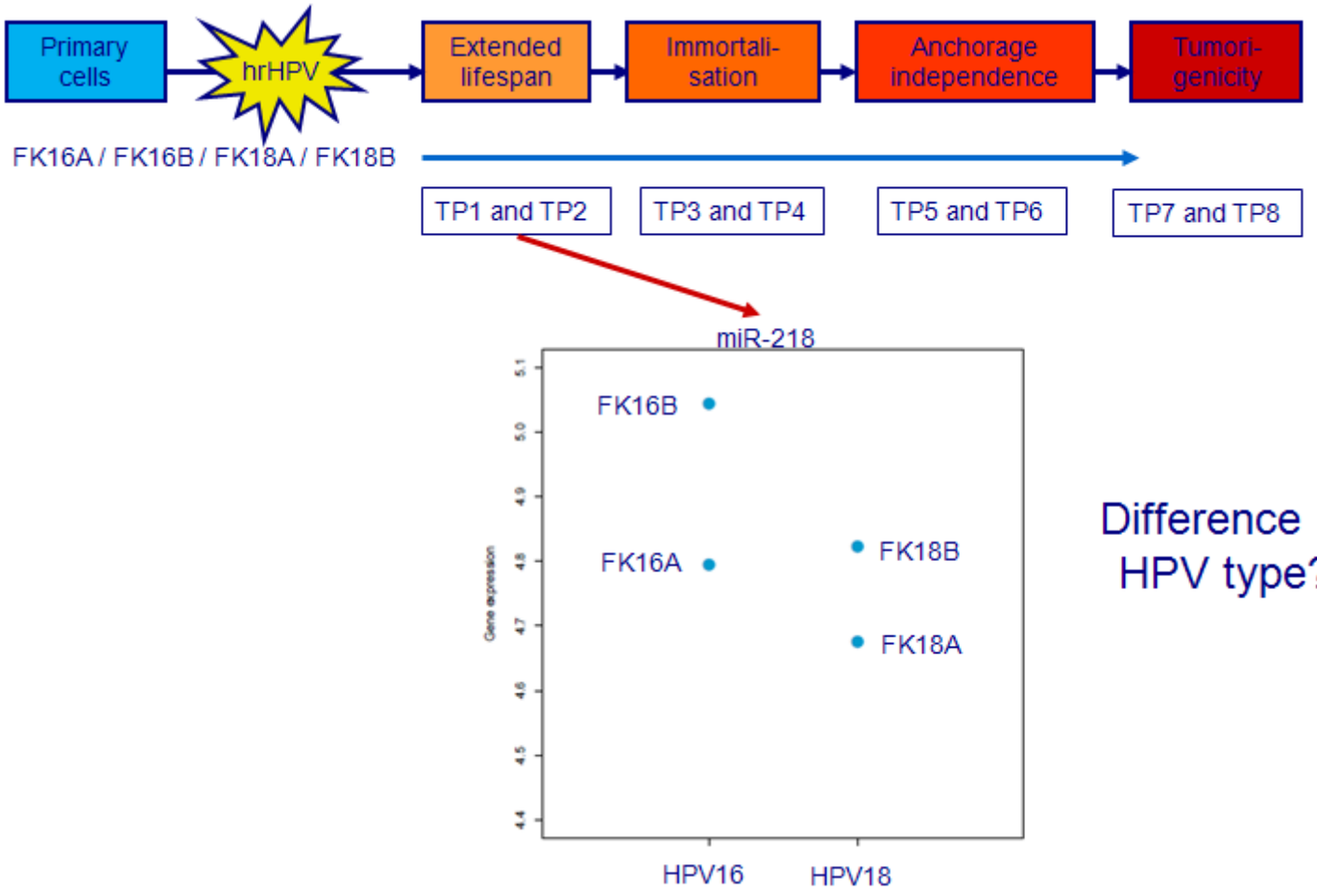
miR-218:

one moment in time: TP1



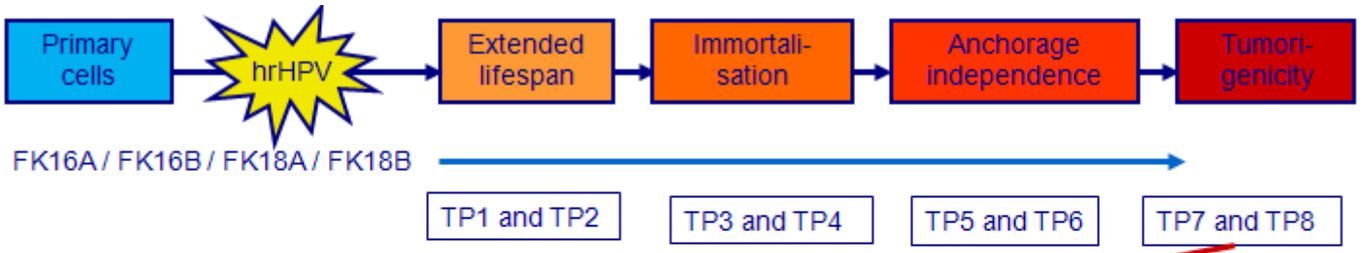
one moment in time: TP8

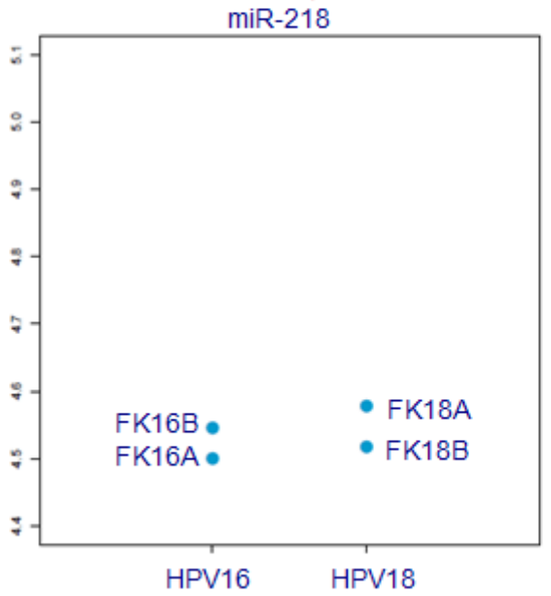
Inference based on TP1



Difference in HPV type?

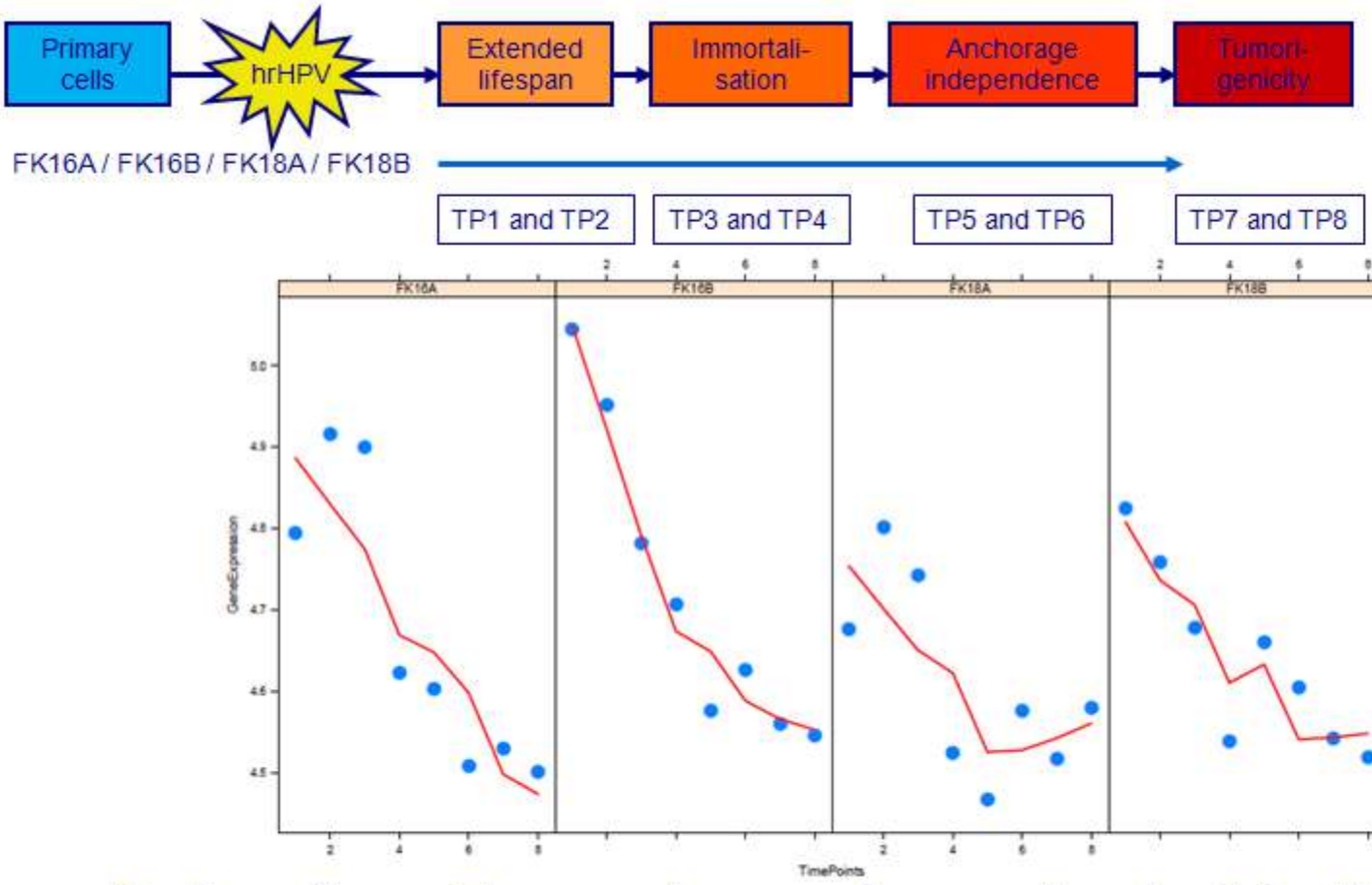
Inference based on TP8





No difference?

Strength of time-course



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

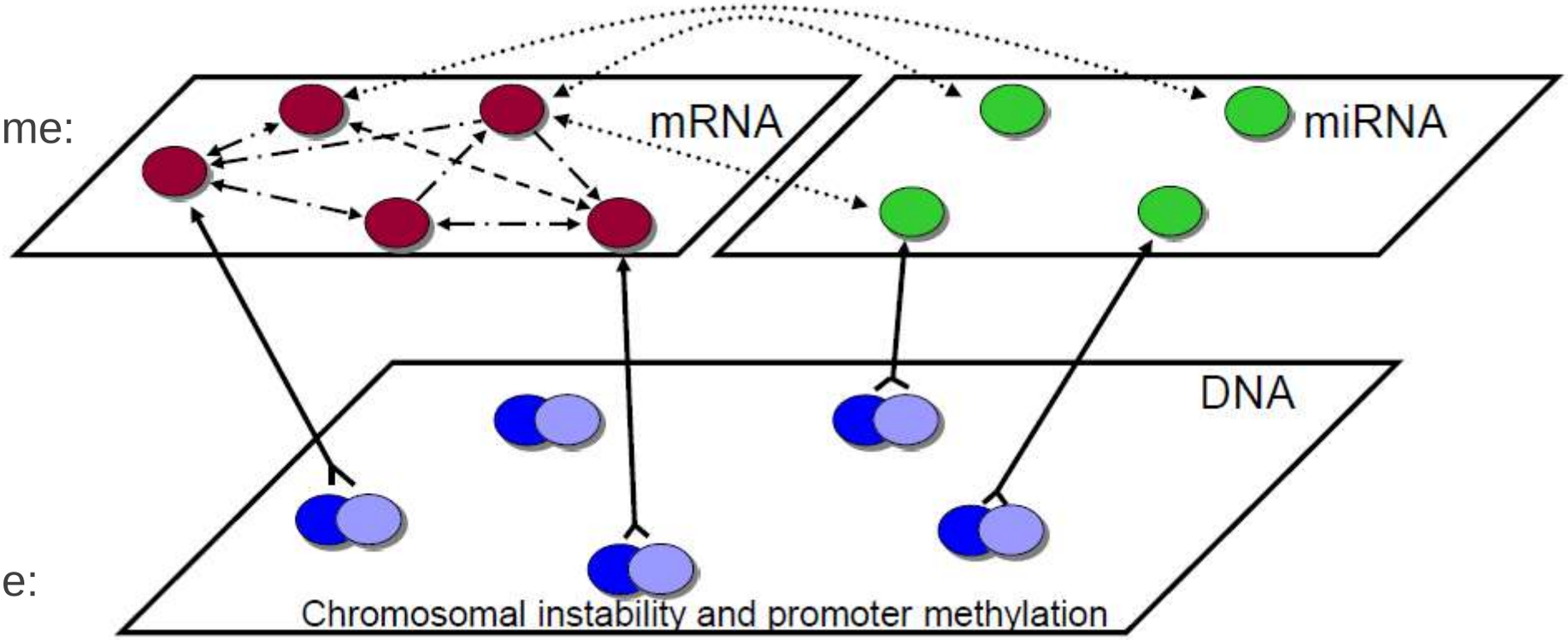
Why integration?

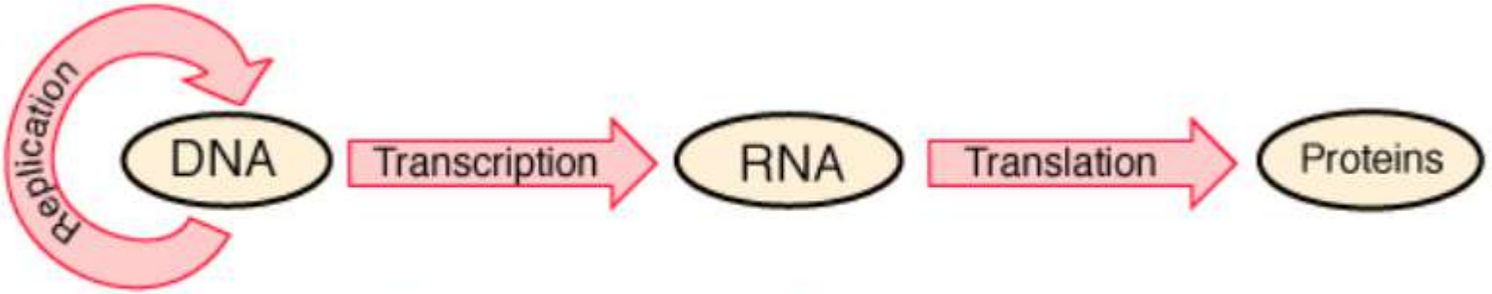
“Blind monks examining an elephant” by Itcho Hanabusa 1888

Multi-omics data integration

Transcriptome:

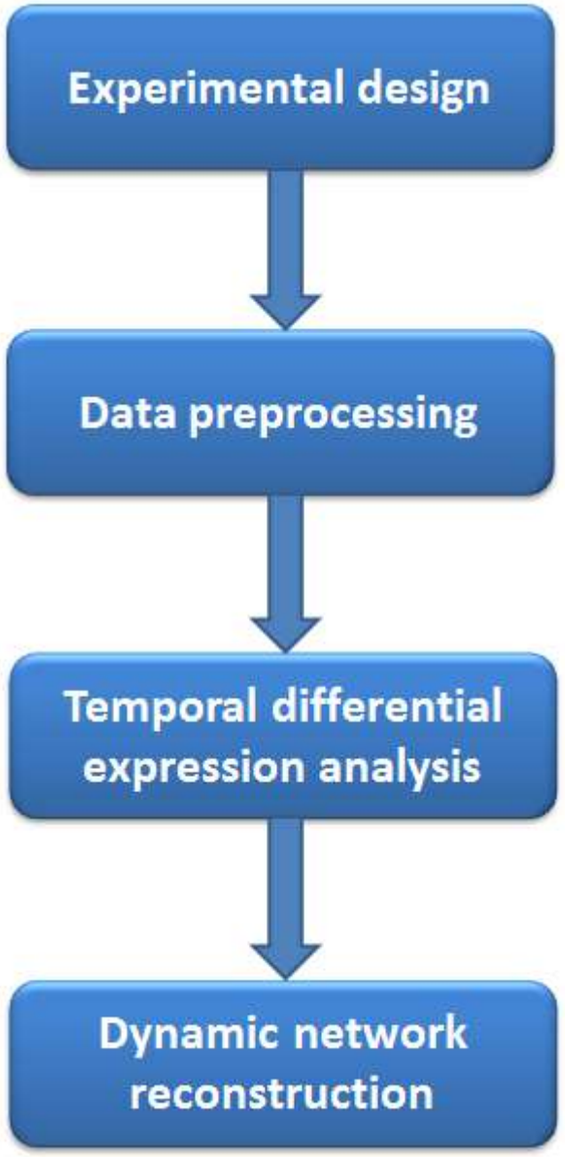
Genome:





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

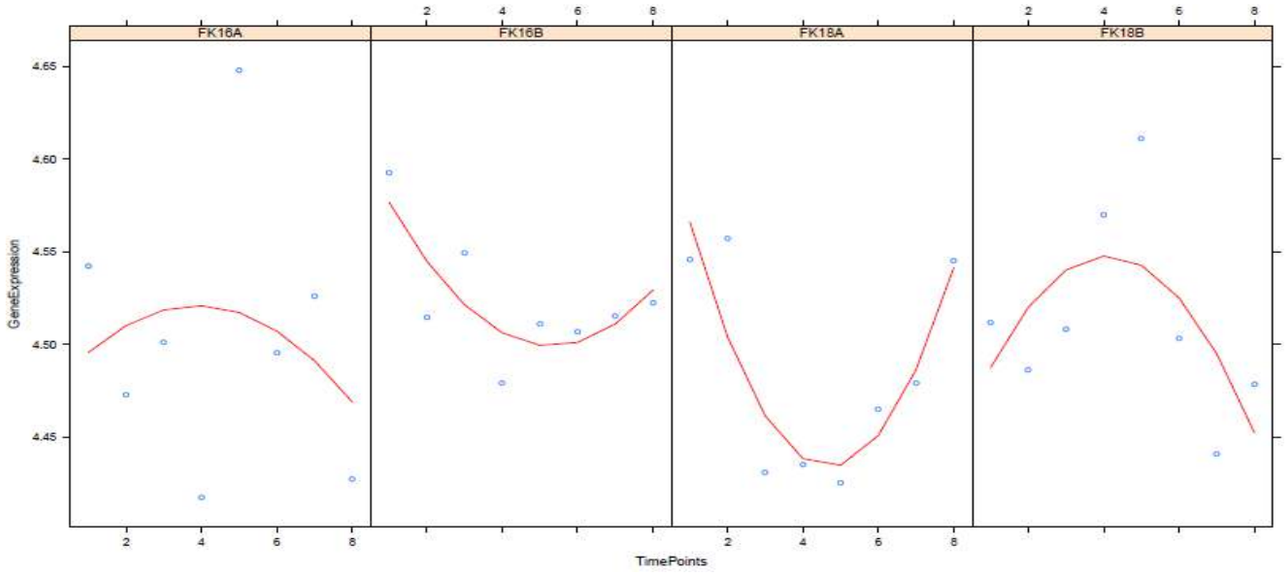
mRNA: 64 genes linked to p53 signaling pathway
miRNA: 106 genes which target mRNA

Temporal differential expression

GE = CL + Time

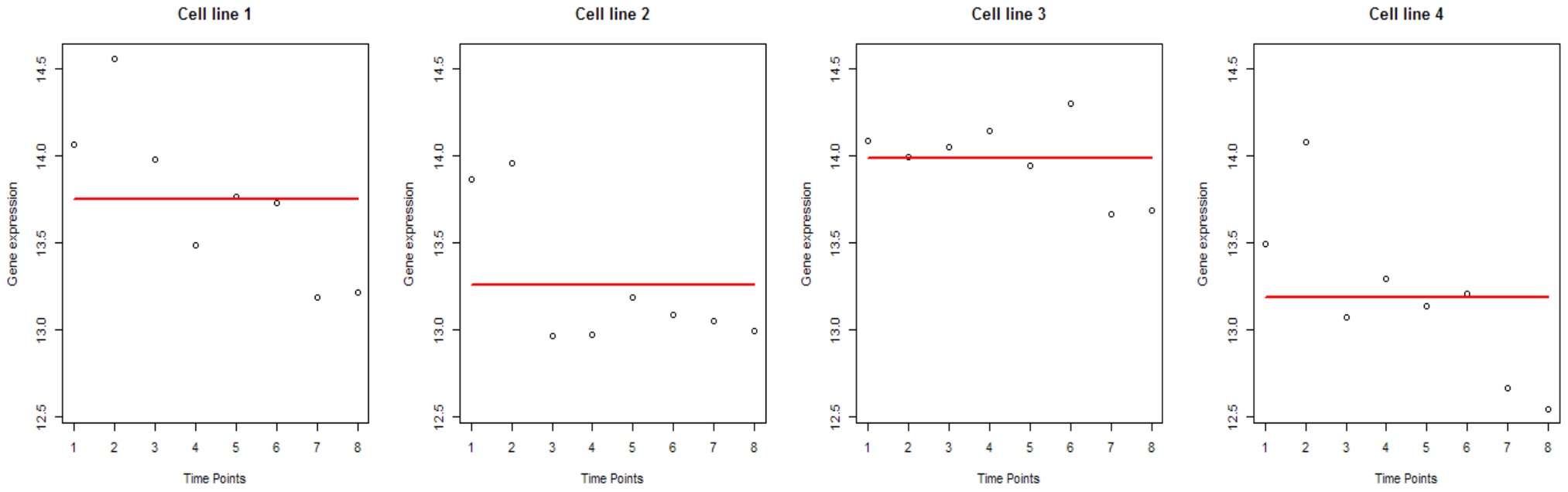
fixed (cell line, CN) effect

random (time) effect



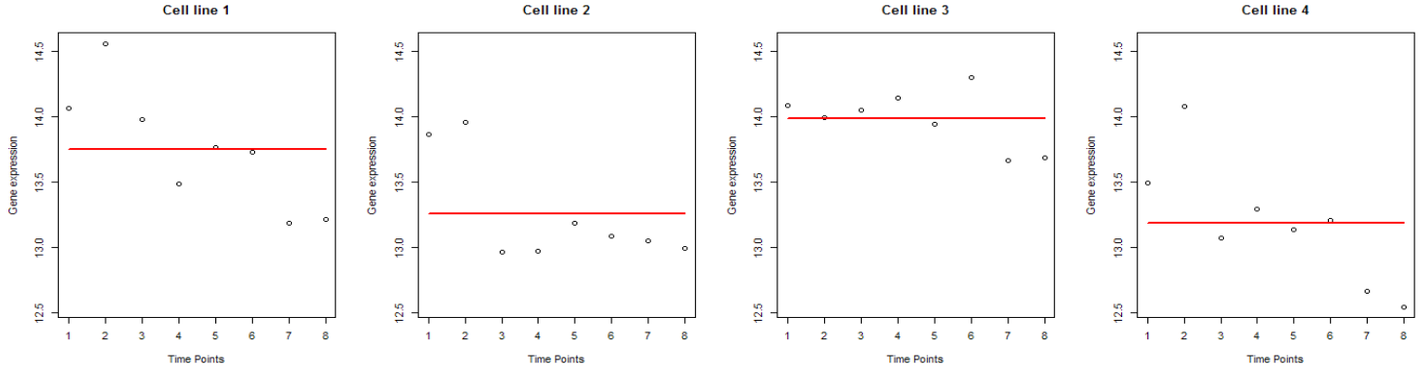
- Cell line effect

SP1



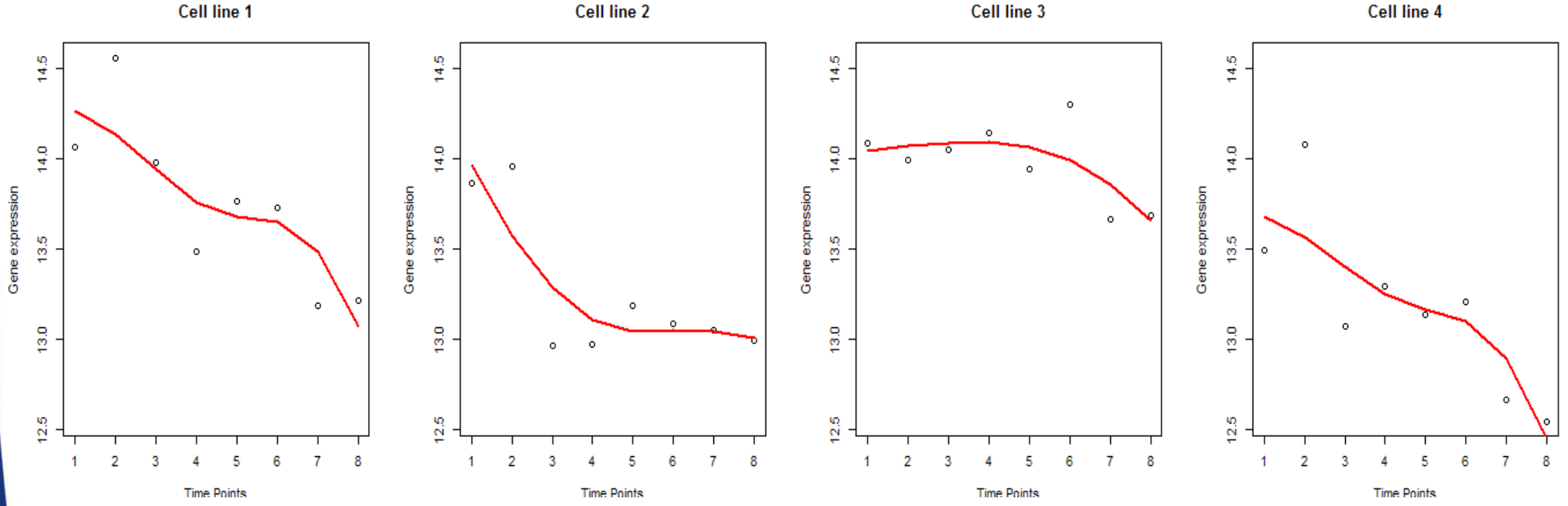
- Cell line effect

SP1



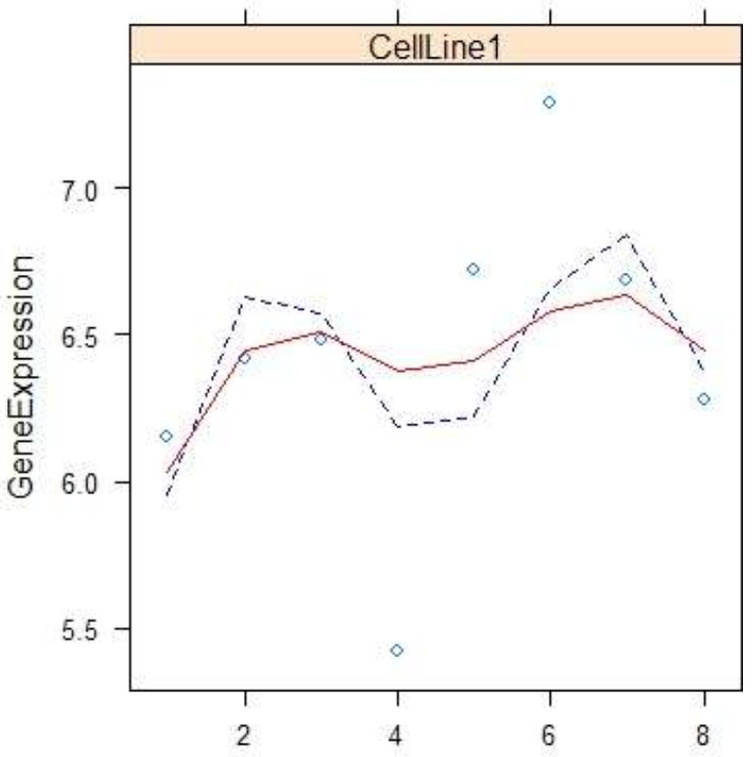
- Cell line and Time effect

SP1



Shrinkage

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



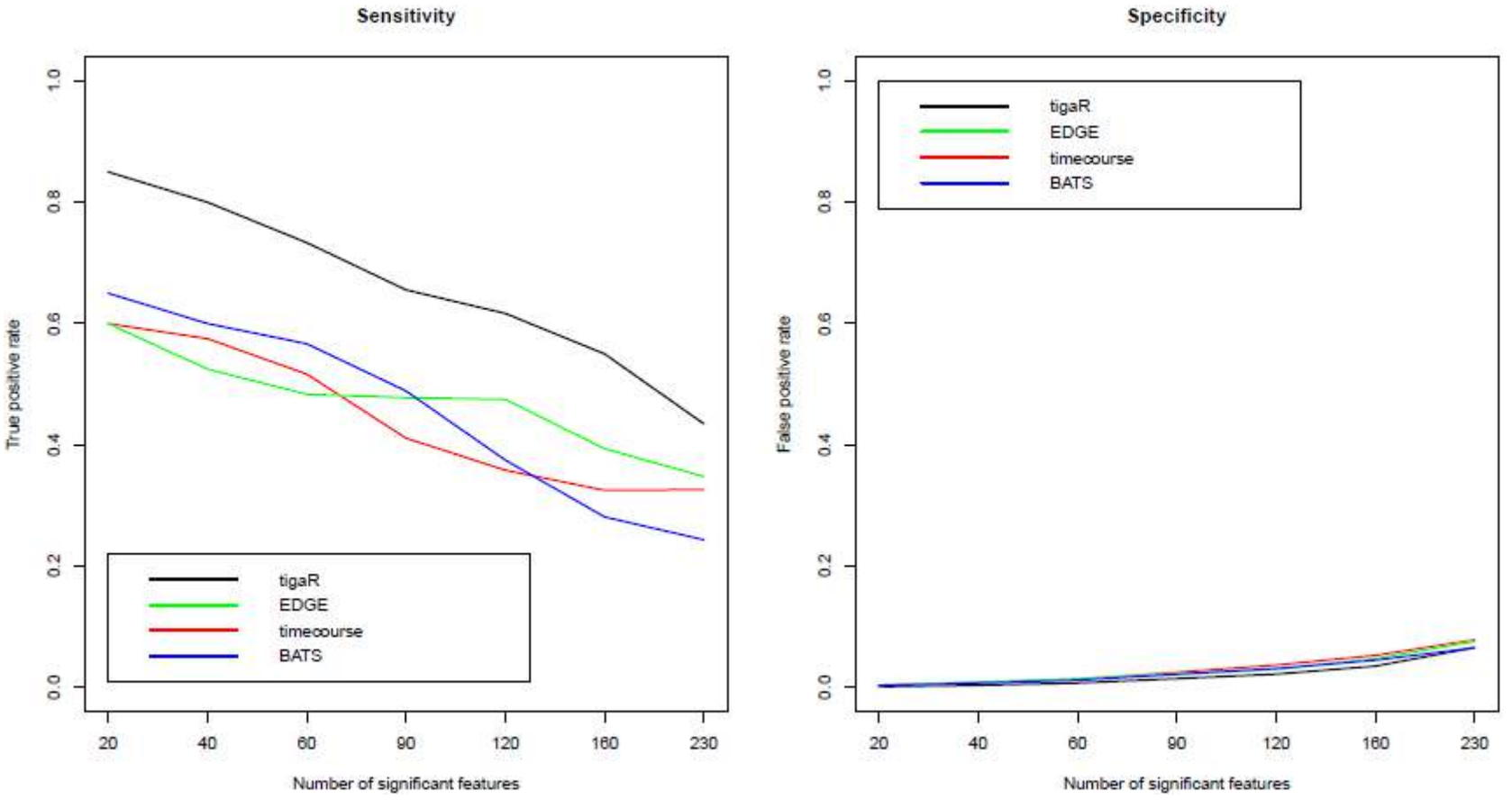
Comparison

- 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.

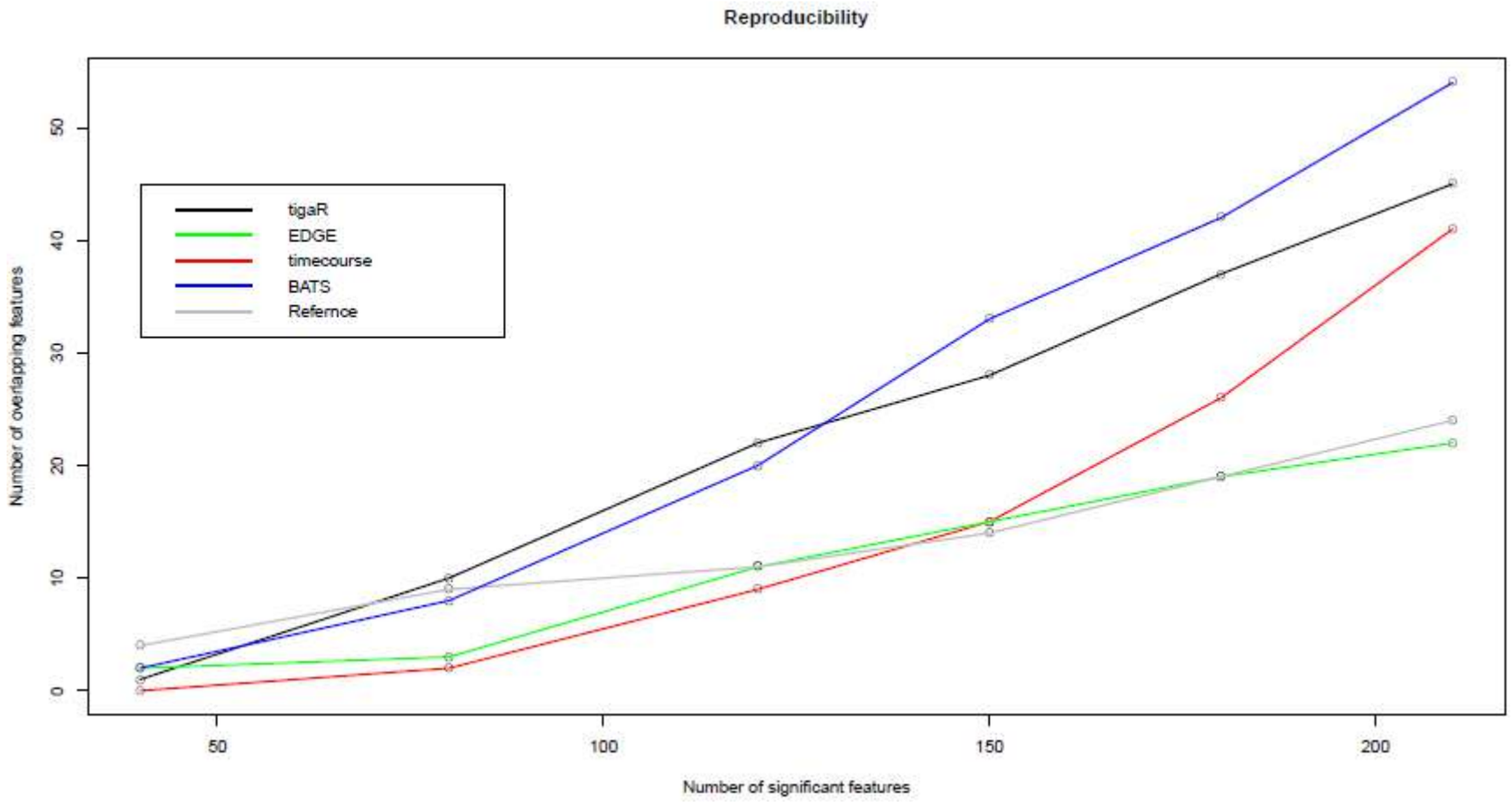
Sensitivity and specificity

- Truth – overlap of significant genes among methods.



Reproducibility

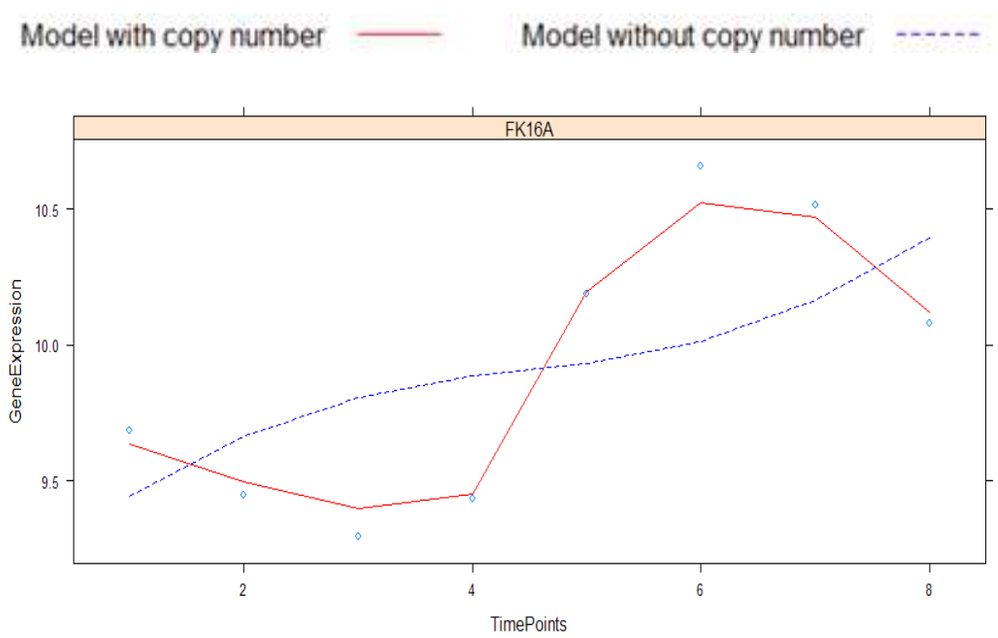
- Equally divided data set in two groups.



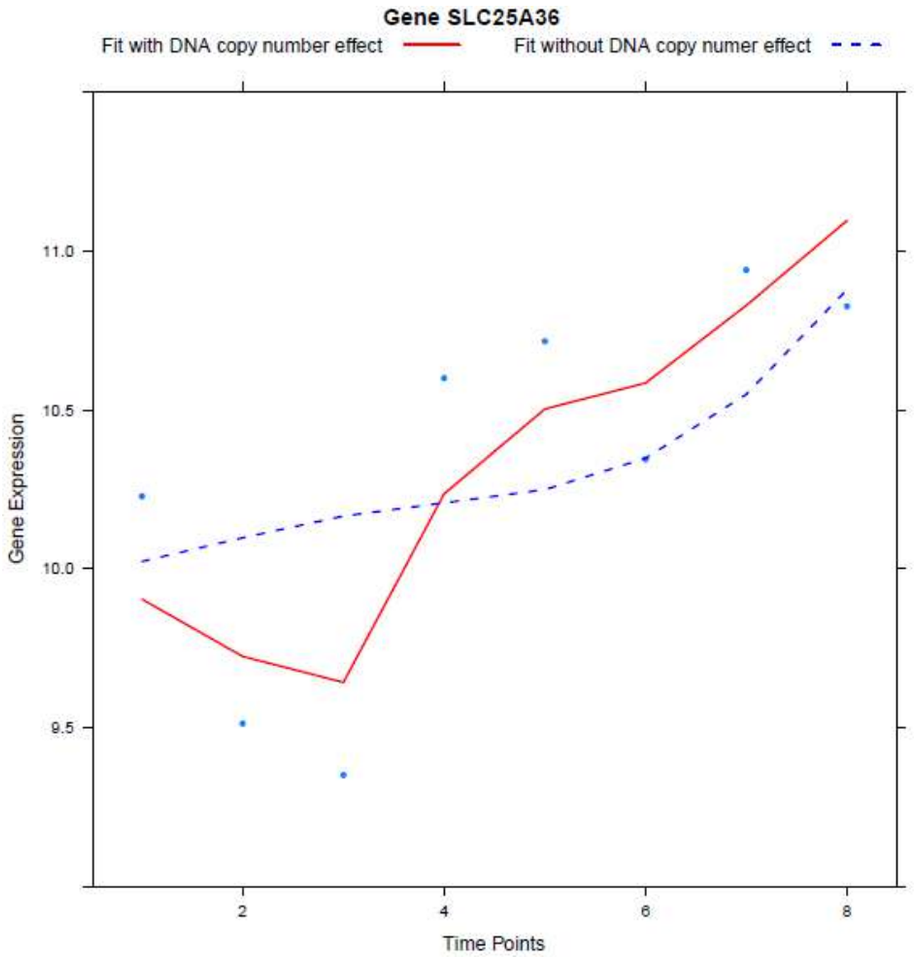
DNA copy number (CN)

GE = CL + CN + Time

Gene GSTM3:

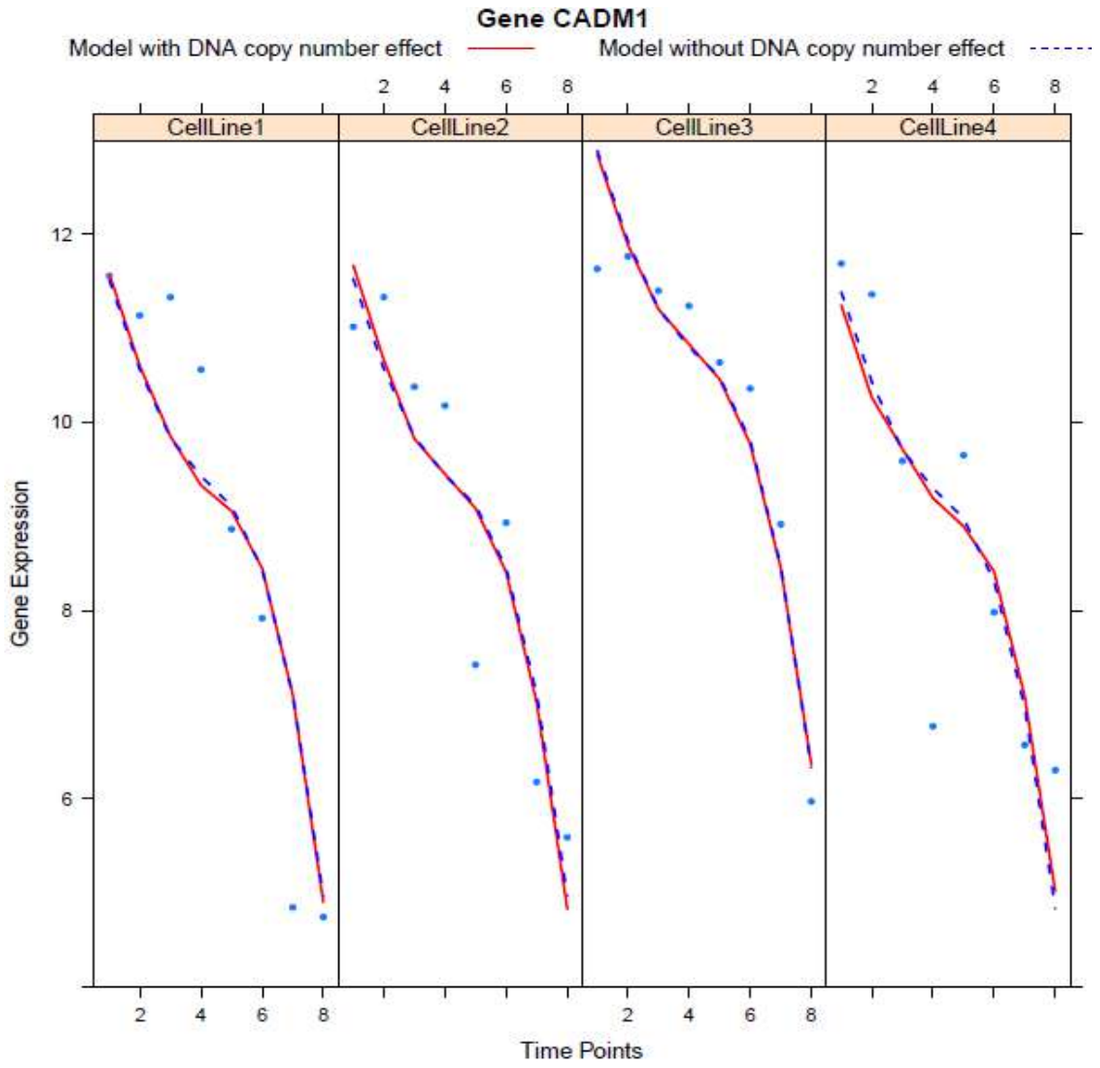


SLC25A36 – gene with CN effect



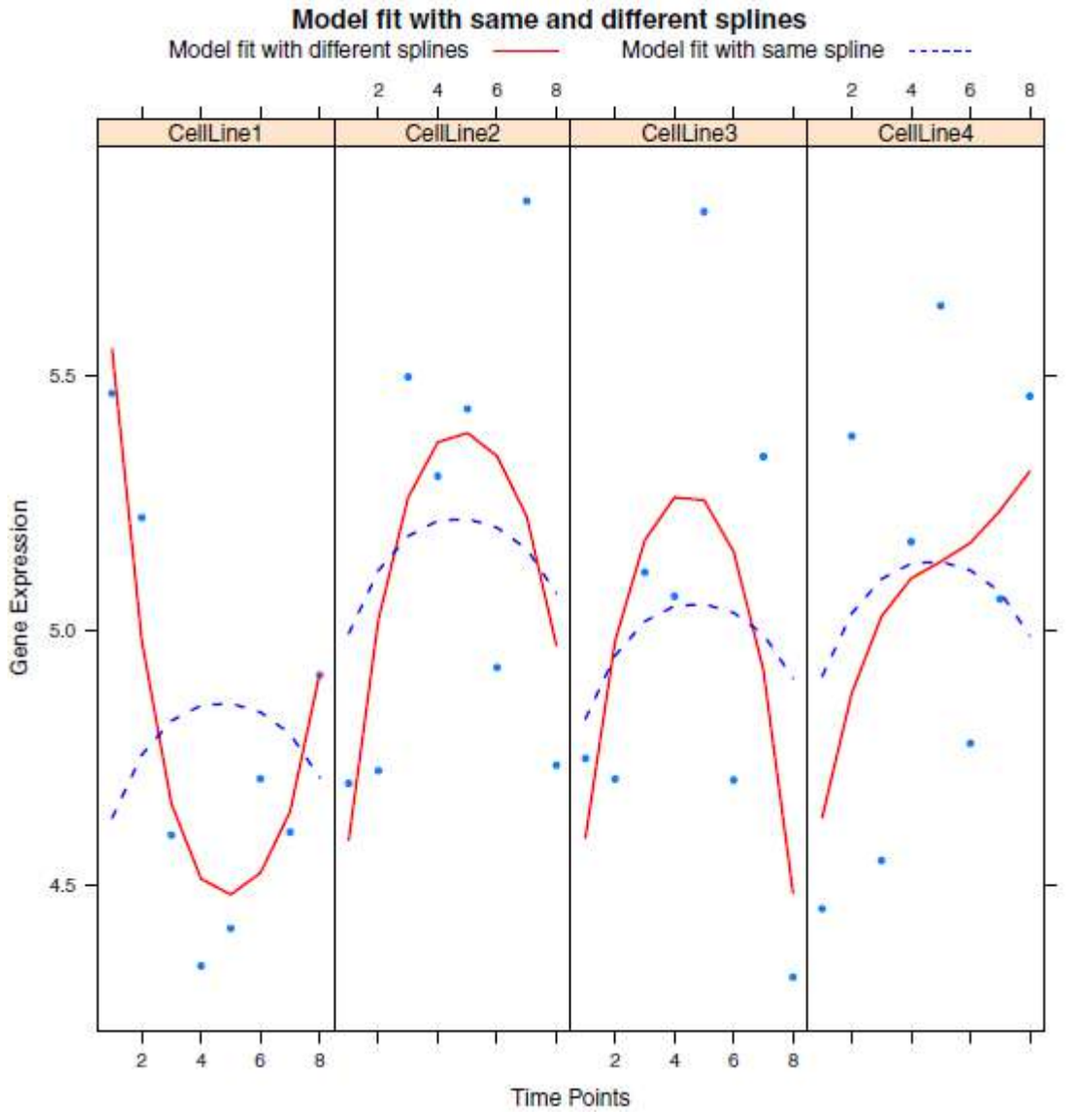
Wilting et al., Genes, Chromosomes and Cancer, 2008.

CADM1- gene without CN effect



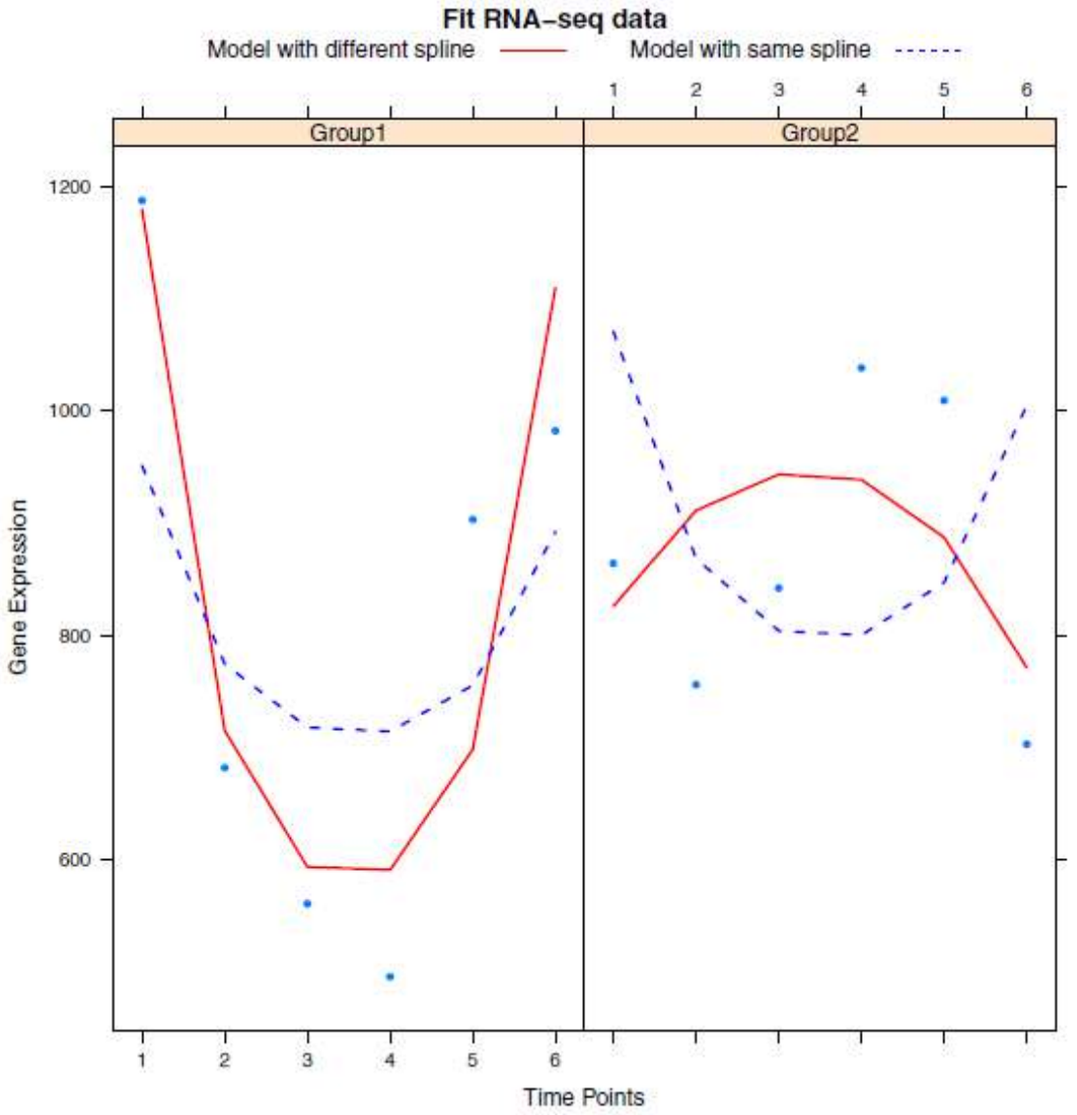
Overmeer et al., J Pathol., 2008.

Fit flexibility

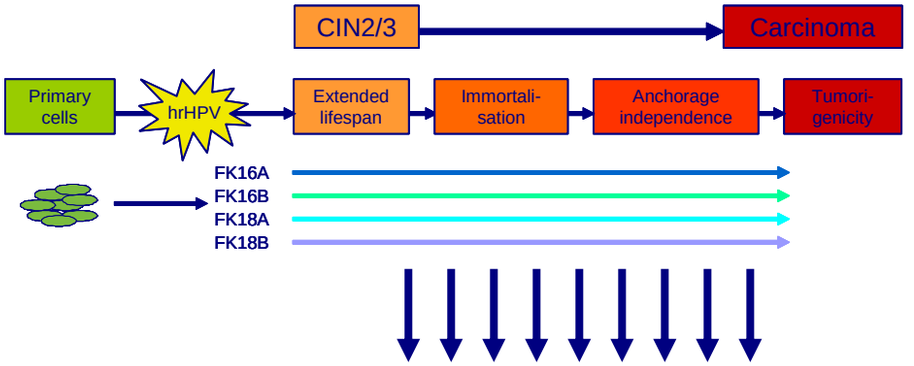


RNA-seq data

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

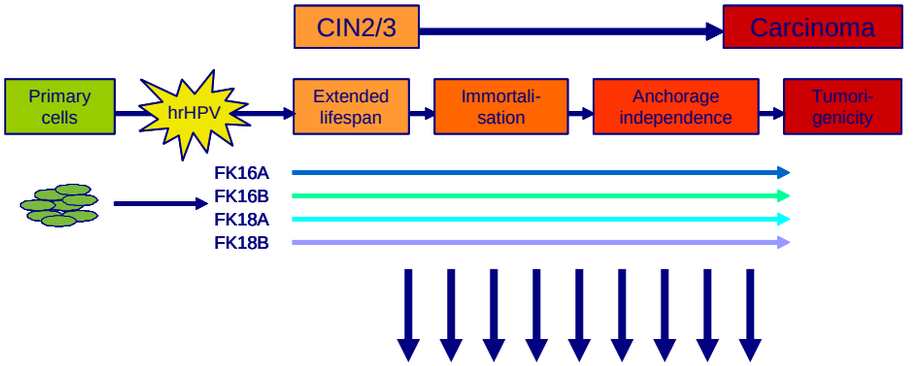


tigaR analysis



Measured expression of 1187 miRNAs and 27637 mRNAs

tigaR analysis

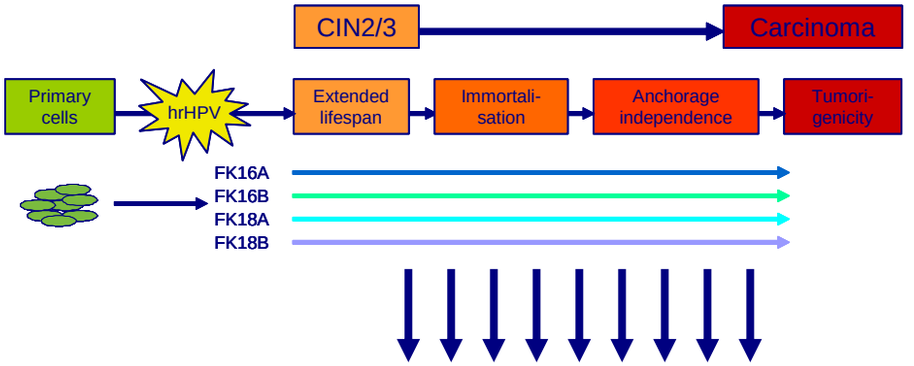


Measured expression of 1187 miRNAs and 27637 mRNAs

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

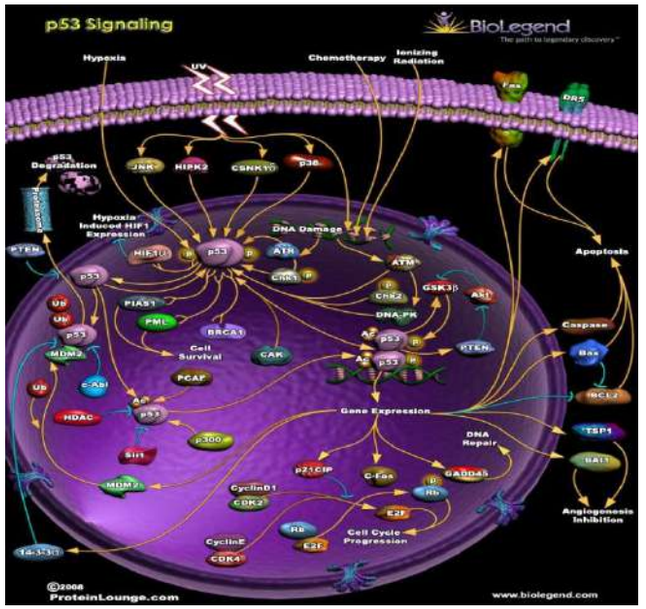
(~34% of altered expression in both cases)

Pathways

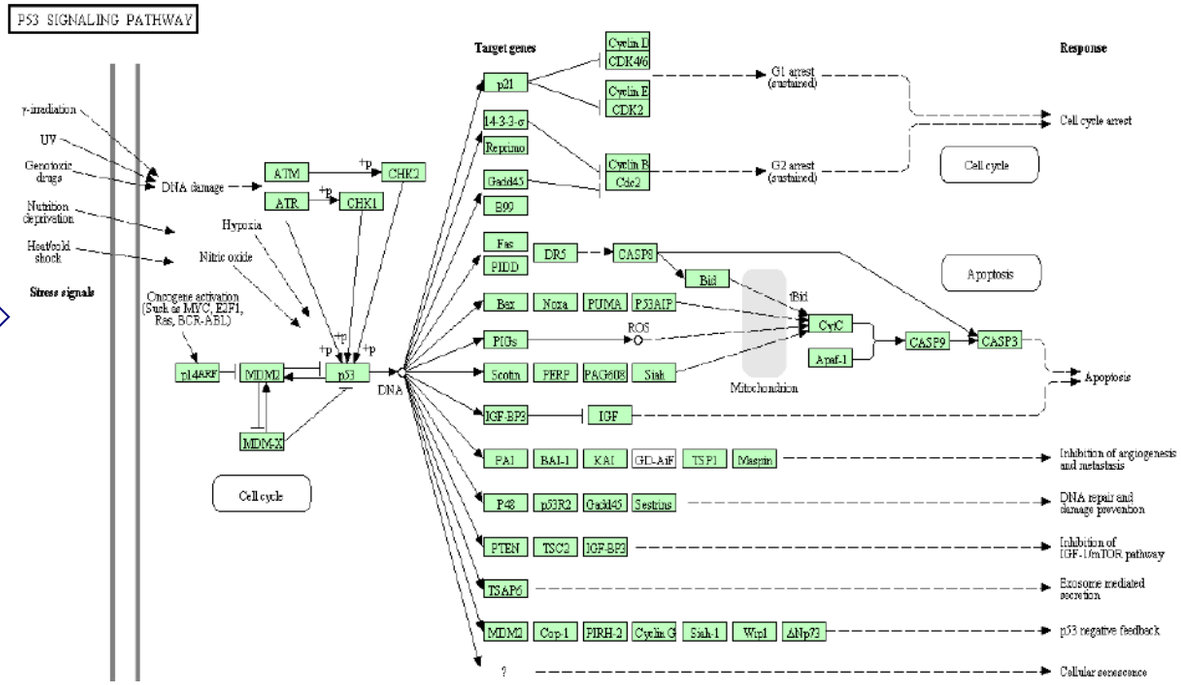
- Pathway are defined using repositories: **KEGG, GO, Reactome**…

- Problems with repositories:
  - Incomplete
  - Mostly well-known pathways
  - Loosely defined

- Reconstruction of the p53 signaling pathway in cervical cancer



KEGG

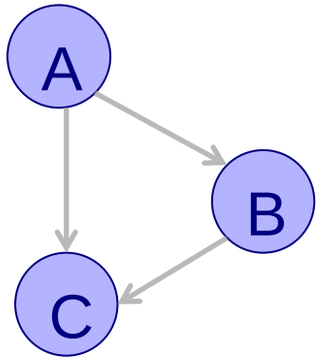


Network

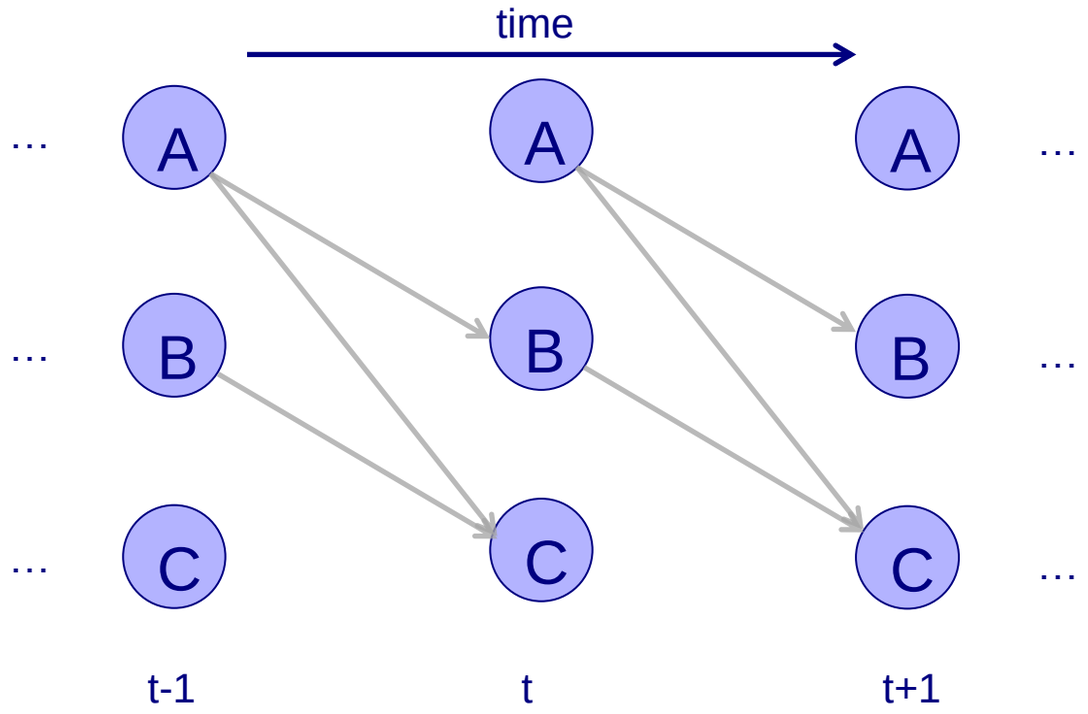
- Pathway can be represented by network or graph

node or vertex, indicate a gene

Edge or arrow, indicating an interaction between two genes

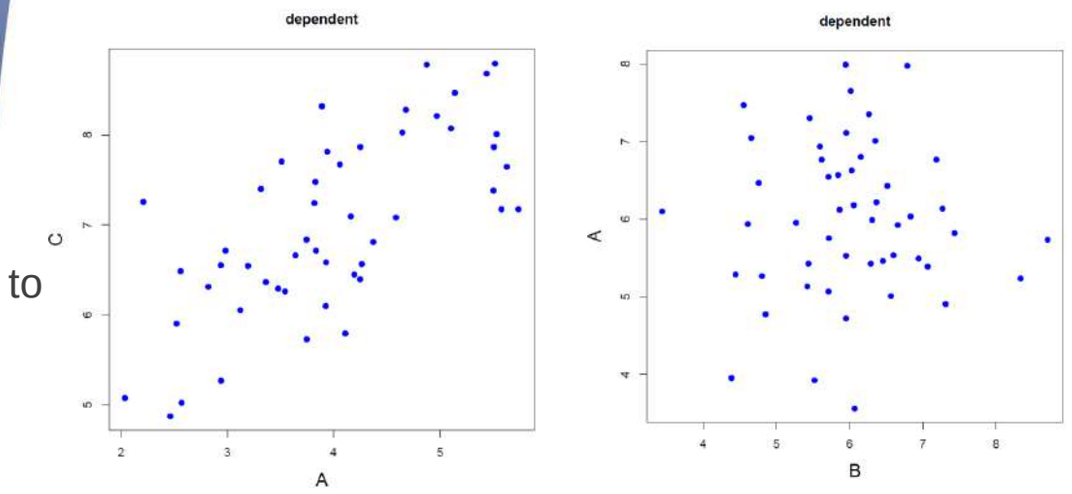


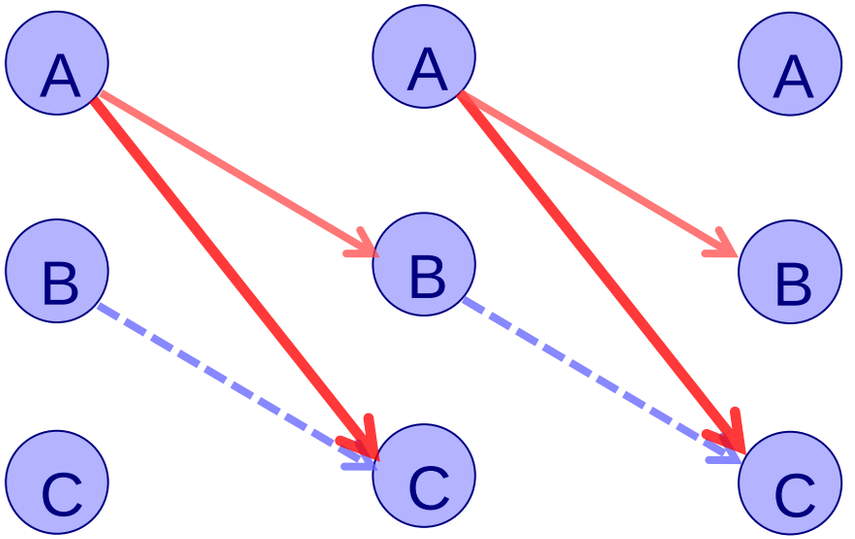
Feed-forward loop



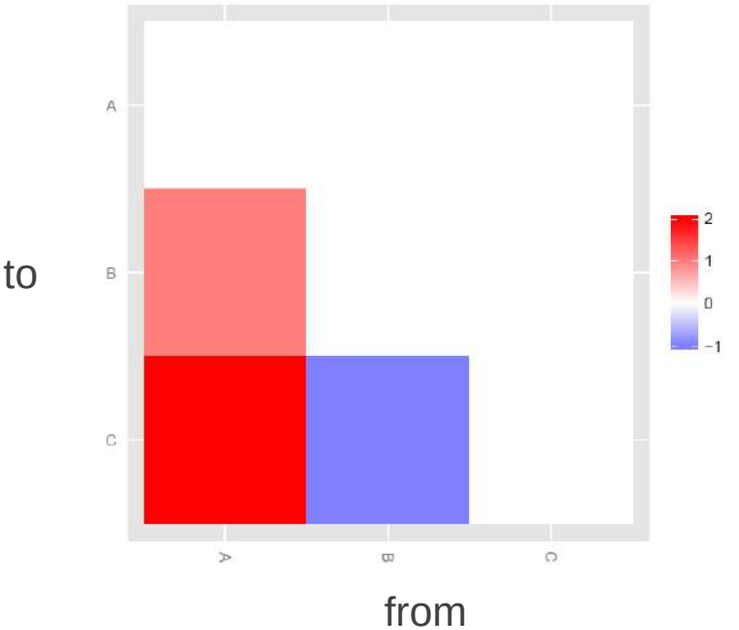
Feed-forward loop (unrolled)

Data, model and network



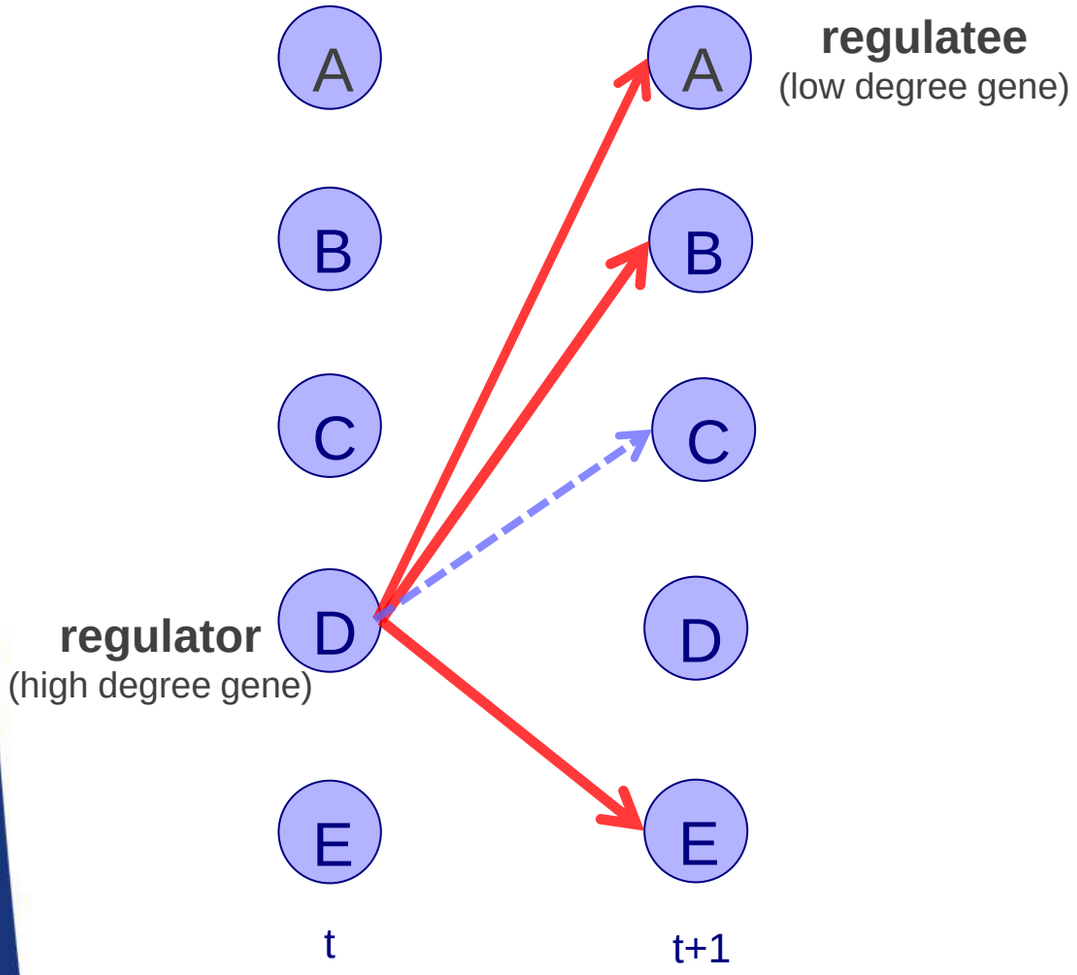


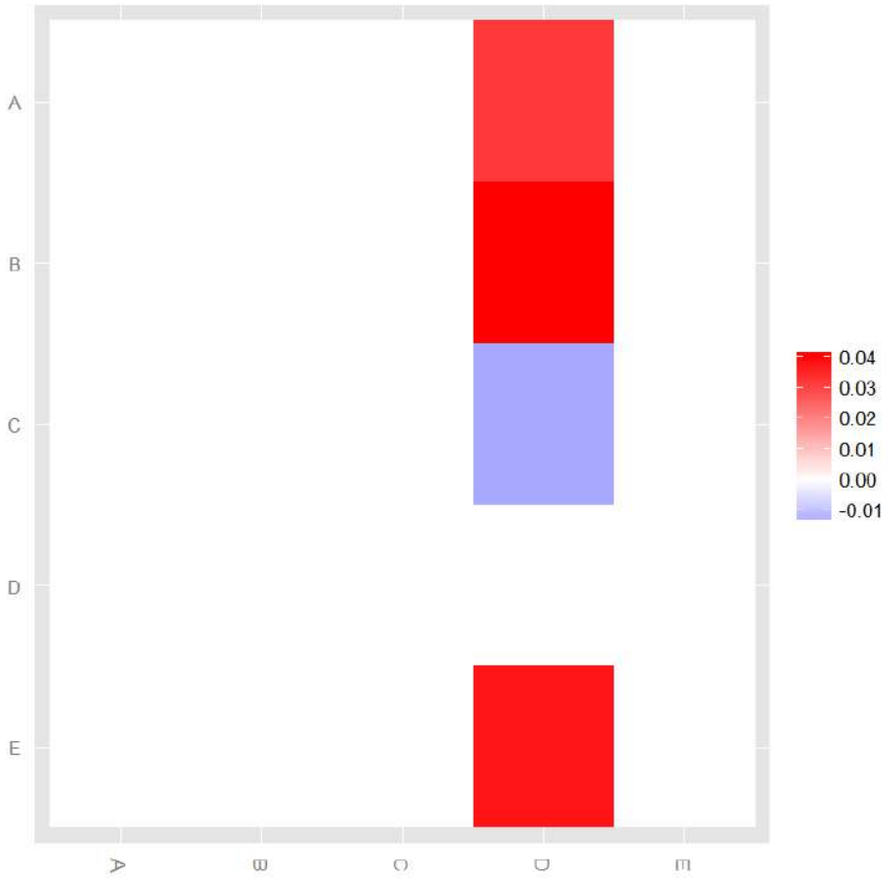
$$
\begin{aligned}
A_t &= &&+\varepsilon_{A,t} \\
B_t &= A_{t-1} &&+\varepsilon_{B,t} \\
C_t &= 2A_{t-1} - B_{t-1} &&+\varepsilon_{C,t}
\end{aligned}
$$



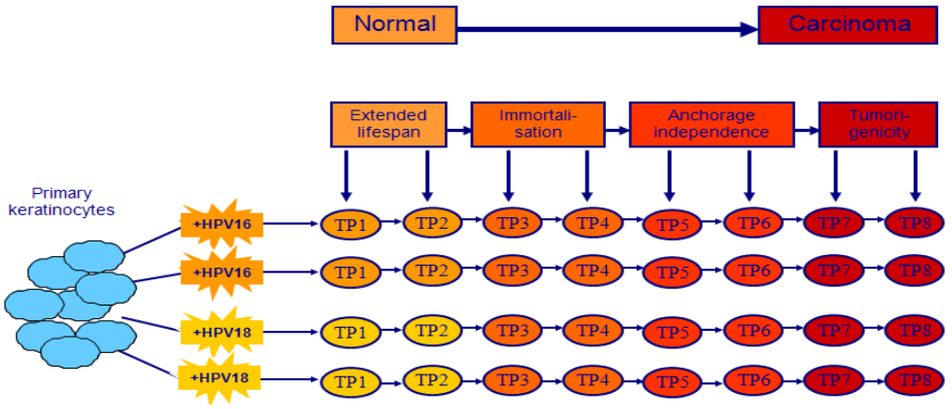
Hub genes

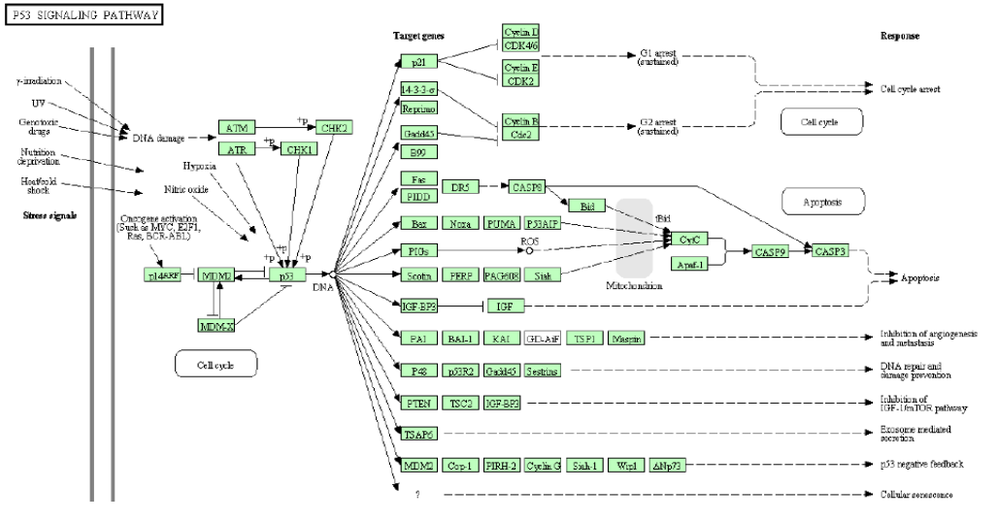
- Hub – gene with many connections - disease genes
- Important roles for diagnosing and therapy





Real data and network





64 mRNAs mapped to p53 signalling pathway

(28% of genes with significant time effect)

CDK6
THBS1
CCNE2
TP53
PMAIP1
IGFBP3
IGF1
SIAH1
ZMAT3
SERPINE1
CDKN2A
SESN3
CDKN1A
CCND2
RPRM
BBC3
GADD45G
TP73
SESN1
SESN2
PERP
TP53AIP1
PTEN
SFN
CCNB3
BID
EI24
CCNG1
RRM2
STEAP3
CASP8
ATR
ATM
FAS
CCNG2
CCNE1
CCND3
TNFRSF10B
RFWD2
SHISA5
GADD45B
APAF1

t

???

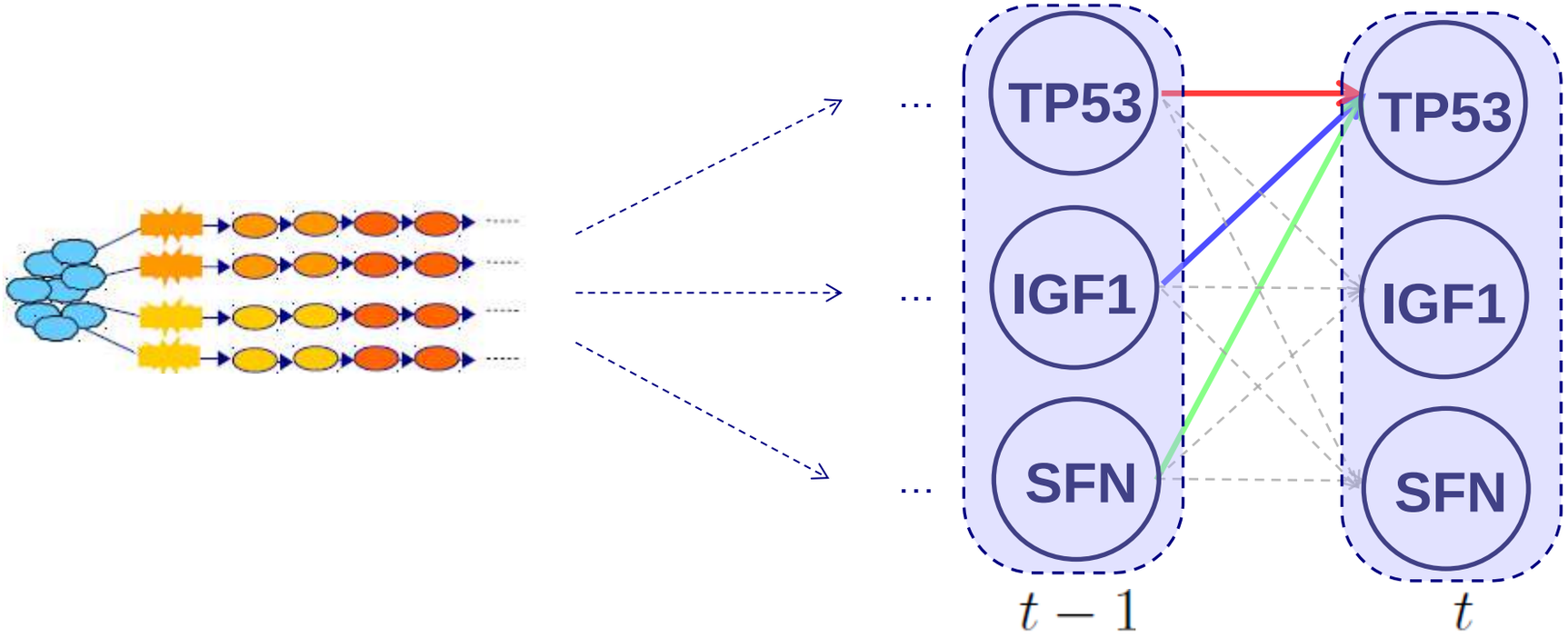
CDK6
THBS1
CCNE2
TP53
PMAIP1
IGFBP3
IGF1
SIAH1
ZMAT3
SERPINE1
CDKN2A
SESN3
CDKN1A
CCND2
RPRM
BBC3
GADD45G
TP73
SESN1
SESN2
PERP
TP53AIP1
PTEN
SFN
CCNB3
BID
EI24
CCNG1
RRM2
STEAP3
CASP8
ATR
ATM
FAS
CCNG2
CCNE1
CCND3
TNFRSF10B
RFWD2
SHISA5
GADD45B
APAF1

t+1

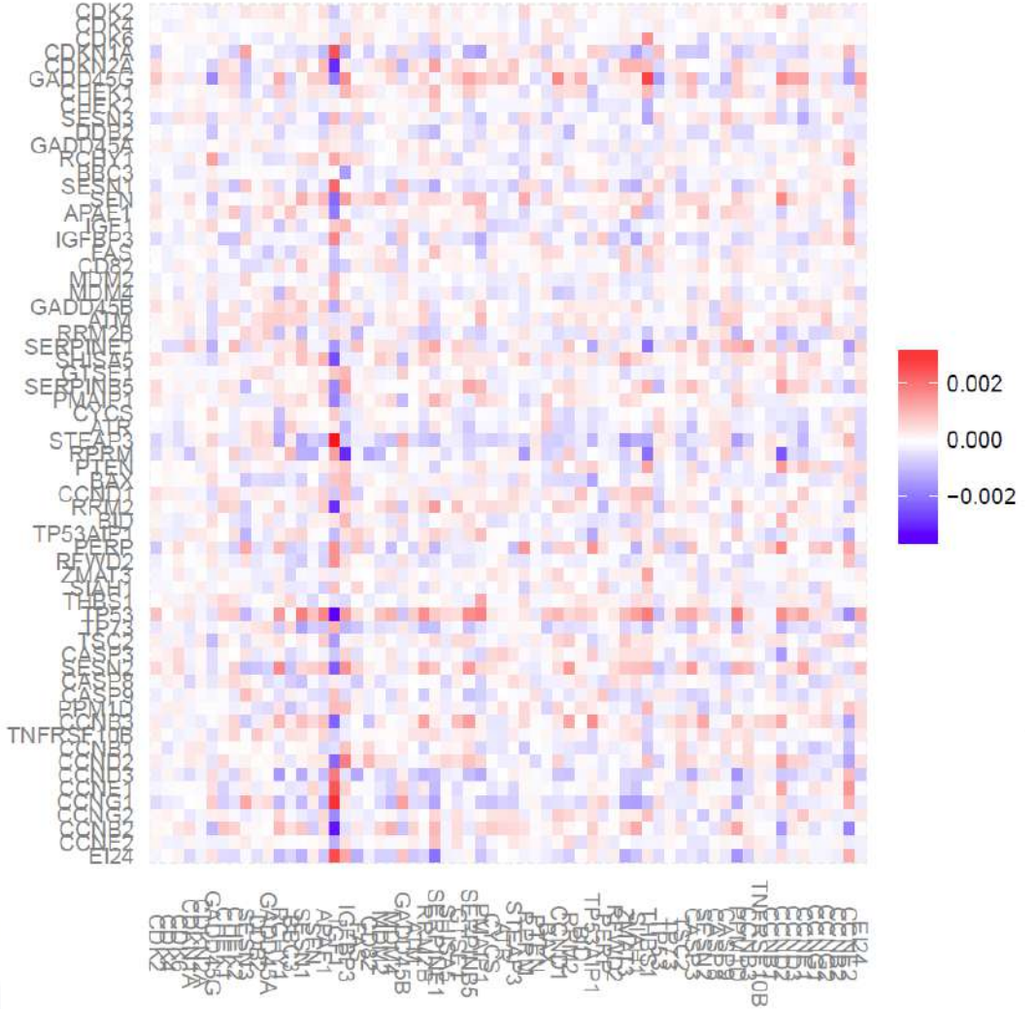
Model

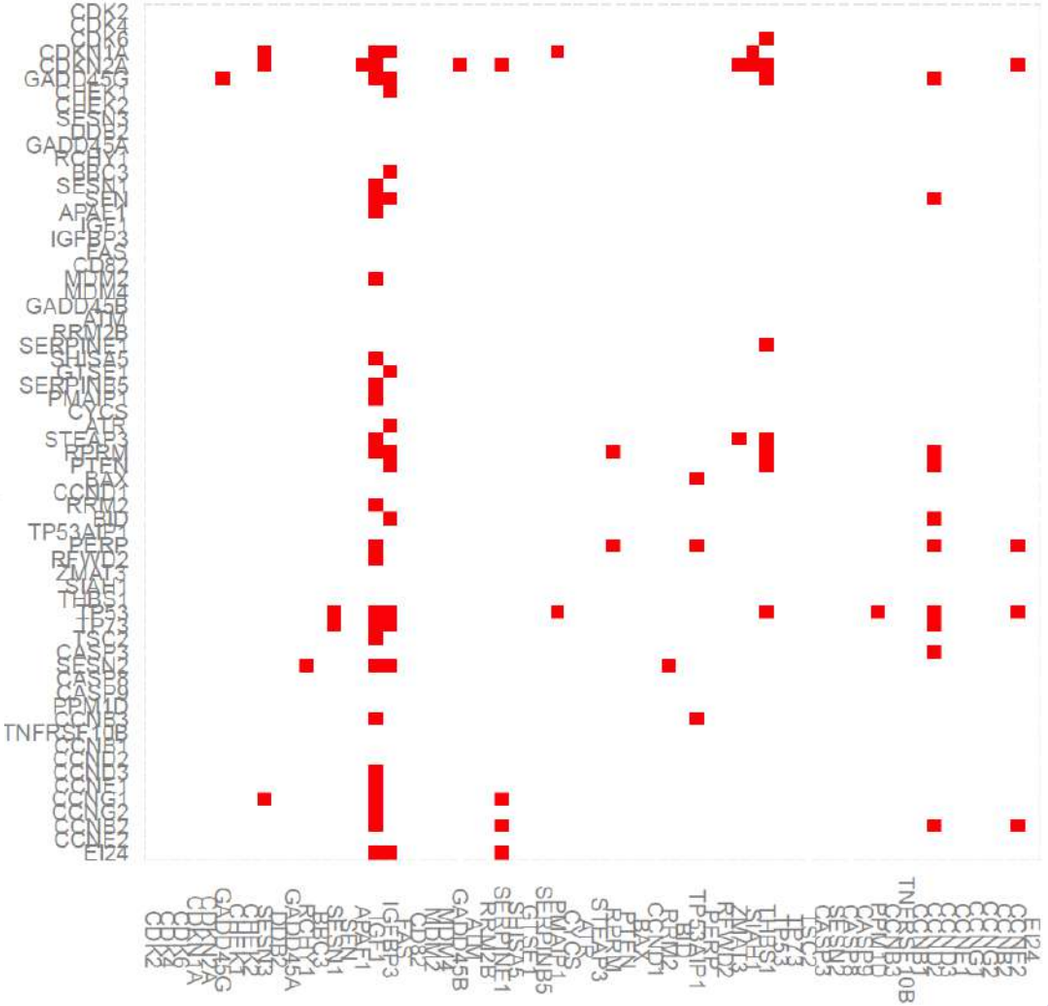
GE(t) = A * GE(t-1)

TP53 gene: $\mathbf{TP53}_t = a_1\mathbf{TP53}_{t-1} + a_2\mathbf{IGF1}_{2,t-1} + a_3\mathbf{SFN}_{t-1} + \varepsilon_{1,t}$



Estimation and sparsification





Prior knowledge and re-estimation

- User-specified prior knowledge on the parameters
- Estimated from pilot study publicly available data

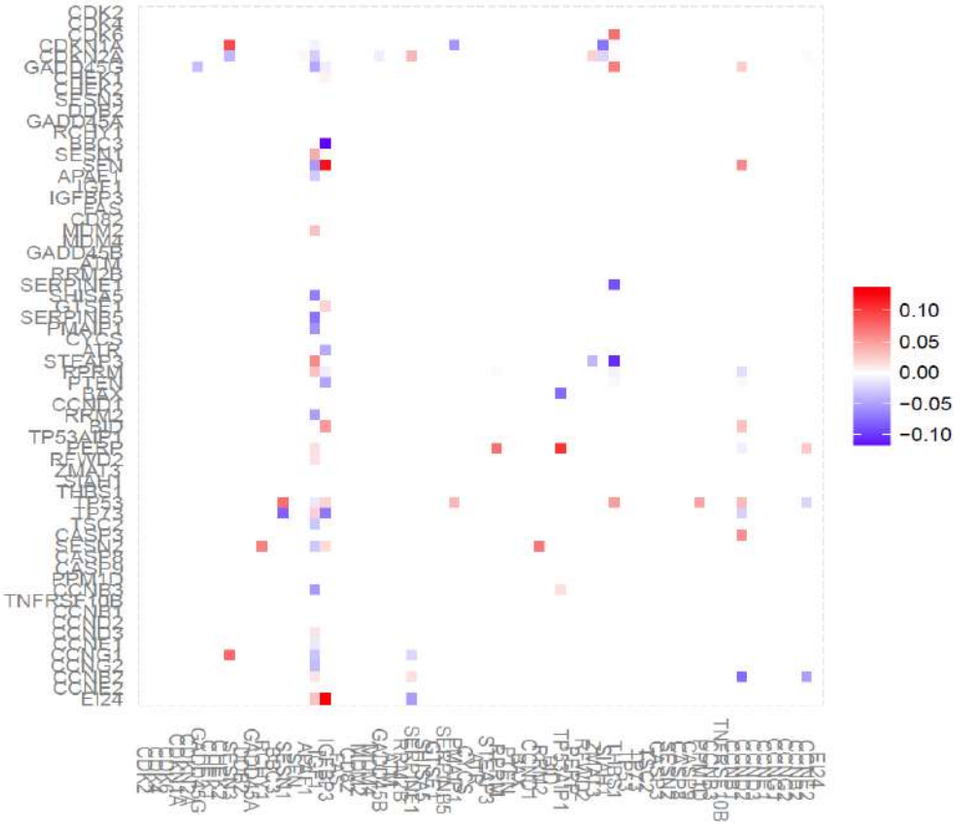
Pilot data + topology

GEO Gene Expression Omnibus

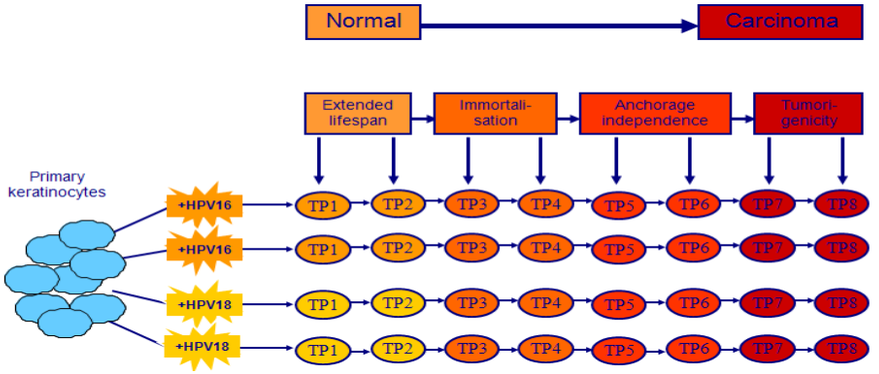
KEGG

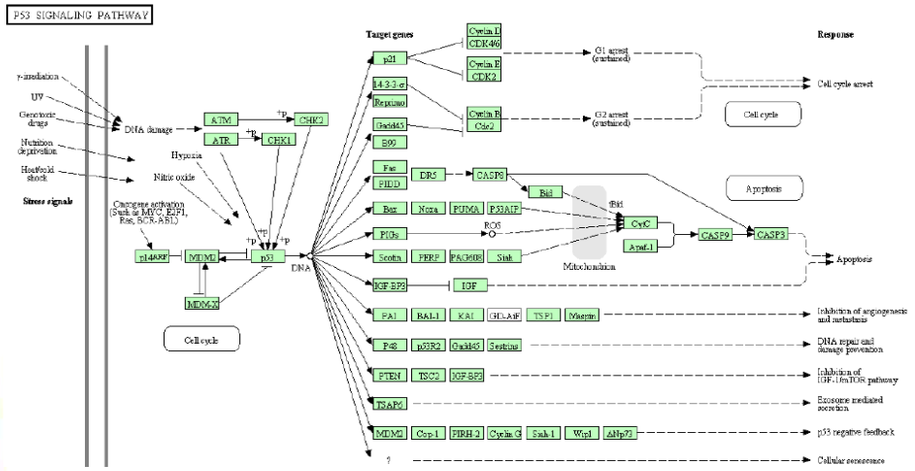
- Analytic solution, efficiently evaluable
- Less biased estimates of **A**

A:



Network



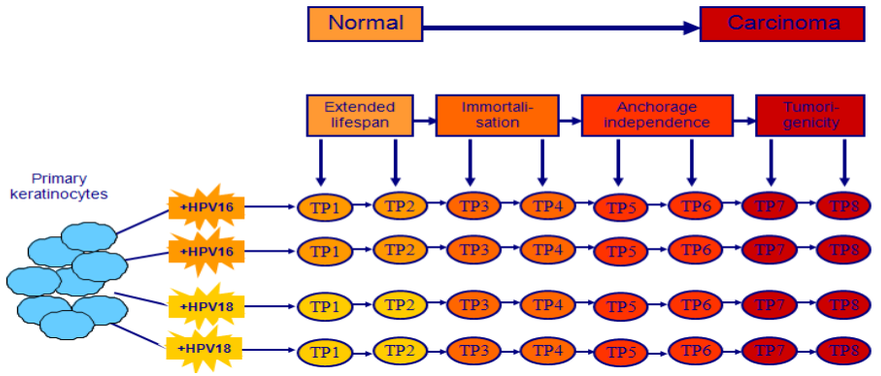


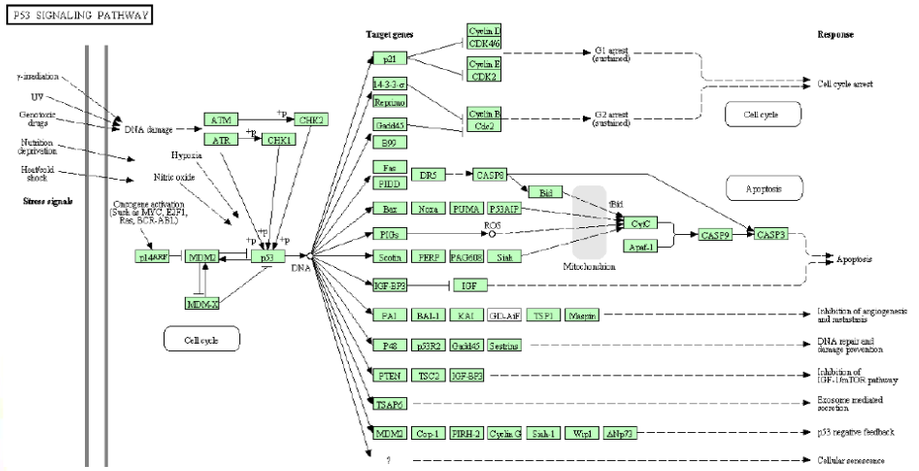
R-package

rags2ridges

(available from CRAN)

Network

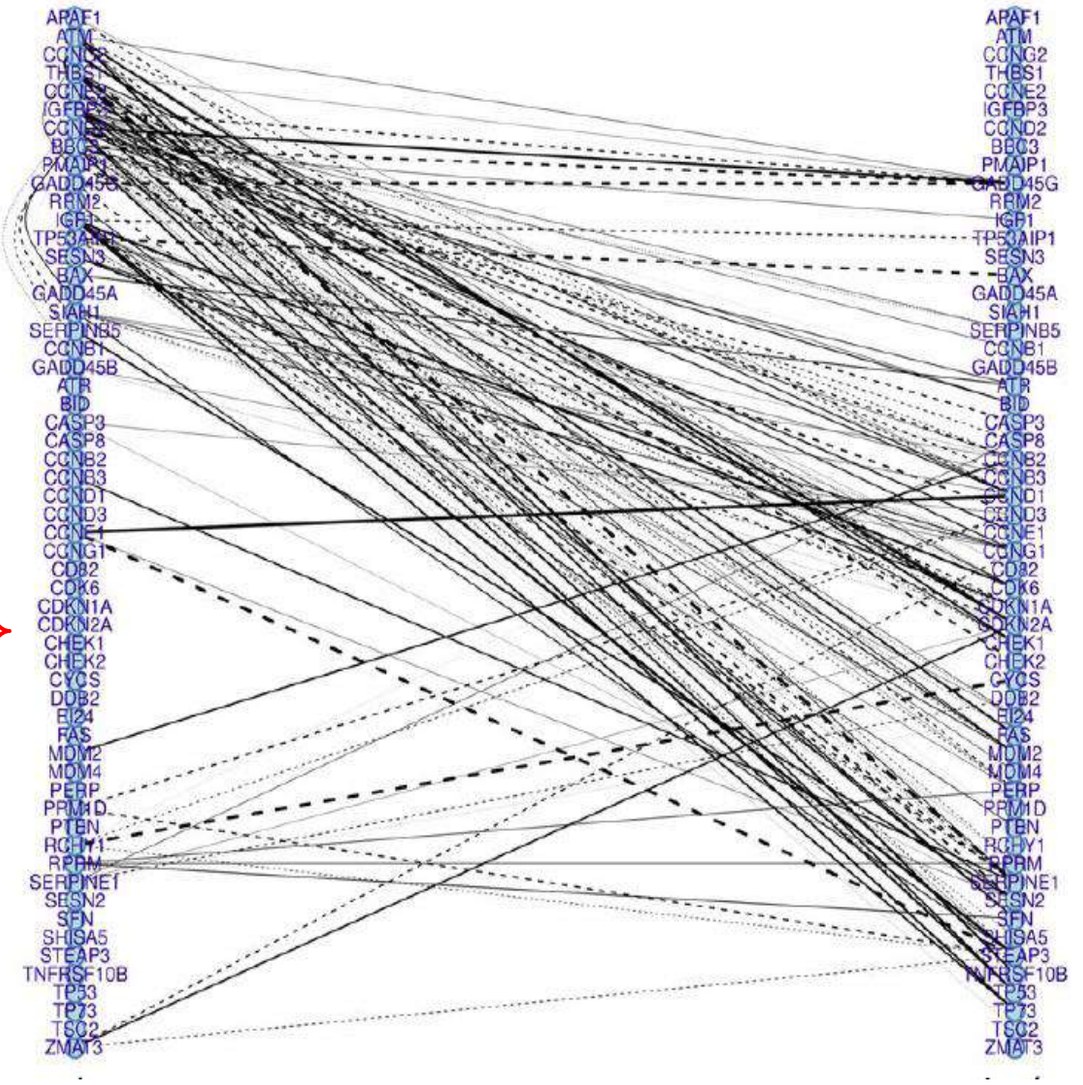




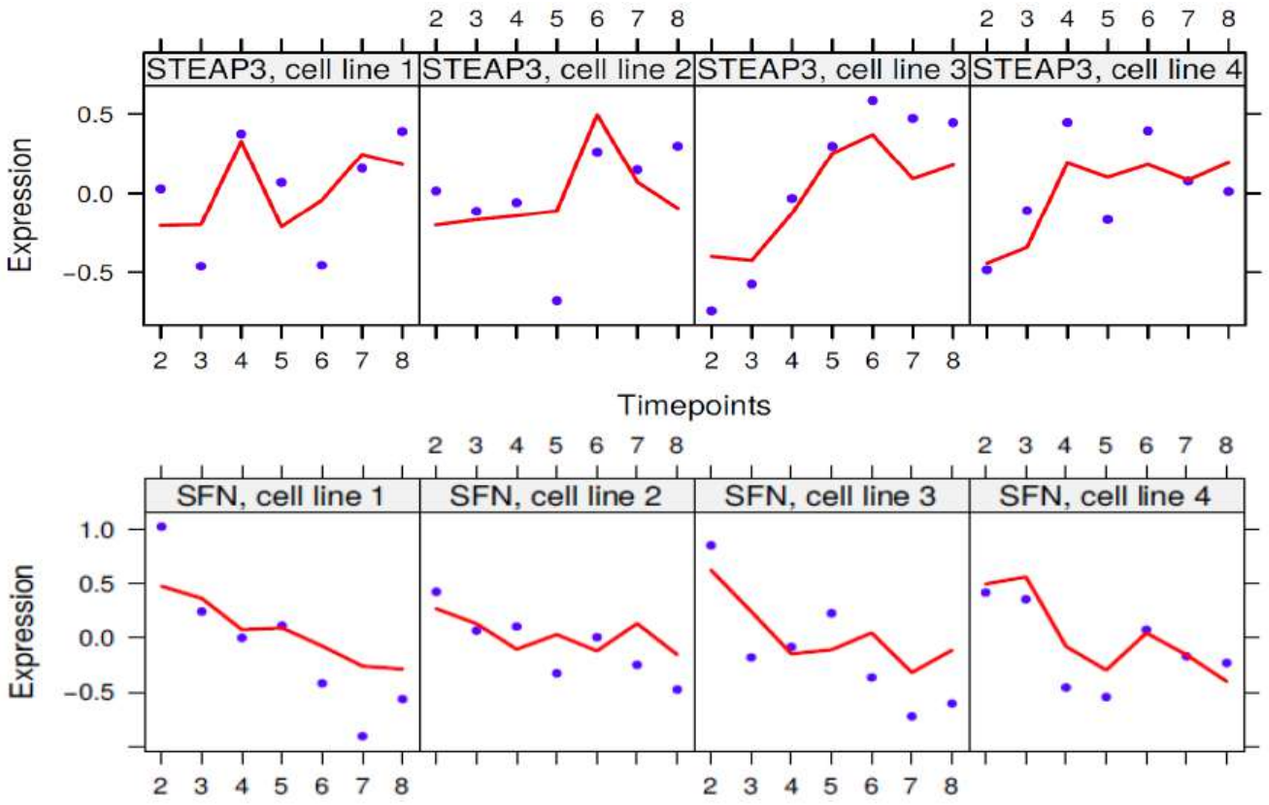
R-package

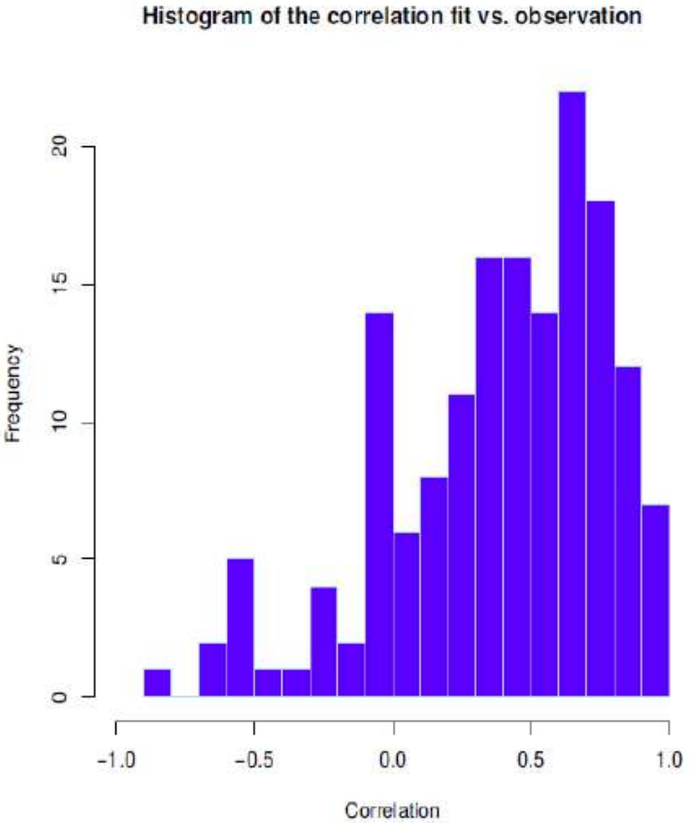
rags2ridges

(available from CRAN)



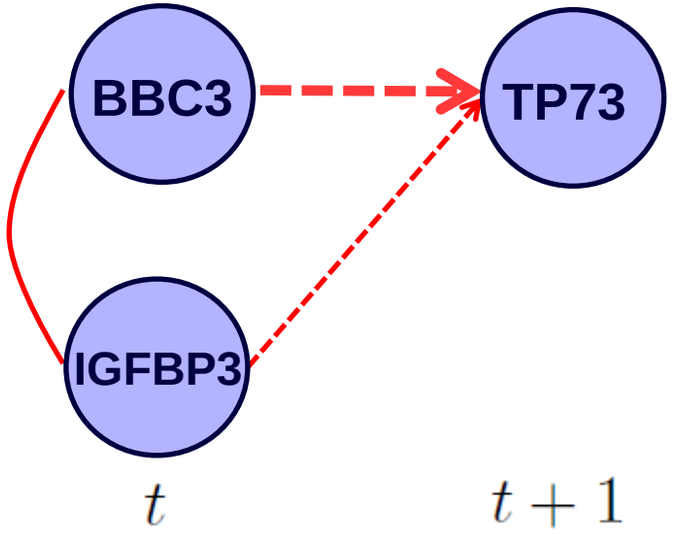
Diagnostics





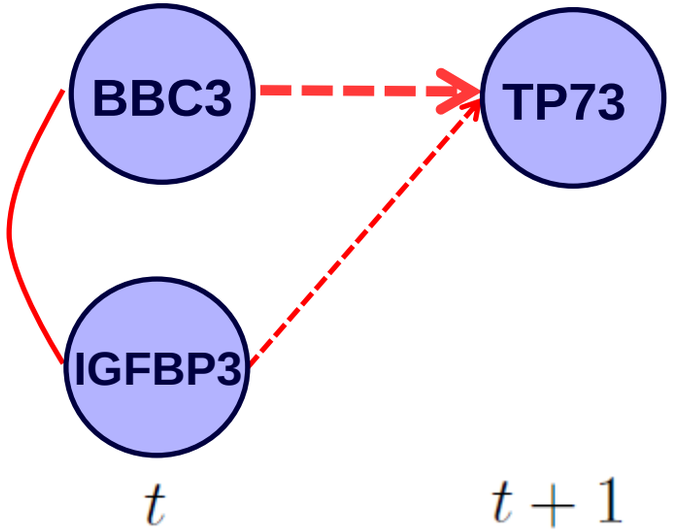
Path decomposition

Contribution between BBC3 and TP73: -0.003168001

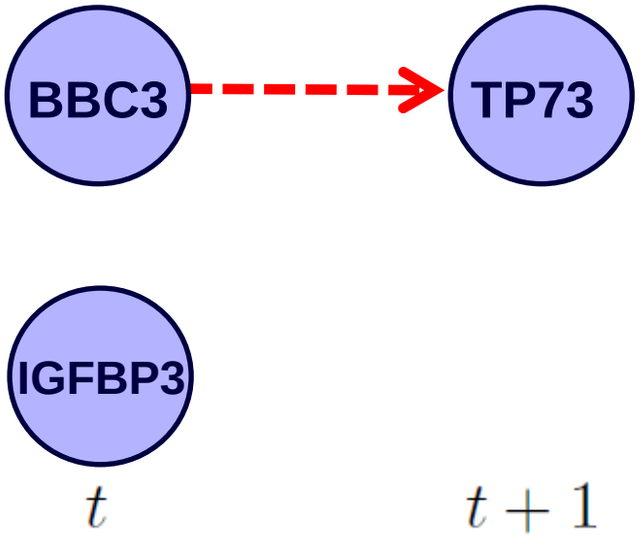


Path decomposition

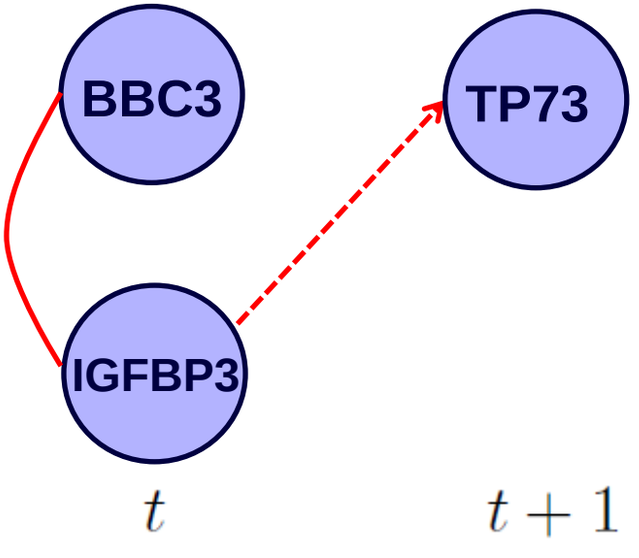
Contribution between BBC3 and TP73: -0.003168001



Contribution path 1: -0.002483485

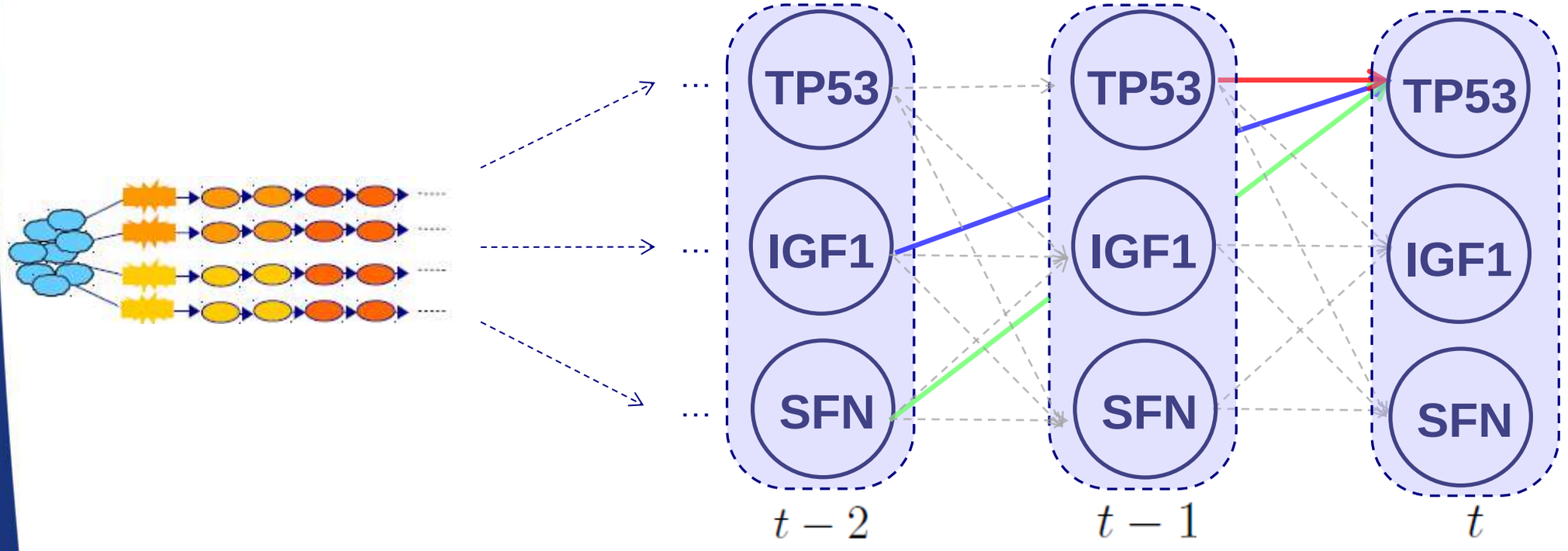


Contribution paths 2: -0.0006845158

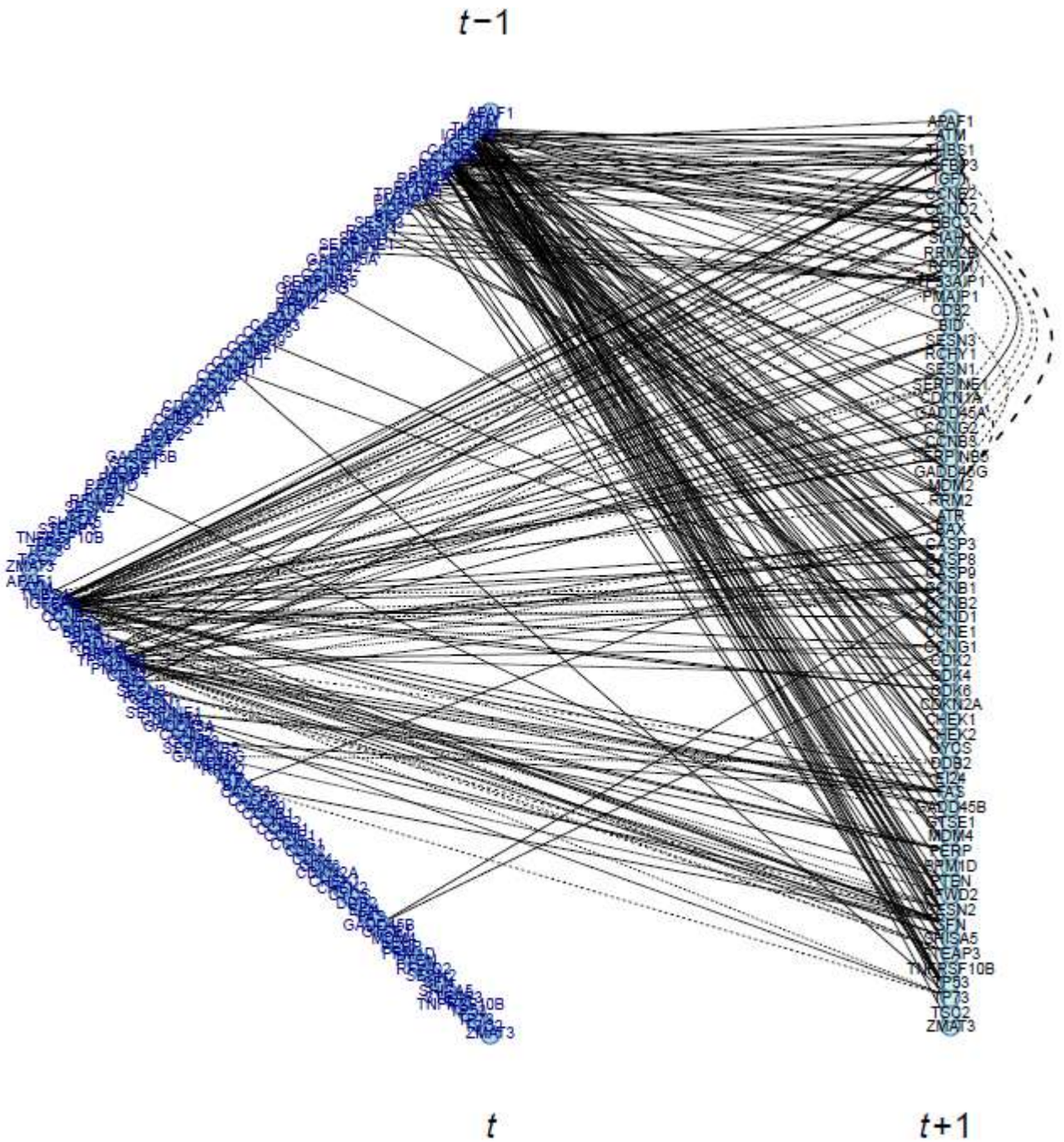


More explanatory time points?

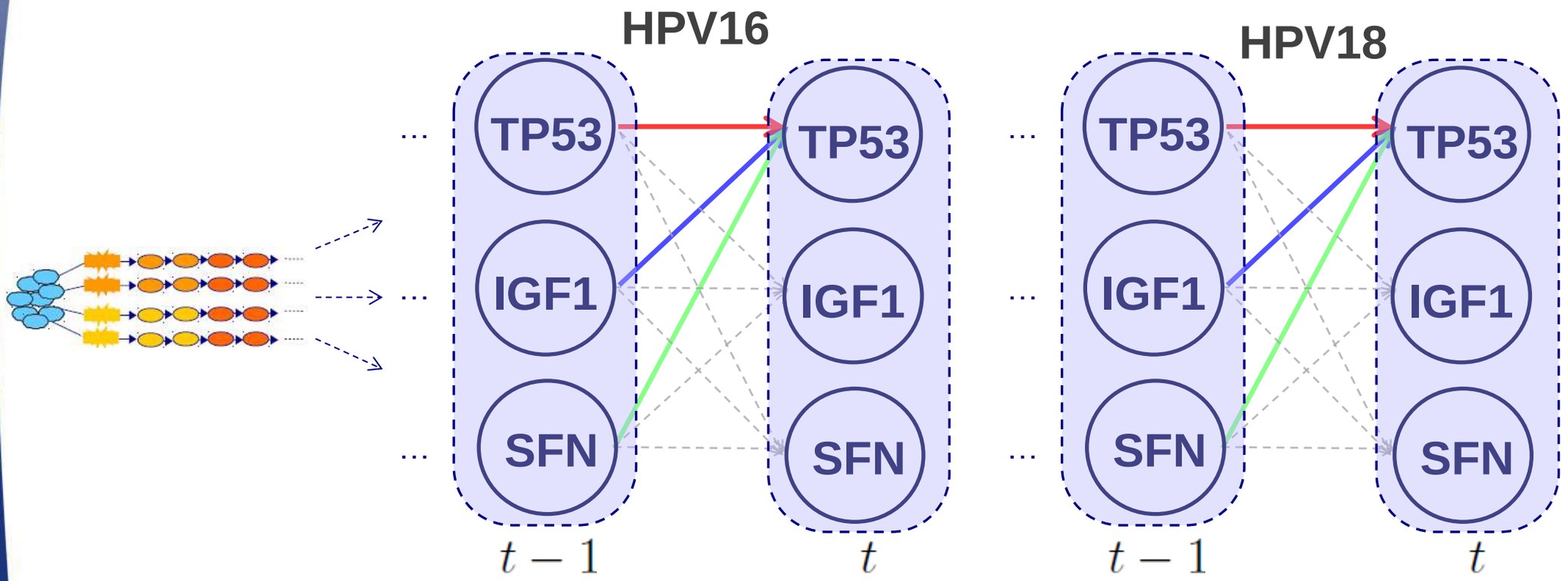
GE(t) = A1 * GE(t-1) + A2 * GE(t-2)



Network

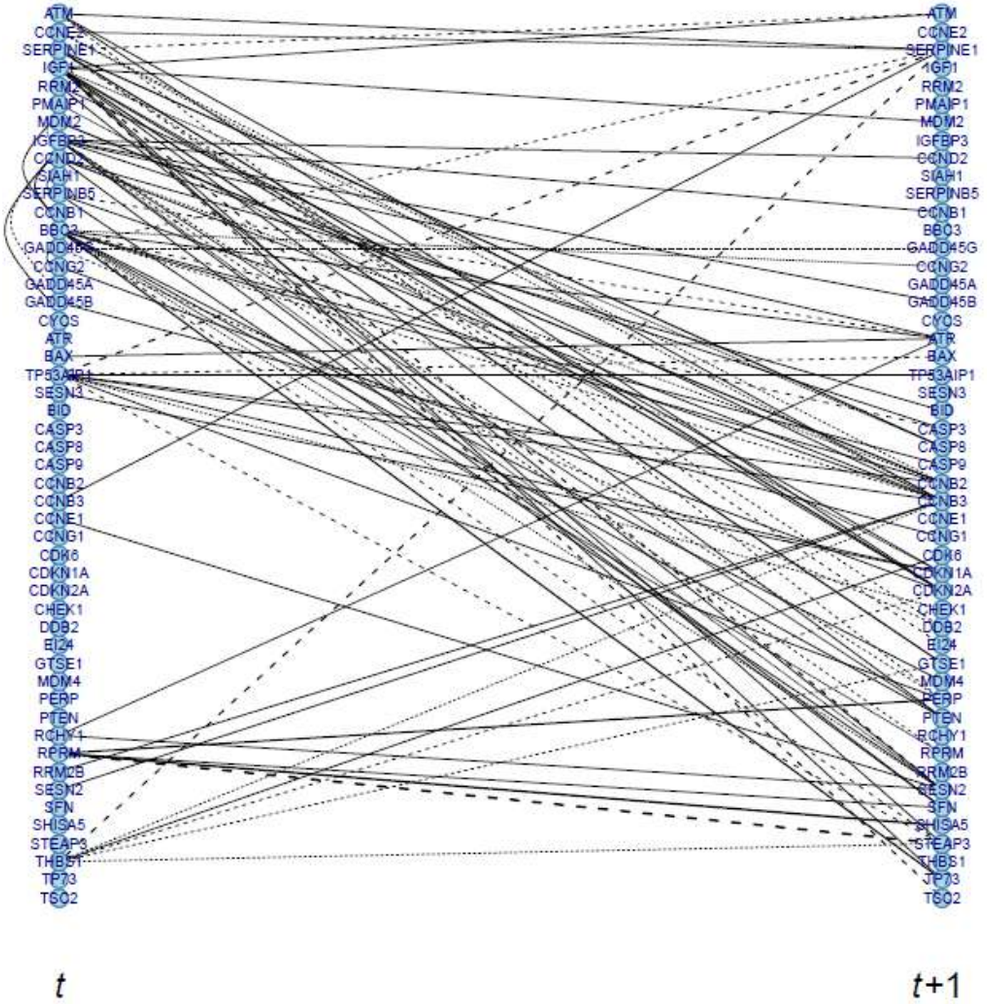


Identify the difference between groups

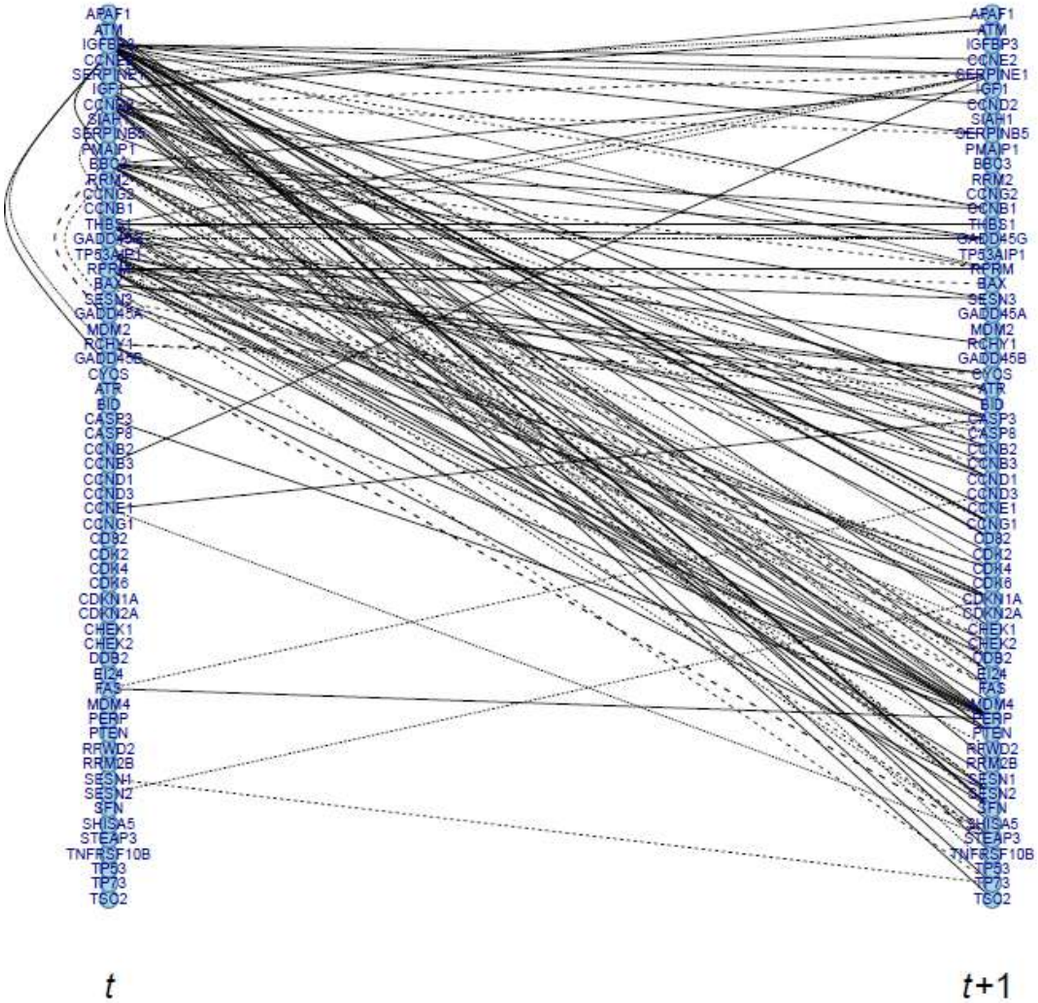


Network

HPV16



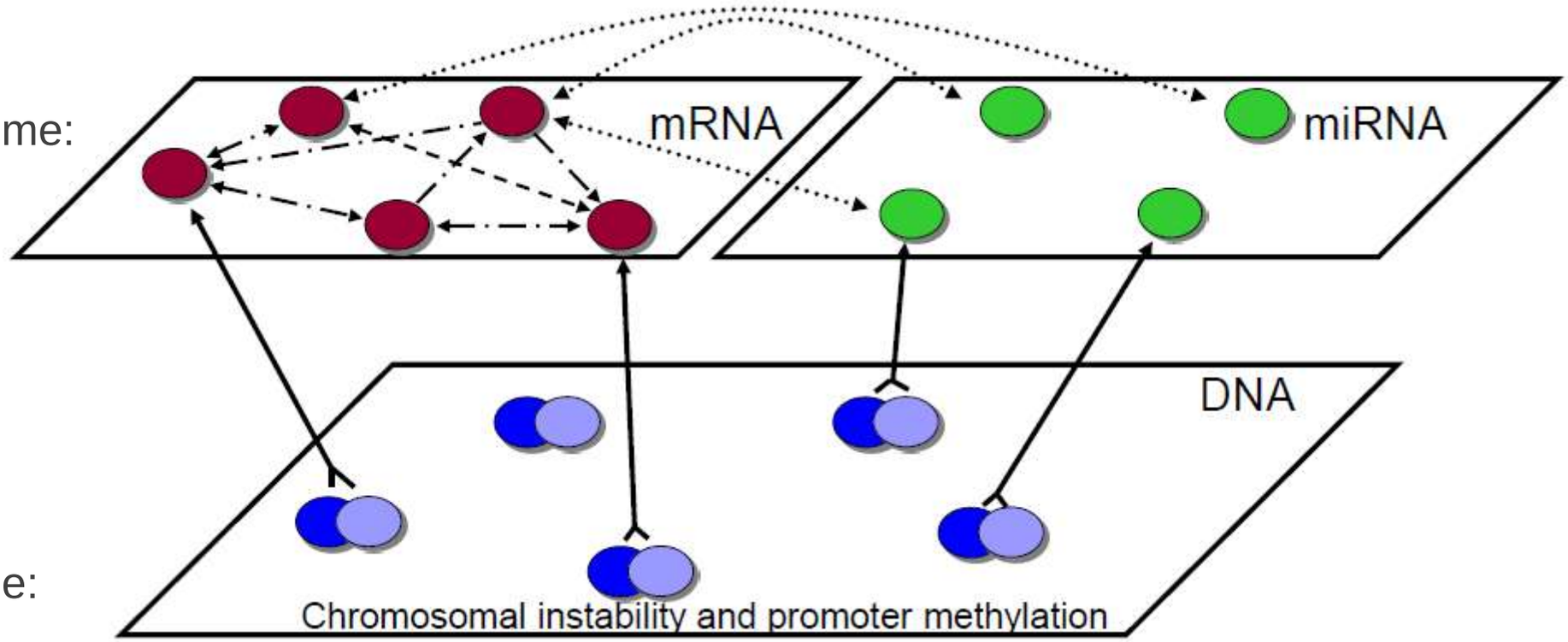
HPV18



Multi-omics data integration

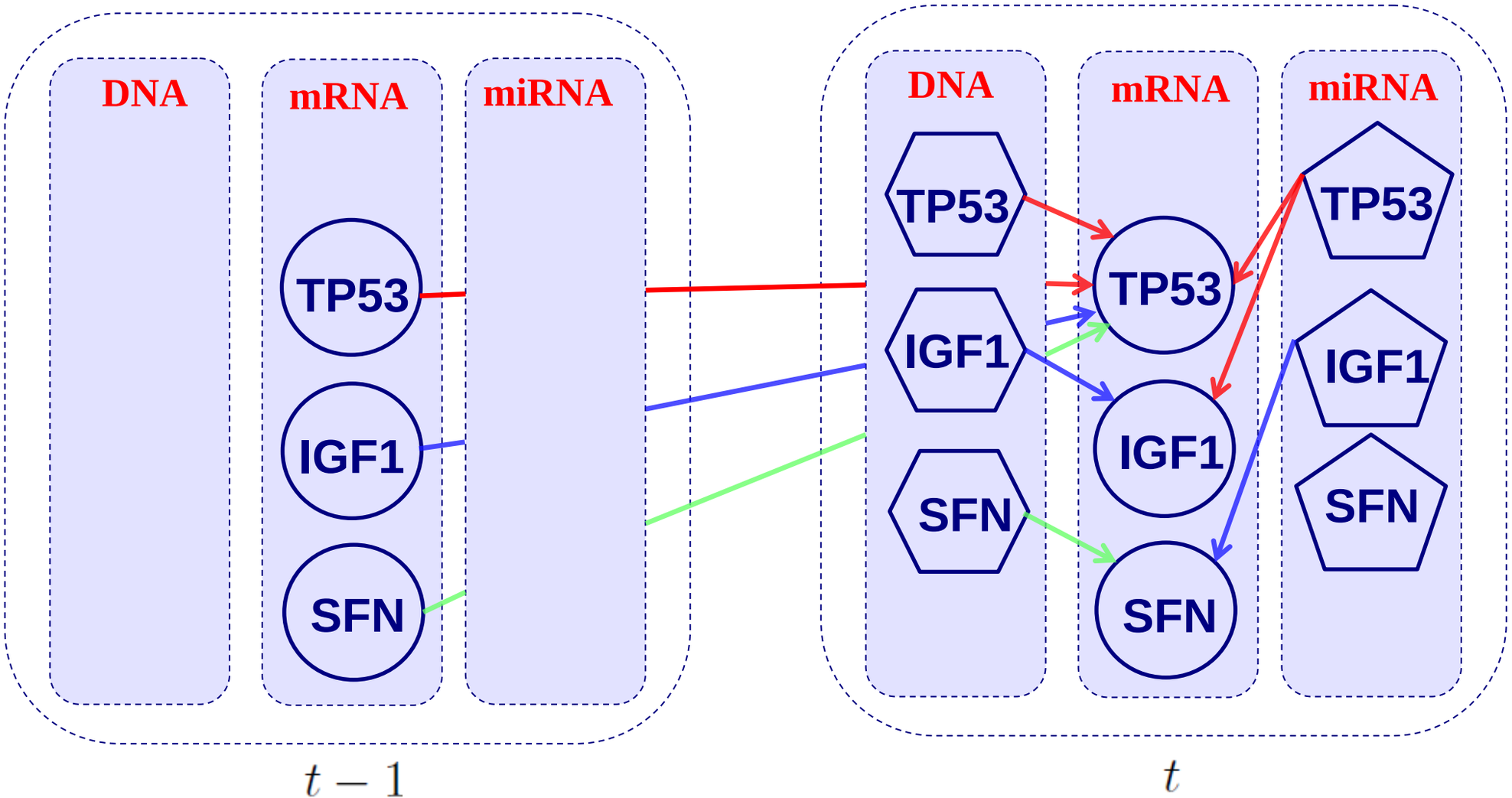
Transcriptome:

Genome:

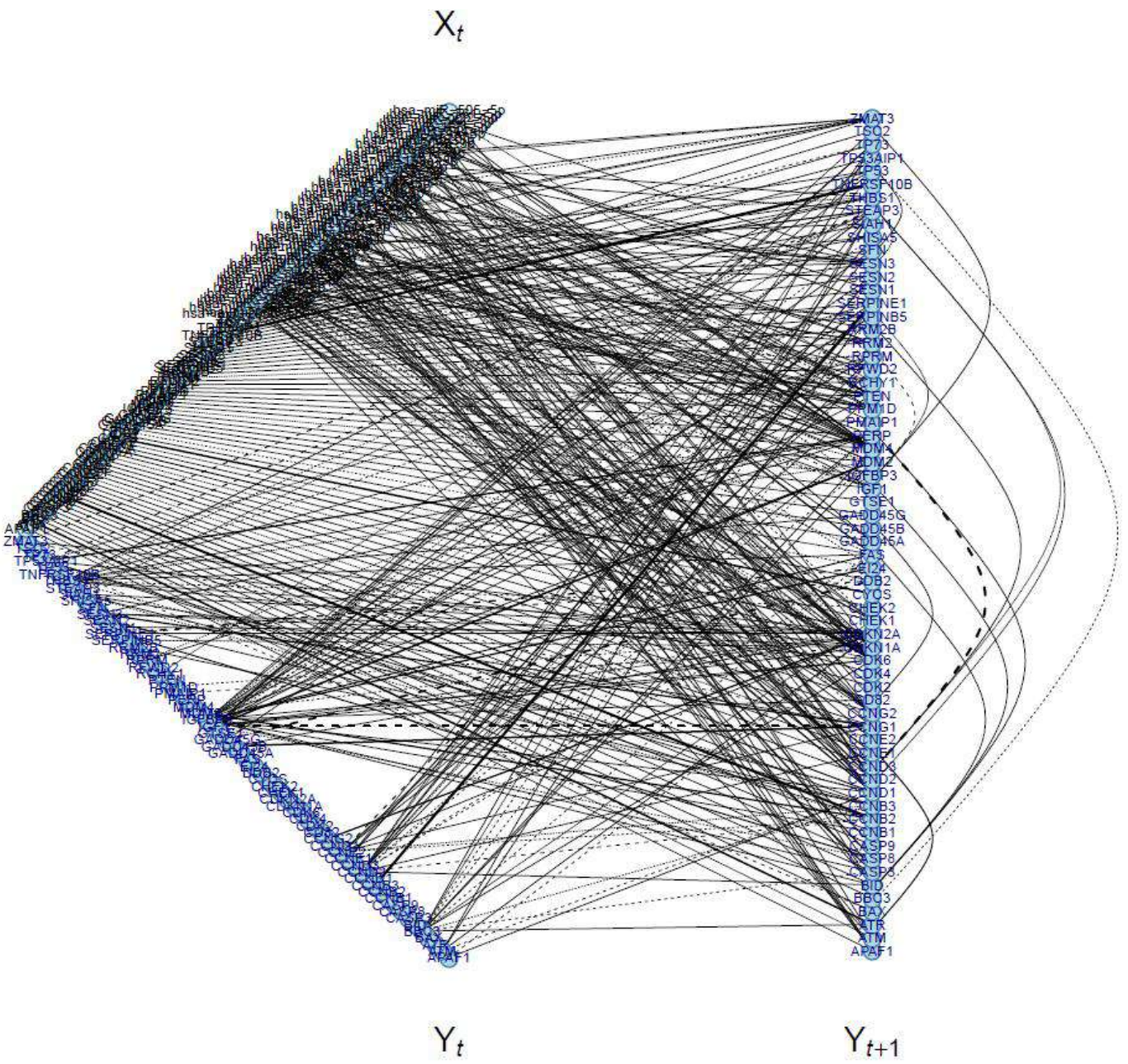


Integration of multi-level dynamic network

$$GE(t) = A1*GE(t-1) + B*[CN,MIR](t-2)$$



Integrated dynamic network



Thank you for your attention!