

WORKSHOP OF BIOINFORMATICS OF GENE REGULATION

A novel information theory-based method for detecting candidate
transcription factors/genes
predicts drivers of altered gene expression in cancer

Darius Wlochowicz

Institute of Bioinformatics,
University Medical Center Göttingen,
University of Göttingen
08.03.2018

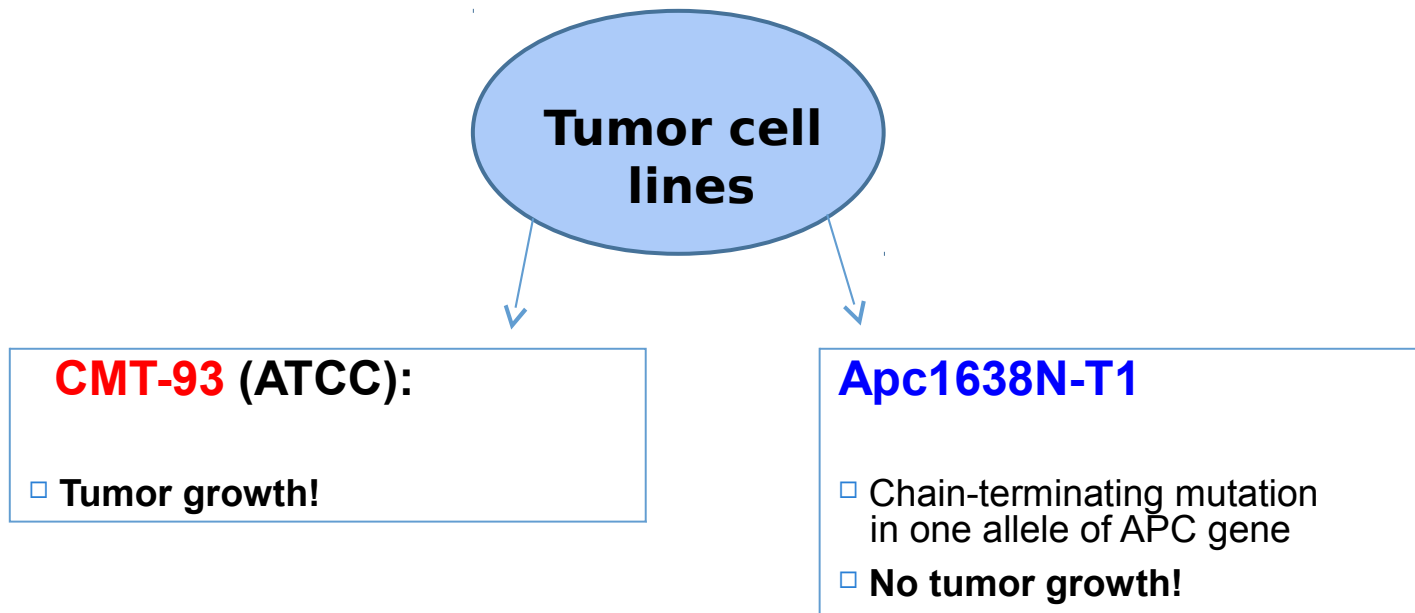
WORKSHOP OF BIOINFORMATICS OF GENE REGULATION

A novel information theory-based method for detecting candidate
transcription factors/genes
predicts drivers of altered gene expression in cancer

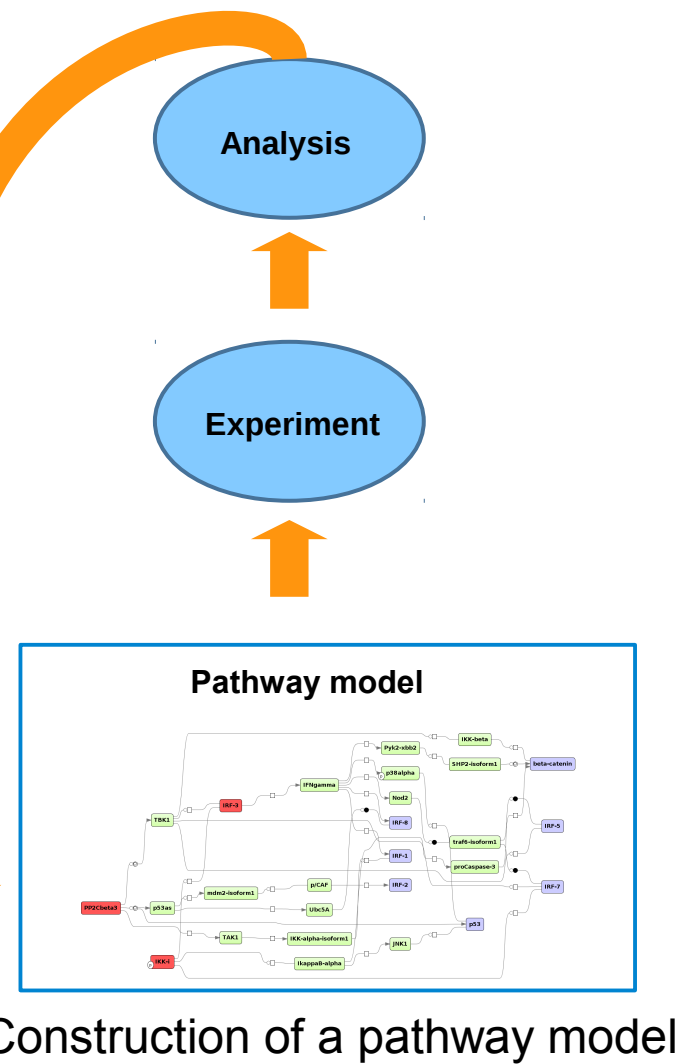
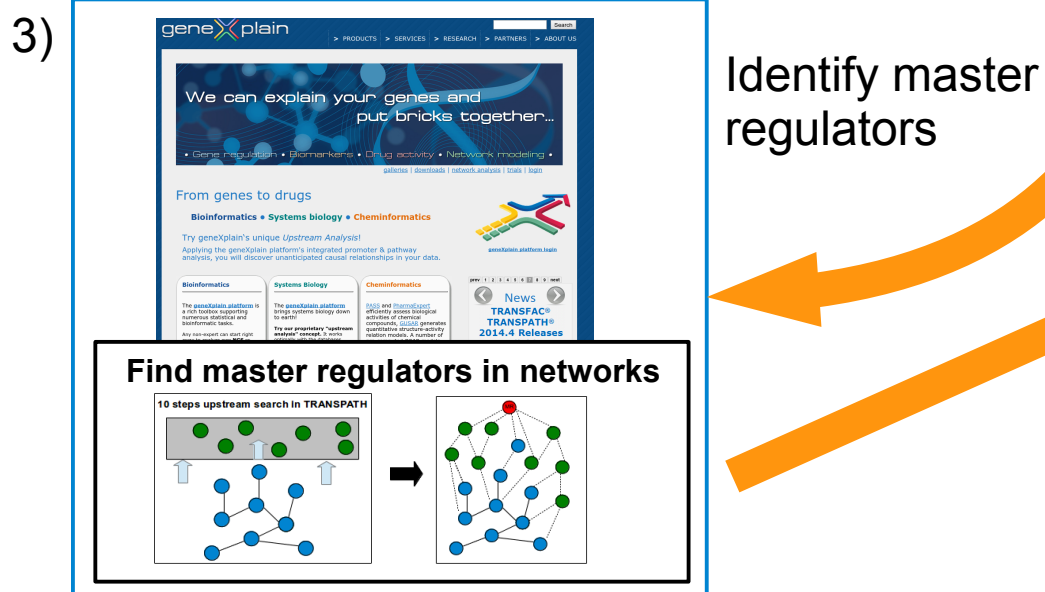
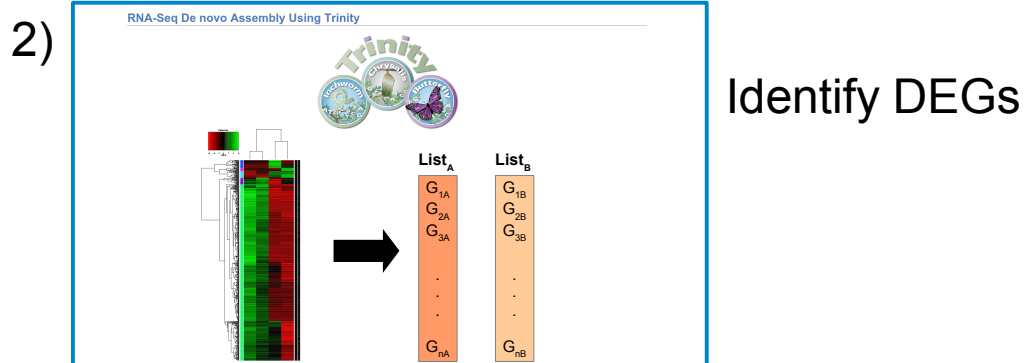
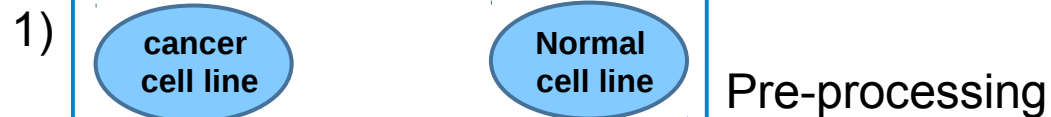
Darius Wlochowicz

Institute of Bioinformatics,
University Medical Center Göttingen,
University of Göttingen
08.03.2018

In silico comparison of different cancer cell lines



Pathway modeling



ORIGINAL RESEARCH ARTICLE

Front. Genet., 05 April 2016 | <https://doi.org/10.3389/fgene.2016.00042>




Download Article



Export citation

Computational Identification of Key Regulators in Two Different Colorectal Cancer Cell Lines

 Darius Wlochowitz^{1*},  Martin Haubrock¹,  Jetcy Arackal²,  Annalen Bleckmann²,  Alexander Wolff²,  Tim Beißbarth²,  Edgar Wingender¹ and  Mehmet Gültas^{1*}

¹Institute of Bioinformatics, University Medical Center Göttingen, Göttingen, Germany

²Department of Hematology/Medical Oncology, University Medical Center Göttingen, Göttingen, Germany

³Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

Transcription factors (TFs) are gene regulatory proteins that are essential for an effective regulation of the transcriptional machinery. Today, it is known that their expression plays an important role in several types of cancer. Computational identification of key players in specific cancer cell lines is still an open challenge in cancer research. In this study, we present a systematic approach which combines colorectal cancer (CRC) cell lines, namely 1638N-T1 and CMT-93, and well-established computational methods in order to compare these cell lines on the level of transcriptional regulation as well as on a pathway level, i.e., the cancer cell-intrinsic pathway repertoire. For this purpose, we firstly applied the Trinity platform to detect signature genes, and then applied analyses of the geneXplain platform to these for detection of upstream transcriptional regulators and their regulatory networks. We created a CRC-specific position weight matrix (PWM) library based on the TRANSFAC database (release 2014.1) to minimize the rate of false predictions in the promoter analyses. Using our proposed workflow, we specifically focused on revealing the similarities and differences in transcriptional regulation between the two CRC cell lines, and report a number of well-known, cancer-associated TFs with significantly enriched binding sites in the promoter regions of the signature genes. We show that, although the signature genes of both cell lines show no overlap, they may still be regulated by common TFs in CRC. Based on our findings, we suggest that canonical Wnt signaling is activated in 1638N-T1, but inhibited in CMT-93 through cross-talks of Wnt signaling with the VDR signaling pathway and/or LXR-related pathways. Furthermore, our findings provide indication of several master regulators being present such as MLK3 and Mapk1 (ERK2) which might be important in cell proliferation, migration, and invasion of 1638N-T1 and CMT-93, respectively. Taken together, we provide new insights into the invasive potential of these cell lines, which can be used for development of effective cancer therapy.

3,539

TOTAL VIEWS



 View Article Impact




Win \$100,000 to host your own conference.

[Submit your Research Topic](#)

SHARE ON



 Open Supplemental Data

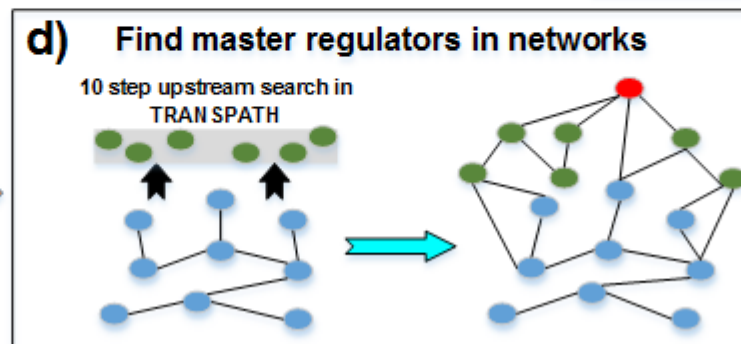
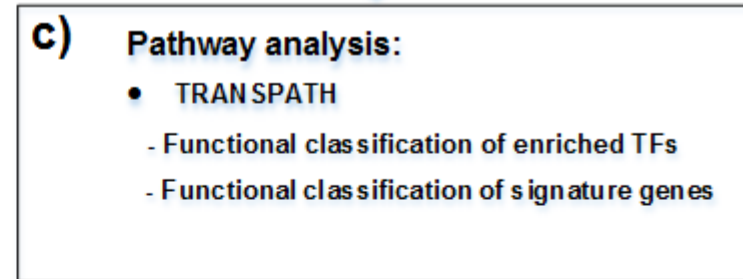
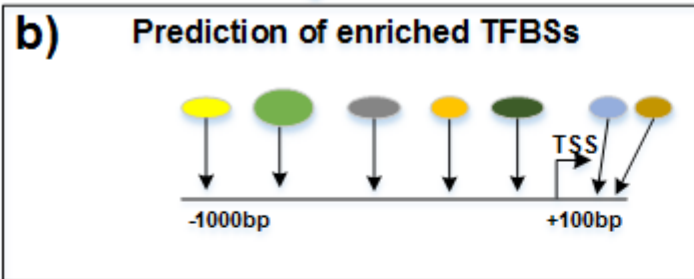
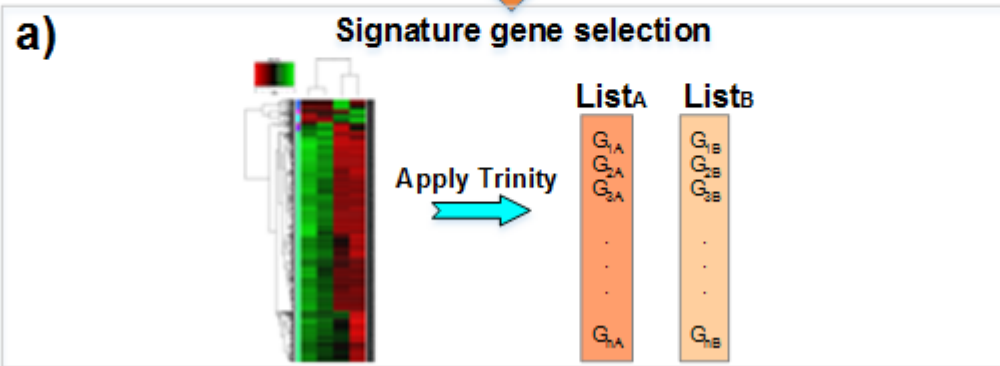
PEOPLE ALSO LOOKED AT

Editorial: Systems Biology of Transcription Regulation

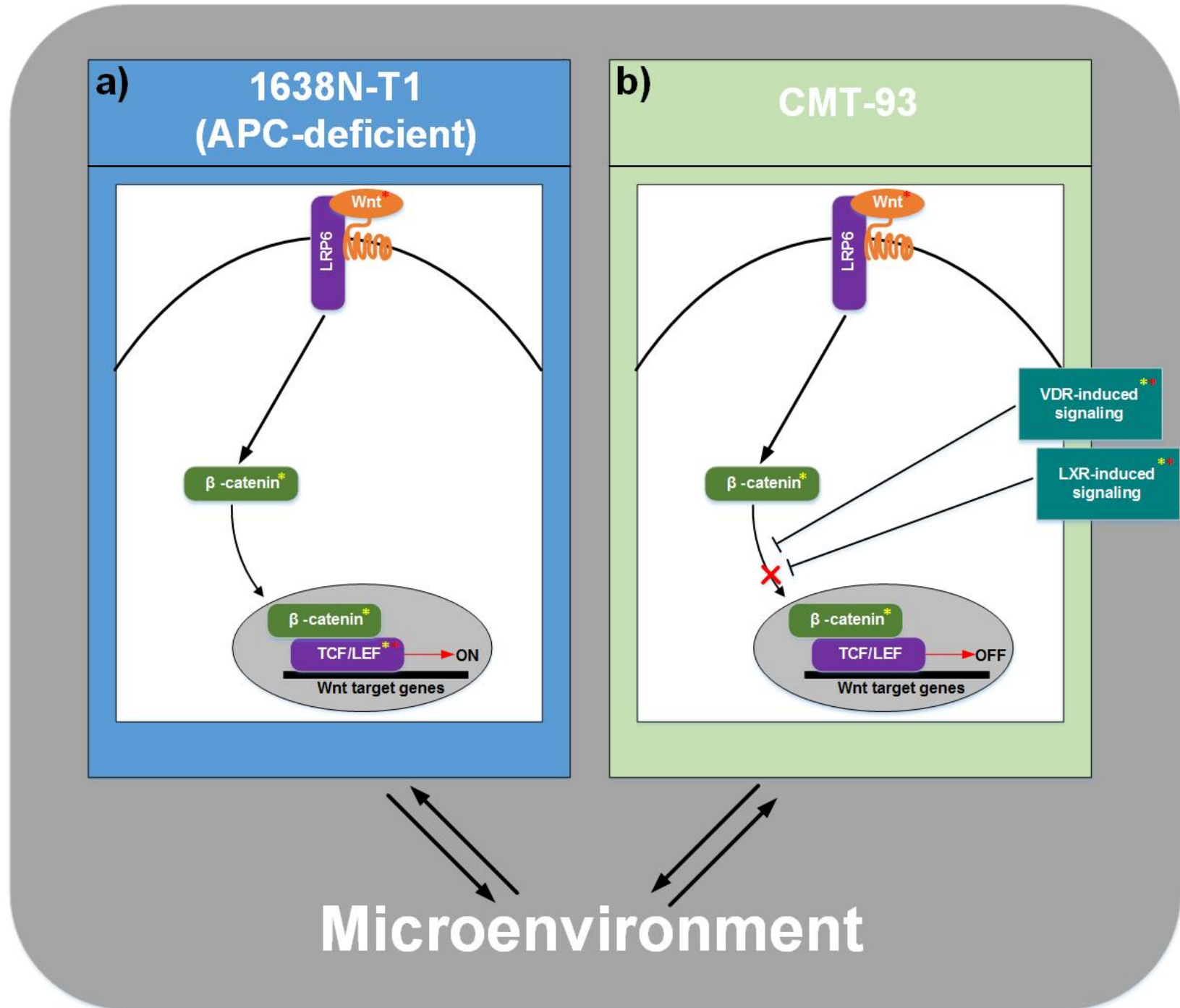
[Ekaterina Shelest and Edgar Wingender](#)

Follow-up analyses: APC vs CMT-93

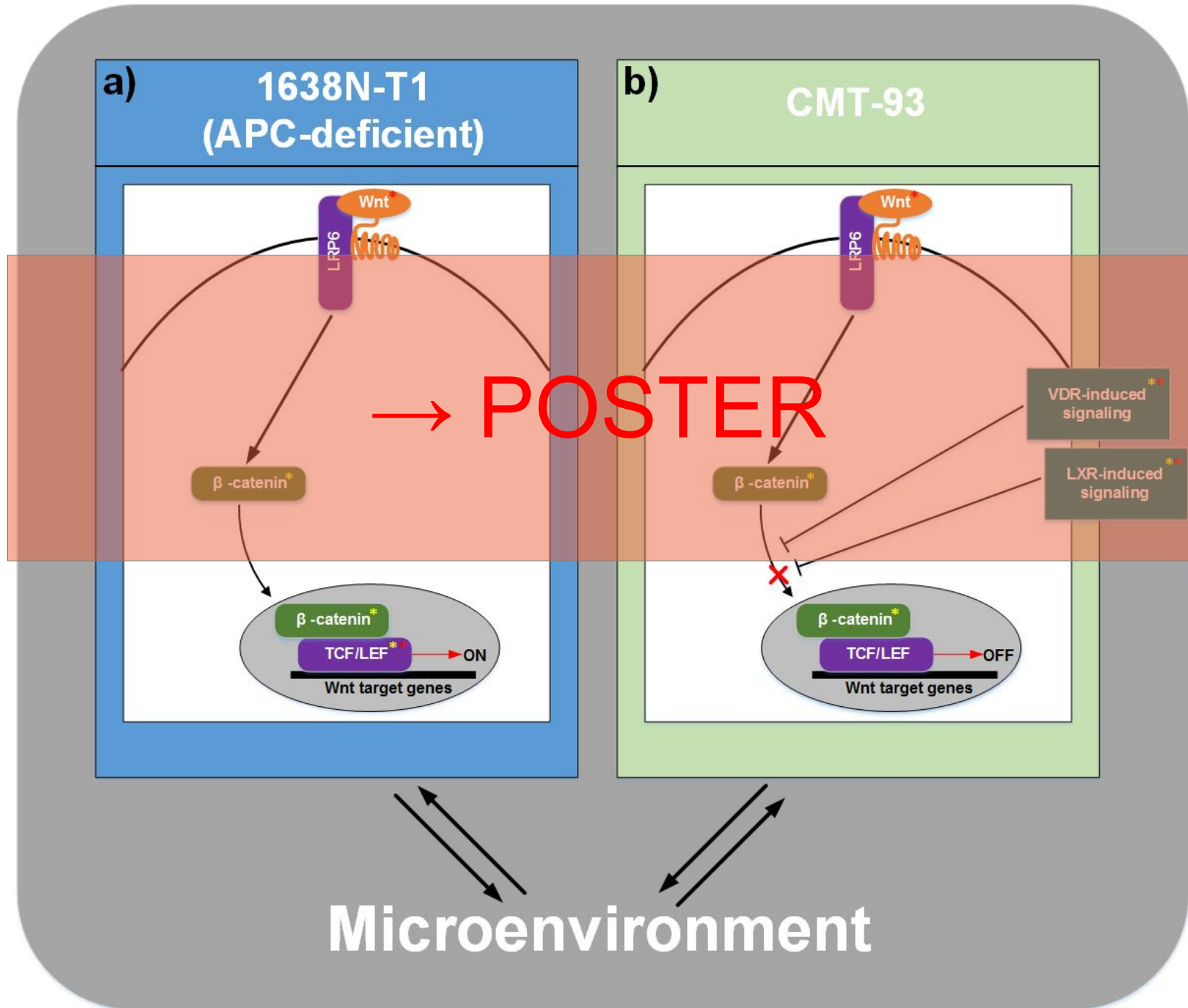
Colorectal cancer cell lines:
1638N-T1 vs. CMT-93



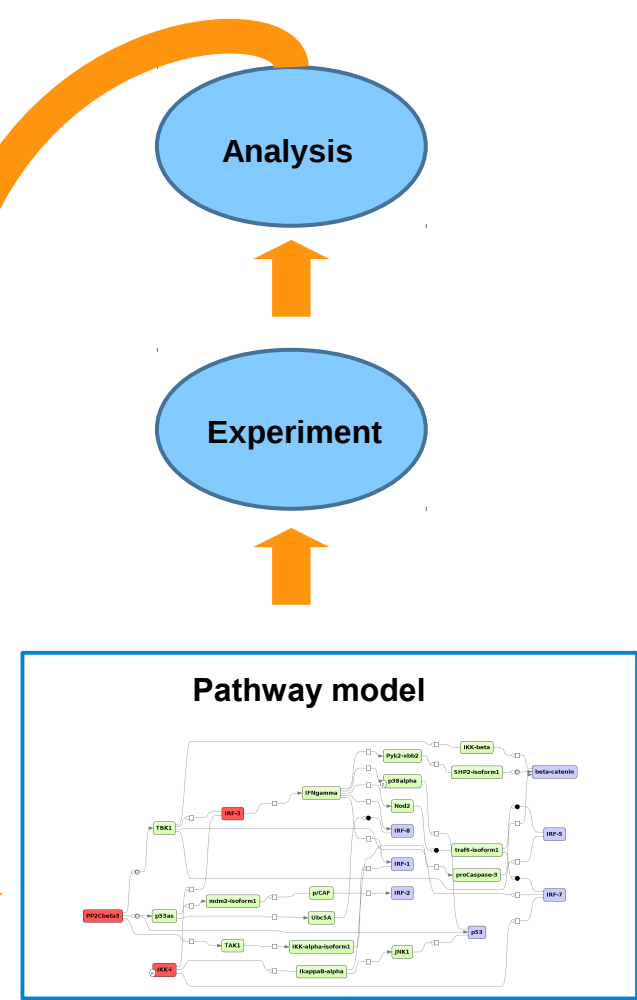
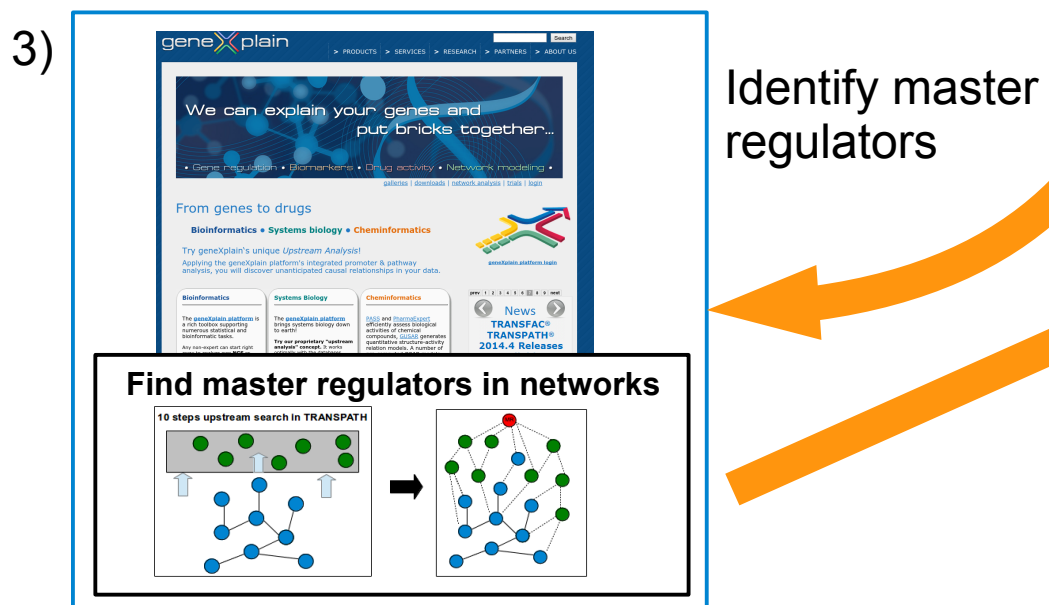
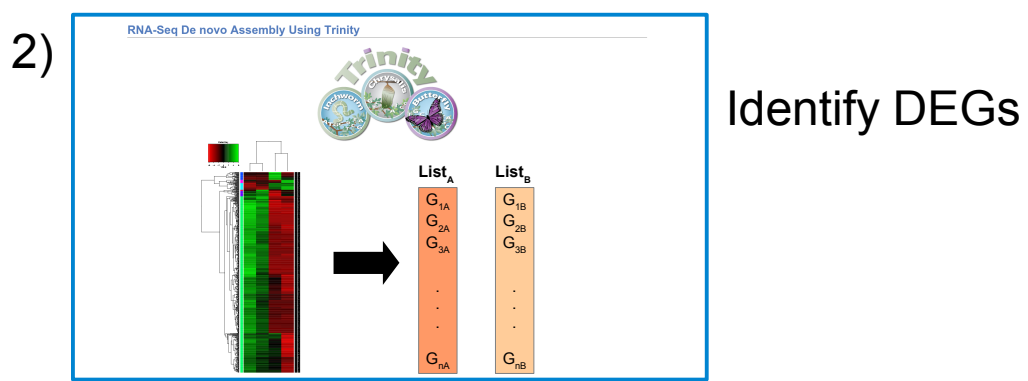
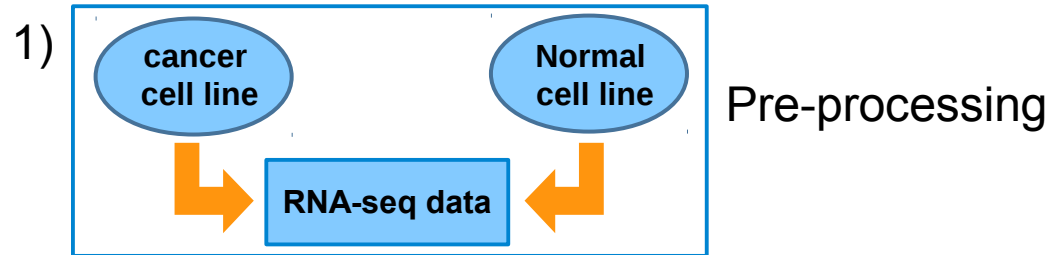
Pathway model: APC vs CMT-93



Pathway model: APC vs CMT-93

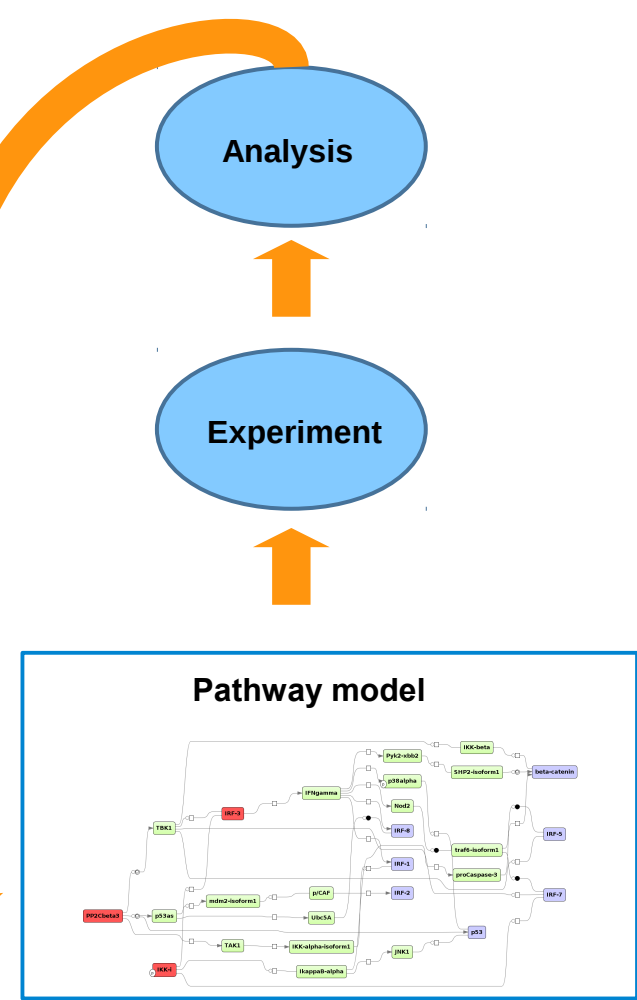
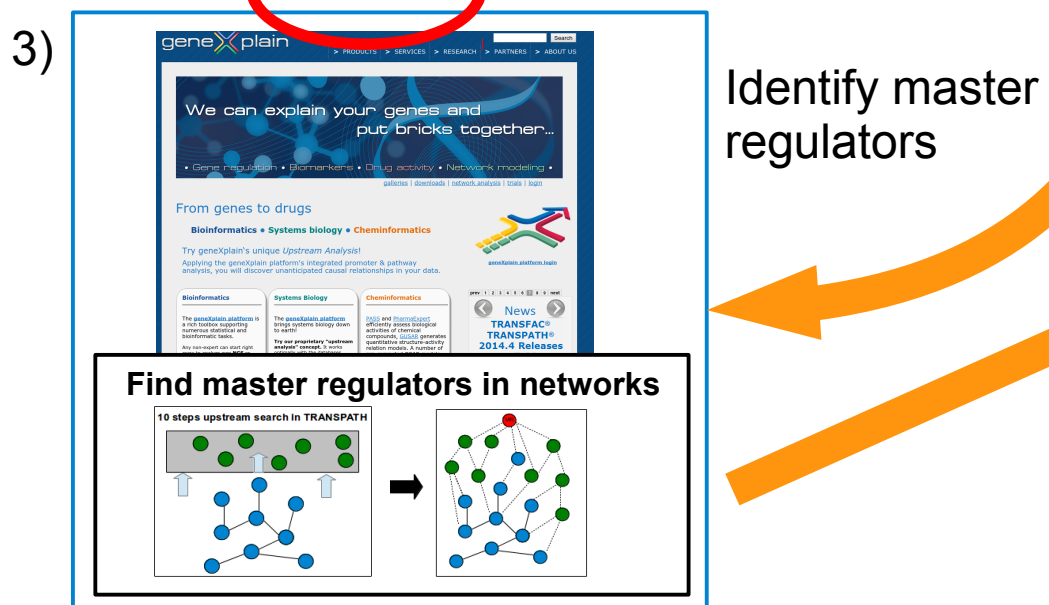
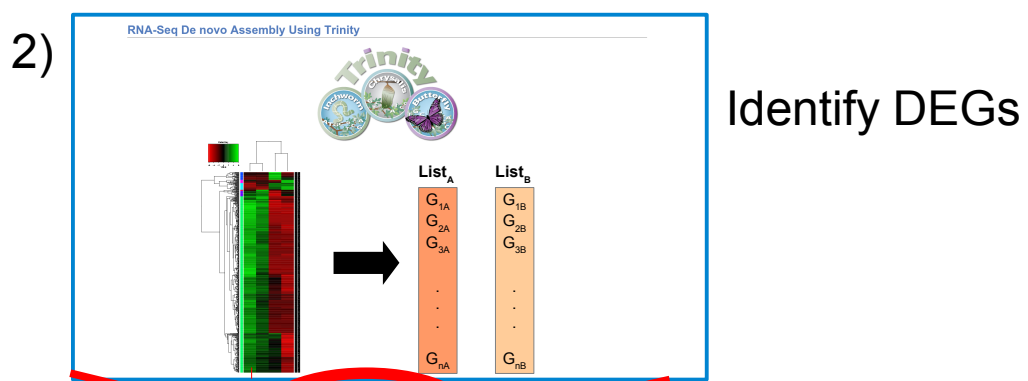
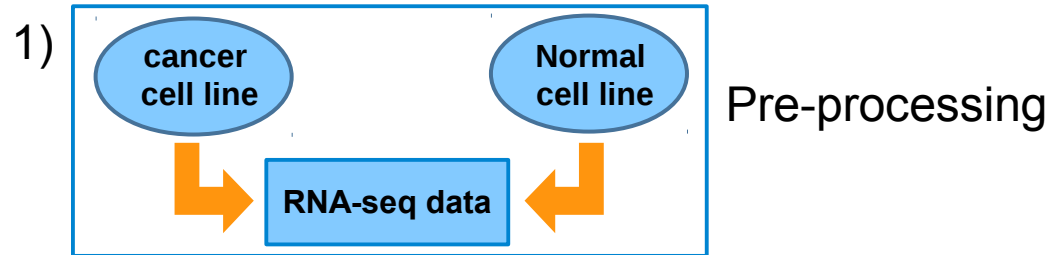


Pathway modeling: How can I improve my model?



Construction of a pathway model

Pathway modeling: How can I improve my model?



Construction of a pathway model

A novel information theory-based method: **GeneSpiker**

- **Concept:** Refine the input data set (DEG list) by identifying the most significantly differentially regulated genes (DRGs)

- **Main steps:**

1. MATCH™ analysis and construction of a transposed TFBS-sequence matrix with affinity-scaled frequency scores
2. Quantification of differences between the distribution of scores using the Jensen-Shannon divergence (JSD)
3. Identification of statistically significant JSD-values
4. Definition of DRGs

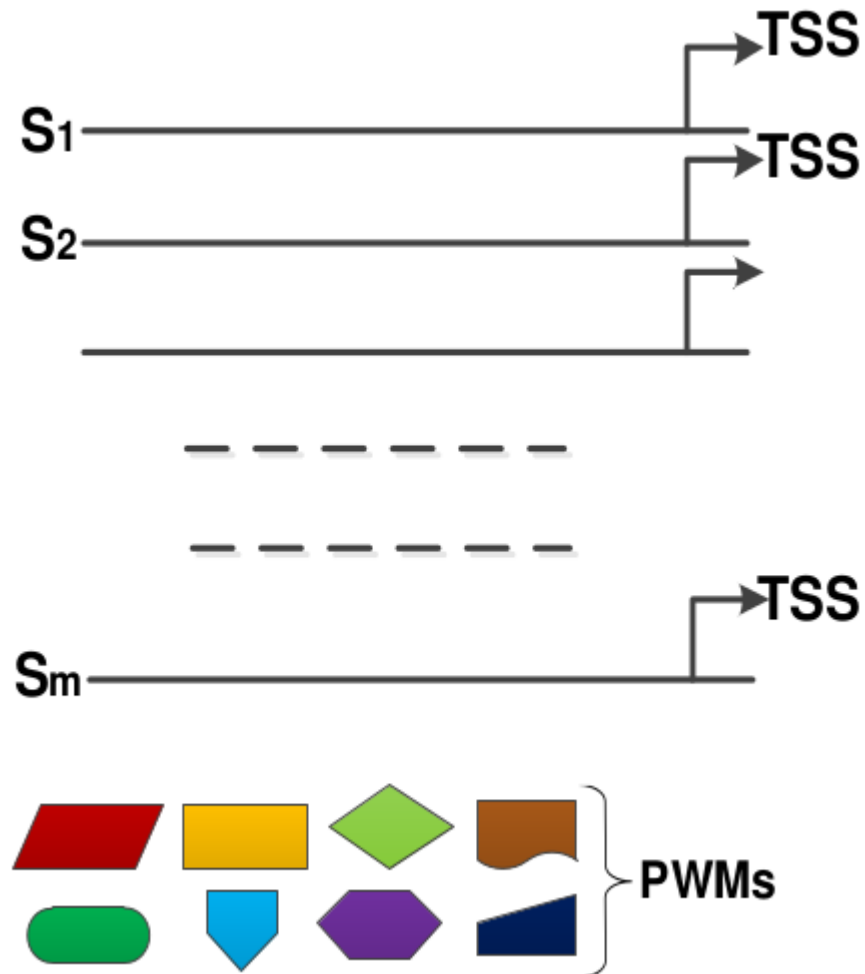
1. Construction of a transposed TFBS-sequence matrix

- MATCH™ analysis with TRANSFAC® PWMs



1. Construction of a transposed TFBS-sequence matrix

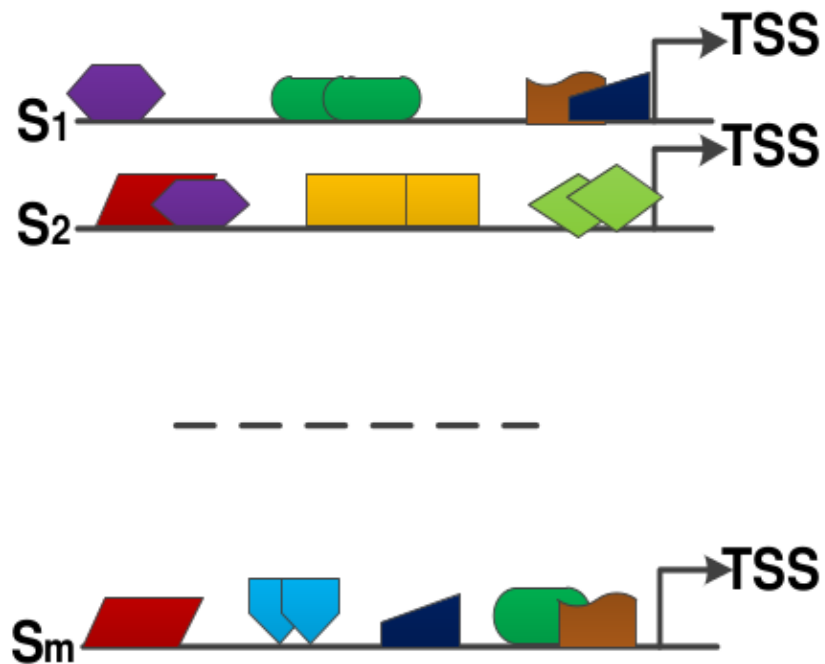
Let $S_i (i = 1, \dots, m)$ be a sequence



1. Construction of a transposed TFBS-sequence matrix

Let $S_i (i = 1, \dots, m)$ be a sequence

Let $t_j (j = 1, \dots, n)$ be a TFBS predicted using PWM j where n is the PWM library size



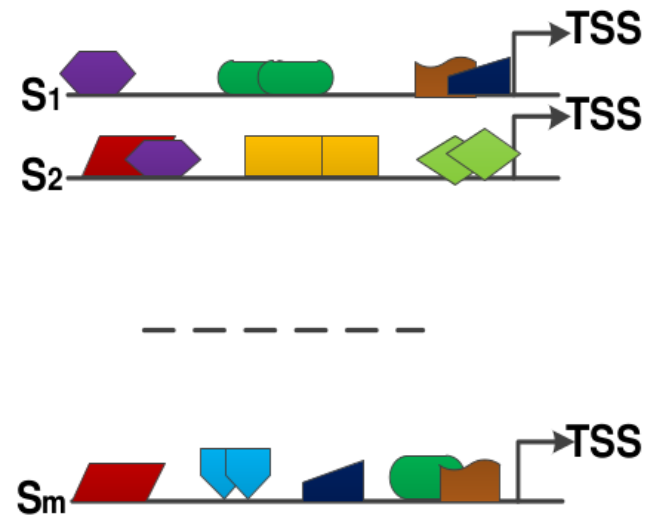
1. Construction of a transposed TFBS-sequence matrix

- Let $\mathbf{s}_i (i = 1, \dots, m)$ be a sequence
- Let $\mathbf{t}_j (j = 1, \dots, n)$ be a TFBS predicted using PWM j where n is the PWM library size

- Construction of the TFBS-based count matrix

$$M_{m,n} = \begin{matrix} & V_1 & V_2 & \dots & V_n \\ \begin{matrix} S_1 \\ S_2 \\ \vdots \\ S_m \end{matrix} & \begin{pmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,n} \\ f_{2,1} & f_{2,2} & \dots & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m,1} & f_{m,2} & \dots & f_{m,n} \end{pmatrix} \end{matrix}$$

⇒ The entry of M at position (i, j) , $f_{i,j}$, is the frequency of \mathbf{t}_j in \mathbf{s}_i



1. Construction of a transposed TFBS-sequence matrix

Concrete example: Matrix M

	PWM1	PWM2	PWM3	PWM4	PWM5	PWM6	PWM7	PWM8	PWM9	PWM10
Gene1	834	677	1621	886	2234	779	1657	965	965	1169
Gene2	4564	6434	5465	3233	7875	9655	8546	6666	8654	4888
Gene3	1621	688	897	1924	966	743	1054	1677	678	767

 transpose

	Gene1	Gene2	Gene3
PWM1	834	4564	1621
PWM2	677	6434	688
PWM3	1621	5465	897
PWM4	886	3233	1924
PWM5	2234	7875	966
PWM6	779	9655	743
PWM7	1657	8546	1054
PWM8	965	6666	1677
PWM9	965	8654	678
PWM10	1169	4888	767

2. Quantification of differences between the binding site distribution of TFs using the Jensen-Shannon divergence (JSD)

Matrix M

	Gene1	Gene2	Gene3
PWM1	834	4564	1621
PWM2	677	6434	688
PWM3	1621	5465	897
PWM4	886	3233	1924
PWM5	2234	7875	966
PWM6	779	9655	743
PWM7	1657	8546	1054
PWM8	965	6666	1677
PWM9	965	8654	678
PWM10	1169	4888	767

- Apply JSD metric to every column pair (k,l) in M

$$JSD(k||l) = H\left(\frac{p_k + p_l}{2}\right) - \frac{1}{2}H(p_k) - \frac{1}{2}H(p_l)$$

- Determine significant JSD values between column pairs

- An *entropy* is a measure of the average uncertainty of an outcome.

- Let X be a discrete random variable with probabilities $p(x_i)$, $i = 1, \dots, n$
⇒ The Shannon entropy is defined by

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

- $H(X) \geq 0$ entropy is always non-negative

2. Quantification of differences between the binding site distribution of TFs using the Jensen-Shannon divergence (JSD)

Matrix M

	Gene1	Gene2	Gene3
PWM1	834	4564	1621
PWM2	677	6434	688
PWM3	1621	5465	897
PWM4	886	3233	1924
PWM5	2234	7875	966
PWM6	779	9655	743
PWM7	1657	8546	1054
PWM8	965	6666	1677
PWM9	965	8654	678
PWM10	1169	4888	767



1. Gene1 - Gene2

- Apply JSD metric to every column pair (k,l) in M

$$JSD(k||l) = H\left(\frac{p_k + p_l}{2}\right) - \frac{1}{2}H(p_k) - \frac{1}{2}H(p_l)$$

- Determine significant JSD values between column pairs

- An *entropy* is a measure of the average uncertainty of an outcome.

- Let X be a discrete random variable with probabilities $p(x_i)$, $i = 1, \dots, n$
⇒ The Shannon entropy is defined by

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

- $H(X) \geq 0$ entropy is always non-negative

2. Quantification of differences between the binding site distribution of TFs using the Jensen-Shannon divergence (JSD)

Matrix M

	Gene1	Gene2	Gene3
PWM1	834	4564	1621
PWM2	677	6434	688
PWM3	1621	5465	897
PWM4	886	3233	1924
PWM5	2234	7875	966
PWM6	779	9655	743
PWM7	1657	8546	1054
PWM8	965	6666	1677
PWM9	965	8654	678
PWM10	1169	4888	767



2. Gene1 - Gene3

- Apply JSD metric to every column pair (k,l) in M

$$JSD(k||l) = H\left(\frac{p_k + p_l}{2}\right) - \frac{1}{2}H(p_k) - \frac{1}{2}H(p_l)$$

- Determine significant JSD values between column pairs

- An *entropy* is a measure of the average uncertainty of an outcome.

- Let X be a discrete random variable with probabilities $p(x_i)$, $i = 1, \dots, n$
⇒ The Shannon entropy is defined by

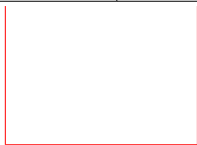
$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

- $H(X) \geq 0$ entropy is always non-negative

2. Quantification of differences between the binding site distribution of TFs using the Jensen-Shannon divergence (JSD)

Matrix M

	Gene1	Gene2	Gene3
PWM1	834	4564	1621
PWM2	677	6434	688
PWM3	1621	5465	897
PWM4	886	3233	1924
PWM5	2234	7875	966
PWM6	779	9655	743
PWM7	1657	8546	1054
PWM8	965	6666	1677
PWM9	965	8654	678
PWM10	1169	4888	767



3. Gene2 - Gene3

- Apply JSD metric to every column pair (k,l) in M

$$JSD(k||l) = H\left(\frac{p_k + p_l}{2}\right) - \frac{1}{2}H(p_k) - \frac{1}{2}H(p_l)$$

- Determine significant JSD values between column pairs

- An *entropy* is a measure of the average uncertainty of an outcome.

- Let X be a discrete random variable with probabilities $p(x_i)$, $i = 1, \dots, n$
⇒ The Shannon entropy is defined by

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

- $H(X) \geq 0$ entropy is always non-negative

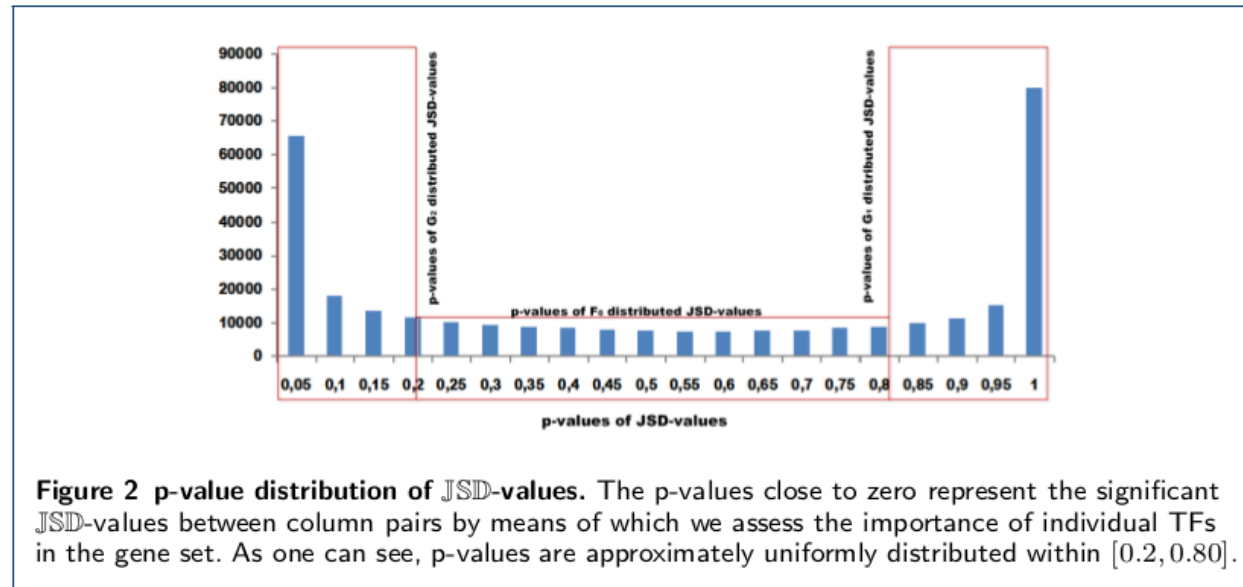
3. Identification of statistically significant JSD-values

- Application of the Jensen-Shannon divergence (JSD) results in

$$\frac{n(n-1)}{2} \text{JSD}(t_a, t_b)\text{-values}$$

,where n = number of genes

- Identification of significant **JSD values** (JSD^{sig})



4. Definition of DRGs

Matrix M

	Gene1	Gene2	Gene3
PWM1	834	4564	1621
PWM2	677	6434	688
PWM3	1621	5465	897
PWM4	886	3233	1924
PWM5	2234	7875	966
PWM6	779	9655	743
PWM7	1657	8546	1054
PWM8	965	6666	1677
PWM9	965	8654	678
PWM10	1169	4888	767

- Apply JSD metric to every column pair (k,l) in M

$$JSD(k||l) = H\left(\frac{p_k + p_l}{2}\right) - \frac{1}{2}H(p_k) - \frac{1}{2}H(p_l)$$

- Determine significant JSD values between column pairs

- An *entropy* is a measure of the average uncertainty of an outcome.

- Let X be a discrete random variable with probabilities $p(x_i)$, $i = 1, \dots, n$
⇒ The Shannon entropy is defined by

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

- $H(X) \geq 0$ entropy is always non-negative

Gene2 exhibits the most different distribution of affinity-scaled frequency scores of the three genes

Gene2 is the most (significantly) differentially regulated gene (DRG)

Advanced Computational Biology Methods Identify Molecular Switches for Malignancy in an EGF Mouse Model of Liver Cancer

Philip Stegmaier, Nico Voss, Tatiana Meier, Alexander Kel, Edgar Wingender, Juergen Borlak 

Published: March 28, 2011 • <https://doi.org/10.1371/journal.pone.0017738>

Article	Authors	Metrics	Comments	Related Content
---------	---------	---------	----------	-----------------

Abstract

Introduction

Results

Discussion

Materials and Methods

Supporting Information

Author Contributions

References

Reader Comments (0)

Media Coverage (0)

Figures

Abstract

The molecular causes by which the epidermal growth factor receptor tyrosine kinase induces malignant transformation are largely unknown. To better understand EGFs' transforming capacity whole genome scans were applied to a transgenic mouse model of liver cancer and subjected to advanced methods of computational analysis to construct de novo gene regulatory networks based on a combination of sequence analysis and entrained graph-topological algorithms. Here we identified transcription factors, processes, key nodes and molecules to connect as yet unknown interacting partners at the level of protein-DNA interaction. Many of those could be confirmed by electromobility band shift assay at recognition sites of gene specific promoters and by western blotting of nuclear proteins. A novel cellular regulatory circuitry could therefore be proposed that connects cell cycle regulated genes with components of the EGF signaling pathway. Promoter analysis of differentially expressed genes suggested the majority of regulated transcription factors to display specificity to either the pre-tumor or the tumor state. Subsequent search for signal transduction key nodes upstream of the identified transcription factors and their targets suggested the insulin-like growth factor pathway to render the tumor cells independent of EGF receptor activity. Notably, expression of IGF2 in addition to many components of this pathway was highly upregulated in tumors. Together, we propose a switch in autocrine signaling to foster tumor growth that was initially triggered by EGF and demonstrate the knowledge gain from promoter analysis combined with upstream key node identification.

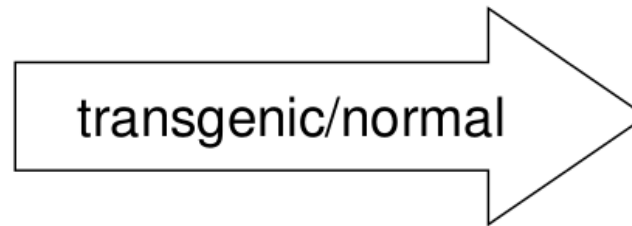
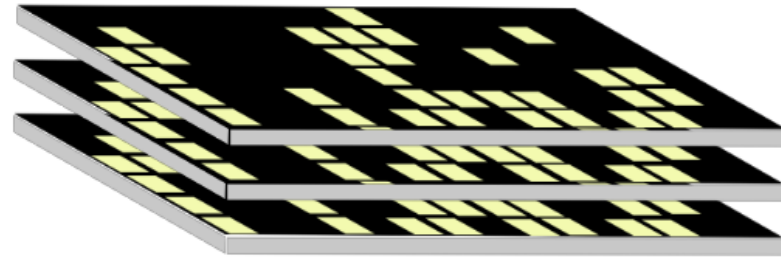
Epidermal Growth Factor induced Carcinogenicity

Philip Stegmaier¹, Alexander Kel¹, Edgar Wingender^{1,2}, and Jürgen Borlak³

Hepatocellular transcriptome data of IgEGF-overexpressing mice



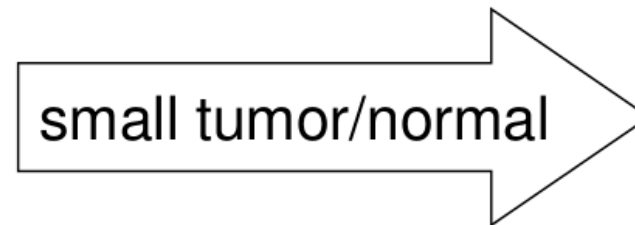
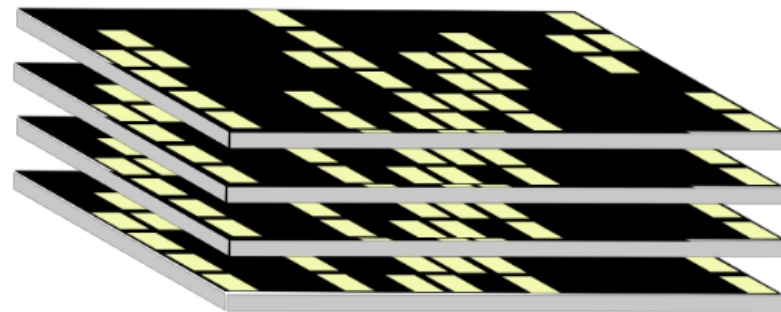
transgenic



Tumoregenic
switch



small tumor



“Walking pathways” and how promoters can help to find new drugs.

(Practical guide to multi-omics and multi-scale data integration)

Alexander Kel

Biosoft.ru, Skolkovo
Moscow

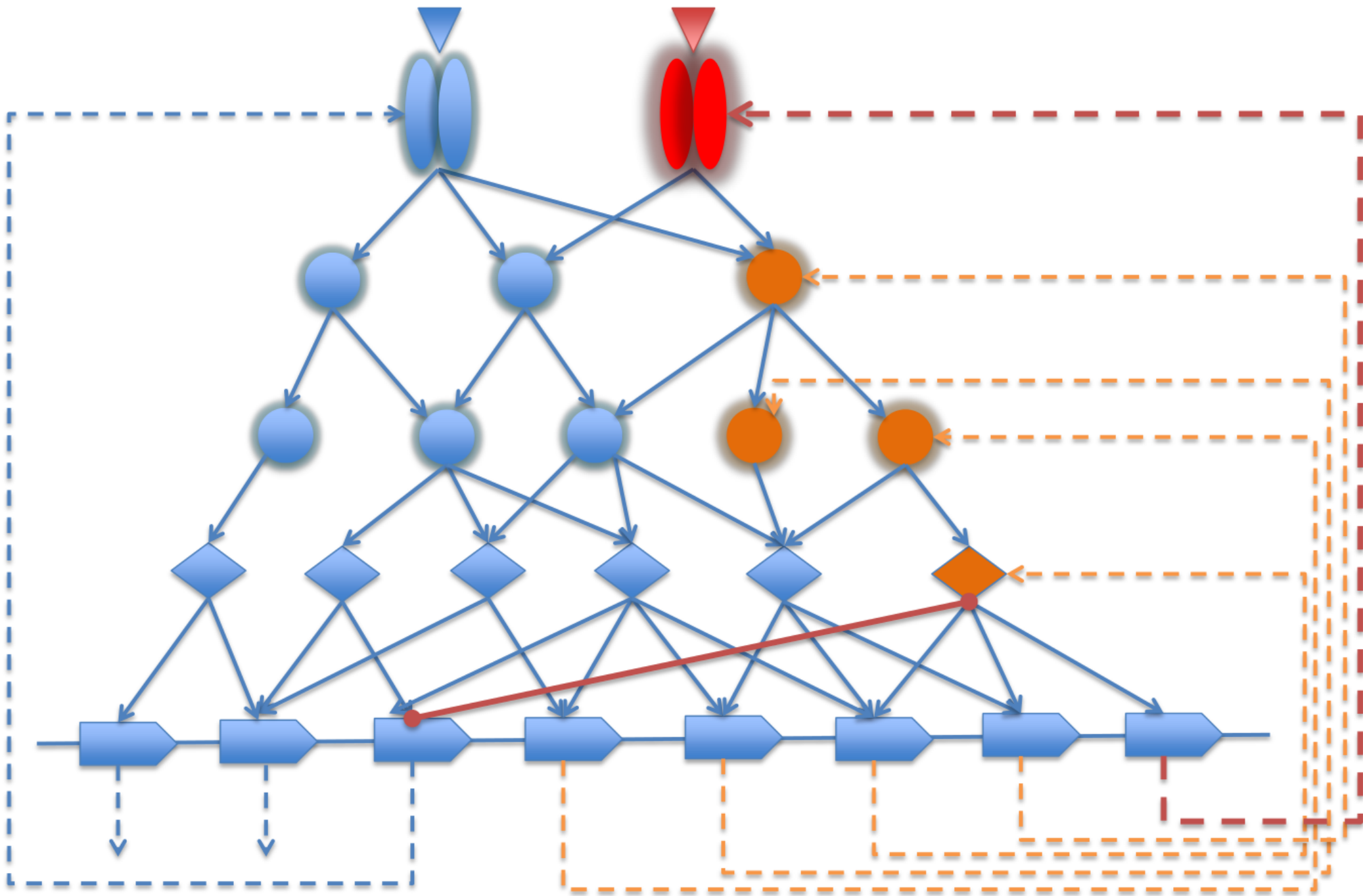


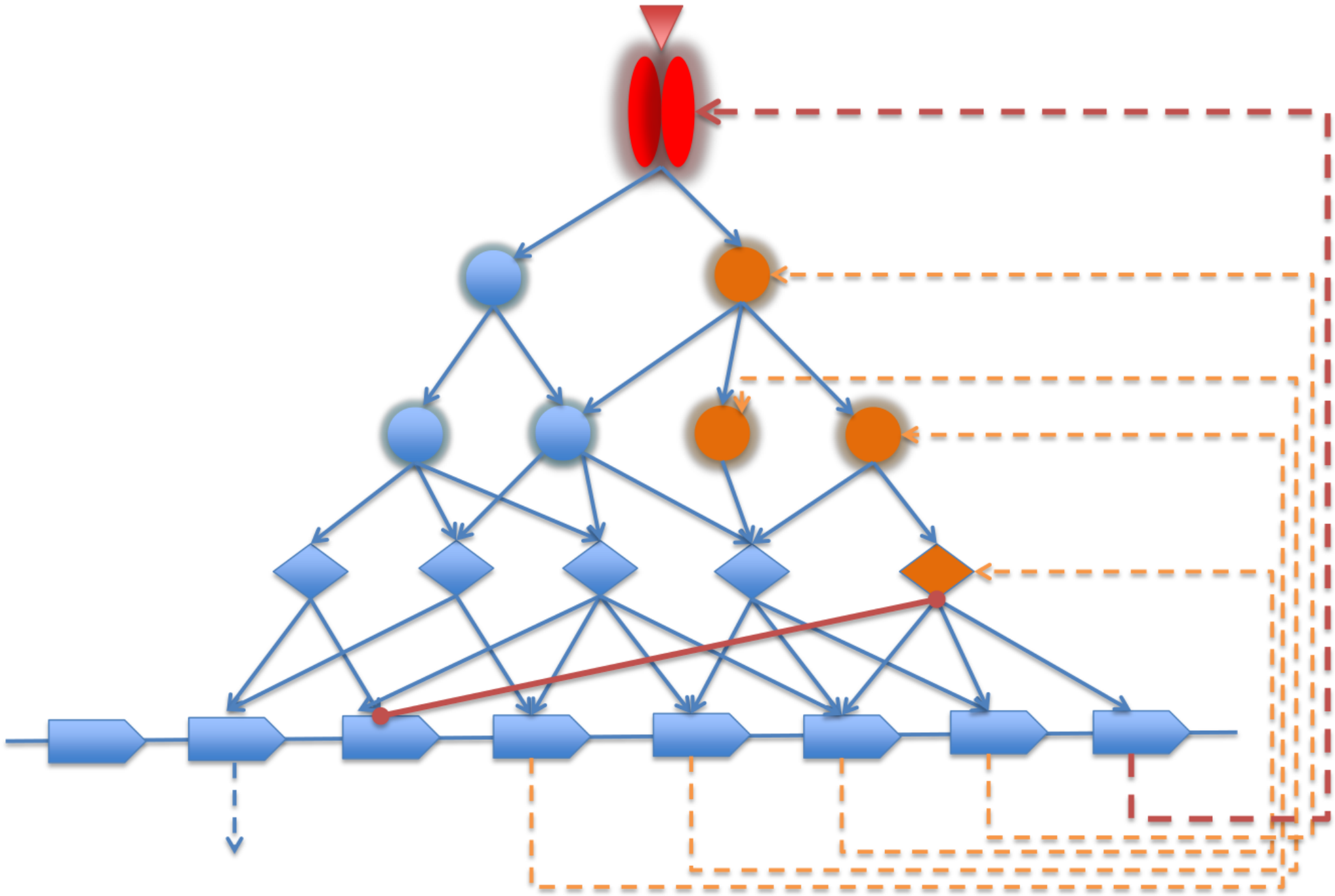
0100100010011101
Institute of Systems Biology

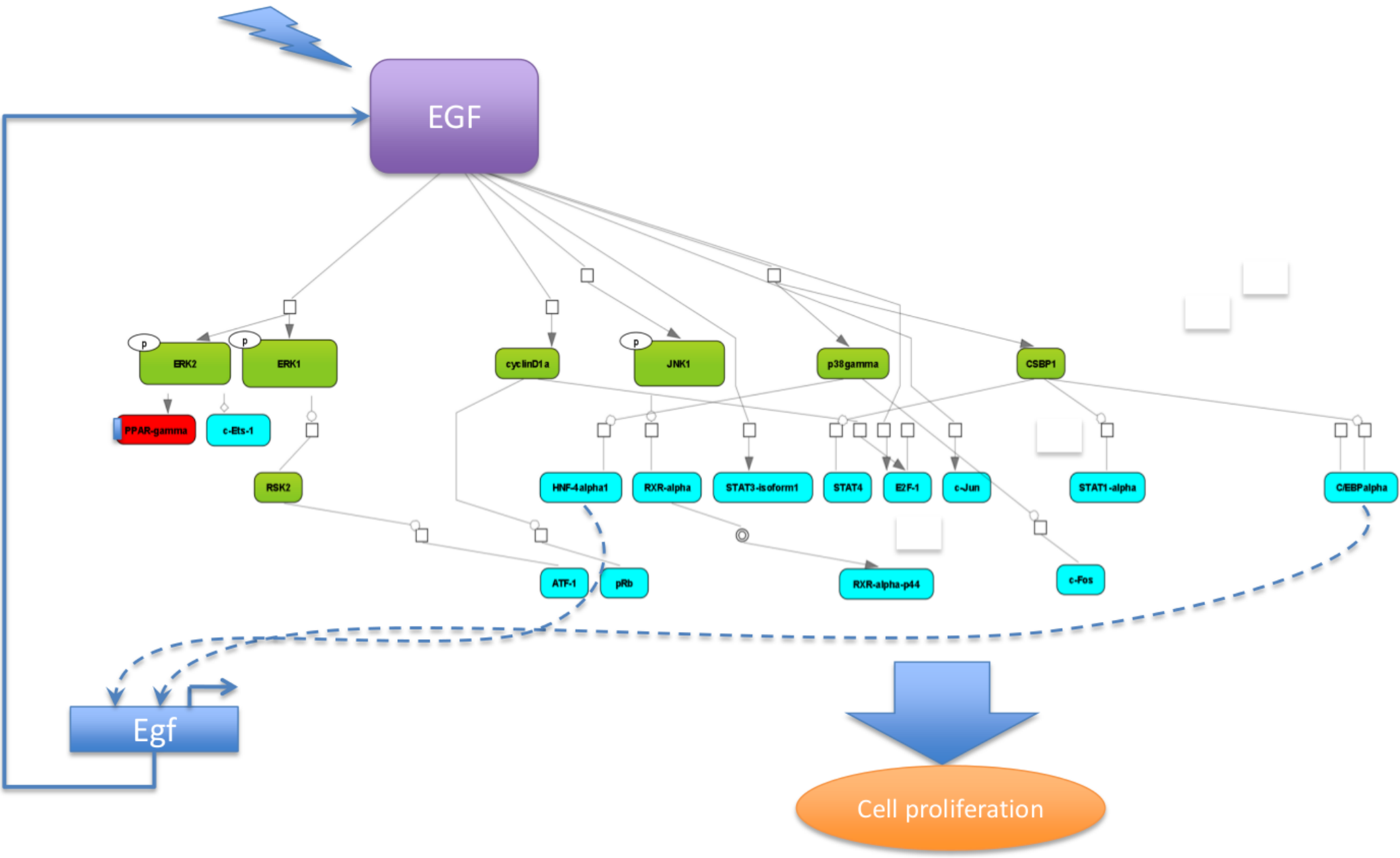
Novosibirsk

geneXplain
Wolfenbüttel

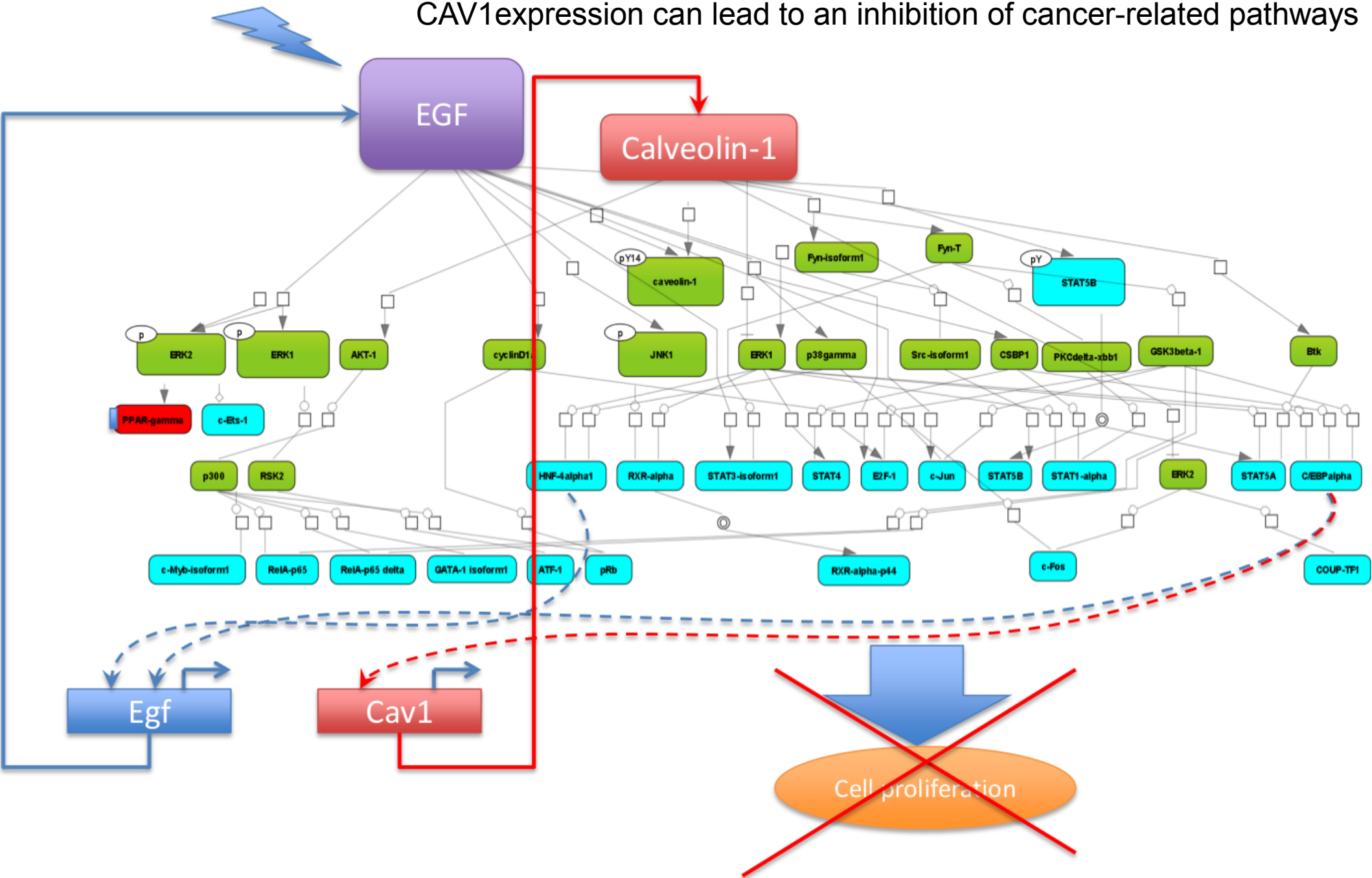
alexander.kel@genexplain.com

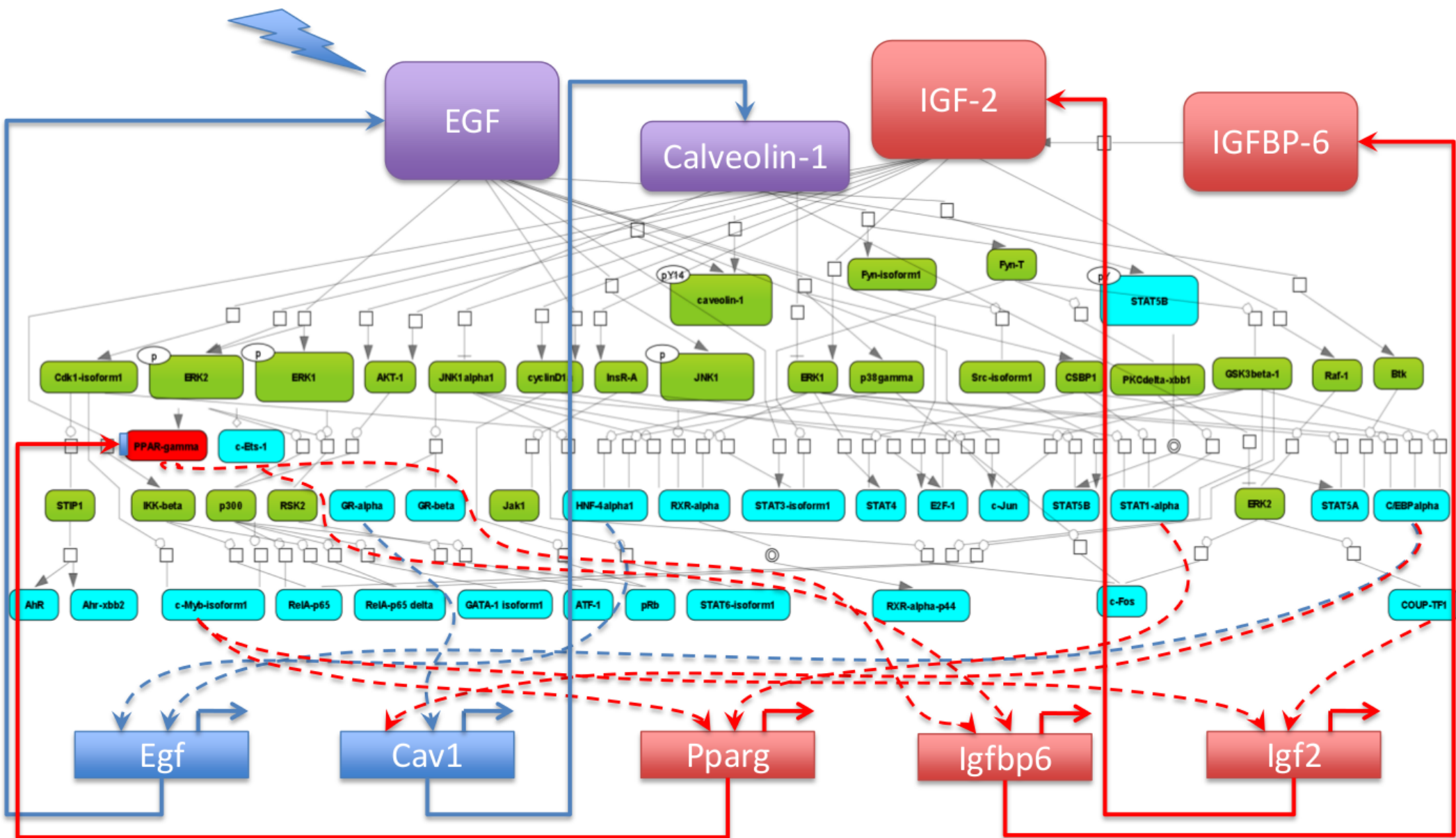


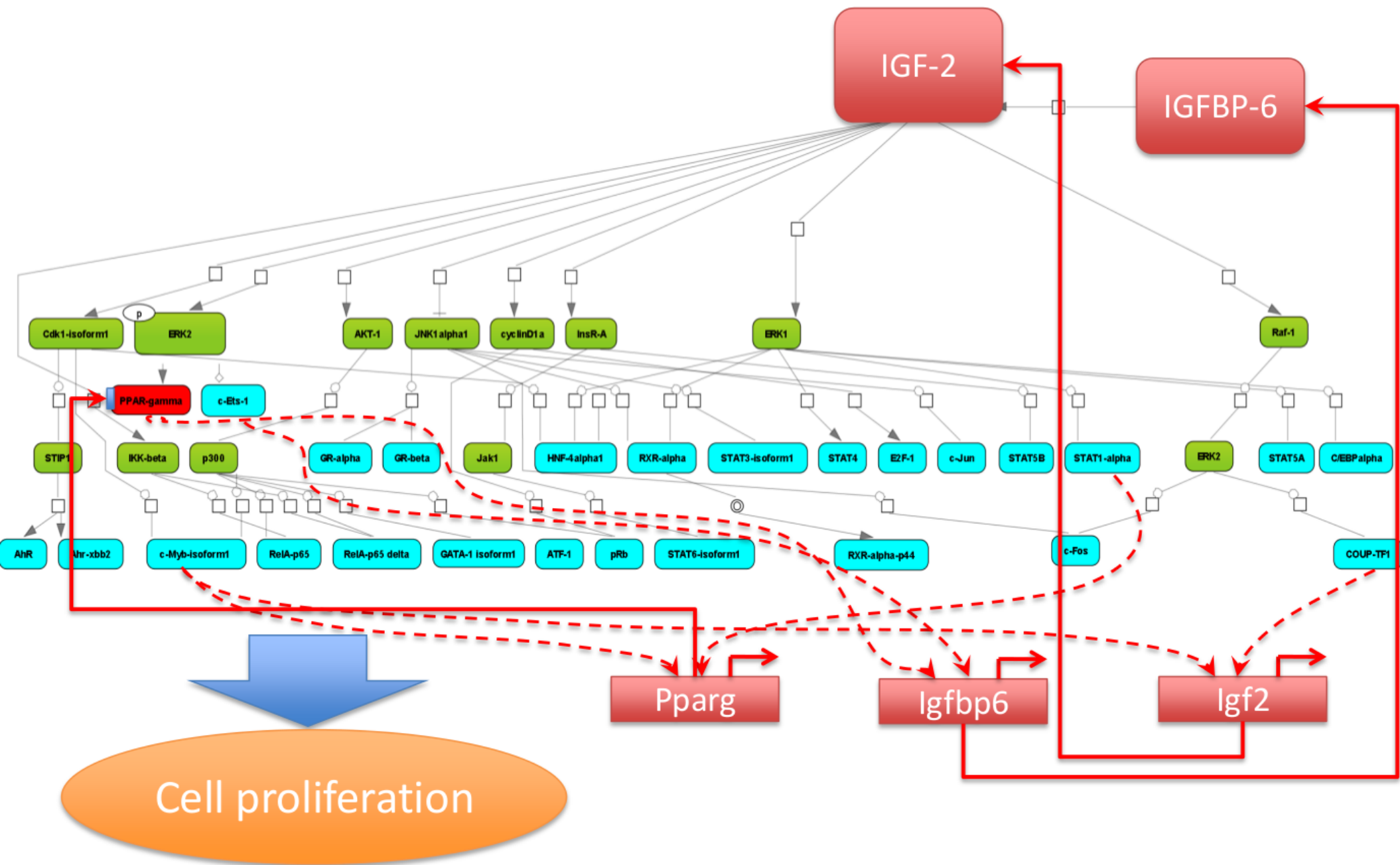




CAV1 expression can lead to an inhibition of cancer-related pathways







Paper: EGF Mouse Model of Liver Cancer

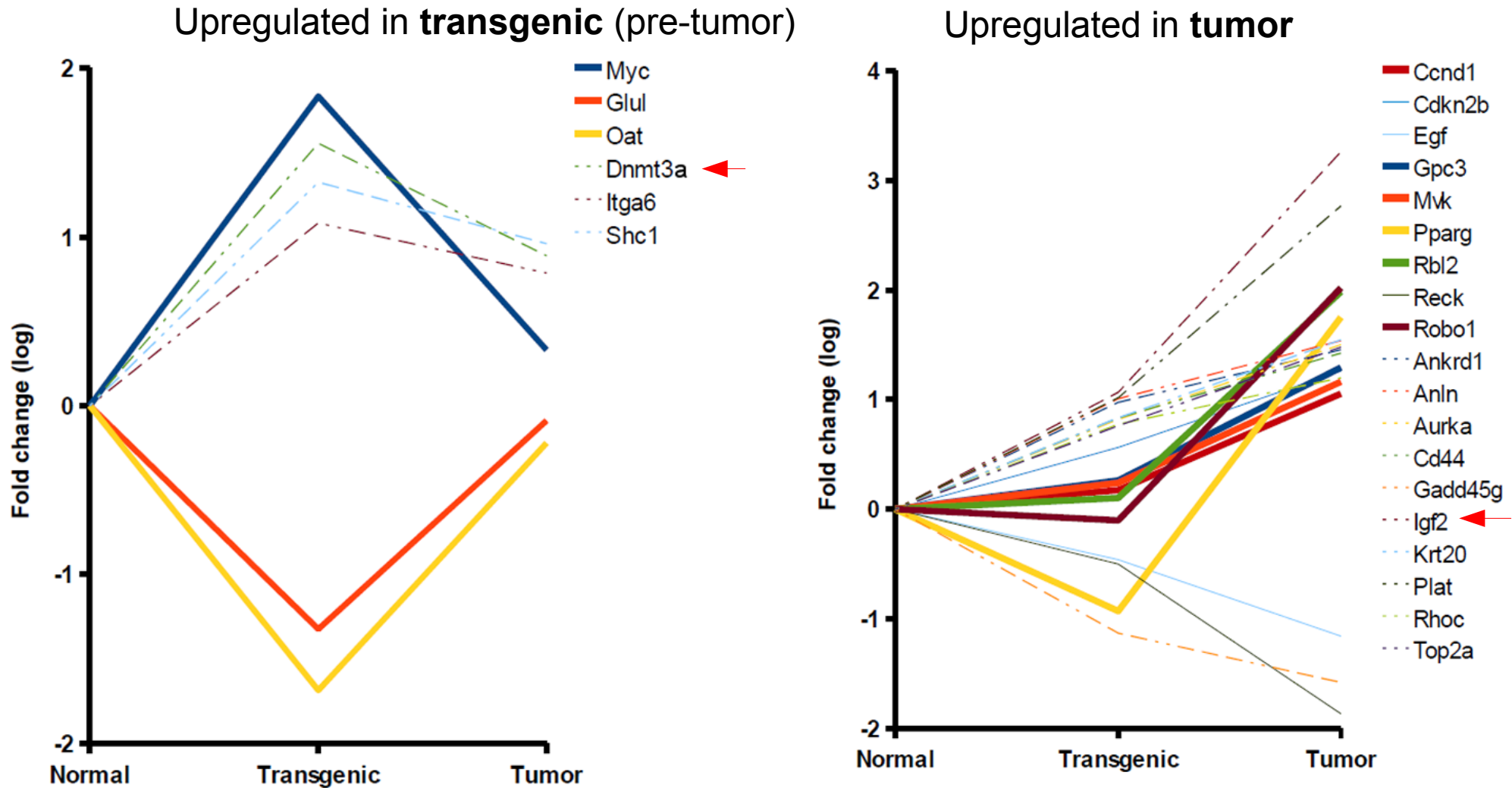


Figure 2. Expression responses of known liver carcinoma/neoplasia biomarkers in EGF-induced carcinogenicity.

Paper: EGF Mouse Model of Liver Cancer

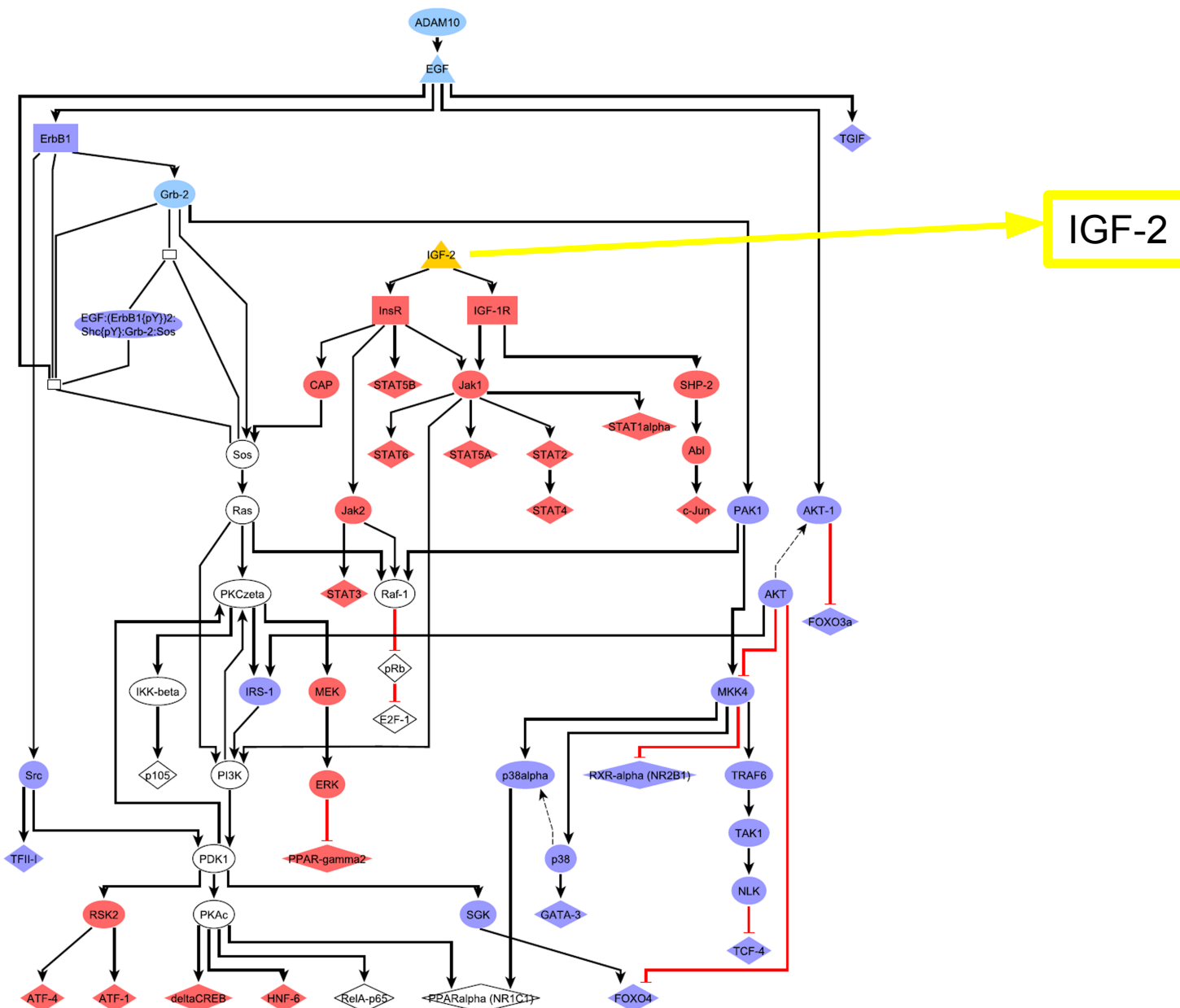
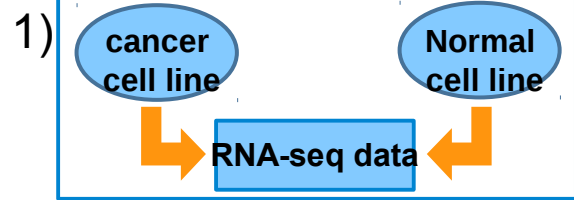
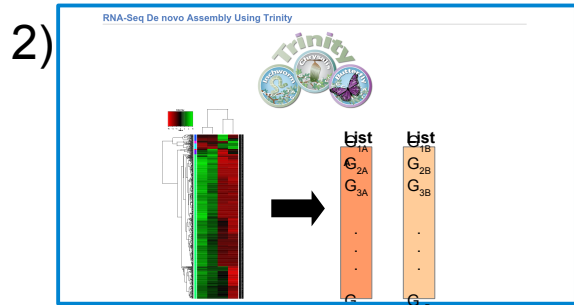


Figure 8. Merged key node networks of EGF and IGF-2 cascades with transcription factors revealed by promoter analysis.

Pathway modeling



Pre-processing



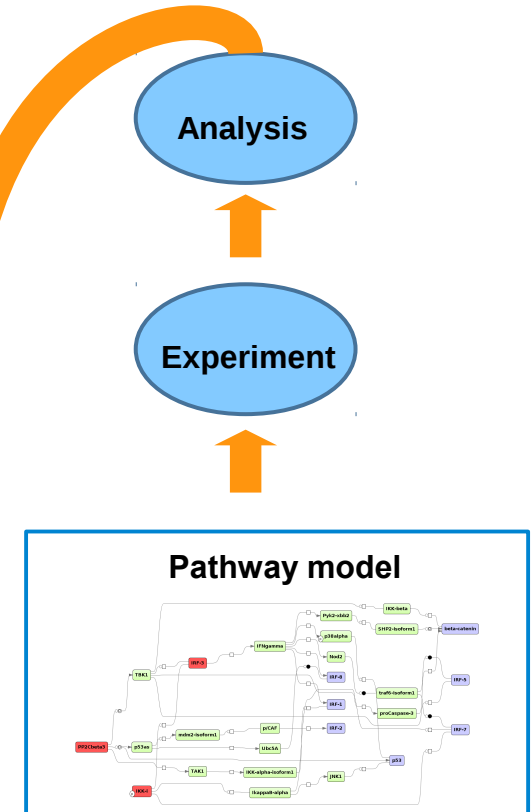
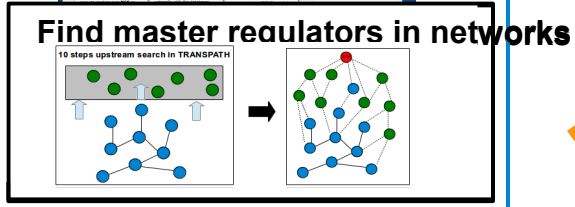
Identify DEGs:
→ Transgenic: 303
→ Tumor: 355



Identify DRGs

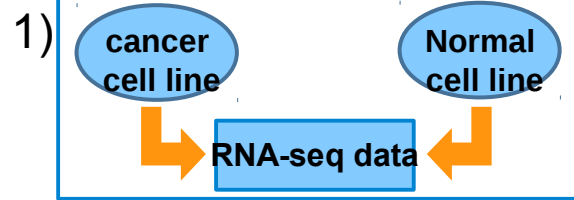


Identify master regulators

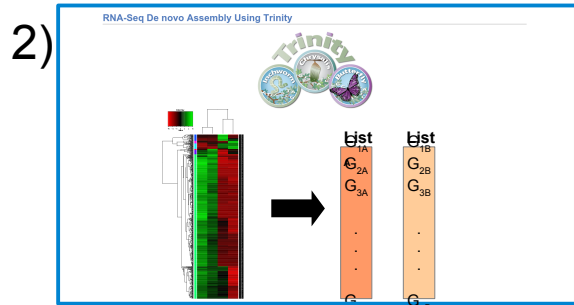


Construction of a pathway model

Pathway modeling



Pre-processing



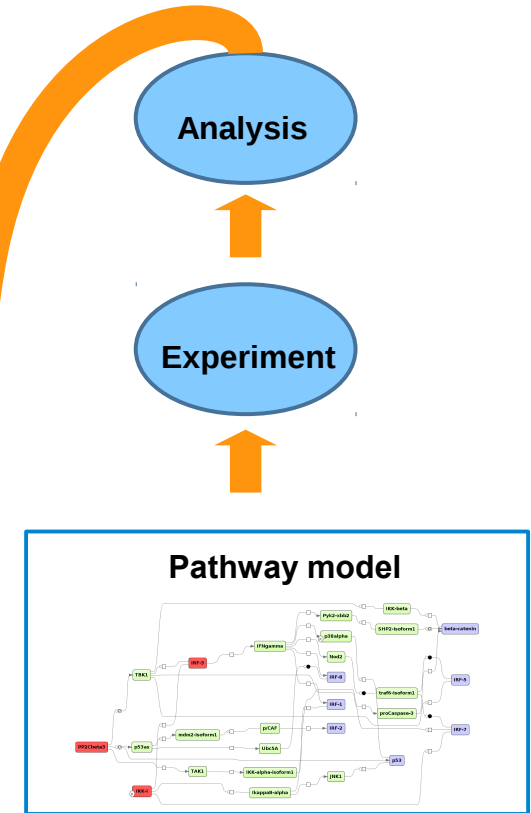
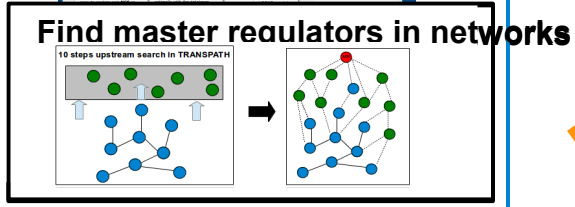
Identify DEGs:
→ Transgenic: 303
→ Tumor: 355



Identify DRGs:
→ Transgenic: 104
→ Tumor: 127



Identify master regulators



Construction of a pathway model

GeneSpiker-significant DRGs

Transgenic: top 20 DRGs

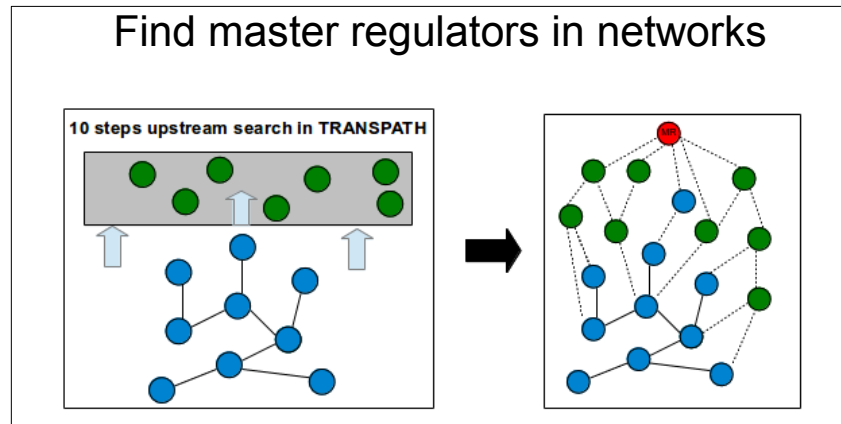
NM_001282943_Ccne2	192
NM_001098203_Hic1	174
NM_001111336_Hsd3b4	152
→ NM_007872_Dnmt3a	152
NM_001289920_E2f3	141
NM_008005_Fgf18	138
NM_011693_Vcam1	135
NM_008680_Enah	132
NR_028441_Vmn2r-ps60	127
NM_010111_Efnb2	127
NM_001146087_Ilf5	118
NM_017480_Icos	112
NM_145515_Mark1	112
NM_010151_Nr2f1*	104
NM_033075_D17H6S56E-5	99
NM_011388_Slc10a2	98
NM_001110783_Ank1	97
NM_008221_Hbb-y	96
NM_010437_Hivep2	96
NM_009697_Nr2f2*	95

Tumor: top 20 DRGs

NM_016768_Pbx3	280
NM_001282943_Ccne2	257
NM_009523_Wnt4	228
→ NM_010514_Igf2	220
NM_001289920_E2f3	209
NR_028441_Vmn2r-ps60	191
NM_010111_Efnb2	184
NM_011146_Pparg	183
NM_010637_Klf4	178
NM_009331_Tcf7	178
NM_009072_Rock2	168
NM_011604_Tlr6	168
NM_033075_D17H6S56E-5	165
NM_007423_Afp	159
→ NM_001122737_Igf2	158
NM_008605_Mmp12	146
NM_001290637_Dab2ip	142
NM_001198833_Ddr1	137
NM_009731_Akr1b7	132
NM_198622_H1fx	132

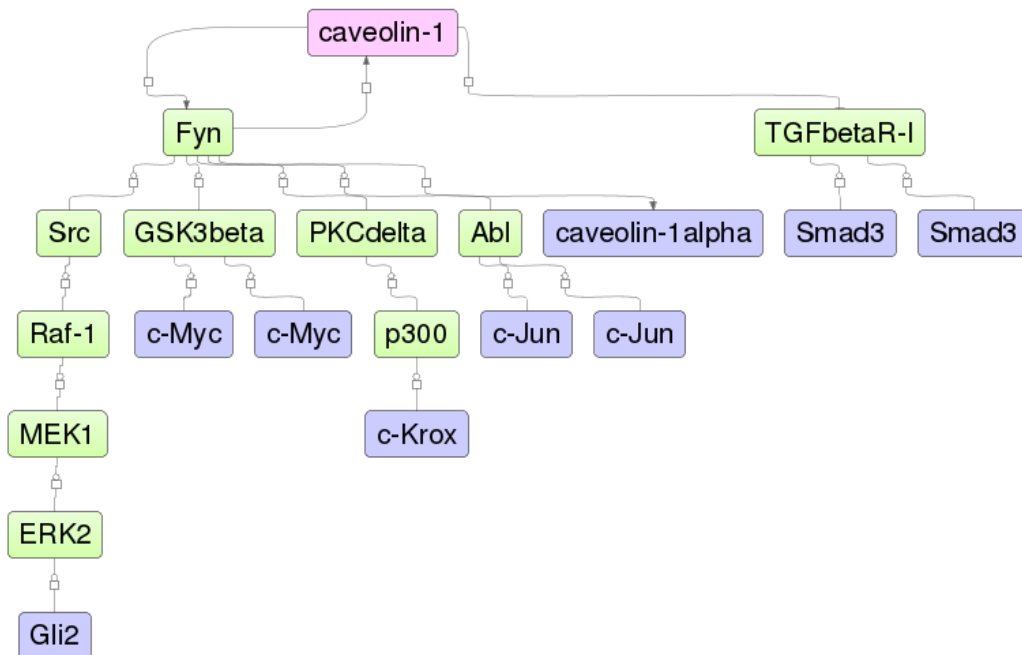
*COUP-TF1/COUP-TF2

Master regulators analysis based on GeneSpiker-significant DRG

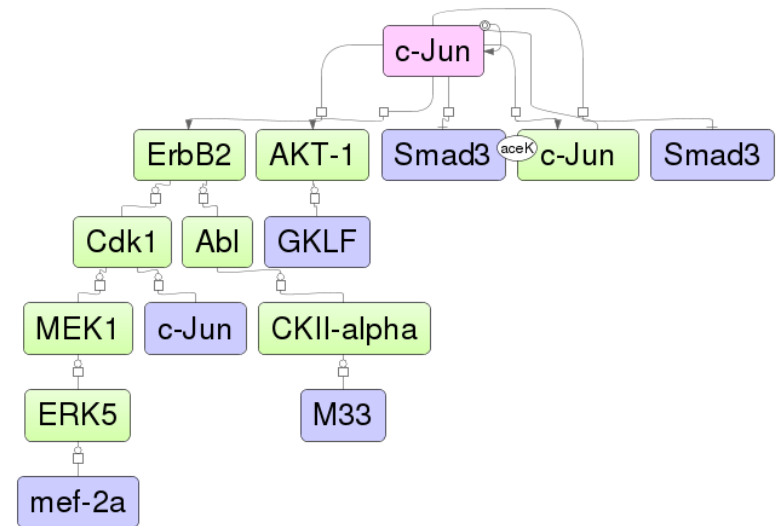


Top 3 Master regulators

Transgenic



Tumor



Conclusion

- Success of your follow-up analyses such as GSEA, TRANSPATH® pathway analysis, TFBS enrichment (F-MATCH) analysis, Master regulator analysis etc. can be enhanced by using more refined input lists (e.g. DEGs)
- **GeneSpiker** is a novel method for the identification of the most significantly differentially regulated genes (DRG)
→Output: a filtered (refined) and ranked gene list
- **GeneSpiker** can identify biomarkers of early/late cancer stage
- **GeneSpiker** in combination with follow-up analyses results in more specific pathway models