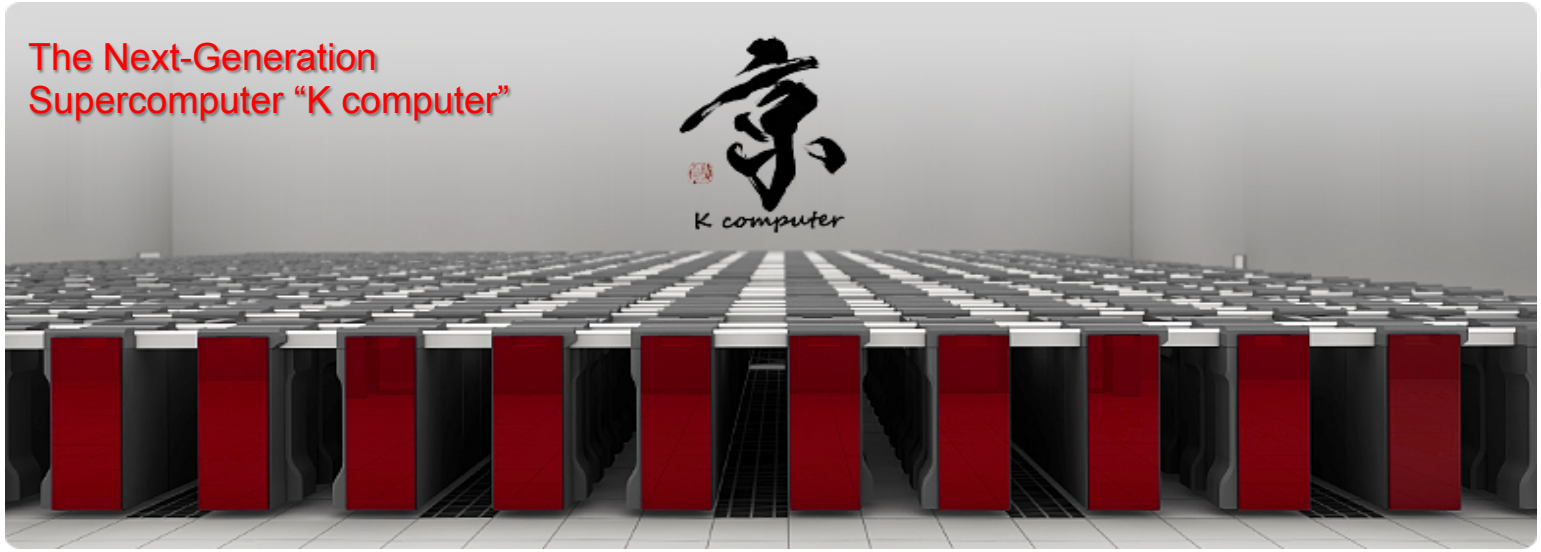
Systems Biology in Supercomputing Environment

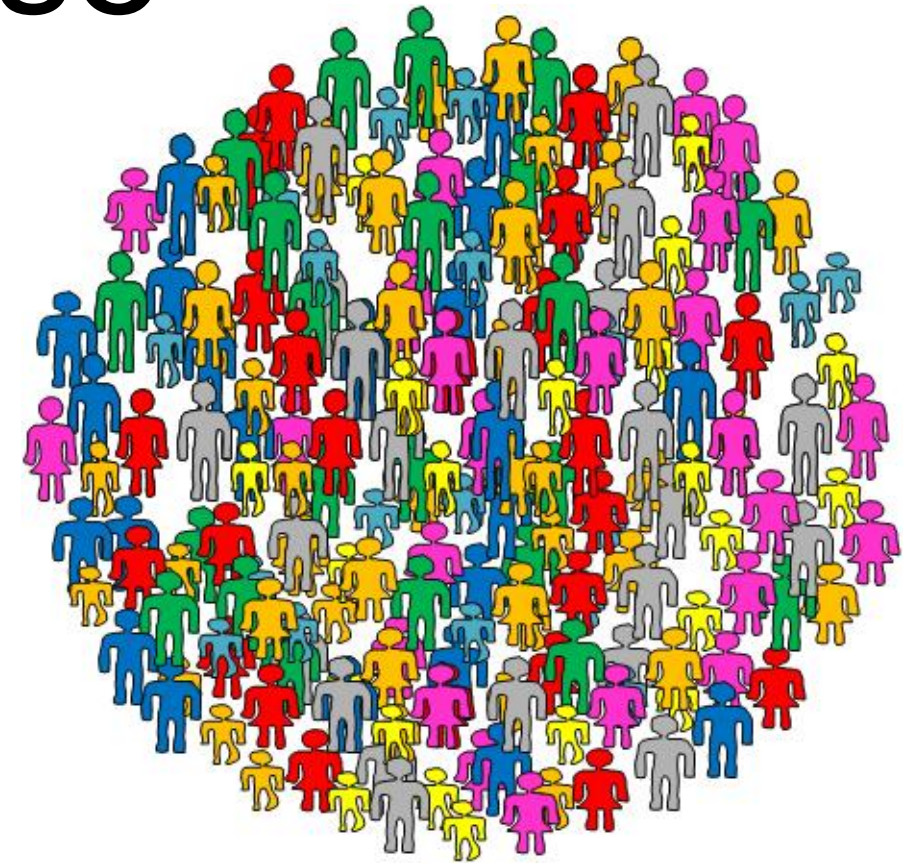
Satoru Miyano
Human Genome Center, Institute of Medical Science, The University of Tokyo
BioPPN 2011
Thistle County Hotel, Newcastle upon Tyne, UK
June 20, 2011



My interest on systems biology is “Cancer”

Because …

Japanese



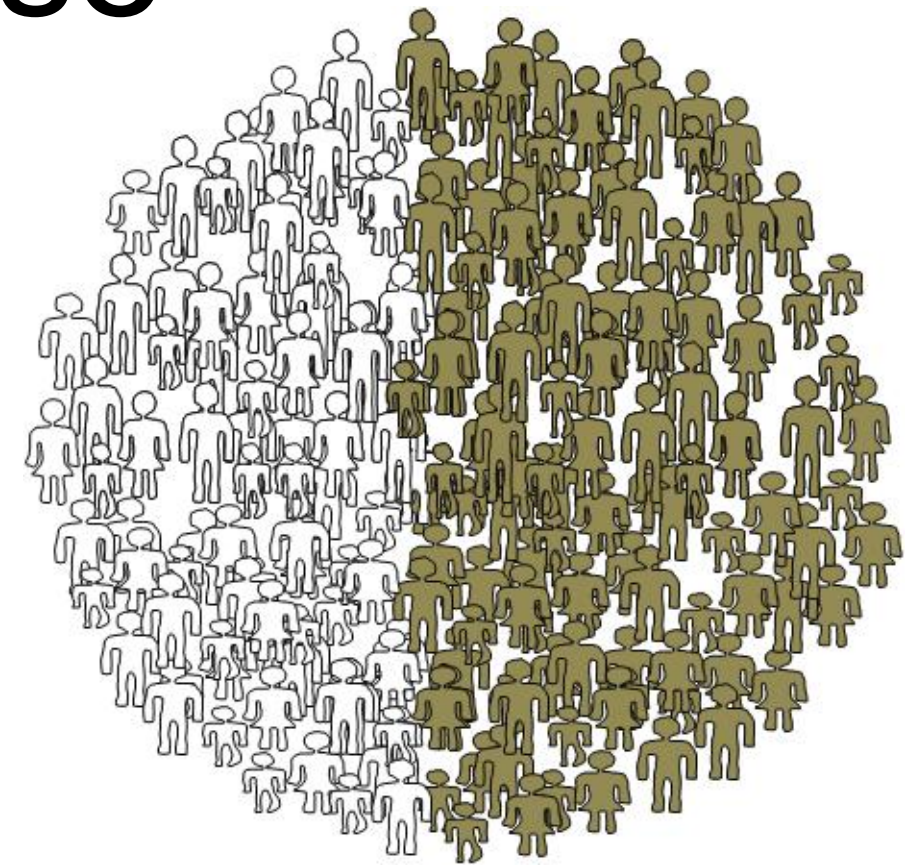
Cancer

No. 1

Cause-Specific Death Rate

Japanese

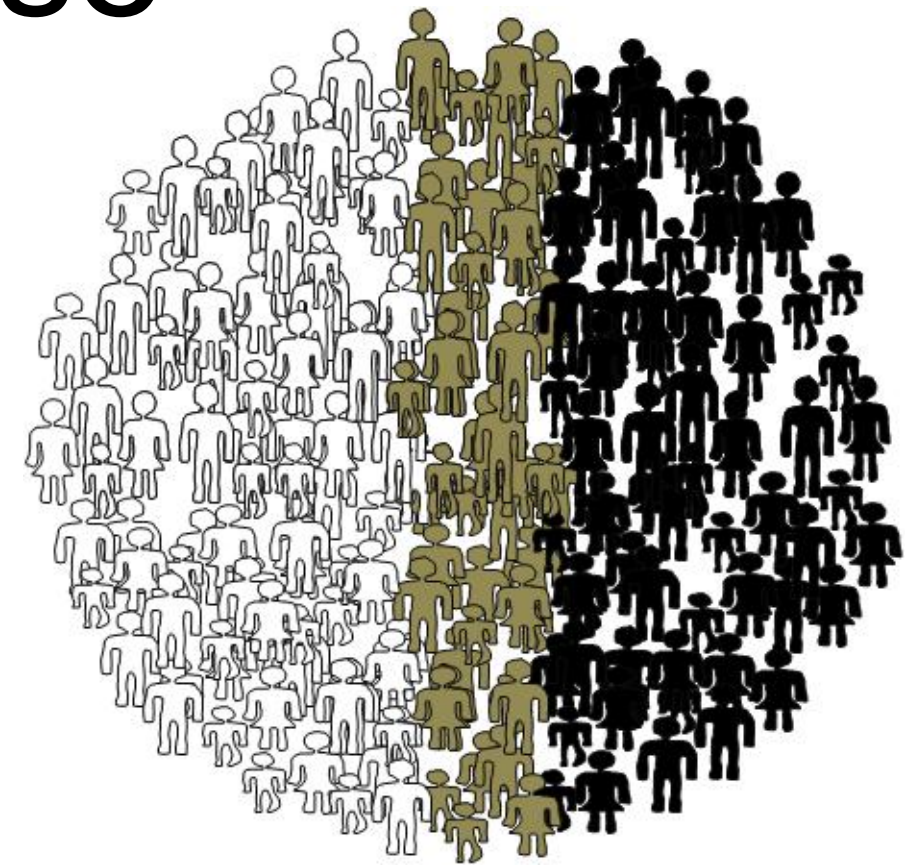
$$\frac{1}{2}$$



Cancer

Japanese

$\frac{1}{3}$

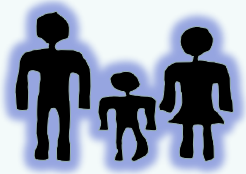


Die of cancer

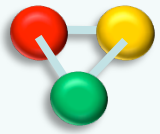
Hack Cancer System with Supercomputer!

Caution: If you hack supercomputer, you will be arrested.

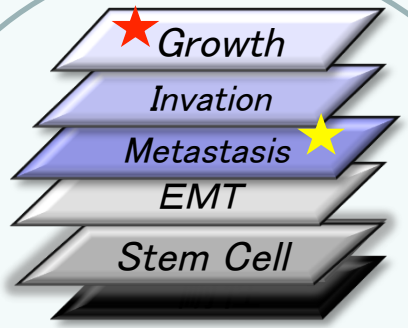
Traditional Cancer Research



Small Number of Samples



2–3 Molecules



Independent Phenomina

Cancer scientists had only slit-views on cancer

However, we have some insights on cancer ・・・

And, molecular targets are discovered!

Lung cancer: Irressa
Breast cancer: Herceptin
Leukemia: Gleevec

But limited use ・・・

NBC

Friday, May 15
9-11PM Pacific Time

Farrah's Story: On A Wing & A Prayer

2001 CML(Leukemia)

2011 Active

2006 Anal Cancer

2009 Died

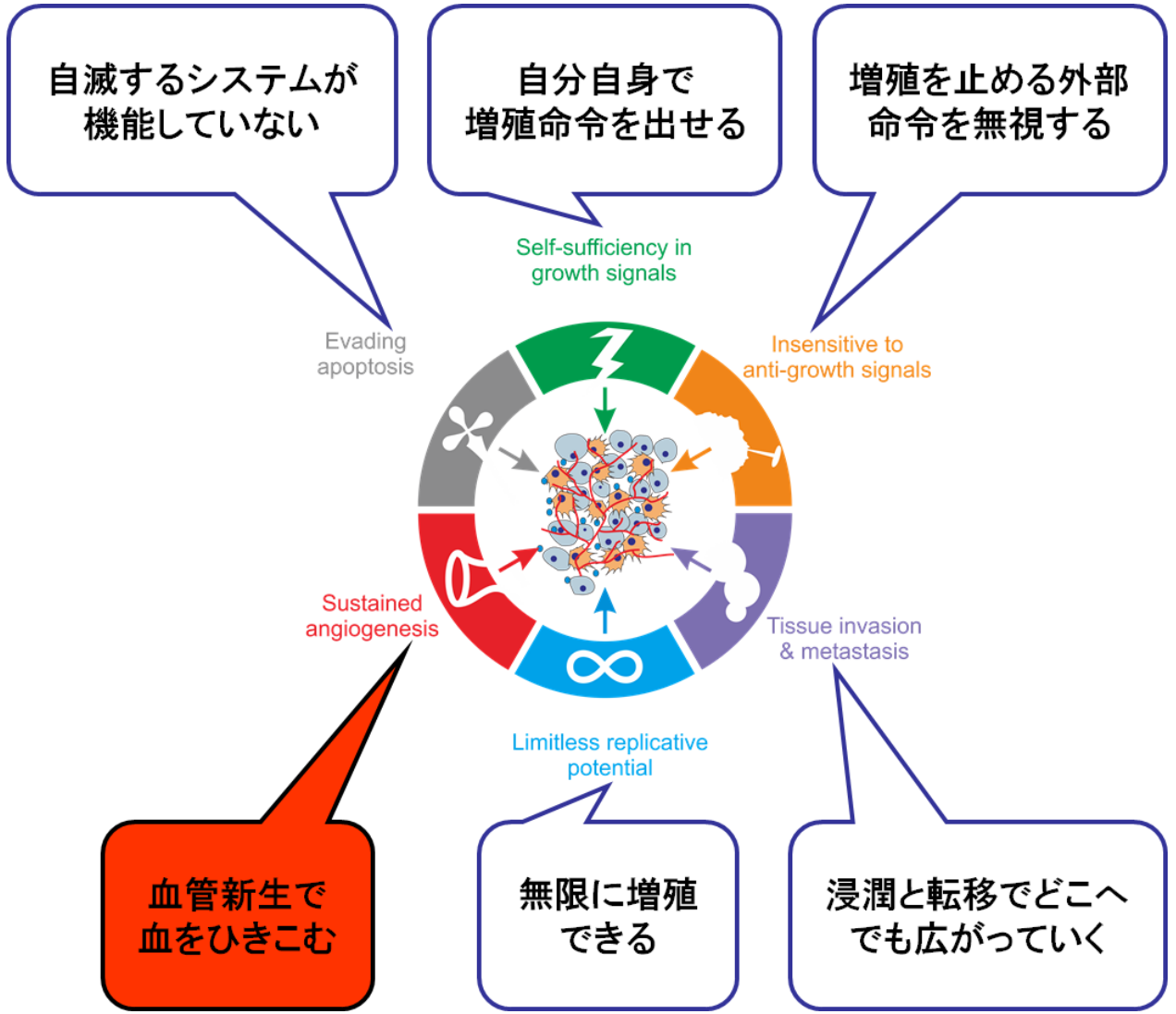
Danjuro Ichikawa – Kabuki Actor

Misako Honda – Actress (1967-1985)

Why not effective to me?

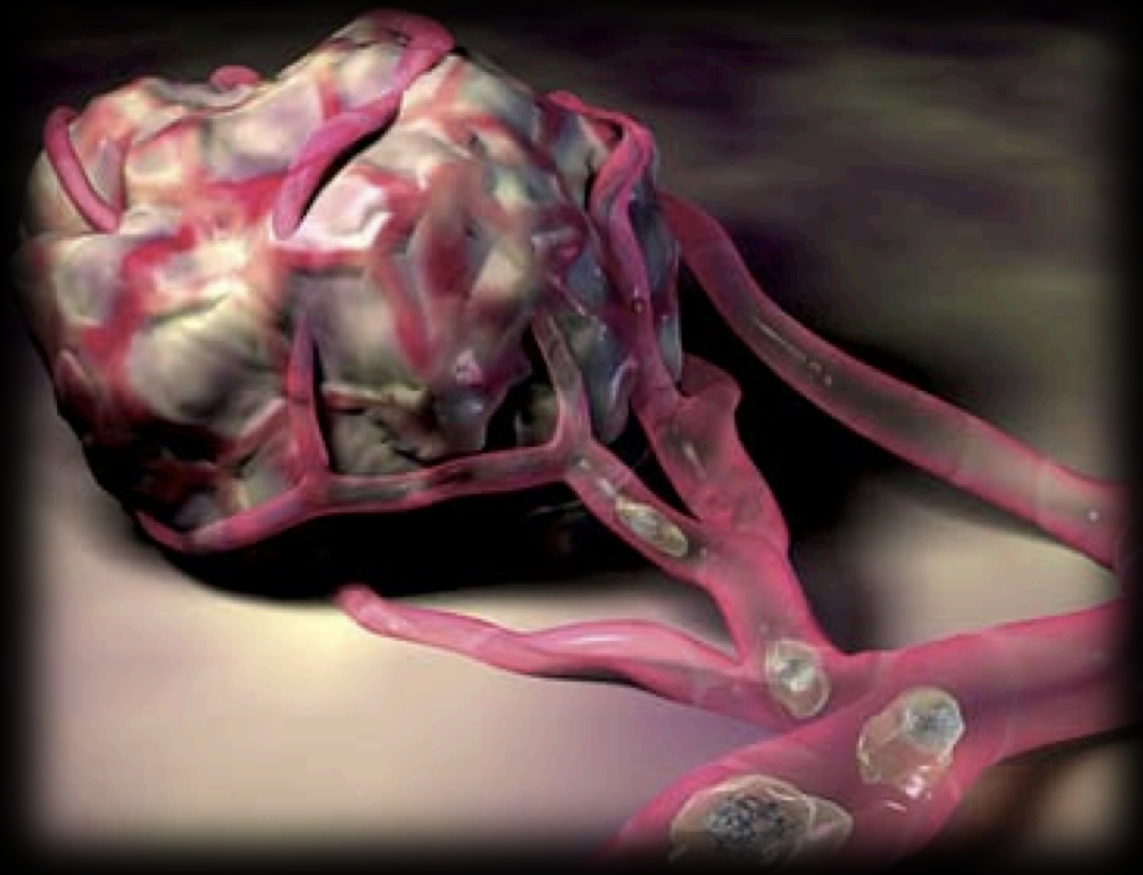
Can we hope new drugs and therapies in the future?

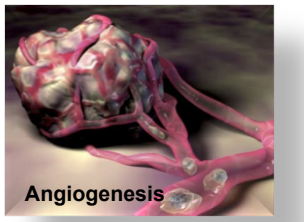
Cancer is Similar to Japanese Bureaucracy System



Modified: “Hanahan & Weinberg, Cell 2000.

Angiogenesis

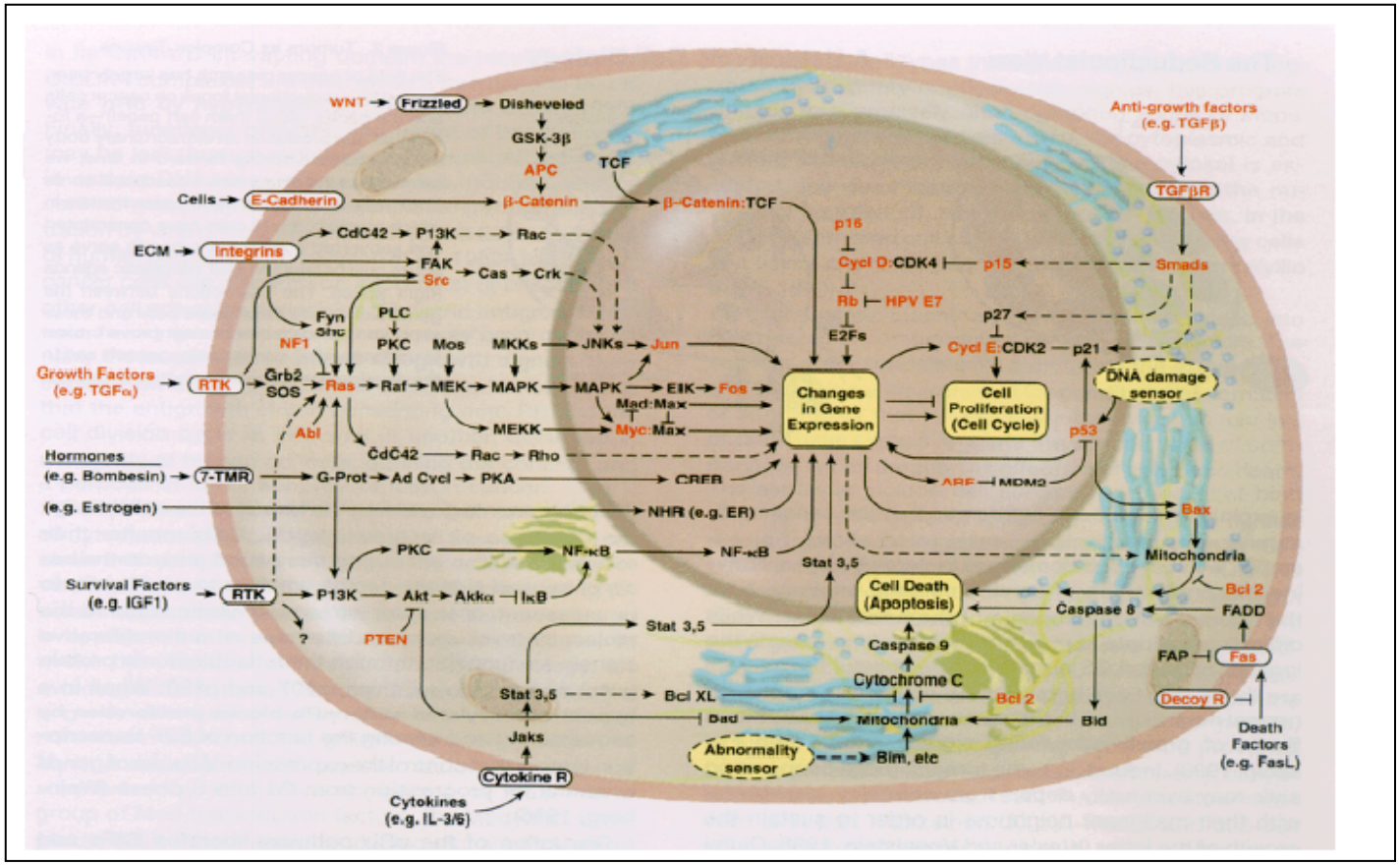




Known mechanisms involved in Angiogenesis is very complex

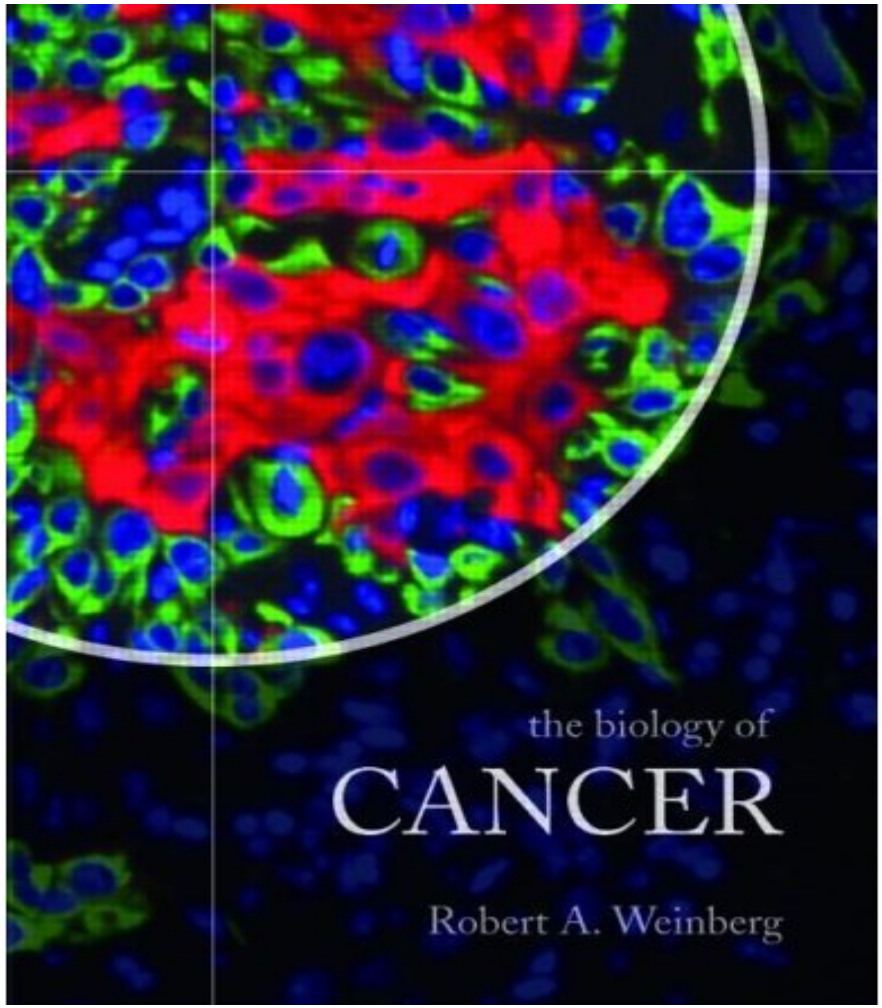
Model made by using Cell Illustrator

Hallmark of Cancer



D. Hanahan and R. A. Weinberg.
Cell., 100(1):57–70 Review, 2000.

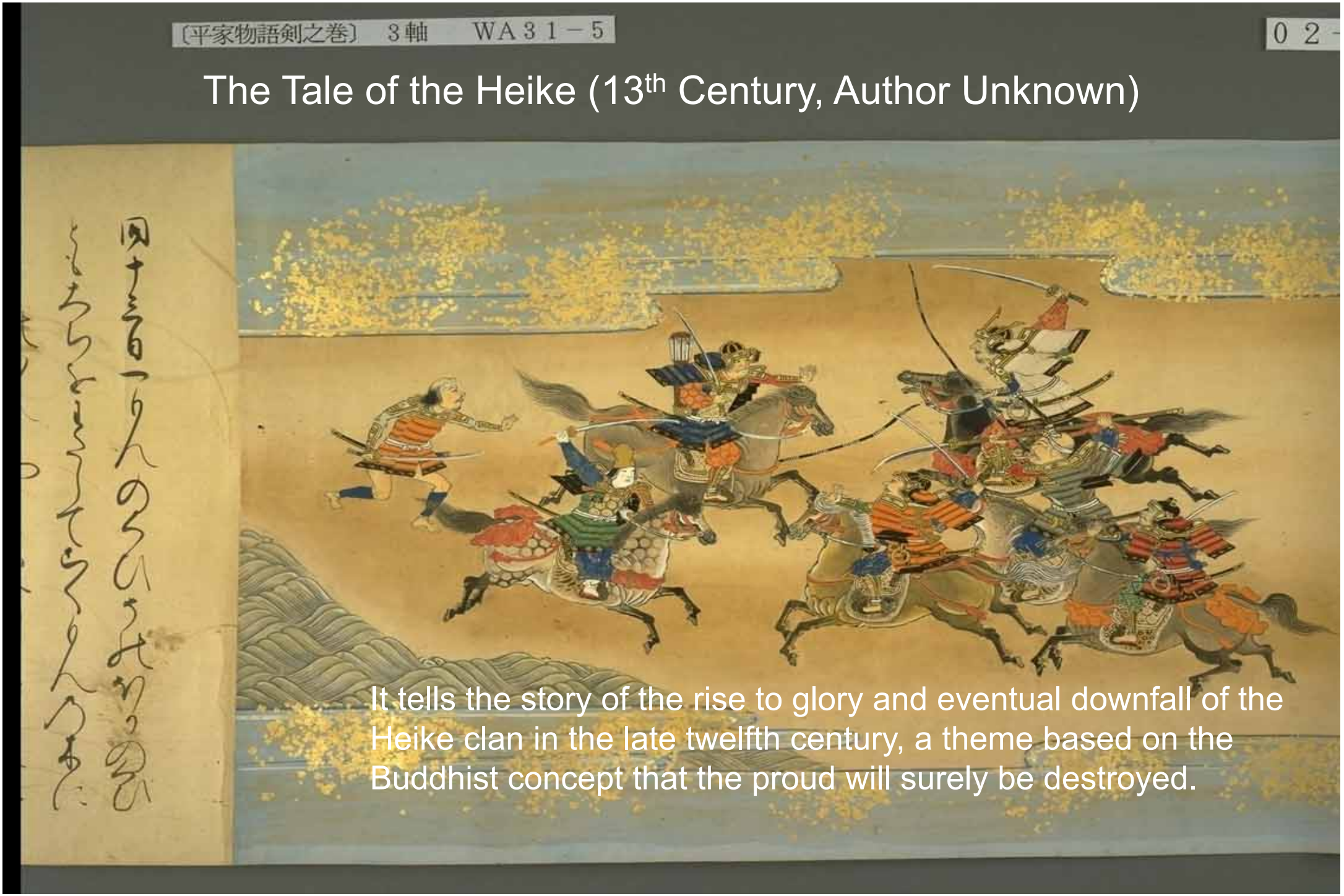
The Biology of Cancer



Robert A. Weinberg

Golden Days of Molecular Biology

- **Phenotypes and Genes**
  –Paradigm of Molecular Biology
- **Super Stars in Biology**
  – Novel Prize Laureates
- **Dramatic History**
  – “The Biology of Cancer”, R.A. Weinberg

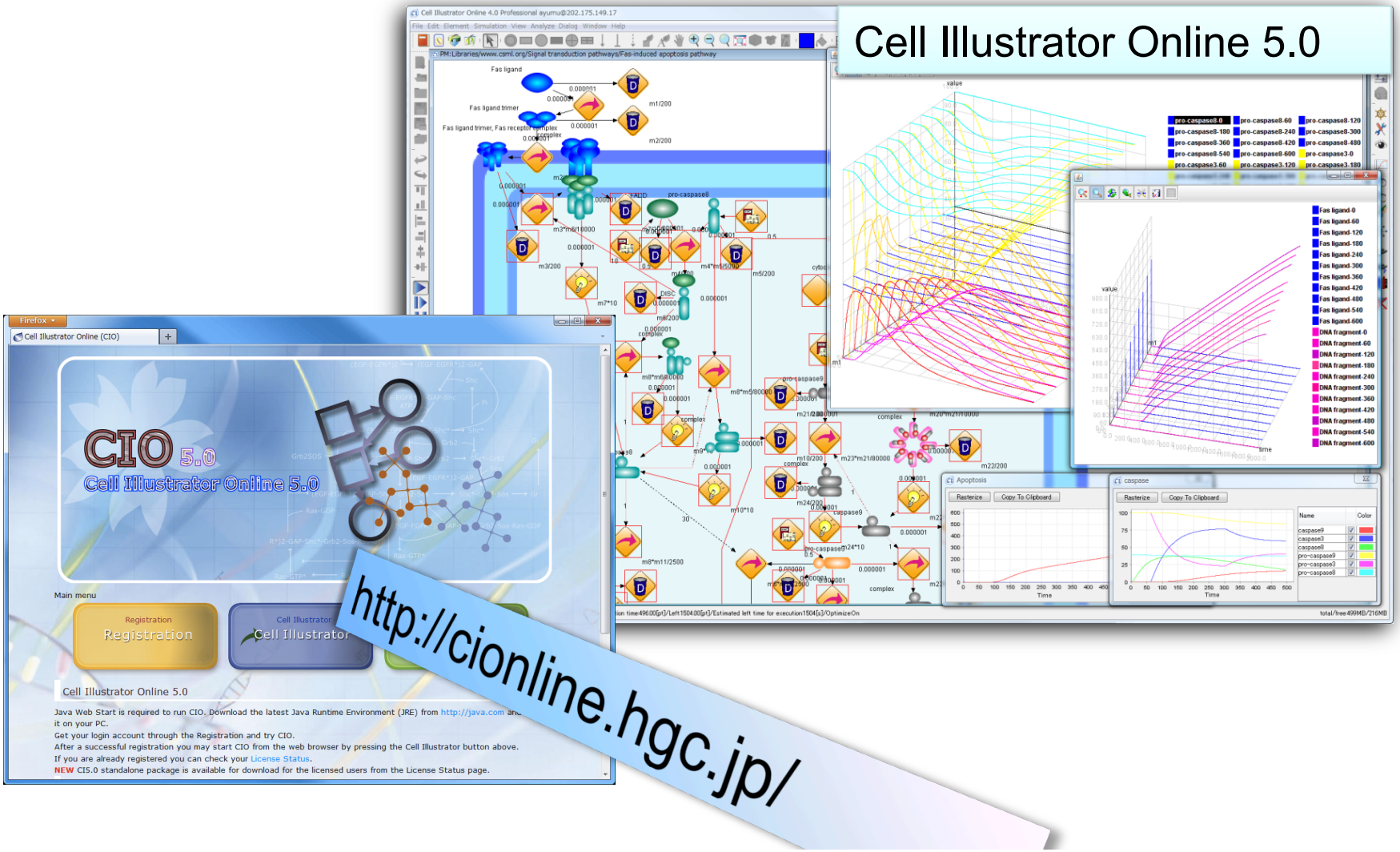

〔平家物語剣之巻〕 3軸 WA31－5
02
The Tale of the Heike (13th Century, Author Unknown)
It tells the story of the rise to glory and eventual downfall of the Heike clan in the late twelfth century, a theme based on the Buddhist concept that the proud will surely be destroyed.

But, Very Naïve Understanding and Representation as Systems

- Draw pictures and add English narrations for biological facts
- “Systems Knowledge” which is unambiguously represented is at most causal relations among molecules
- Huge gap between “what is represented” and “what is to be represented”

Cell Illustrator Online 5.0

Java Web Start Application



Hybrid Functional Petri Net with extension (HFPNe)

Nagasaki M, Doi A, Matsuno H, Miyano S. A versatile Petri net based architecture for modeling and simulation of complex biological processes. *Genome Informatics*. 15(1):180-97, 2005.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Computational modeling of biological processes with Petri net based architecture. In "*Bioinformatics Technologies*" (Y.P. Chen, ed.). Springer Press. 179-243, 2005

Extension

- Real (continuous)
- Integer (discrete)

+

- Object (universal)

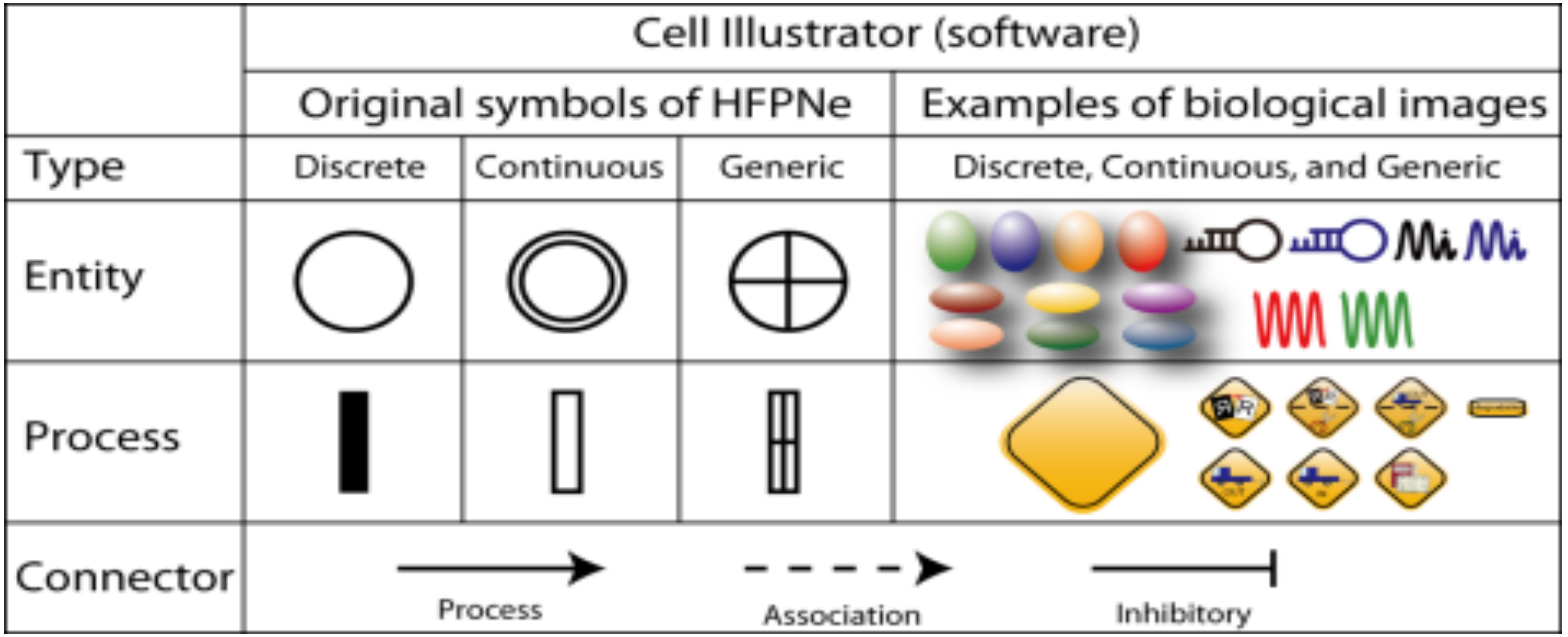
Real, Integer, Boolean, String, vector, etc.

- Dynamic cellular localization information
- Dynamic multi-cellular behavior
- Detailed molecular mechanism

User's Abilities Required for Cell Illustrator

- Biology
- Mobile phone
- Math at the level of junior high school

Place, Transition, Arc in Petri Net



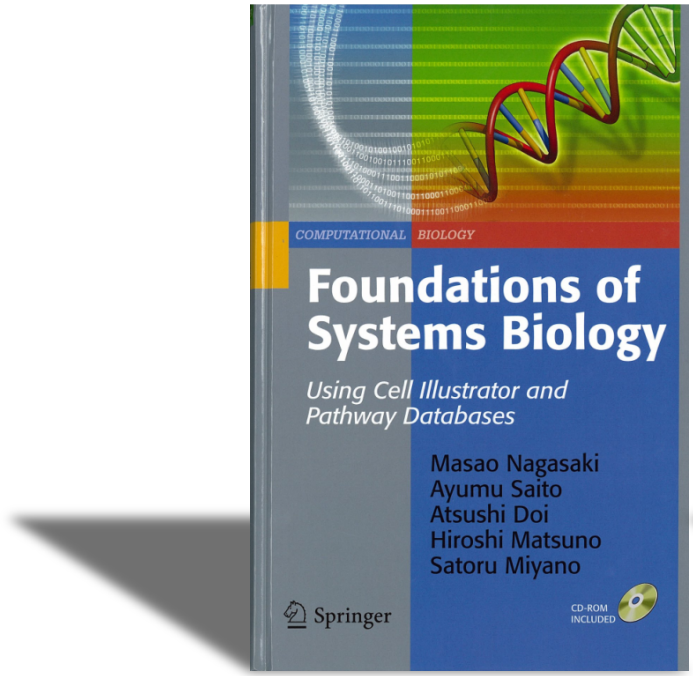
Foundations of Systems Biology

Using Cell Illustrator and Pathway Databases

Series: Computational Biology , Vol. 13

Nagasaki, M., **Saito**, A., **Doi**, A., **Matsuno**, H., **Miyano**, S.

2009, Approx. 170 p. 145 illus. in color. With CD-ROM., Hardcover
ISBN: 978-1-84882-022-7 26,95 €



2. Knowledge Representation of Dynamic Biopathways

Cell System Ontology (CSO 3.0)

Cell System Markup Language (CSML 3.0)

http://www.csml.org/

CSML 3.0 & CSO 3.0

- Native XML format for Cell Illustrator
- XML and Ontology for Biopathway
  - Modeling
  - Simulation
  - Visualization

Features of CSO 3.0

Differ from BioPAX

- System-dynamics centered ontology.
- The ontology is implemented with Web Ontology Language (OWL), which enables semantic validation and provide complete and consistent biological pathway models.

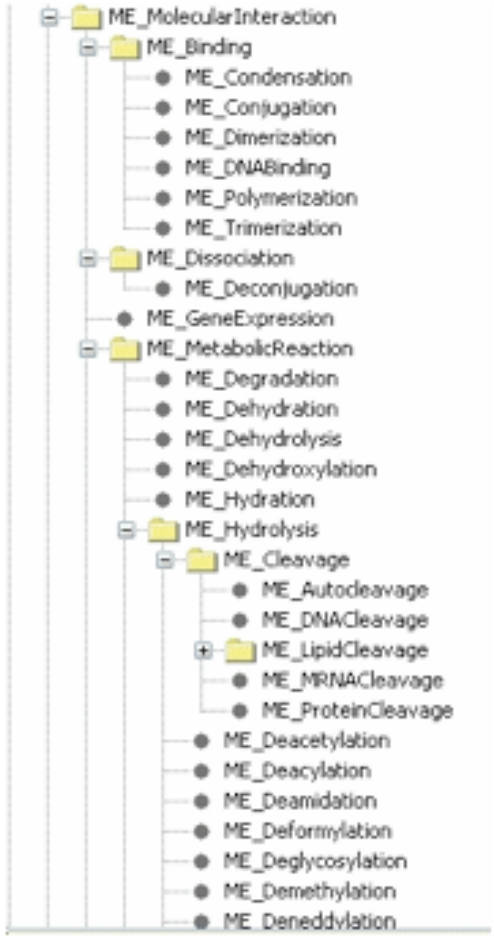
Differ from CellML and SBML

Cell System Ontology: Representation for Modeling, Visualizing, and Simulating Biological Pahtways, Euna Jeong※ Masao Nagasaki※, Ayumu Saito, Miyano Satoru, *In Silico Biology*, 2007; 7: 0055.

Jeong E, Nagasaki M, Miyano S. Conversion from BioPAX to CSO for System Dynamics and Visualization of Biological Pathway. *Genome Informatics.* 2007; 18: 225-236.

Features of CSO 3.0

- CSO equips mature core vocabularies (more than 350)
- Each core vocabulary has at least one standard icon



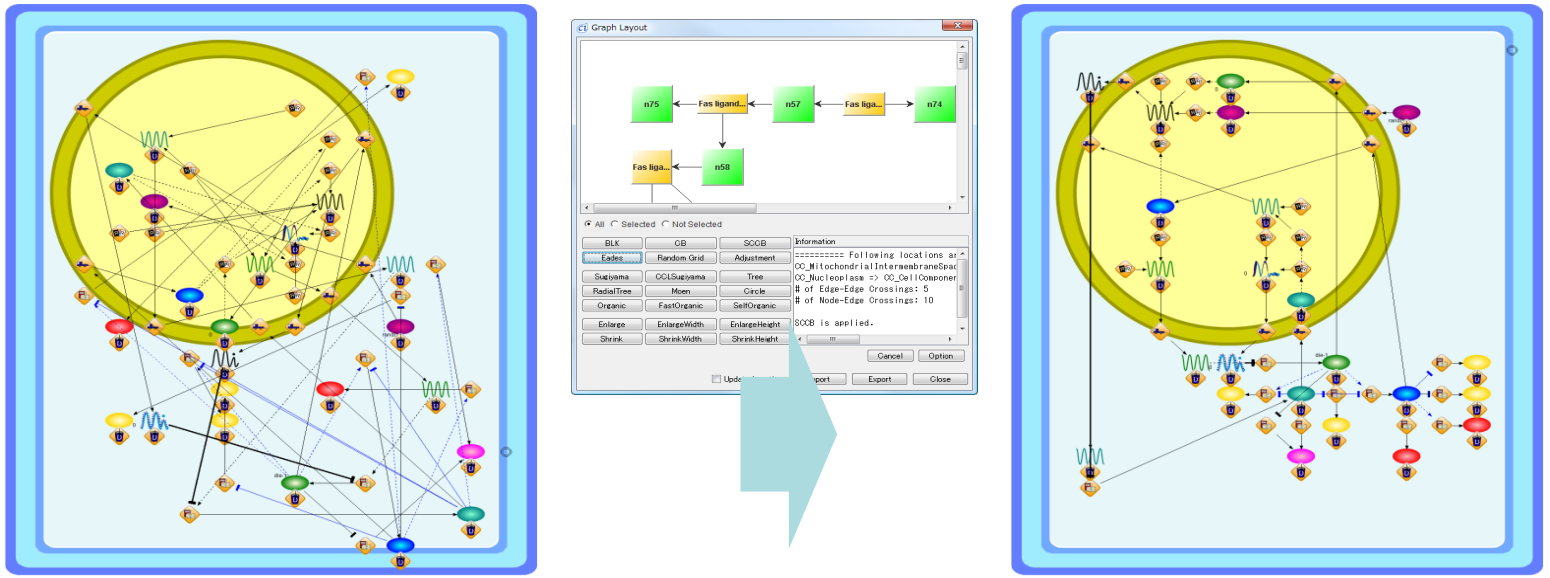
Features of CSML 3.0

1. Hybrid Functional Petri net with extension (HFPNe)
2. Logic based descriptions, *e.g.,* temporal logic, can be defined with the format.
3. User can create sub-models for a model with a filtering concept.
4. User can define more than one views for a model, *e.g.,* gene network view, simulation view.
5. All terms in CSML 3.0 has the background of Ontology: Cell System Ontology (CSO) 3.0.

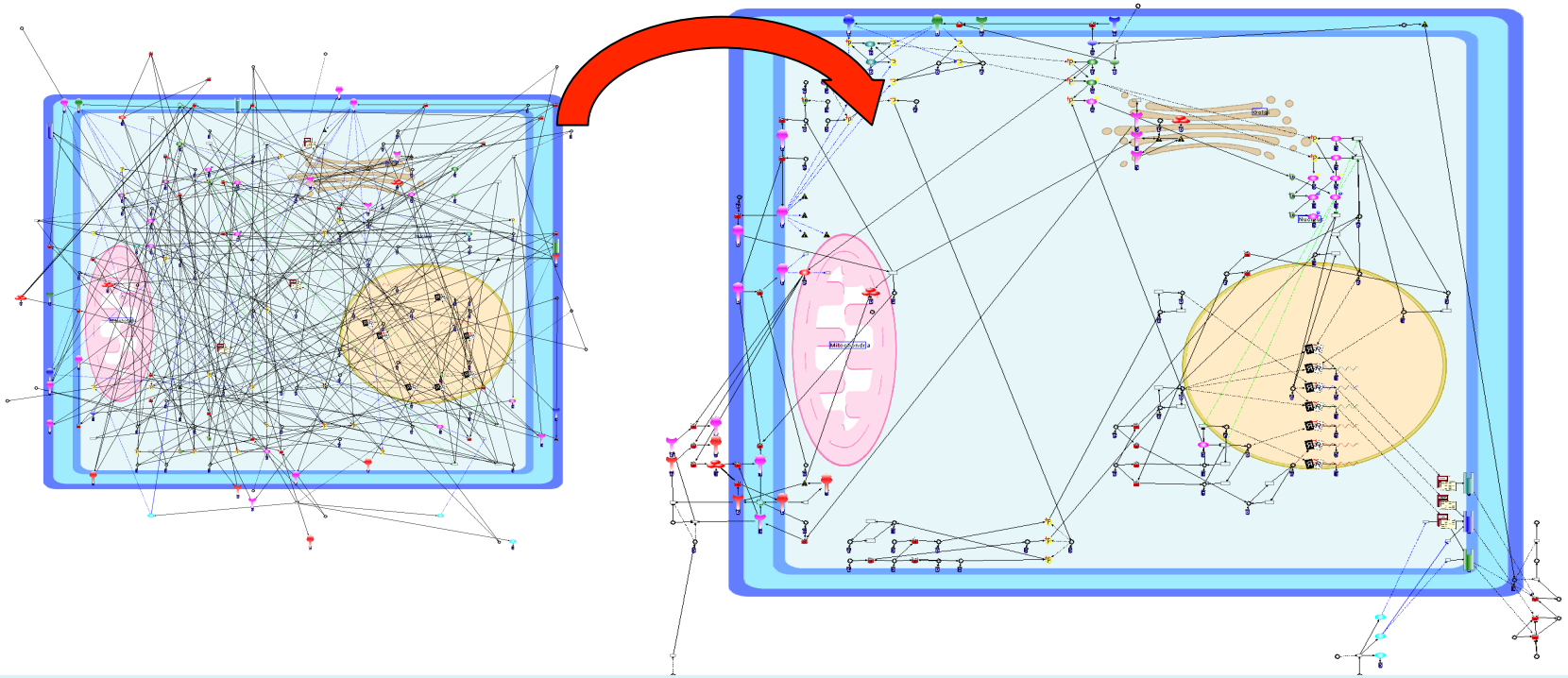
These parts were missing in SBML and CellML.
SMBM and CellML are subsets of CSML3.0

3. Biopathway Layout

- Biologically sophisticated pathway layout algorithms are required
- Cell Illustrator Online 5.0 has more than 20 layout algorithms



Fast Grid Layout Algorithms for Biological Pathways



Kojima, K., Nagasaki, M.*, Miyano, S. Fast grid layout algorithm for biological networks with sweep calculation. Bioinformatics. 24(12): 1426-1432, 2008.

Kojima K, Nagasaki M, Jeong E, Kato M, Miyano S. An efficient grid layout algorithm for biological networks utilizing various biological attributes. *BMC Bioinformatics*. 2007; 8:76.

4. Transforming Pathway Databases for Cell Illustrator

XML Format

- TRANSPATH2CSML
- SBML2CSML
- CellML2CSML

Ontology Format

- BioPAX2CSO

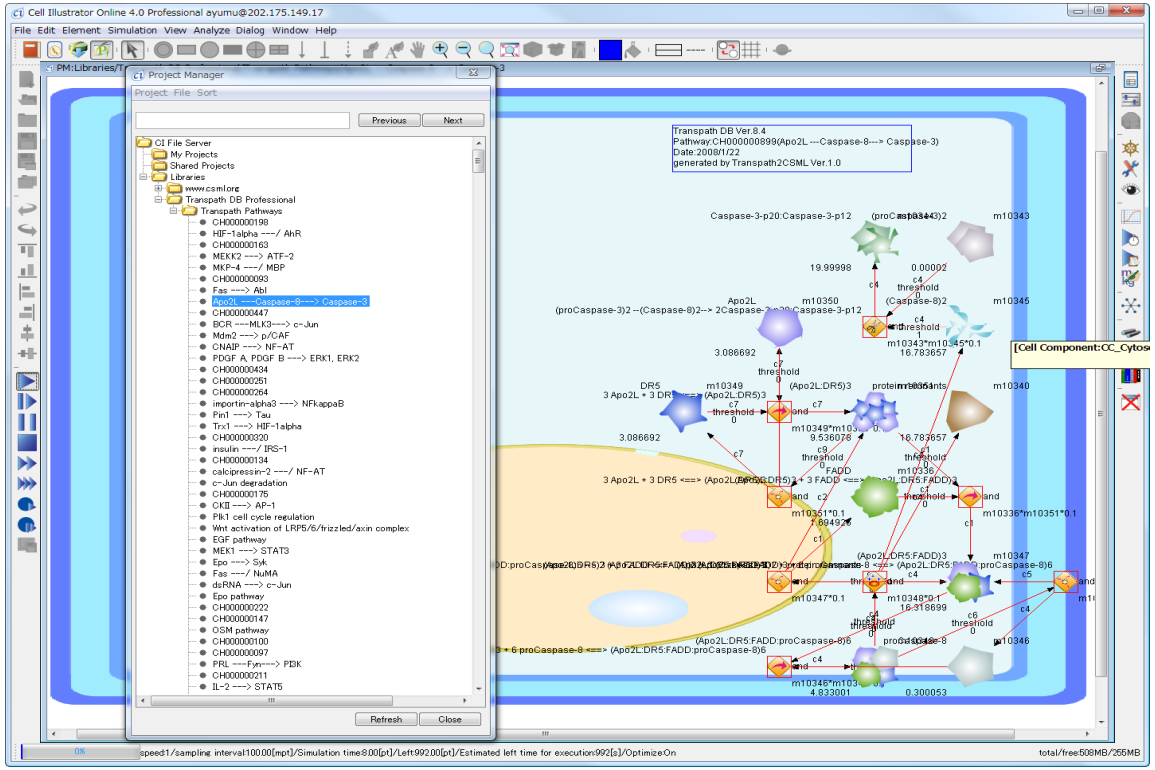
TRANSPATH to CSML

- 16 modeling rules based on Hybrid Functional Petri Net with extension (HFPNe).
- TRANSPATH (Biobase): More than 115,000 cellular events in humans, mice, and rats, collected from over 31,500 publications.
- Petri net element is incorporated with Cell System Ontology (CSO) to enable semantic interoperability of models.
- 97% of the reactions in TRANSPATH are converted into simulation-based models in CSML.

Nagasaki M, Saito A, Li C, Jeong E, Miyano S. Systematic reconstruction of TRANSPATH data into Cell System Markup Language. BMC Systems Biology. 2:53, 2008.

TRANSPATH Pathway Library Module

- More than 1,000 TRANSPATH pathways (Signal Transduction Pathway and Gene Regulatory Network) are supplied. All pathways can be loaded, edited, saved and simulated on CIO.
  - Support pathways supplied in TRANSPATH 8.4 (BIOBASE).
  - Academic user can register and use the academic version of TRANSPATH.
  - Curated 100,000 reactions and 100,000 molecules in Human and Mouse.

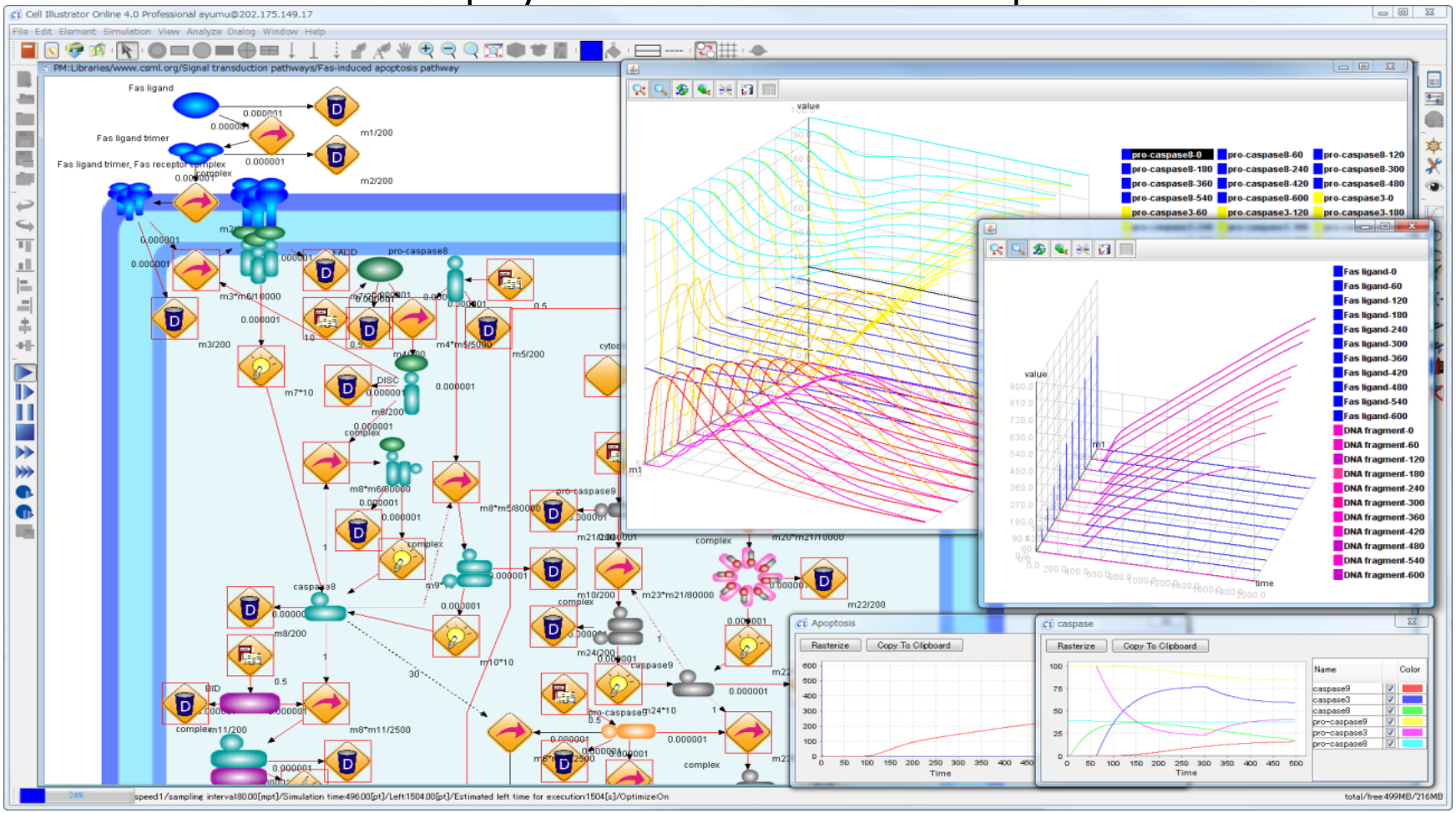


Pathway Parameter Search Module

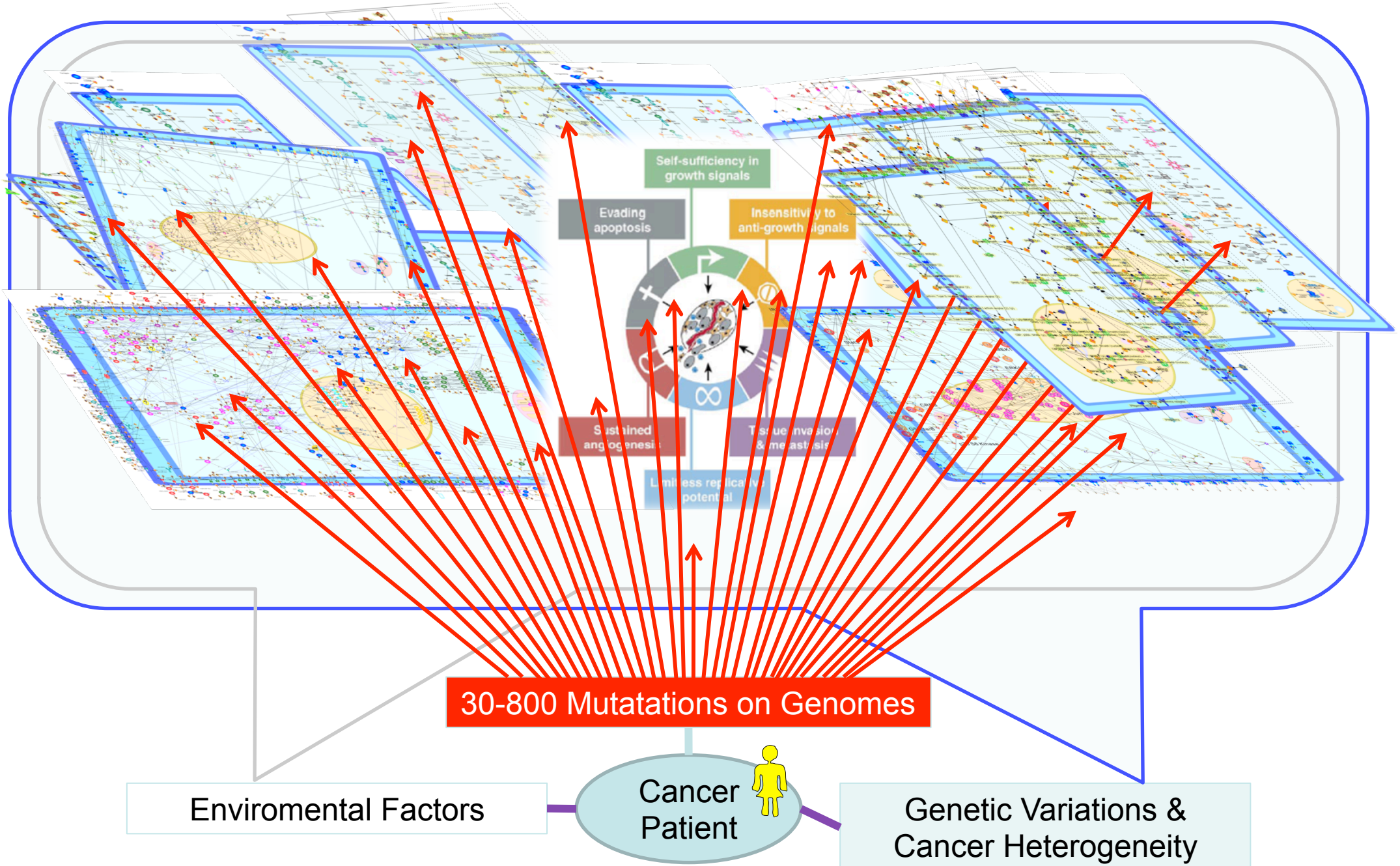
Data Assimilation Module

→ Li, Kuroyanagi et al.

- For a CIO pathway model, the module executes the user specified multiple initial conditions at once and displays the result with 2D or 3D plots.



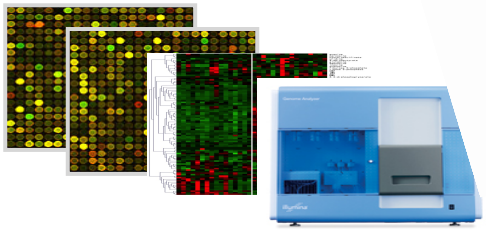
Back to Cancer


Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
30-800 Mutatations on Genomes
Enviromental Factors
Cancer Patient
Genetic Variations & Cancer Heterogeneity

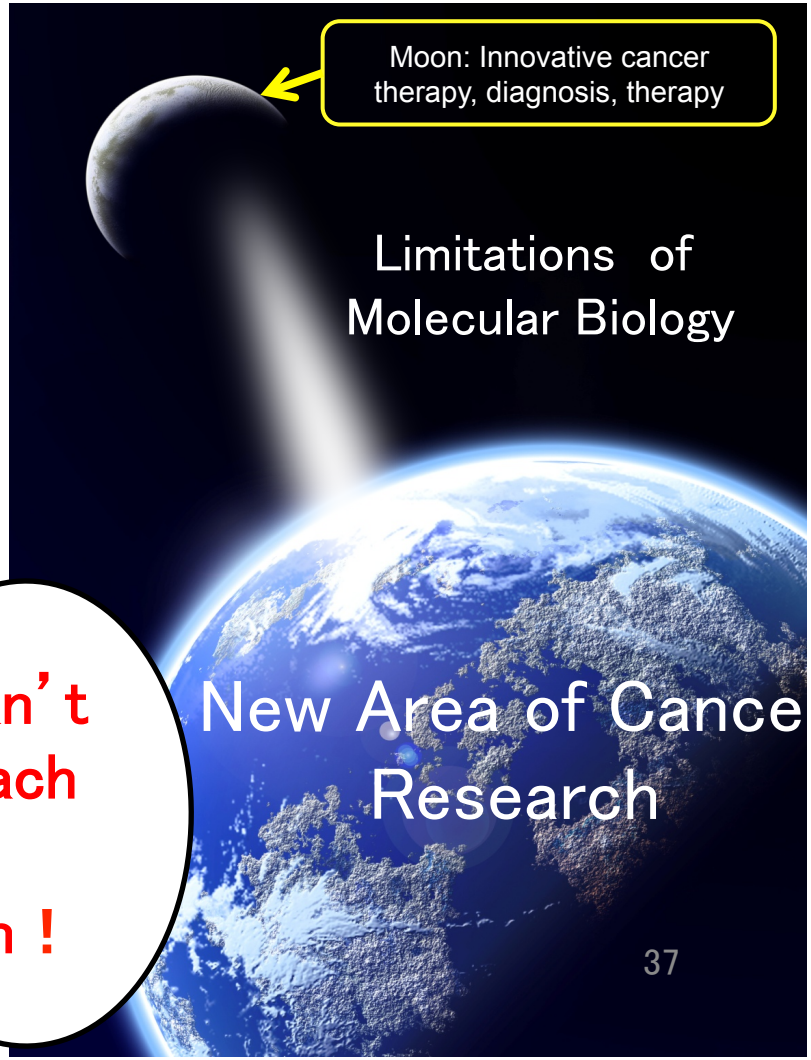
Personal Genomes
Integration of Omics Data

$10^3 \sim 10^4$

Many patient samples

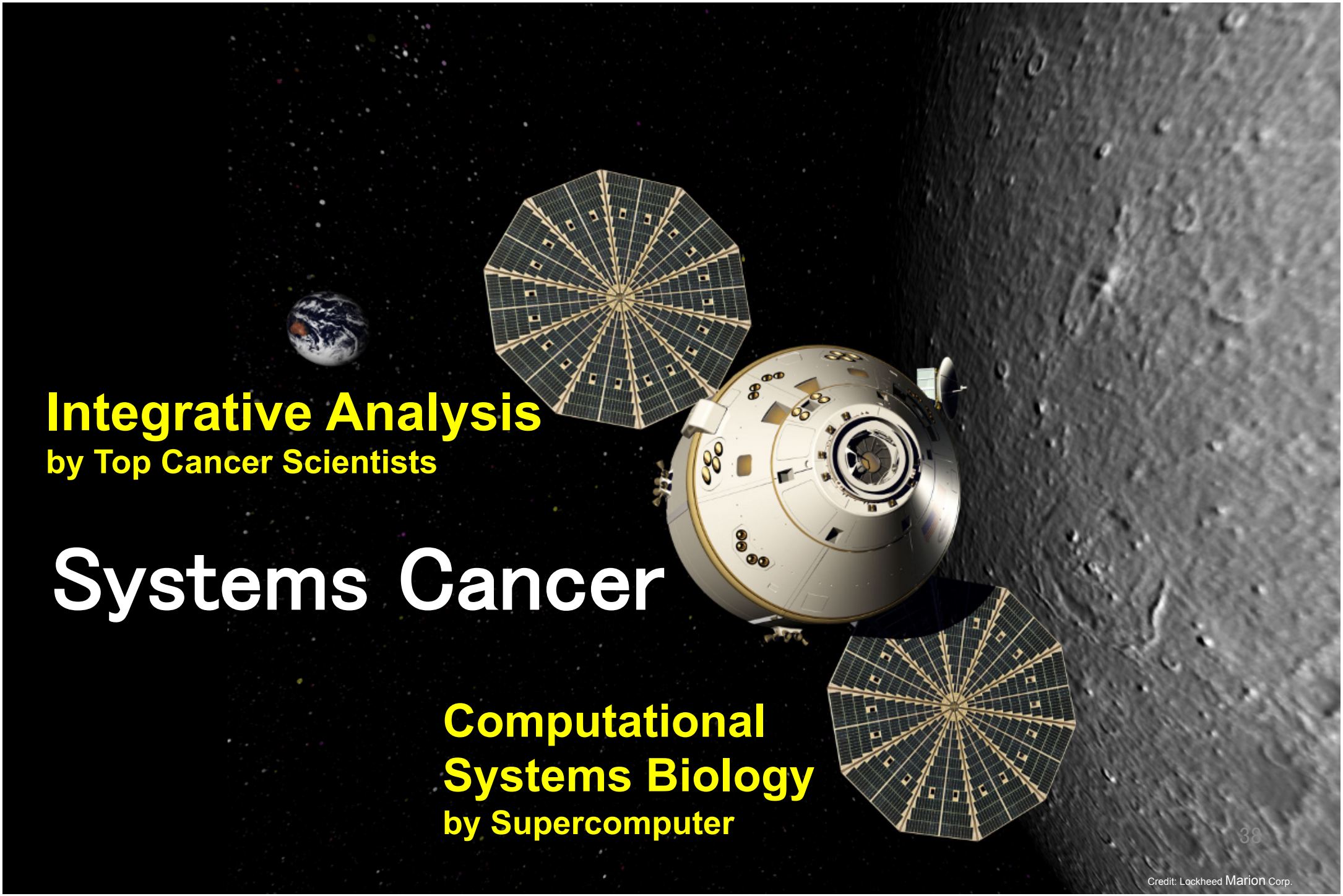


Can't reach by run !



High-Throughput Technology

Microarray, ChIP-Chip, CGH array, SNP array, DNA-Seq, Exome-Seq, RNA-Seq, ChIP-Seq, …


Integrative Analysis
by Top Cancer Scientists
Systems Cancer
Computational
Systems Biology
by Supercomputer
38
Credit: Lockheed Marion Corp.

Systems Cancer Research Project by MEXT (2010-2014)

Integrative Systems Understanding for Advanced Diagnosis, Therapy and Prevention

Oncology & Cancer Biology

Scientific research on innovative areas

個人のがんシステムをデジタル化するには——
スーパーコンピューターインフラと最先端の生命システム解析技術を生かし、
現在のがん研究が直面している限界を超えて、がん研究の水準を飛躍的に向上・強化させる。

Computational Systems Biolog

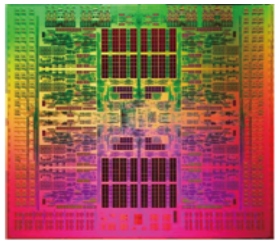
Supercomputer

http://systemscancer.hgc.jp

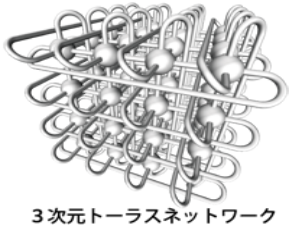

Shunpei
Ishikawa,MD
Tomoyoshi Soga
Takashi
Takahashi,MD
Satoru Miyano
Johji
Inazawa,MD
Seishi
Ogawa,MD
Tatsuhiko Tsunoda
FMACS
PAC-9

Next-Generation Supercomputer Project

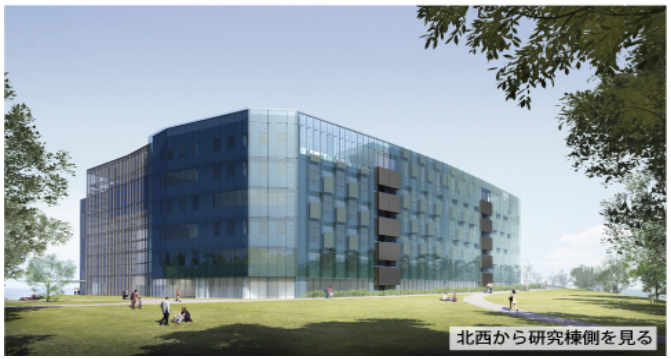
1,15000,000,000JPY for 7 years (2006-2012)



SPARC64™ VIIIfx
©Fujitsu Limited

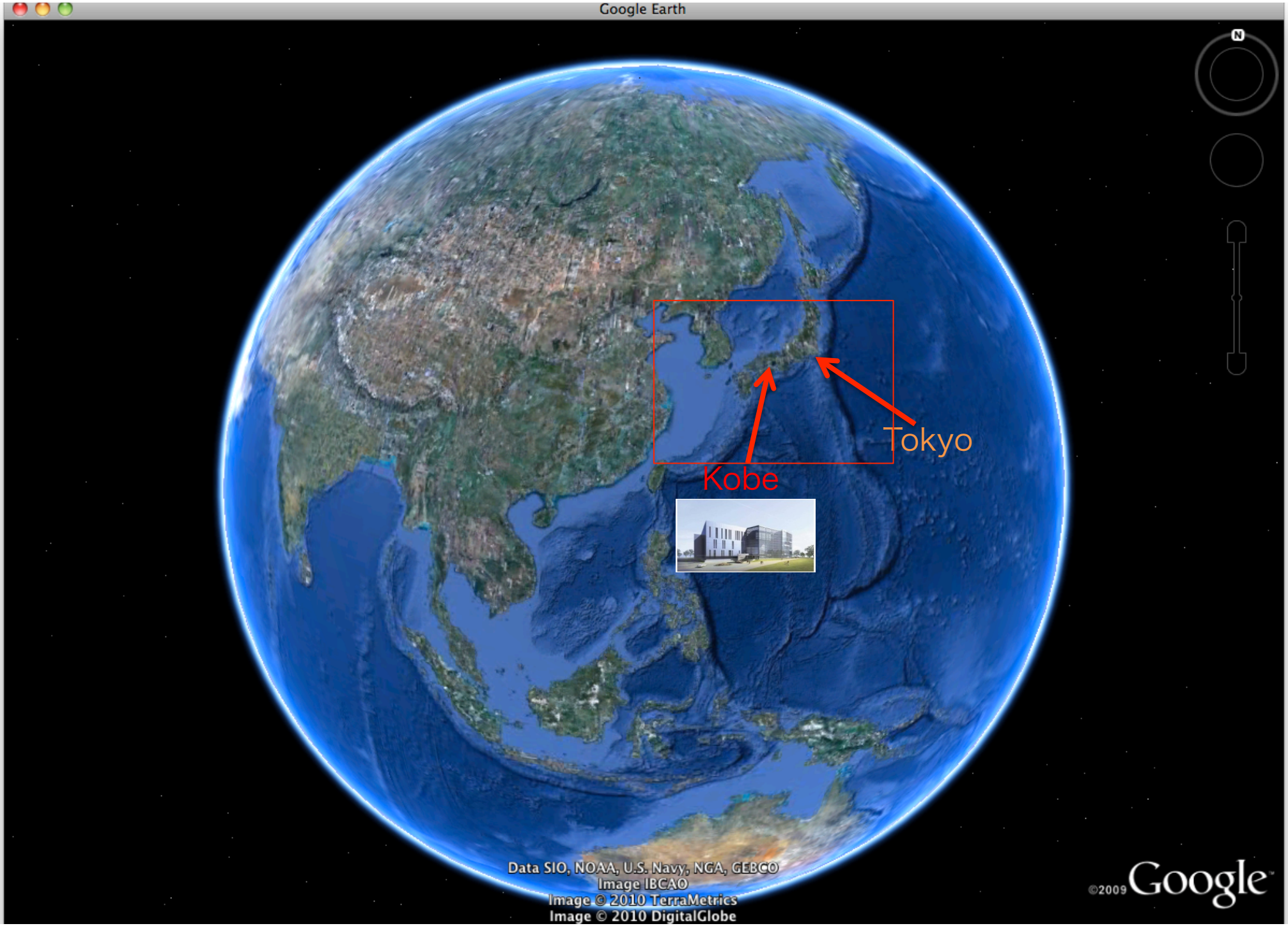


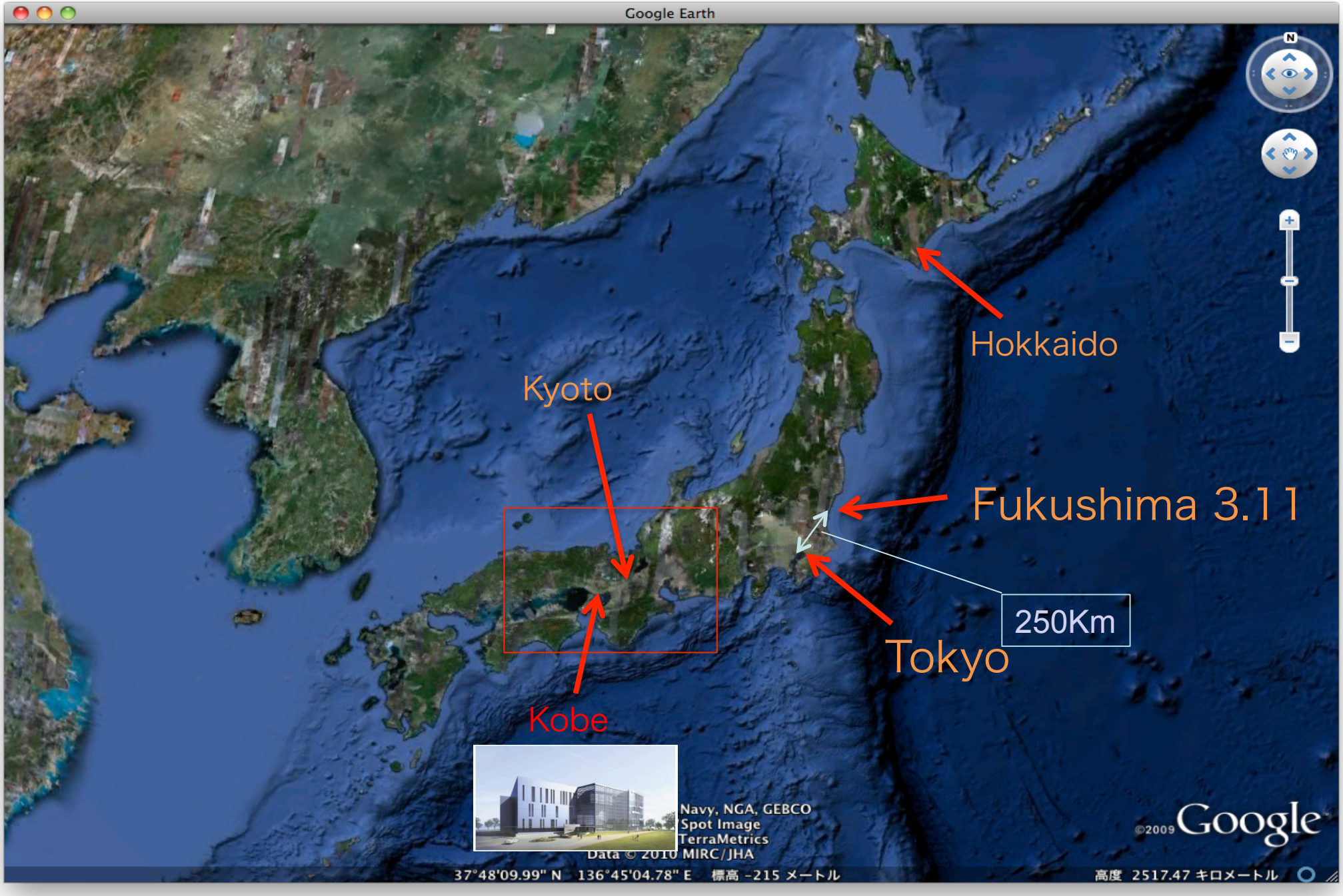
3次元トーラスネットワーク
(概念図)
©Fujitsu Limited

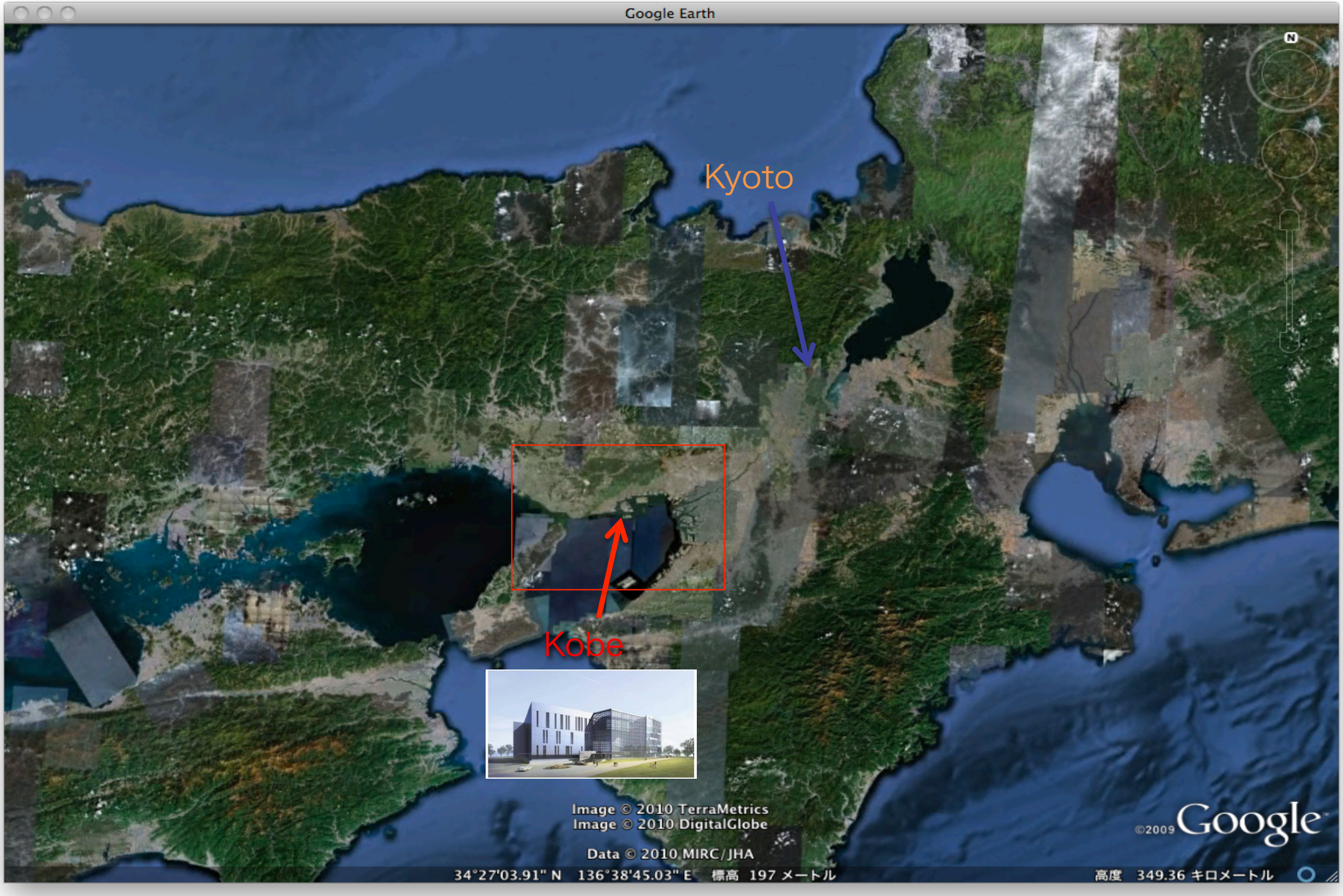


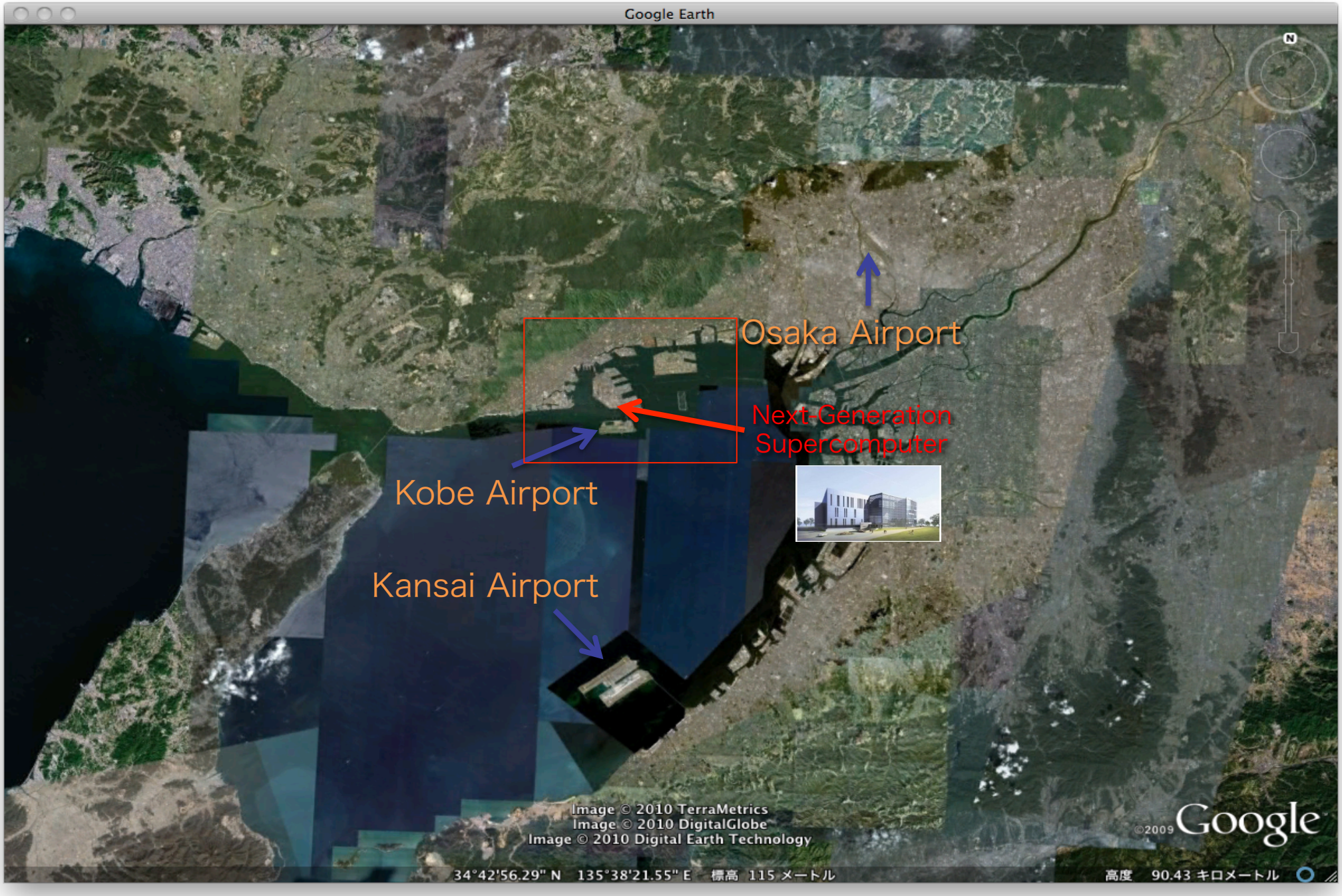
- Design, build, and set up the **Next-Generation Supercomputer** with a speed of **10 PETA FLOPS**.
- Over 80,000 nodes & 640,000 cores.

- Kobe（神戸）


Google Earth
Tokyo
Kobe
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image IBCAO
Image © 2010 TerraMetrics
Image © 2010 DigitalGlobe
©2009 Google


Google Earth
Hokkaido
Kyoto
Fukushima 3.11
250Km
Tokyo
Kobe
Navy, NGA, GEBCO
Spot Image
TerraMetrics
Data © 2010 MIRC/JHA
©2009 Google
37°48'09.99" N 136°45'04.78" E 標高 -215 メートル
高度 2517.47 キロメートル


Google Earth
Kyoto
Kobe
Image © 2010 TerraMetrics
Image © 2010 DigitalGlobe
Data © 2010 MIRC/JHA
©2009 Google
34°27'03.91" N 136°38'45.03" E 標高 197 メートル
高度 349.36 キロメートル


Google Earth
Osaka Airport
Next-Generation
Supercomputer
Kobe Airport
Kansai Airport
Image © 2010 TerraMetrics
Image © 2010 DigitalGlobe
Image © 2010 Digital Earth Technology
©2009 Google
34°42'56.29" N 135°38'21.55" E 標高 115 メートル
高度 90.43 キロメートル


Google Earth
“Rokko Island”
(artificial island)
“Port Island”
(artificial island)
Next-Generation
Supercomputer
Kobe Airport
Image © 2010 Digital Earth Technology
Image © 2010 DigitalGlobe
©2009 Google
画像取得日: 2007年3 月31日
34°41'45.44" N 135°16'32.62" E 標高 7 メートル
高度 13.43 キロメートル


Google Earth
Next-Generation
Supercomputer
Kobe Airport
Image © 2010 DigitalGlobe
©2009 Google
画像取得日: 2007年3 月31日
34°38'39.57" N 135°13'33.42" E 標高 -3 メートル
高度 3.26 キロメートル

Google Earth

Next-Generation
Supercomputer

©2009 Google

画像取得日: 2007年3 月31日　34°39'03.37" N　135°13'18.02" E　標高　5 メートル　高度　698 メートル

Cancer System Hacker


京

K computer No.1
International Supercomputing 2011
Hamburg, June 19, 2011

80% K computer (68,544 Nodes)
(8.774PFLOPS at Peak)

LINPACK 8.162PFLOPS（93.0%）

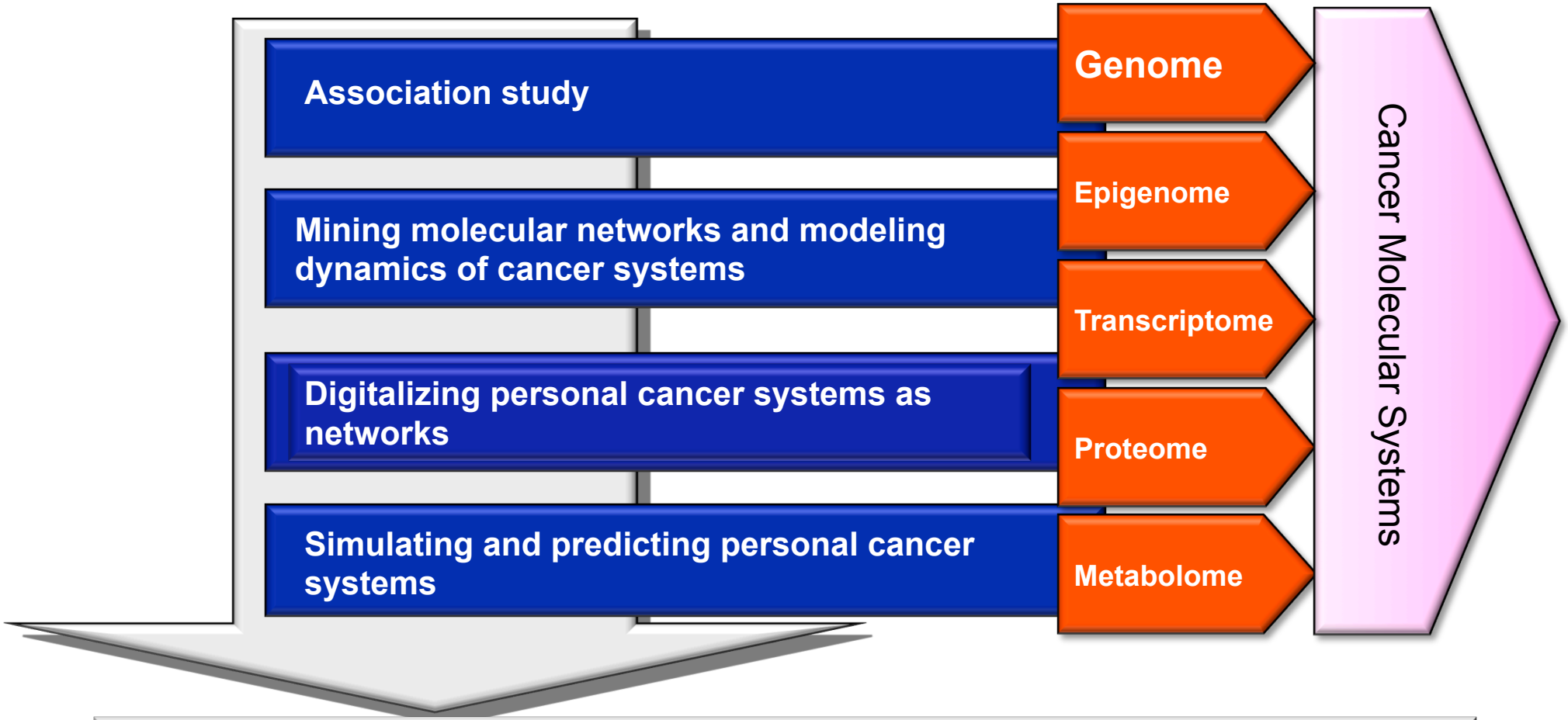
Big Contributor

Supercomputer System for 2009-2014

Human Genome Center
Institute of Medical Science, University of Tokyo

- **January 2009**: 75 TFLOPS at peak & 1 PB Disk Space
  - PC Cluster (Sun Microsystems)
  - Large Shared Memory Machine (SGI Altix)
  - Lustre File System (Sun Microsystems)
- **January 2012**: 225 TFLOPS at peak & 4PB Disk Space

How to Hack Cancer Systems



Integrative Systems Understanding of Cancer for Advanced Diagnosis, Therapy and Prevention

How to Hack Cancer Systems

- Digitalizing Heterogeneity/Characteristics of Cancer Systems
- Bayesian Gene Networks of Cancer Systems
- Modeling Dynamics in Cancer Systems with State Space Model
- Comparing Networks in Cancer Systems under Different Conditions
- Extracting Functional Modules in Cancer Systesm
- Software Platform – eXtensible integrative Pipeline & Cell Illustrator

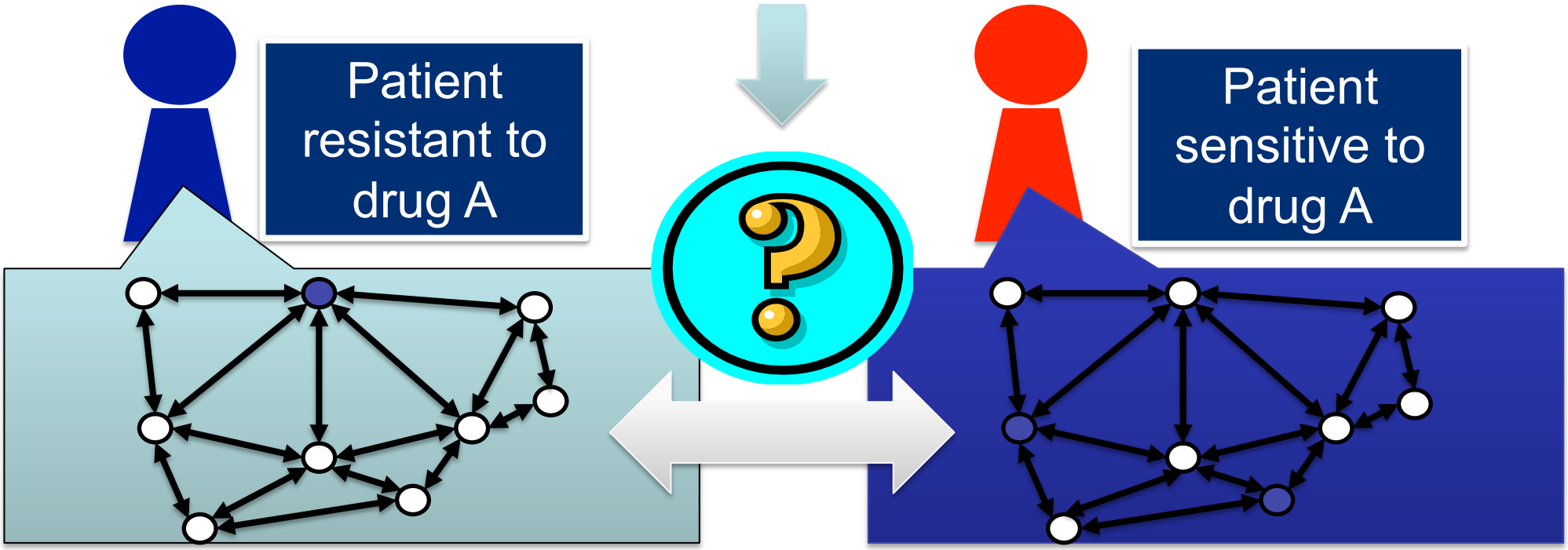
1. Digitalizing Cancer Systems Towards Personal Gene Networks

Teppei Shimamura, Seiya Imoto, Yukako Shimada, Yasuyuki Hosono, Atsushi Niida, Masao Nagasaki, Rui Yamaguchi, Takashi Takahashi, Satoru Miyano, PLoS ONE. 6(6): e20804, 2011.

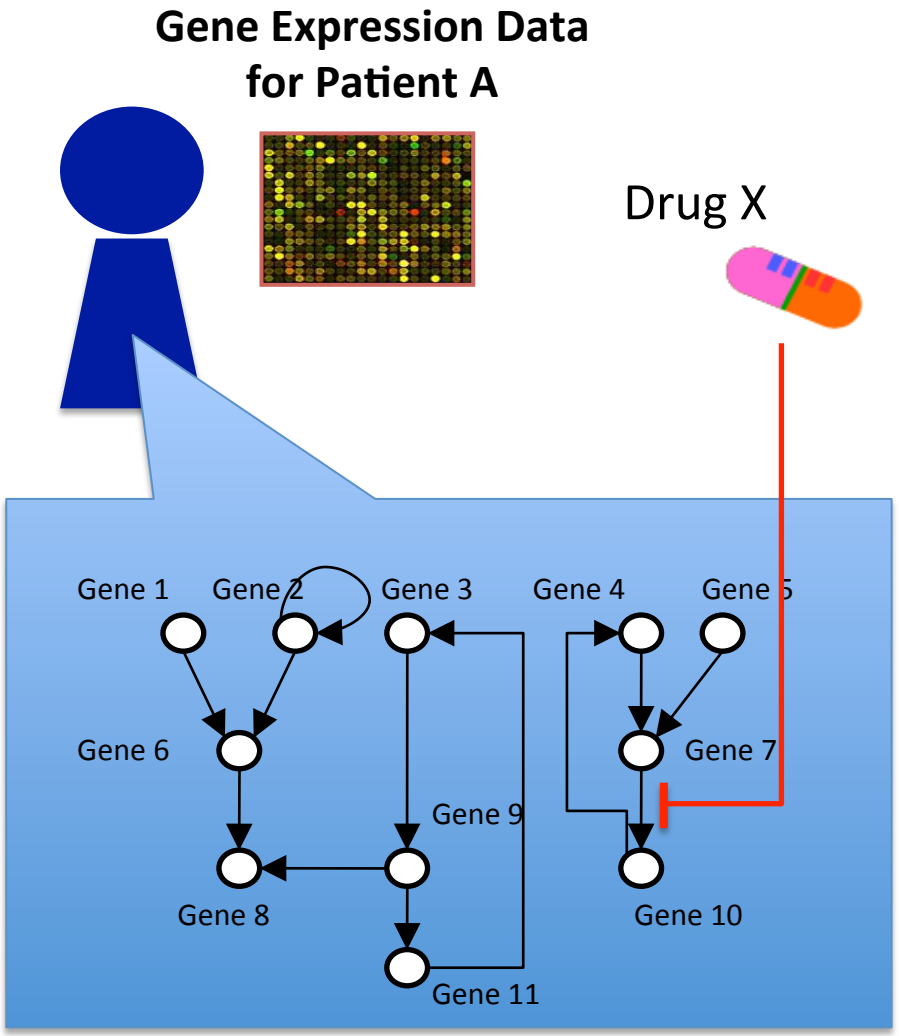
Cancer Heterogeneity and Individual Variation

High-Throughput Technology

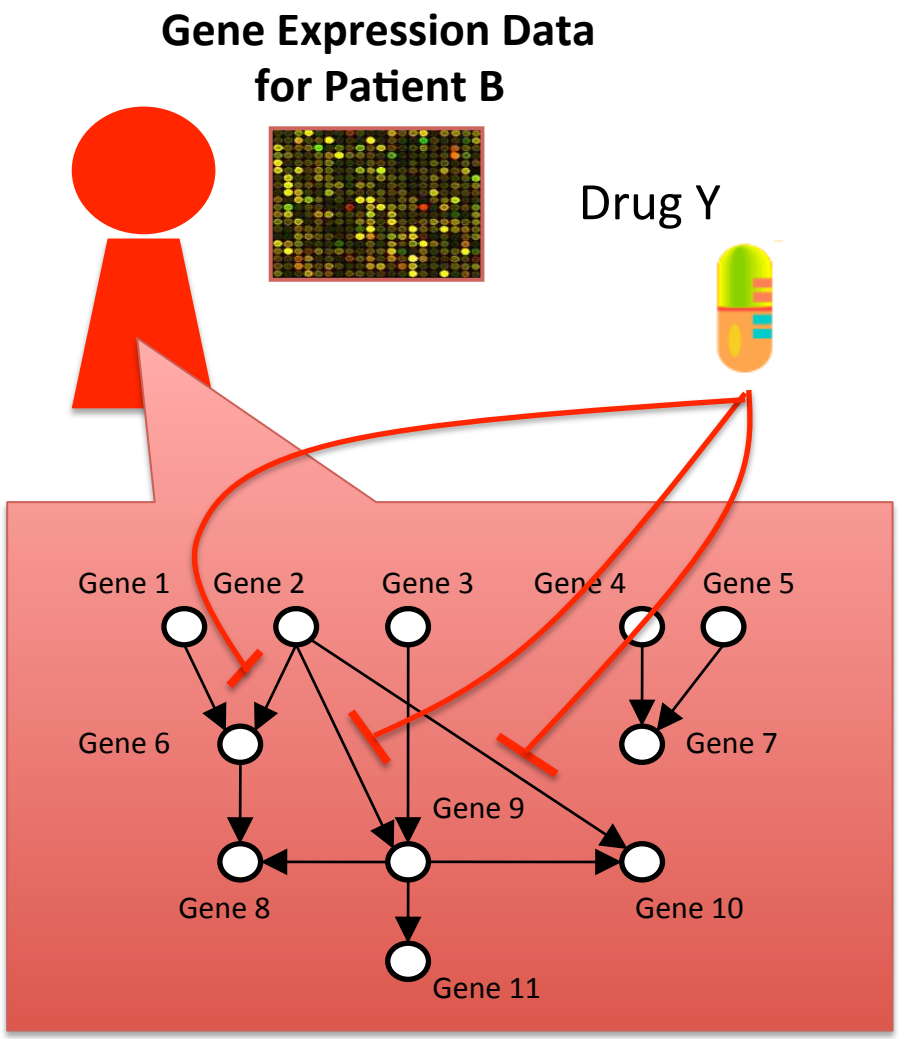
Microarray, ChIP-Chip, CGH array, SNP array, DNA-Seq, Exome-Seq, RNA-Seq, ChIP-Seq, …



Patient-Specific Gene Network



Cellular System for Patient A



Cellular System for Patient B

Traditional Gene Network Estimation Problem

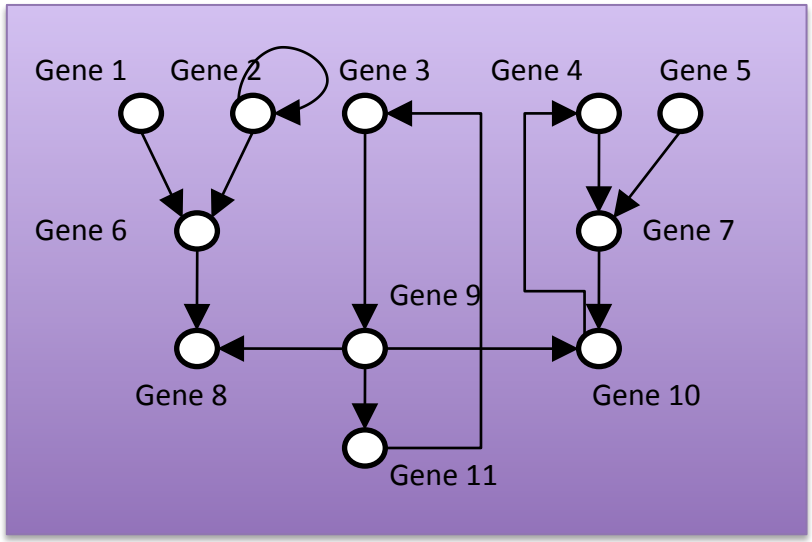
Knock-down Microarray Data for Sample A

	sample 1	sample 2	sample 3	....
gene 1	1.5	5.2	1.4	....
gene 2	5.2	6.3	0.4	....
gene 3	3.4	9.3	0.3	....
gene 4	2.9	0.3	6.4	....
:	:	:	:	:
:	:	:	:	:

Time-series Microarray Data for Sample A

	time 1	time 2	time 3	....
gene 1	1.5	5.2	1.4	....
gene 2	5.2	6.3	0.4	....
gene 3	3.4	9.3	0.3	....
gene 4	2.9	0.3	6.4	....
:	:	:	:	:
:	:	:	:	:

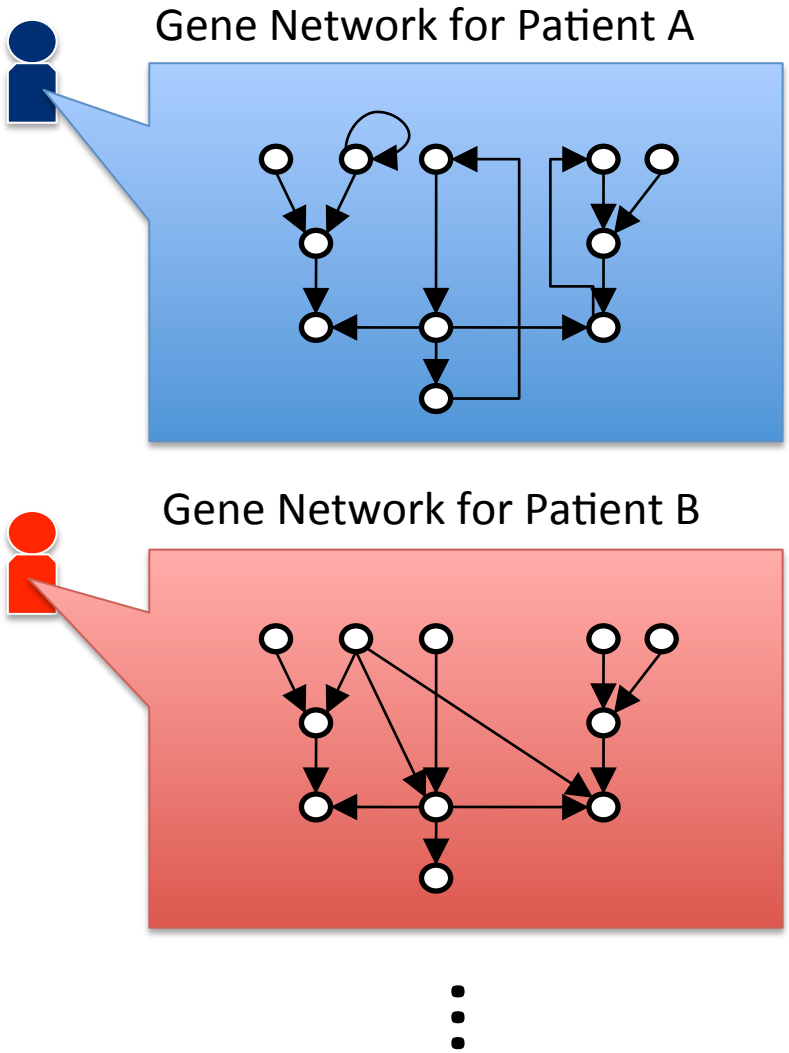
Gene Network for Sample A



Patient-Specific Gene Network Estimation

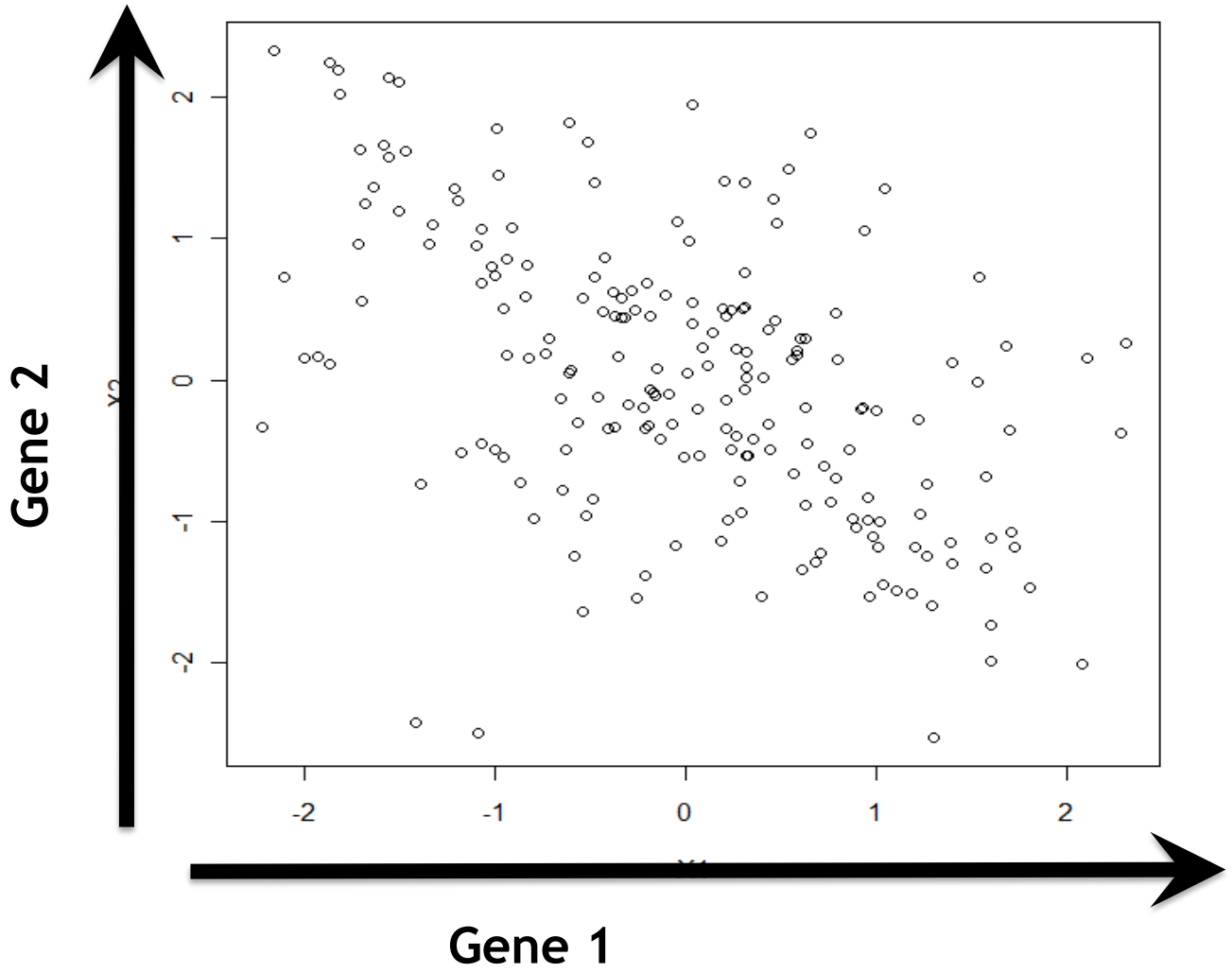
Microarray Data for Cancer Patients

	patient 1	patient 2	patient 3	....
gene 1	1.5	5.2	1.4	....
gene 2	5.2	6.3	0.4	....
gene 3	3.4	9.3	0.3	....
gene 4	2.9	0.3	6.4	....
:	:	:	:	:
:	:	:	:	:

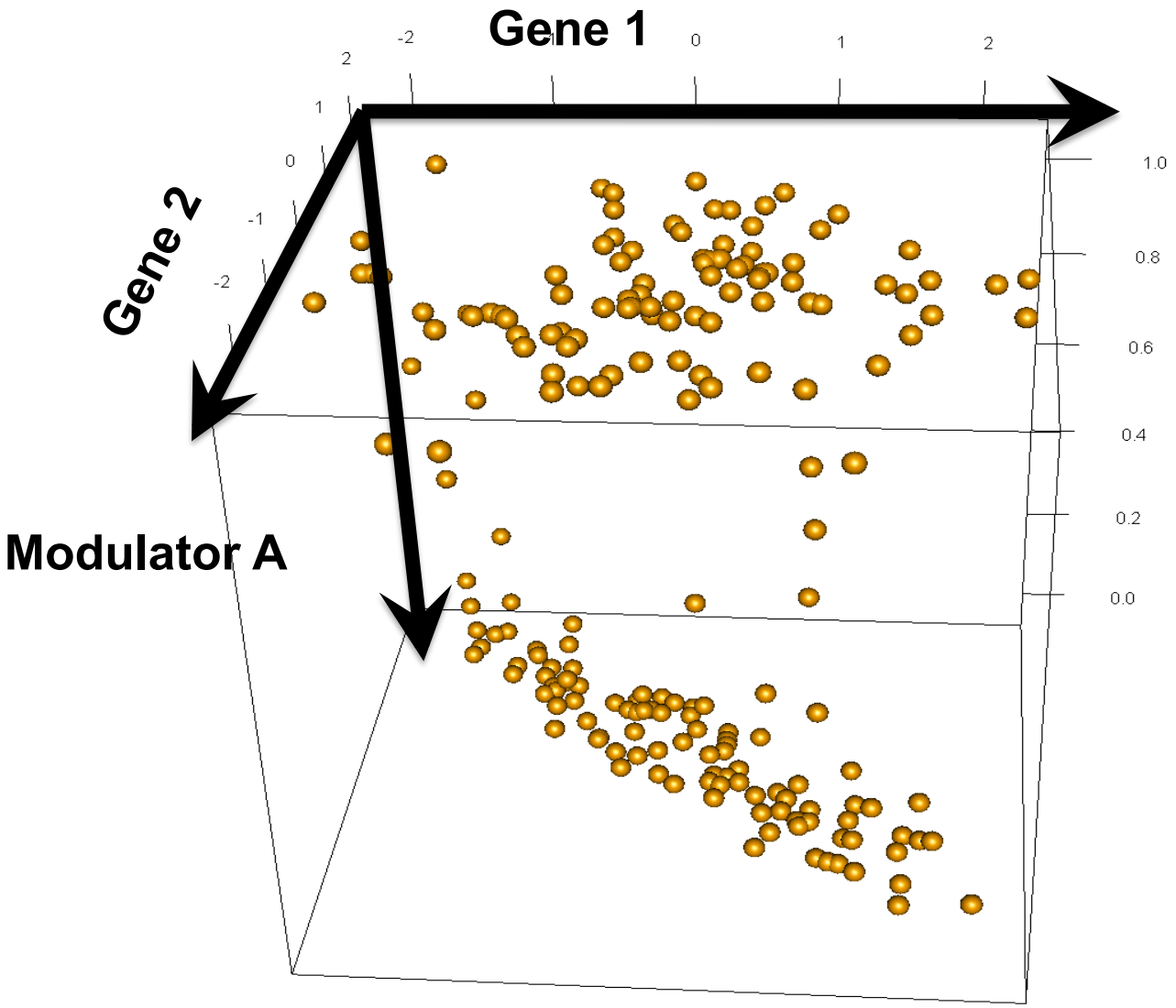


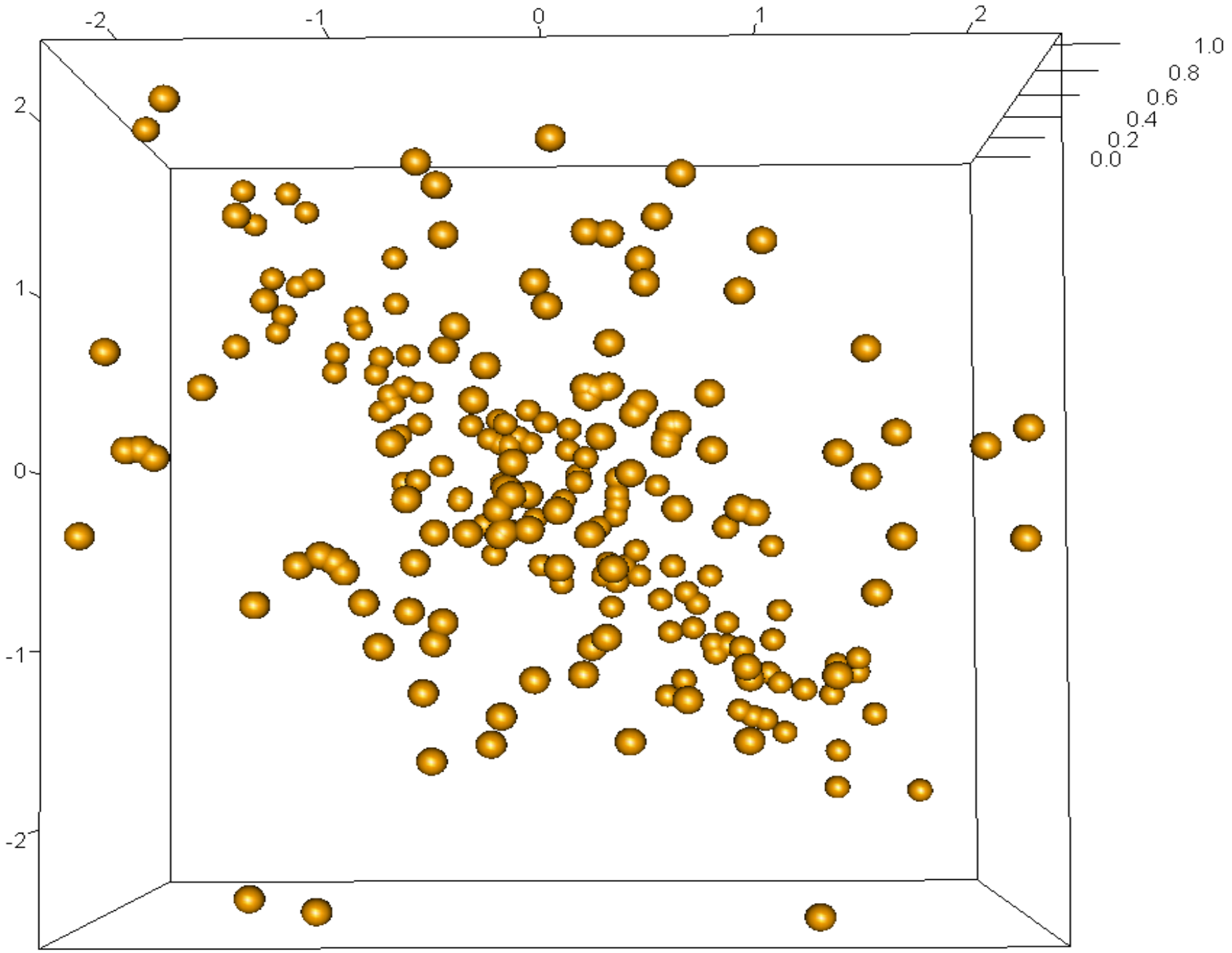
Corelation between 2 genes

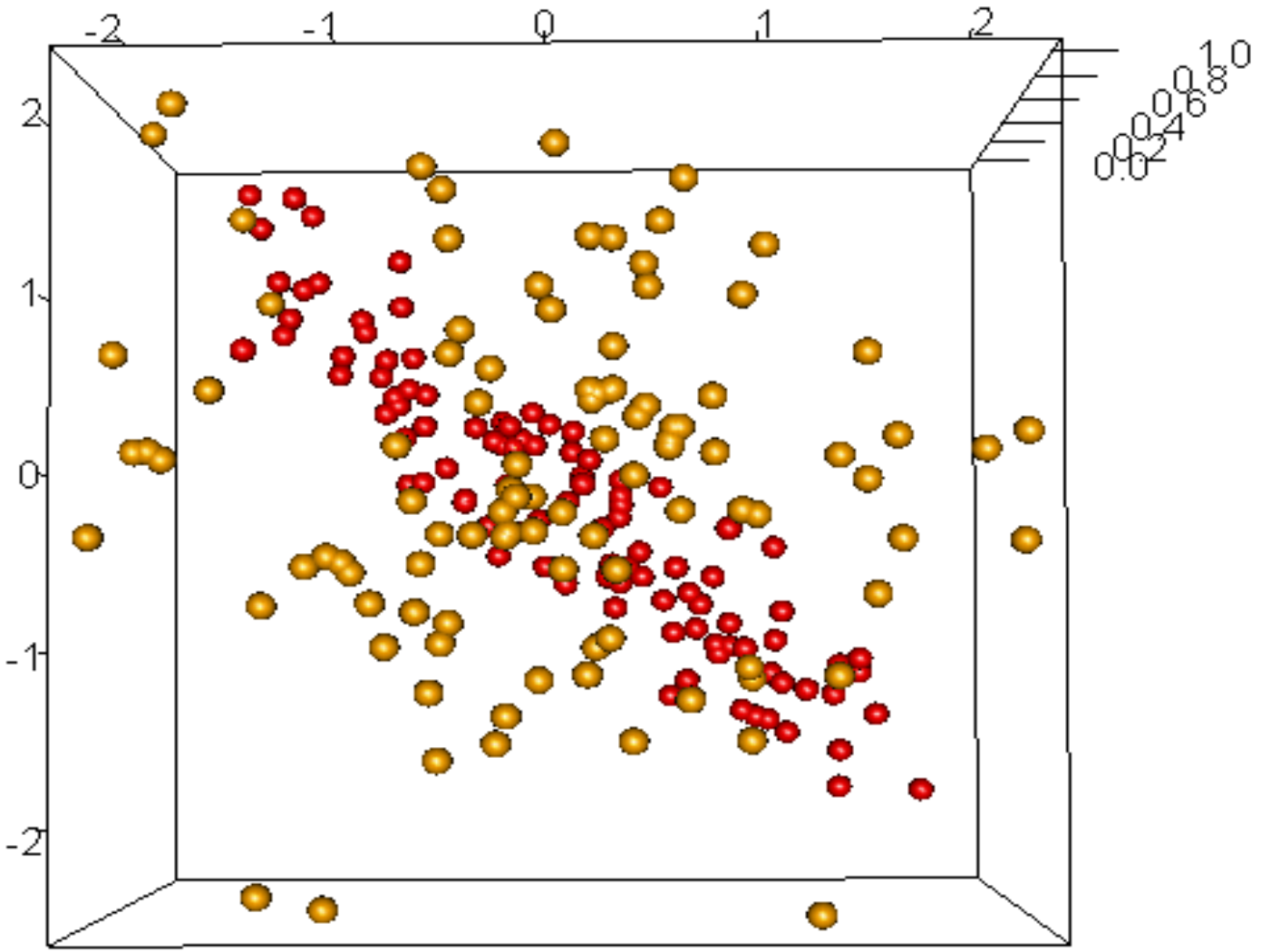
No relation?



Change can be found if we look with a modulator



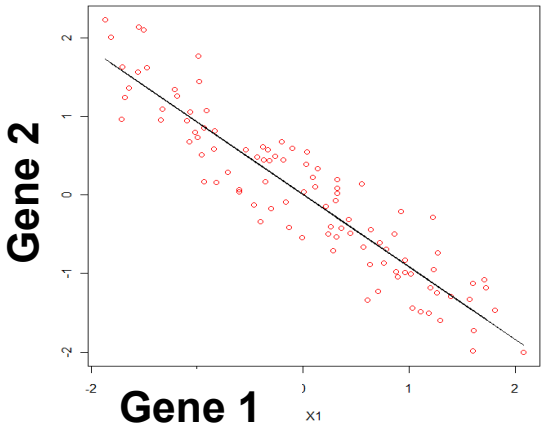

-2
-1
0
1
2
2
1
0
-1
-2
1.0
0.8
0.6
0.4
0.2
0.0

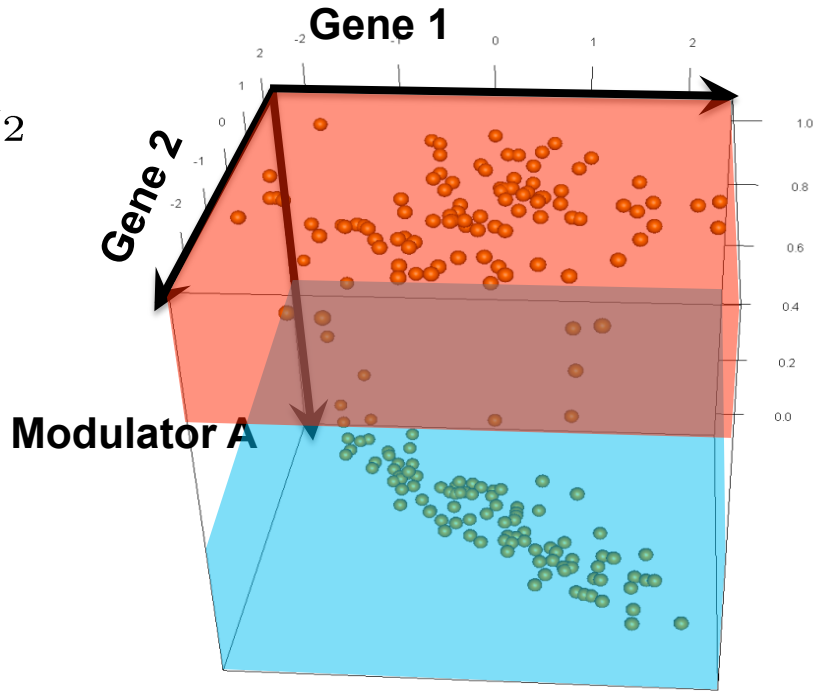

-2
-1
0
1
2
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0

Finding Gene Relation by Data Stratification

$$X_2 = -0.92 \times X_1 + \epsilon_2$$

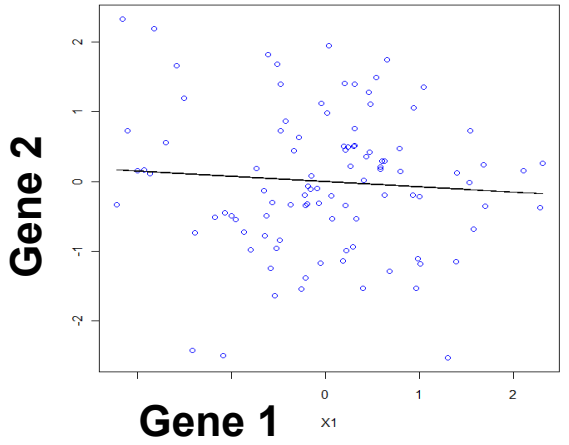
Modulator A is low



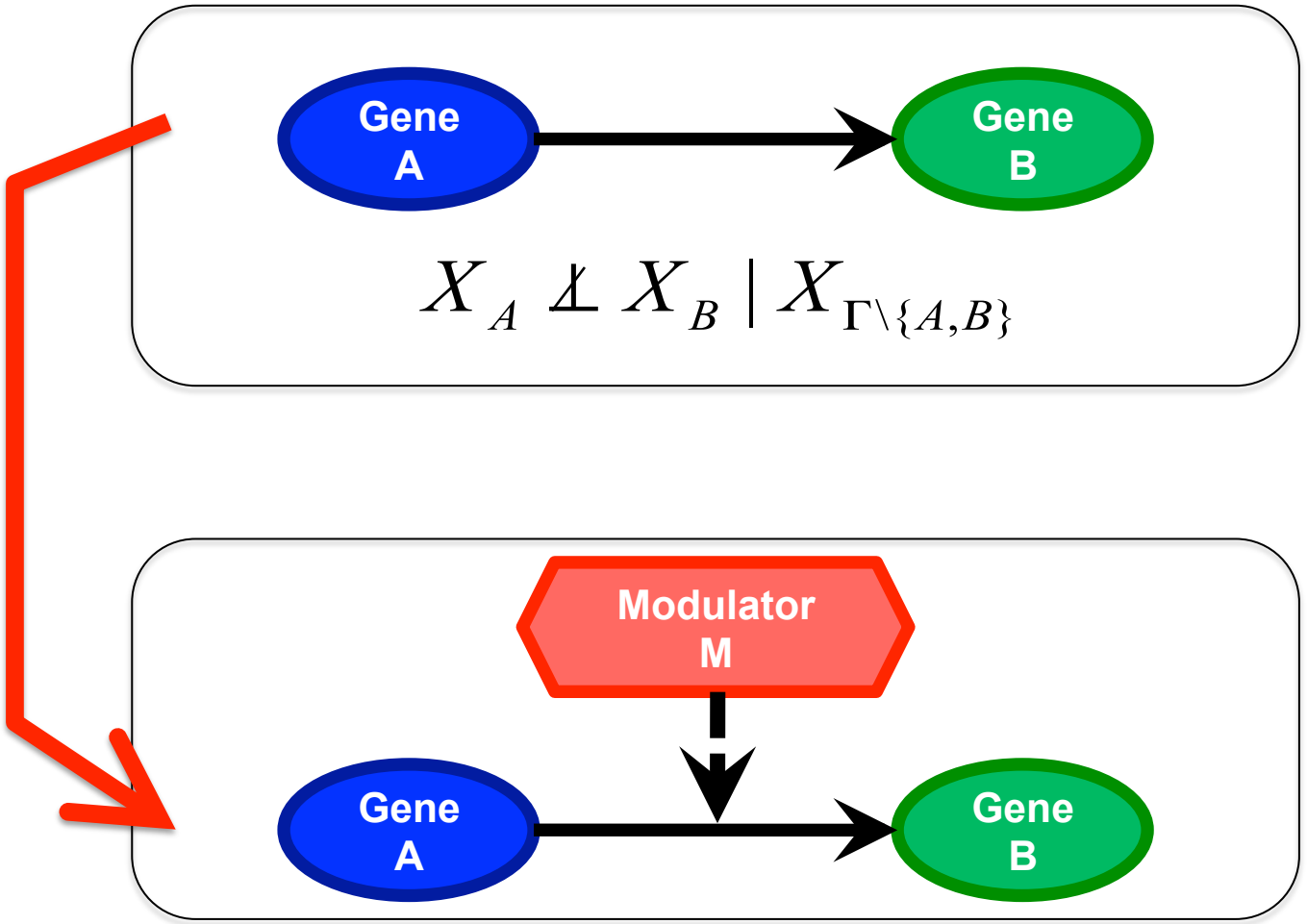


$$X_2 = -0.07 \times X_1 + \epsilon_2$$

Modulator A is high

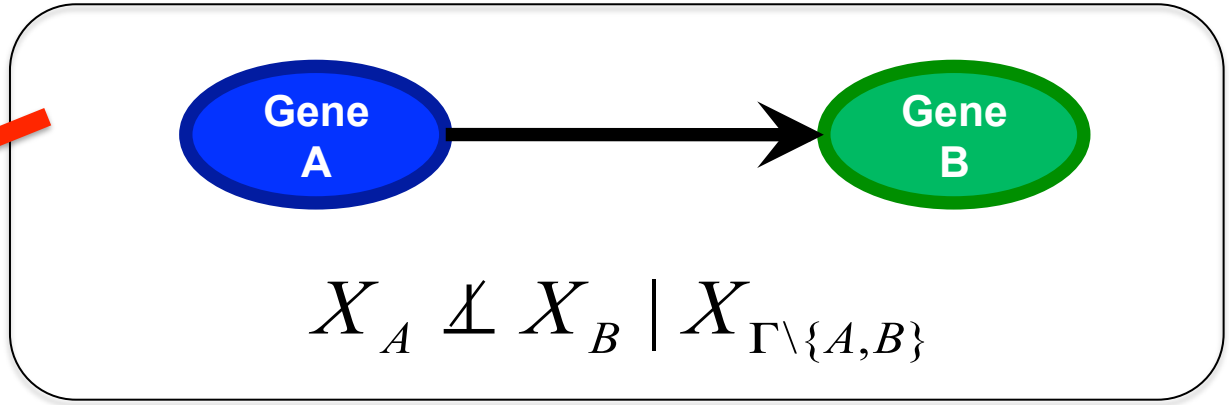


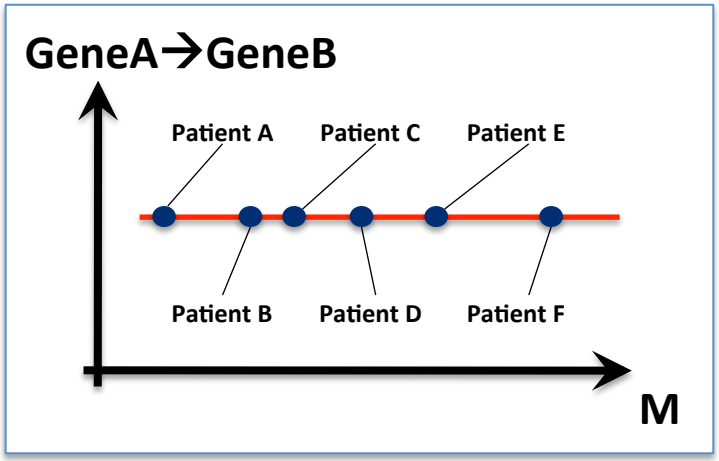
Stratification of Data


Gene
A
Gene
B
$X_A \not\perp X_B \mid X_{\Gamma\setminus\{A,B\}}$
Modulator
M
Gene
A
Gene
B

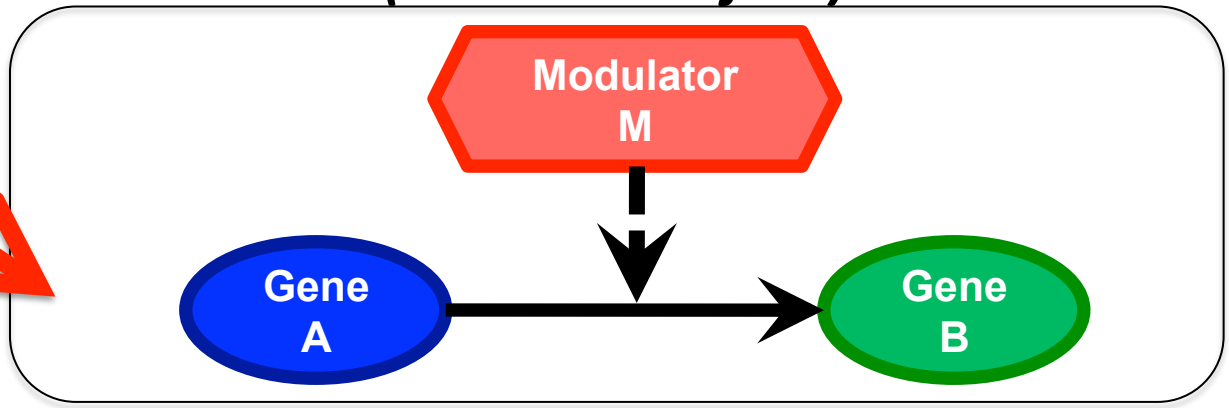
Concept of NetworkProfiler

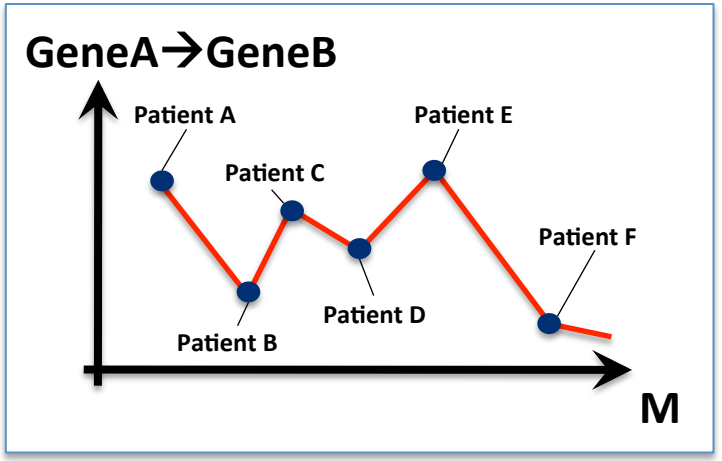
Usual graphical model





Modulator-dependent graphical model (NetworkProfiler)



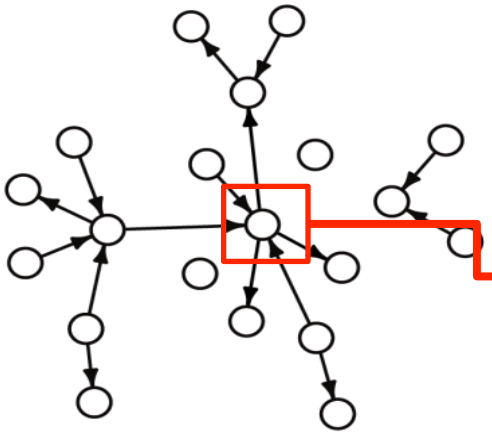


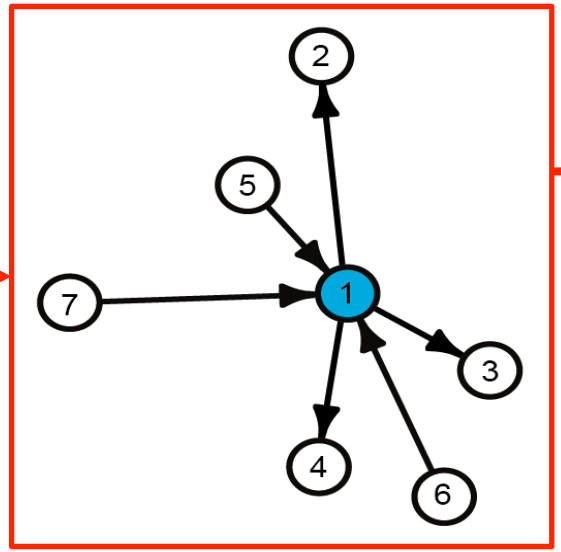
Modulator: **Cofactor** modulating relationship between genes A and B

Examples of Modulator

- Tumor progression (Stage I, Stage II, …)
- Drug sensitivity (IC50, GI50, …)
- Disease-free survival
- Molecular characteristics (Metastasis, EMT…)
- Pathway activity
- …

What is System?



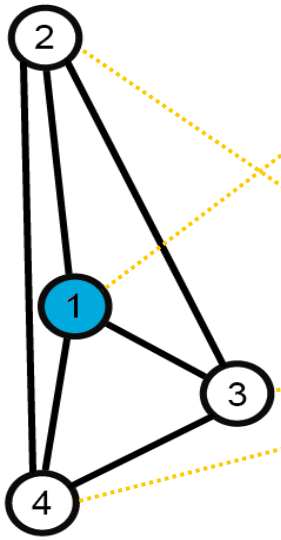


Structural Equations

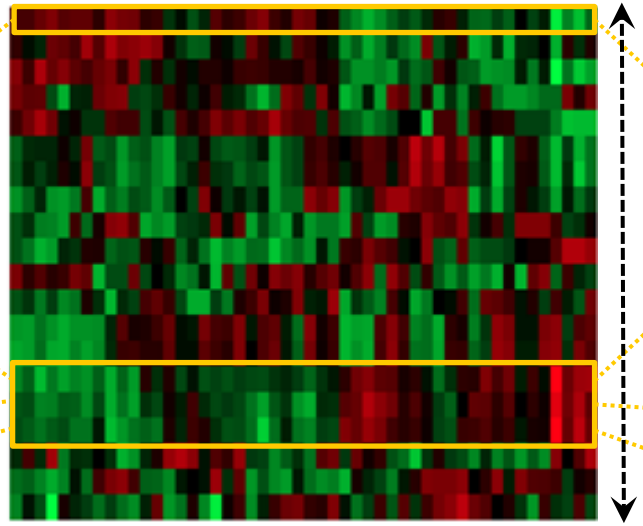
$$
\begin{aligned}
x_1 &= 4.6 \times x_5 + 4.2 \times x_6 + 0.7 \times x_7 + \epsilon_1 \\
x_2 &= -2.7 \times x_1 + 0.95 \times x_8 + \epsilon_2 \\
x_3 &= -1.4 \times x_1 + \epsilon_3 \\
x_4 &= -0.5 \times x_1 + \epsilon_4
\end{aligned}
$$

Node: **gene transcript**
Edge: **conditional dependence**
(not equal to correlation)

Co-expression Network

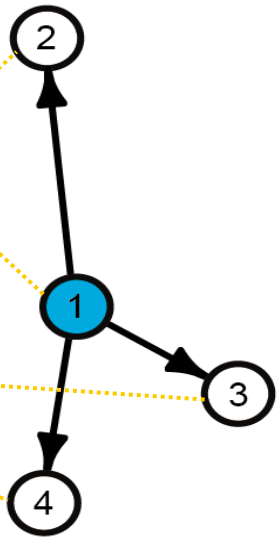


Samples



Genes

Gene Network



NetworkProfiler

p (genes) × n (patients) gene expression data matrix

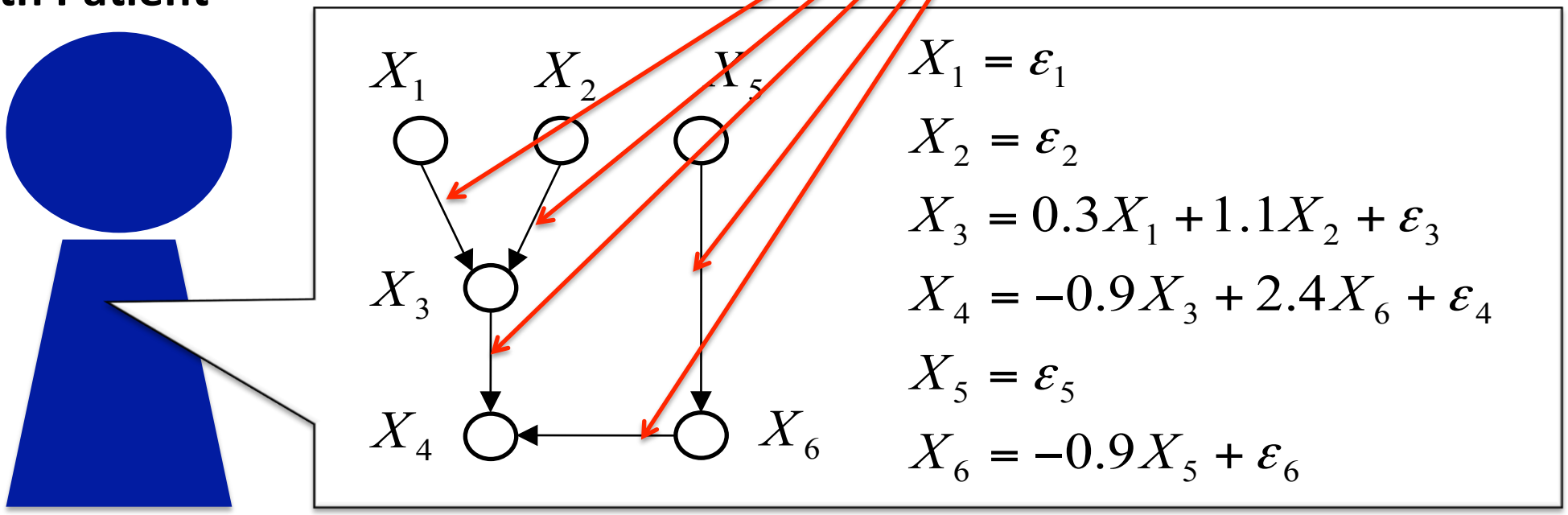
$x_{\alpha j}$; expression value of j-th gene for α-th patient

m_{α} ; modulator value for α-th patient

Structural equation model of *k*-th gene for α-th patient

$$x_{\alpha k} = \sum_{j=0, j\neq k}^{p} \beta_{jk\alpha} x_{\alpha j} + \varepsilon_{\alpha k}, \quad \beta_{jk\alpha} = \beta_{jk}(m_{\alpha})$$

α-th Patient



NetworkProfiler

p (genes) × n (patients) gene expression data matrix

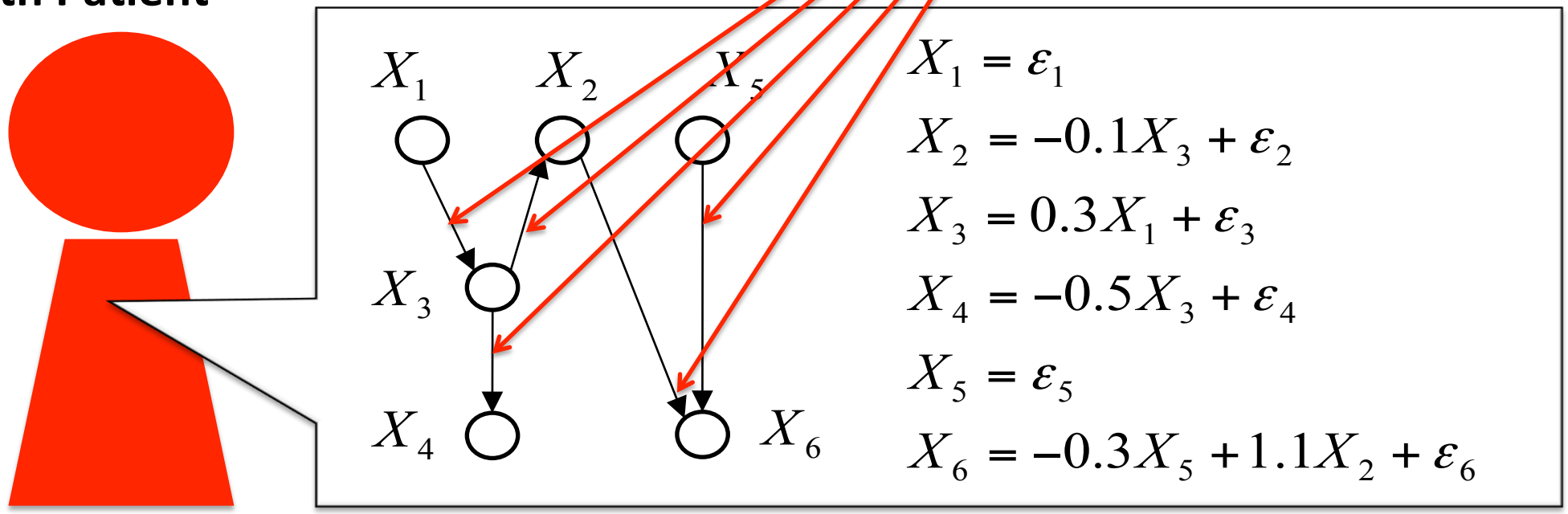
$x_{\beta j}$; expression value of j-th gene for β-th patient

m_{β} ; modulator value for β-th patient

Structural equation model of *k*-th gene for *β*-th patient

$$x_{\beta k} = \sum_{j=0, j\neq k}^{p} \beta_{jk\beta} x_{\beta j} + \varepsilon_{\beta k}, \quad \beta_{jk\beta} = \beta_{jk}(m_{\beta})$$

***β*-th Patient**

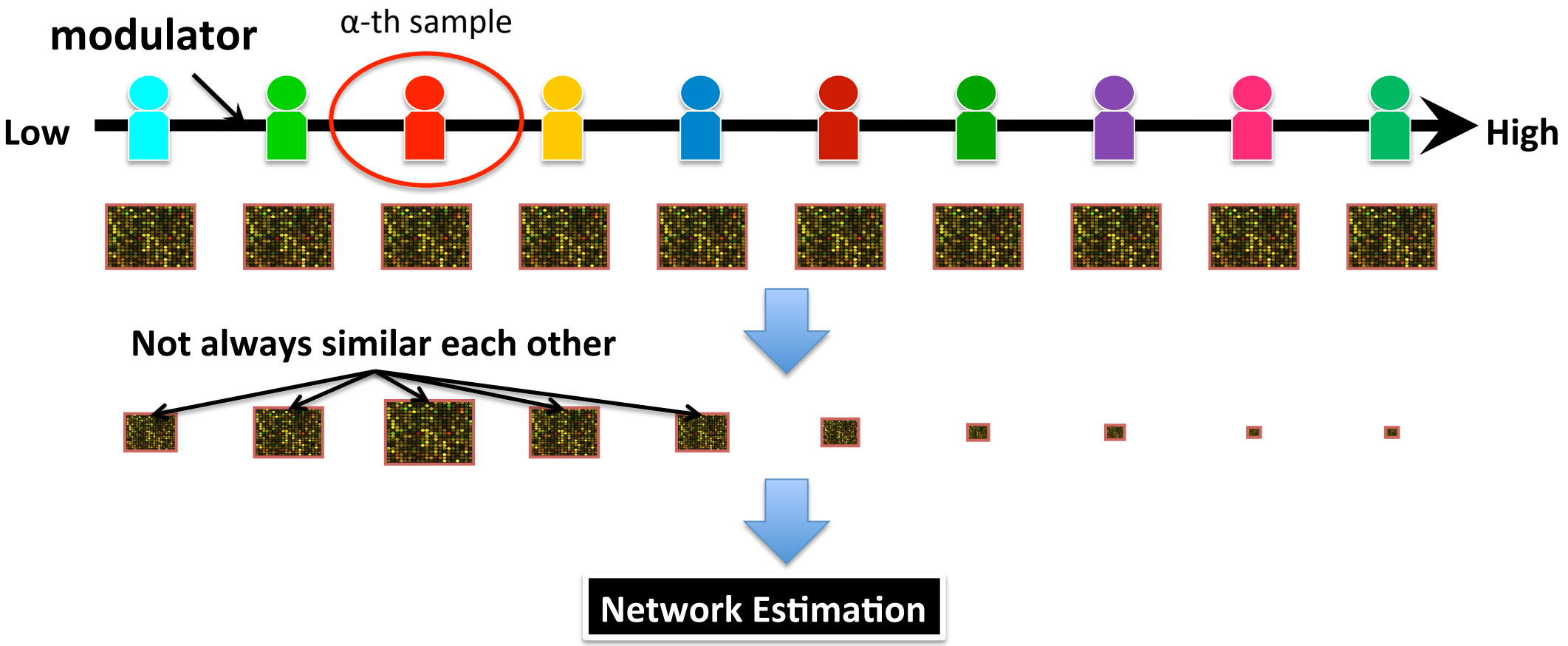


Concept of NetworkProfiler

Purpose: estimation of **coefficients** for α-th sample $\beta_{jk\alpha}$

Problem: microarray for α-th sample is only **one**!!

Idea: data integration by **sample weighting**

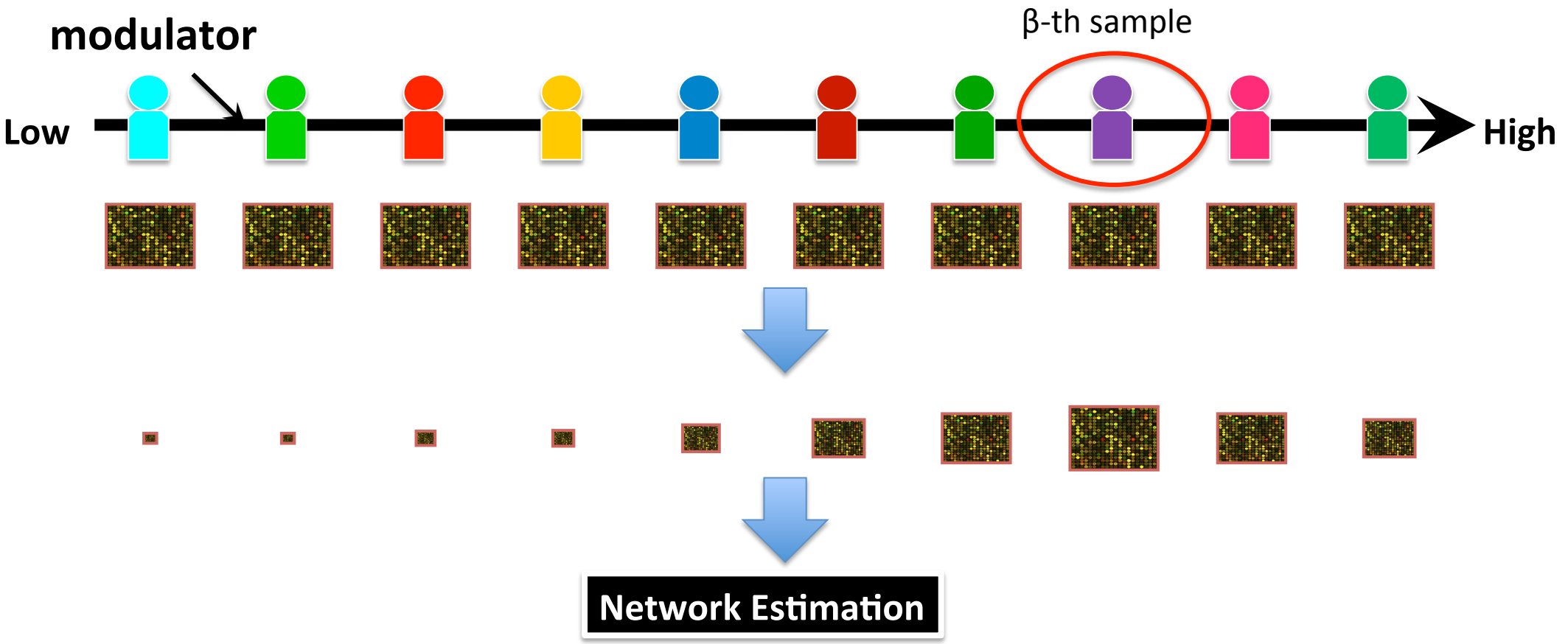


Concept of NetworkProfiler

Purpose: estimation of **coefficients** for β-th sample $\beta_{jk\beta}$

Problem: microarray for β-th sample is only **one**!!

Idea: data integration by **sample weighting**



NetworkProfiler

Structural equation model of *k*-th gene for α-th patient

$$x_{\alpha k} = \sum_{j=0, j\neq k}^{p} \beta_{jk\alpha} x_{\alpha j} + \varepsilon_{\alpha k}, \quad \beta_{jk\alpha} = \beta_{jk}(m_\alpha)$$

Elastic net-type regularized weighted loss function

$$S(\beta_{1k\alpha},\ldots,\beta_{pk\alpha} \mid m_\alpha) = \sum_{i=1}^{n} K_h(m_i - m_\alpha)\left\{ x_{ik} - \sum_{j\neq k}^{p} \beta_{jk\alpha} x_{ij} \right\}^2 + \lambda_1 \sum_{j\neq k} w_{jk\alpha} \left|\beta_{jk\alpha}\right| + \frac{\lambda_2}{2} \sum_{j\neq k} \beta_{jk\alpha}^2$$

$K_h(m_i - m_\alpha)$: Gaussian kernel function (m_α: center, h: width)
λ_1, λ_2: regularization parameters

$$\hat{\boldsymbol{\beta}}_{k\alpha} = \left(\hat{\beta}_{1k\alpha},\ldots,\hat{\beta}_{pk\alpha}\right)^T = \arg\min_{\boldsymbol{\beta}_{k\alpha}} \; S(\beta_{1k\alpha},\ldots,\beta_{pk\alpha} \mid m_\alpha)$$

The neighborhood samples in terms of modulator have **similar network structure**

Model Selection

Performance of NetworkProfiler → Selection of λ_1, λ_2, and h

- Select λ_1 *and* λ_2 based on **WAICc** (Shimamura et al., 2010b)
- Select h based on **cross-validation**

$$S^{(-i)}(\beta_{1k\alpha},\ldots,\beta_{1k\alpha} \mid m_\alpha) = \sum_{i\neq\alpha} K_h(m_i - m_\alpha)\left\{x_{ik} - \sum_{j\neq k}^{p} \beta_{jk\alpha} x_{ij}\right\}^2 + \lambda_1 \sum_{j\neq k} w_{jk\alpha}\left|\beta_{jk\alpha}\right| + \frac{\lambda_2}{2}\sum_{j\neq k}\beta_{jk\alpha}^2$$

$$\hat{\boldsymbol{\beta}}_{k\alpha}^{(-i)} = \arg\min\left\{S^{(-i)}(\beta_{1k\alpha},\ldots,\beta_{1k\alpha} \mid m_\alpha)\right\}$$

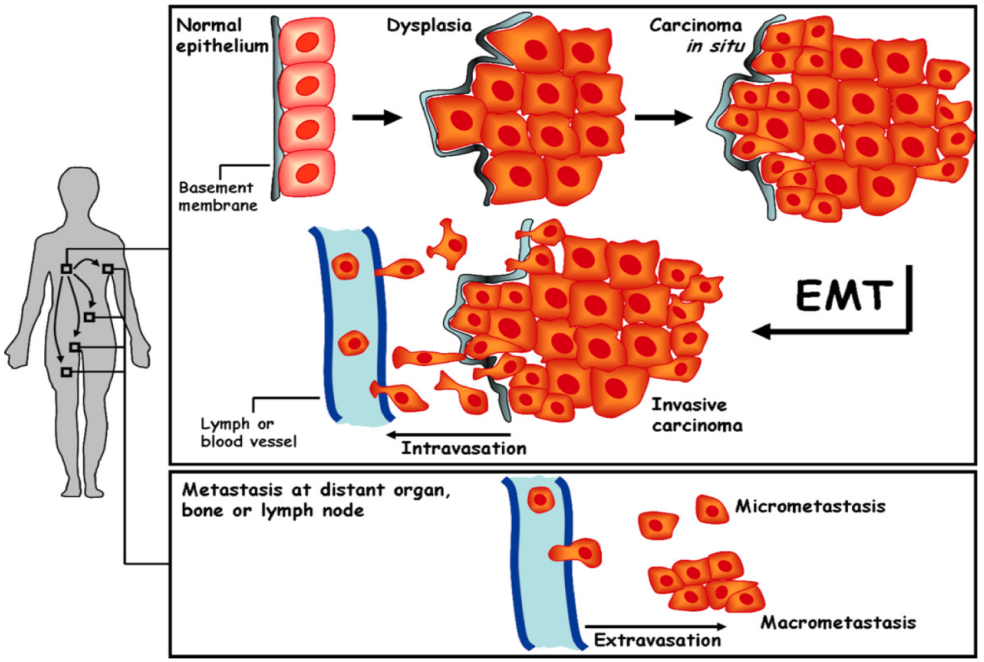
$$CV_k(h) = \sum_{\alpha=1}^{n}\left\{x_{\alpha k} - \sum_{j\neq k}\hat{\beta}_{jk\alpha}^{(-\alpha)} x_{\alpha j}\right\}^2$$

Select h minimizing $CV_k(h)$

Epithelial-Mesenchymal Transition (EMT)

- Key developmental remodeling program, where cells alternate between Epithelial-like and Mesenchymal-like phenotypes
- Relate to tumor grade and metastasis
- Contribute to increasing in drug resistance

Chua et al. (2008)

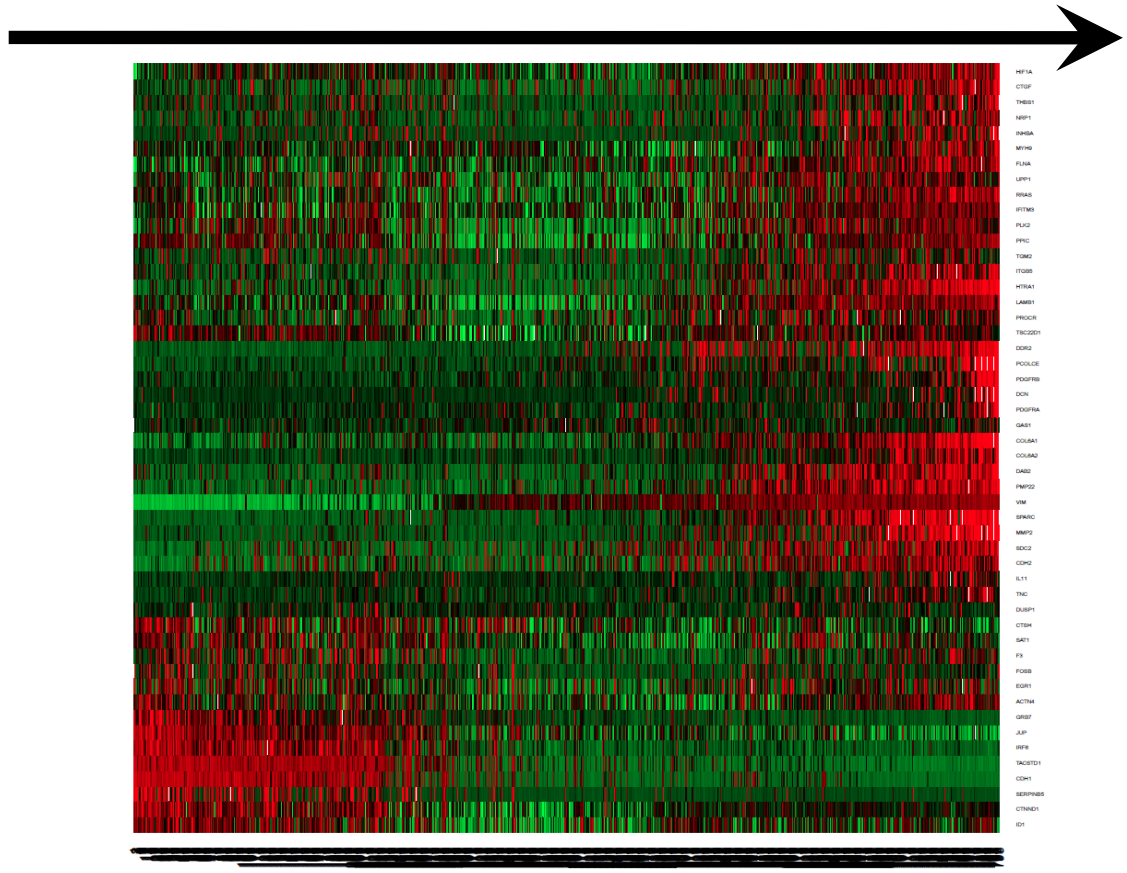


System related to EMT is a "**black box**"

Modulator for EMT

We selected coherent 50 genes from 121 EMT signature genes to define the modulator for EMT (EEM, Niida et al., 2009)

Epithelial Modulator **Messenchymal**

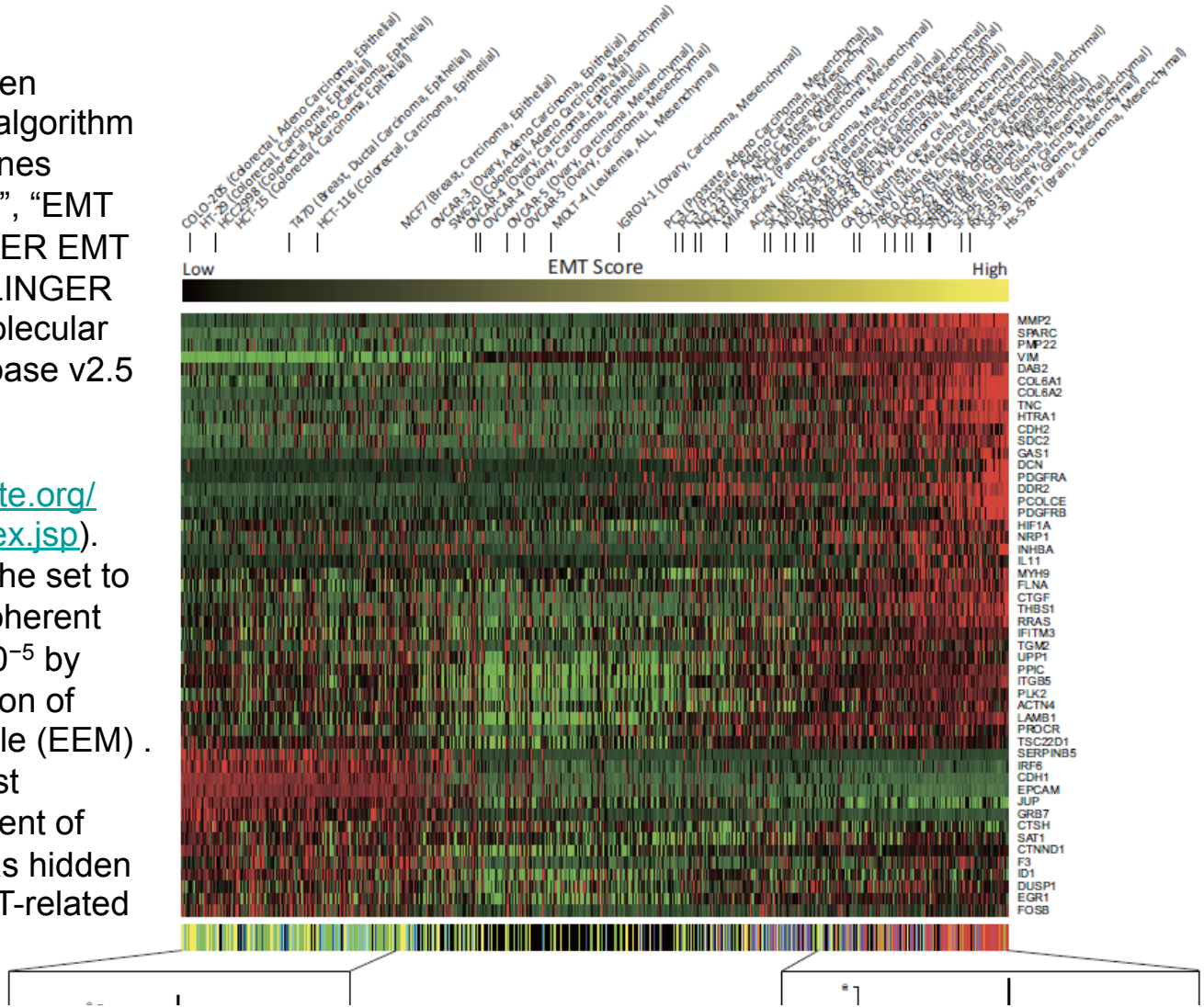


Modulator for EMT

We selected coherent 50 genes from 122 EMT signature genes to define the modulator for EMT (EEM, Niida et al., 2009)

Signature-based hidden modulator extraction algorithm

1. Selected 122 genes labeled “EMT UP”, “EMT DN”, “JECHLINGER EMT UP”, and “JECHLINGER EMT DN”from Molecular Signatures Database v2.5 ([6]; http://www.broadinstitute.org/gsea/msigdb/index.jsp).
2. Then, narrowed the set to 50 functionally coherent genes with $p < 10^{-5}$ by using the extraction of expression module (EEM) .
3. Computed the first principal component of these 50 genes as hidden values of the EMT-related modulator



Elucidating Systems Responsible for EMT

Input

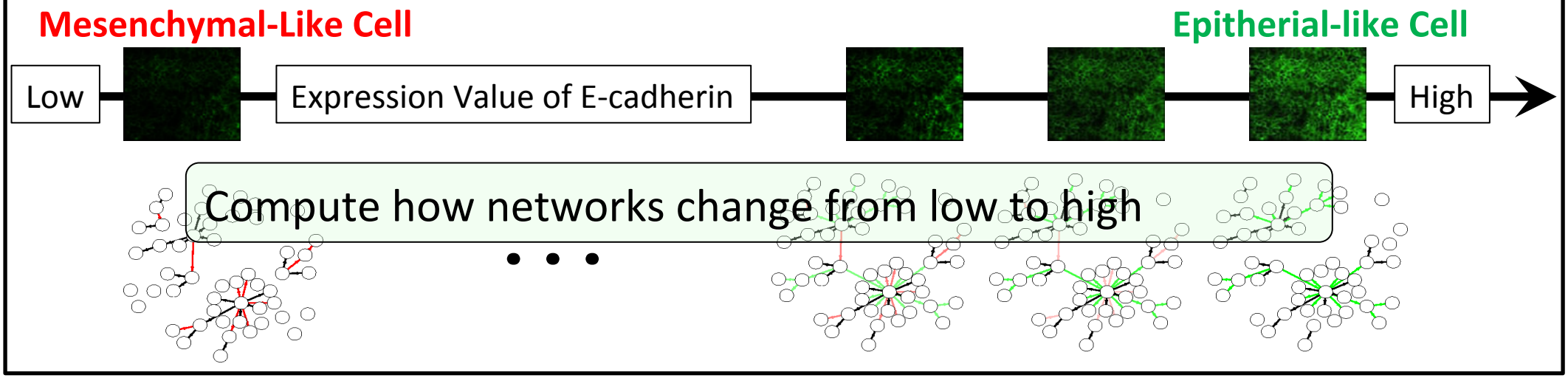
- Transcriptome data of 762 cancer cells
  (22,283 transcripts=22277 mRNA+581 miRNA)
- EMT Modulator

Computation with 1,024 cores (12.3 TFLOPS)

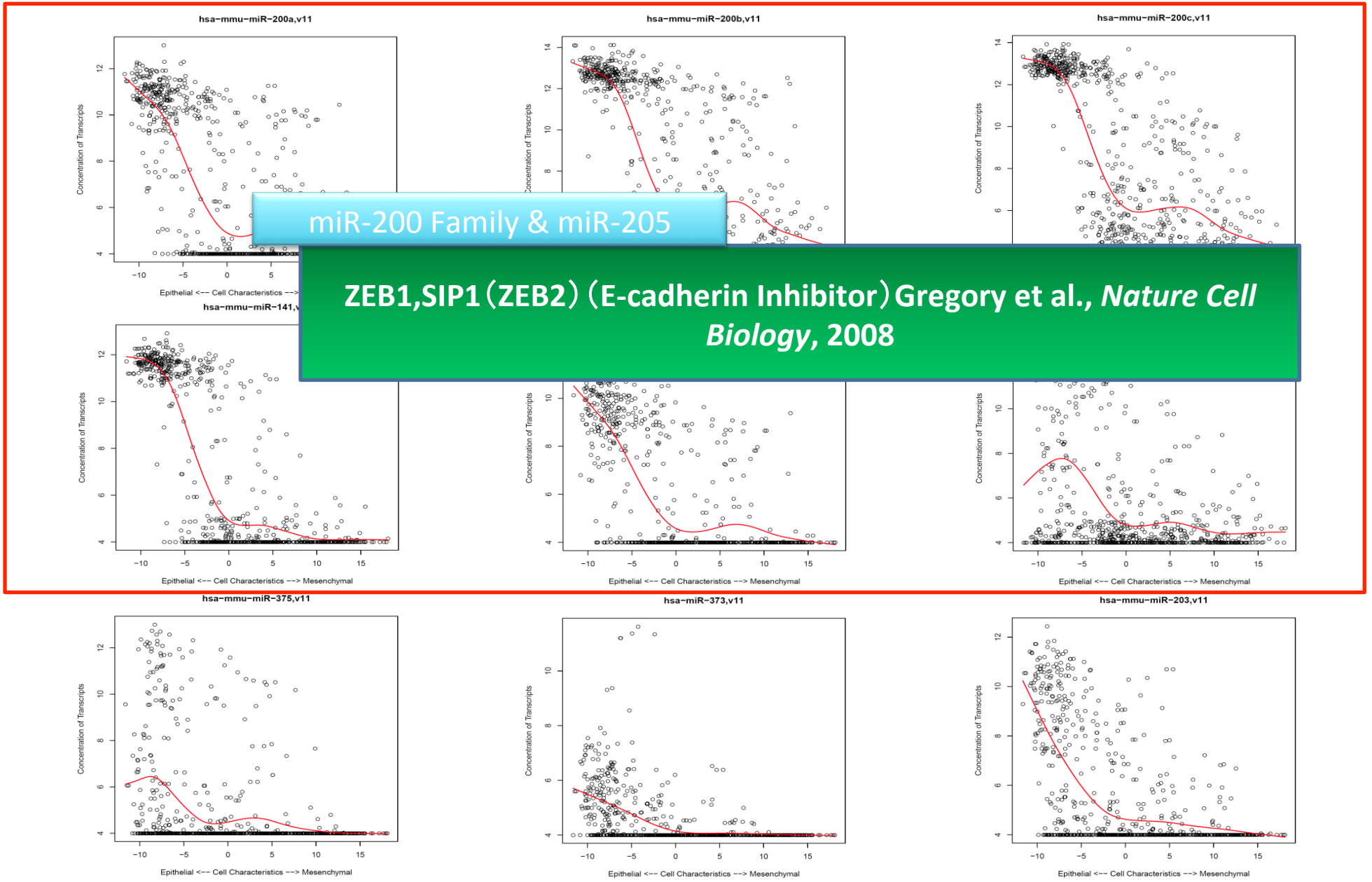
3 months

Output

762 structural varying systems with 22,283 genes related to EMT

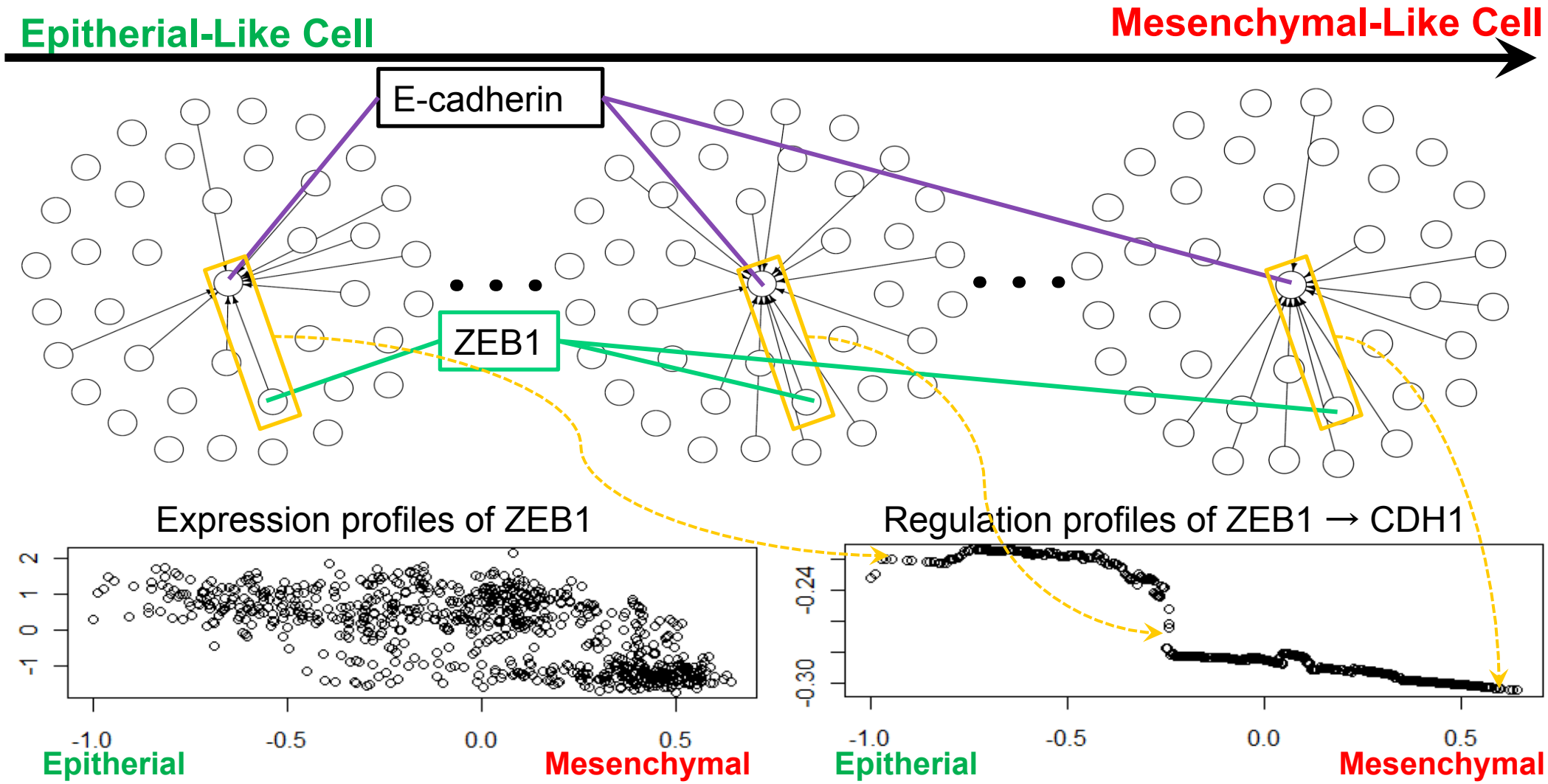


miRNAs highly expressed in Epithelial



System Changes Related to EMT

- Functional loss of E-cadherin = a hallmark of EMT
- Focus on regulators of E-cadherin



Upstream Regulatory Changes of E-cadherin

Coefficient profiles of E-cadherin regulators through EMT

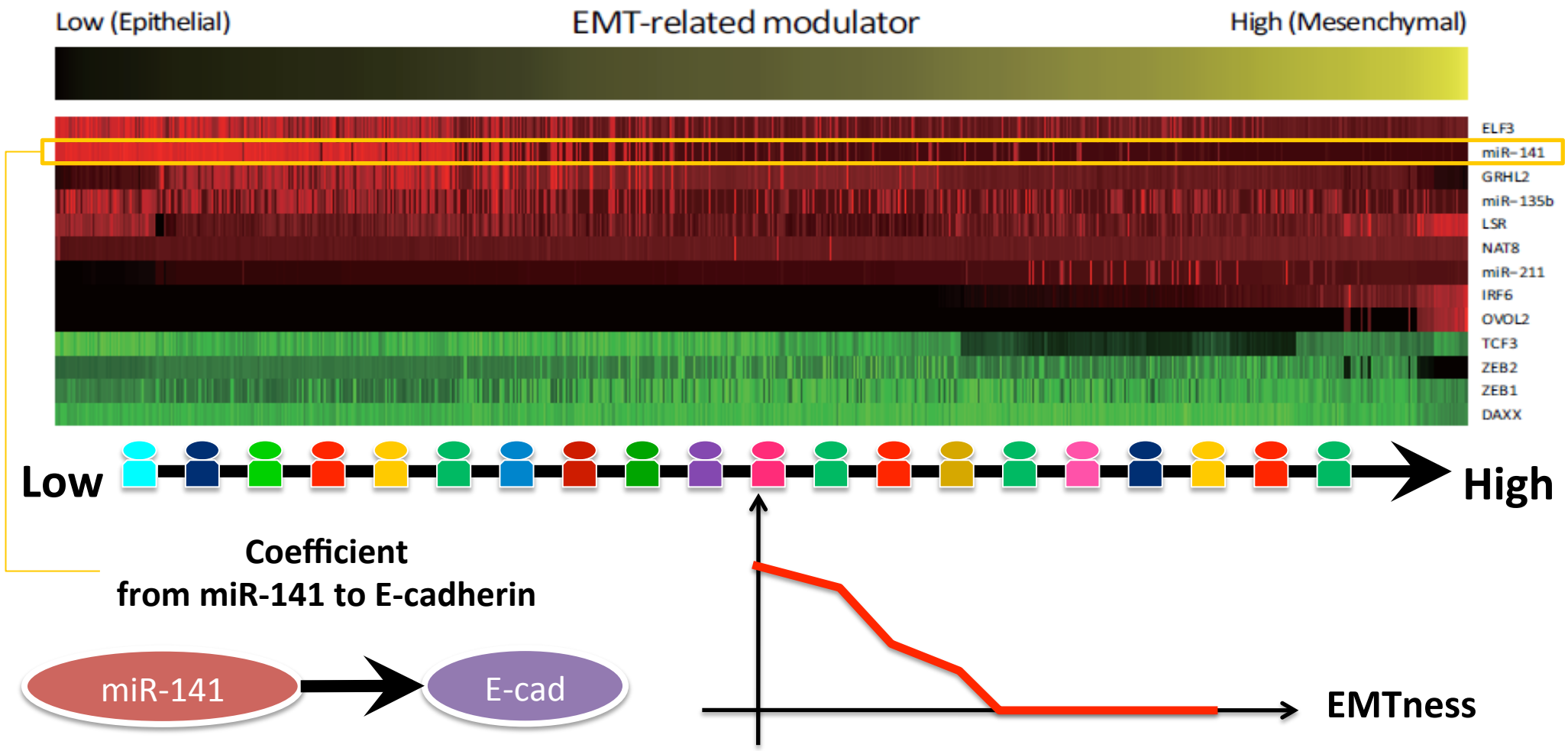
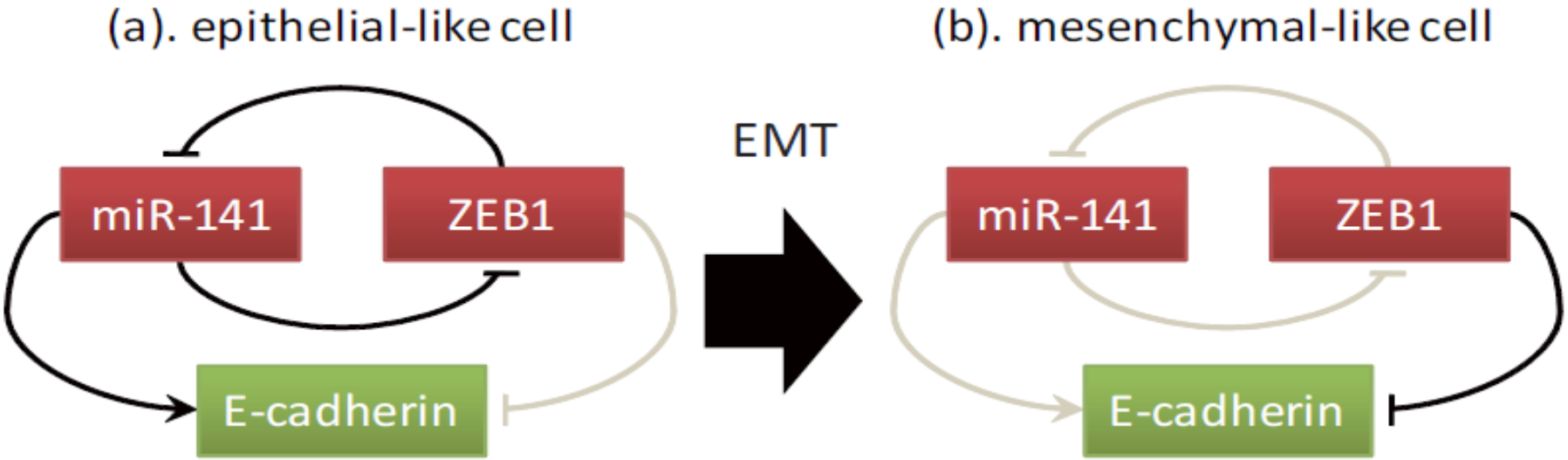


Table 1. 25 top-ranked regulators of E-cadherin for the change in the regulatory effect change among the EMT with published evidence

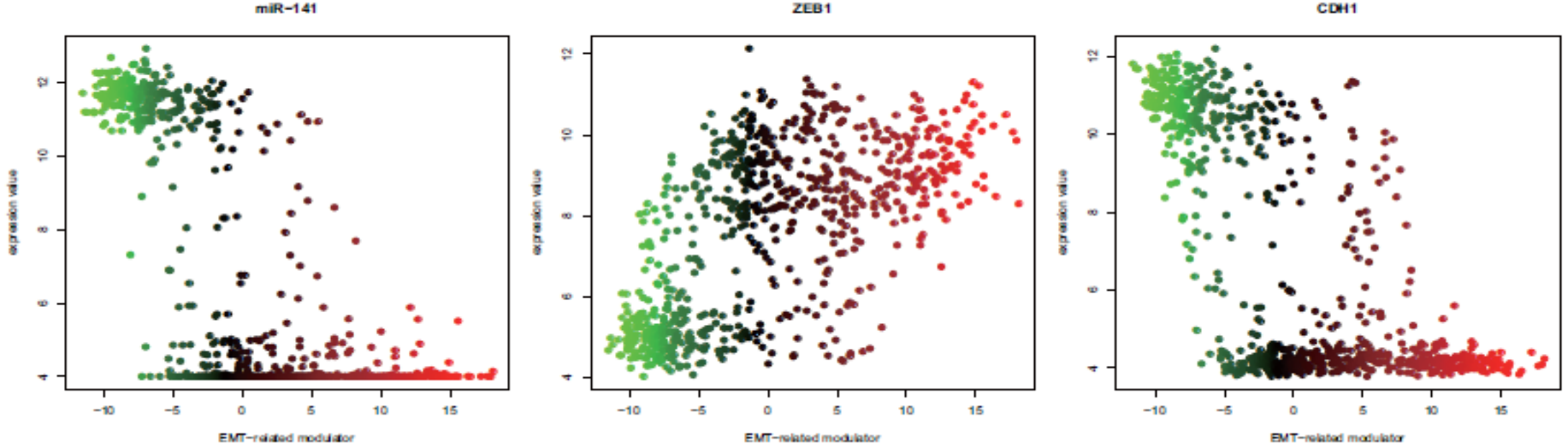
regulator	type	regulatory effect change	Evidence
IRF6	A	101.04	
miR-141	A	87.58	[8]
GRHL2	A	64.13	
ZEB1 (SIP1)	I	50.72	[9]
LSR	I	46.89	
miR-200b	A	31.55	[8]
KLF4	A	26.28	[10]
OVOL2	A	22.08	
miR-200a	A	17.70	[8]
FOXA2	A	17.26	[11]
TCF4 (E2.2)	I	14.15	[12]
ELF3	A	13.58	
ZNF217	A	13.53	
MYB	A	12.50	
KLF5	A	12.42	
miR-192	A	12.30	[13, 14]
FOXA1	A	11.69	[11]
ZNF165	A	11.39	
NKX2-1	A	11.21	
HNF1B	A	11.08	
TFE3	A	11.01	
ZEB2 (δEF)	I	10.66	[15]
TRIM29	I	9.87	
SNAI2	I	9.74	[16]

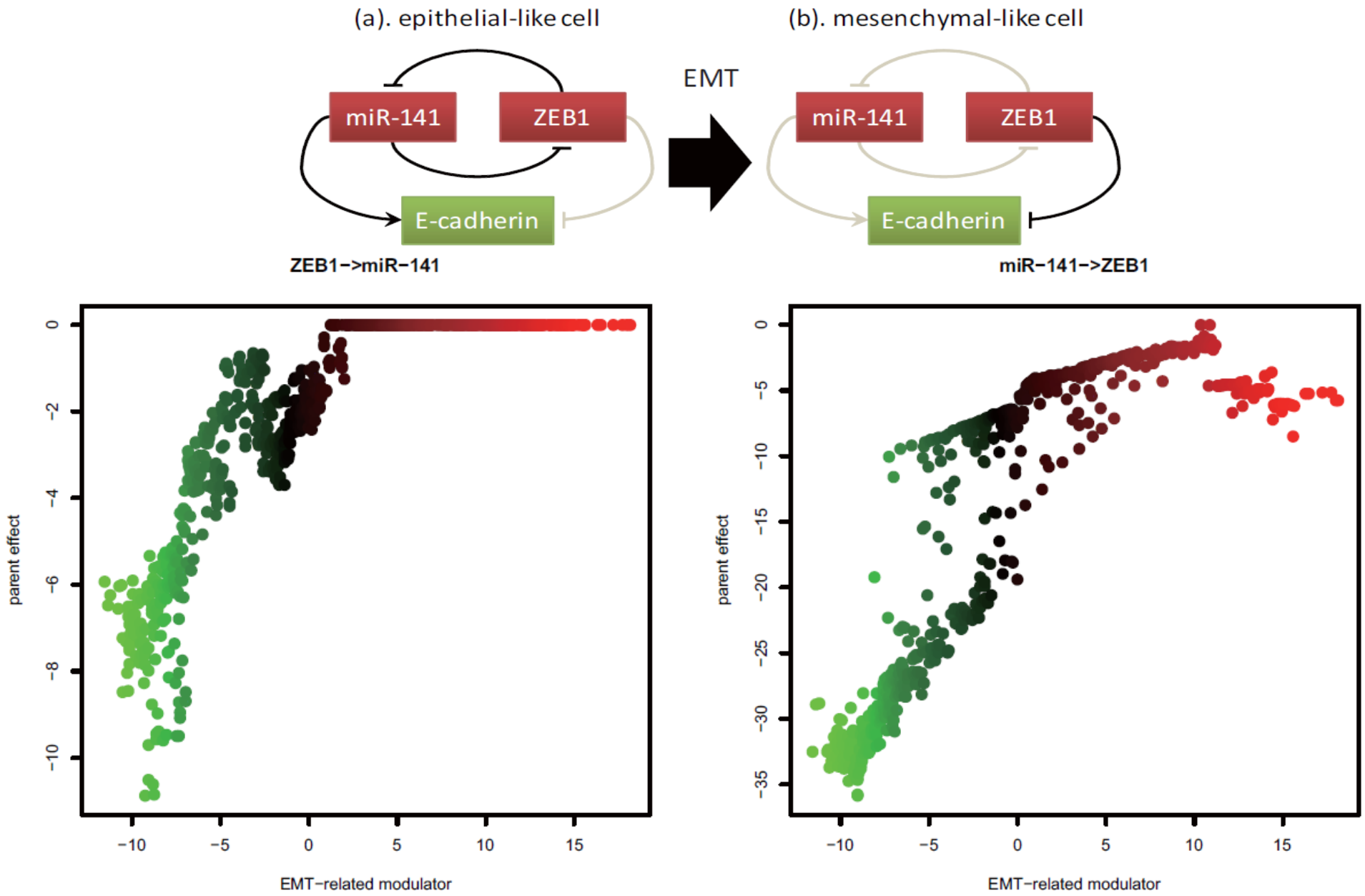
A: Activator
I: Inhibitor

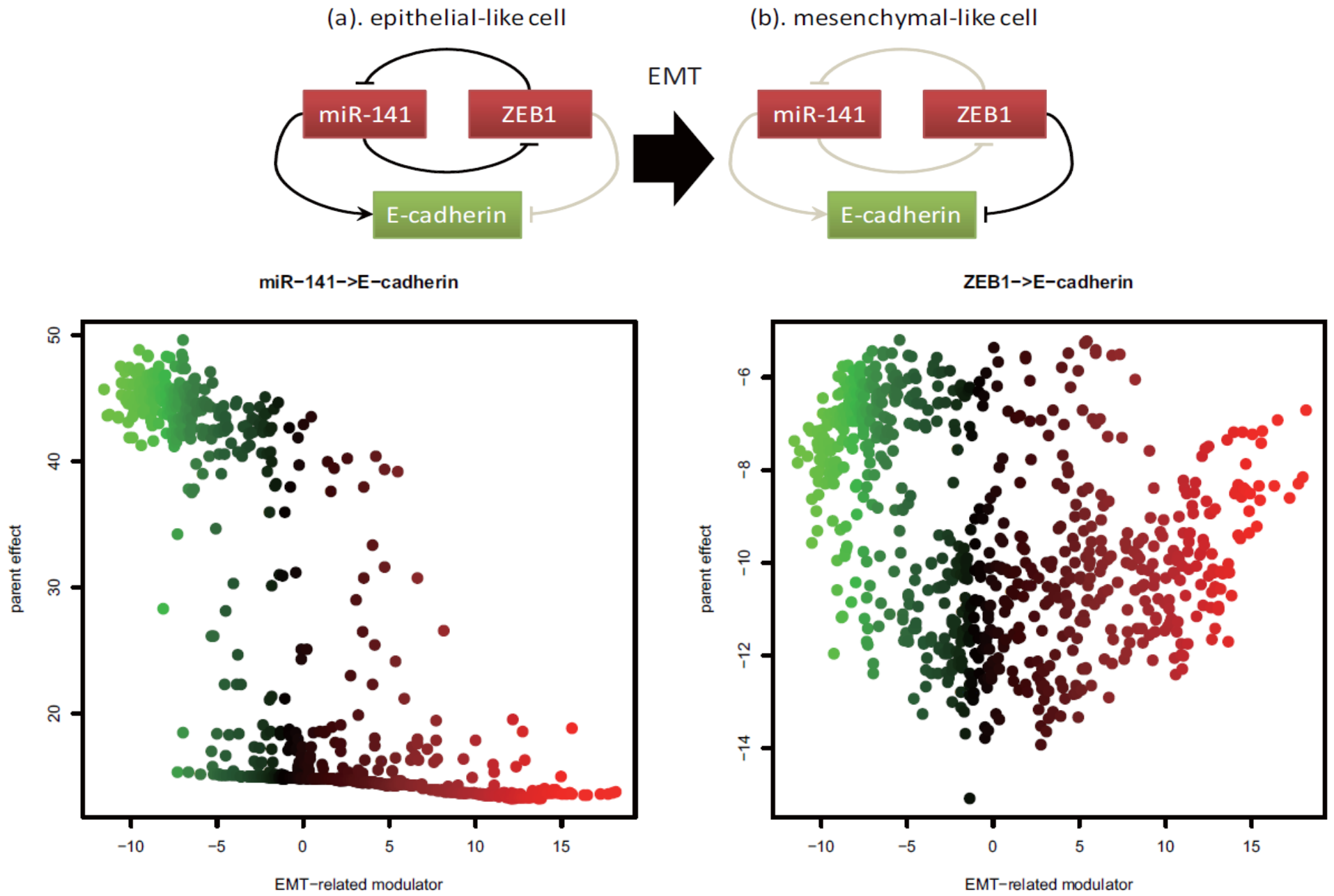
Upstream Regulatory Changes of E-cadherin



(c). The green and red colors indicate epithelial- and mesenchymal-like cells, respectively.




(a). epithelial-like cell
(b). mesenchymal-like cell
EMT
miR-141
ZEB1
E-cadherin
miR-141
ZEB1
E-cadherin
ZEB1−>miR−141
miR−141−>ZEB1
parent effect
EMT−related modulator
parent effect
EMT−related modulator


(a). epithelial-like cell
(b). mesenchymal-like cell
EMT
miR-141
ZEB1
E-cadherin
miR-141
ZEB1
E-cadherin
miR−141−>E−cadherin
ZEB1−>E−cadherin
parent effect
EMT−related modulator
parent effect
EMT−related modulator

miR-141, ZEB1, and E-cadherin

- NetworkProfiler revealed regulatory changes in *miR-141*, *ZEB1*, and E-cadherin.
- Specifically, it suggested that decreased expression of *miR-141* in mesenchymal cells disrupts the negative feedback loop between*miR-141* and *ZEB1*, which would allow *ZEB1* to decrease the expression of E-cadherin during the EMT.

- We predicted 45 EMT-dependent putative master regulators that control sets of genes involved in cell adhesion, migration, invasion and metastasis, namely,
- 17 of which are in the downstream of TGFB1, a master switch of the EMT, in our prediction.

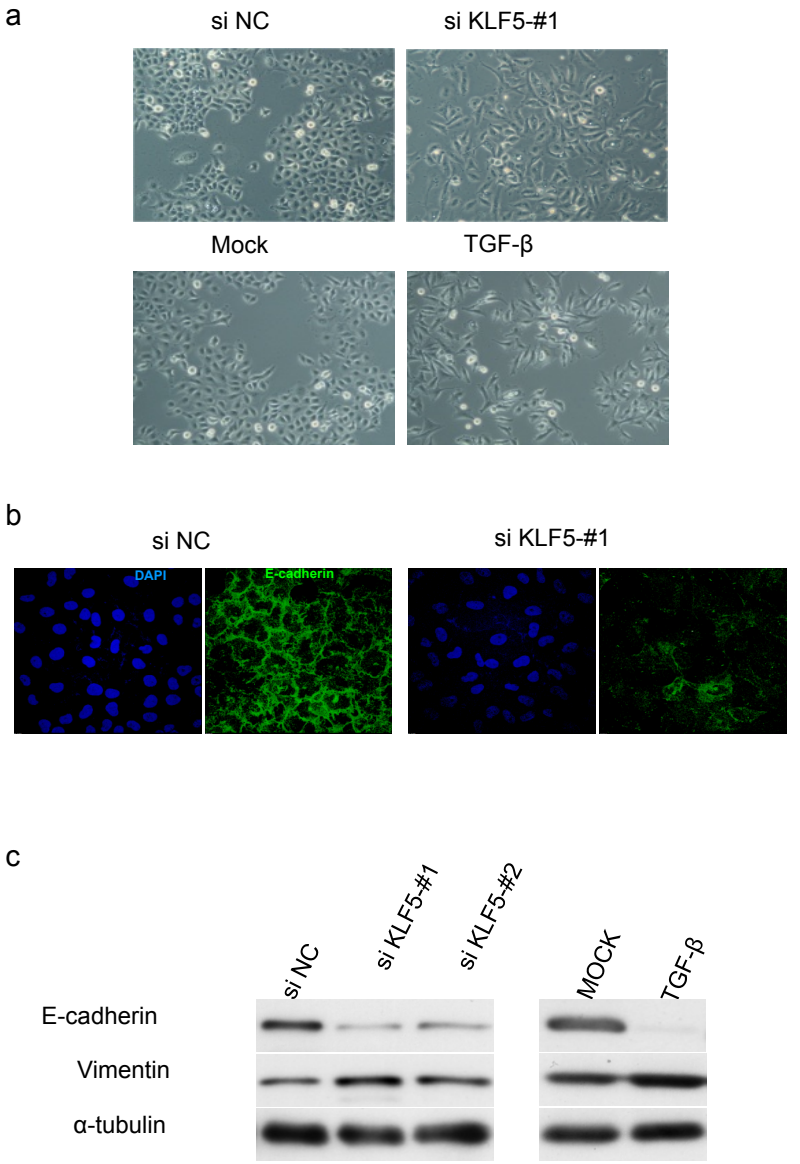
Table 1. 25 top-ranked regulators of E-cadherin for the change in the regulatory effect change among the EMT with published evidence

regulator	type	regulatory effect change	Evidence
IRF6	A	101.04	
miR-141	A	87.58	[8]
GRHL2	A	64.13	
ZEB1 (SIP1)	I	50.72	[9]
LSR	I	46.89	
miR-200b	A	31.55	[8]
KLF4	A	26.28	[10]
OVOL2	A	22.08	
miR-200a	A	17.70	[8]
FOXA2	A	17.26	[11]
TCF4 (E2.2)	I	14.15	[12]
ELF3	A	13.58	
ZNF217	A	13.53	
MYB	A	12.50	
KLF5	A	12.42	
miR-192	A	12.30	[13, 14]
FOXA1	A	11.69	[11]
ZNF165	A	11.39	
NKX2-1	A	11.21	
HNF1B	A	11.08	
TFE3	A	11.01	
ZEB2 (δEF)	I	10.66	[15]
TRIM29	I	9.87	
SNAI2	I	9.74	[16]

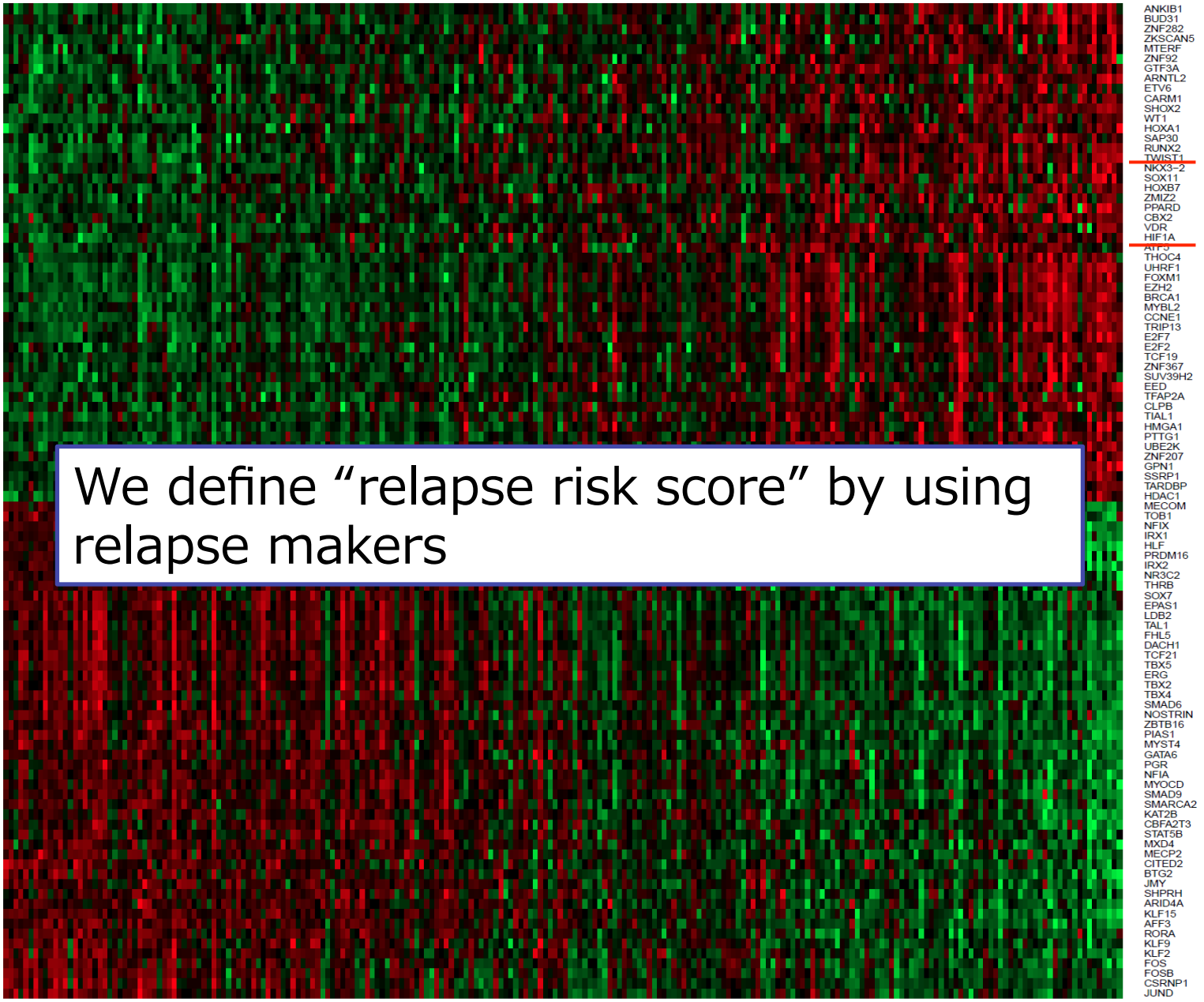
A: Activator
I: Inhibitor

A novel regulator KLF5 of EMT

- Krueppel-like factor 5 (KLF5) from a list of the remaining candidate regulators and conducted *in vitro* validation experiments.
- As a result, we found that knockdown of KLF5 by siRNA significantly decreased E-cadherin expression and induced morphological changes characteristic of EMT.

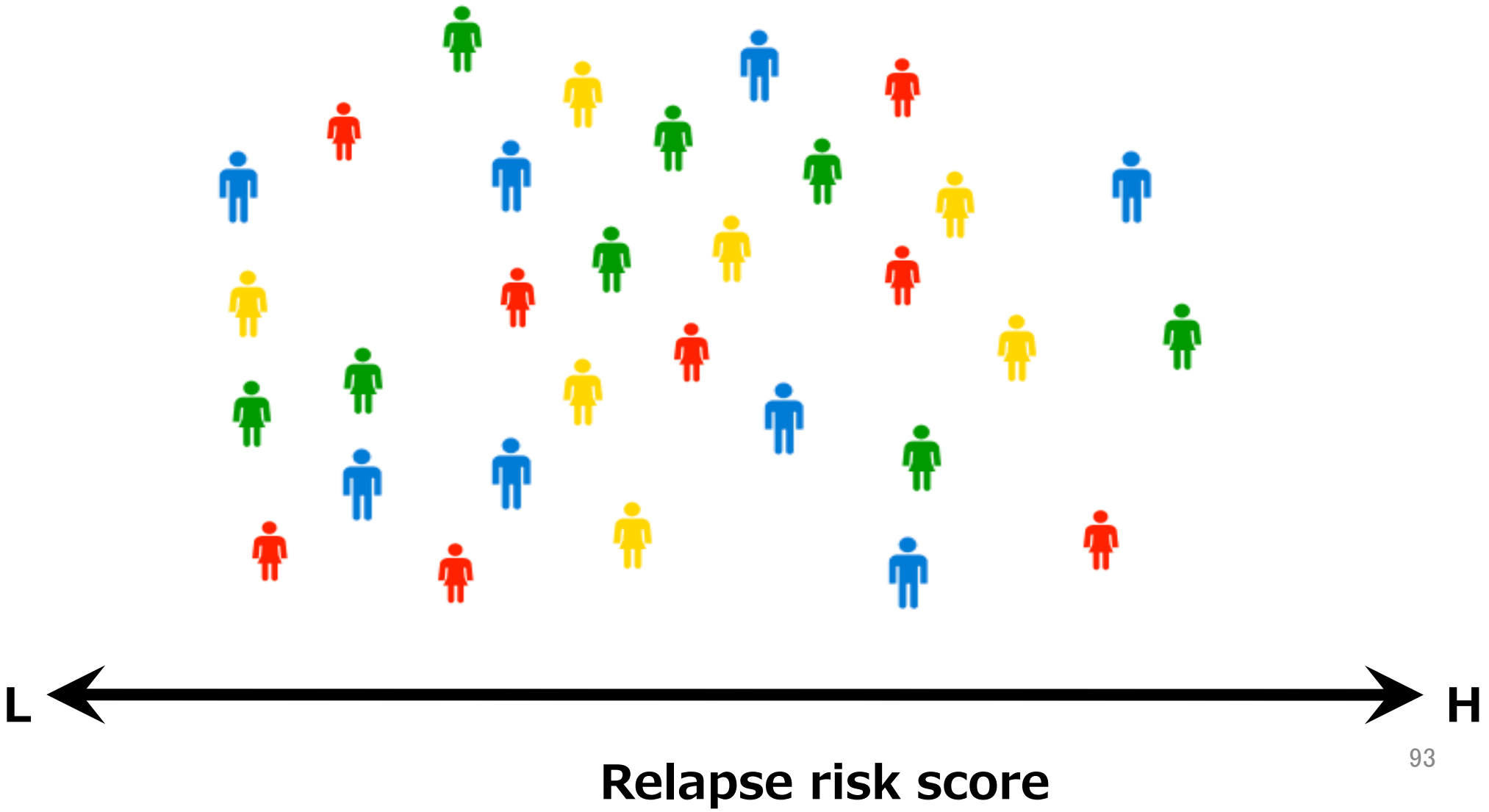


low ← **Relapse risk** → high



We define “relapse risk score” by using relapse makers

Patient Samples (Microarray Data)

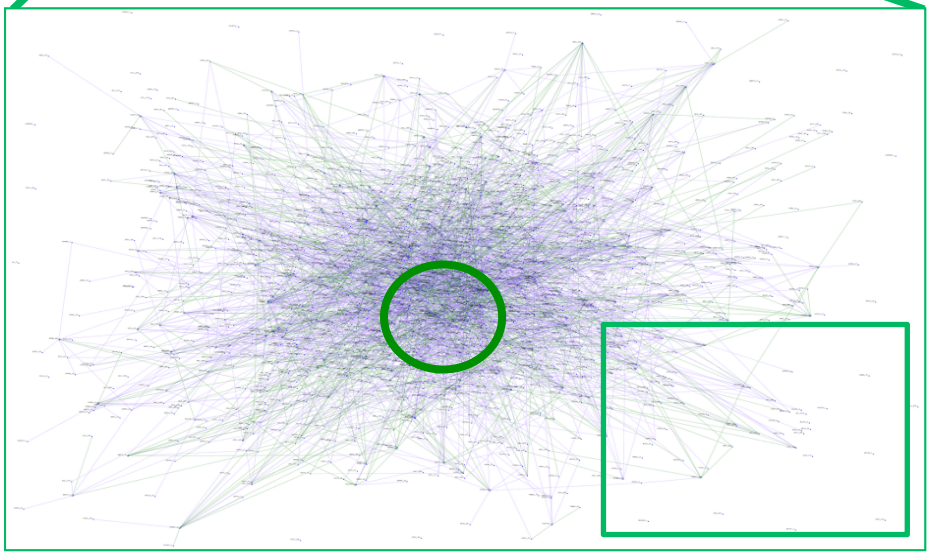


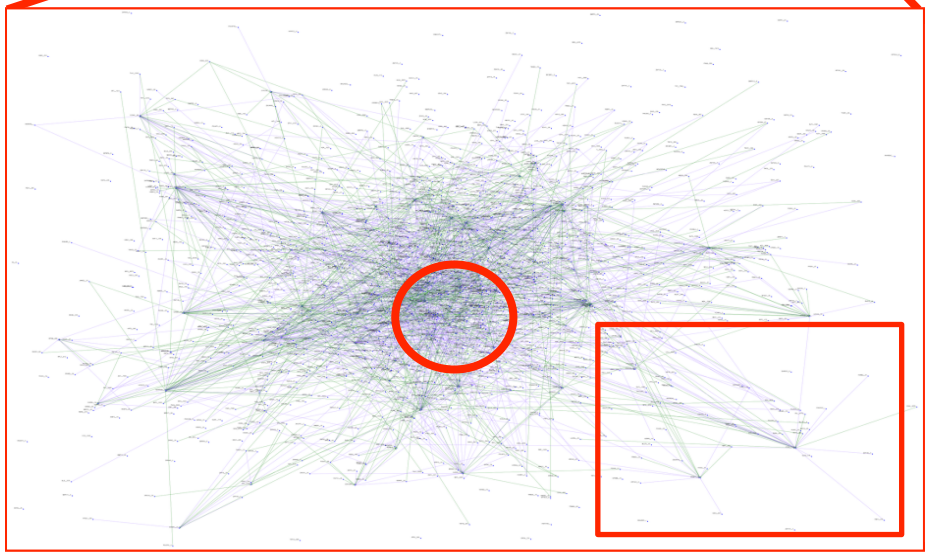
Network Profiling

Replase risk

H ⟷ H

Each patient network is computed

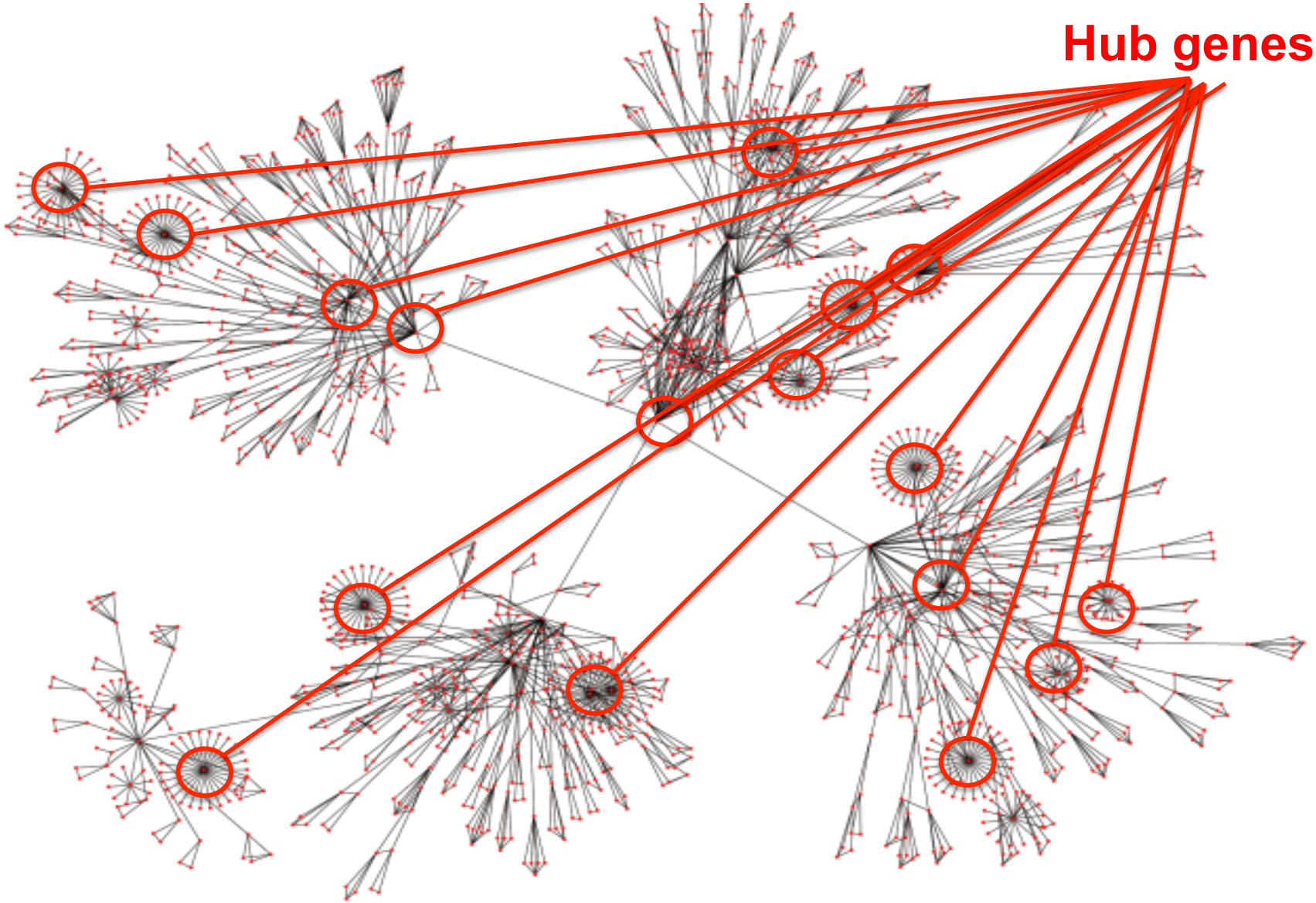




Network of the lowest risk patient

Network of the highest risk patient

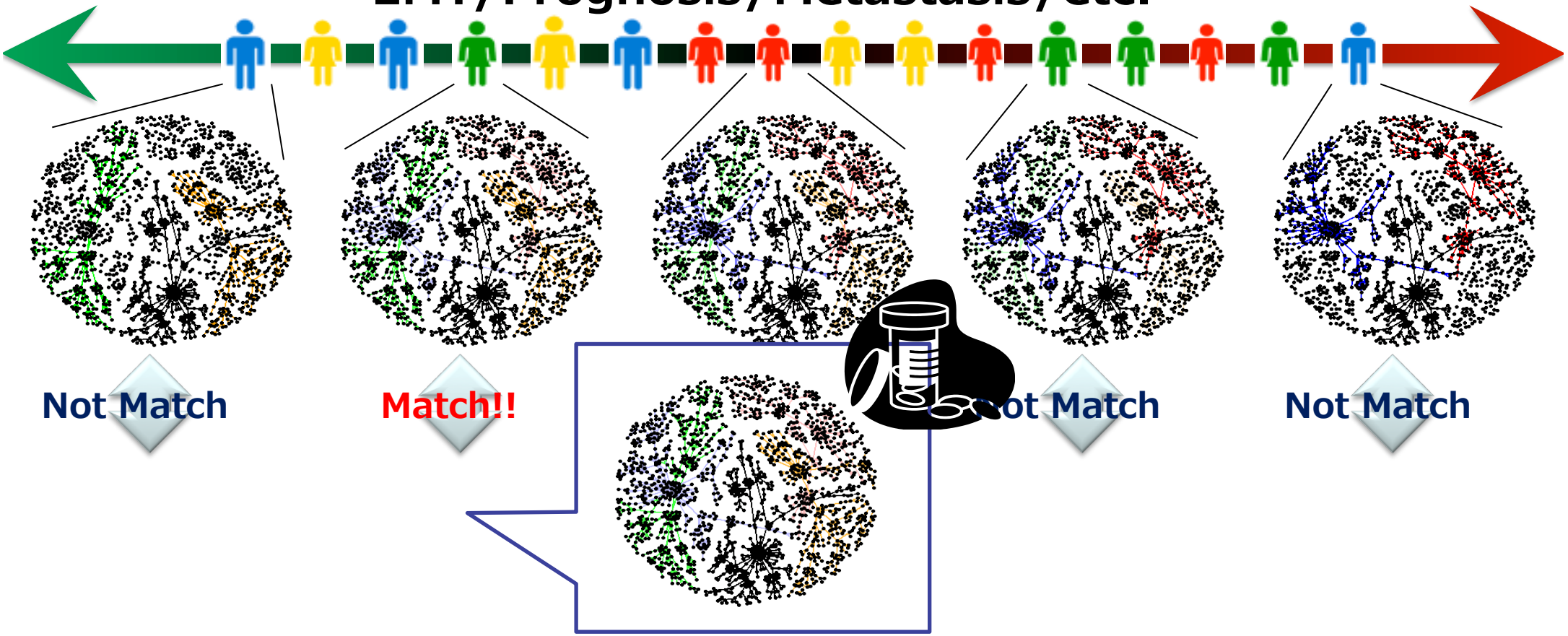
Differences of Hubness Suggest Key Genes



System-Oriented Personalized Medicine

Networks

EMT/Prognosis/Metastasis/etc.



Thorough Analysis Requires 500 Times

25 yeas with 1000 cores
10 days with 1,000,000 cores

Prologue

Mapmakers in Systems Biology

伊能忠敬 Tadataka Inoh
A man who walked 40,000,000 steps

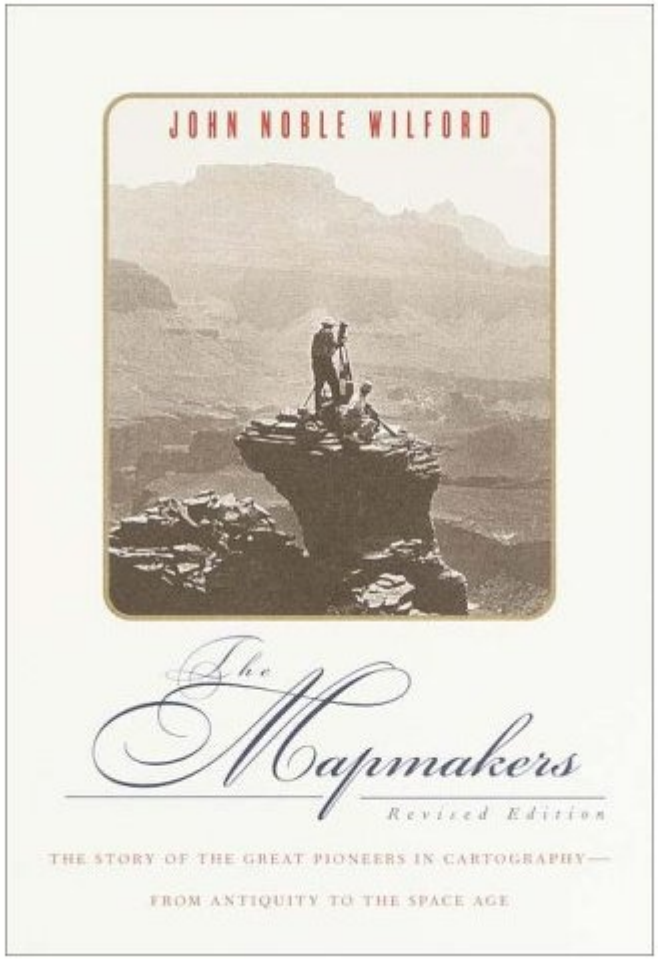
- 伊能忠敬は，江戸幕府の事業として，1800年から1816年にかけて全国を歩いて測量をし，1821年に「大日本沿海輿地全図」が幕府に納められたといいます．伊能忠敬はその完成を見ずに1818年に死去ましたが，その後，仕上げの編纂作業が行われ，全部で21年の歳月をかけてこの地図は完成しています．

2001年のNHKの正月時代劇「四千万歩の男・伊能忠敬」（原作：井上ひさし「四千万歩の男」（講談社））として放映

The Mapmakers

The story of the great pioneers in cartography – from antiquity to the space age

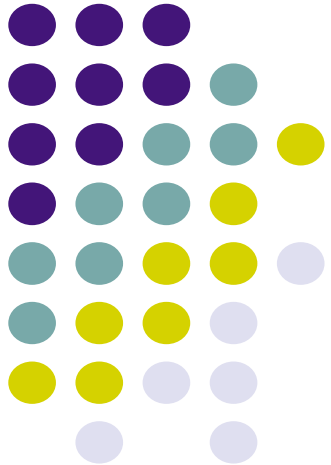
- John Noble Wilford



2. Mining Gene Networks from Gene Expression Profiles

Mapmaking of Molecular Networks

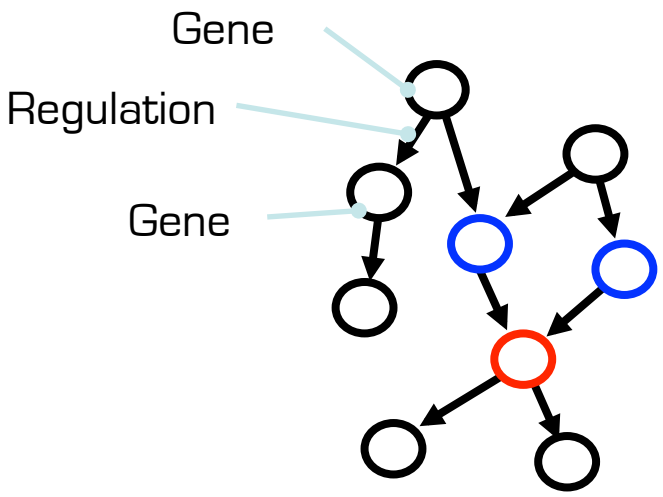
- Bayesian Network Gene Networks
  - Gene knock-down/knock-out
  - Various shocks
  - Time-course data



Yoshinori Tamada;Seiya Imoto;Masao Nagasaki;Satoru Miyano

2. Bayesian Networks + Nonparametric Regression

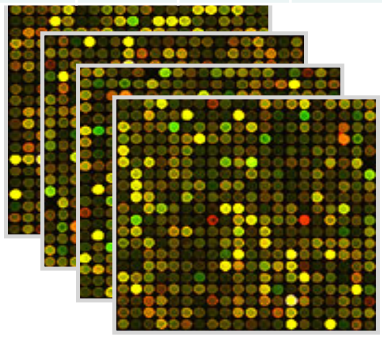
- Gene network: model for transcriptome level gene–gene regulation using directed graphs.
- We want to estimate gene networks from high-throughput biological data e.g. gene expression data.



Gene Network

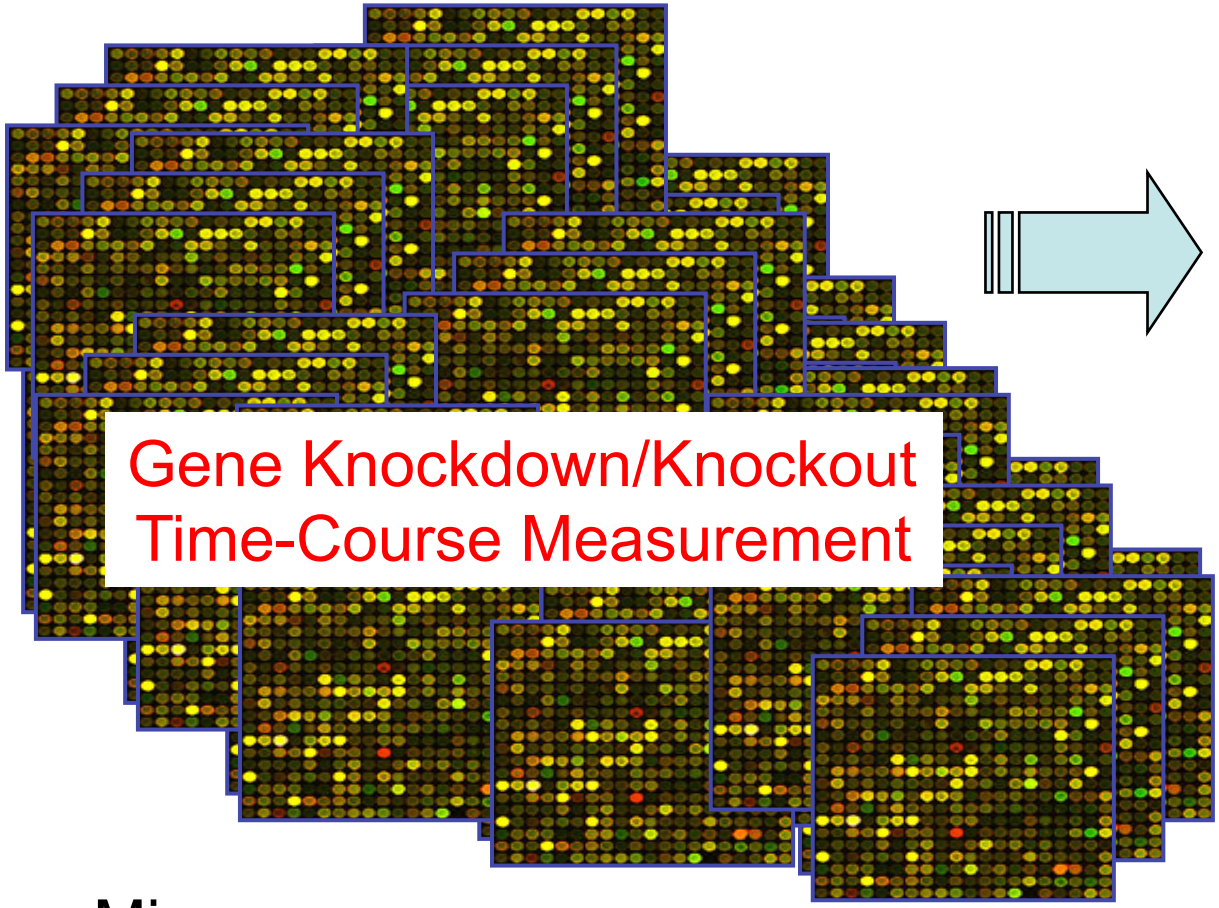
Estimation

Gene	KD1	KD2	KD3	…
Gene 1	1.45	-1.54	1.23	…
Gene 2	3.21	-2.1	1.44	…
…	…	…	…	…

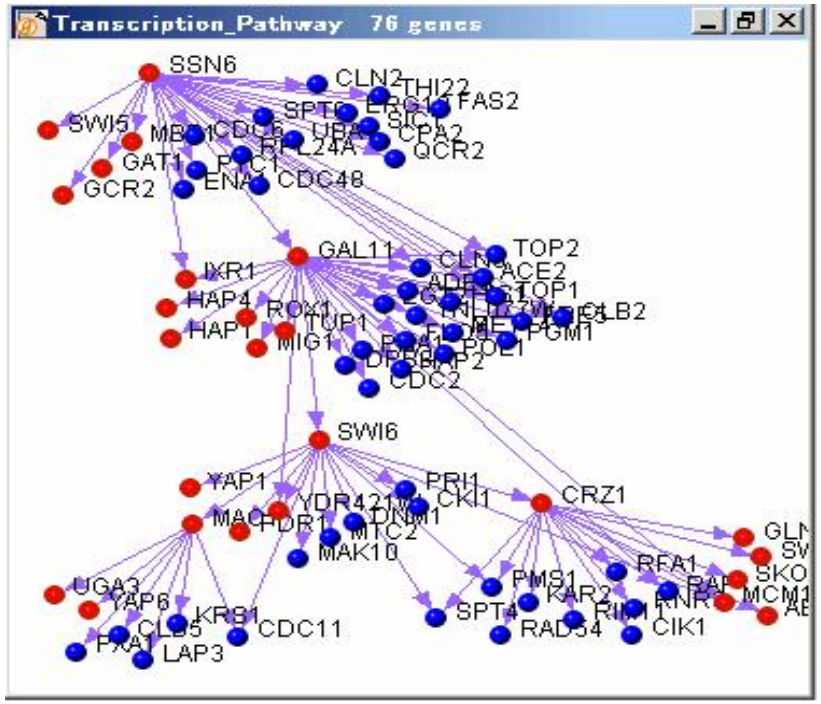


Gene Expression Data

What we wanted to do



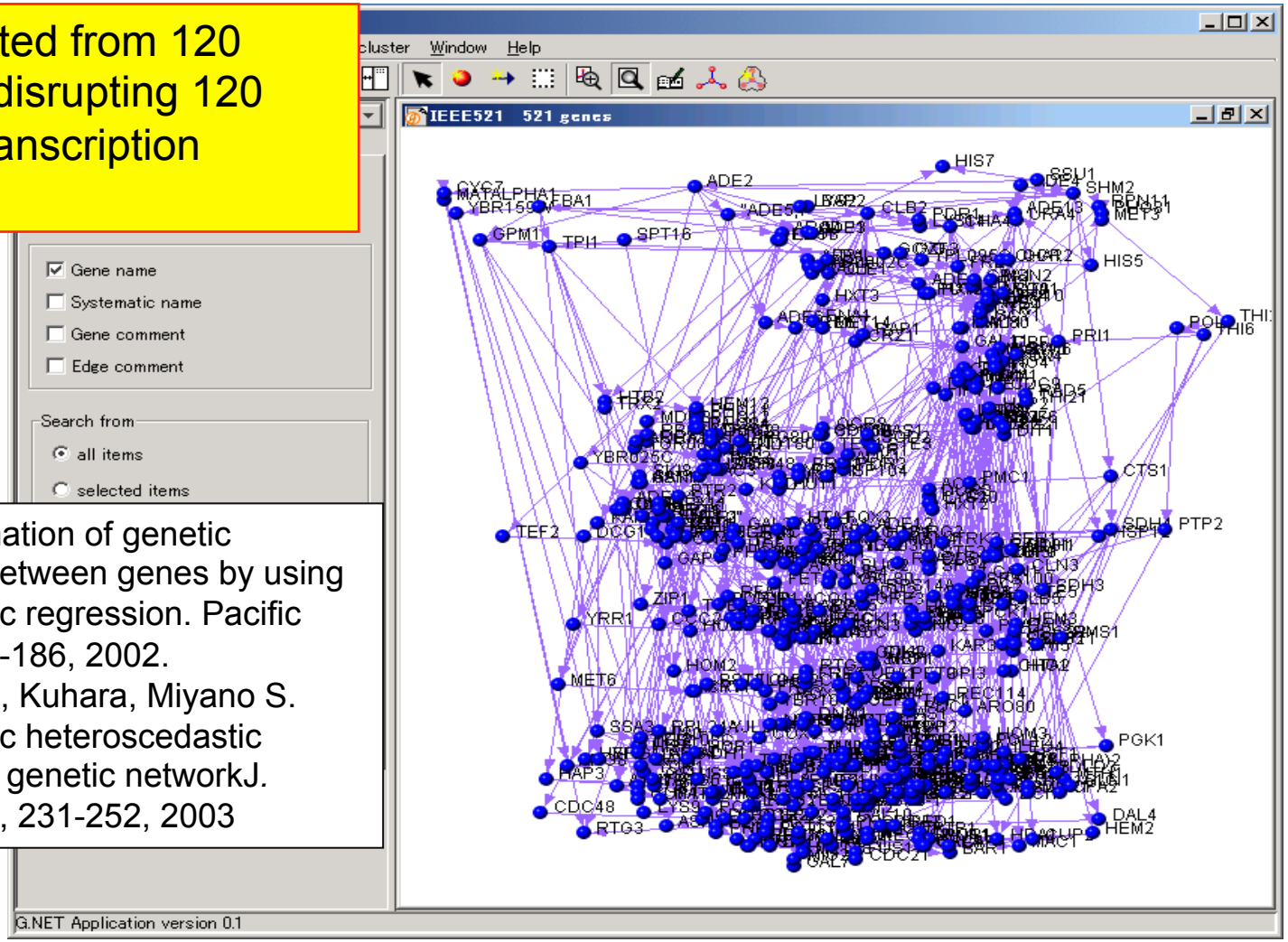
Microarray gene expression data



Gene network

Bayesian Network and Nonparametric Regression

Network of 521 genes constructed from 120 yeast microarrays obtained by disrupting 120 genes, where 78 of them are transcription factors.



1. Imoto, S., Goto, T., Miyano, S. Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. Pacific Symposium on Biocomputing. 7:175-186, 2002.
2. Imoto, Kim, Goto, Aburatani, Tashiro, Kuhara, Miyano S. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic networkJ. *Bioinformatics and Comp. Biol.*, 1(2), 231-252, 2003

Nonlinear Bayesian network model

$$f(x_{i1},\ldots,x_{ip};\boldsymbol{\theta}_G)=\prod_{j=1}^{p} f_j(x_{ij}\mid \mathbf{p}_{ij};\boldsymbol{\theta}_j),$$

$$f_j(x_{ij}\mid \mathbf{p}_{ij};\boldsymbol{\theta}_j)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x_{ij}-\mu_{ij})^2}{2\sigma_j^2}\right\}$$

$$\mu_{ij}=m_1(p_{i1}^{(j)})+\cdots+m_{q_j}(p_{iq_j}^{(j)})$$

$$=\sum_{k=1}^{q_j}\sum_{m=1}^{M_{jk}}\gamma_{mk}b_{mk}^{(j)}(p_{ik}^{(j)})$$

Criterion for Selecting Good Networks

BNRC Score

Bayesian Network and Nonparametric Regression Criterion

$$\mathrm{BNRC}(G) = -2\log\pi_G \int \prod_{i=1}^{n} f(\mathbf{x}_i;\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G \mid \boldsymbol{\lambda})d\boldsymbol{\theta}_G$$

$$= -2\log\pi_G - r\log(2\pi n^{-1})$$

$$+ \log\left|J_\lambda(\hat{\boldsymbol{\theta}}_G)\right| - 2nl_\lambda(\hat{\boldsymbol{\theta}}_G \mid \mathbf{X}_n)$$

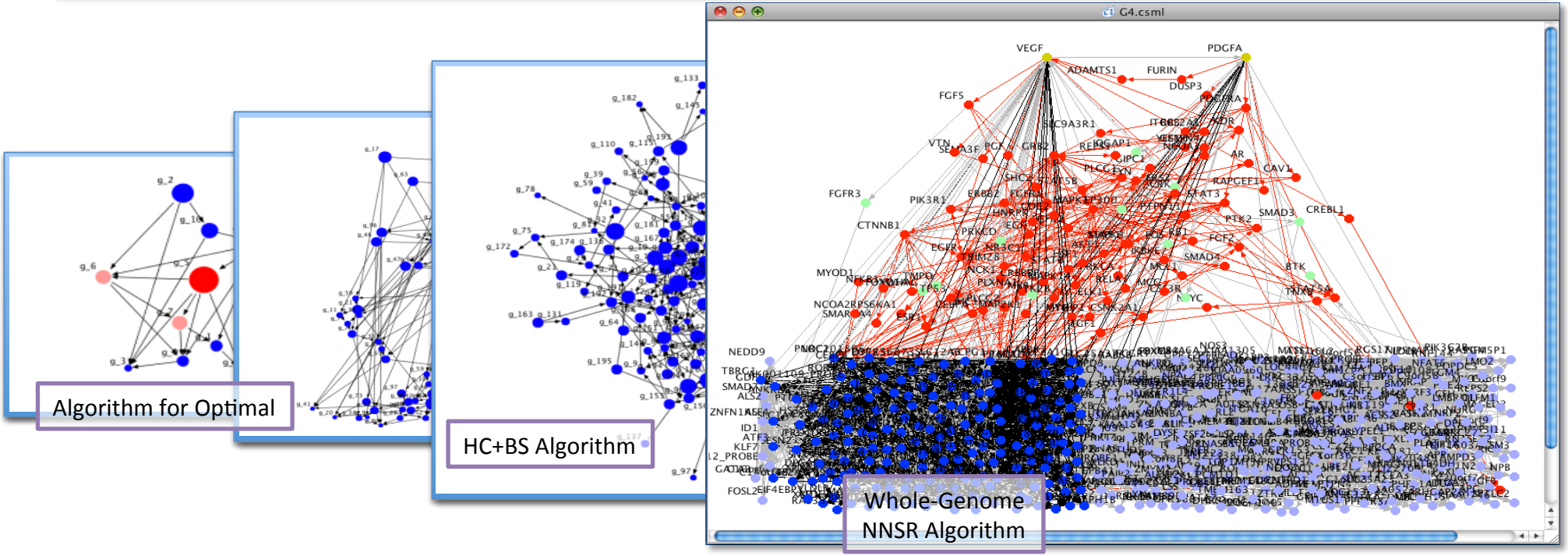
We choose the graph that minimizes the value of the BNRC score.

A Series of Programs on Supercomputer for mining gene networks of size from 30 to 20,000 (genome-wide)

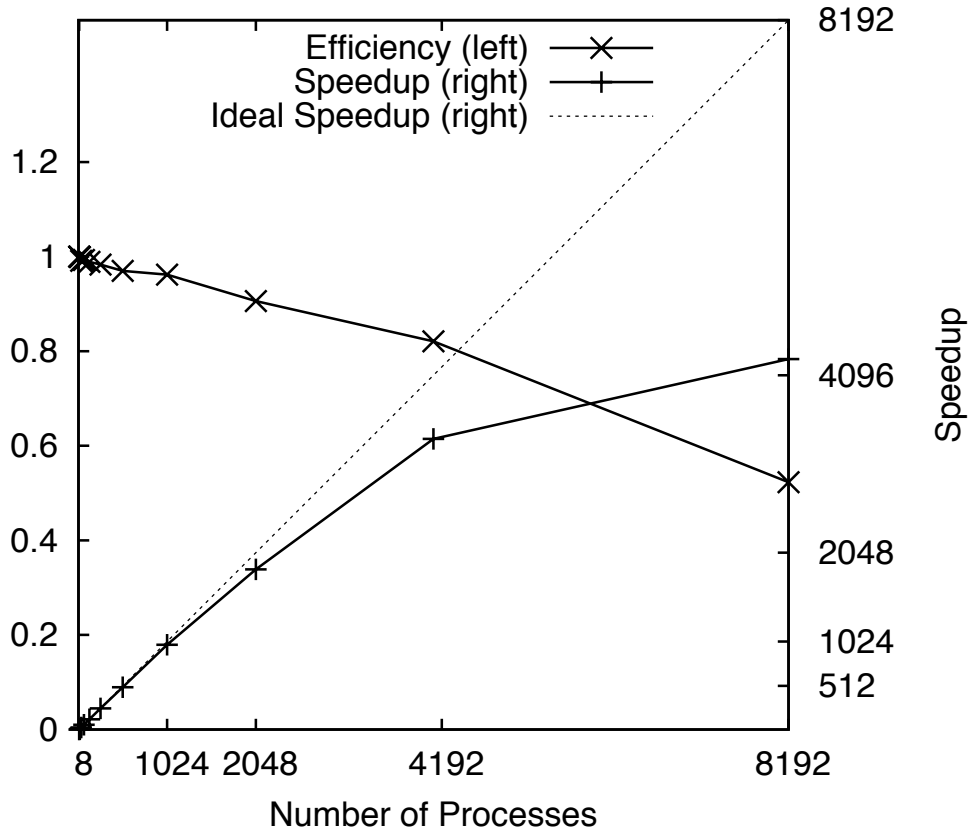
Optimal to Locally Optimal

WR: Optimal Bayesian Networks of 32 Nodes

April 2010

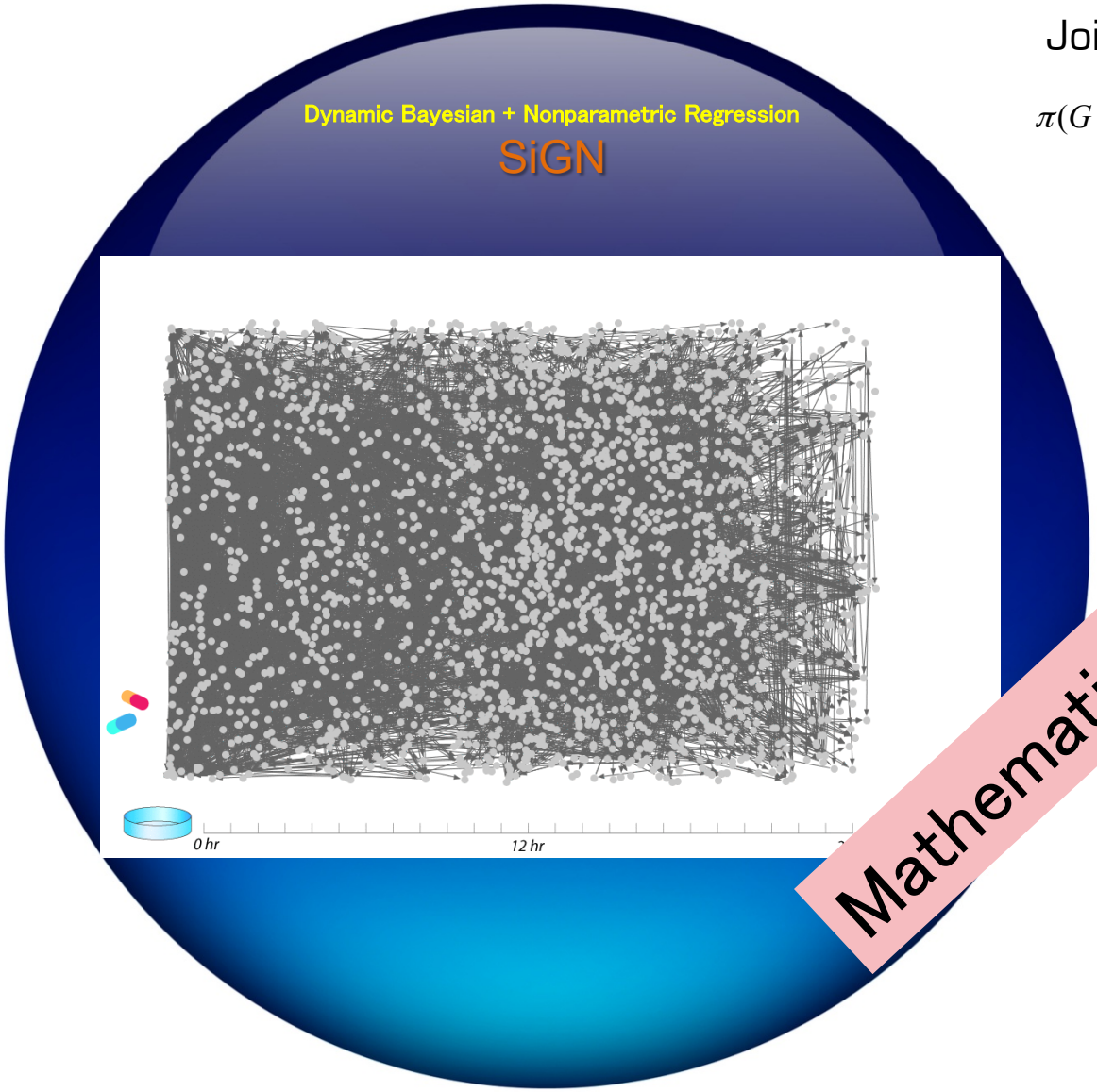


Parallel Computing with 8192 Cores



Genome-Wide Bayesian Network Computation

SiGN
Extracting Dynamic Changes in Cancer Gene Network by Dynamic Bayesian Network



Joint Probability by a DAG

$$\pi(G \mid \mathbf{X}) \propto \pi(G) \int \prod_{i=1}^{n} f(x_{i1},..,x_{ip} \mid \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G \mid$$

= BNRC(G): Net… …re

G : …work (DAG)

… …ssion data

Nonparam… …gression by B-spline

$$\ldots + m_{jq_j}(p_{iq_j}^{(j)}) + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0,\sigma_j^2)$$

$$\ldots n_{jk}(p_{ik}^{(j)}) = \sum_{l=1}^{M_{jk}} \gamma_{lk} b_{lk}^{(j)}(p_{ik}^{(j)})$$

…work Estimation

$$G = \arg\min_G \mathrm{BNRC}(G) = \arg\min_G \sum_{j=1}^{n} \mathrm{BNRC}_j(G)$$

$$\mathrm{BNRC}_j(G) = 2(q_j+1) - \left(\sum_{k=1}^{q_j} M_{jk} + 1\right)\log\left(\frac{2\pi}{n}\right) + n\log(2\pi\hat{\sigma}_j^2) + n$$

$$+ \sum_{k=1}^{q_j}\left\{\log|\Lambda_{jk}| - M_{jk}\log(n\hat{\sigma}_j^2)\right\} - \log(2\hat{\sigma}_j^4)$$

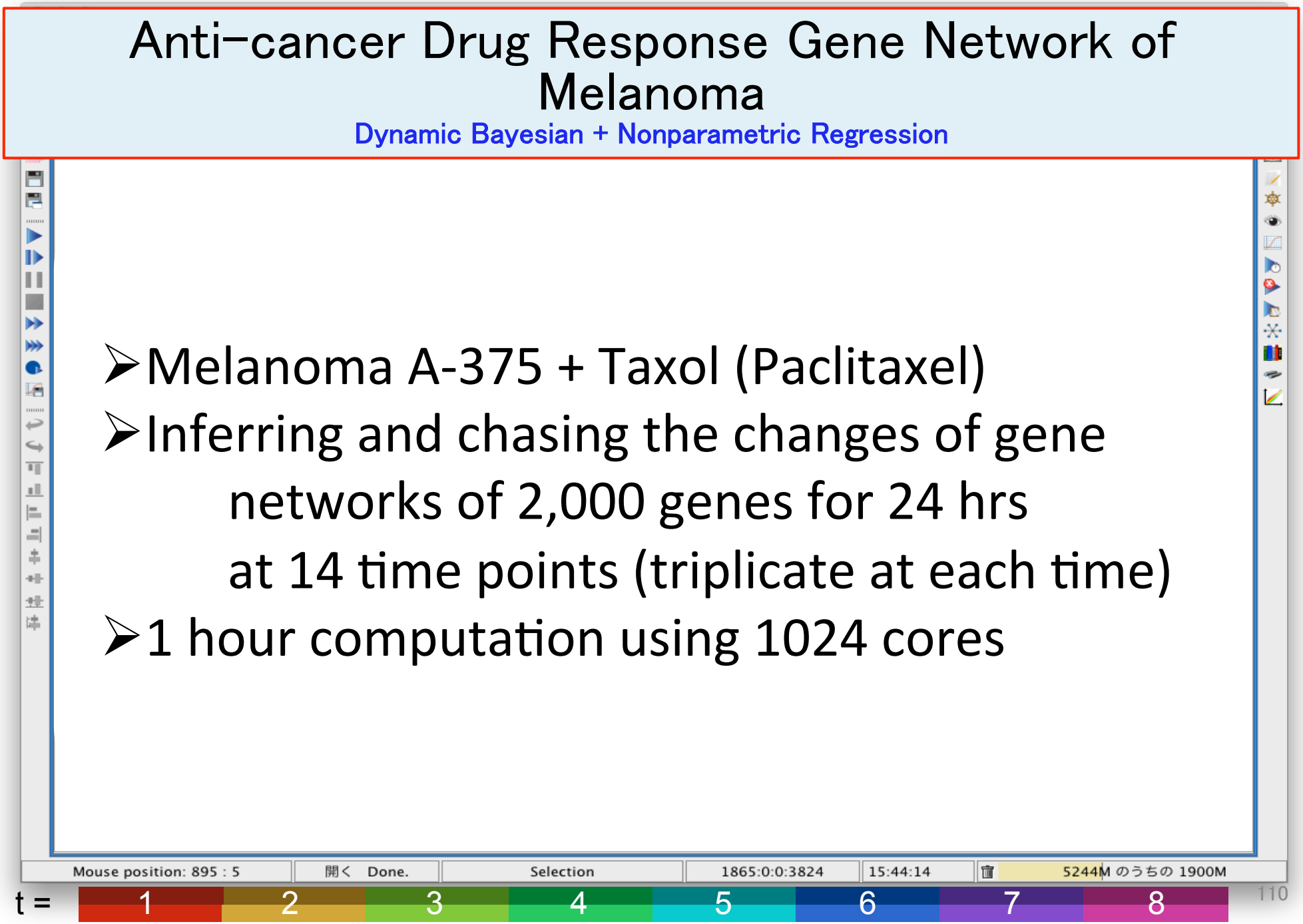
$$+ \sum_{k=1}^{q_j}\left\{(M_{jk}-2)\log\left(\frac{2\pi\hat{\sigma}_j^2}{n\beta_{jk}}\right) - \log|K_{jk}|_+ + \frac{n\beta_{jk}}{\hat{\sigma}_j^2}\hat{\gamma}_{jk}^T K_{jk}\hat{\gamma}_{jk}\right\}$$

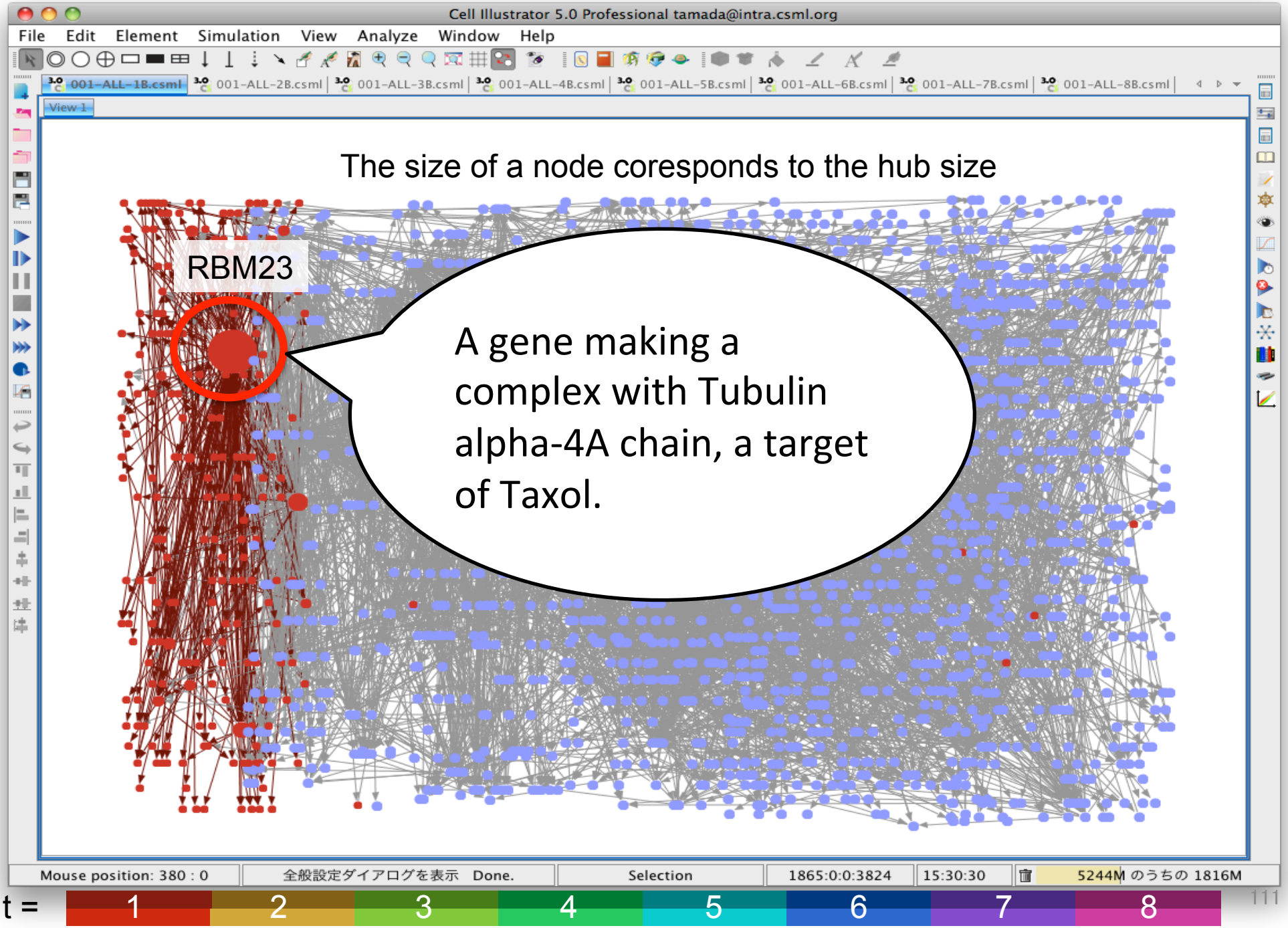
Mathematics is harmful to your health.

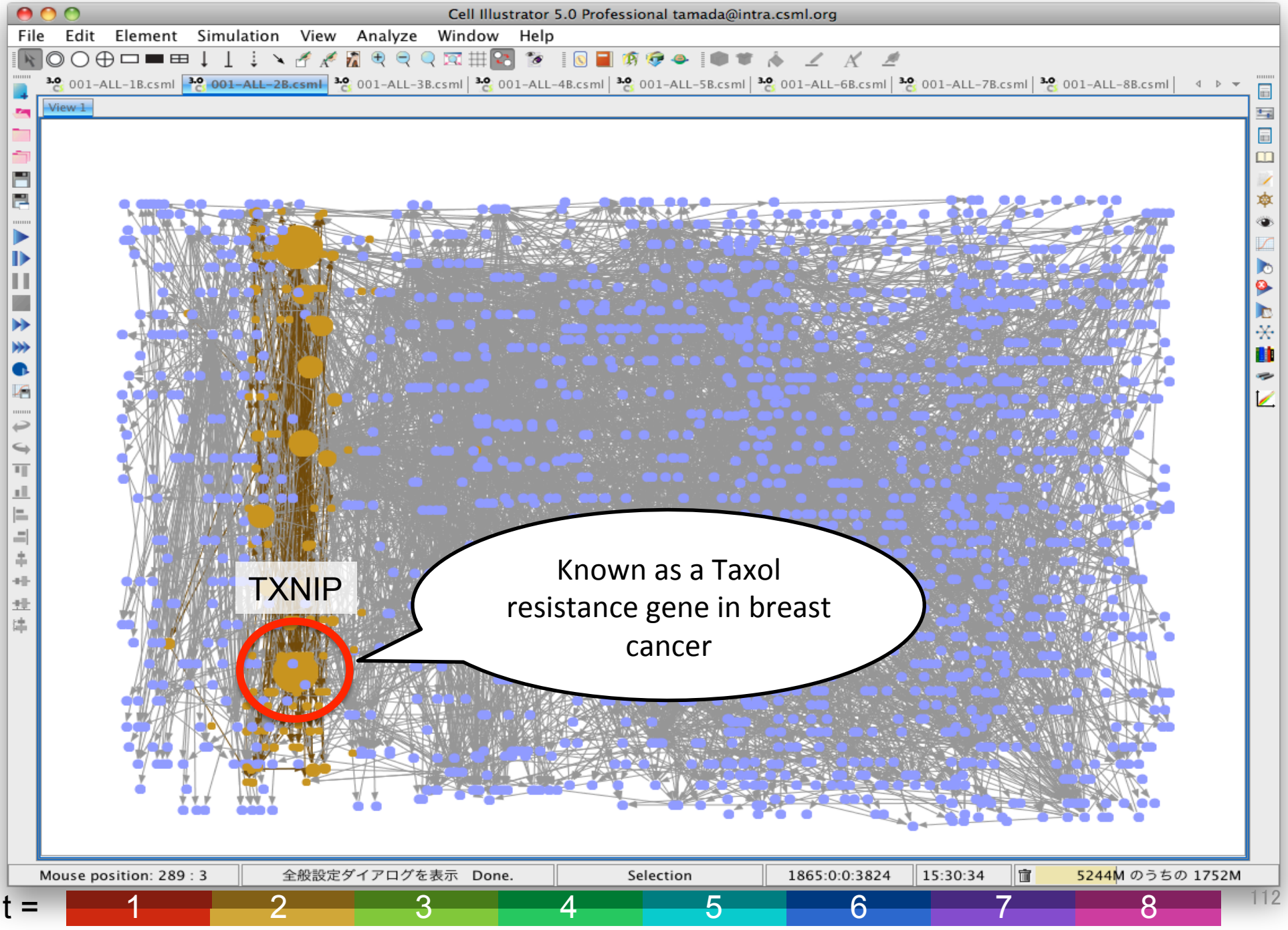
Anti-cancer Drug Response Gene Network of Melanoma

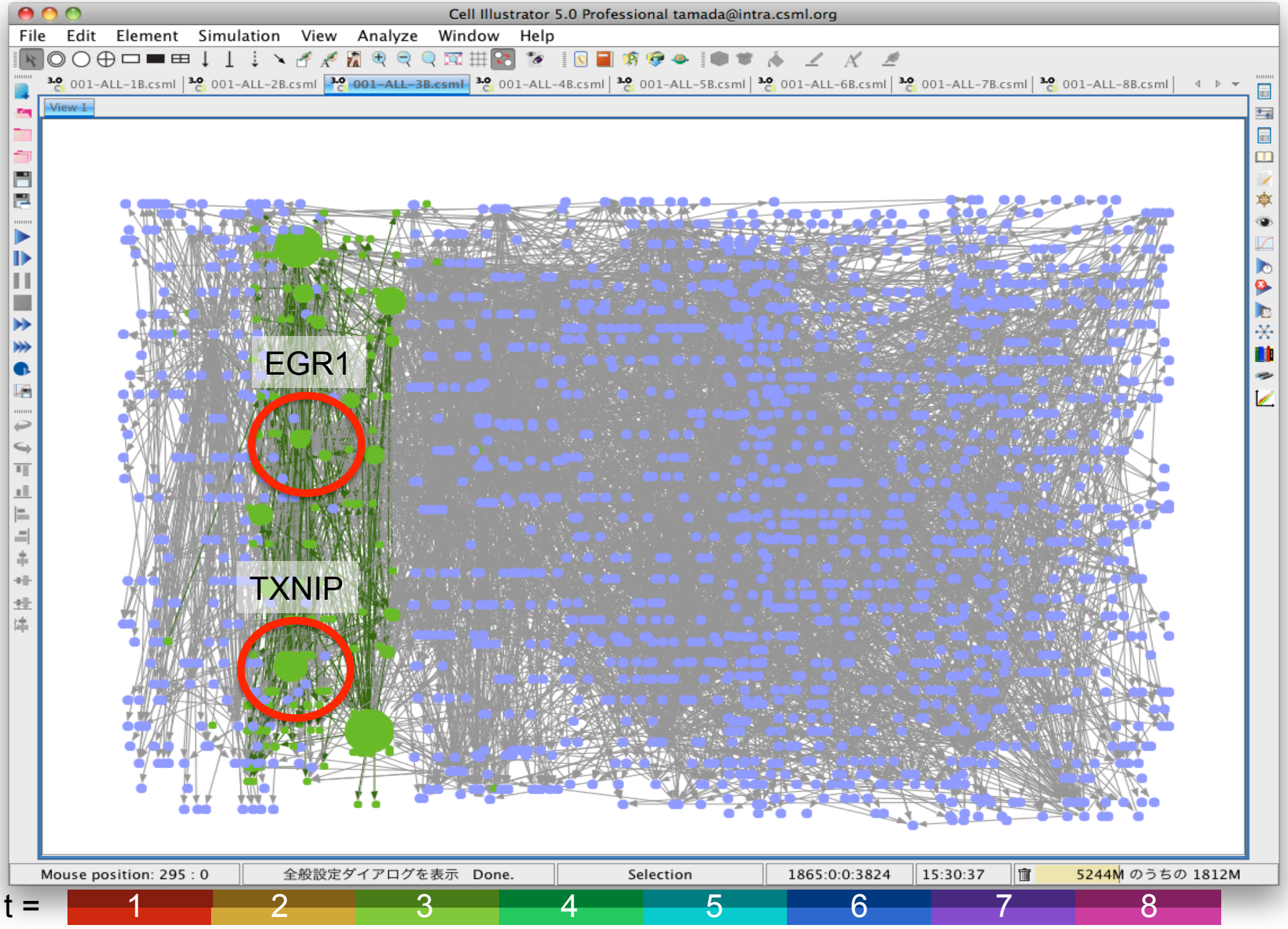
Dynamic Bayesian + Nonparametric Regression

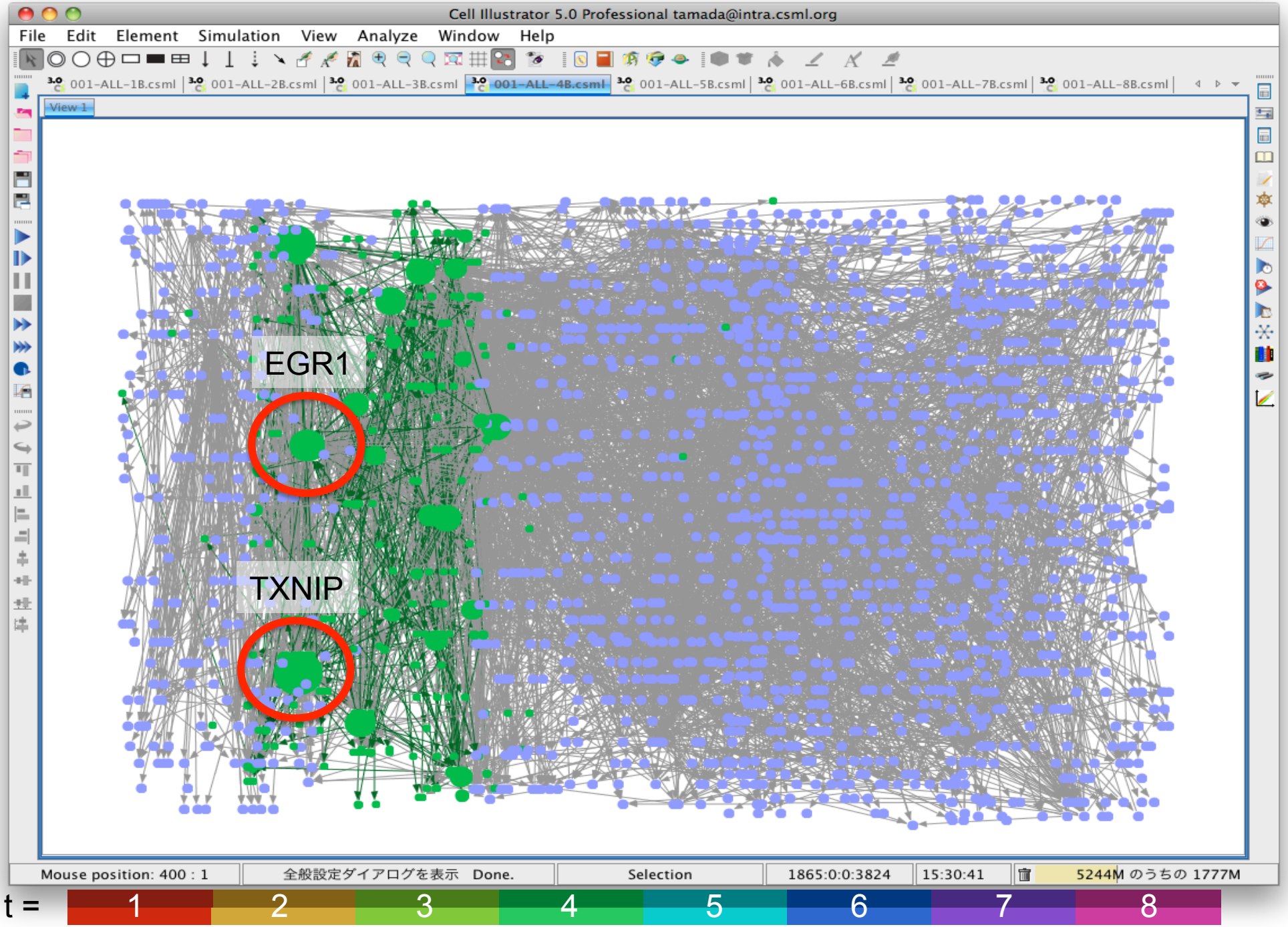
- Melanoma A-375 + Taxol (Paclitaxel)
- Inferring and chasing the changes of gene networks of 2,000 genes for 24 hrs at 14 time points (triplicate at each time)
- 1 hour computation using 1024 cores

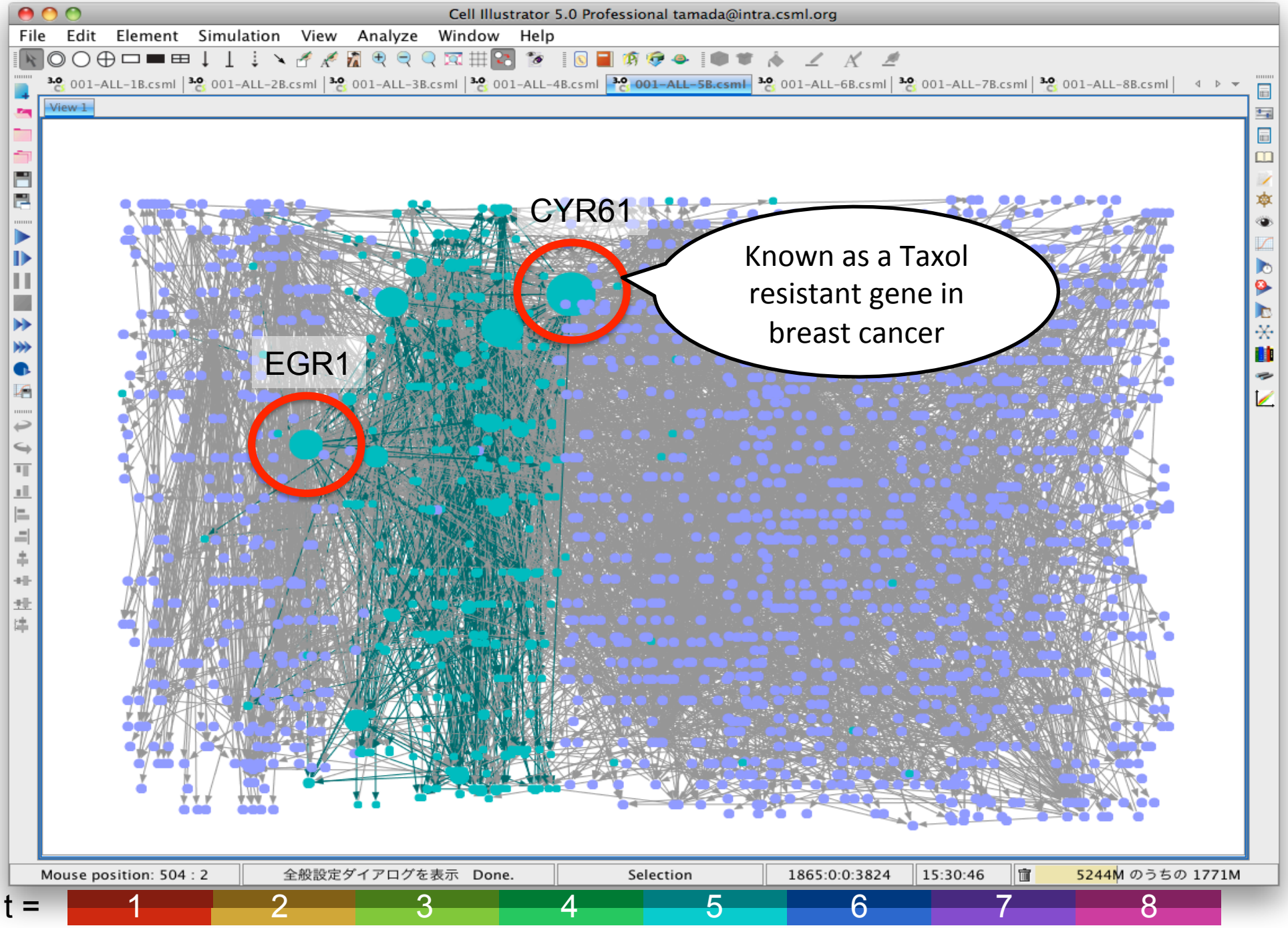


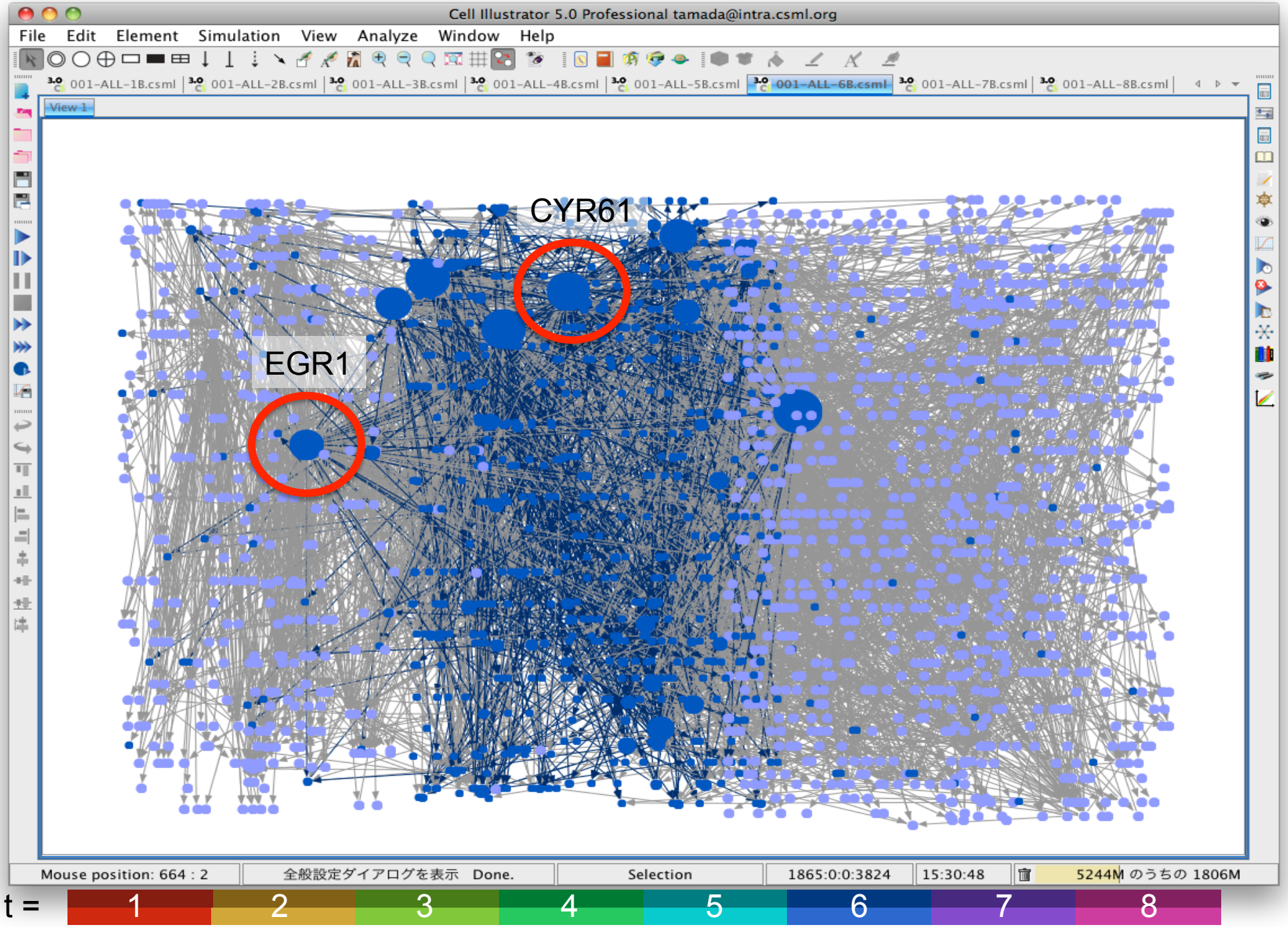

Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml 001-ALL-2B.csml 001-ALL-3B.csml 001-ALL-4B.csml 001-ALL-5B.csml 001-ALL-6B.csml 001-ALL-7B.csml 001-ALL-8B.csml
View 1
The size of a node coresponds to the hub size
RBM23
A gene making a complex with Tubulin alpha-4A chain, a target of Taxol.
Mouse position: 380 : 0
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:30
5244M のうちの 1816M
t =
1
2
3
4
5
6
7
8

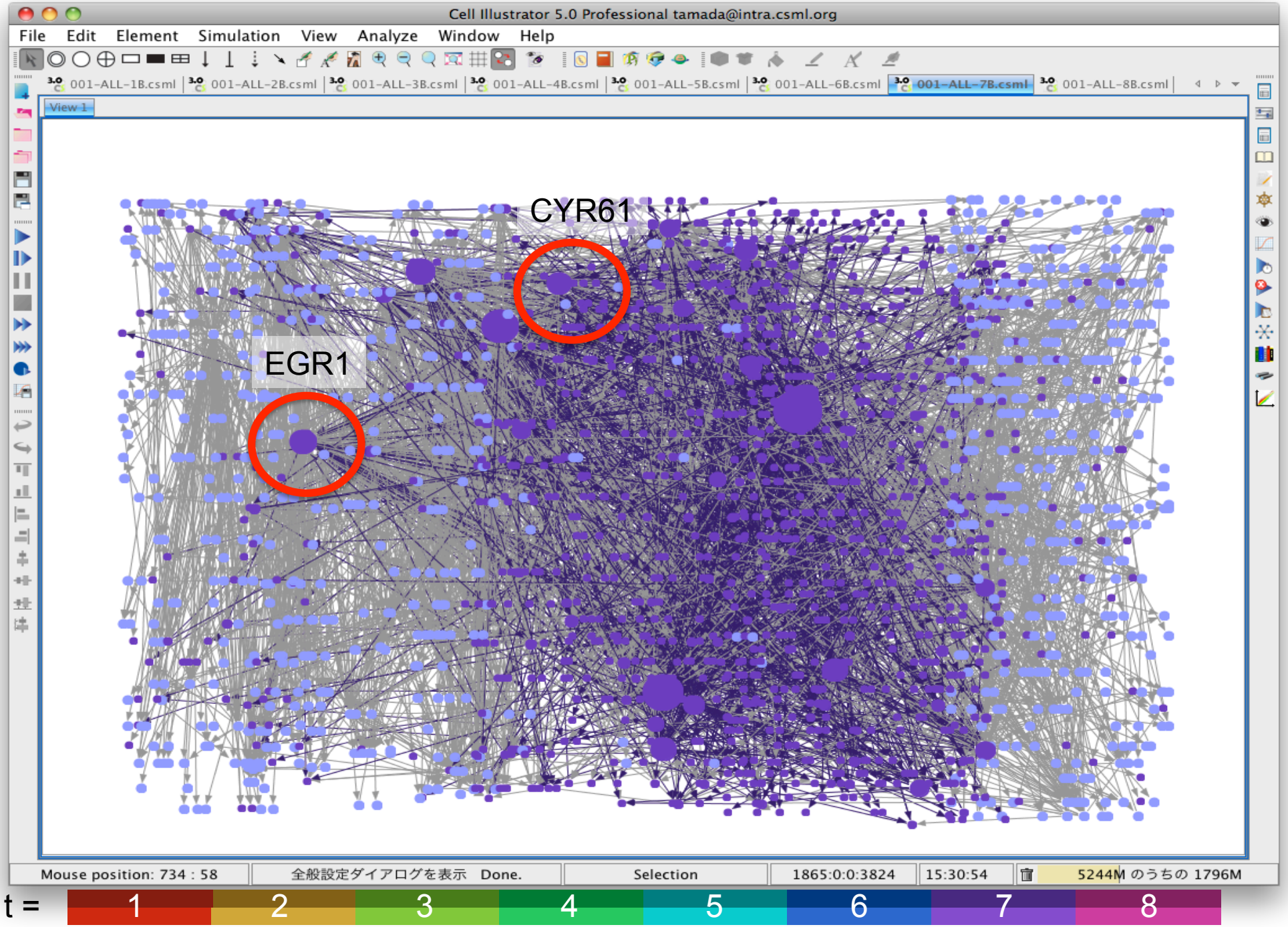

Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml
001-ALL-2B.csml
001-ALL-3B.csml
001-ALL-4B.csml
001-ALL-5B.csml
001-ALL-6B.csml
001-ALL-7B.csml
001-ALL-8B.csml
View 1
TXNIP
Known as a Taxol resistance gene in breast cancer
Mouse position: 289 : 3
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:34
5244M のうちの 1752M
t =
1
2
3
4
5
6
7
8

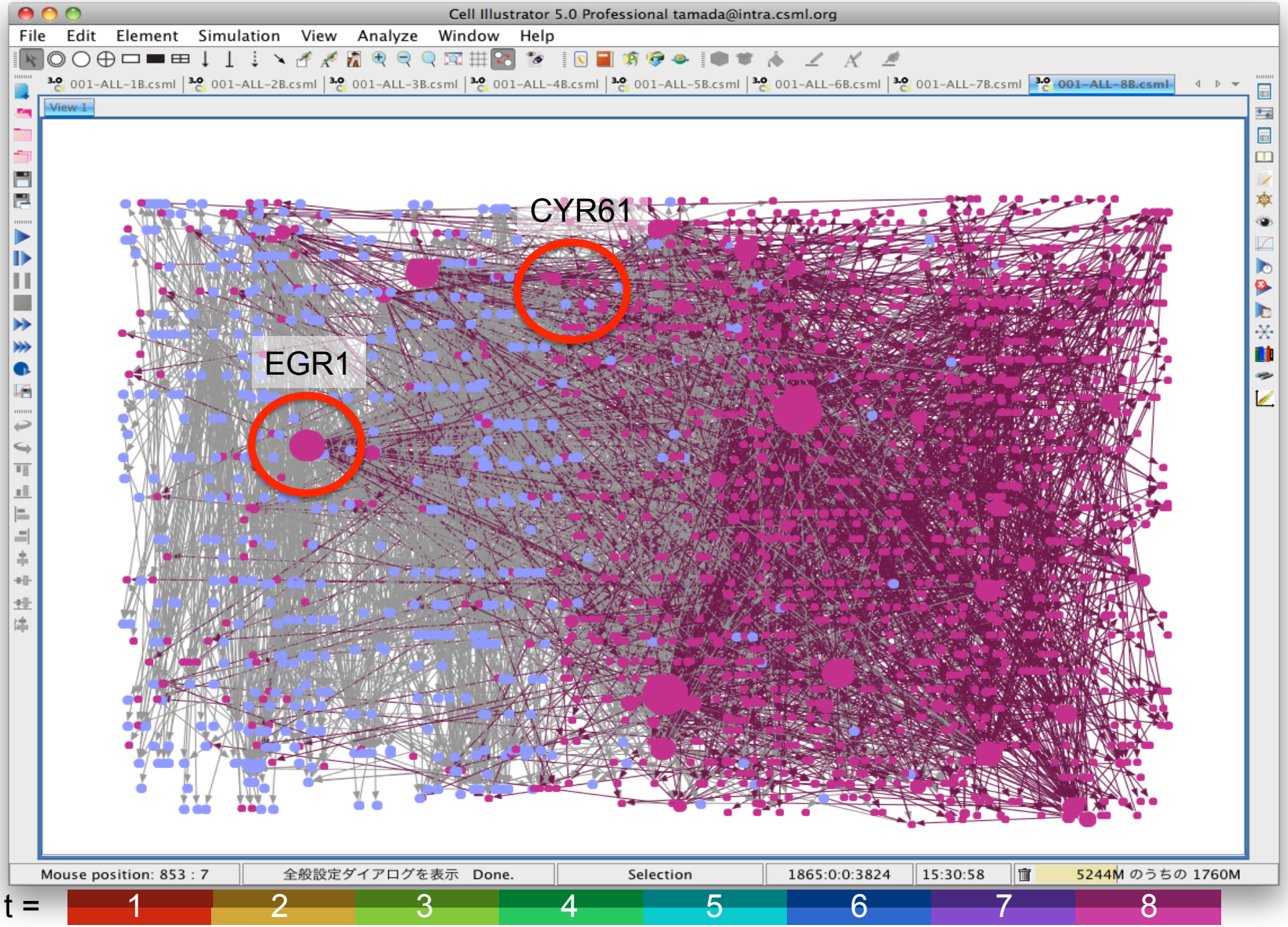

Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml
001-ALL-2B.csml
001-ALL-3B.csml
001-ALL-4B.csml
001-ALL-5B.csml
001-ALL-6B.csml
001-ALL-7B.csml
001-ALL-8B.csml
View 1
EGR1
TXNIP
Mouse position: 295 : 0
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:37
5244M のうちの 1812M
t =
1
2
3
4
5
6
7
8

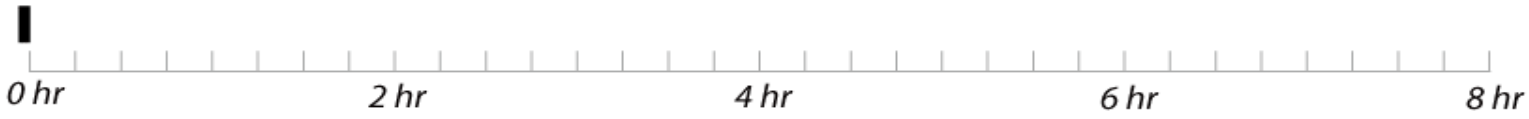

Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml
001-ALL-2B.csml
001-ALL-3B.csml
001-ALL-4B.csml
001-ALL-5B.csml
001-ALL-6B.csml
001-ALL-7B.csml
001-ALL-8B.csml
View 1
EGR1
TXNIP
Mouse position: 400 : 1
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:41
5244M のうちの 1777M
t =
1
2
3
4
5
6
7
8


Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml
001-ALL-2B.csml
001-ALL-3B.csml
001-ALL-4B.csml
001-ALL-5B.csml
001-ALL-6B.csml
001-ALL-7B.csml
001-ALL-8B.csml
View 1
CYR61
EGR1
Known as a Taxol resistant gene in breast cancer
Mouse position: 504 : 2
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:46
5244M のうちの 1771M
t =
1
2
3
4
5
6
7
8


Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml
001-ALL-2B.csml
001-ALL-3B.csml
001-ALL-4B.csml
001-ALL-5B.csml
001-ALL-6B.csml
001-ALL-7B.csml
001-ALL-8B.csml
View 1
CYR61
EGR1
Mouse position: 664 : 2
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:48
5244M のうちの 1806M
t =
1
2
3
4
5
6
7
8


Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml 001-ALL-2B.csml 001-ALL-3B.csml 001-ALL-4B.csml 001-ALL-5B.csml 001-ALL-6B.csml 001-ALL-7B.csml 001-ALL-8B.csml
View 1
CYR61
EGR1
Mouse position: 734 : 58
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:54
5244M のうちの 1796M
t =
1
2
3
4
5
6
7
8


Cell Illustrator 5.0 Professional tamada@intra.csml.org
File Edit Element Simulation View Analyze Window Help
001-ALL-1B.csml 001-ALL-2B.csml 001-ALL-3B.csml 001-ALL-4B.csml 001-ALL-5B.csml 001-ALL-6B.csml 001-ALL-7B.csml 001-ALL-8B.csml
View 1
CYR61
EGR1
Mouse position: 853 : 7
全般設定ダイアログを表示 Done.
Selection
1865:0:0:3824
15:30:58
5244M のうちの 1760M
t =
1
2
3
4
5
6
7
8

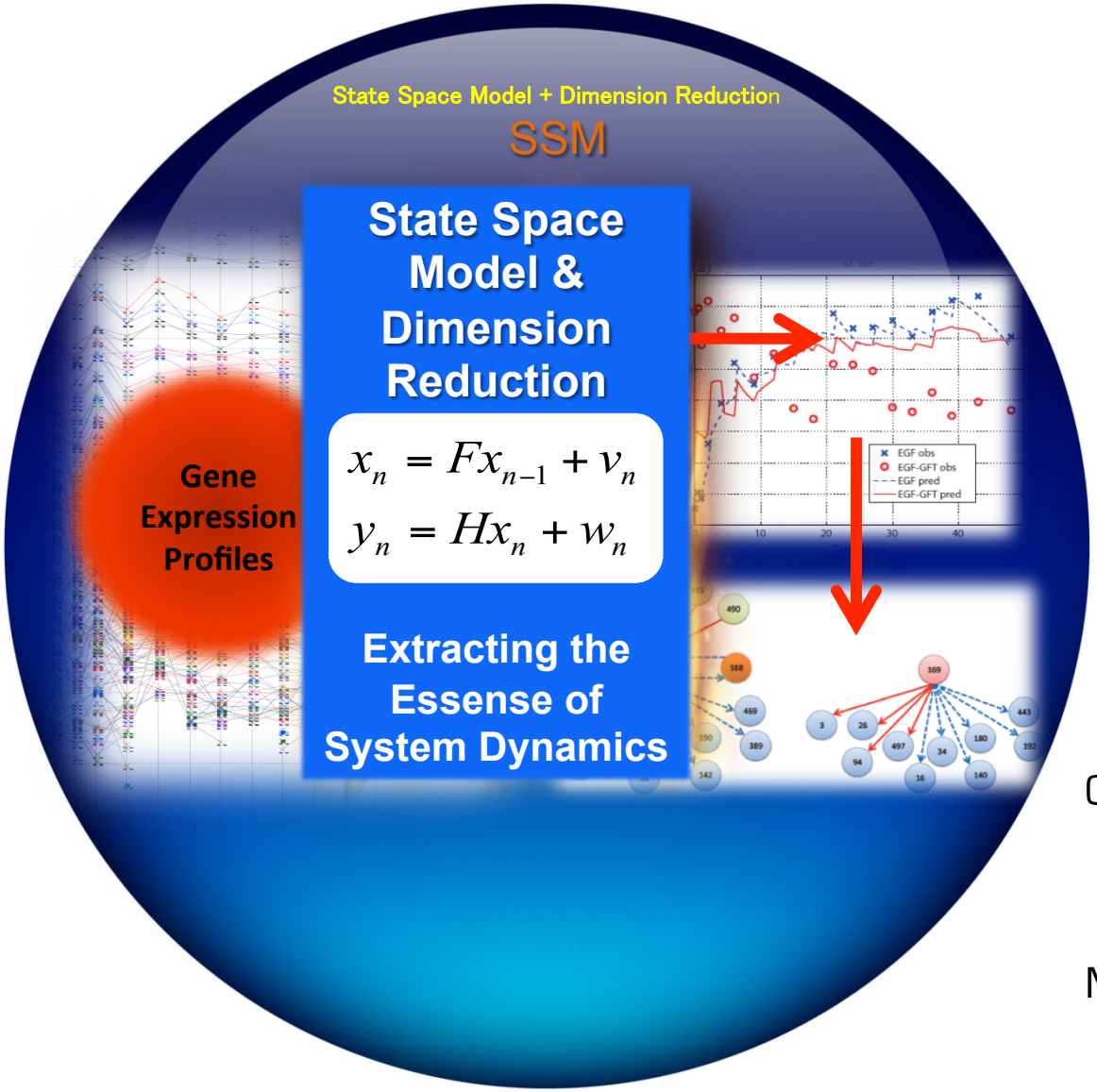

0 hr
2 hr
4 hr
6 hr
8 hr

3. Gene Networks of Small Airway Epithelial Cell and Gefitinib - State Space Model -

Growth Factor Signaling Systems Identify Critical Genes for Survival Prediction in Early Stage Lung Adenocarcinoma

Yamaguchi, R., Imoto, S., Yamauchi, M., Nagasaki, M., Yoshida, R., Shimamura, T., Hatanaka, Y., Ueno, K., Higuchi, T., Gotoh, N., Miyano, S.

Modeling Cancer Systems with State Space Model



State Space Model:

$$x_n = Fx_{n-1} + v_n \in R^k \quad \text{System Model}$$

$$y_n = Hx_n + w_n \in R^p \quad \text{Observation Model}$$

High-Dimensional Short Time Series Data:

$$Y = \{y_1, \cdots, y_{N_{\text{obs}}}\} \qquad N_{\text{obs}} << p \approx 10^3$$

System Estimation with Dimension Reduction:

$$\dim(x_n) = k < \dim(y_n) = p$$

$$L(\theta) = \int p(Y, X \mid \theta) dX$$

$$H^T R^{-1} H = \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_k)$$

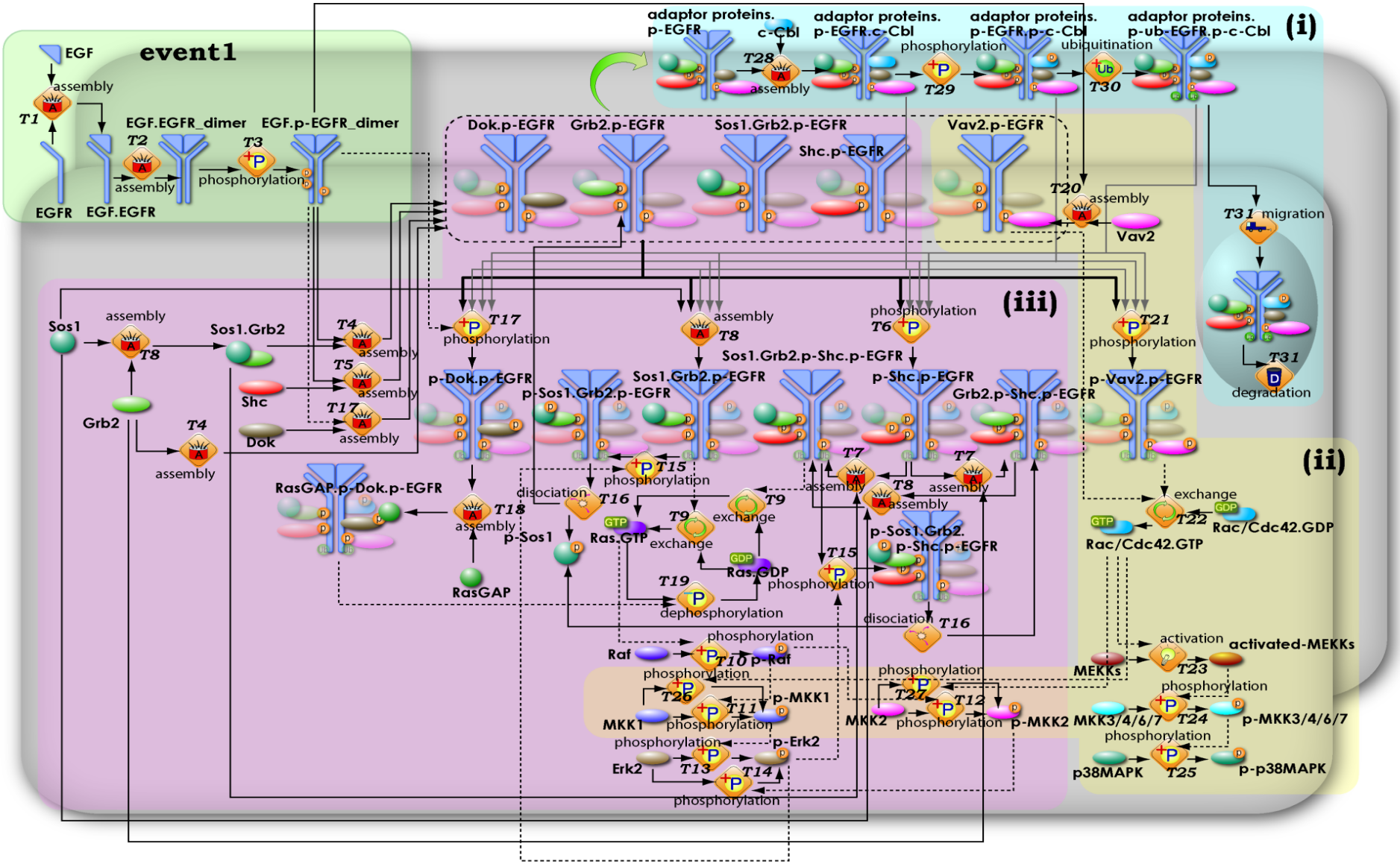
Gene Expression Prediction:

$$y_{n|n-1} = H \int x_n p(x_n \mid y_1, \cdots, y_{n-1}, \theta) dx_n$$

Module-Based Gene Network Estimation:

$$R^{-1/2}(y_n - w_n) = \Psi R^{-1/2}(y_{n-1} - w_{n-1}) + R^{-1/2} H v_n$$

Epidermal Growth Factor Receptor Pathway



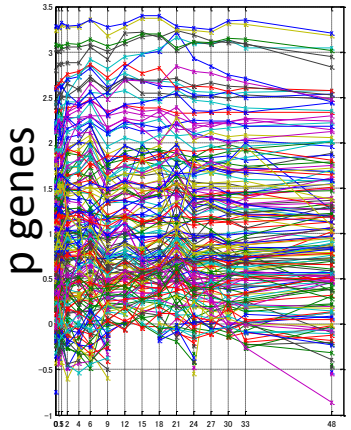
State Space Model for Inferring Transcriptional Module Networks from Time-Course Gene Expression Data

SSM

$$x_n = Fx_{n-1} + v_n,\ n \in \mathcal{N},$$
$$y_n = Hx_n + w_n,\ n \in \mathcal{N}_{\text{obs}},$$

$$\text{SSM}(\theta) \quad \theta = \{H, F, R, x_0\}$$

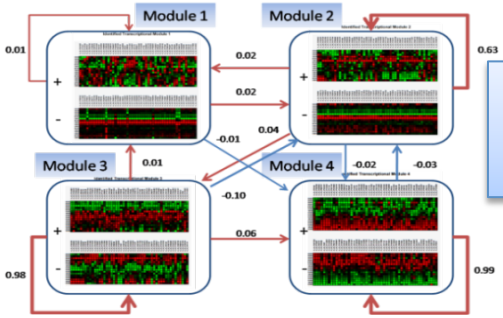
Time-Series Data



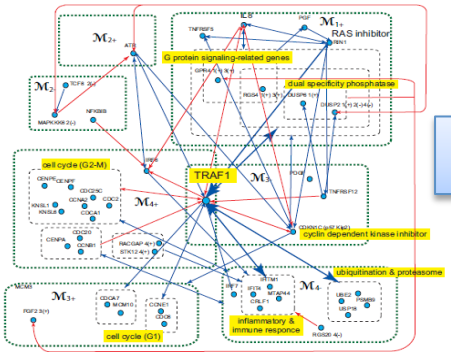
N_{obs} Time points

Short- and High-dimensional Data

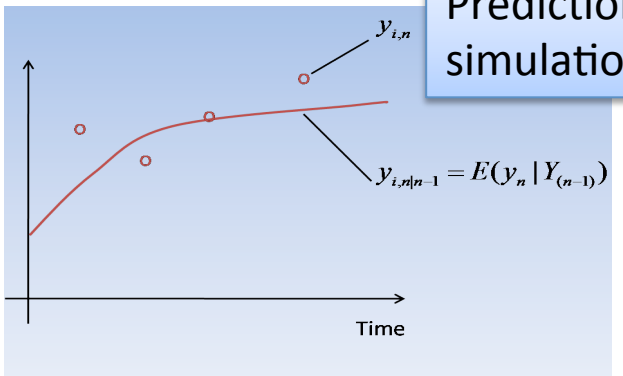
$$Y_{\mathcal{N}_{\text{obs}}} = \{y_n\},\ n \in \mathcal{N}_{\text{obs}}$$



Module Network



Gene Network

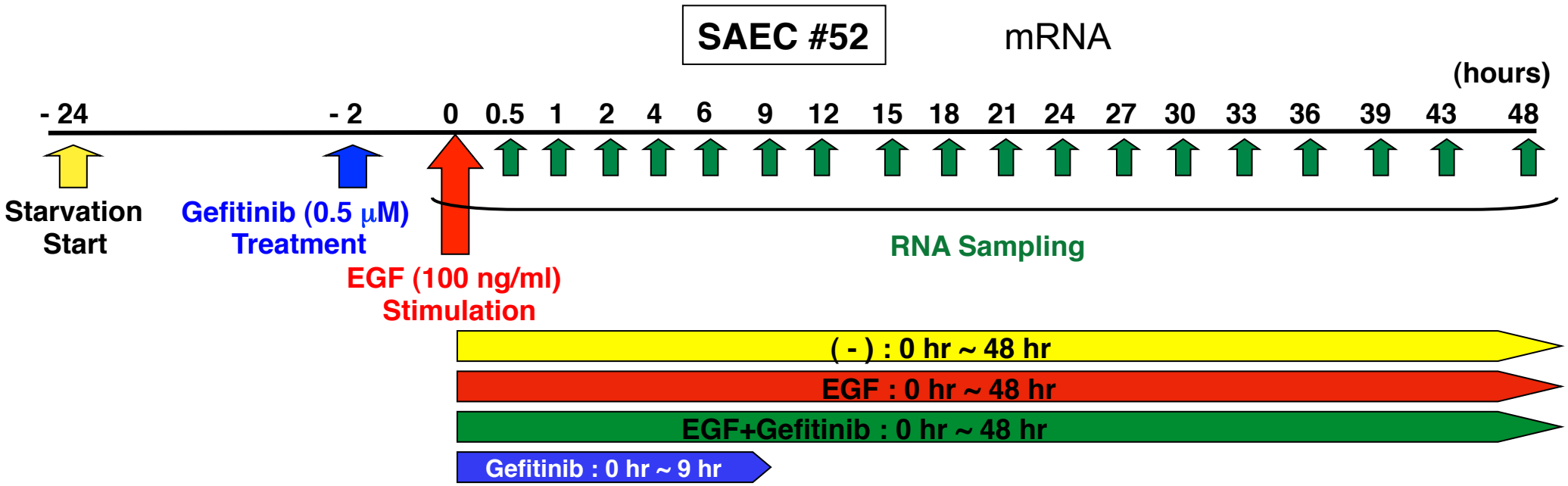


Prediction by simulation

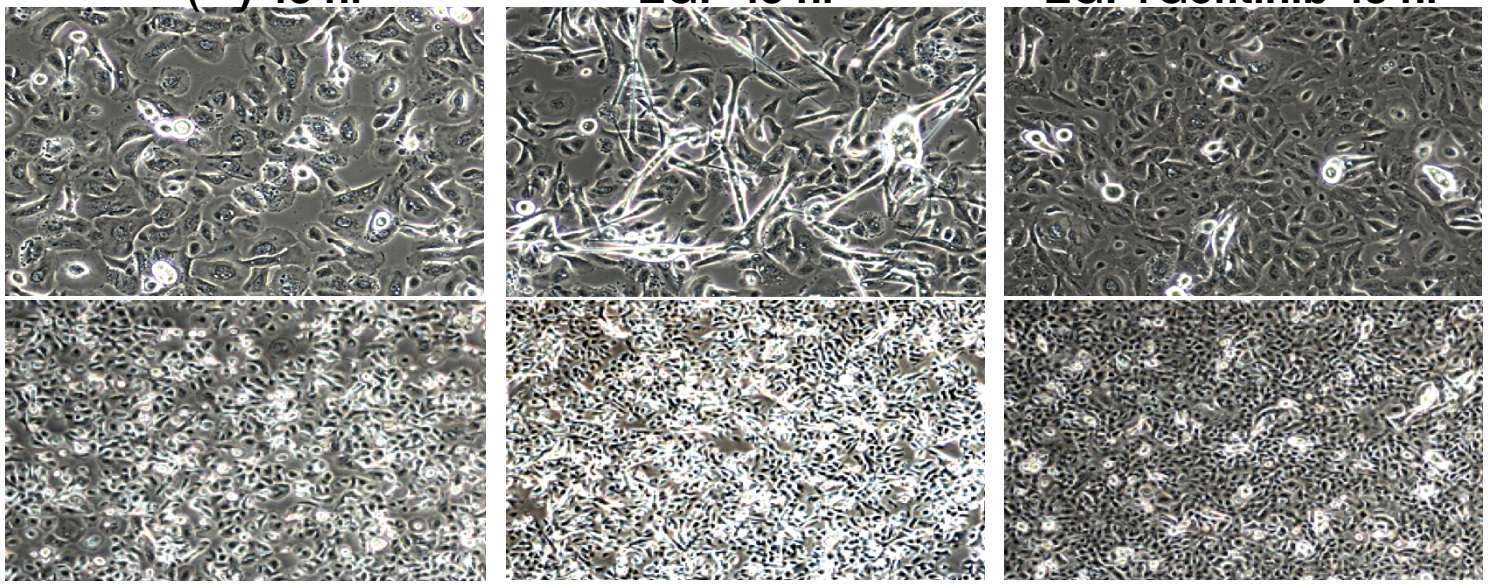
SSM Application

Predicting Differences in Gene Regulatory Systems

- Focus: EGFR pathway
- Data
  - Time Course Gene Expression Microarray Data (20K)
    - 19 time points during 48 hours
  - Human Small Airway Epithelial Cell (SAEC)
  - Two Conditions (Different Drugs)
    - EGF Stimulation (Control Data)
    - EGF + Gefitinib Dosed (Case Data)

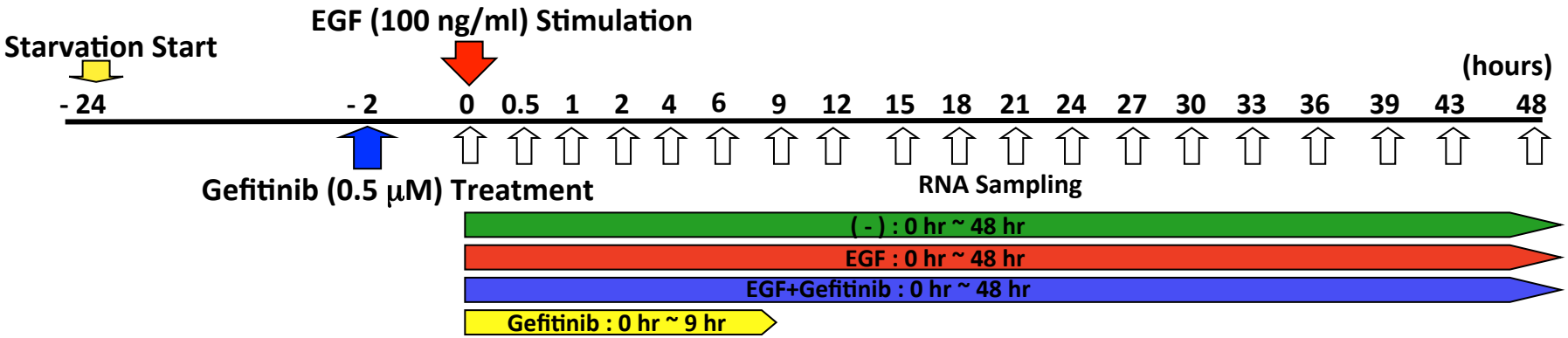


Cell Morphology

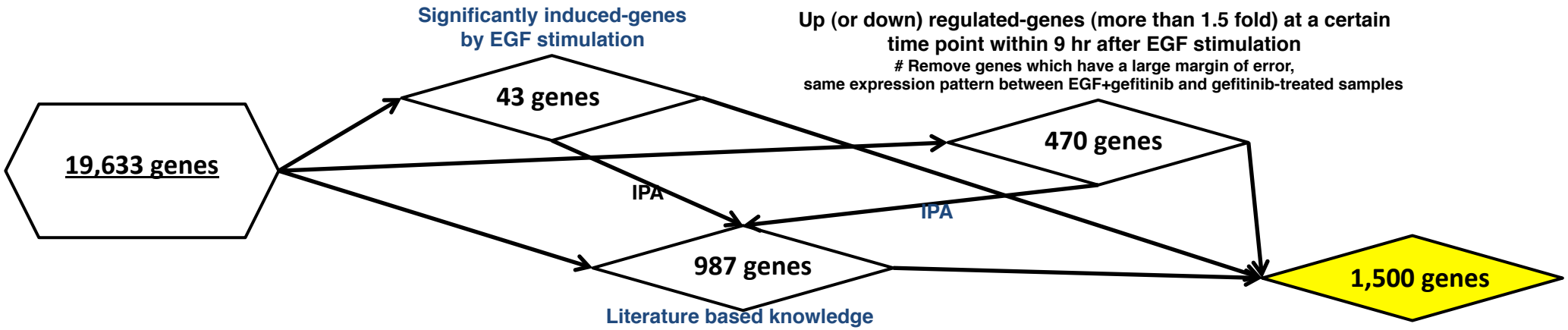


Selection 1500 Genes for Network Analysis

a



b



Time-Course of Drug-Stimulated Human Normal Lung Cell

Cell: SAEC (Human Small Airway Epithelial Cell)
Conditions: Control EGF (mimic of Cancer Cells)
Case EGF＋Gefitinib (mimic of Cancer Cells + Drug)
Time-Course: 19 Time Points in 48 hours

Gene Selection for SSM Analysis → **1500 Genes: Literature DB (Ingenuity) + Variation Filter (variance)**

Selection of System Dimension: k
Parameter Estimation: θ → **Learning SSM with EGF data (Control's System)**
k = 8 minimizes prediction errors for hold-out sample.

Meta Analysis of Transcriptional Modules

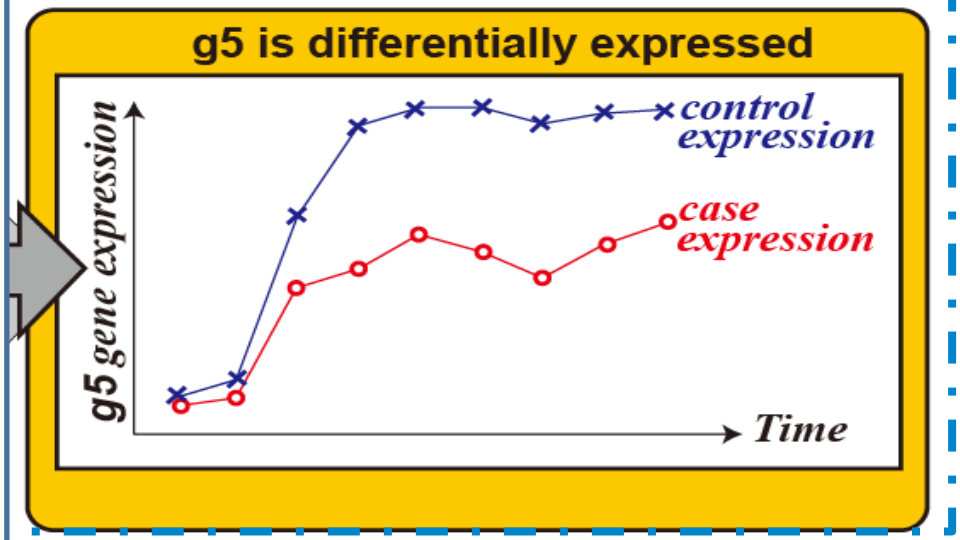
Construction of Gene Network

Prediction of EGF-Gefitinib data by EGF-Learned SSM
Obtaining *p*-value for each Gene

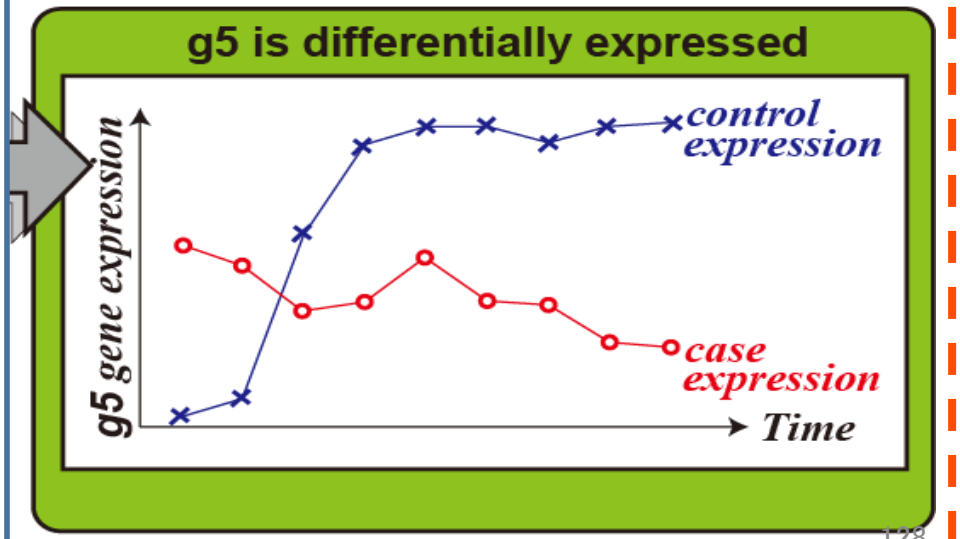
Genes with small p-values are considered to be induced differential regulations by Gefitinib

Differentially-Expressed and Differentially-Regulated Genes

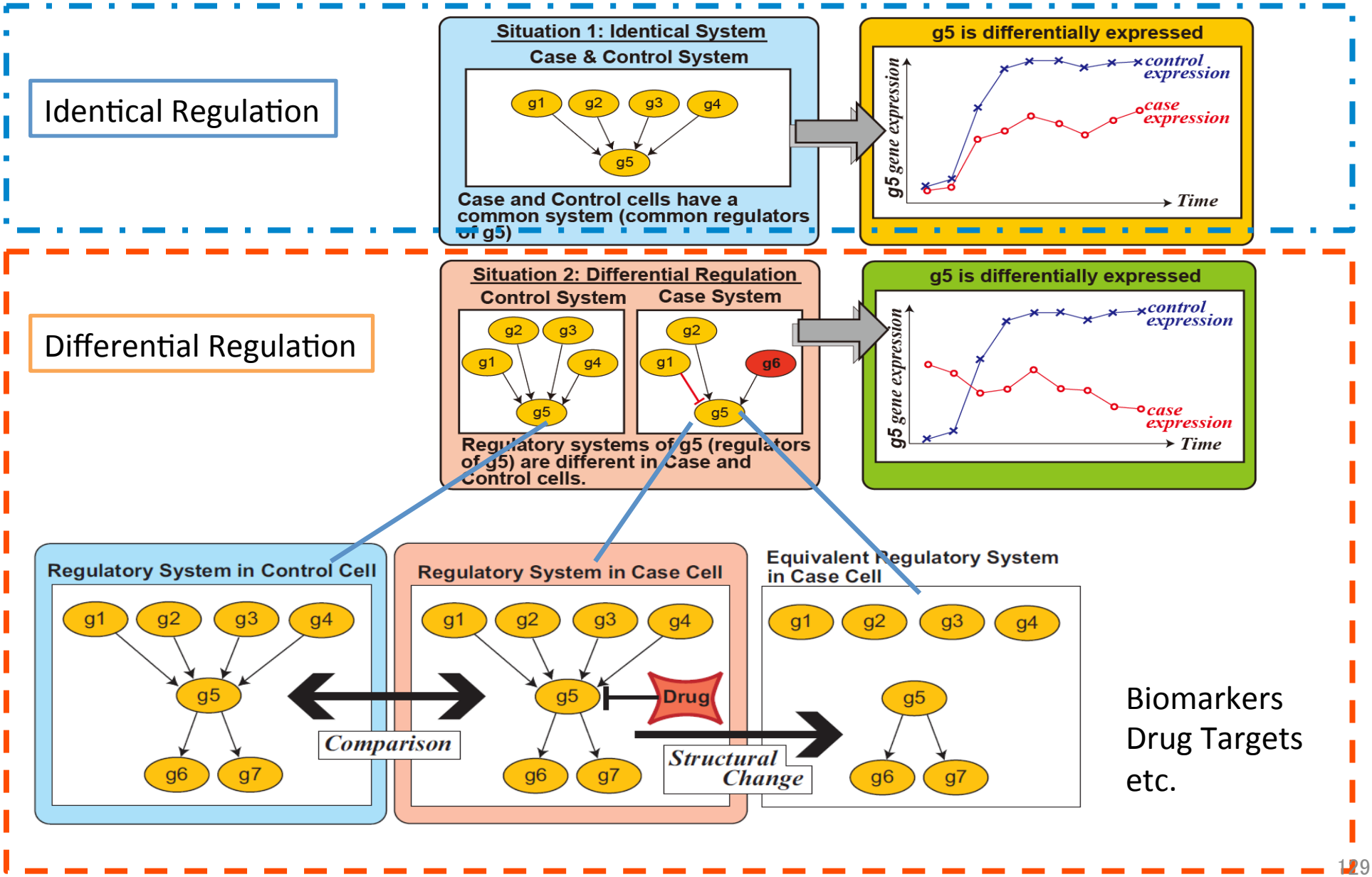
Time-Course Gene Expression Profiles from a Case-Control Experiment.



What kind of systems are behind differential expressions?

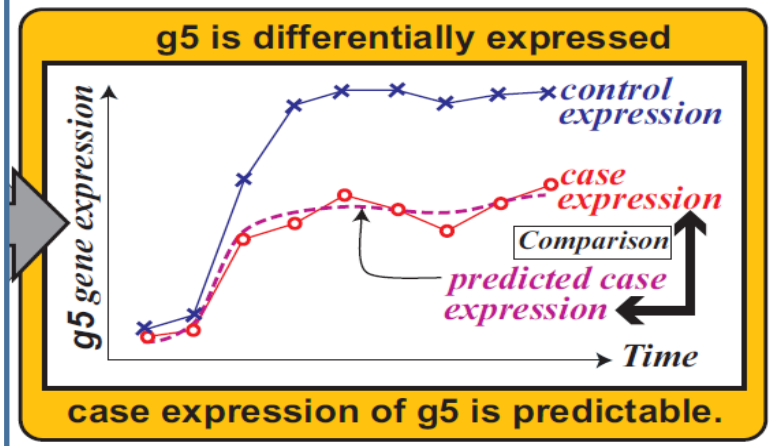


Differential Regulations by Drug Dosing

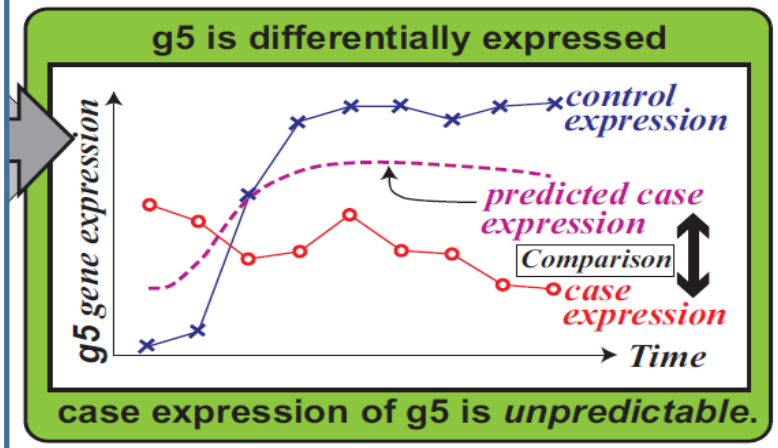


Predicting Case Data by Control's System to Discriminate the Two Situations

A Thought Experiment: If we know the Control's System, we use it to predict the Case data.



Case is **predictable** from Control's System

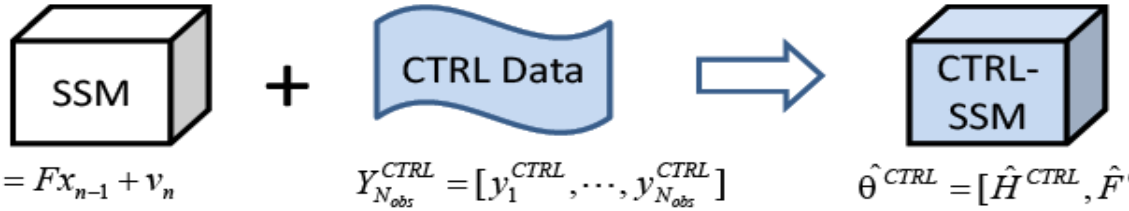


Case is **unpredictable** From Control's System

We use a statistical model for inferring gene regulatory systems.

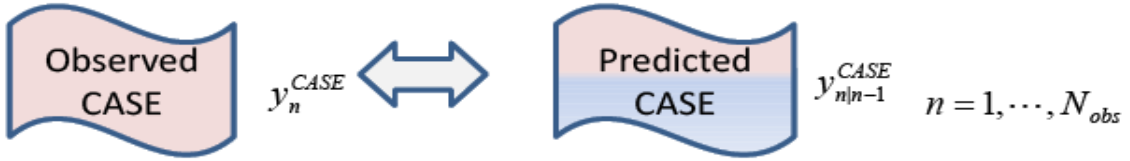
Strategy to Predict Differentially Regulated Genes

1. Train SSM by CTRL time-course data and Estimate an CTRL-SSM System



2. Predict CASE time-course data by the CTRL-SSM System

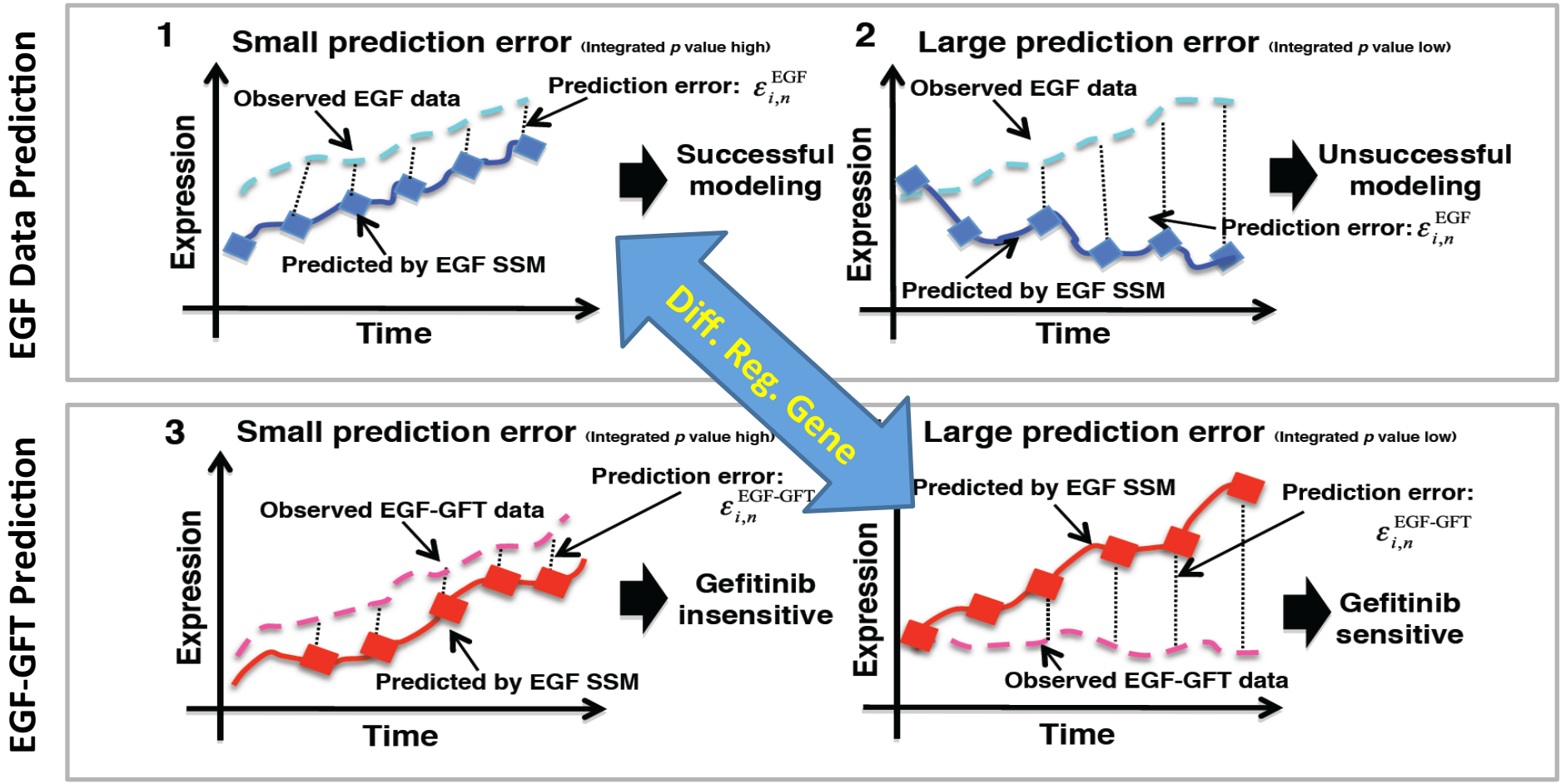
3. Compare Predicted and Observed Time Courses



Candidates of Differentially Regulated Genes:
Unpredictable genes in the Case data by the Control's SSM.

Prediction of EGF Data and EGF-GFT Data by SSM of EGF System

EGF Data Prediction: Check for Model Accuracy
EGF+GFT Data Prediction: Exploration for Differences between Systems

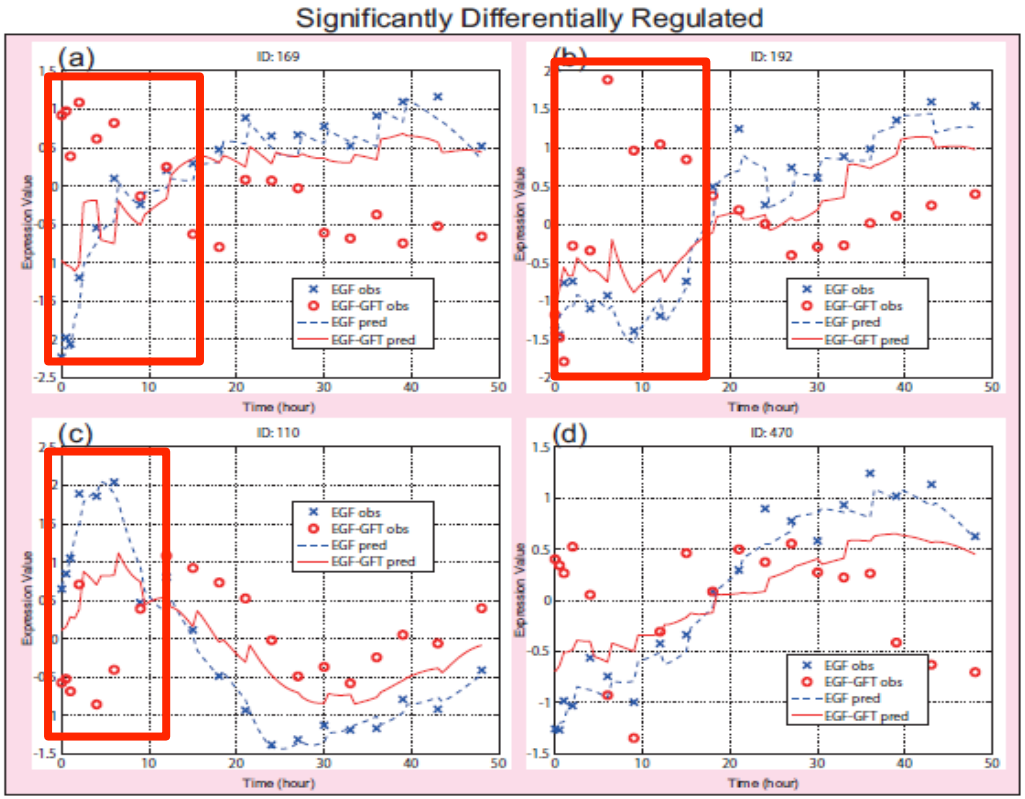


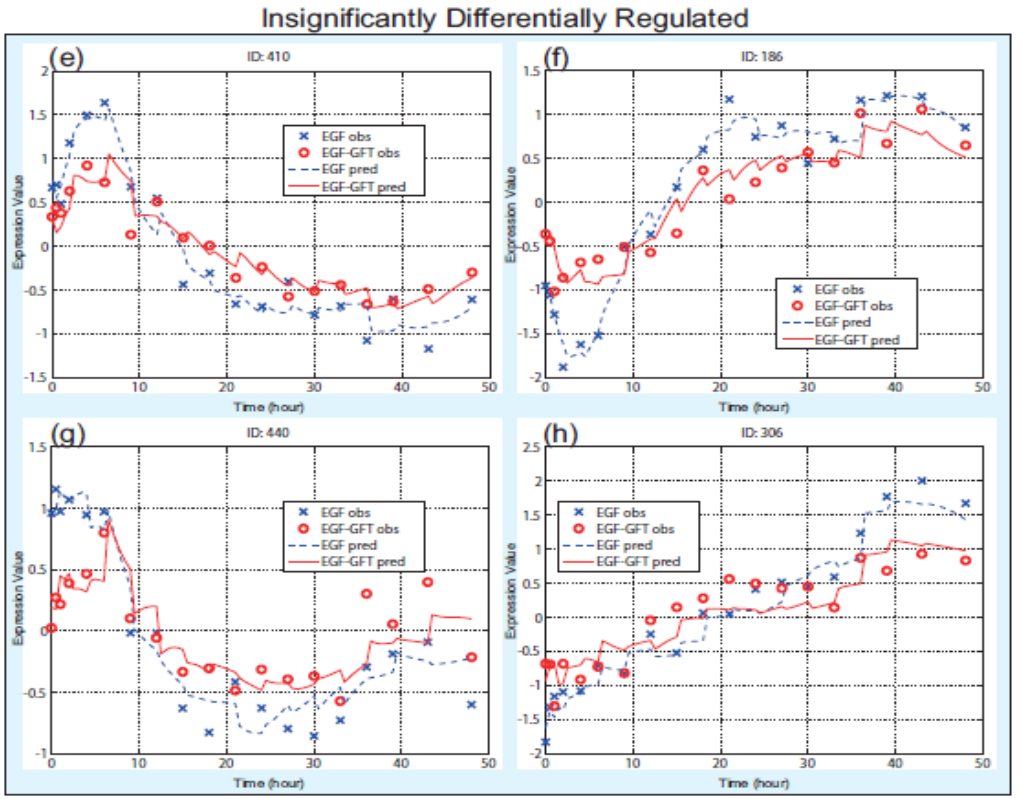
X : EGF obs (Control)
O : EGF+GFT obs (Case)
--- : EGF pred by SSM(EGF)
— : EGF+GFT pred by SSM(EGF)

Examples: Selected Diff Reg Genes

Differentially Regulated Genes

Similarly Regulated Genes

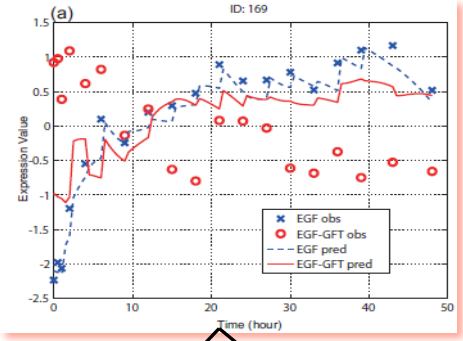


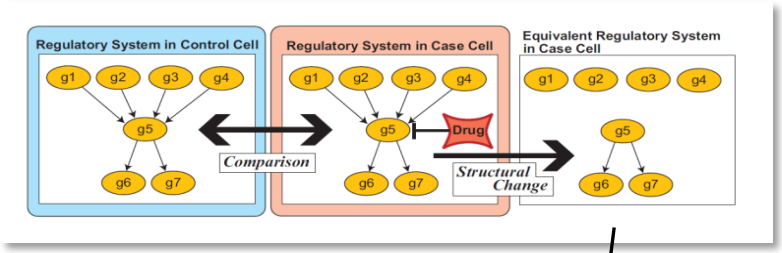


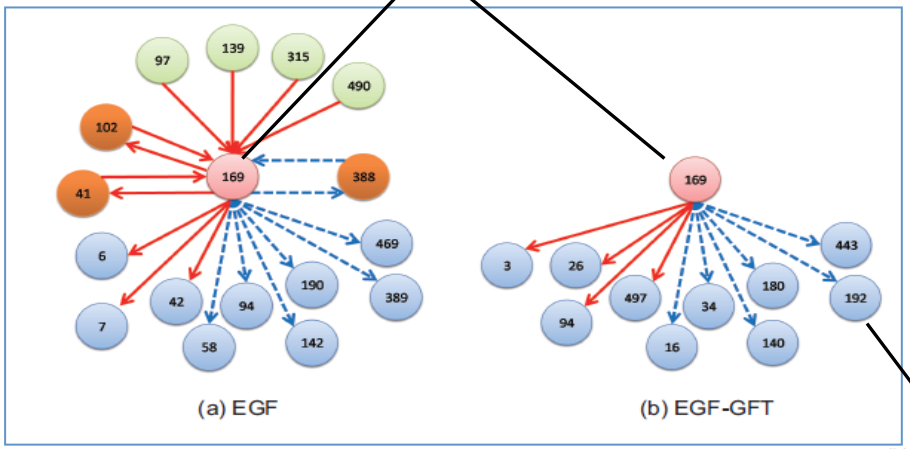
Estimated Gene Networks with SSMs

$$R^{-1/2}(y_n - w_n) = \Psi R^{-1/2}(y_{n-1} - w_{n-1}) + R^{-1/2} H v_n$$

Diff gene
(gene 162)





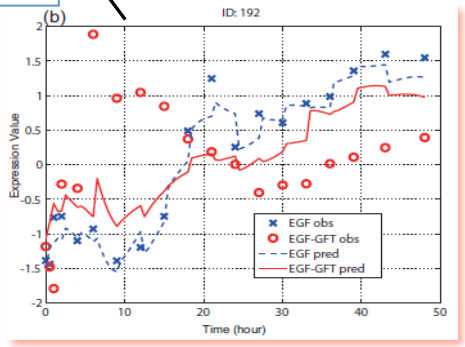


Predicted
disappearance
of parent nodes
by drug dosing

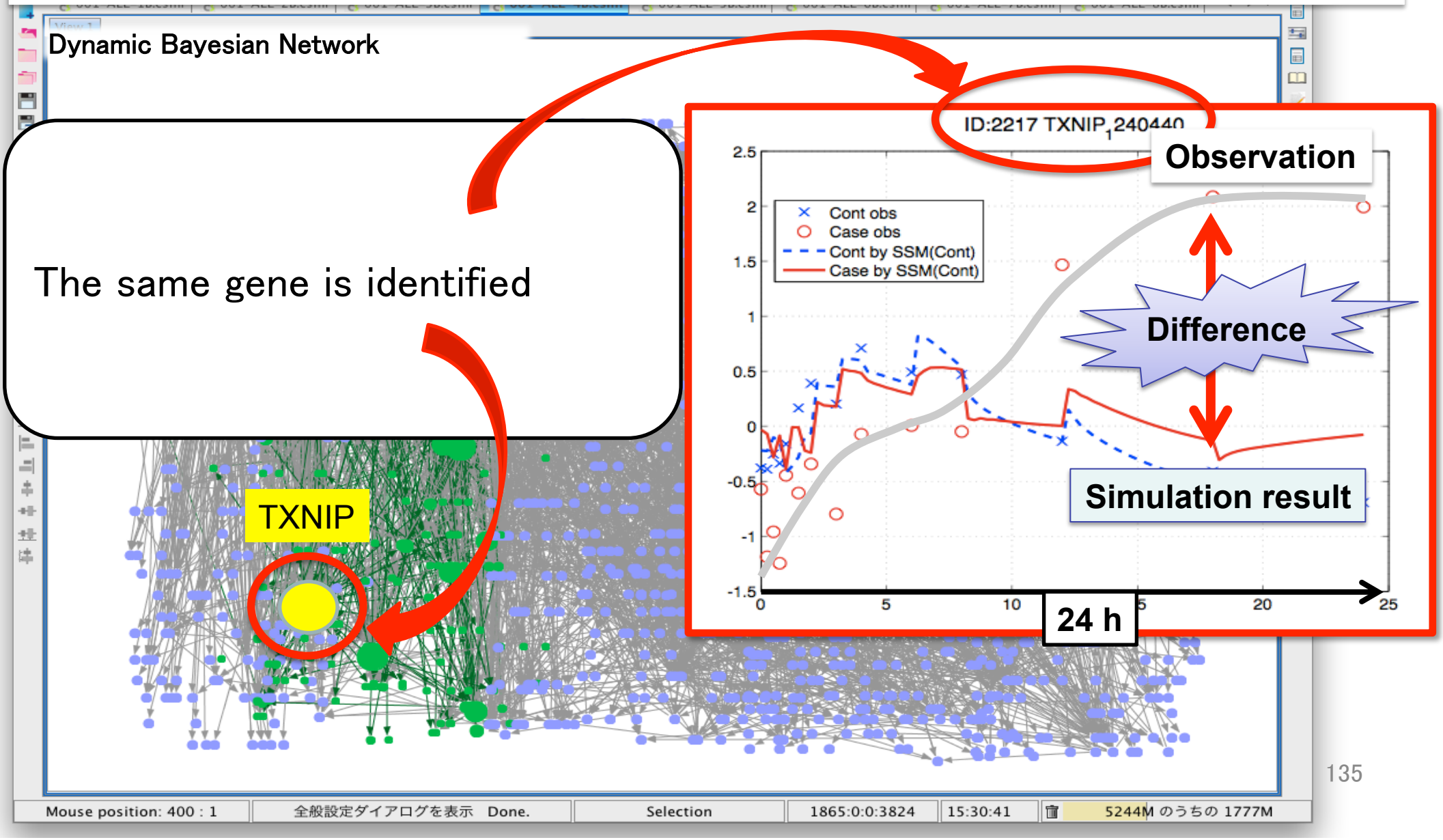
Control
(EGF)

Case
(EGF+GFT)

Diff gene
(gene 192)



State Space Model computed from Melanoma + Taxol



ARTICLES

nature medicine

Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study

Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma*[,1]
Kerby Shedden[2,3,17], Jeremy M G Taylor[3,4,17], Steven A Enkemann[5,17], Ming-Sound Tsao[6,17], Timothy J Yeatman[5,17], William L Gerald[7,17], Steven Eschrich[5,17], Igor Jurisica[6,17], Thomas J Giordano[8], David E Misek[3,9], Andrew C Chang[3,9], Chang Qi Zhu[6], Daniel Strumpf[6], Samir Hanash[3], Frances A Shepherd[6], Keyue Ding[10], Lesley Seymour[10], Katsuhiko Naoki[11], Nathan Pennell[11], Barbara Weir[11], Roel Verhaak[11], Christine Ladd-Acosta[12], Todd Golub[12], Michael Gruidl[5], Anupama Sharma[5], Janos Szoke[7], Maureen Zakowski[7], Valerie Rusch[7], Mark Kris[7], Agnes Viale[7], Noriko Motoi[7], William Travis[7], Barbara Conley[13], Venkatraman E Seshan[14,17], Matthew Meyerson[11,12,17], Rork Kuick[3,17], Kevin K Dobbin[15,17], Tracy Lively[16,17], James W Jacobson[16,17] & David G Beer[3,9,17]

Although prognostic gene expression signatures for survival in early-stage lung cancer have been proposed, for clinical application, it is critical to establish their performance across different subject populations and in different laboratories. Here we report a large, training–testing, multi-site, blinded validation study to characterize the performance of several prognostic models based on gene expression for 442 lung adenocarcinomas. The hypotheses proposed examined whether microarray measurements of gene expression either alone or combined with basic clinical covariates (stage, age, sex) could be used to predict overall survival in lung cancer subjects. Several models examined produced risk scores that substantially correlated with actual subject outcome. Most methods performed better with clinical data, supporting the combined use of clinical and molecular information when building prognostic models for early-stage lung cancer. This study also provides the largest available set of microarray data with extensive pathological and clinical annotation for lung adenocarcinomas.

In the United States and in many Western countries, lung cancer represents the leading cause of cancer-related death[1]. The 5-year, overall survival rate is 15% and has not improved over many decades. This is mainly because approximately two-thirds of lung cancers are discovered at advanced stages, for which cure by surgical resection is no longer an option. Furthermore, even among early-stage patients who are treated primarily by surgery with curative intent, 30–55% will develop and die of metastatic recurrence. Recent multinational clinical trials (IALT, JBR10, ANITA, UFT, LACE) conducted in several continents have demonstrated that adjuvant chemotherapy significantly improves the survival of patients with early-stage (IB–II) disease[2]. Nevertheless, it is clear that a proportion of patients with stage I disease have poorer prognosis and may benefit significantly from adjuvant chemotherapy, whereas some with stage II disease with relatively good prognoses may not benefit significantly from adjuvant chemotherapies. It remains possible, however, that the latter patients could derive additional benefit from adjuvant targeted therapies[3–4]. Therefore, there is an urgent need to establish new diagnostic paradigms and validate in clinical trials methods for improving the selection of stage I–II patients who are most likely to benefit from adjuvant chemotherapy.

Global gene-expression profiling using microarray technologies has helped to improve our understanding of the histological heterogeneity of non–small cell lung cancer (NSCLC) and has identified potential biomarkers and gene signatures for classifying patients with significantly different survival outcomes[5–11]. However, the performance and general applicability of published classifiers has not been easy to establish because of small numbers of subjects examined and inclusion of heterogeneous tumor types. Furthermore, there have not been uniform criteria for sample inclusion, annotation, sample processing and data analyses. To address these concerns and to generate a large microarray database of NSCLC samples that have been collected and studied using a common protocol[12], we conducted a large retrospective, multi-site, blinded study. The study included a blinded validation step to characterize the performance of several newly developed prognostic models using a total of 442 lung adenocarcinomas, the specific type of lung cancer that is increasing in incidence[13].

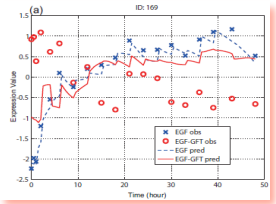
To ensure scientific validity of the results, subject samples along with all relevant clinical, pathological and outcome data were collected by investigators at four institutions using data from six lung-cancer treatment sites with subject inclusion criteria defined a priori. Gene

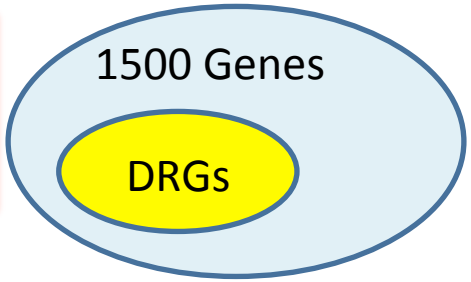
*The complete list of affiliations appears at the end of the article. Correspondence should be addressed to J.W.J. (jacobsonj@ctep.nci.nih.gov) or D.G.B. (dgbeer@umich.edu).

Received 8 January; accepted 12 June; published online 20 July 2008; doi:10.1038/nm.1790

Survival Predictions for Lung Cancer Patients with Gene Sets Identified by SSM analysis for SAEC data

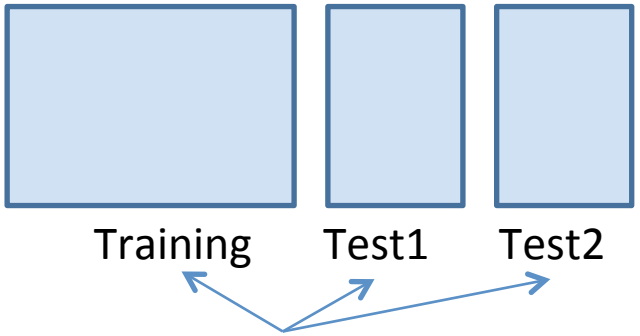
1. DRGs Identified from SAEC data by SSM





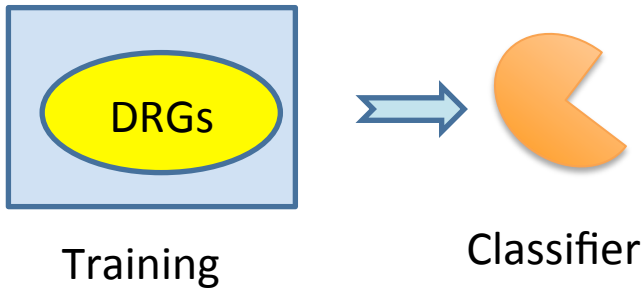
Q: Can Expressions of DRGs in Lung Cancer Patients Predict Prognosis? (Esp. for Stage I)

2. Lung Cancer Patients Data
(Shedden *et al.* 2008)

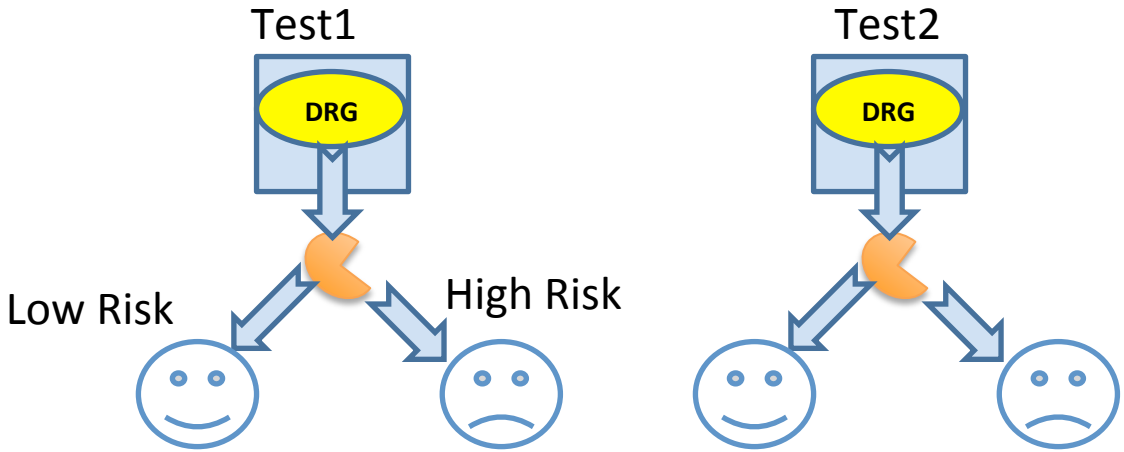


Survival Time & Gene Expression

3. Classifier Construction with Expressions of the DRGs of Patients in Training Sample



4. Stratifications of Test Patients by Predicted Risk Scores from the Classifier



DRG: Differentially Regulated Gene

Classifier: Risk Score Function

- Partial Cox Regression Model
  - Hazard Function: $\lambda(t, X_i) = \lambda_0(t)\exp\left[f(X_i)\right]$
  - Risk Score Function

$$f\left(X_i^{\text{Valid}}\right) = \sum_{j=1}^{p} \beta_j^{*\text{Train}} \left(X_{ij}^{\text{Valid}} - \bar{x}_j^{\text{Train}}\right)$$

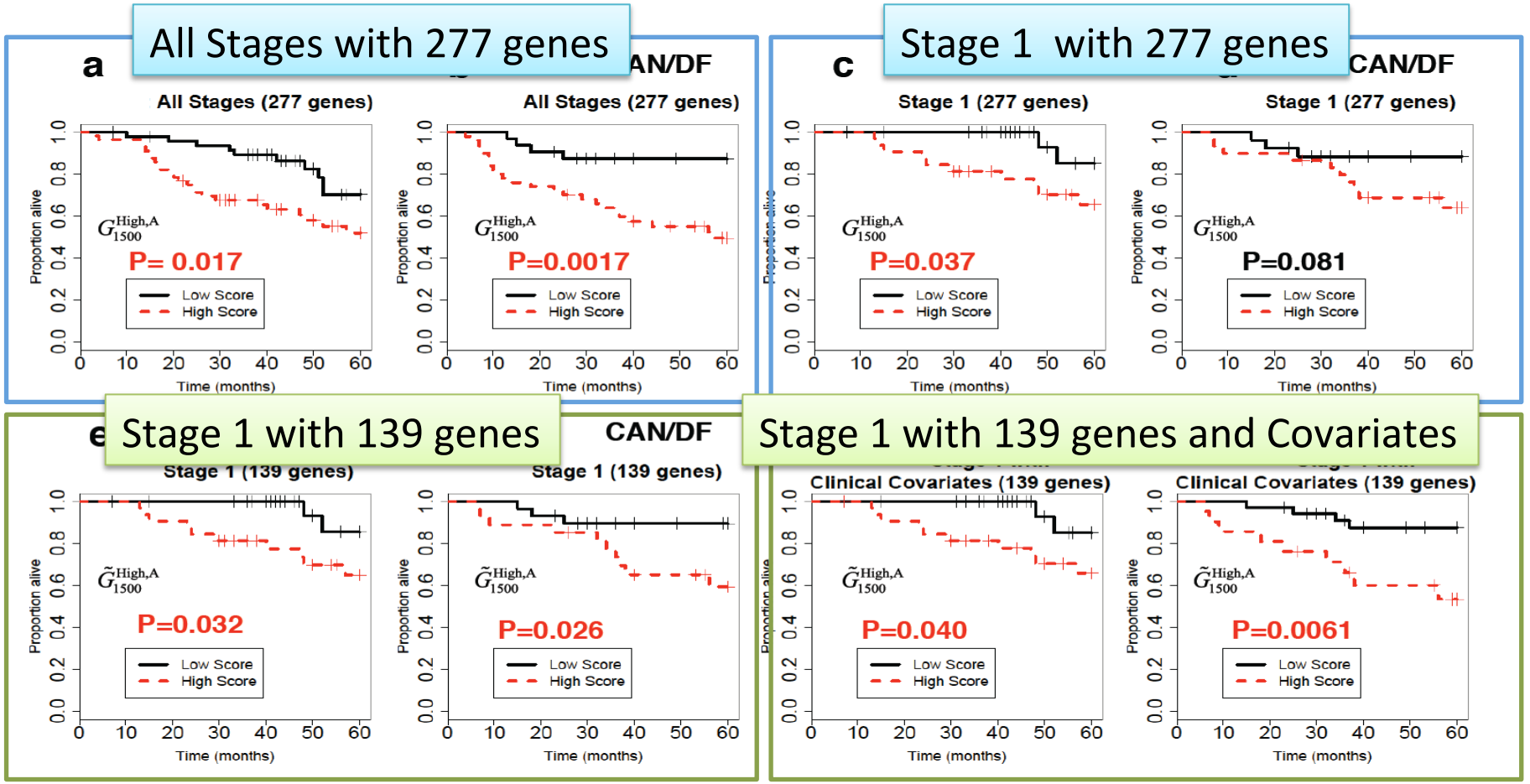
where $X_i^{\text{Valid}} = \left[X_{i1}^{\text{Valid}}, \text{L}\ , X_{ip}^{\text{Valid}}\right]^T$ The p gene expressions for the *i*th patient in a validation set

The *i*th patient is classified into a high risk group when $f(X_i^{\text{Valid}}) > 0$

or a low risk group when $f(X_i^{\text{Valid}}) \leq 0$

Li and Gui, 2004

Survival Predictions for Lung Cancer Patients with Gene Sets Identified by SSM analysis for SAEC data



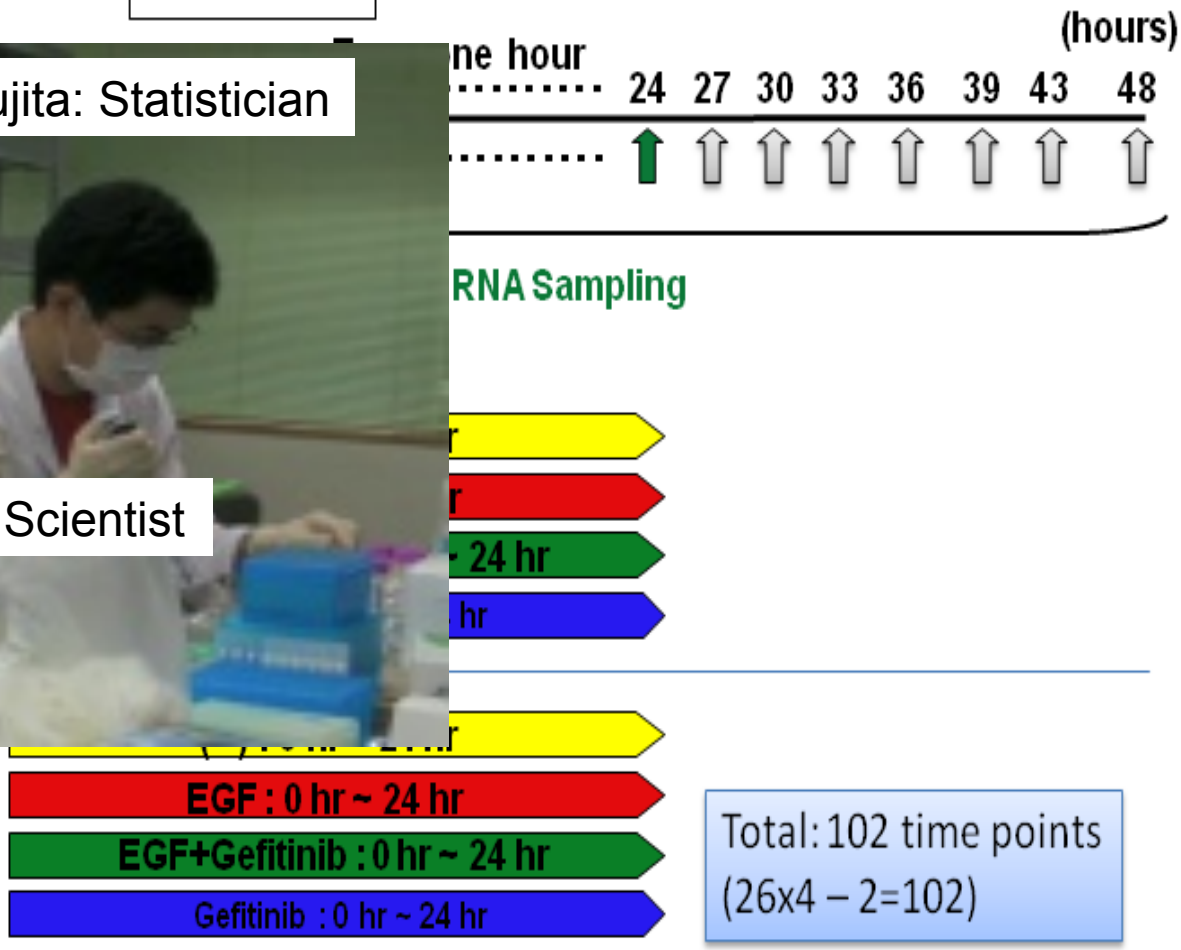
4. Gene Networks of Lung Cancer

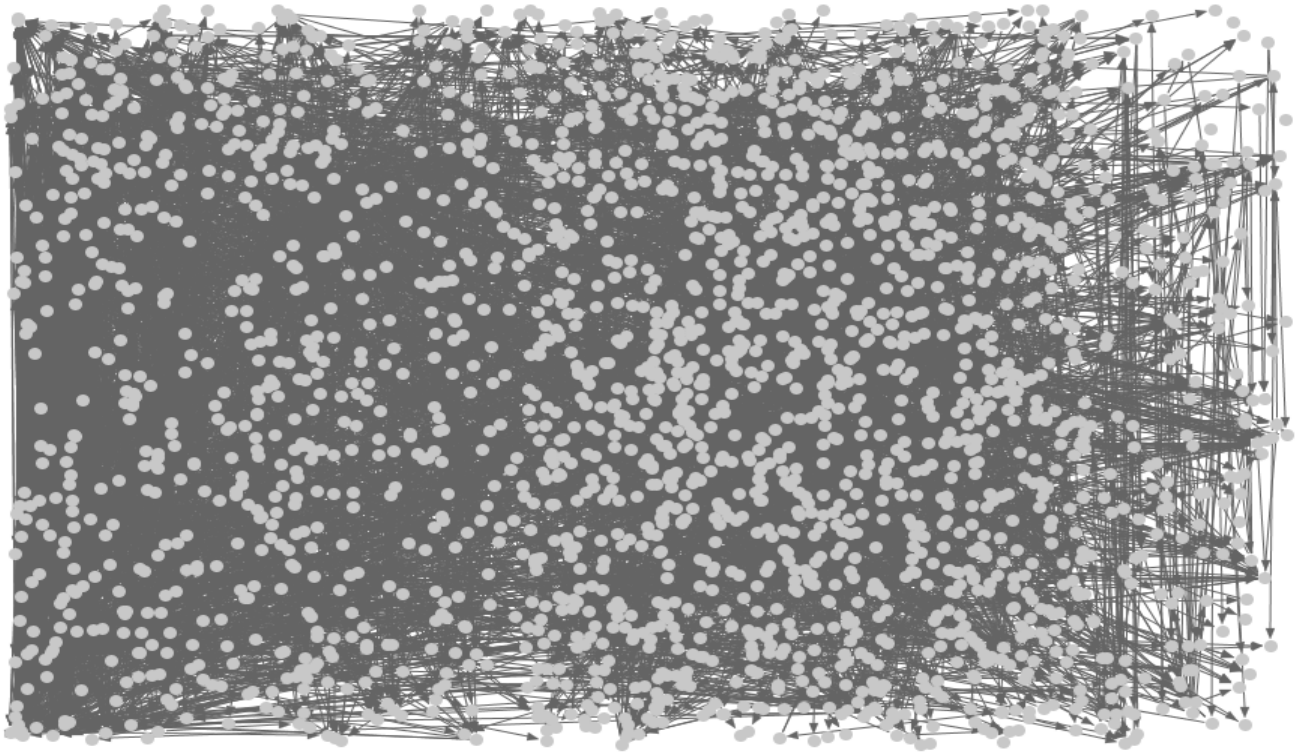
Gene Networks of
Gefitinib Sensitive PC9 and
Gefitinib Resistant PC9GR2

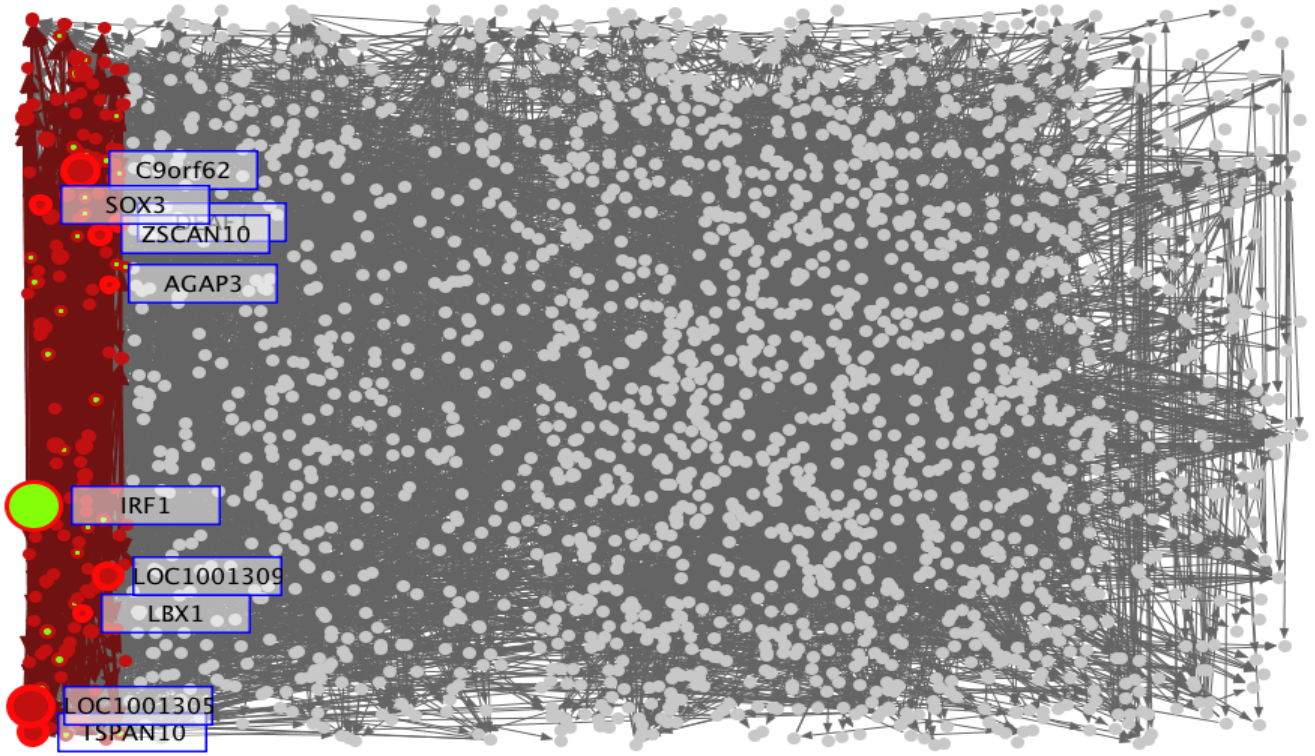
PC9GR2

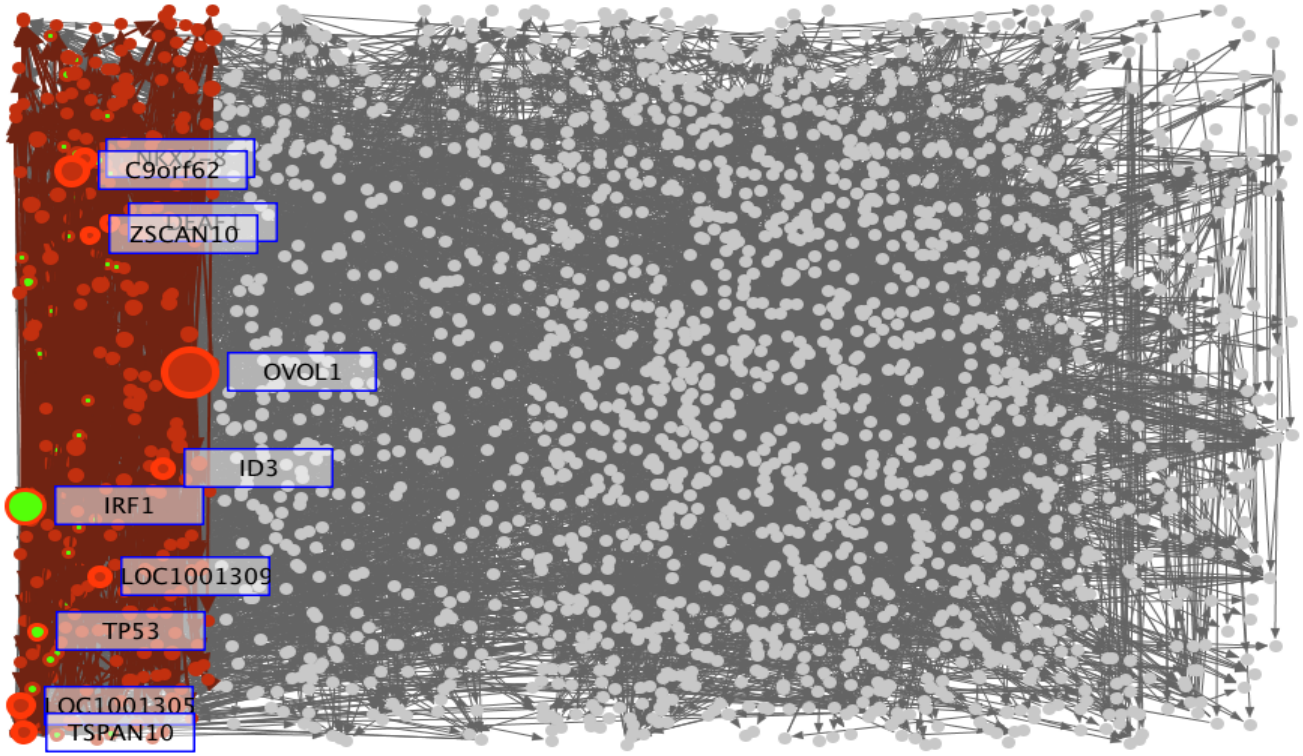
(Lung Cancer Cell Line with Gefitinib resistance)

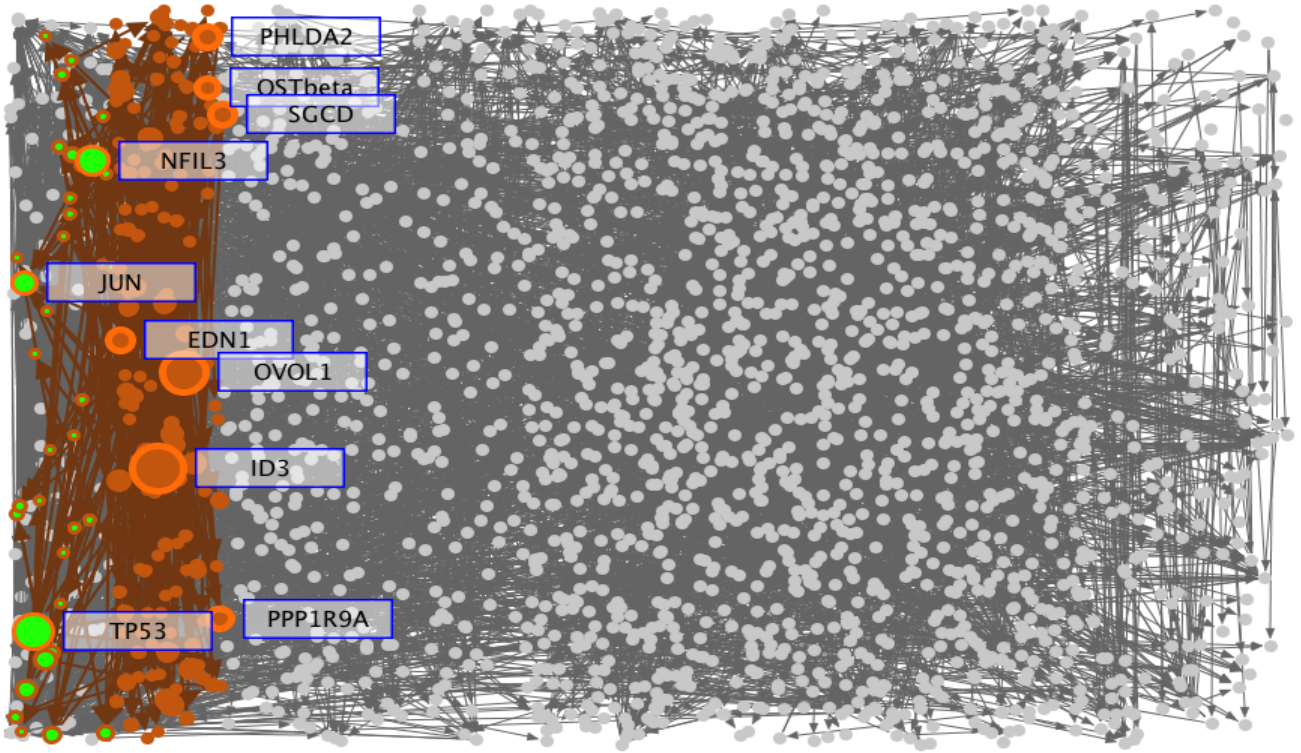
PC9GRM2

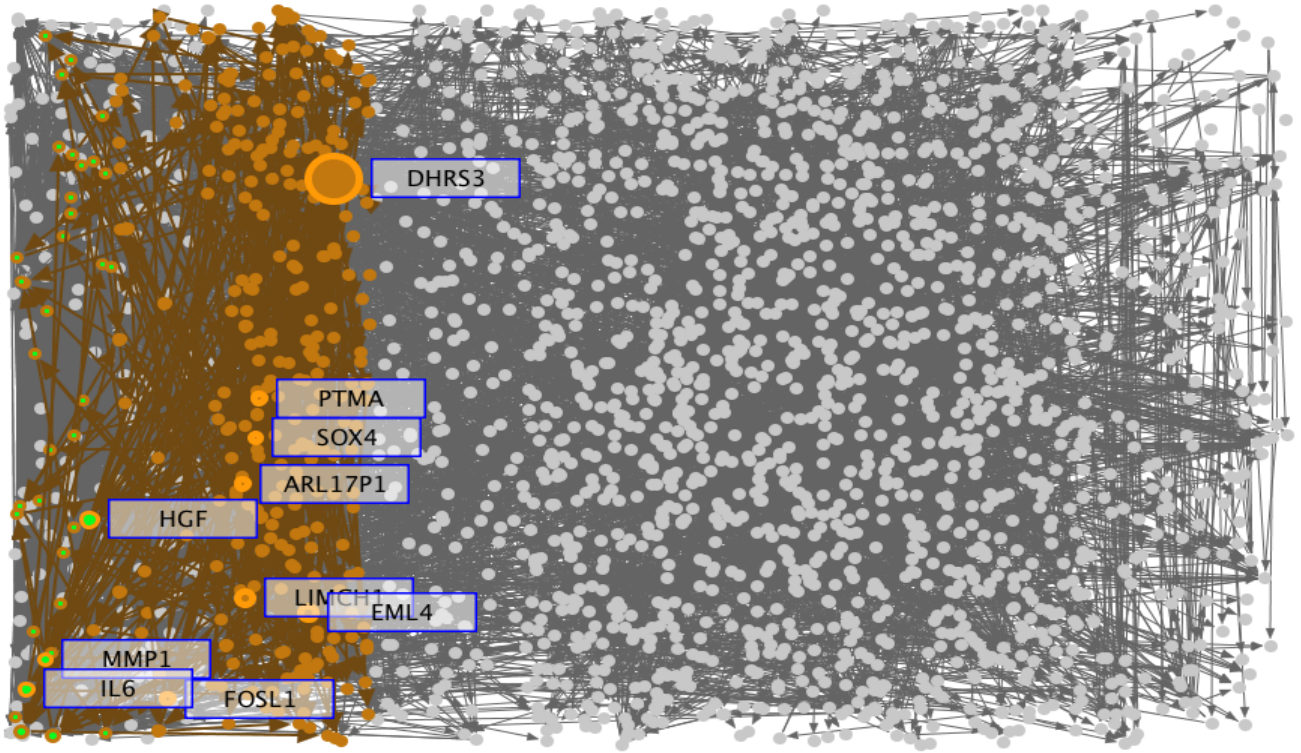


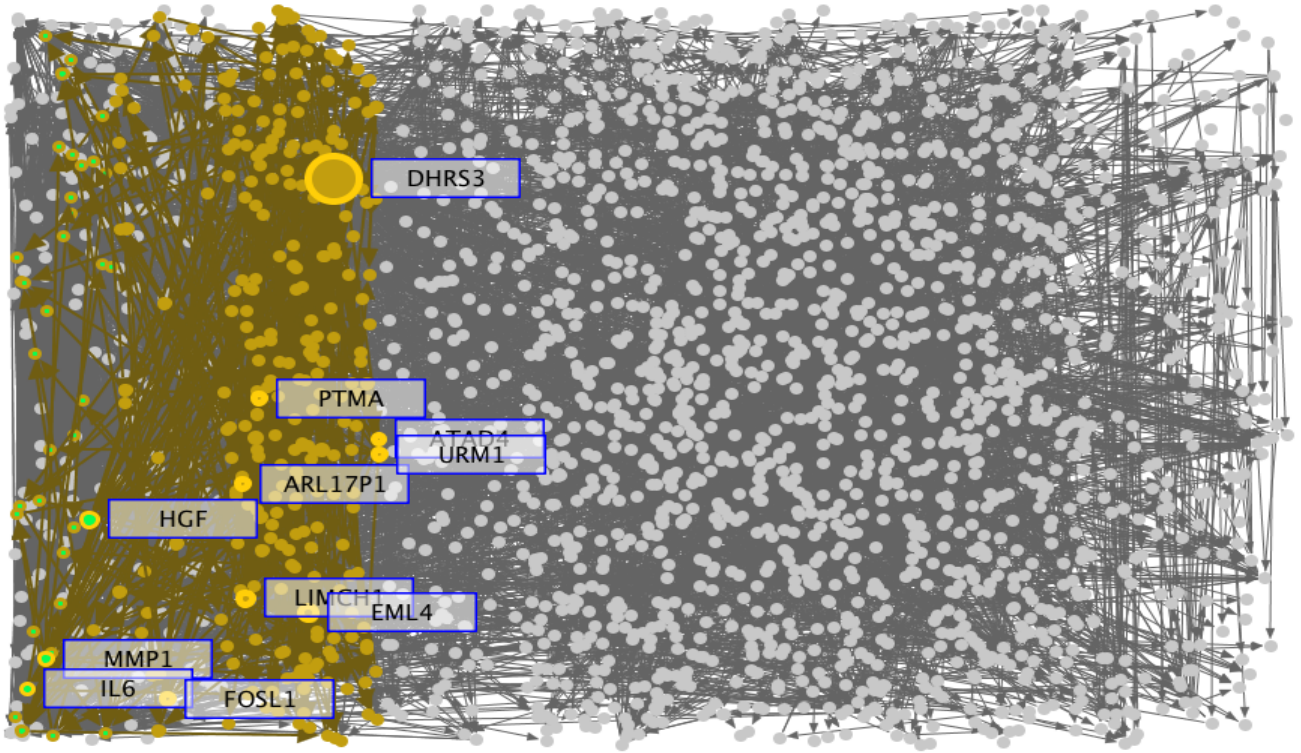


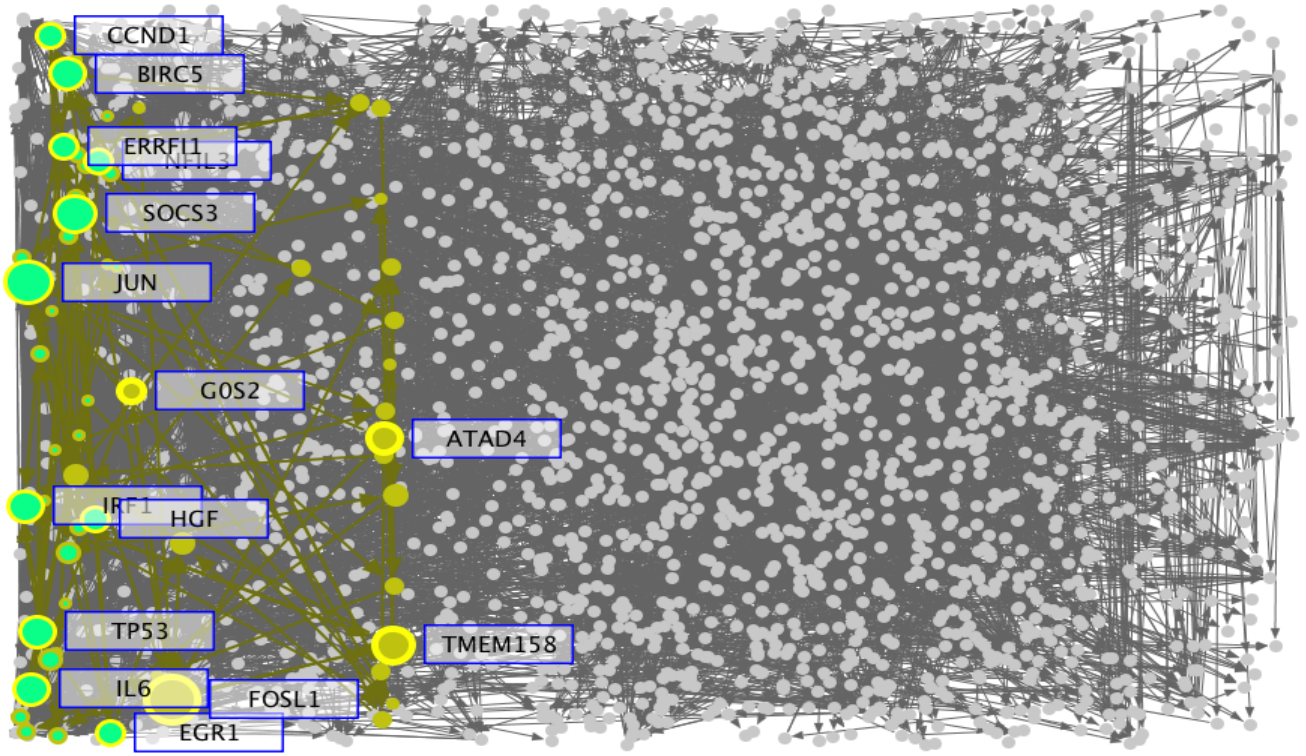

C9orf62
SOX3
ZSCAN10
AGAP3
IRF1
LOC1001309
LBX1
LOC1001305
TSPAN10

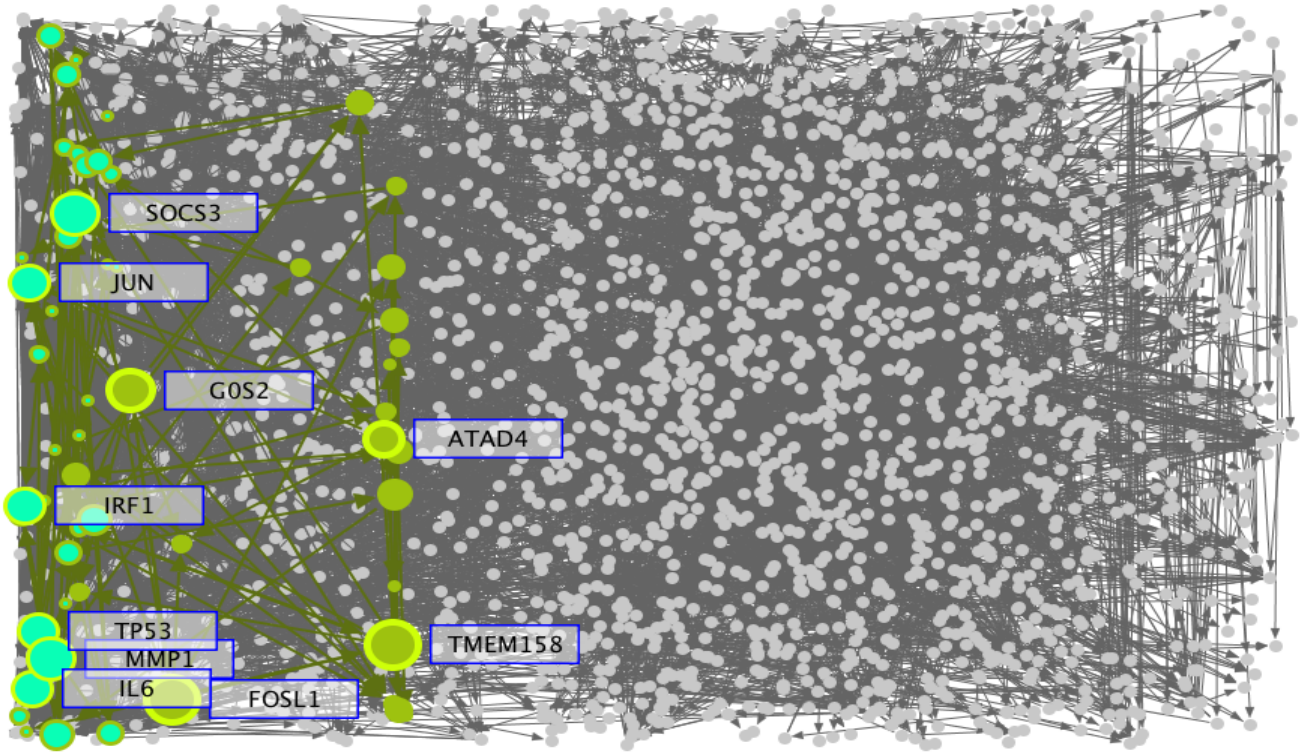

C9orf62
ZSCAN10
OVOL1
ID3
IRF1
LOC1001309
TP53
LOC1001305
TSPAN10

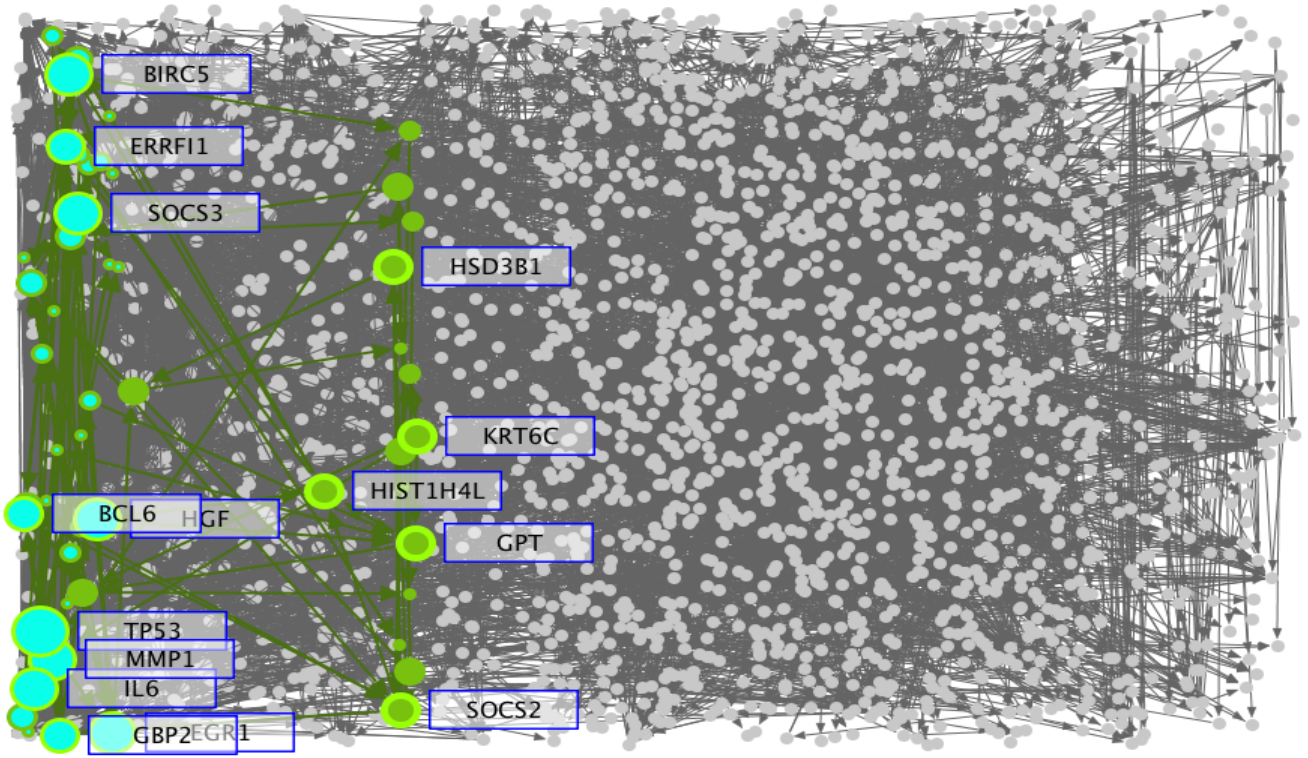

PHLDA2
OSTbeta
SGCD
NFIL3
JUN
EDN1
OVOL1
ID3
TP53
PPP1R9A

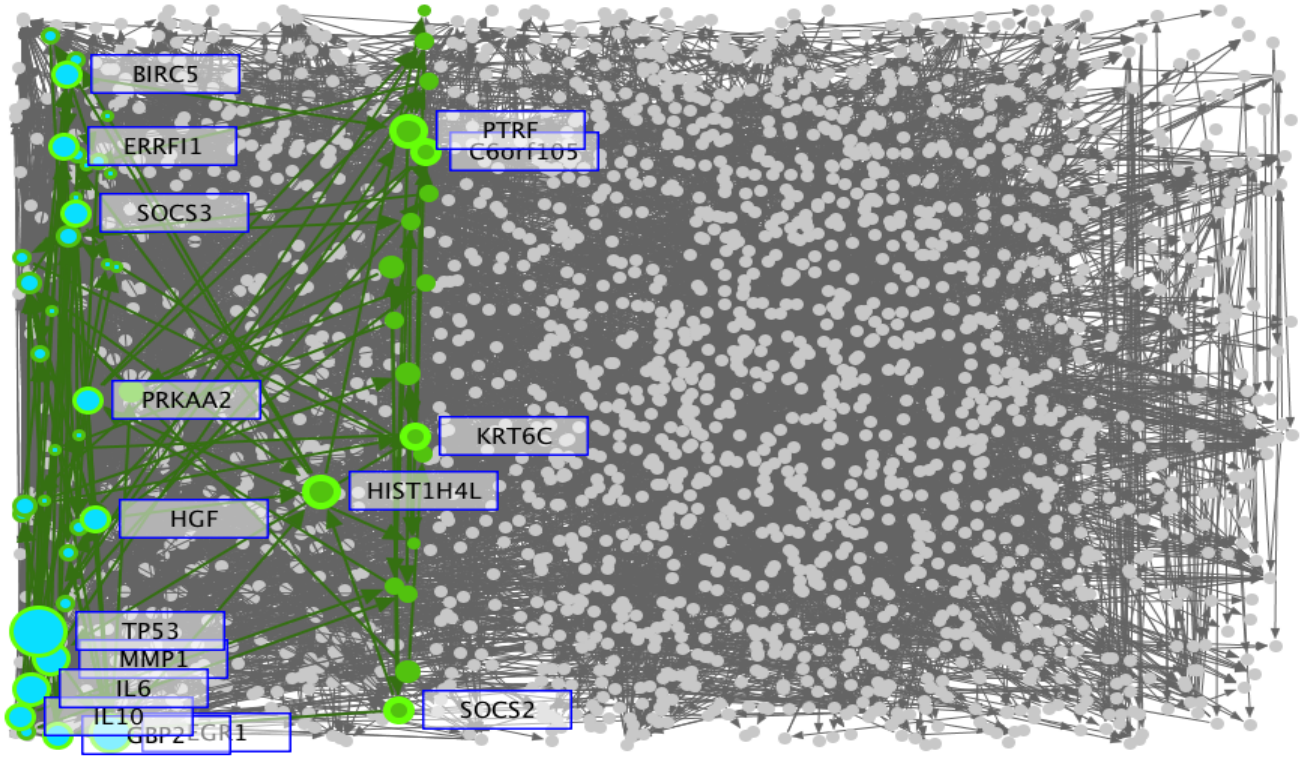

DHRS3
PTMA
SOX4
ARL17P1
HGF
LIMCH1
EML4
MMP1
IL6
FOSL1

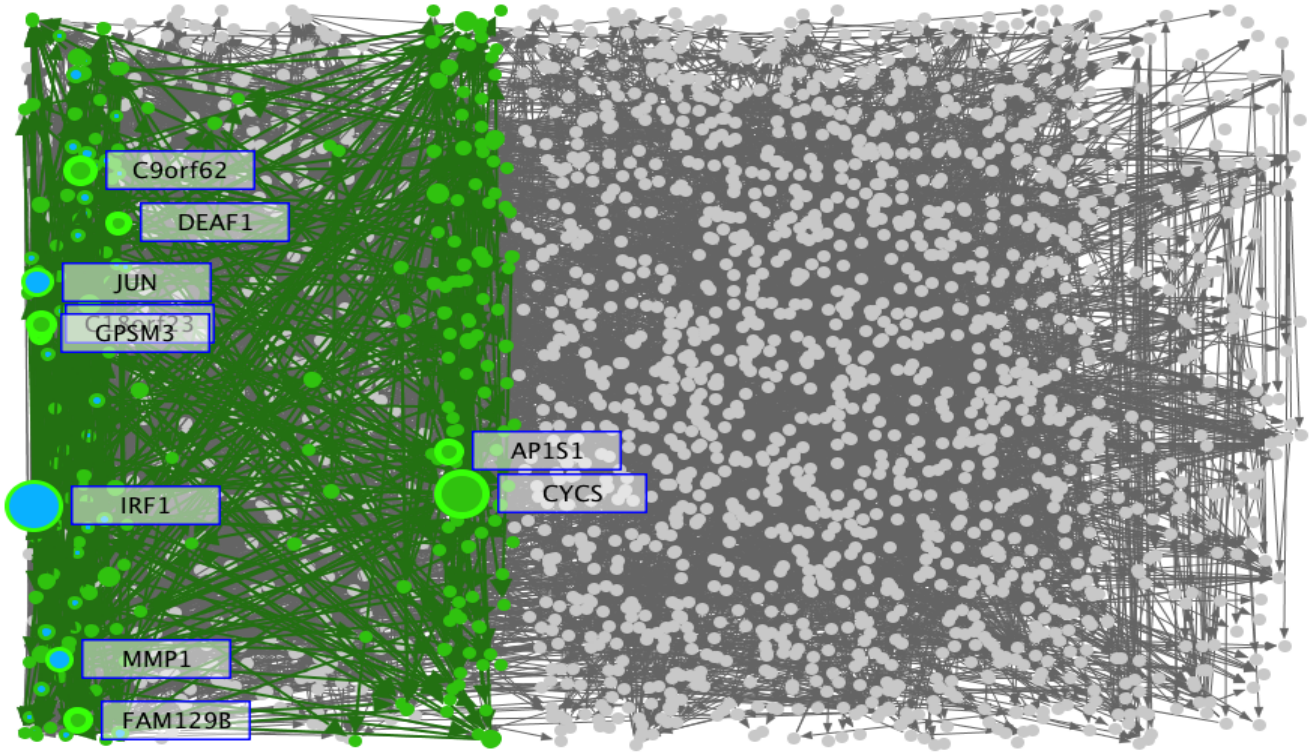

DHRS3
PTMA
ATAD4
URM1
ARL17P1
HGF
LIMCH1
EML4
MMP1
IL6
FOSL1

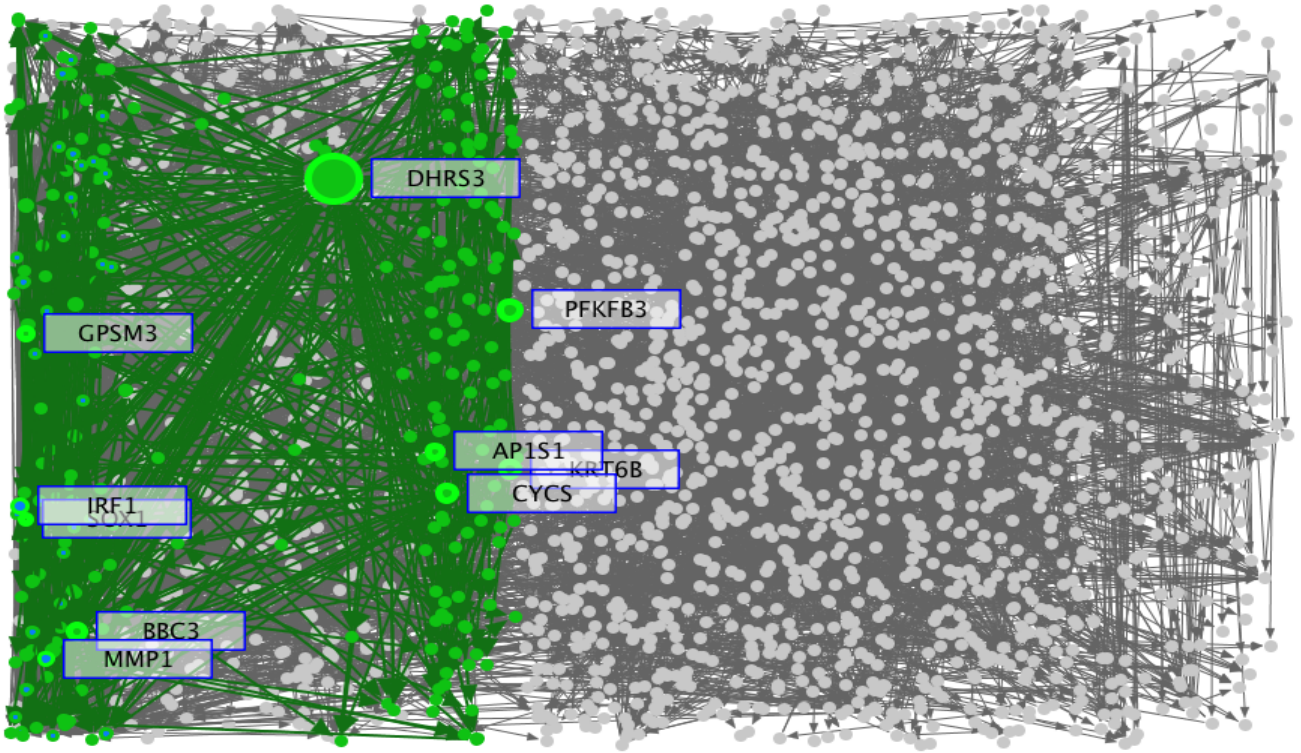

CCND1
BIRC5
ERRFI1
SOCS3
JUN
G0S2
ATAD4
HGF
TP53
TMEM158
IL6
FOSL1
EGR1

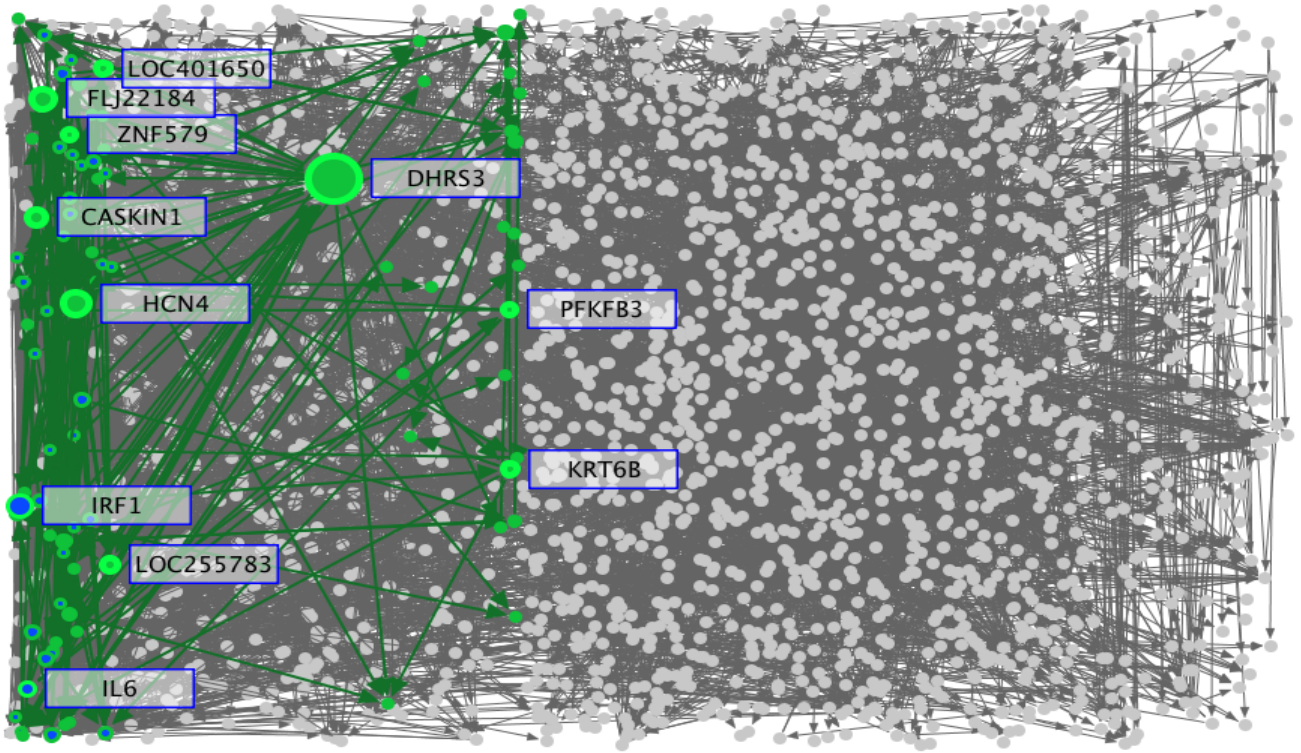

SOCS3
JUN
G0S2
ATAD4
IRF1
TP53
MMP1
IL6
FOSL1
TMEM158

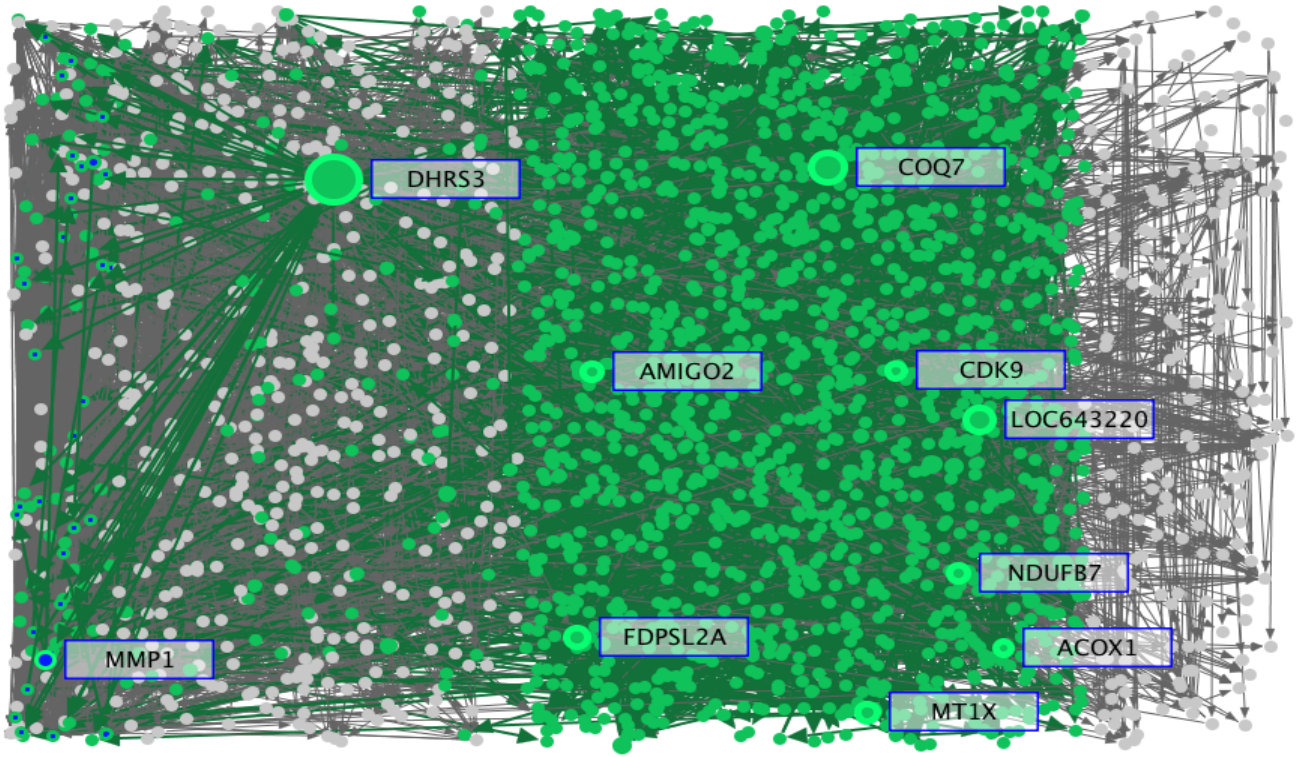

BIRC5
ERRFI1
SOCS3
HSD3B1
KRT6C
HIST1H4L
BCL6
HGF
GPT
TP53
MMP1
IL6
GBP2
EGR1
SOCS2

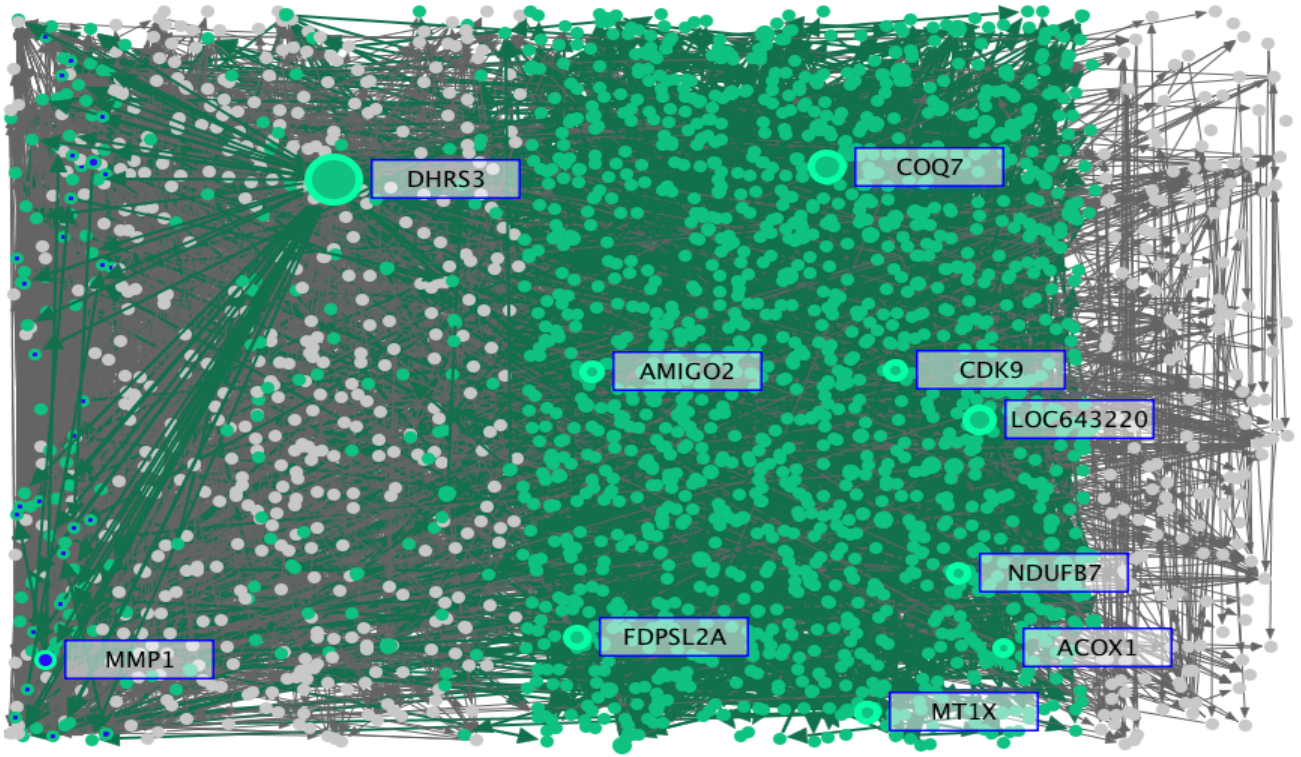

BIRC5
ERRFI1
SOCS3
PRKAA2
HGF
TP53
MMP1
IL6
IL10
PTRF
KRT6C
HIST1H4L
SOCS2

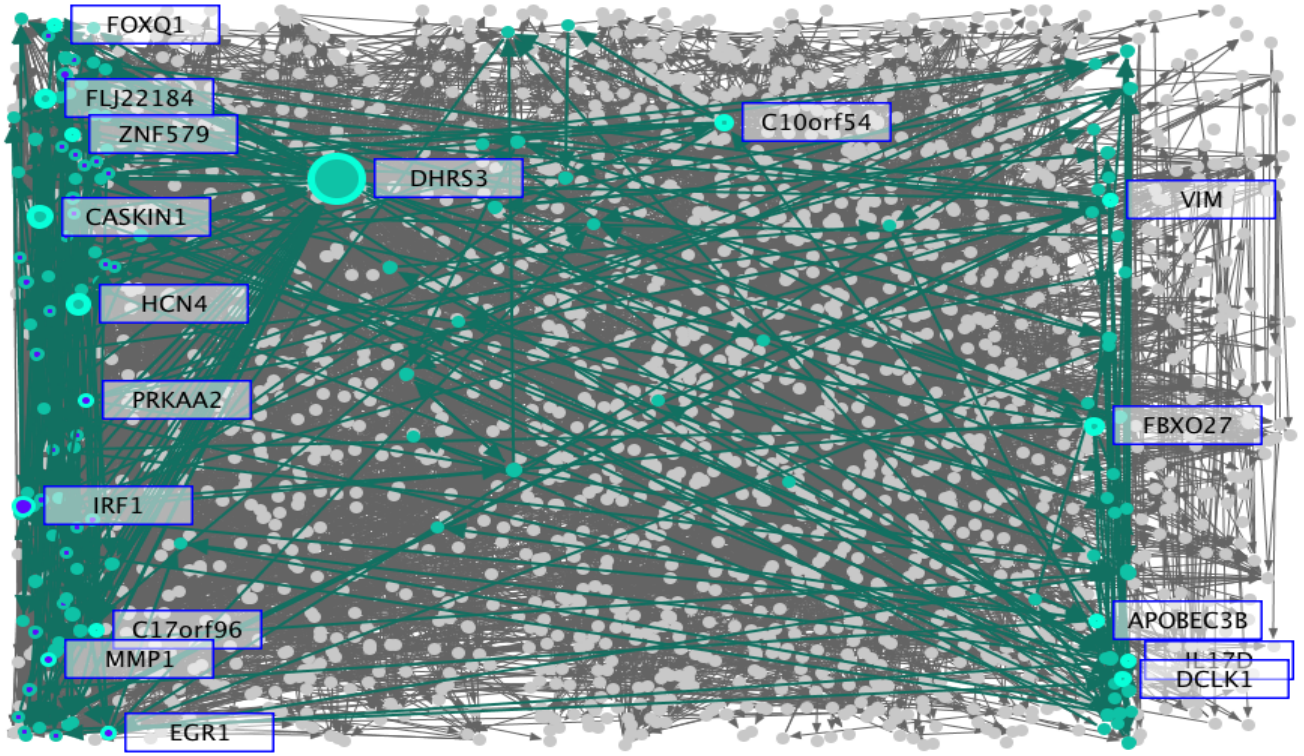

C9orf62
DEAF1
JUN
GPSM3
AP1S1
CYCS
IRF1
MMP1
FAM129B

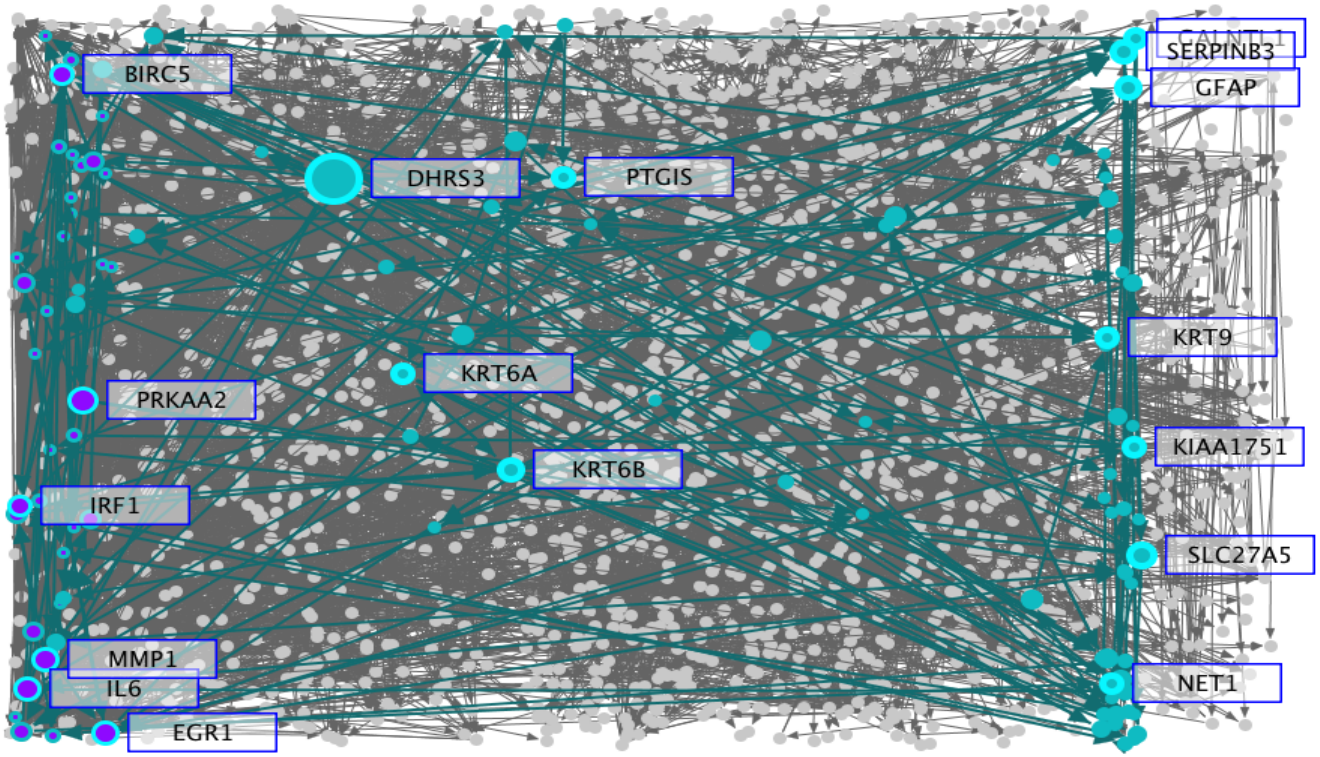

DHRS3
PFKFB3
GPSM3
AP1S1
KRT6B
CYCS
IRF1
SOX1
BBC3
MMP1

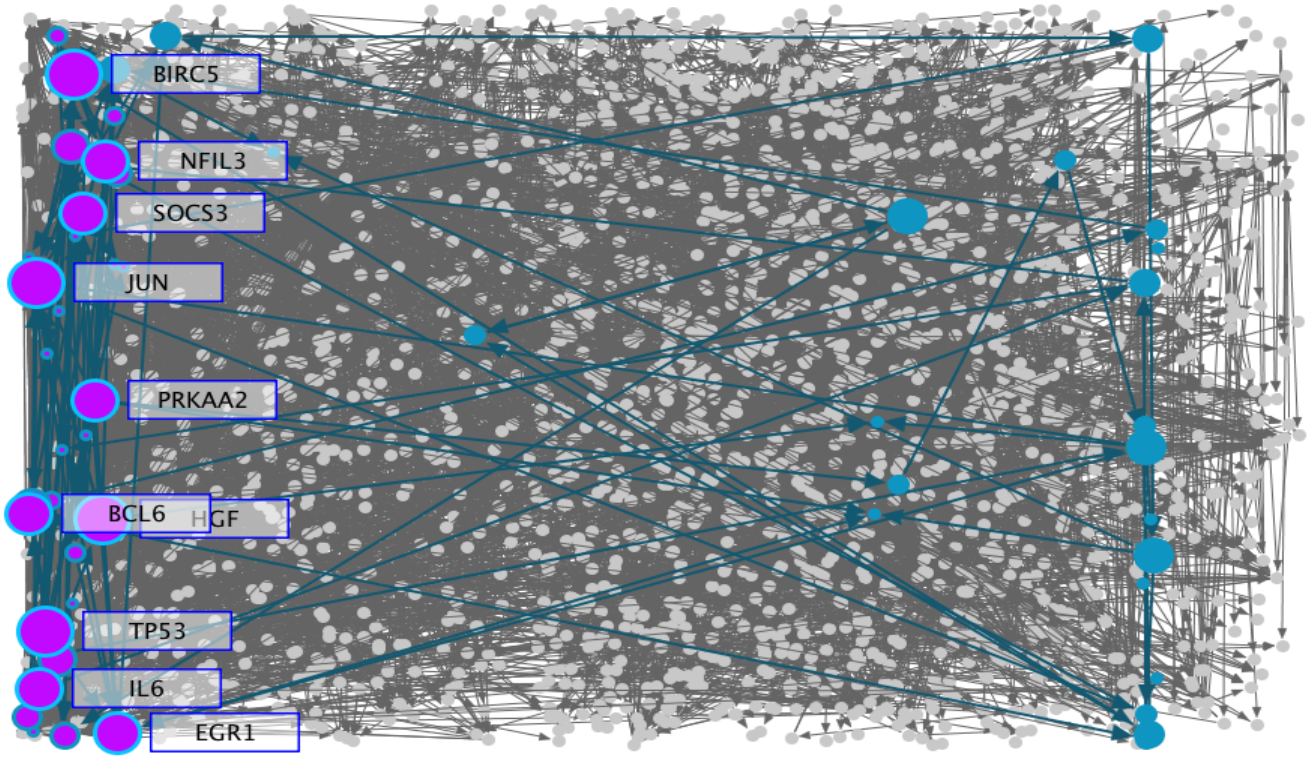

LOC401650
FLJ22184
ZNF579
DHRS3
CASKIN1
HCN4
PFKFB3
KRT6B
IRF1
LOC255783
IL6

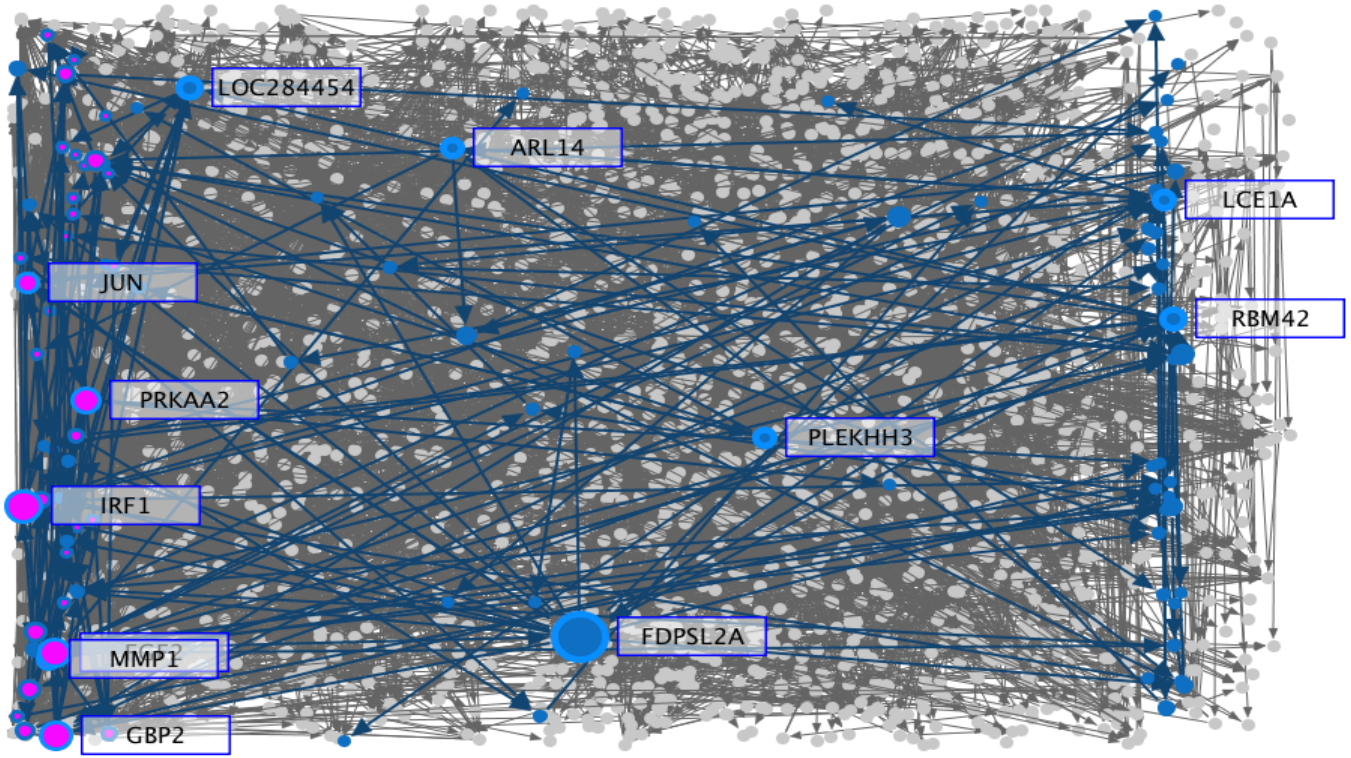

DHRS3
COQ7
AMIGO2
CDK9
LOC643220
NDUFB7
FDPSL2A
ACOX1
MMP1
MT1X

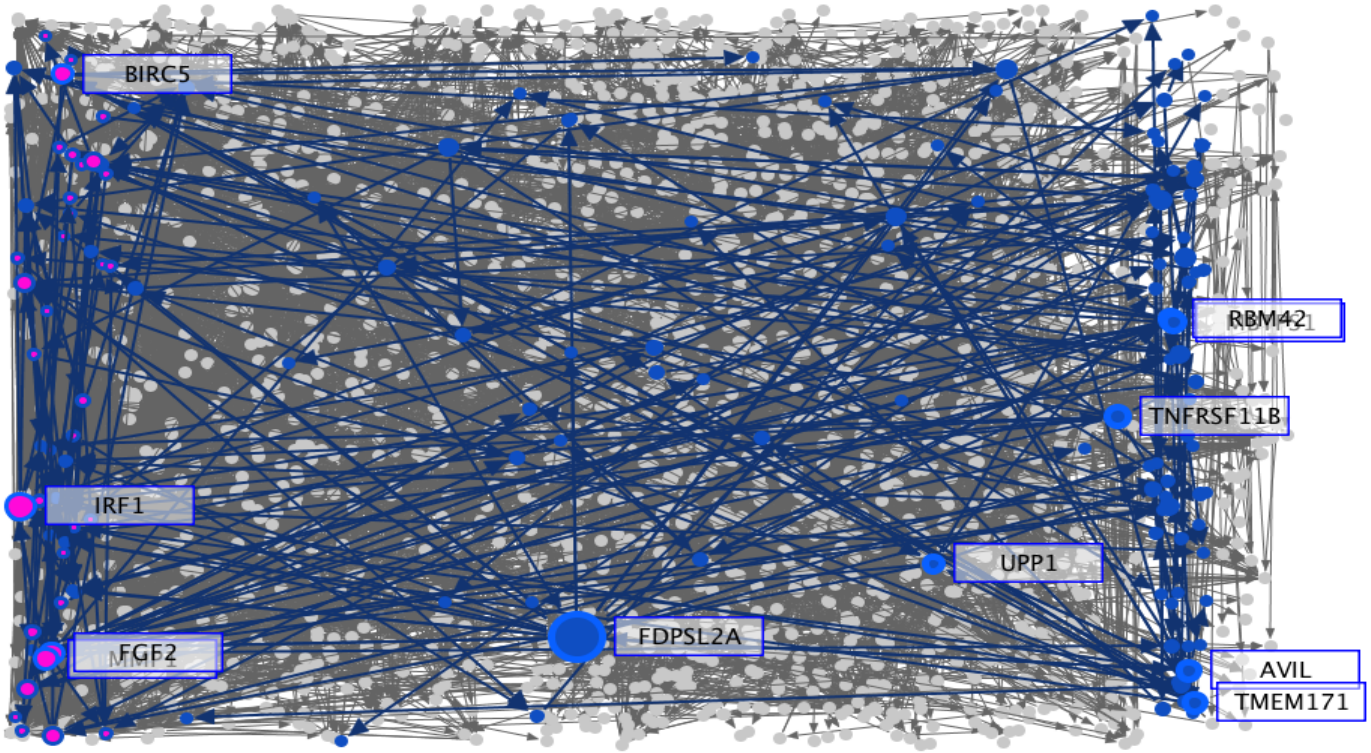

DHRS3
COQ7
AMIGO2
CDK9
LOC643220
NDUFB7
FDPSL2A
ACOX1
MMP1
MT1X

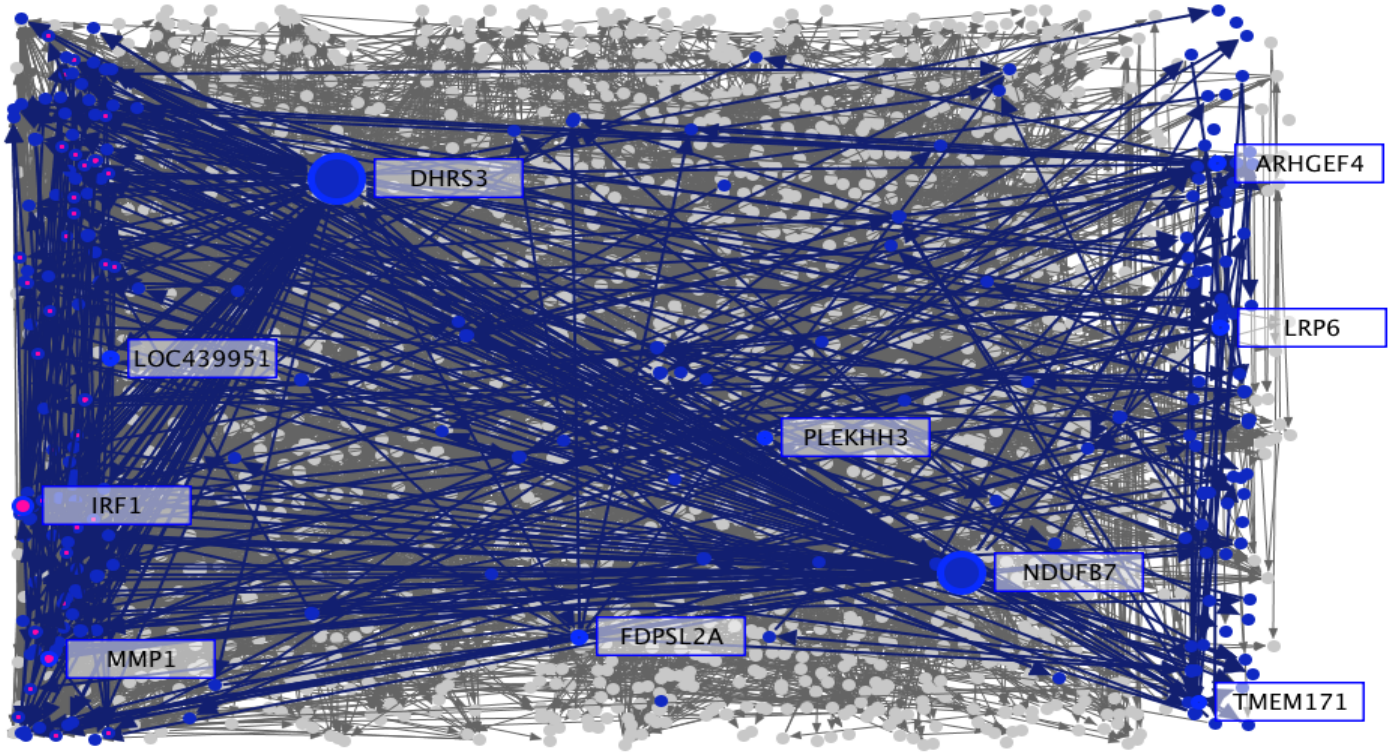

FOXQ1
FLJ22184
ZNF579
CASKIN1
HCN4
PRKAA2
IRF1
C17orf96
MMP1
EGR1
DHRS3
C10orf54
VIM
FBXO27
APOBEC3B
IL17D
DCLK1

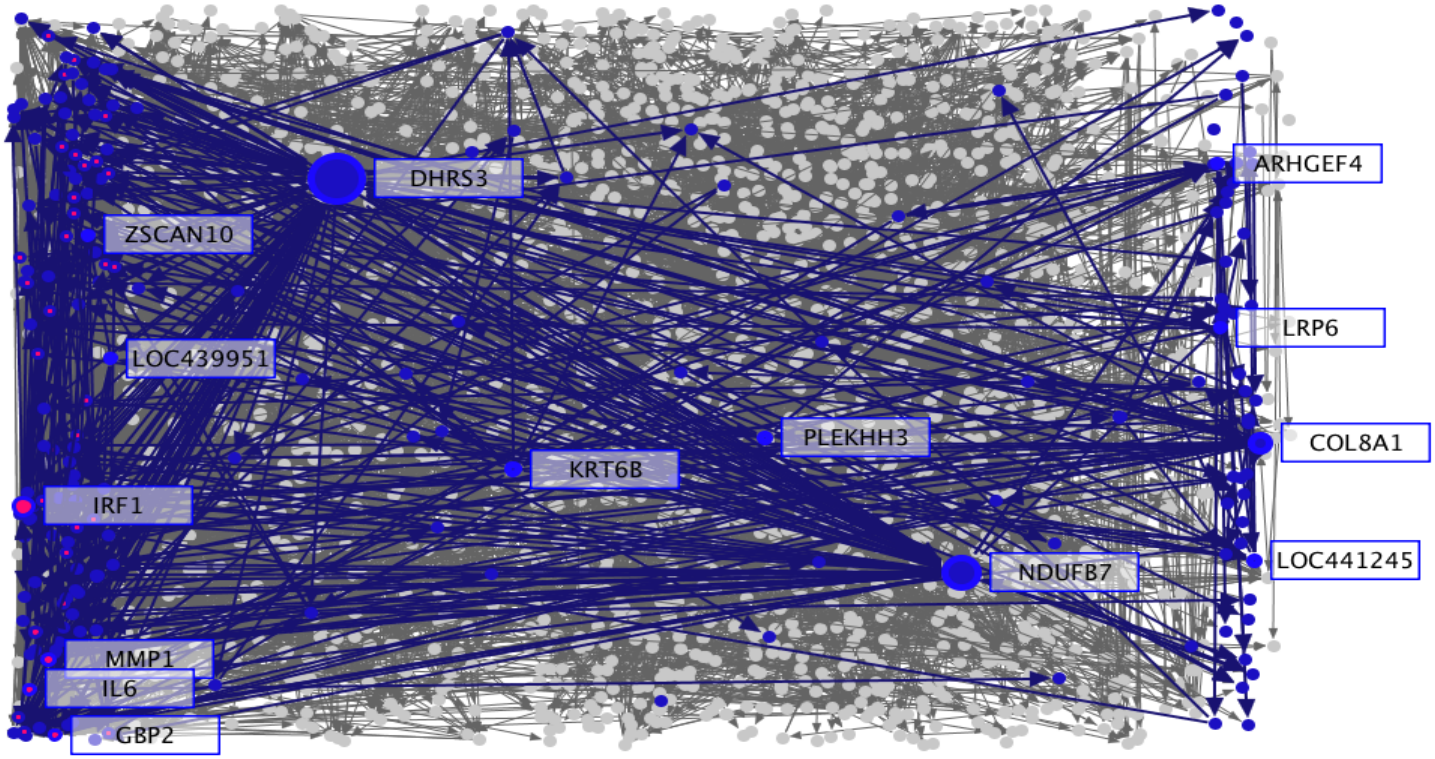

BIRC5
DHRS3
PTGIS
KRT6A
PRKAA2
KRT6B
IRF1
MMP1
IL6
EGR1
CALNTL1
SERPINB3
GFAP
KRT9
KIAA1751
SLC27A5
NET1

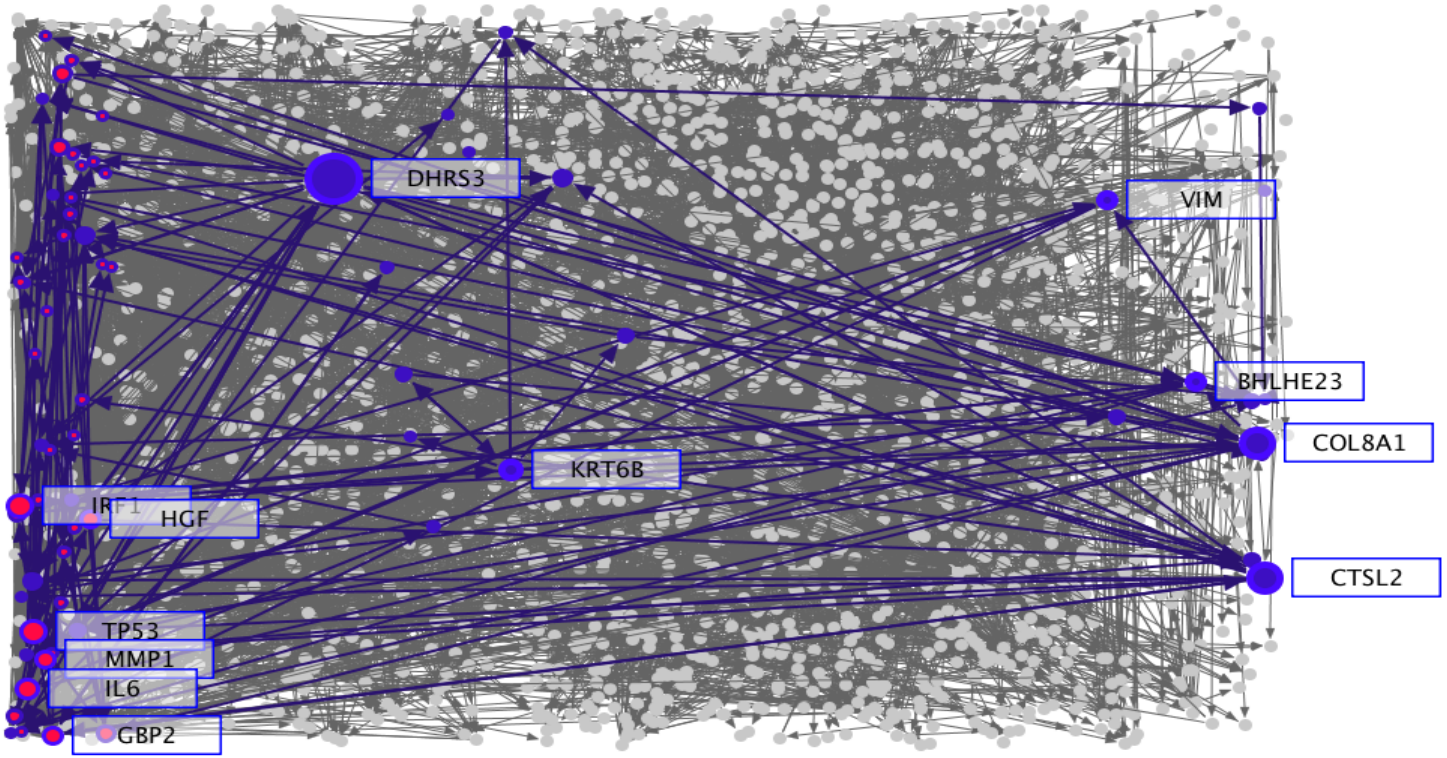

BIRC5
NFIL3
SOCS3
JUN
PRKAA2
BCL6
HGF
TP53
IL6
EGR1

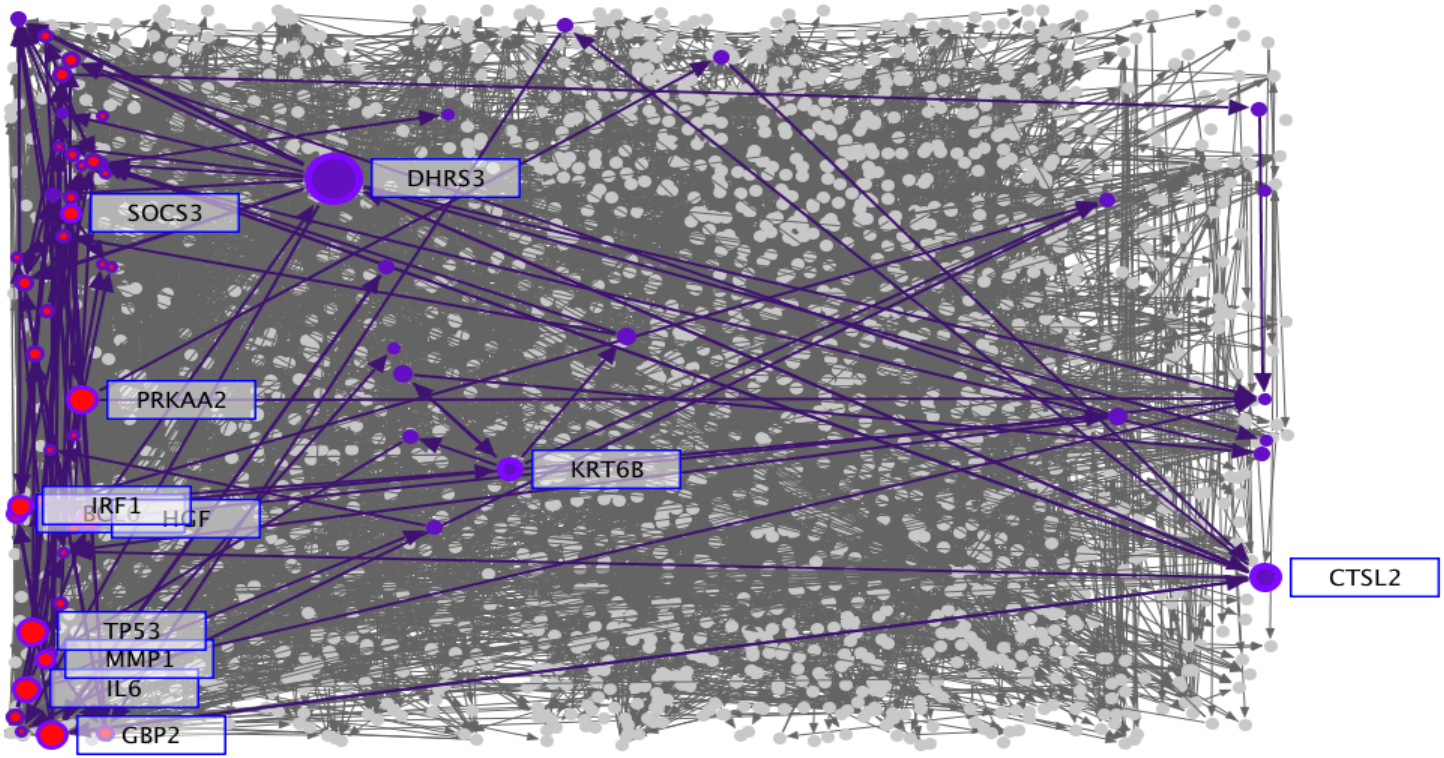

LOC284454
ARL14
LCE1A
JUN
RBM42
PRKAA2
PLEKHH3
IRF1
FDPSL2A
MMP1
GBP2

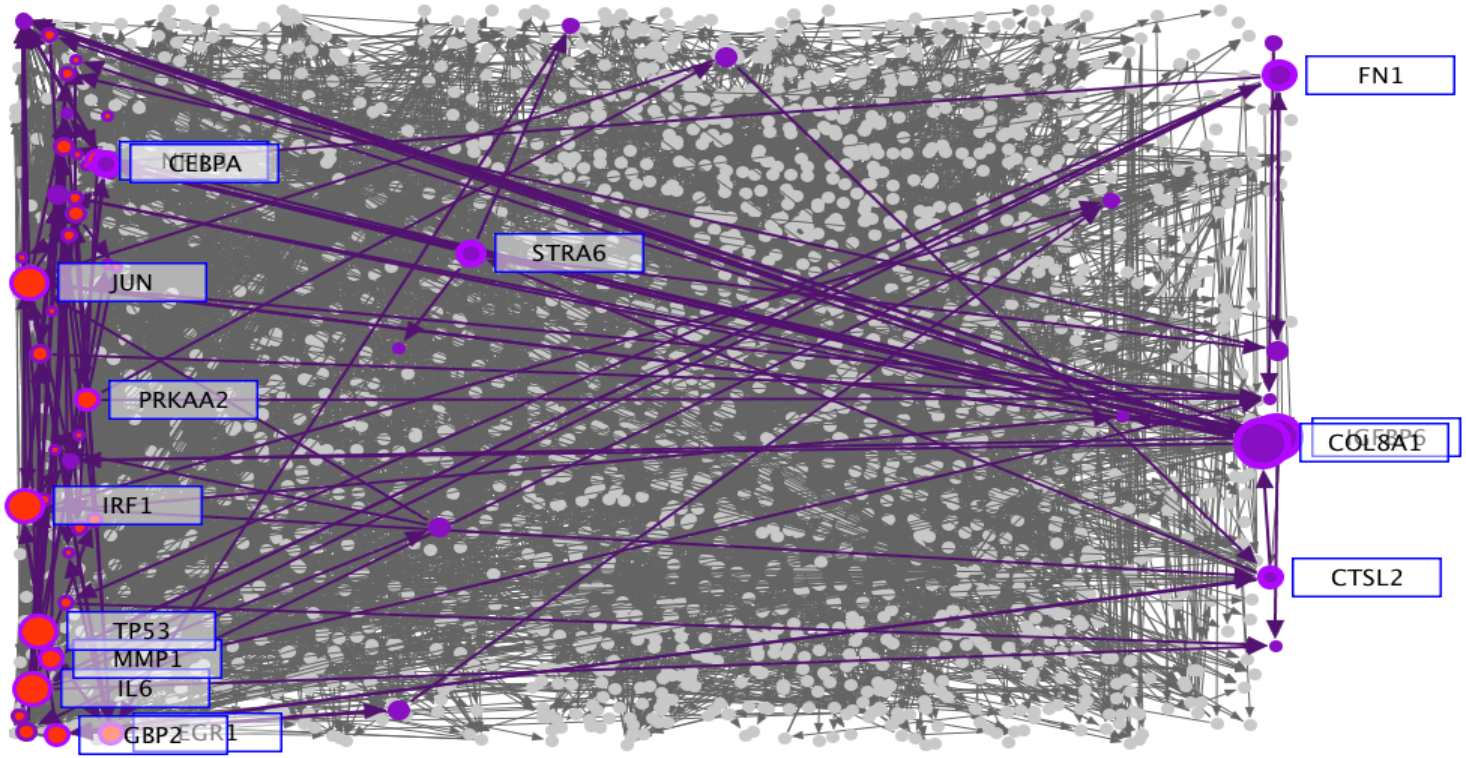

BIRC5
RBM42
TNFRSF11B
IRF1
UPP1
FDPSL2A
FGF2
AVIL
TMEM171

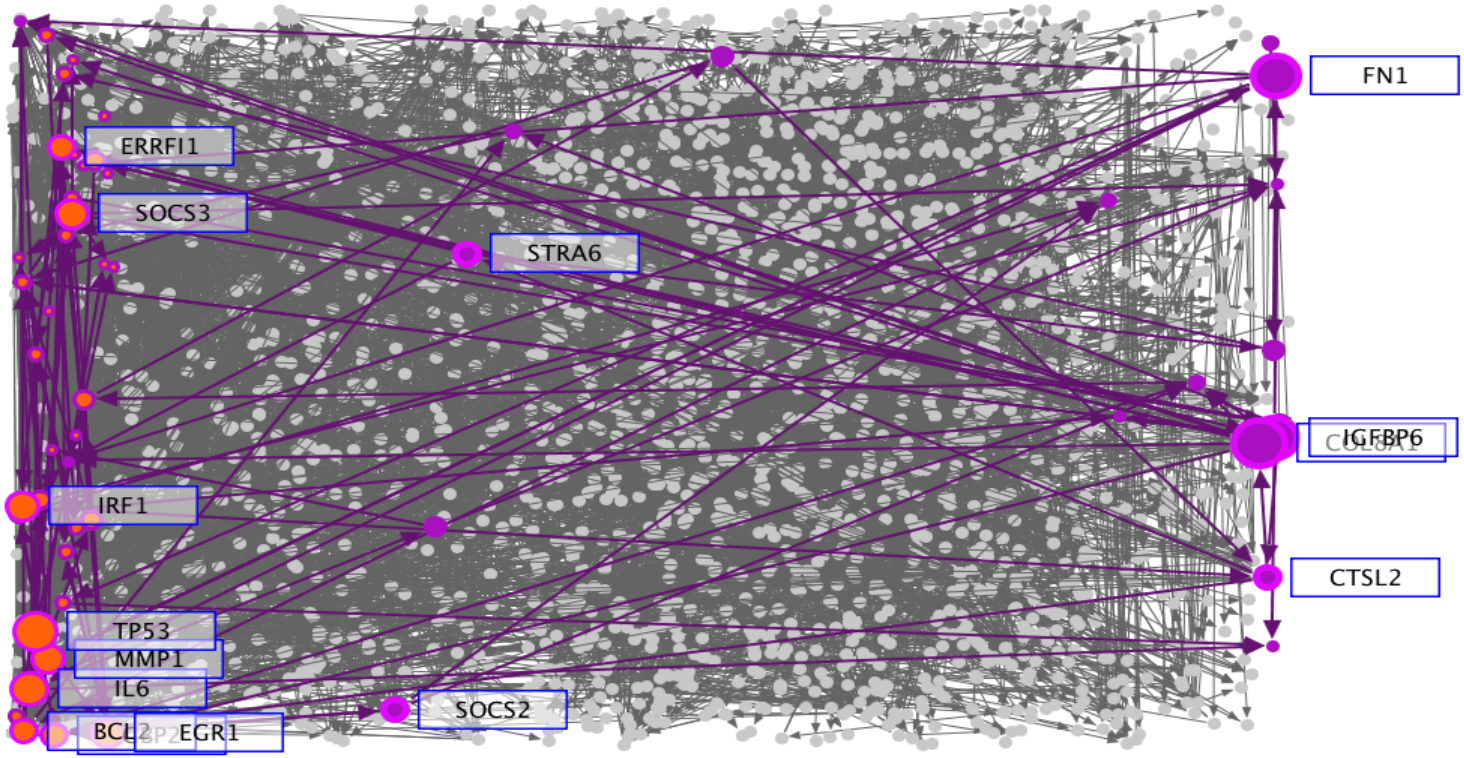

DHRS3
ARHGEF4
LRP6
LOC439951
PLEKHH3
IRF1
NDUFB7
FDPSL2A
MMP1
TMEM171

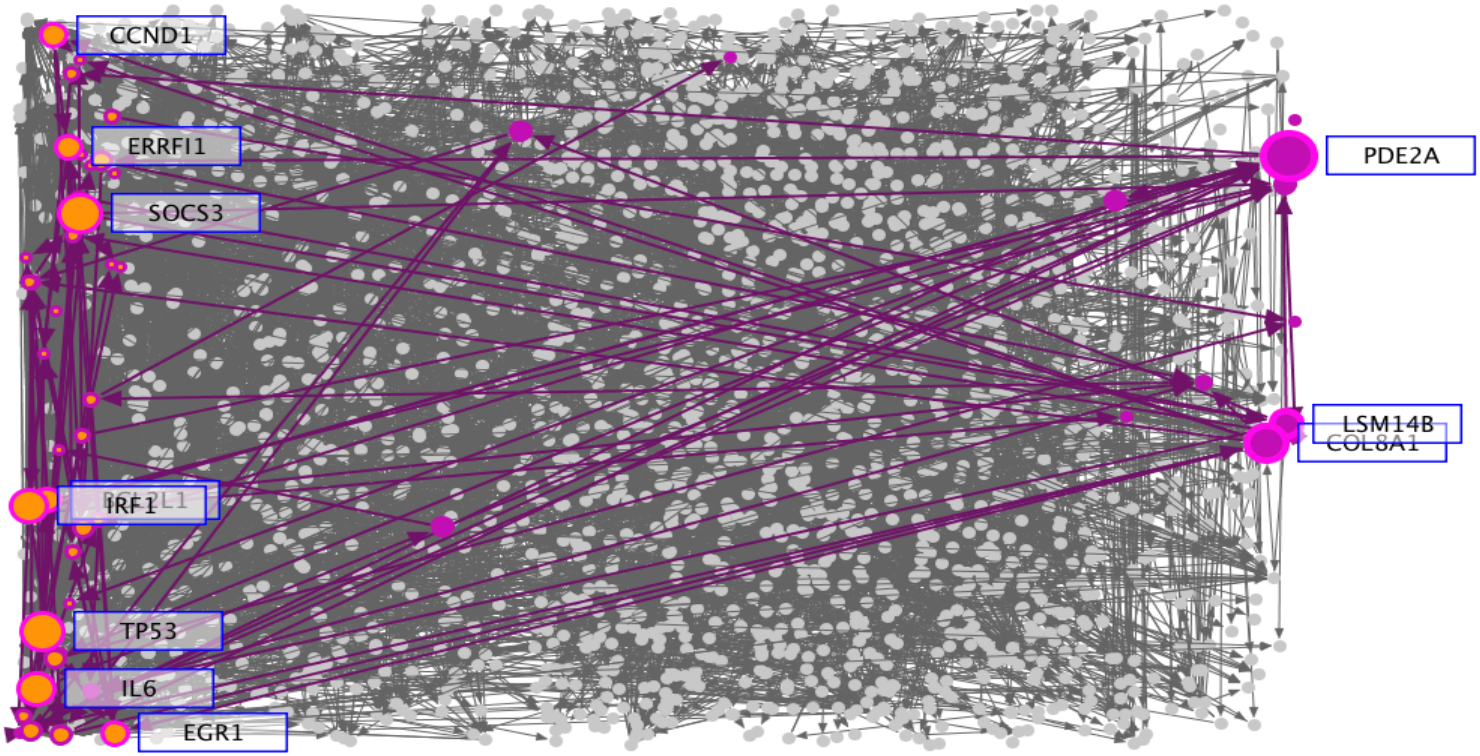

DHRS3
ARHGEF4
ZSCAN10
LRP6
LOC439951
PLEKHH3
COL8A1
KRT6B
IRF1
NDUFB7
LOC441245
MMP1
IL6
GBP2


DHRS3
VIM
BHLHE23
COL8A1
KRT6B
IRF1
HGF
CTSL2
TP53
MMP1
IL6
GBP2

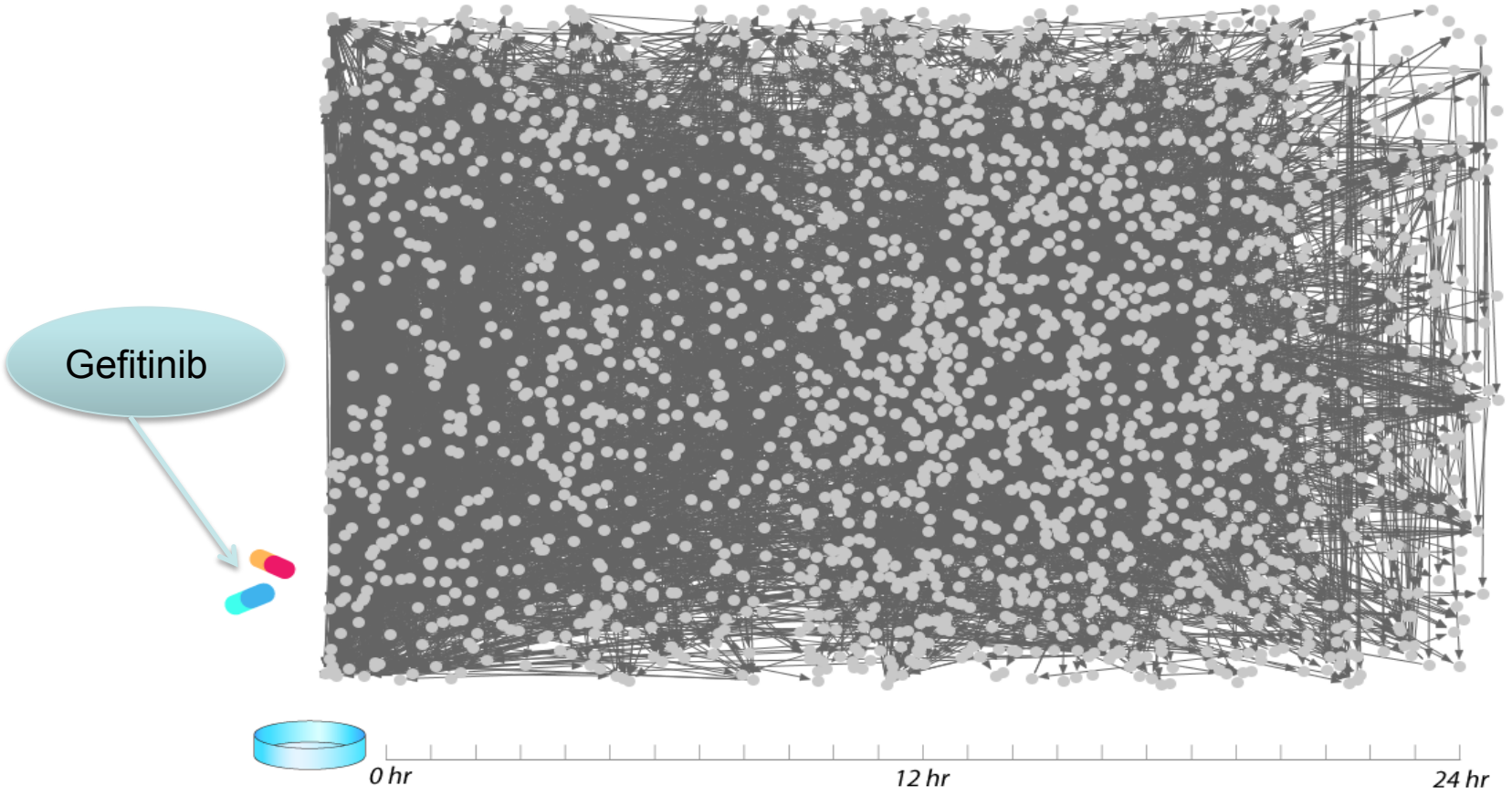

DHRS3
SOCS3
PRKAA2
KRT6B
IRF1
HGF
CTSL2
TP53
MMP1
IL6
GBP2

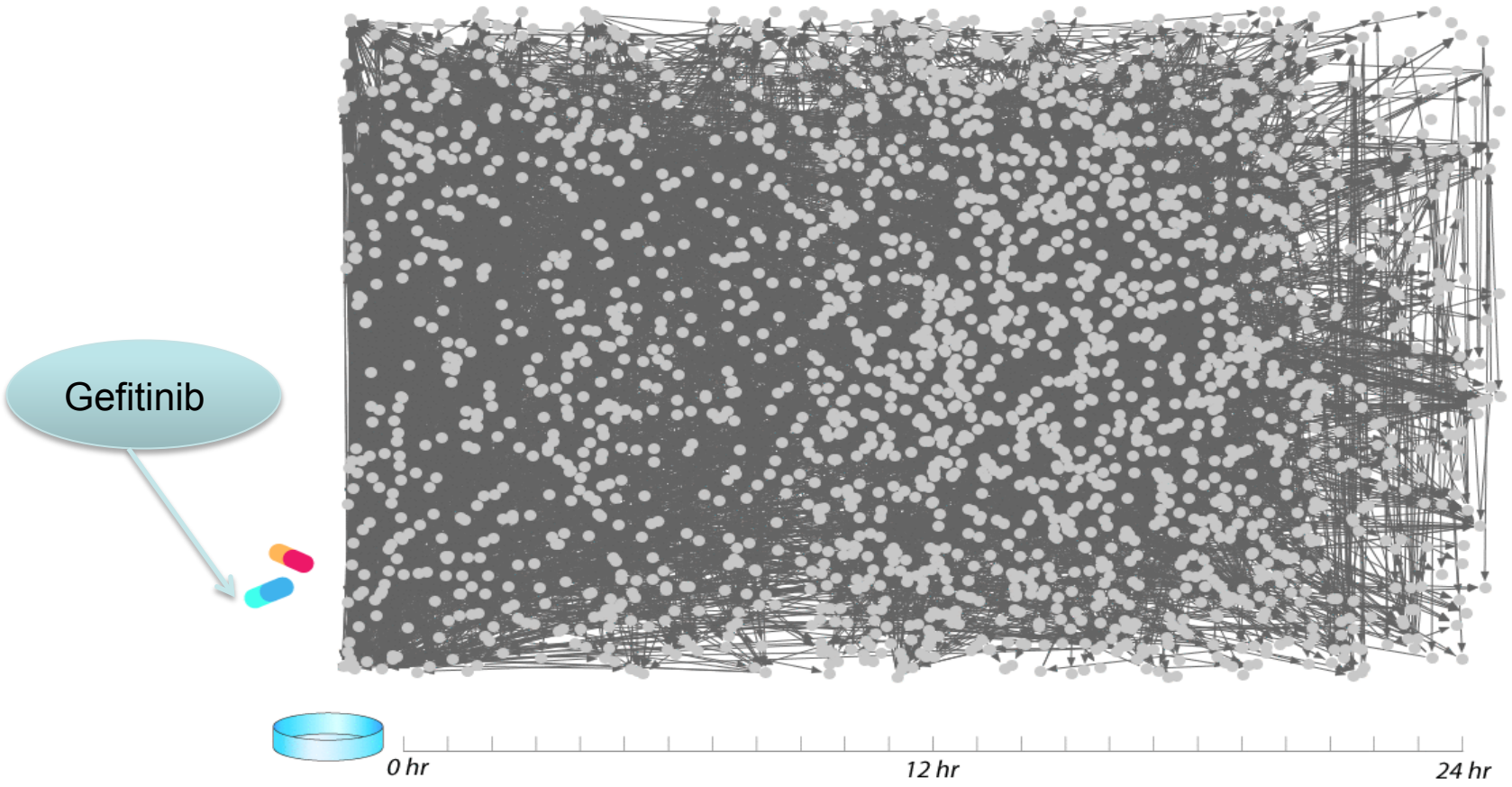

FN1
CEBPA
STRA6
JUN
PRKAA2
COL8A1
IRF1
CTSL2
TP53
MMP1
IL6
GBP2

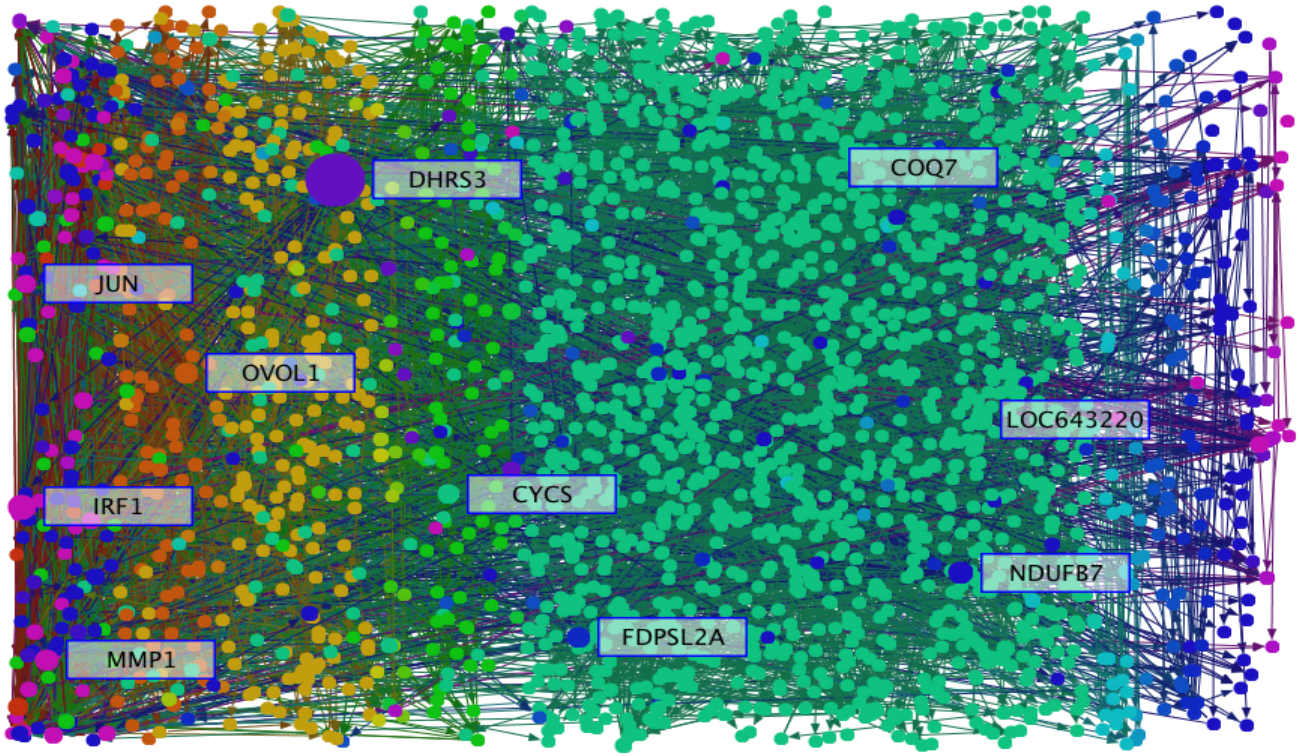

FN1
ERRFI1
SOCS3
STRA6
IGFBP6
IRF1
CTSL2
TP53
MMP1
IL6
SOCS2
BCL2
EGR1


CCND1
ERRFI1
SOCS3
IRF1
TP53
IL6
EGR1
PDE2A
LSM14B
COL8A1

Dynamic Bayesian Network + Nonparametric Regression




Gefitinib
0 hr
12 hr
24 hr

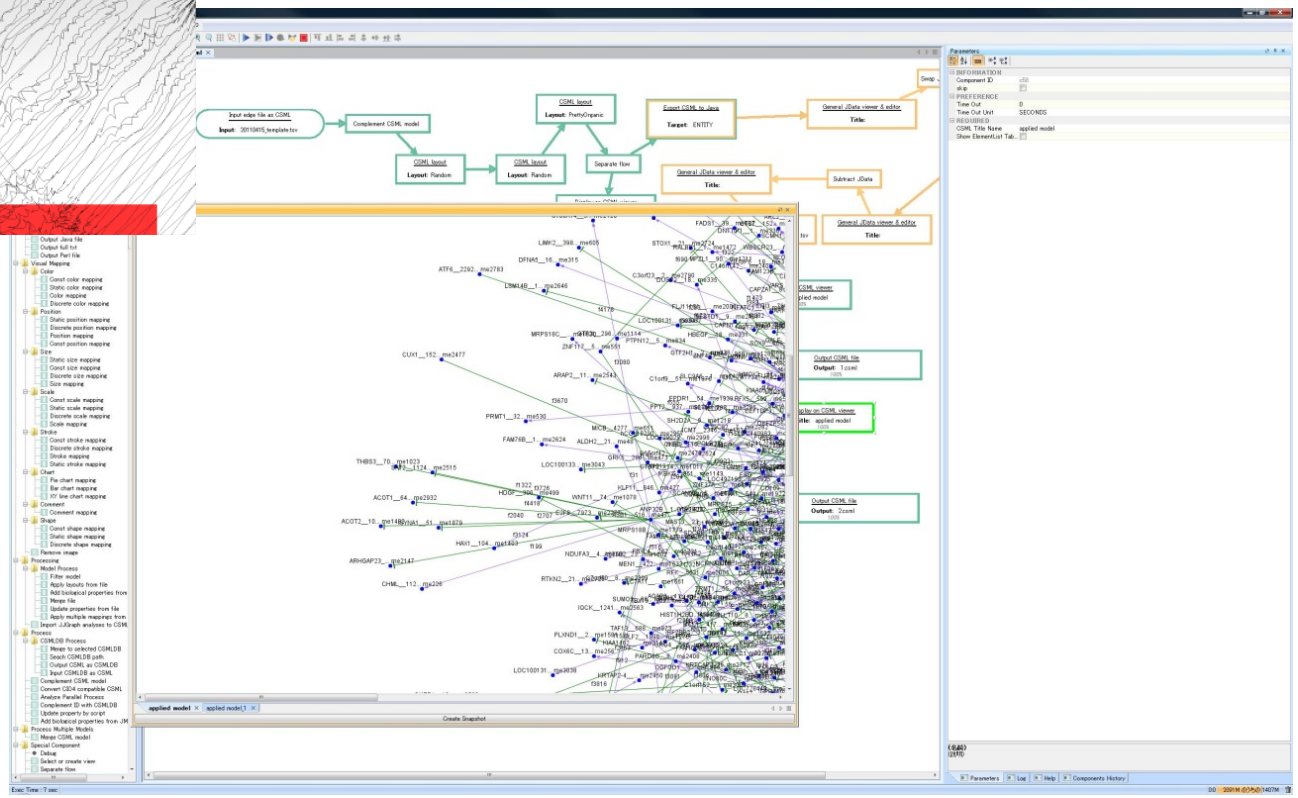

DHRS3
COQ7
JUN
OVOL1
LOC643220
IRF1
CYCS
NDUFB7
FDPSL2A
MMP1

5. Building Data Analysis and Simulation Pipeline at ONE STOP

XiP (eXtensible integrative Pipeline)

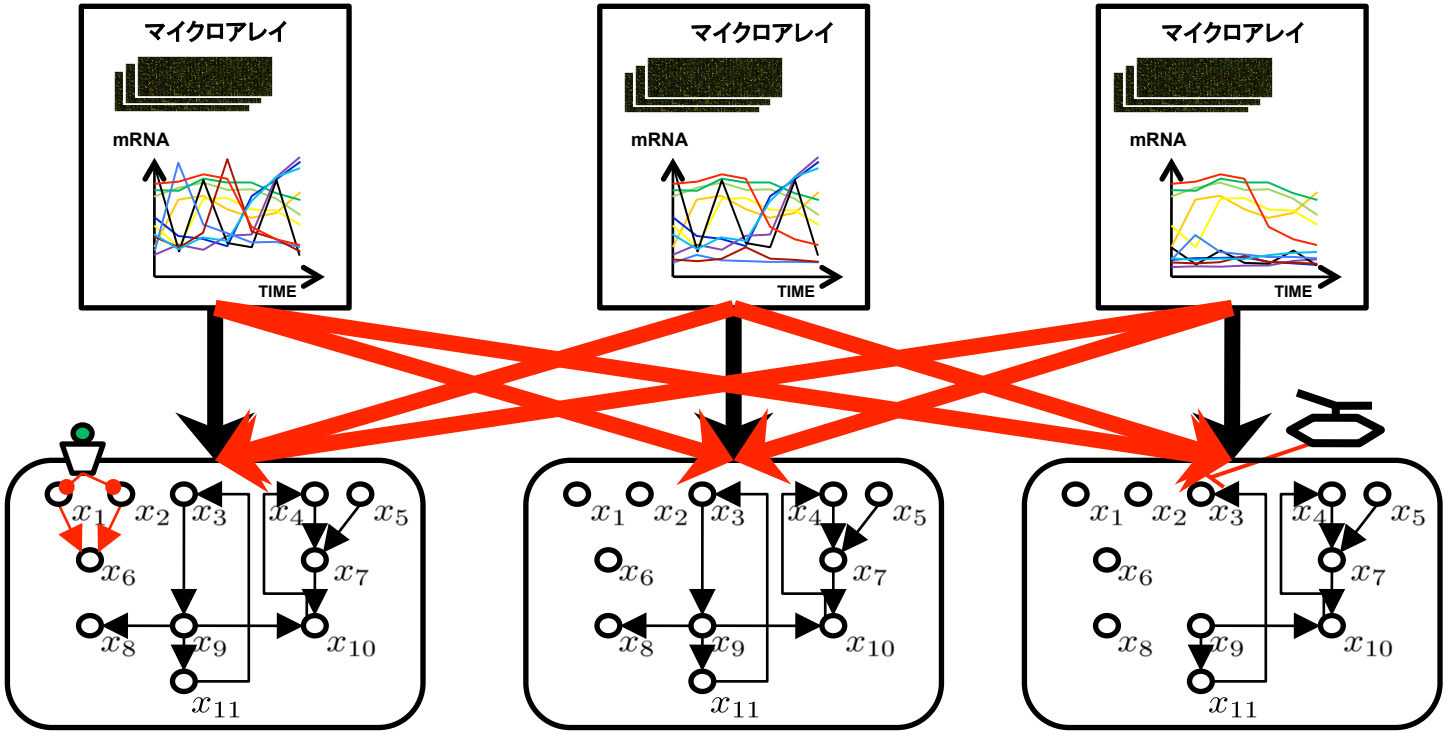
Systems Biology integrative Pipeline

NetComparator analysis pipeline

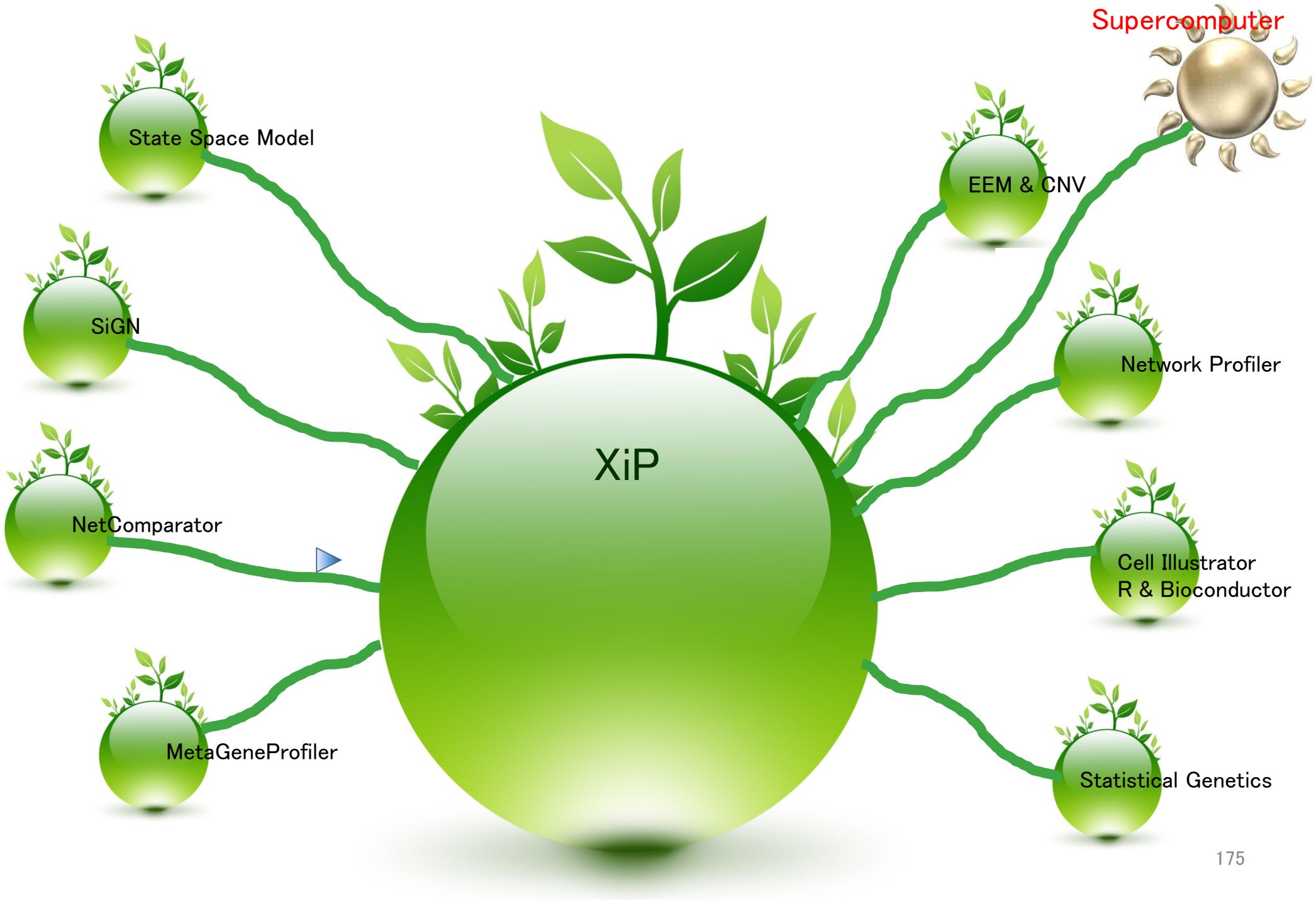


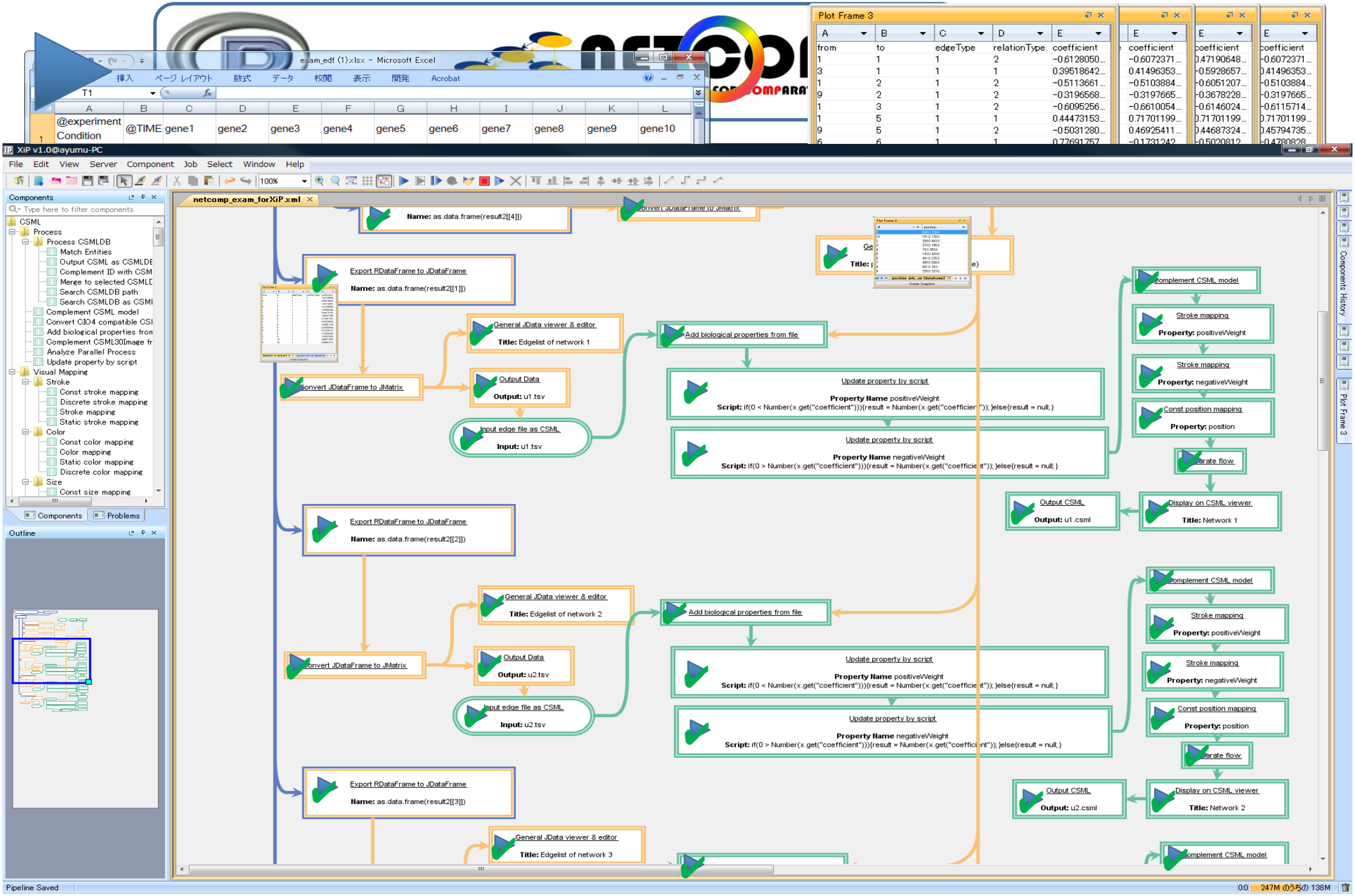
NetComparator

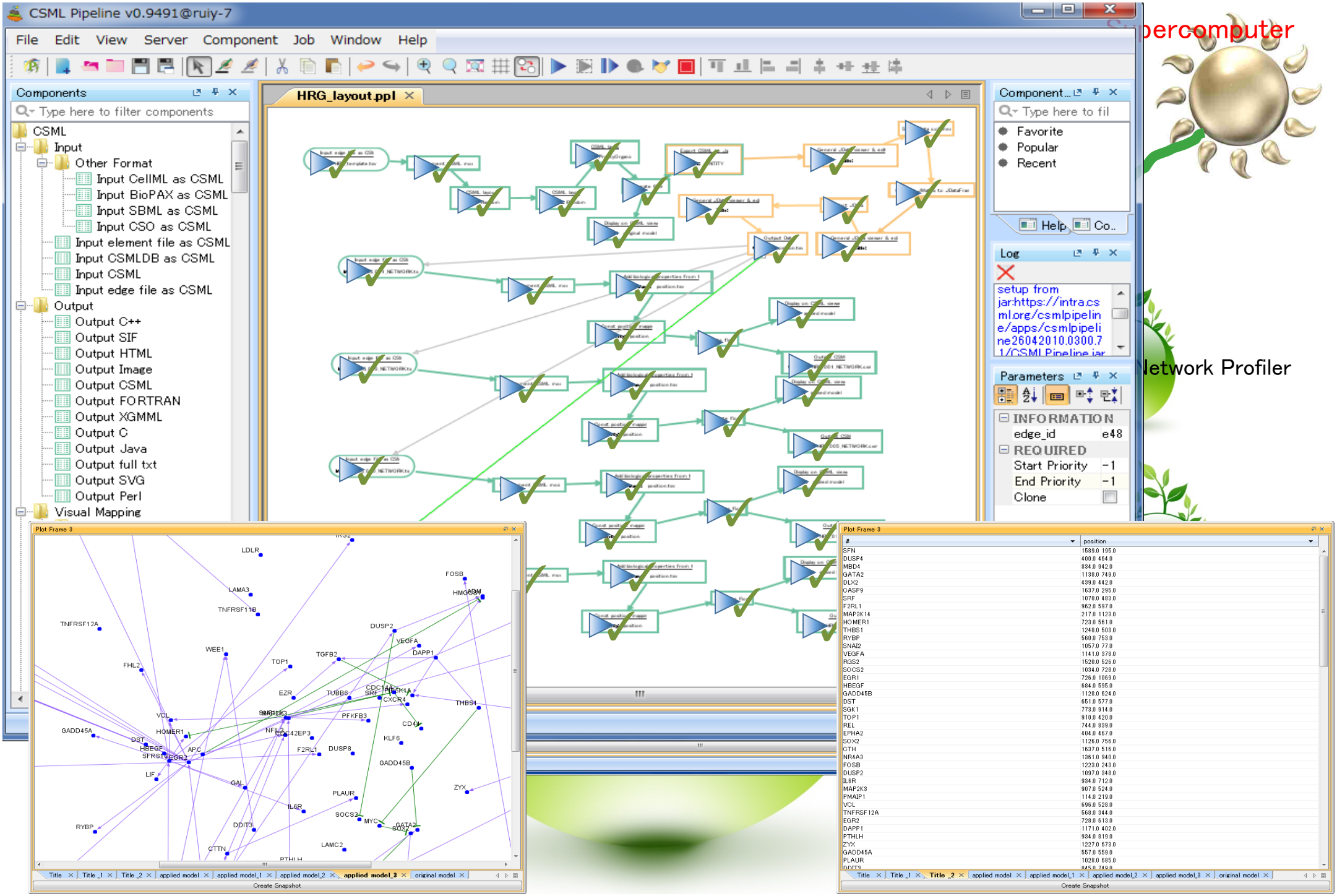
Shimamura, T., Imoto, S., Yamaguchi, R., Nagasaki, M., Miyano, S. Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. Bioinformatics. 26(8):1064-1072, 2010.



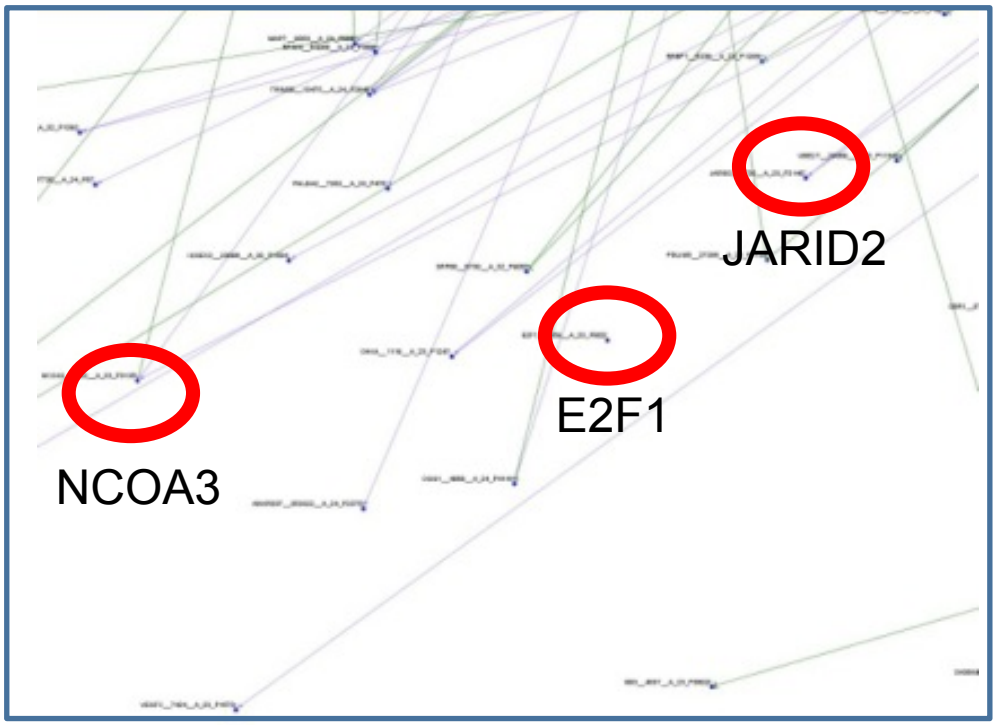
Regularized Weighted Recursive Elastic Net

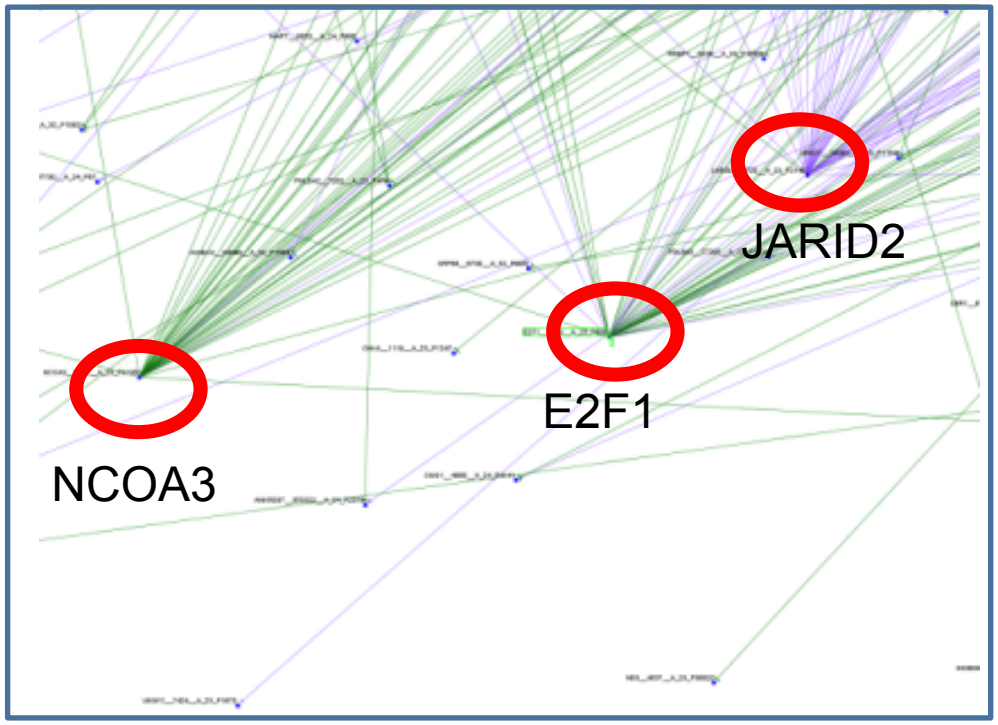

Supercomputer
State Space Model
SiGN
NetComparator
MetaGeneProfiler
XiP
EEM & CNV
Network Profiler
Cell Illustrator
R & Bioconductor
Statistical Genetics


exam_edf (1).xlsx - Microsoft Excel
@experiment Condition
@TIME
gene1
gene2
gene3
gene4
gene5
gene6
gene7
gene8
gene9
gene10
Plot Frame 3
from
to
edgeType
relationType
coefficient
XiP v1.0@ayumu-PC
File Edit View Server Component Job Select Window Help
netcomp_exam_forXiP.xml
Components
Outline
Export RDataFrame to JDataFrame
Name: as.data.frame(result2[[1]])
Name: as.data.frame(result2[[2]])
Name: as.data.frame(result2[[3]])
Name: as.data.frame(result2[[4]])
General JData viewer & editor
Title: Edgelist of network 1
Title: Edgelist of network 2
Title: Edgelist of network 3
Convert JDataFrame to JMatrix
Output Data
Output: u1.tsv
Output: u2.tsv
Input edge file as CSML
Input: u1.tsv
Input: u2.tsv
Add biological properties from file
Update property by script
Property Name positiveWeight
Property Name negativeWeight
Complement CSML model
Stroke mapping
Property: positiveWeight
Property: negativeWeight
Const position mapping
Property: position
Separate flow
Output CSML
Output: u1.csml
Output: u2.csml
Display on CSML viewer
Title: Network 1
Title: Network 2
Pipeline Saved


CSML Pipeline v0.9491@ruiy-7
File Edit View Server Component Job Window Help
Components
Type here to filter components
CSML
Input
Other Format
Input CellML as CSML
Input BioPAX as CSML
Input SBML as CSML
Input CSO as CSML
Input element file as CSML
Input CSMLDB as CSML
Input CSML
Input edge file as CSML
Output
Output C++
Output SIF
Output HTML
Output Image
Output CSML
Output FORTRAN
Output XGMML
Output C
Output Java
Output full txt
Output SVG
Output Perl
Visual Mapping
HRG_layout.ppl
Component...
Type here to fil
Favorite
Popular
Recent
Help
Co..
Log
setup from jar:https://intra.csml.org/csmlpipeline/apps/csmlpipeline26042010.0300.71/CSMLPipeline.jar
Parameters
INFORMATION
edge_id e48
REQUIRED
Start Priority -1
End Priority -1
Clone
Plot Frame 3
LDLR
LAMA3
TNFRSF11B
TNFRSF12A
FOSB
DUSP2
VEGFA
DAPP1
TGFB2
TOP1
WEE1
FHL2
EZR
TUBB6
CXCR4
THBS1
PFKFB3
CD44
KLF6
GADD45A
HOMER1
DST
HBEGF
APC
F2RL1
DUSP8
GADD45B
LIF
GAL
PLAUR
IL6R
SOCS2
MYC
ZYX
RYBP
DDIT3
CTTN
LAMC2
Title
Title_1
Title_2
applied model
applied model_1
applied model_2
applied model_3
original model
Create Snapshot
Plot Frame 3
#
position
SFN 1589.0 195.0
DUSP4 400.0 464.0
MBD4 834.0 942.0
GATA2 1138.0 749.0
DLX2 439.0 442.0
CASP9 1637.0 295.0
SRF 1070.0 483.0
F2RL1 962.0 597.0
MAP3K14 217.0 1123.0
HOMER1 729.0 561.0
THBS1 1248.0 503.0
RYBP 560.0 753.0
SNAI2 1057.0 77.0
VEGFA 1141.0 378.0
RGS2 1520.0 526.0
SOCS2 1034.0 728.0
EGR1 726.0 1069.0
HBEGF 684.0 595.0
GADD45B 1128.0 624.0
DST 651.0 577.0
SGK1 773.0 914.0
TOP1 910.0 420.0
REL 744.0 839.0
EPHA2 404.0 467.0
SOX2 1126.0 756.0
CTH 1637.0 516.0
NR4A3 1361.0 940.0
FOSB 1223.0 243.0
DUSP2 1097.0 348.0
IL6R 934.0 712.0
MAP2K3 907.0 524.0
PMAIP1 114.0 219.0
VCL 696.0 528.0
TNFRSF12A 568.0 344.0
EGR2 728.0 613.0
DAPP1 1171.0 402.0
PTHLH 934.0 819.0
ZYX 1227.0 673.0
GADD45A 567.0 559.0
PLAUR 1028.0 685.0
DDIT3 845.0 749.0
Title
Title_1
Title_2
applied model
applied model_1
applied model_2
applied model_3
original model
Create Snapshot

Difference between Gefitinib Sensitive Lung Cancer and Resistant Lung Cancer





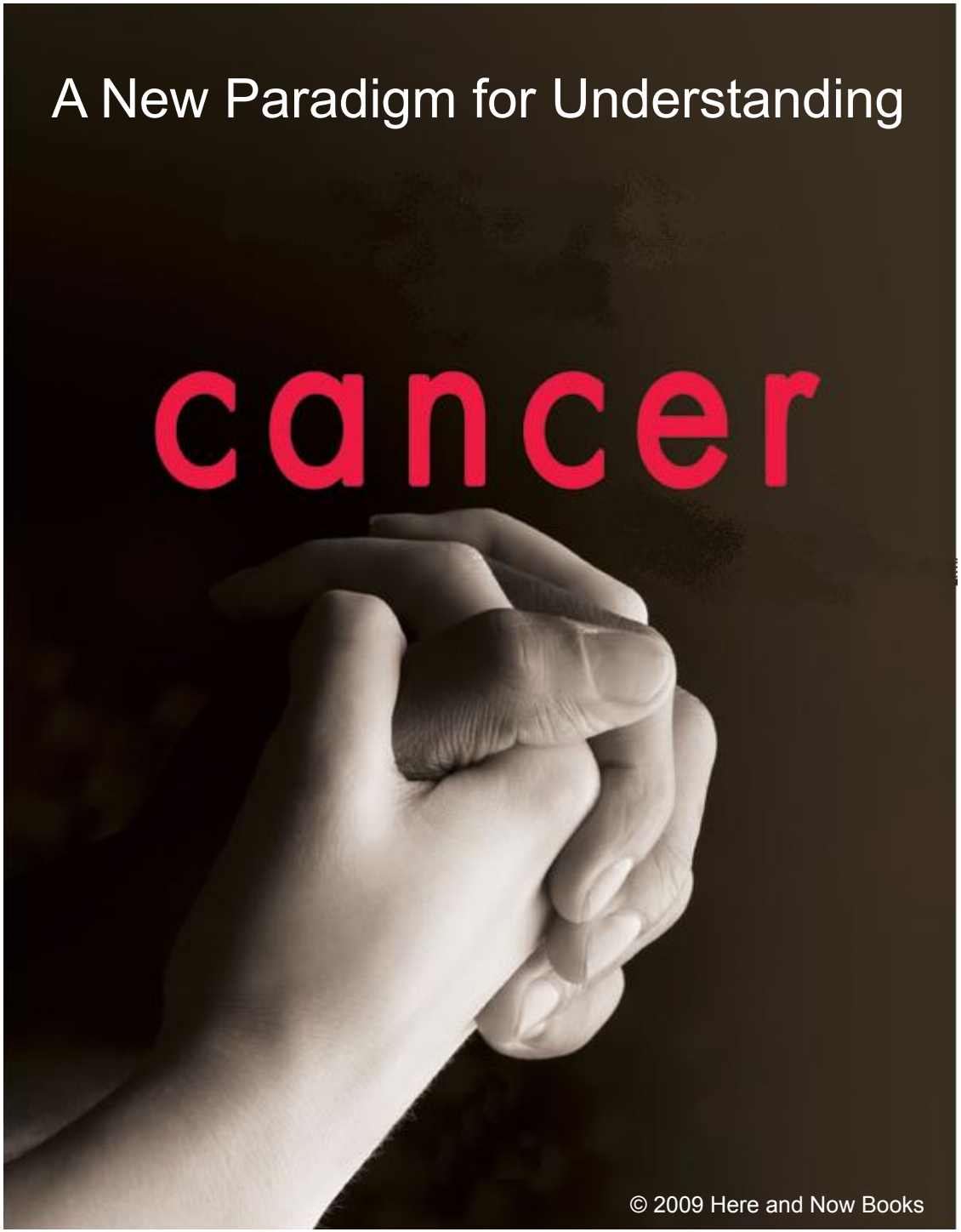
E2F1 is knows as a TF regulating apoptosis and cell cycle

Gefitinib Sensitive Lung Cancer

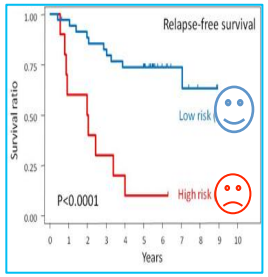
Gefitinib Resistant Lung Cancer

NCOA3: a nuclear receptor coactivator that interacts with nuclear hormone receptors toenhance their transcriptional activator functions.

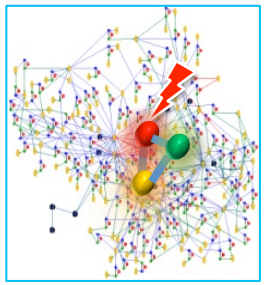
A New Paradigm for Understanding



Super-early biomarker



Prediction of efficacy and relapse



New molecular targets

Mechanism of drug resistance

Systems Understanding of Cancer