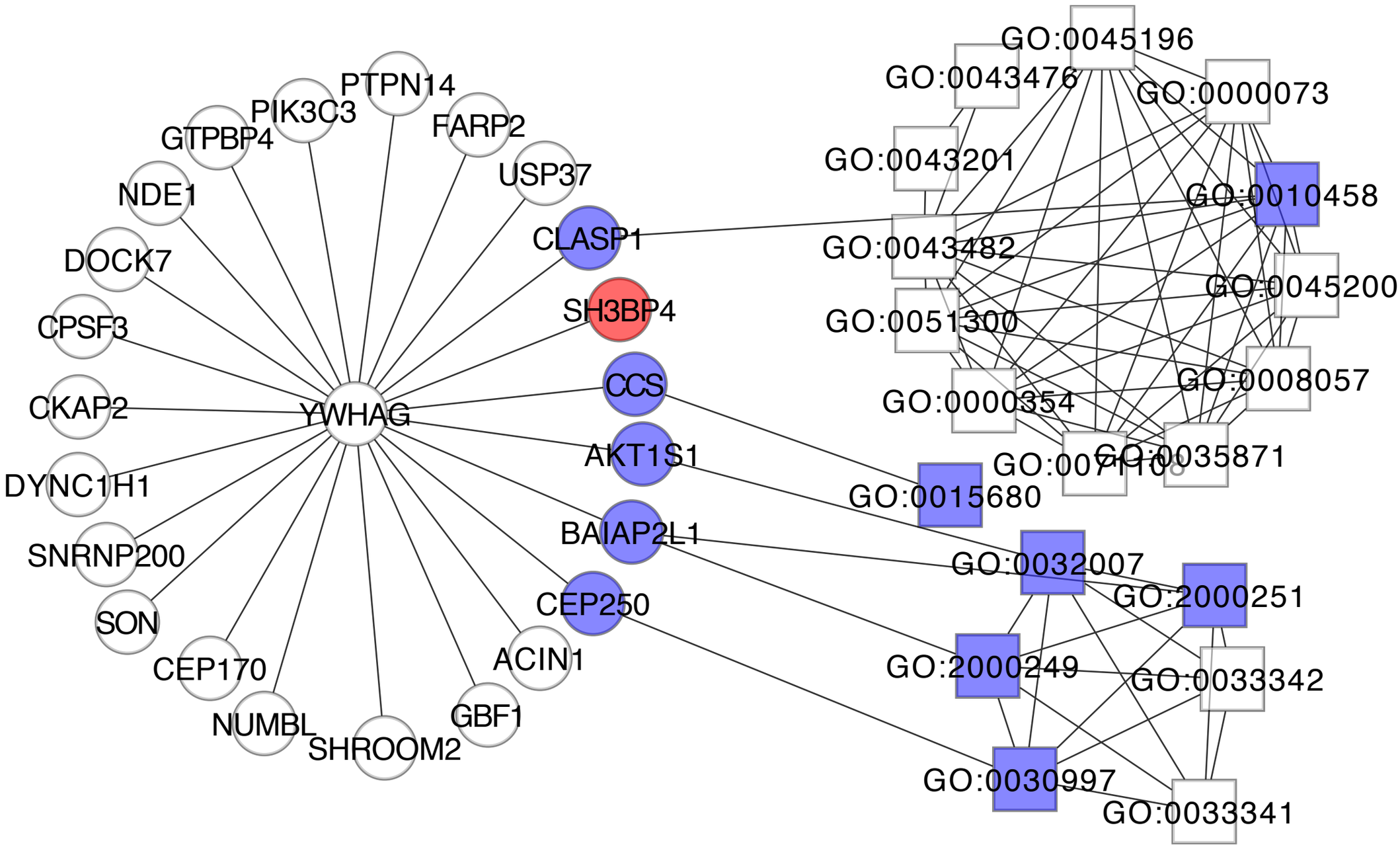
SH3BP4 function prediction



GO:0032007 negative regulation of TOR signaling cascade

- Discovery of SH3BP4 as a potential candidate of tumor suppressor that functions in the Rag GTPase-mTORC1 signaling [*Molecular Cell 2012*]

Network biology methods integrating genomic data with biological prior knowledge for cancer genomics

Tae Hyun Hwang, Ph.D.
Biostatistics and Bioinformatics, Masonic Cancer Center
University of Minnesota Twin-Cities

UNIVERSITY OF MINNESOTA
Driven to Discover℠

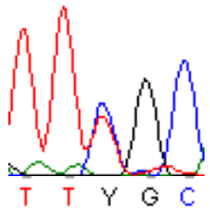
통계유전학 워크샵

NCI CCC
A Comprehensive Cancer Center Designated by the National Cancer Institute

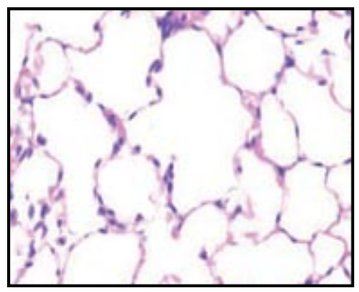
Motivation

- A catalogue of molecular aberrations that cause cancer is critical for developing and deploying therapies that will improve patients' lives

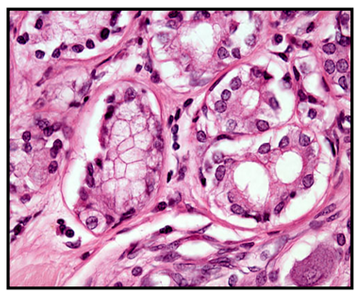
SEQUENCE MUTATIONS



TRANSLOCATIONS

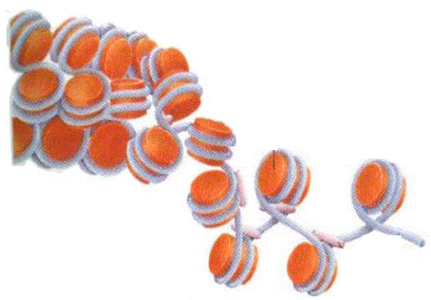


NORMAL CELLS

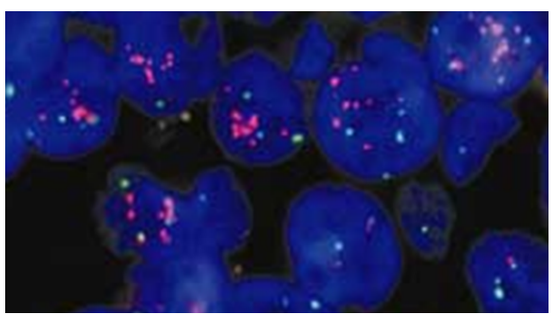


CANCER CELLS

EPIGENETIC ALTERATIONS

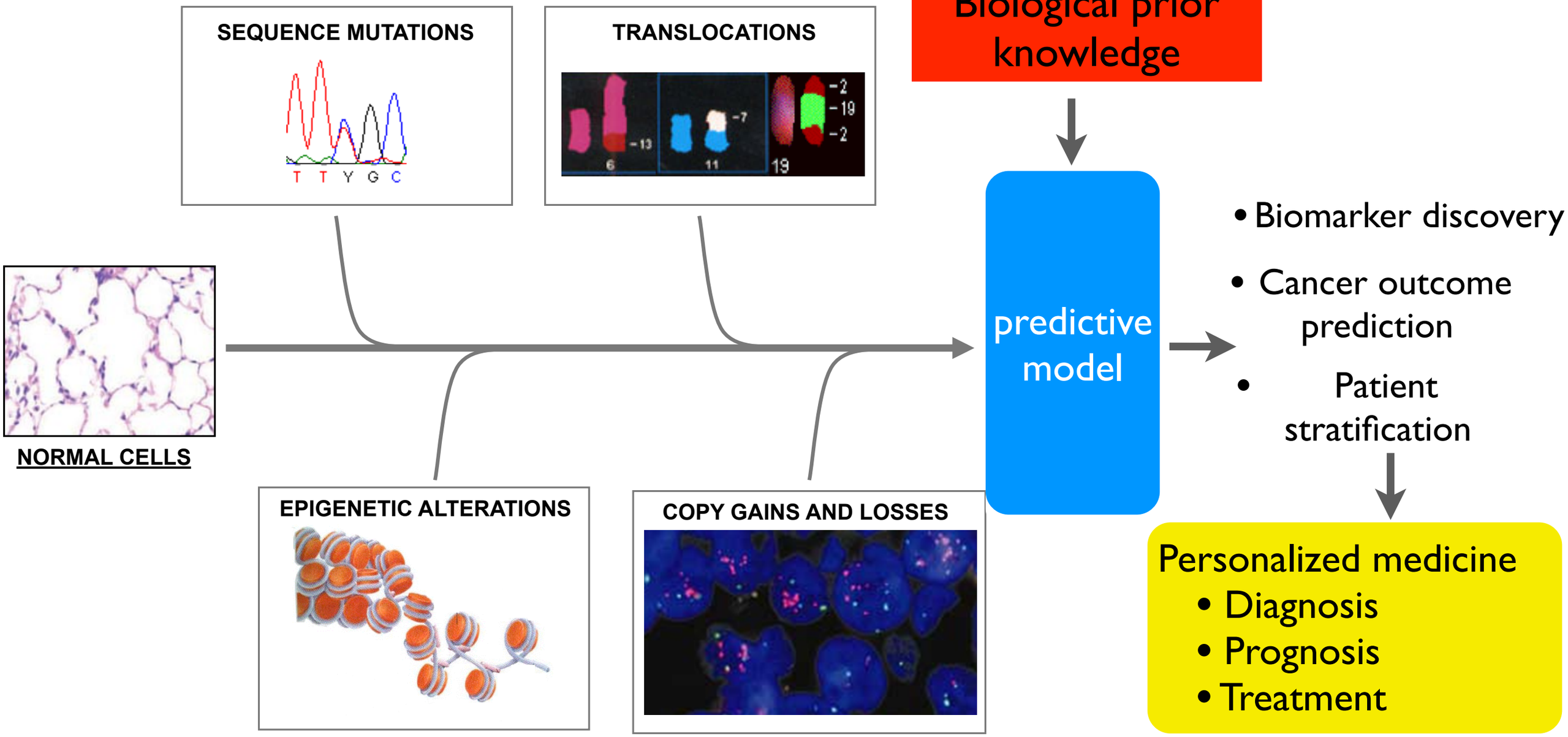


COPY GAINS AND LOSSES



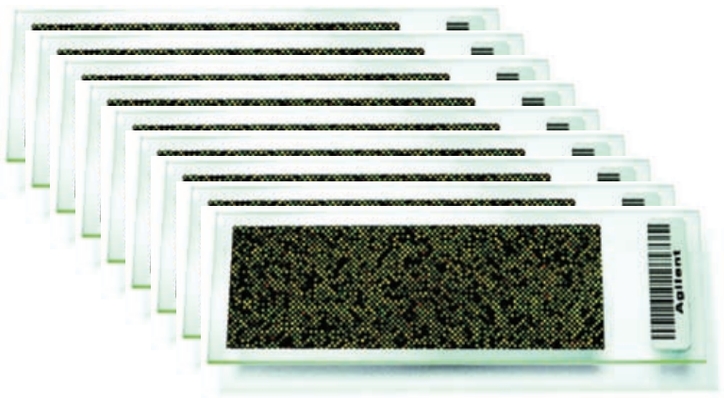
Motivation

- Integrating data with prior knowledge to build reliable predictive models for the development of drug targets and efficient therapeutic strategies is one of key challenges in cancer genomics

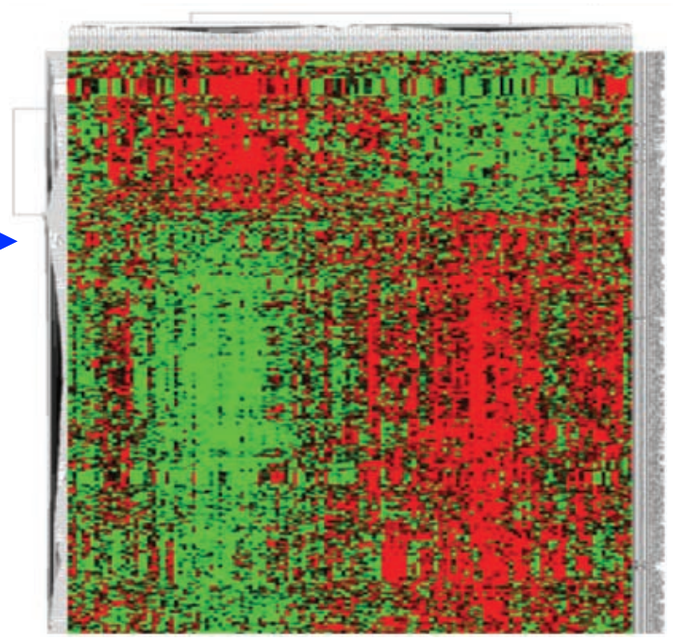


Gene expression biomarker

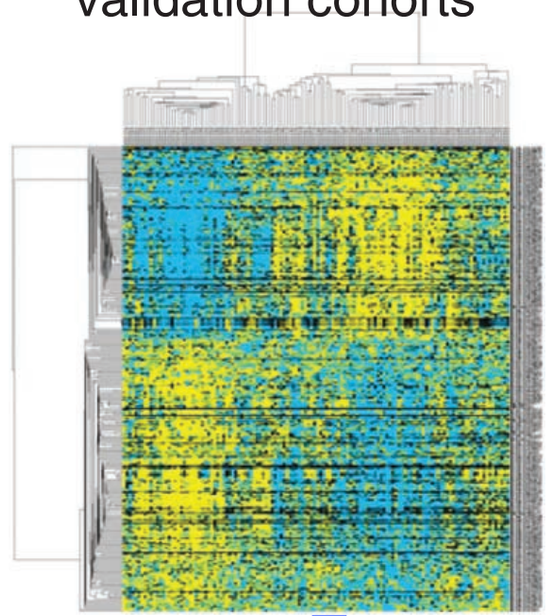
a Whole-genome analysis



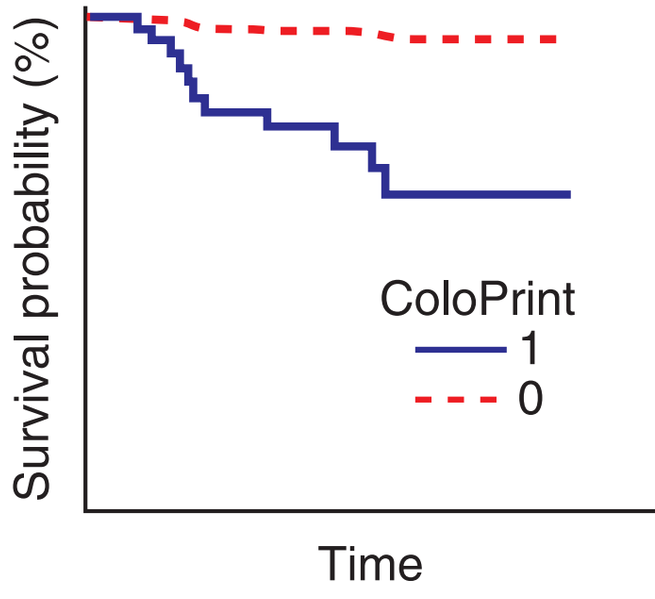
b Generate signature of genes that correlate with clinical parameter



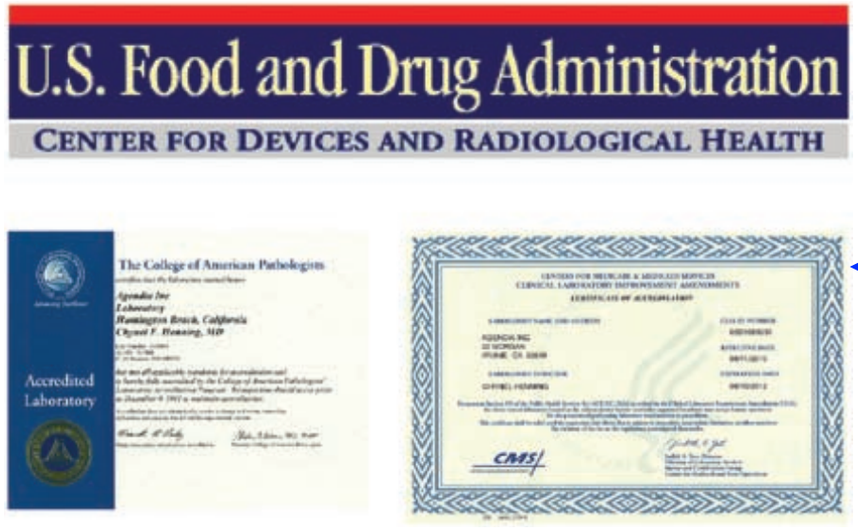
c Apply to independent validation cohorts



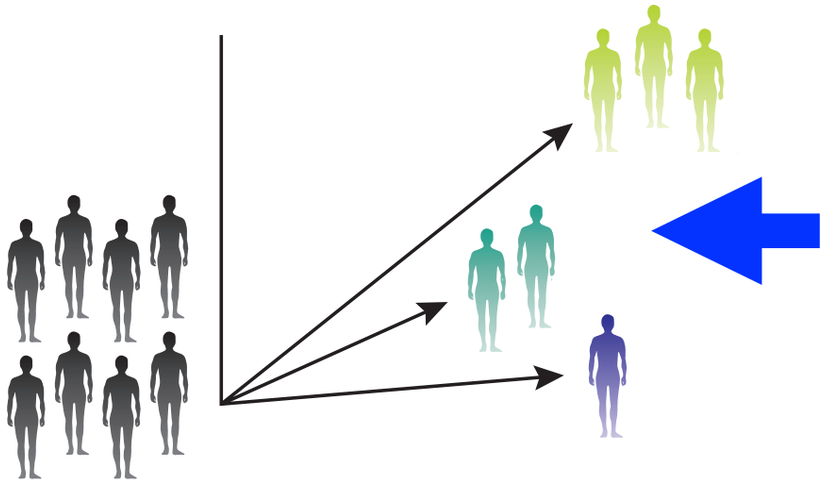
d Evaluate clinical performance



e Regulatory approval



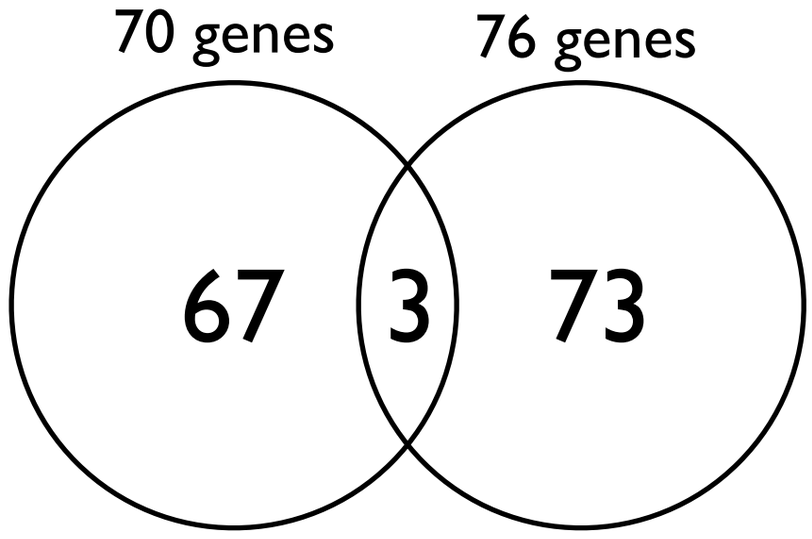
f Clinical implementation



Ian J Majewski & René Bernards.," Taming the dragon: genomic biomarkers to individualize the treatment of cancer", Nature Medicine 2011

Challenge

- Inconsistent biomarker discovery
    - Only three common biomarkers from two breast cancer studies



van't veer et.al, Nature 2002 Wang et. al, Lancet 2005

- high dimensions but low sample size
- different platform (Agilent vs. Affymetrix)
- noise and etc.

Network biology methods

- Network-based learning methods

  - Represent data as objects (i.e. patients, genes, or disease) and edges (i.e., interactions, co-expressions or associations)

    - Capture the **dependency** (i.e. interactions, co-expression, or co-occurrences) of genes, SNPs, and Copy Numbers Variations (CNVs)

    ✓ **Interpretability**

      - Biologically interpretable

- Data Integration
- Generalizability and scalability

  - Efficient optimization

Our approach

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 24 no. 18 2008, pages 2023–2029
doi:10.1093/bioinformatics/btn383

Gene expression

Robust and efficient identification of biomarkers by classifying features on graphs

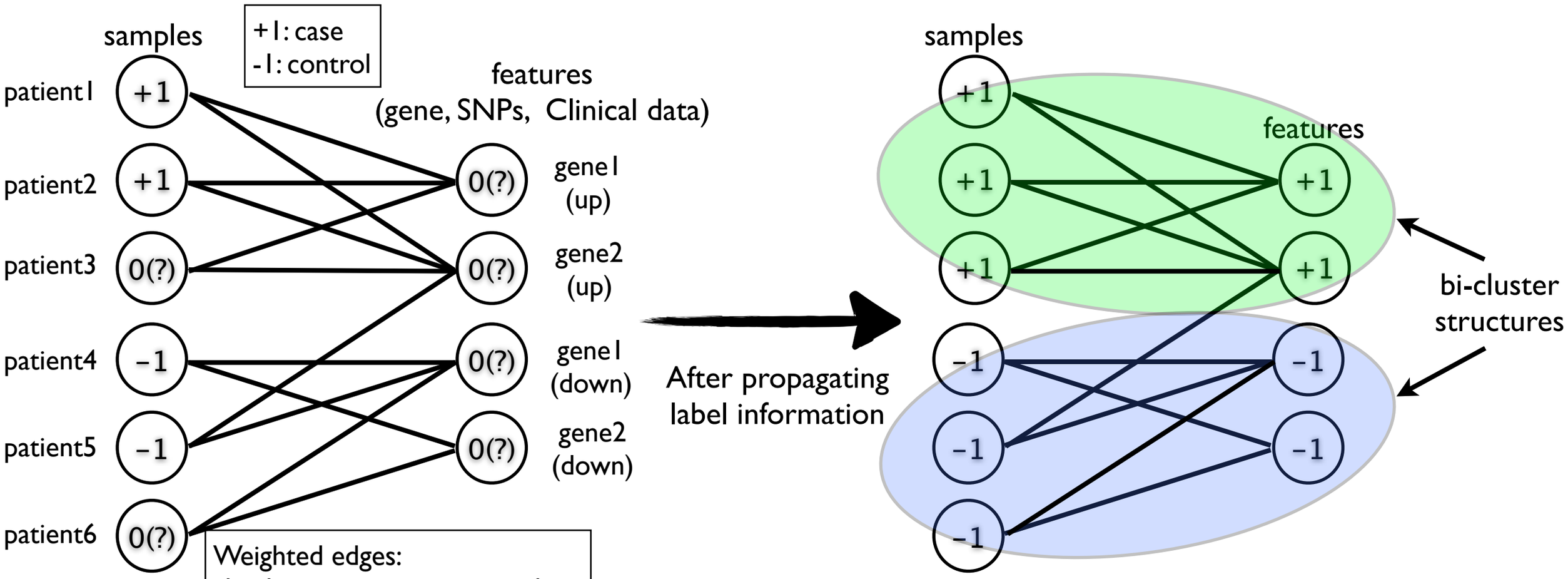
TaeHyun Hwang[1], Hugues Sicotte[2], Ze Tian[1], Baolin Wu[3], Jean-Pierre Kocher[2], Dennis A. Wigle[4], Vipin Kumar[1] and Rui Kuang[1,*]

[1]Department of Computer Science and Engineering, University of Minnesota, Twin Cities, [2]Bioinformatics Core, Mayo Clinic College of Medicine, Rochester, [3]Division of Biostatistics, School of Public Health, University of Minnesota, Twin Cities and [4]Division of General Thoracic Surgery, Mayo Clinic Cancer Center, Rochester, MN, USA

***Joint work w/ Mayo Clinic and IBM TJ Watson**

Network Propagation

✓ Use labeled samples to classify **unlabeled samples** and **genes** by exploring bi-cluster structures of the graph



$$\Omega(f) = \sum_{(v,u)\in E} w(v,u)\left(\frac{f(v)}{\sqrt{d(v)}} - \frac{f(u)}{\sqrt{d(u)}}\right)^2 + \varrho\sum_{v\in V}(f(v)-y(v))^2 + \varrho\sum_{u\in U}(f(u)-y(u))^2,$$

Graph $G = (U, V, E, w)$

U : Sample nodes

V : Feature nodes

E : Weighted edges

$f(x)$: Label assignment

$w(v,u)$: Edge weight between u and v

$d(u) = \sum_{(v,u)\in E} w(v,u)$

$y(x)$: Initial label

T. Hwang et. al, **Bioinformatics** 2008

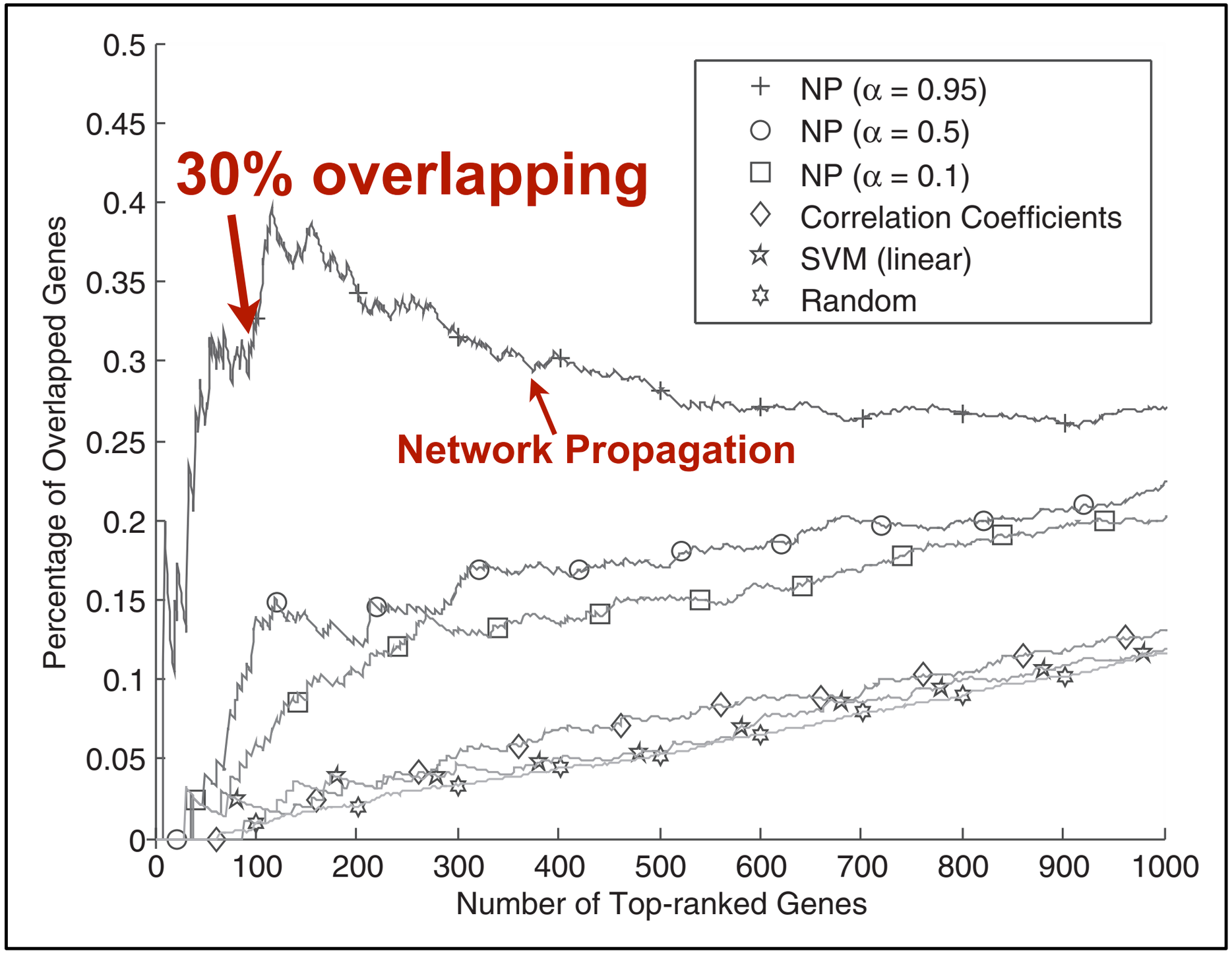
Classification results

Algorithms	Rosetta		Vijver	Wang
	Clinical	Genes	Genes	Genes
(A) Classification results on three datasets				
Network propagation	**0.788**	**0.740**	**0.667**	**0.564**
SVM (linear)	0.773	0.730	**0.662**	0.536
SVM (RBF)	0.783	0.737	0.661	**0.568**
Naïve Bayes	**0.795**	0.617	0.476	0.554
LDA	0.579	**0.740**	0.648	0.502
(B) Comparison between network propagation and the baseline algorithms				
NP versus SVM (linear)	278/31/191	247/27/226	242/86/172	309/25/166
NP versus SVM (RBF)	248/44/208	214/124/162	254/81/165	137/130/233
NP versus Naïve Bayes	144/106/250	393/10/97	466/3/31	261/24/215
NP versus LDA	460/8/32	232/36/232	297/61/142	359/15/126

***The classification performance of all methods are evaluated using area under the receiver operating characteristics (ROC) score.**

T. Hwang et. al, **Bioinformatics** 2008

Reproducible biomarker



T. Hwang et. al, **Bioinformatics** 2008

Take home message

- We proposed a novel network-based learning algorithm to classify genes and patients in the bipartite graph
- Exploring the cluster structure of the bipartite graph (e.g., co-expression) could help to accurately predict cancer outcome and identify reproducible biomarker
- The proposed method is general method, and applicable for other genomic data (e.g., SNPs, copy number variation, and clinical data)
- No improvement has been achieved from simple data integration

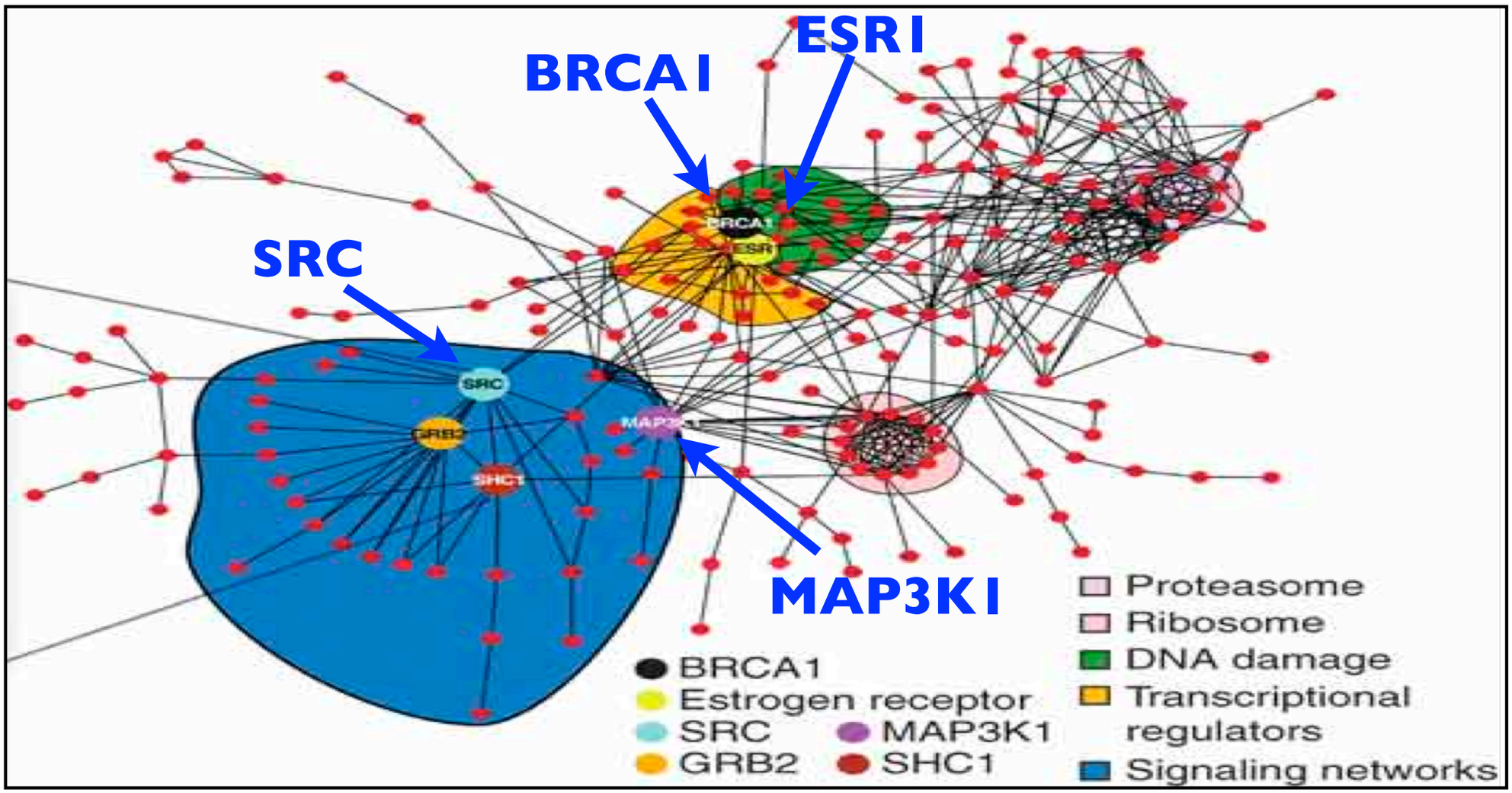
•**T. Hwang**, H. Sicotte, Z. Tian, B. Wu, JP Kocher, D. Wigle, V. Kumar, R. Kuang, "Robust and efficient identification of biomarker by classifying features on graph", **Bioinformatics** 2008

•**T. Hwang**, and R. Kuang. "A Comparative study of breast cancer microarray gene expression profiles using label propagation", **SDM** 2008

•**T. Hwang**, H. Sicotte, JP Kocher, D. Wigle, V. Kumar, R. Kuang, "Identifying clinical and genetic markers of human disease by classifying features on graphs", **Technical Report UMN-CS-07-021** 2007
  - Chronic Fatigue Syndrome (SNPs, Gene Expression Data)

Biological prior knowledge

- Protein-protein interaction networks can provide
  - Modular structures of genes having similar functions, and involved in same pathways
  - Cancer genes tend to interact with each other in protein-protein interaction networks (PPI)



Ian W Taylor et al., Dynamic modularity in protein interaction networks predicts breast cancer outcome, Nature Biotechnology 2009

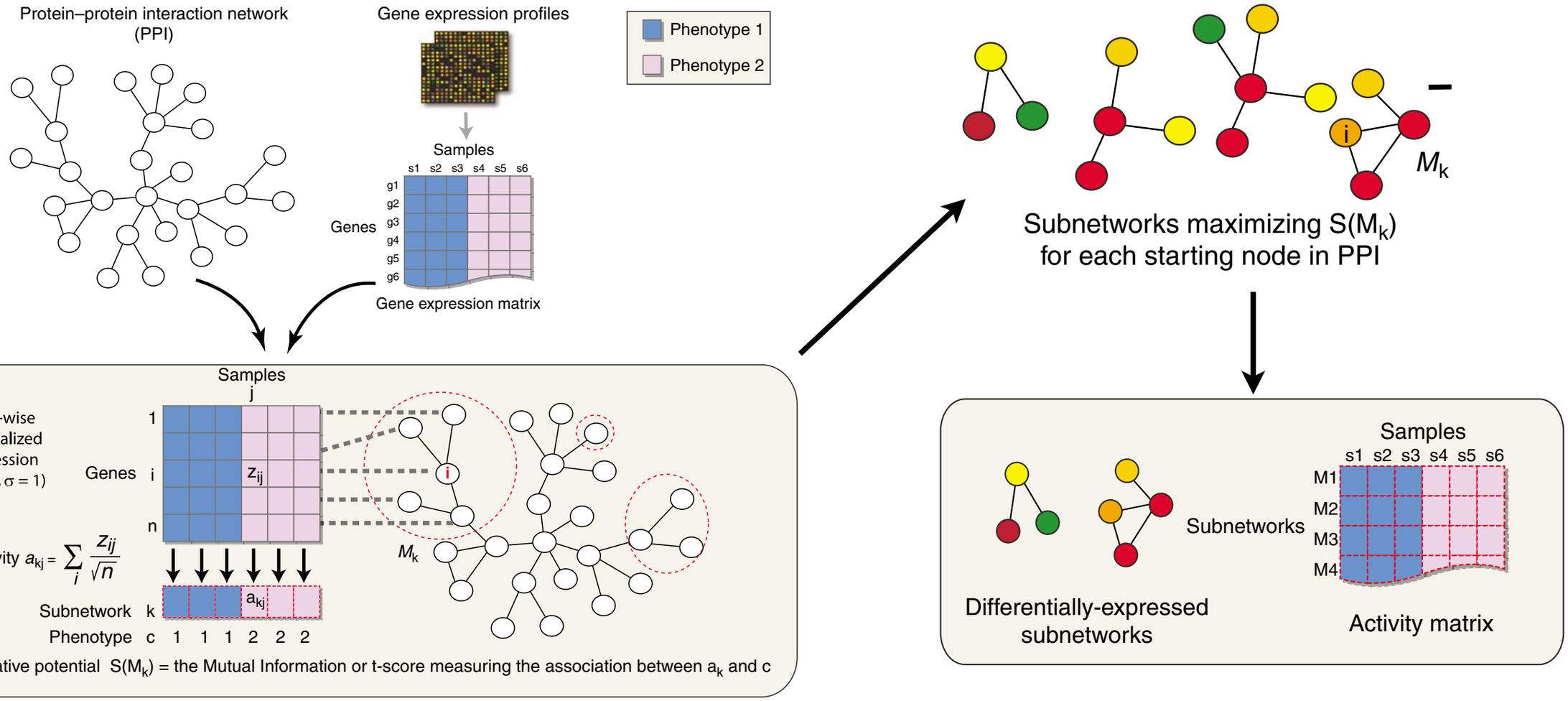
Network-based method

- Two step approach

: Best available approaches are often two step approaches:

1) Use seed genes from data and identify subnetworks

2) Use classifiers (e.g., SVM) with selected genes (member genes in the subnetworks) to predict clinical outcomes



Chuang et al, Molecular System Biology 2007

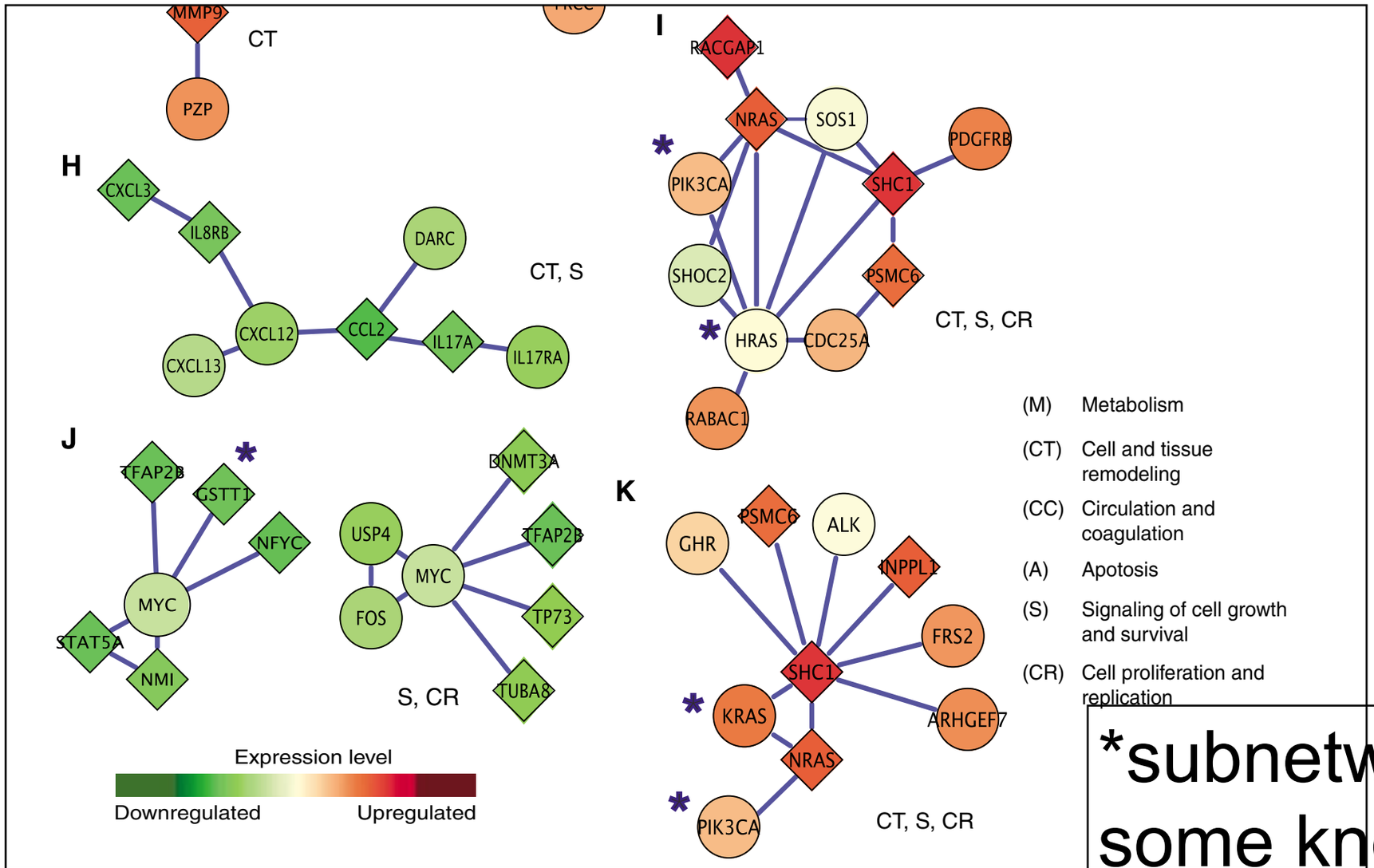
Network-based method

- Two step approach

: Best available approaches are often two step approaches:

1) Use seed genes from data and identify subnetworks

2) Use classifiers (e.g., SVM) with selected genes (member genes in the subnetworks) to predict clinical outcomes



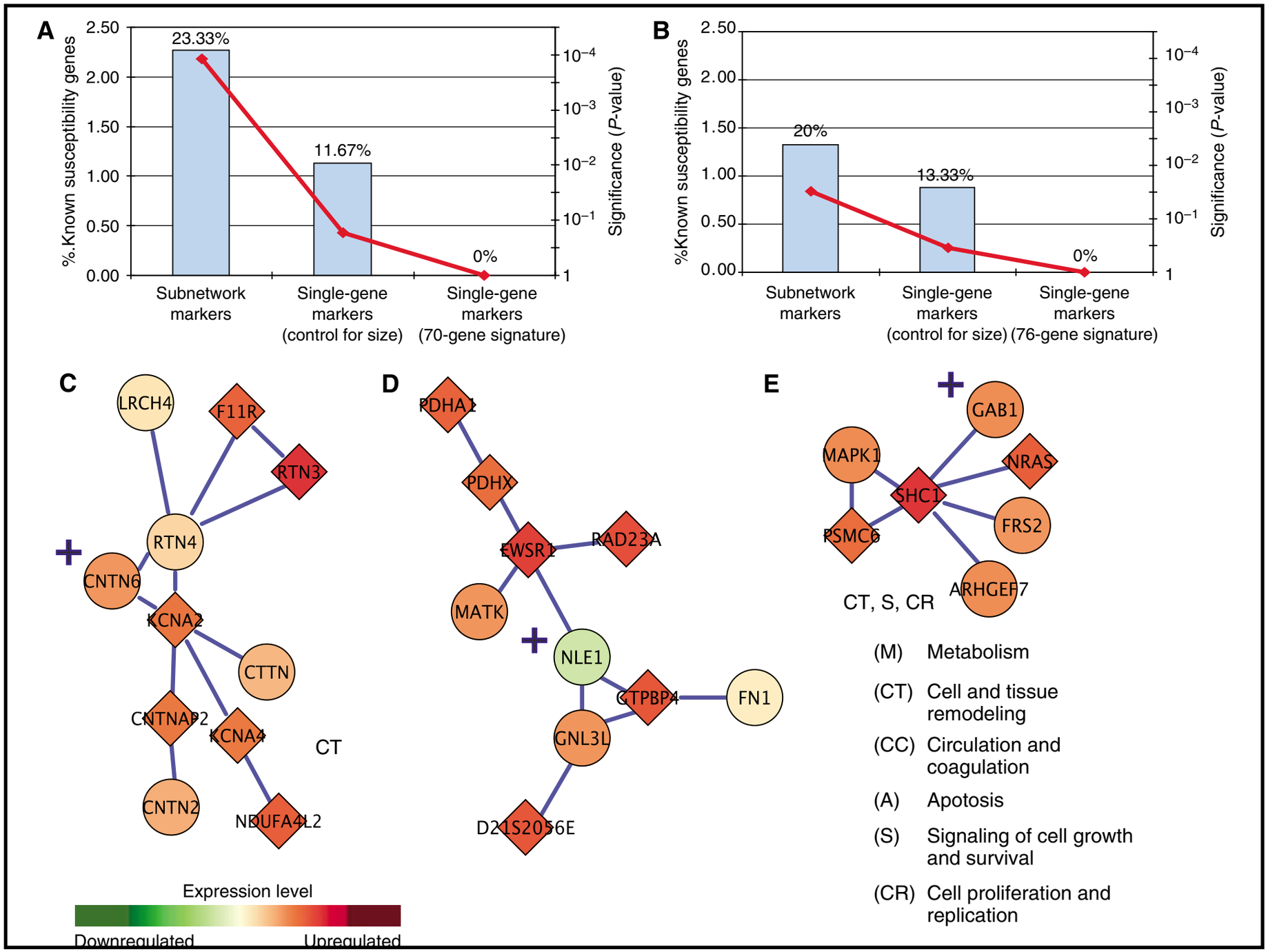
*subnetworks include some known oncogenes!

Chuang et al, Molecular System Biology 2007

Network-based method

- Two step approach

  : More reproducible biomarker & accurate cancer outcome prediction!



Chuang et al, Molecular System Biology 2007

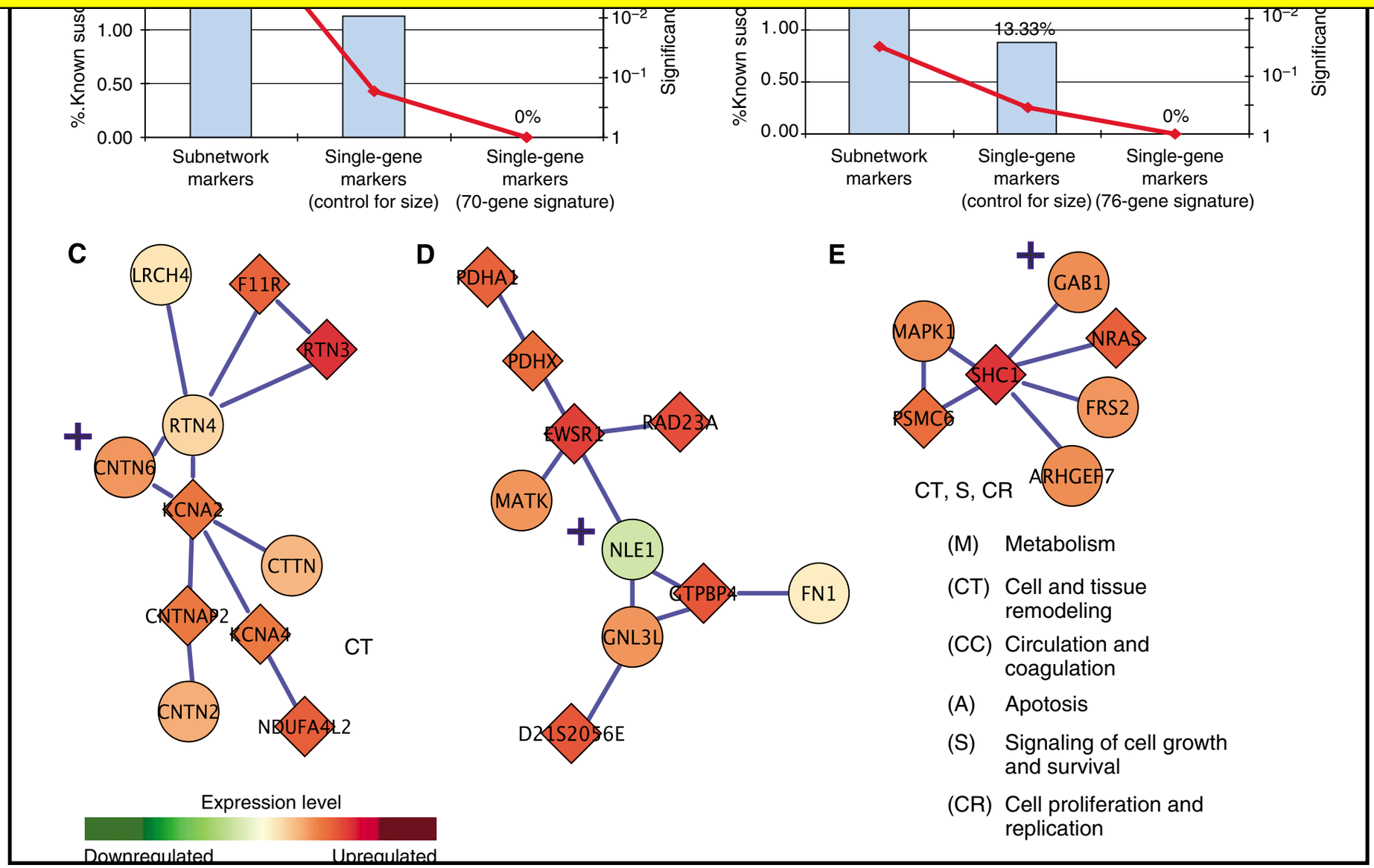
Network-based method

- Two step approach

Disadvantage:

- Use heuristic function to identify subnetworks

- Do not utilize interactions between genes when perform classification



Chuang et al, Molecular System Biology 2007

Network-based method

2008 Eighth IEEE International Conference on Data Mining

Learning on Weighted Hypergraphs to Integrate Protein Interactions and Gene Expressions for Cancer Outcome Prediction

TaeHyun Hwang*, Ze Tian*, and Rui Kuang†
Department of Computer Science and Engineering
University of Minnesota Twin Cities
thwang, tianze, kuang@cs.umn.edu

Jean-Pierre Kocher
Bioinformatics Core
Mayo Clinic College of Medicine
Kocher.JeanPierre@mayo.edu

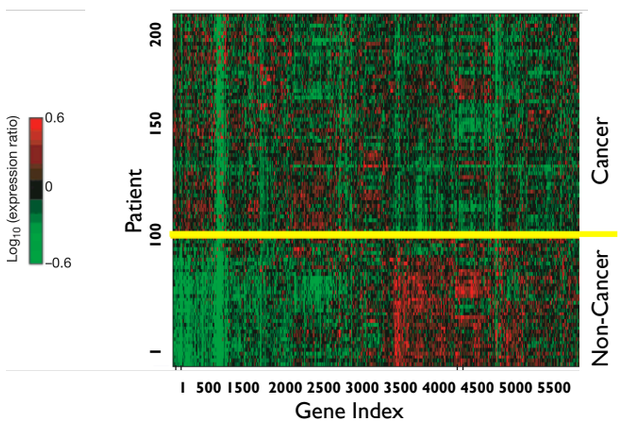
BIOINFORMATICS **ORIGINAL PAPER**
Vol. 25 no. 21 2009, pages 2831–2838
doi:10.1093/bioinformatics/btp467

Systems biology

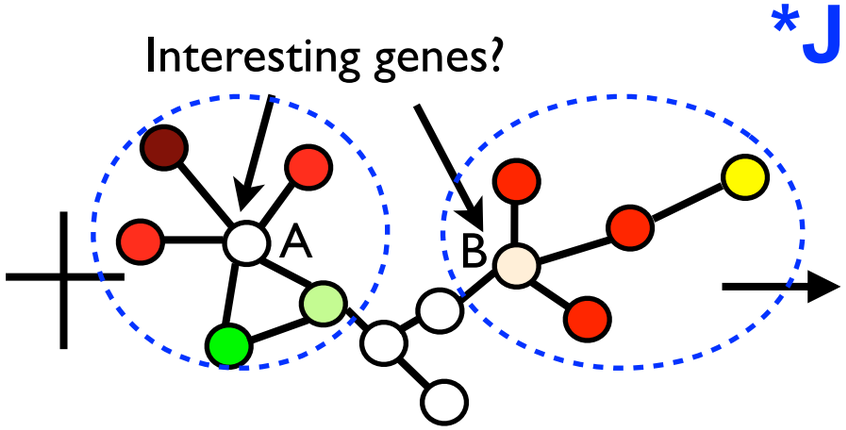
A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge

Ze Tian†, TaeHyun Hwang† and Rui Kuang*
Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA

***Joint work with Mayo Clinic**

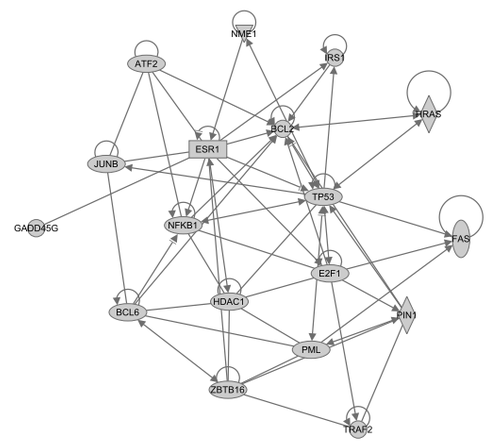


Gene expression/Copy Number



protein-protein interaction networks

- Subnetwork marker
- Cancer outcome prediction

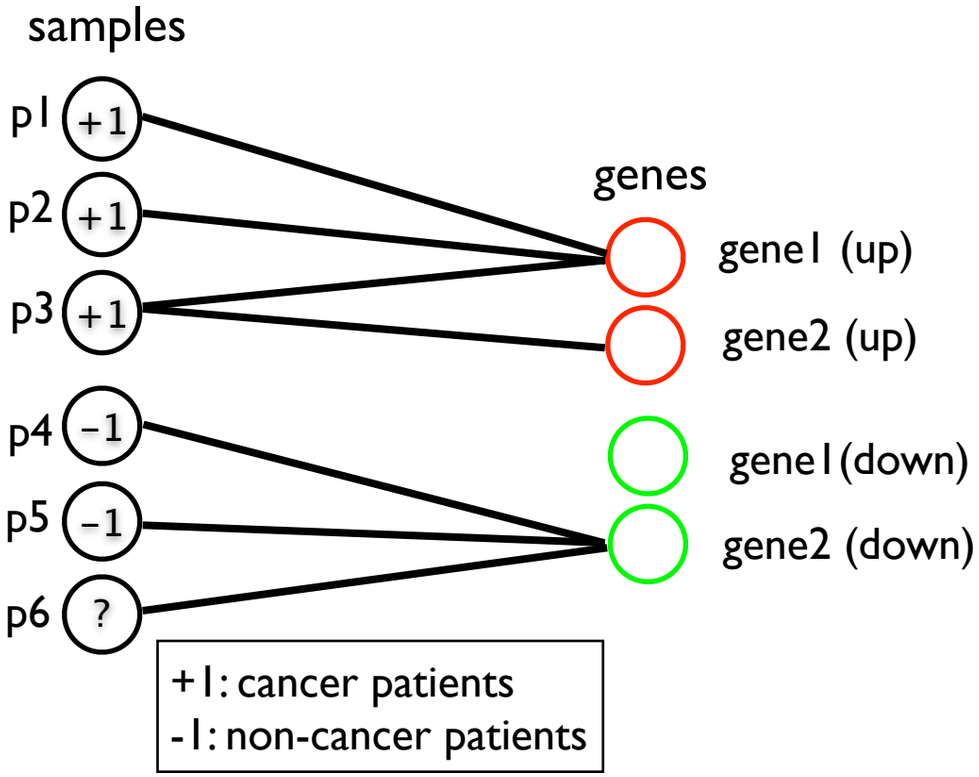


Hypergraph vs. normal graph

Microarray gene expression data

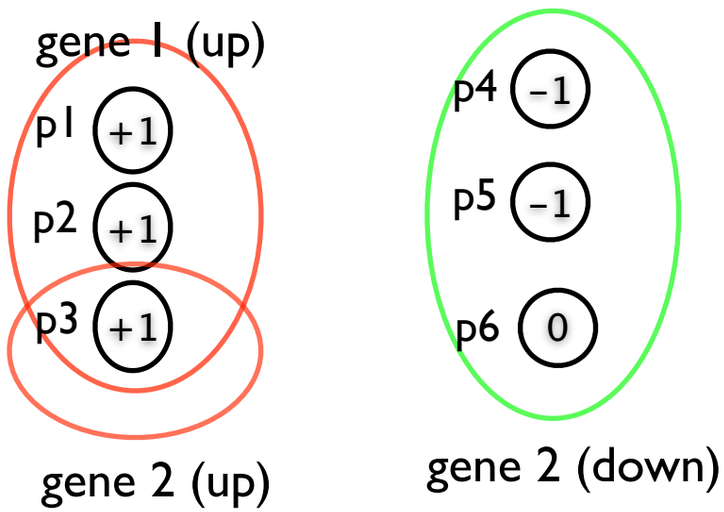
Sample	Disese status	Gene1 (up)	Gene2 (up)	Gene1 (down)	Gene2 (down)
Patient1	Cancer	1	0	0	0
Patient2	Cancer	1	0	0	0
Patient3	Cancer	1	1	0	0
Patient4	Normal	0	0	0	1
Patient5	Normal	0	0	0	1
Patient6	?	0	0	0	1

bi-partite graph



VS

hypergraph



T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Regularization framework

$$\min_{f,w} \Phi(f,w) = \Omega(f,w) + \mu \| f - y \|^2 + \rho\Psi(w)$$

Learning labels **Learning weights of genes**

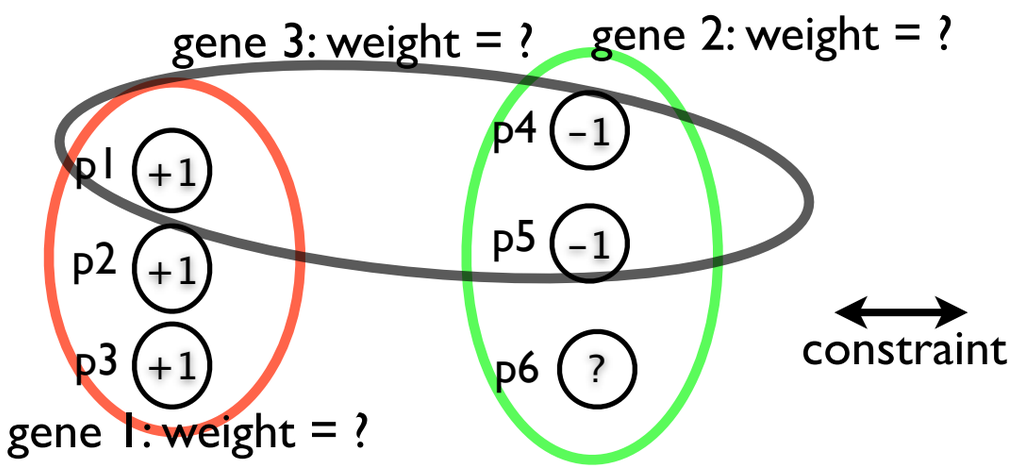
Cost 1: Highly connected samples should have the same label:

$$\Omega(f, w) = \frac{1}{2}\sum_{e\in E}\frac{w(e)}{d(e)}\sum_{u,v\in e}\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2$$

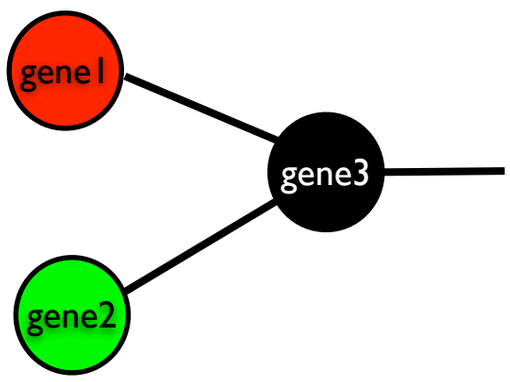
Cost 2: Supervised learning - the prediction should be consistent with initial labeling $\sum_i (f(u_i) - y(u_i))^2$

Cost 3: Genes that interact with each other should have similar weights:

$$\Psi(w) = \frac{1}{2}\sum_{i,j=1}^{|E|}\delta_{ij}\left(\frac{w(e_i)}{\sqrt{\sigma(e_i)}} - \frac{w(e_j)}{\sqrt{\sigma(e_j)}}\right)^2$$



constraint



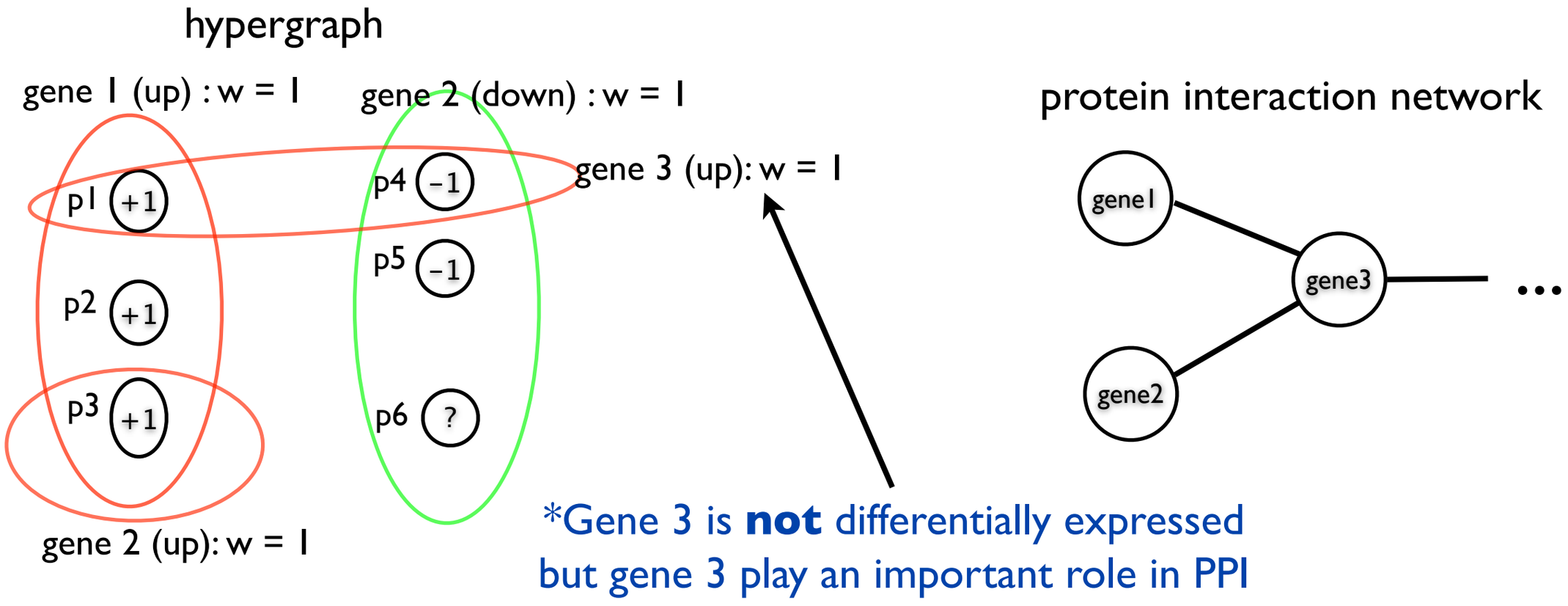
protein-protein interaction

Cost 1 and 2 **Cost 3**

$f(u)$: Predicted label of sample u
$y(u)$: Initial label of sample u
$w(e)$: Weight of hyperedge e
$d(u)$: Degree of sample u in hypergraph
$d(e)$: Degree of hyperedge e in hypergraph
$u, v \epsilon e$: Vertex u and v are connected by hyperedge e
$\sigma(e)$: Degree of hyperedge e in protein-protein interaction network
δ_{ij}: Interaction between hyperedge e_i and e_j

Working example

- Q: Classify patient 6, and identify biomarkers (two step iterative method)



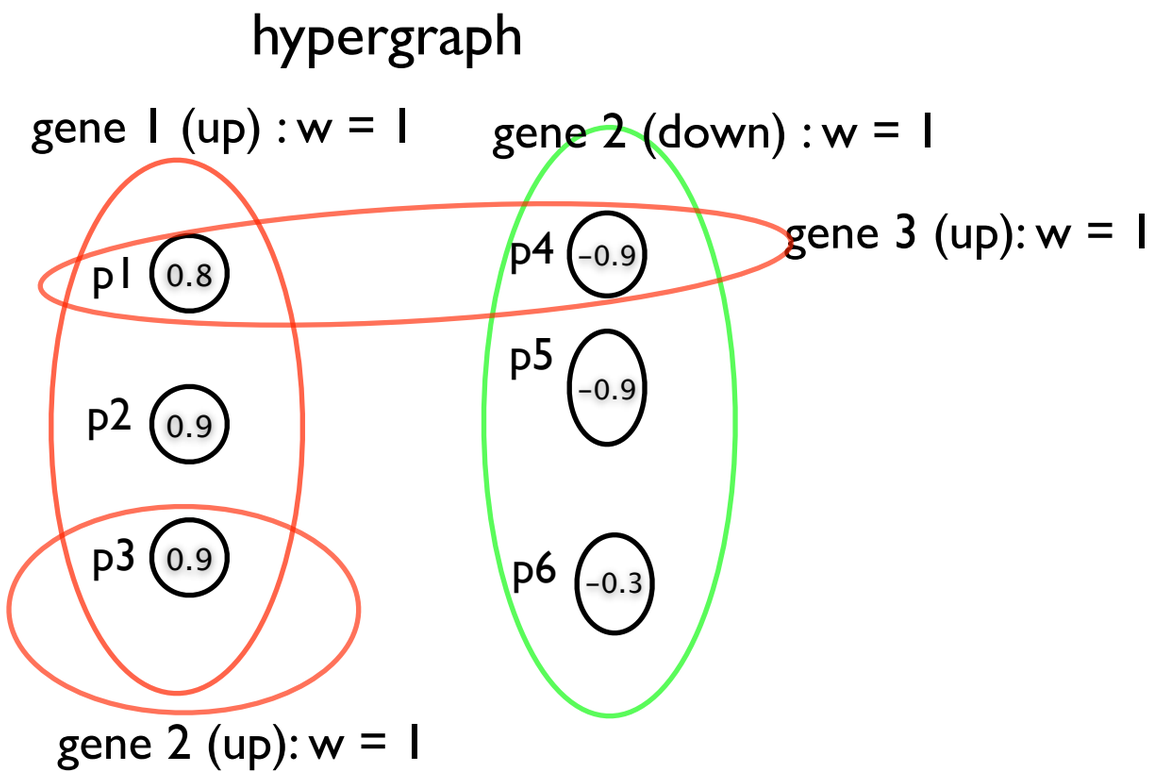
Working example

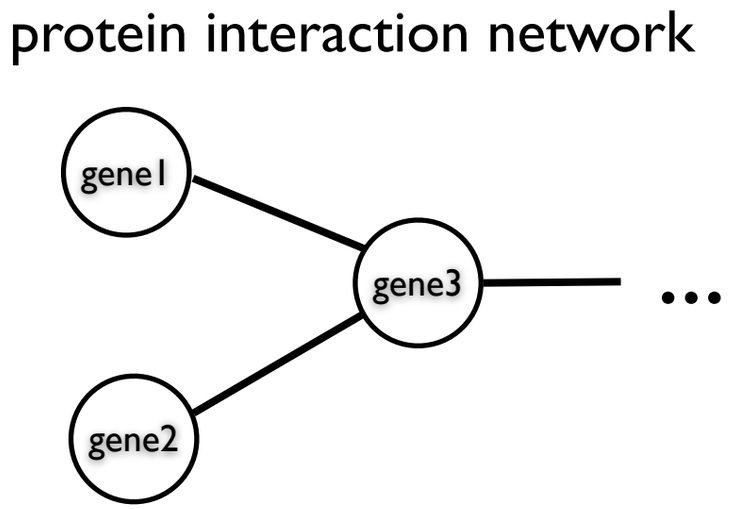
- Q: Classify patient 6, and identify biomarkers (two step iterative method)

1. **Sample classification: (initial weights of genes are uniform)**

   a) Highly connected samples should have same label

   b) The prediction should be consistent with initial labeling



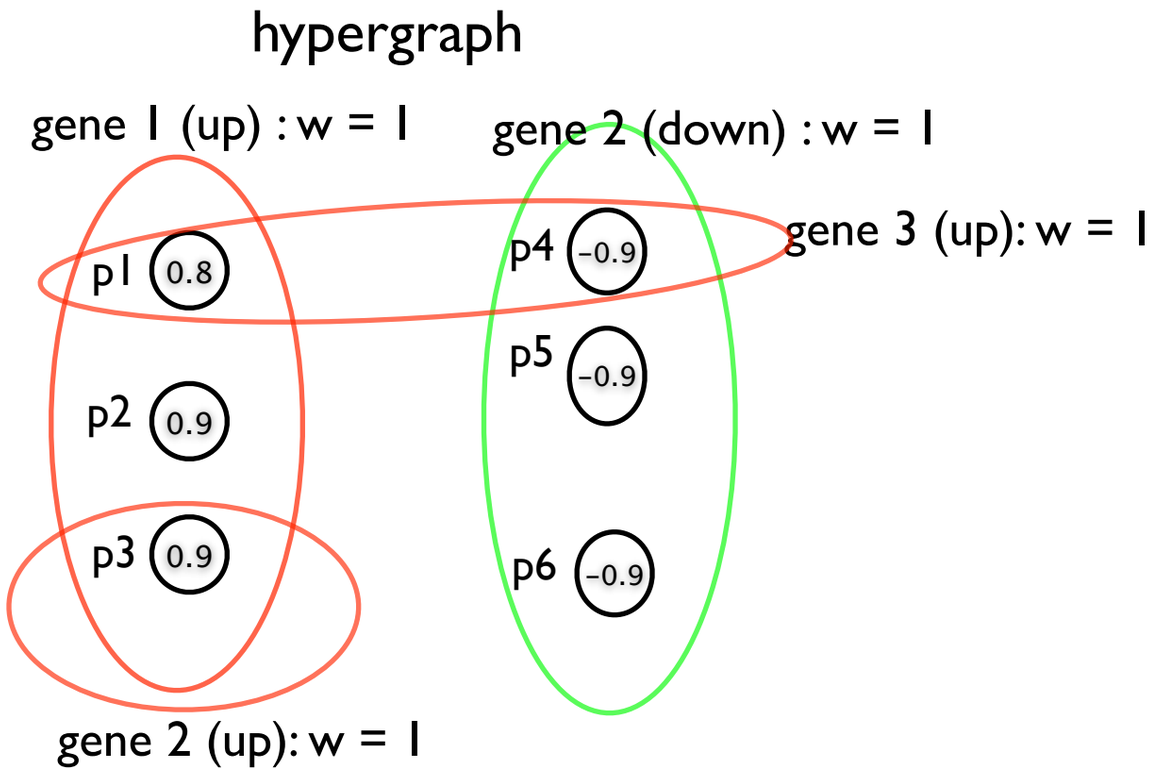


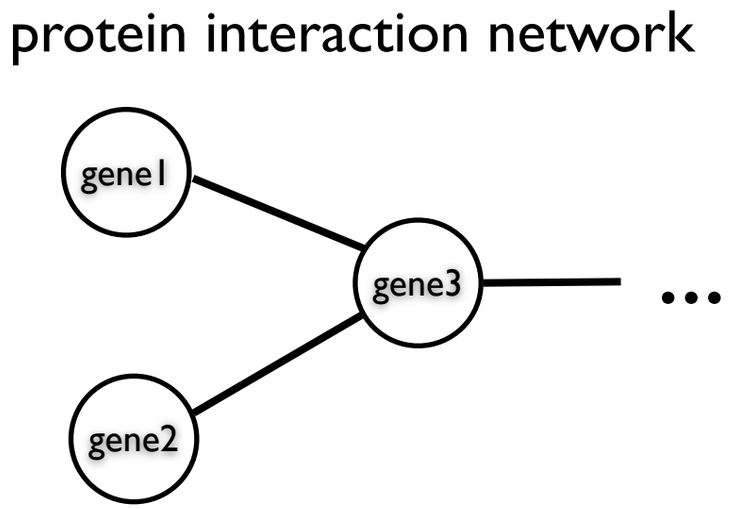
Working example

- Q: Classify patient 6, and identify biomarkers (two step iterative method)

1. Sample classification
2. **Learning weights of hyperedges**

   a) Fix current label information, and learn weights of hyperedges

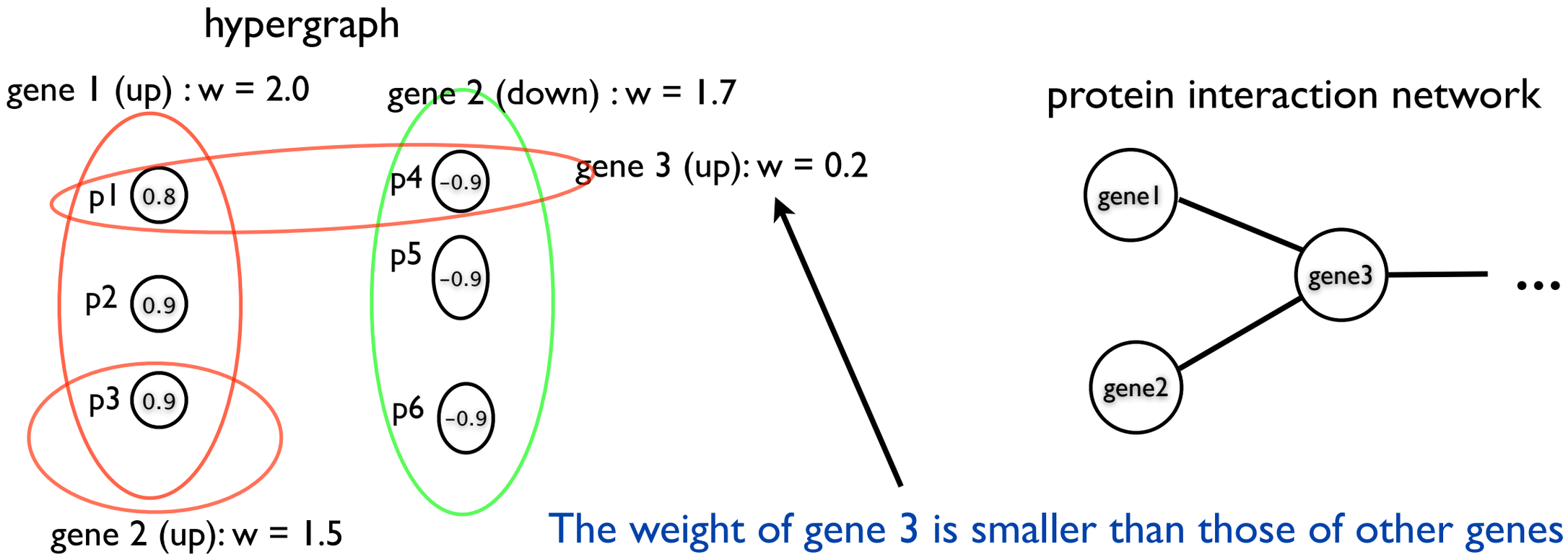




Working example

- Q: Classify patient 6, and identify biomarkers (two step iterative method)

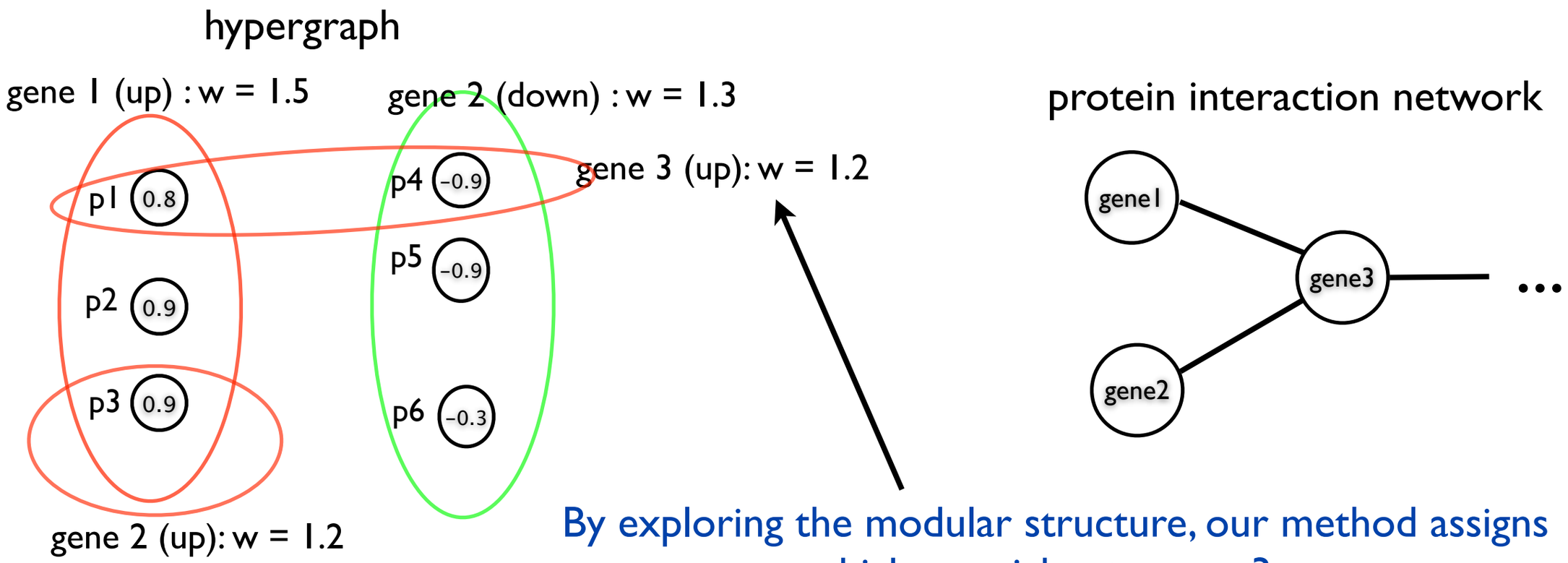
1. Sample classification
2. **Learning weights of hyperedges**
   a) Fix current label information, and learn weights of hyperedges
   b) Genes that interact with each other in should have similar weights



Working example

- Q: Classify patient 6, and identify biomarkers (two step iterative method)

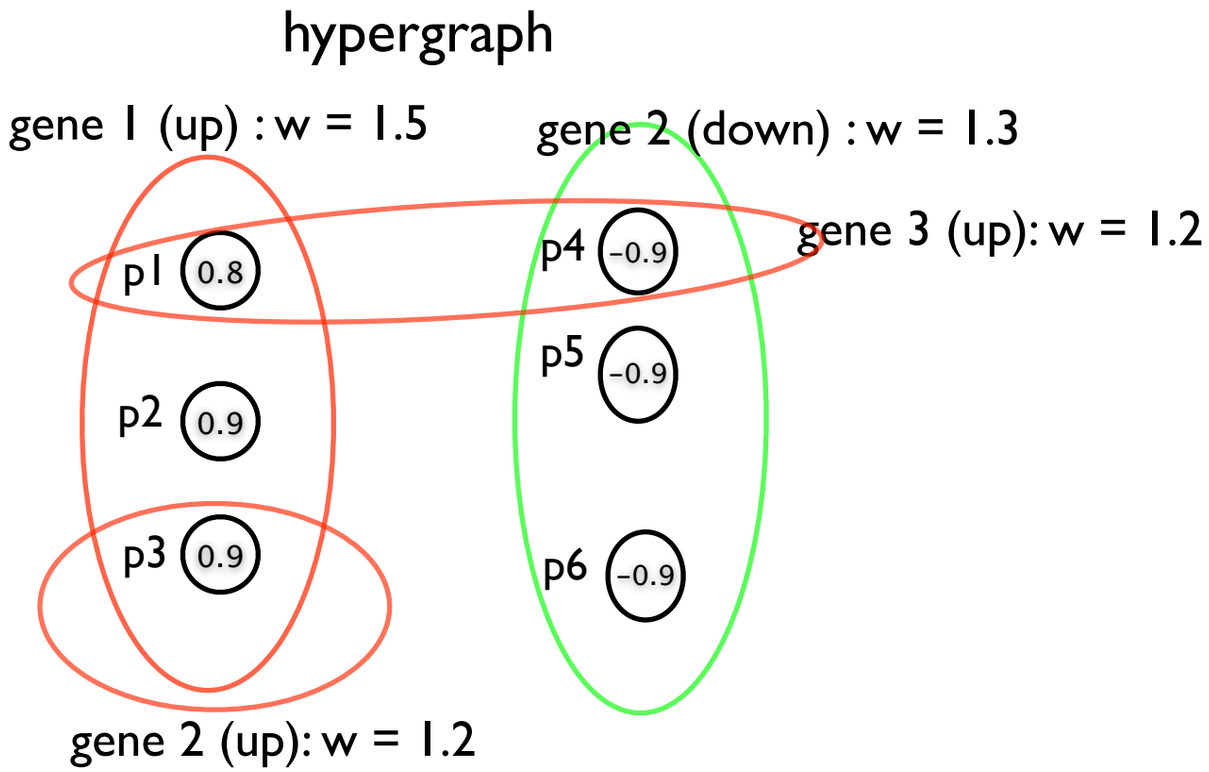
1. Sample classification
2. **Learning weights of hyperedges**
   a) Fix current label information, and learn weights of hyperedges
   b) Genes that interact with each other in should have similar weights

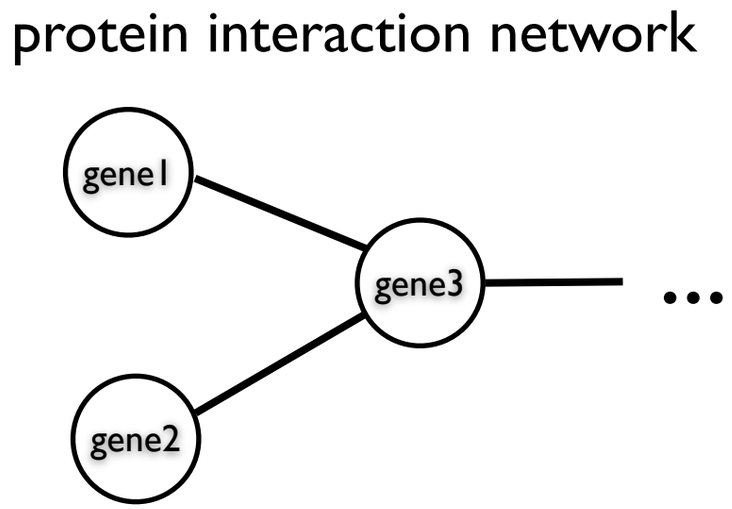


Working example

- Q: Classify patient 6, and identify biomarkers (two step iterative method)

1. Sample classification
2. Learning weights of hyperedges
3. Repeat step 1 and 2 until stopping criteria satisfies

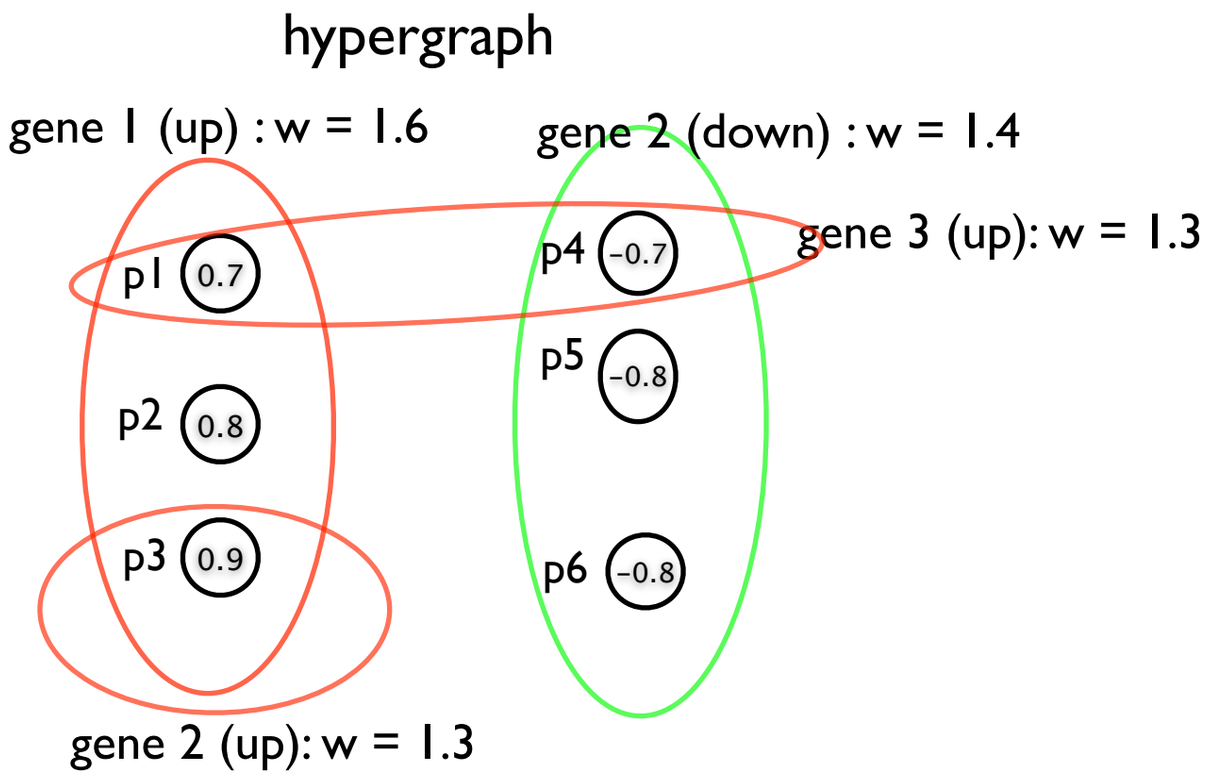


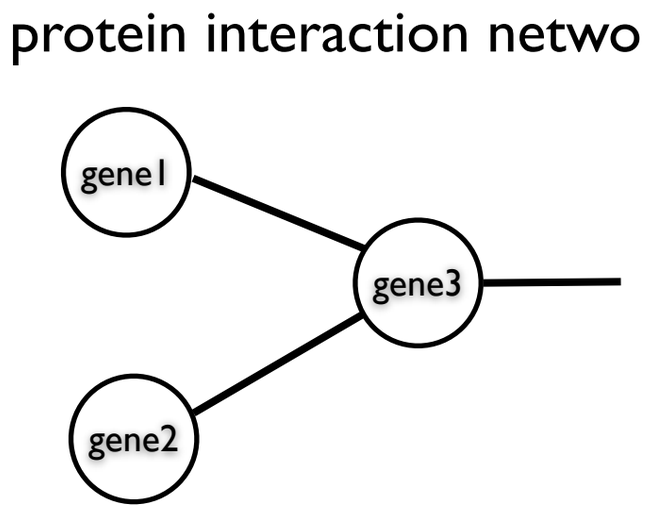


Working example

- Q: Classify patient 6, and identify biomarkers.

1. Sample classification
2. Learning weights of hyperedges
3. Repeat step 1 and 2 until stopping criteria satisfies
4. Rank hyperedges based on their weights: Highly ranked hyperedges can be potential biomarkers





T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Experiments 1

- Baselines
  - Support Vector Machines (SVMs) with linear and RBF kernels
  - Rapaport et al, BMC bioinformatics 2007
  - Li and Li, Bioinformatics 2008
  - Hypergraph
  - HyperPrior-LP
  - HyperPrior-NB

- Task
  - Cancer outcome prediction + Biomarker identification
- Dataset (Gene expression)
  - Two groups (metastasis vs non-metastasis)
  1. van't Veer et al, Nature 2002
     - 78 samples + 19 samples
  2. van de Vijver et al, New Engl. J. Med 2002
     - 295 samples (5 folds cross validation)
  3. Protein interaction networks

***The classification performance of all methods are evaluated using area under the receiver operating characteristics (ROC) score.**

T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Classification results

	van 't Veer *et al.*	van de Vijver *et al.*	
Algorithms	231 genes	326 genes	1464 genes
SVM (linear)	0.857	0.676	0.671
SVM (RBF)	0.857	0.681	0.667
Rapaport *et al.*	0.869	0.682	0.665
Li and Li	0.833	0.695	0.657
Hypergraph	0.857	0.687	0.685
HyperPrior-LP	**0.881**	**0.697**	**0.692**
HyperPrior-NB	0.869	**0.697**	**0.692**

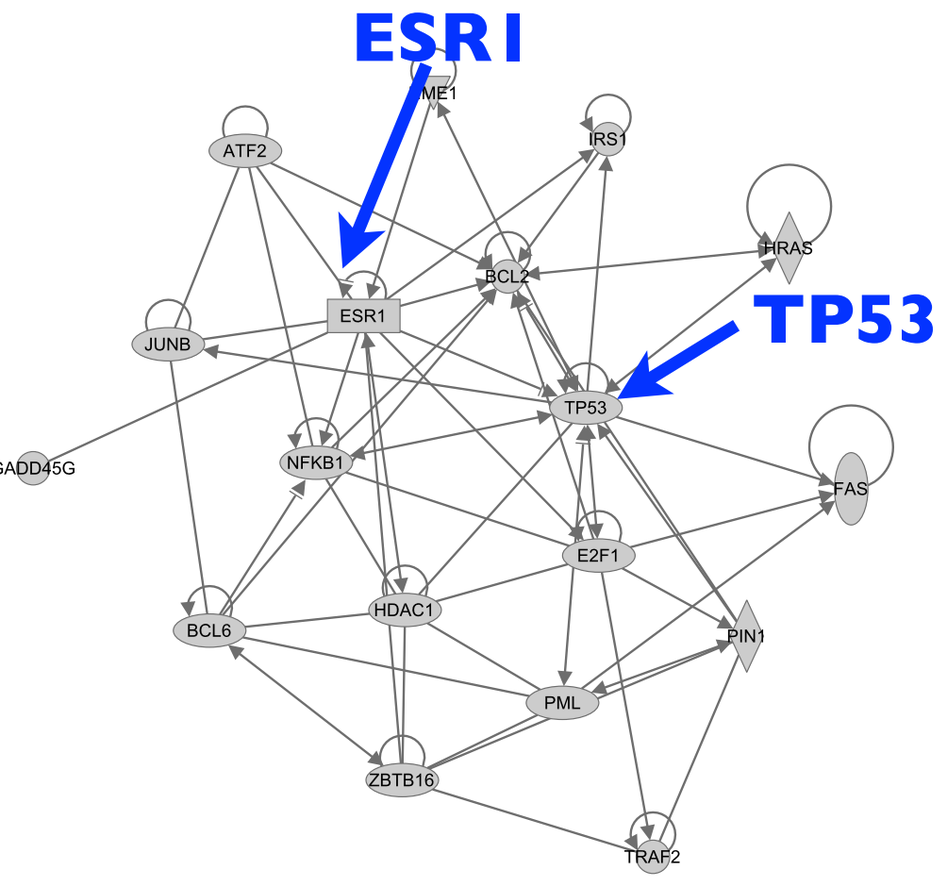
On the van 't Veer *et al.* dataset, the AUC on the 19-patient test set is reported. On the van de Vijver *et al.* dataset, over the random 5-fold cross-validations (50 times on both the 326 genes and the 1464 genes), the mean AUCs are reported.

*231 genes reported in van't Veer *et* al. are used.
*326 and 1,464 cancer related genes collected from *Ingenuity* and *Memorial Sloan Kettering Cancer Gene lists* are used in the second experiments

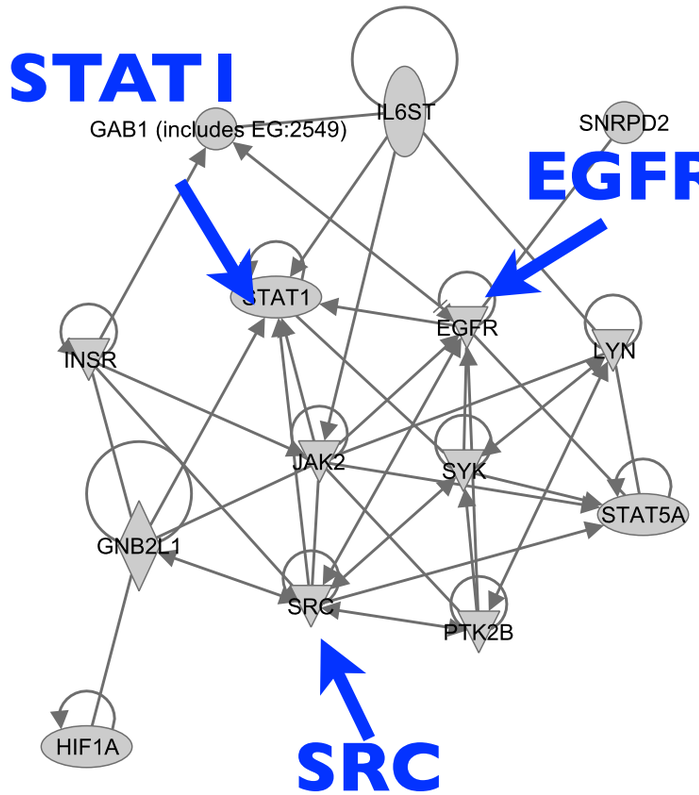
T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Subnetwork identification

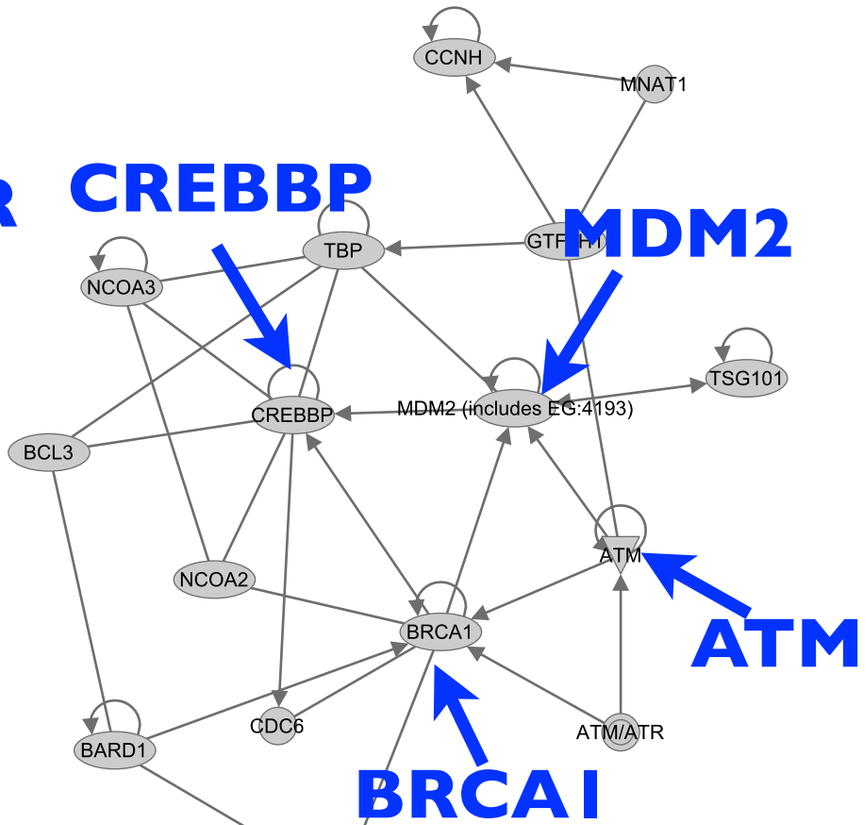
- Data integration can help to identify **breast cancer-related subnetworks**



TP53 subnetwork



STAT1 subnetwork



BRCA1 subnetwork

T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Biomarker discovery

Known Disease Gene	**Gene Ranking** HyperGene α=0.5, ρ=1	HyperGene α=0.5, ρ=0.001	CC
TP53	1	2	601
BRCA1	14	19	629
ESR1	17	22	208
BARD1	51	72	562
ATM	75	77	1054
HRAS	96	81	437
AKT1	99	154	1024
TGFB1	130	152	760
CASP8	142	201	1221
PTEN	157	198	725
PPM1D	182	60	266
KRAS	183	257	1267
SERPINE1	207	118	973
BRCA2	227	299	924
PIK3CA	415	363	712
STK11	632	609	773

The ranking of known breast cancer (OMIM#114480) susceptibility genes

T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Experiments 2

- Baselines
  - Support Vector Machines (SVMs) with linear and RBF kernels
  - L_1-Support Vector Machines (SVMs)
  - Rapaport et al, bioinformatics 2008 (Fused-SVM)
  - Hypergraph
  - HyperPrior-LP
  - HyperPrior-NB

- Task
  - Cancer outcome prediction + Biomarker identification
- Dataset (Copy number)
  - Two groups (by grade, stage, and metastasis)
    1. bladder tumor
       - 12 grade1 vs 45 grade 2&3
       - 16 stage T1 vs 32 stage T2+
    2. melanoma tumor
       - 35 metastasis vs 43 no-metastasis

***The classification performance of all methods are evaluated using area under the receiver operating characteristics (ROC) score.**

T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Classification results

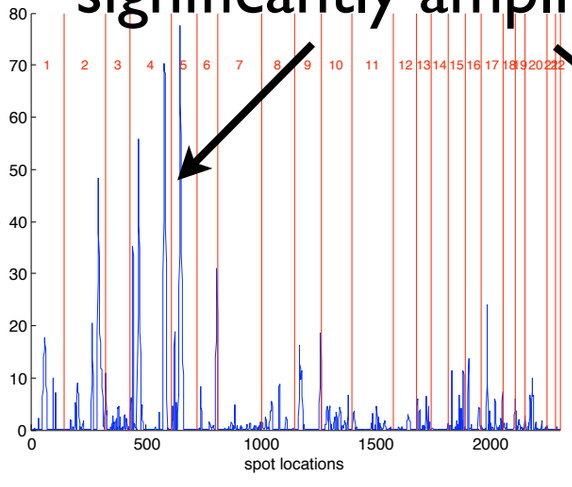
Table 1. Classification performance on arrayCGH data

LOO errors	SVM (linear)	SVM (RBF)	L_1-SVM	Fused SVM	Hypergraph	*HyperPrior-LP*	*HyperPrior-NB*
Bladder tumors (by grade)	9	9	12	7	11	**6**	**6**
Bladder tumors (by stage)	9	9	13	7	9	**5**	6
Melanoma tumors	10	10	8	**7**	**7**	**7**	**7**

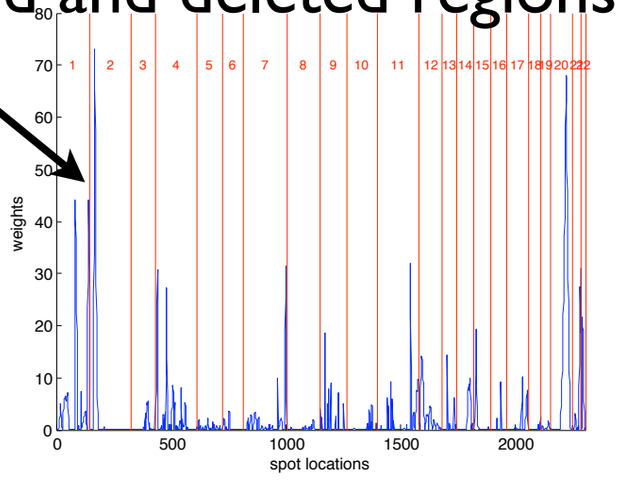
This table shows the number of misclassified samples in the LOO cross-validation on the bladder cancer dataset with two different labeling schemes (by tumor grade or by cancer stage) and the melanoma cancer dataset.

Our methods achieved overall best performances!

significantly amplified and deleted regions

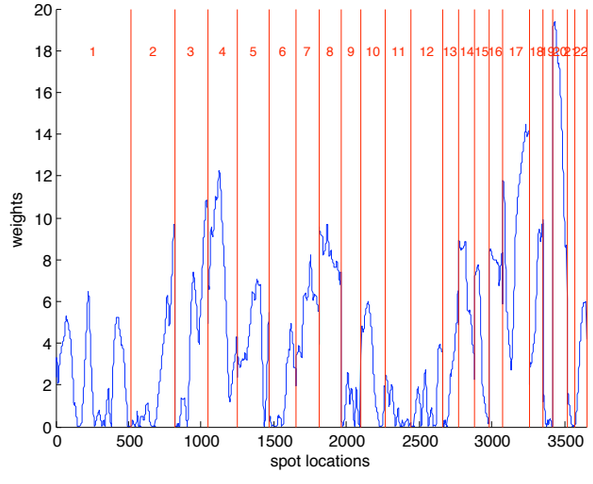


weights of DNA amplifications

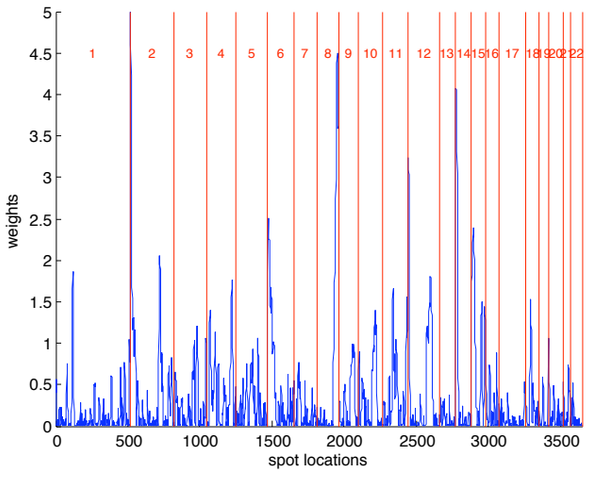


weights of DNA deletions

(A) Bladder cancer



weights of DNA amplifications



weights of DNA deletions

(B) Melanoma cancer

Our methods found cancer-related copy number regions!

T. Hwang et. al, **ICDM** 2008
Z. Tian, **T. Hwang**, R. Kuang, **Bioinformatics** 2009

Take home message

- Our proposed method that integrates genomic data with biological prior knowledge can help to improve cancer outcome prediction and discover cancer-related subnetworks in breast cancer
- Our proposed method also found cancer-related copy number variations with aCGH data experiments in melanoma and bladder cancer
- One should be careful to interpret results from network-based methods

- **T. Hwang***, Z. Tian*, JP Kocher, R. Kuang, “Learning on Weighted Hypergraphs for Integrating Protein Interactions and Gene Expressions”, IEEE International Conference on Data Mining, **ICDM 2008**
- Z. Tian*, **T. Hwang***, and R. Kuang. “A Hypergraph-based Learning Algorithm for Classifying Gene Expression and arrayCGH data with Prior Knowledge”, **Bioinformatics 2009**
- Z. Tian*, **T. Hwang***, and R. Kuang. “A Hypergraph-based Learning Algorithm for Classifying arrayCGH data with Spatial Prior Knowledge”, Proc. of IEEE International Workshop on Genomic Signaling Processing and Statistics, **GENSIPS 2009**

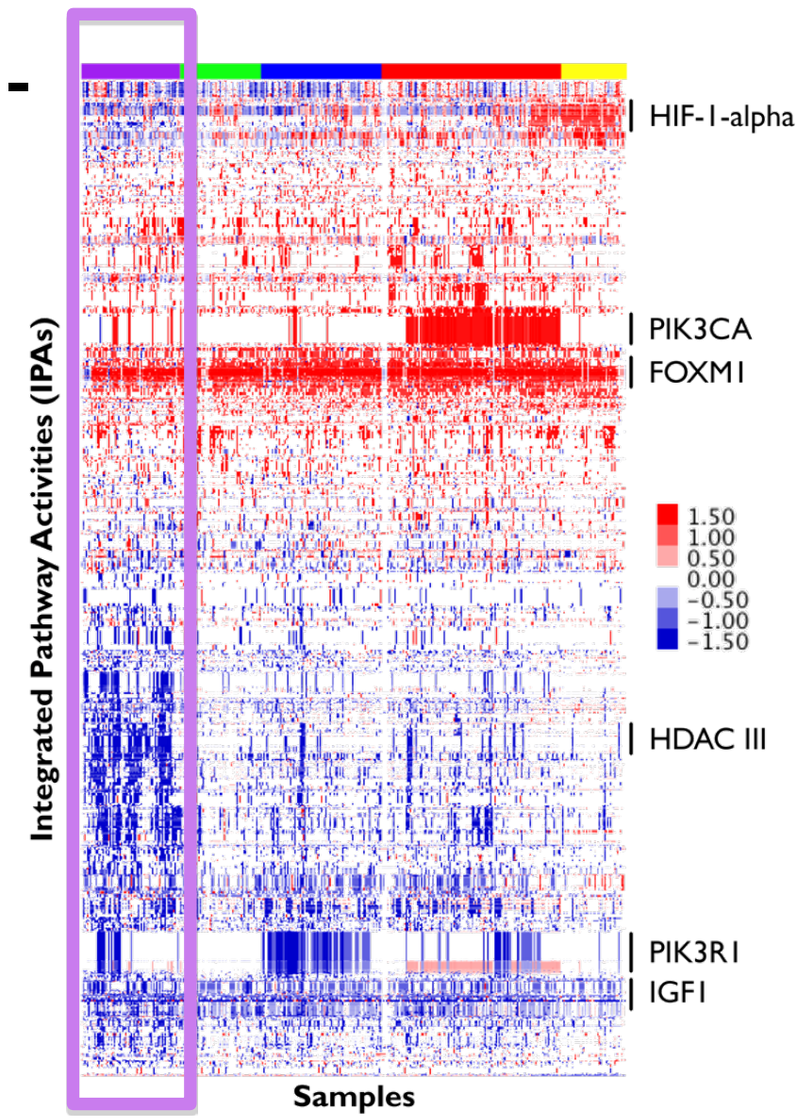
***Joint first author**

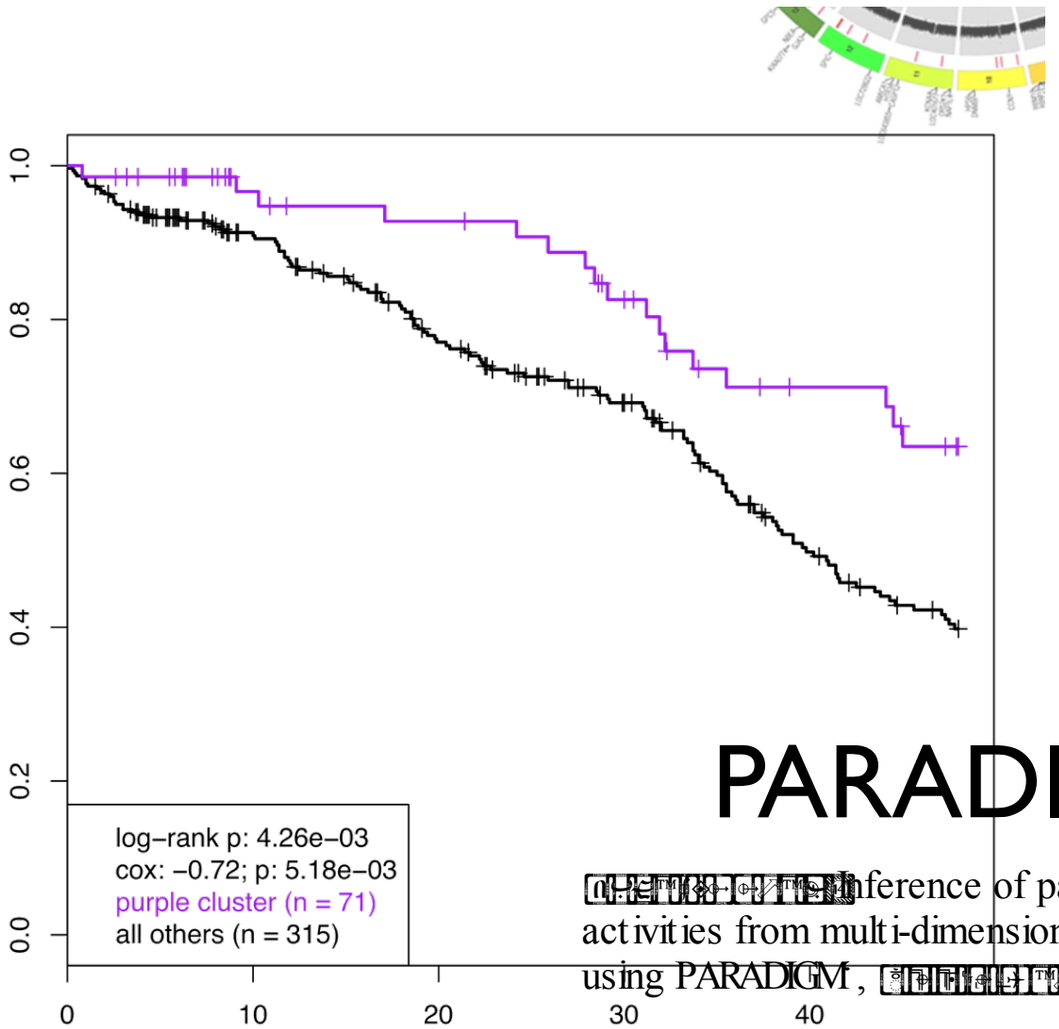
Network/pathway based methods for patient stratification

Motivation

- Somatic mutation, and copy number alternations (CNAs) at the distinct loci of the human genome may contribute to the development of cancers

- The systematic characterization of disrupted pathways by genomic alterations in human cancer can help to establish the refined genetic landscape of cancer



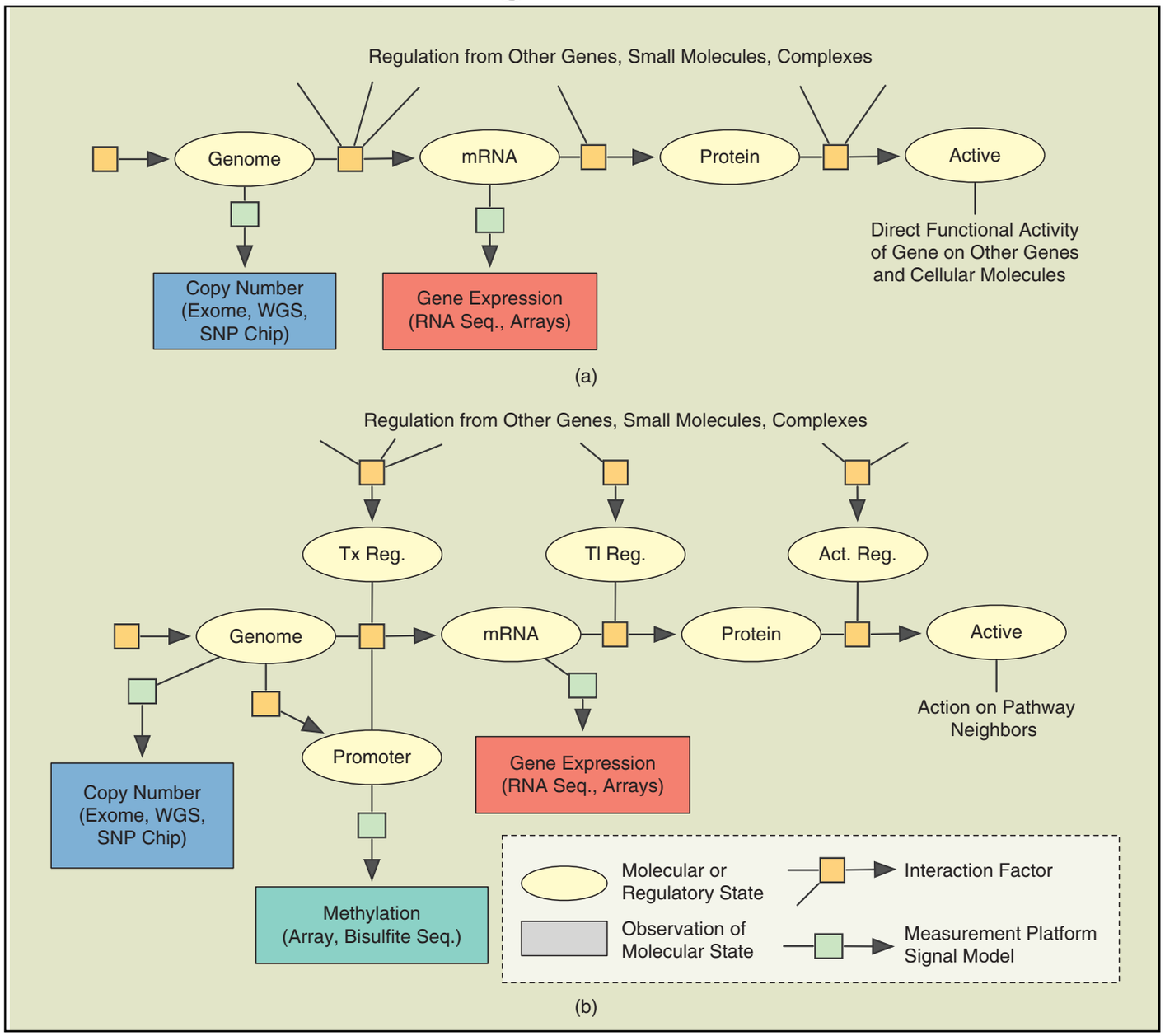


PARADIGM

[redacted] inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM , [redacted]

PARADIGM

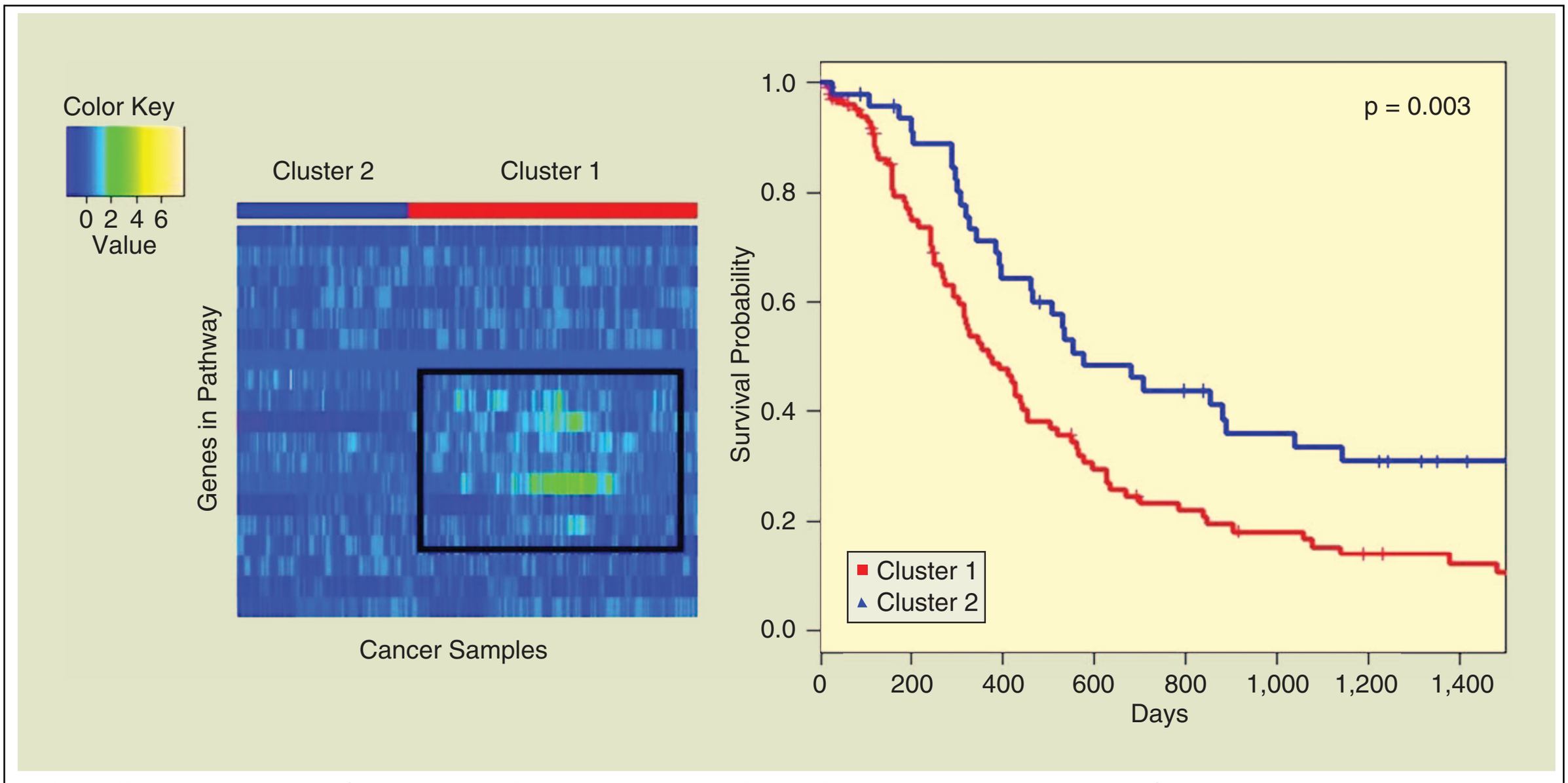
- Given: genomic data (e.g., mutation, copy number, gene expression and etc), and pathway
- Task: Identify pathway activity of patient
  - Input: genomic data and pathway
  - Output: pathway activity (e.g., active or inactive)



Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM,

PARADIGM

- Pathway activities could be used to identify patient subgroups having different survival outcome
- Pathway activities could guide a clinical decision for efficient therapy



[FIG11] Clustering analysis of ovarian samples using Paradigm shows that patients with lower IPAs for genes in the E-cadherin adherens junction pathway are associated with better response to platinum therapy ($p = 0.003$).

Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM,

PARADIGM

- Pathway activities could be used to identify patient subgroups having different survival outcome
- Pathway activities could guide a clinical decision for efficient therapy

Limitation
- rely on existing pathway database
  - current knowledge of pathway is still incomplete (~4000 genes annotated with current existing pathway database)
- need to use independent algorithms to cluster patient samples
  - step 1) identify pathway activities
  - step 2) use pathway activities to discover patient subgroups

[illegible] inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, [illegible]

HotNet

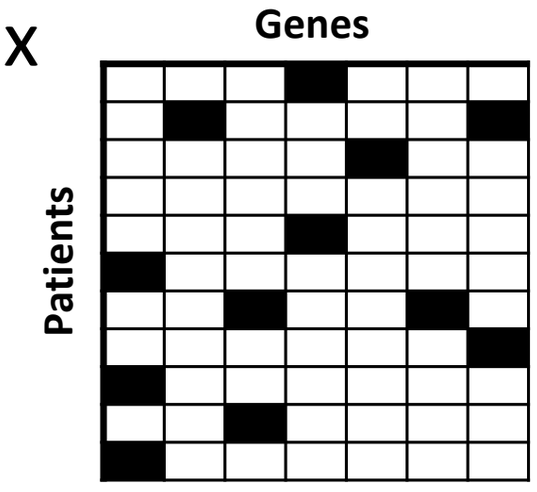
- Given: genomic data (e.g., mutation, or copy number), and protein-protein interaction networks
- Task: Identify significantly mutated subnetworks
  - Input: genomic data and protein interaction networks
  - Output: subnetwork

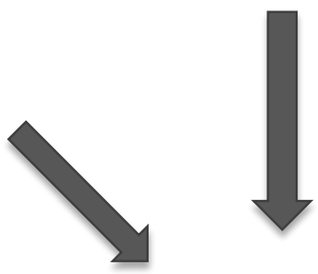
Given:

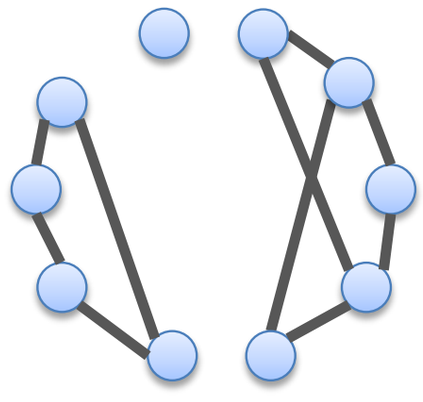
1. Network $G = (V, E)$

   V = genes. E = interactions b/w genes

2. Binary mutation matrix







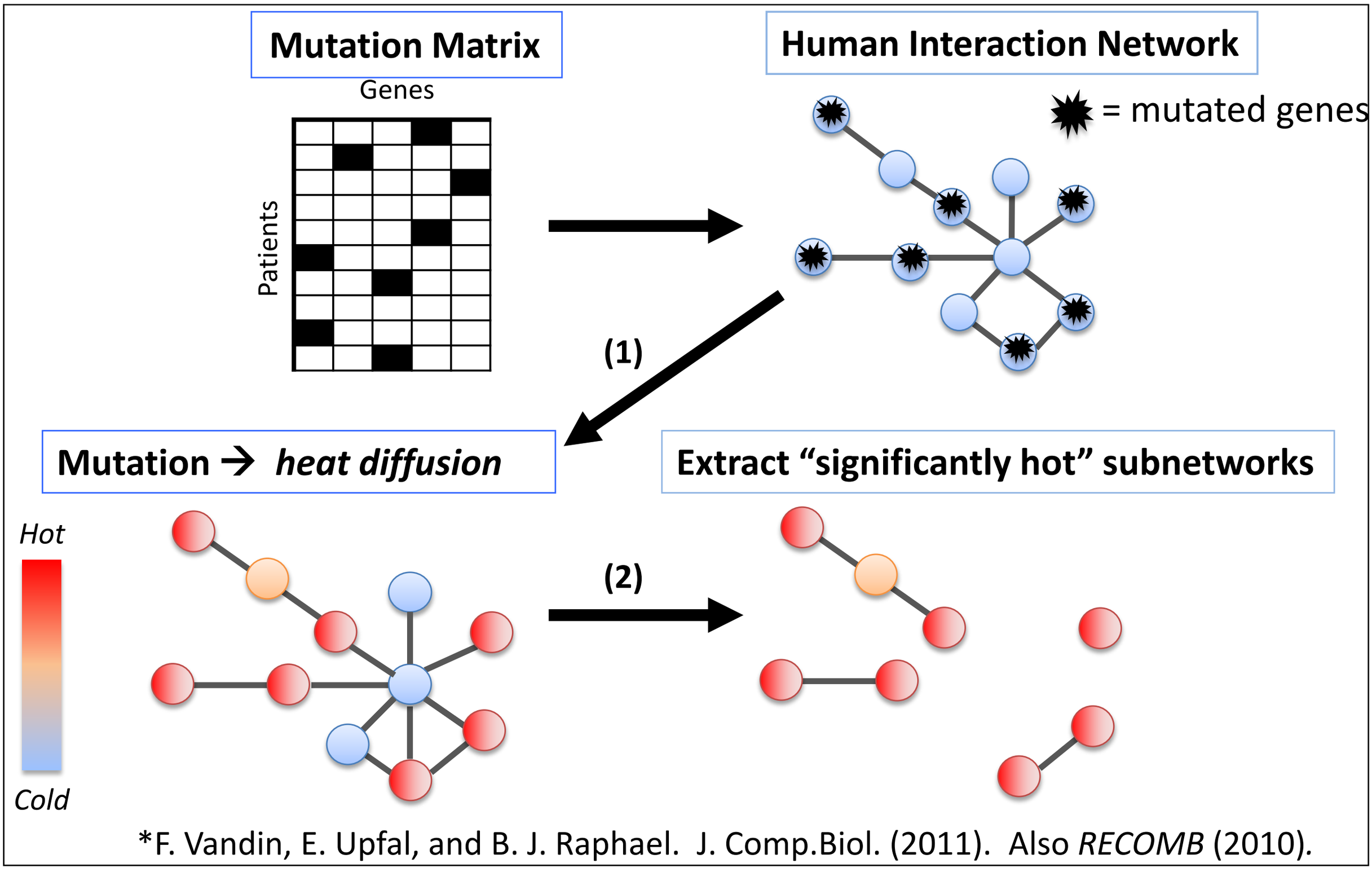
Find: *Connected subnetworks* mutated in a significant number of patients

- *mutated* in patient if ≥ 1 gene mutated in patient

slide courtesy of Dr. Ben Raphael

HotNet

- Workflow



*F. Vandin, E. Upfal, and B. J. Raphael. J. Comp.Biol. (2011). Also *RECOMB* (2010).

slide courtesy of Dr. Ben Raphael

HotNet

- Results

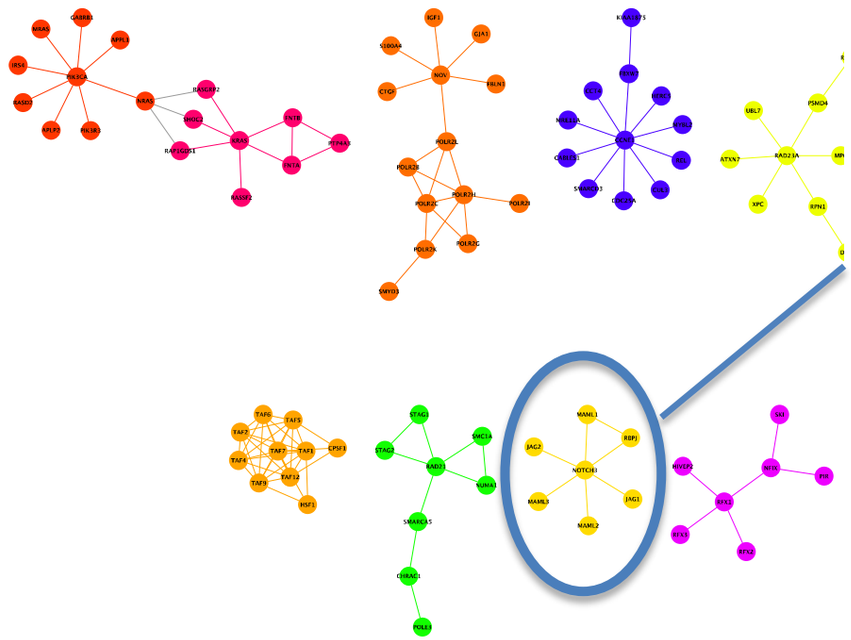
Ovarian Subnetworks

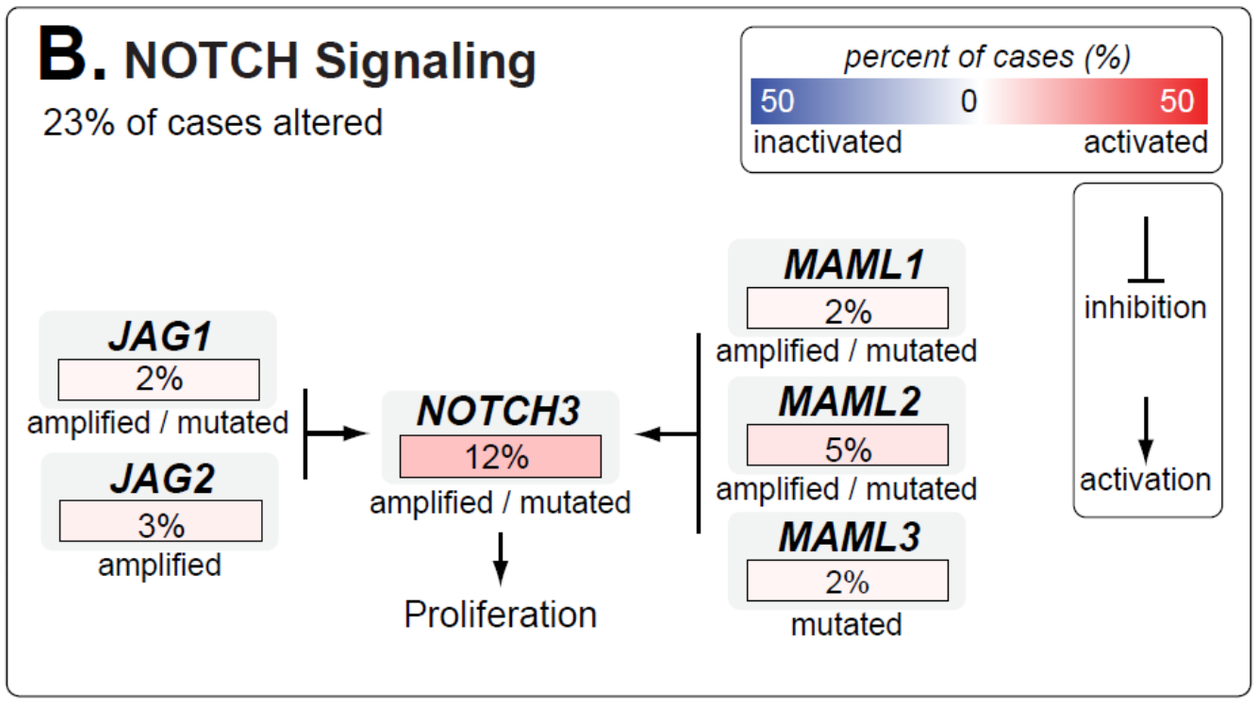
ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*





Kegg Pathway

Notch signaling ($p < 6x10^{-7}$)

12/27 subnetworks significantly overlap known pathways (KEGG) or protein complexes (PINdb)

TCGA. *Nature* (2011)

slide courtesy of Dr. Ben Raphael

HotNet

- Results

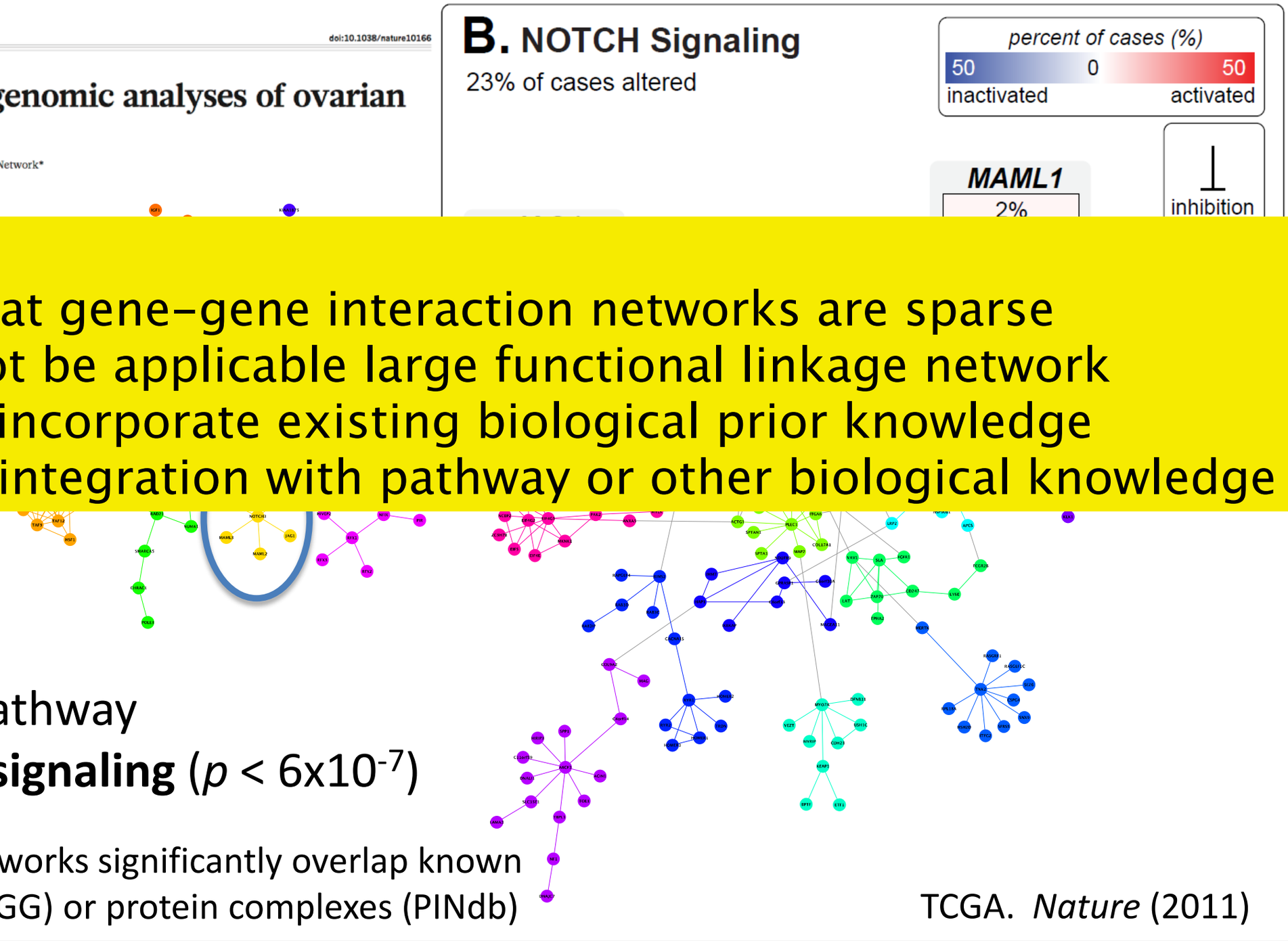
Ovarian Subnetworks

ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*



Kegg Pathway

Notch signaling ($p < 6x10^{-7}$)

12/27 subnetworks significantly overlap known pathways (KEGG) or protein complexes (PINdb)

TCGA. *Nature* (2011)

Limitation

- assume that gene-gene interaction networks are sparse
  - could not be applicable large functional linkage network
- could not incorporate existing biological prior knowledge
  - no data integration with pathway or other biological knowledge

slide courtesy of Dr. Ben Raphael

MEMo

- Workflow

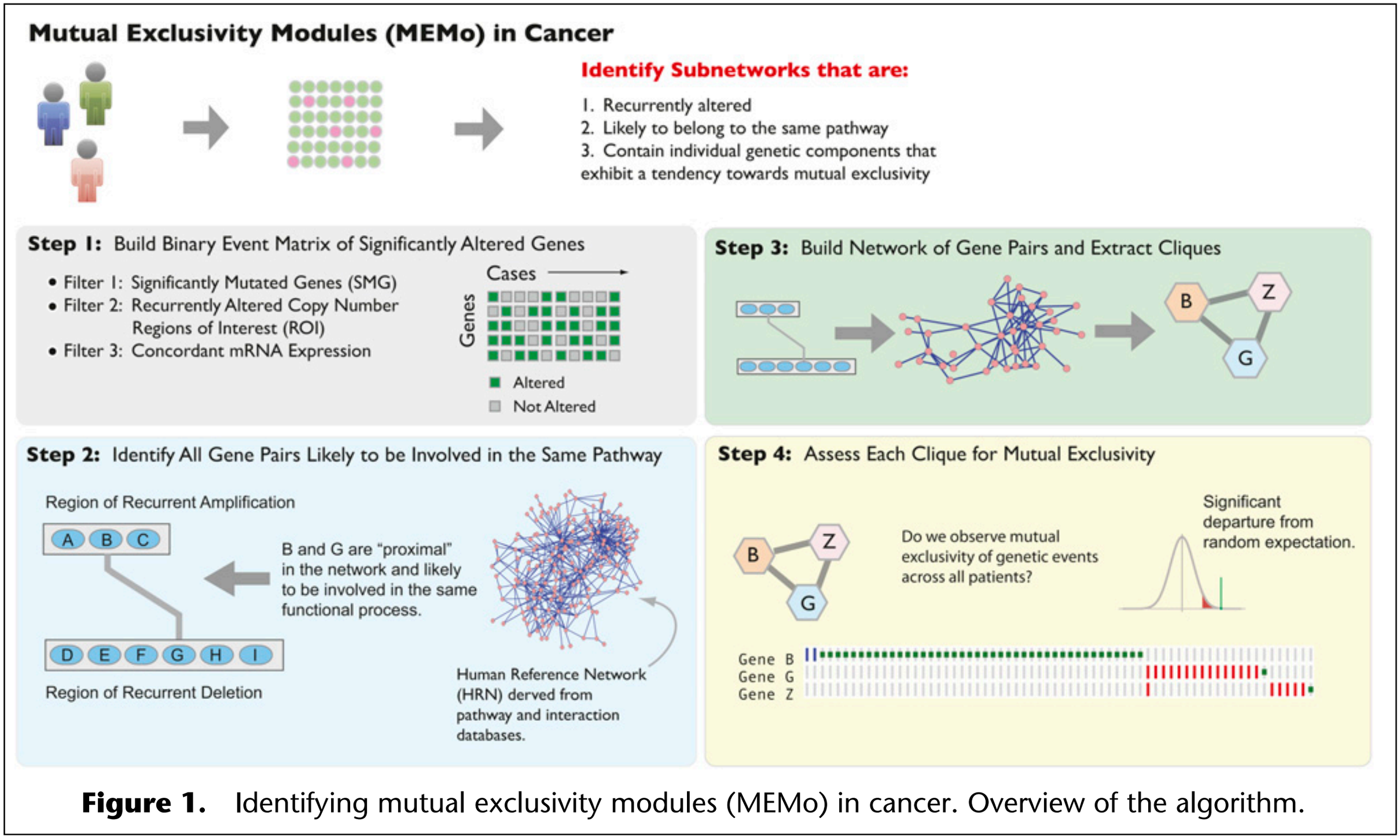


Figure 1. Identifying mutual exclusivity modules (MEMo) in cancer. Overview of the algorithm.

Ciriello et. al., "Mutual exclusivity analysis identifies oncogenic network modules", Genome research 2012

MEMo

- Results

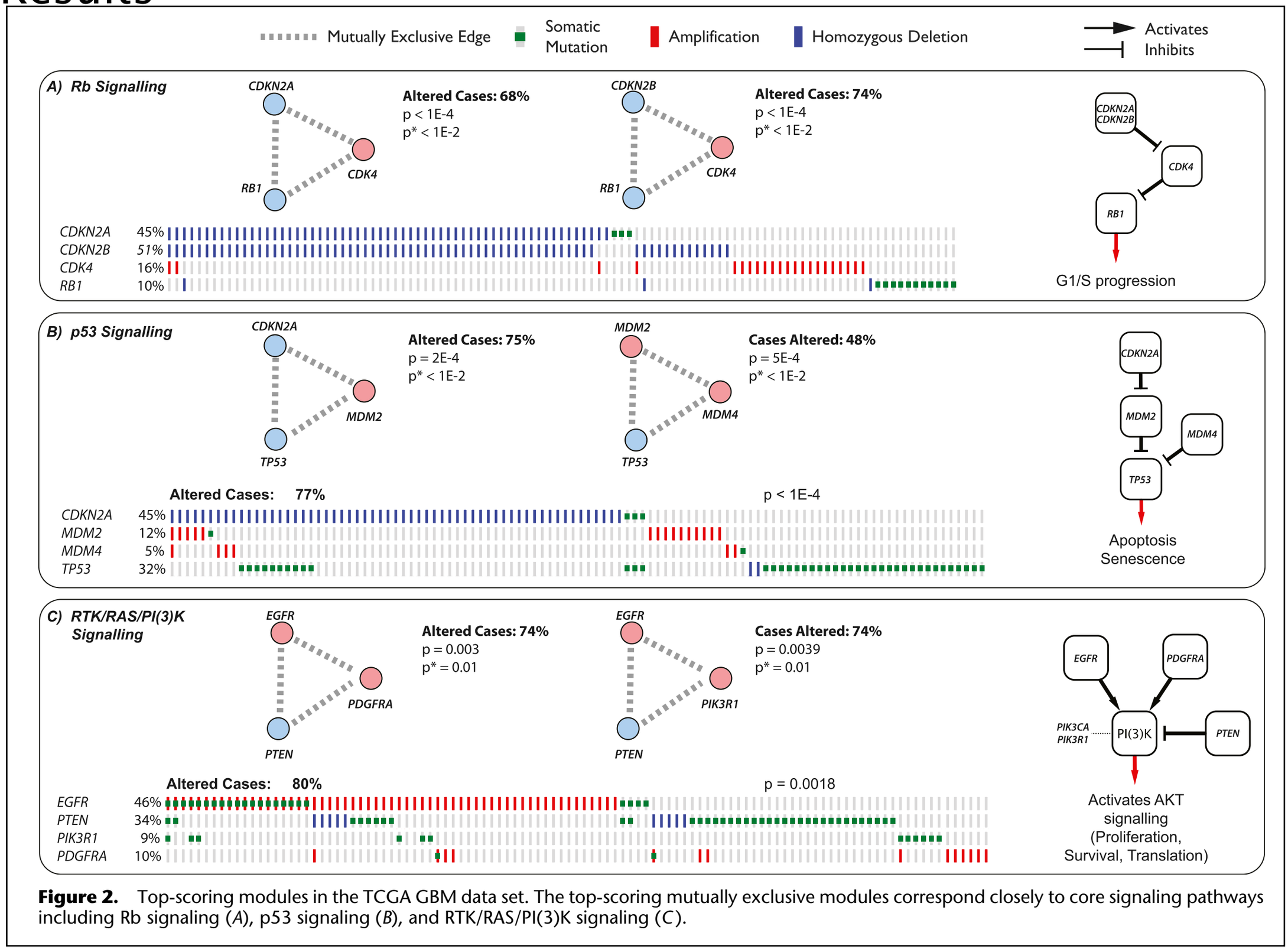
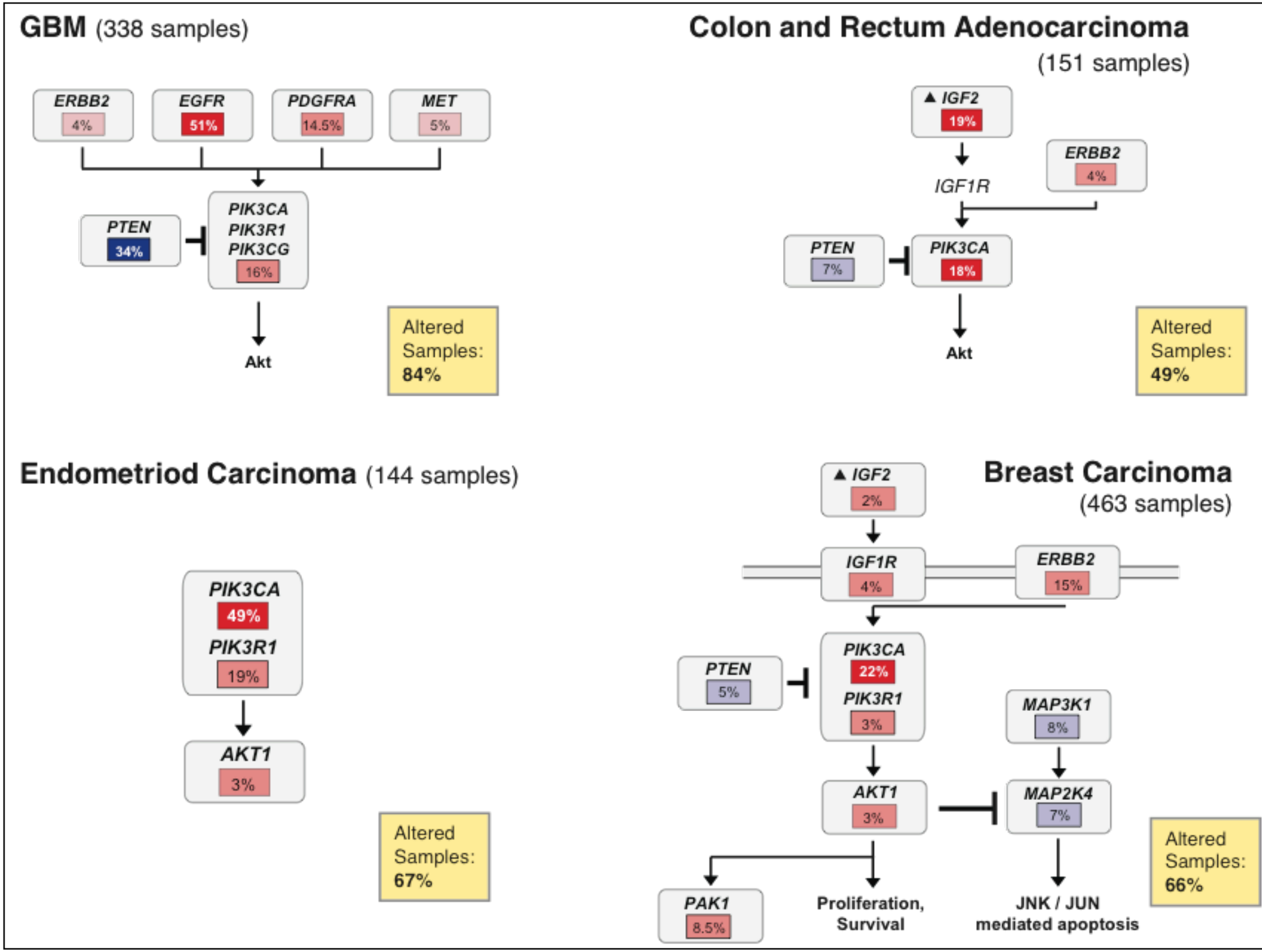


Figure 2. Top-scoring modules in the TCGA GBM data set. The top-scoring mutually exclusive modules correspond closely to core signaling pathways including Rb signaling (*A*), p53 signaling (*B*), and RTK/RAS/PI(3)K signaling (*C*).

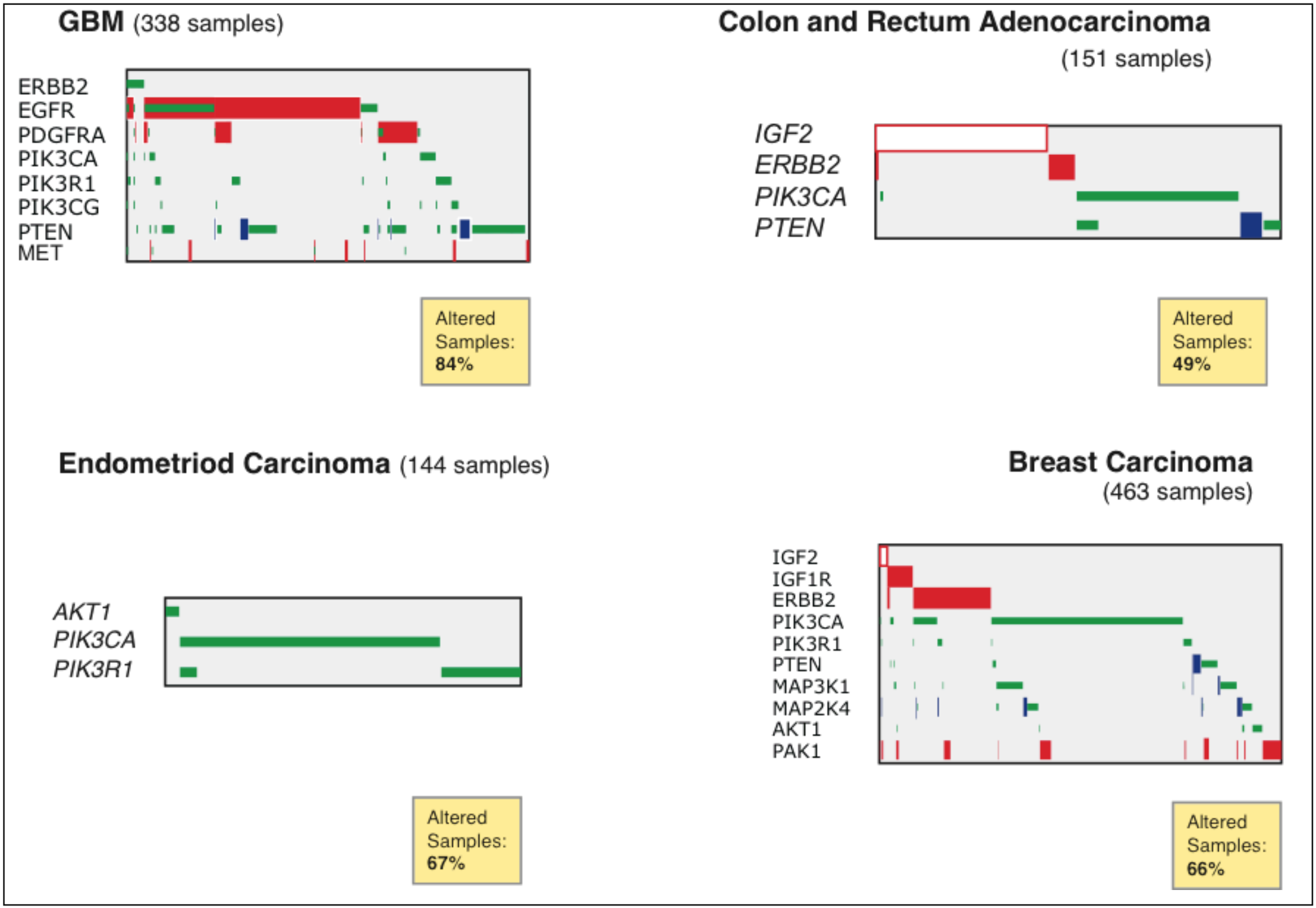
Ciriello et. al., "Mutual exclusivity analysis identifies oncogenic network modules", Genome research 2012

MEMo



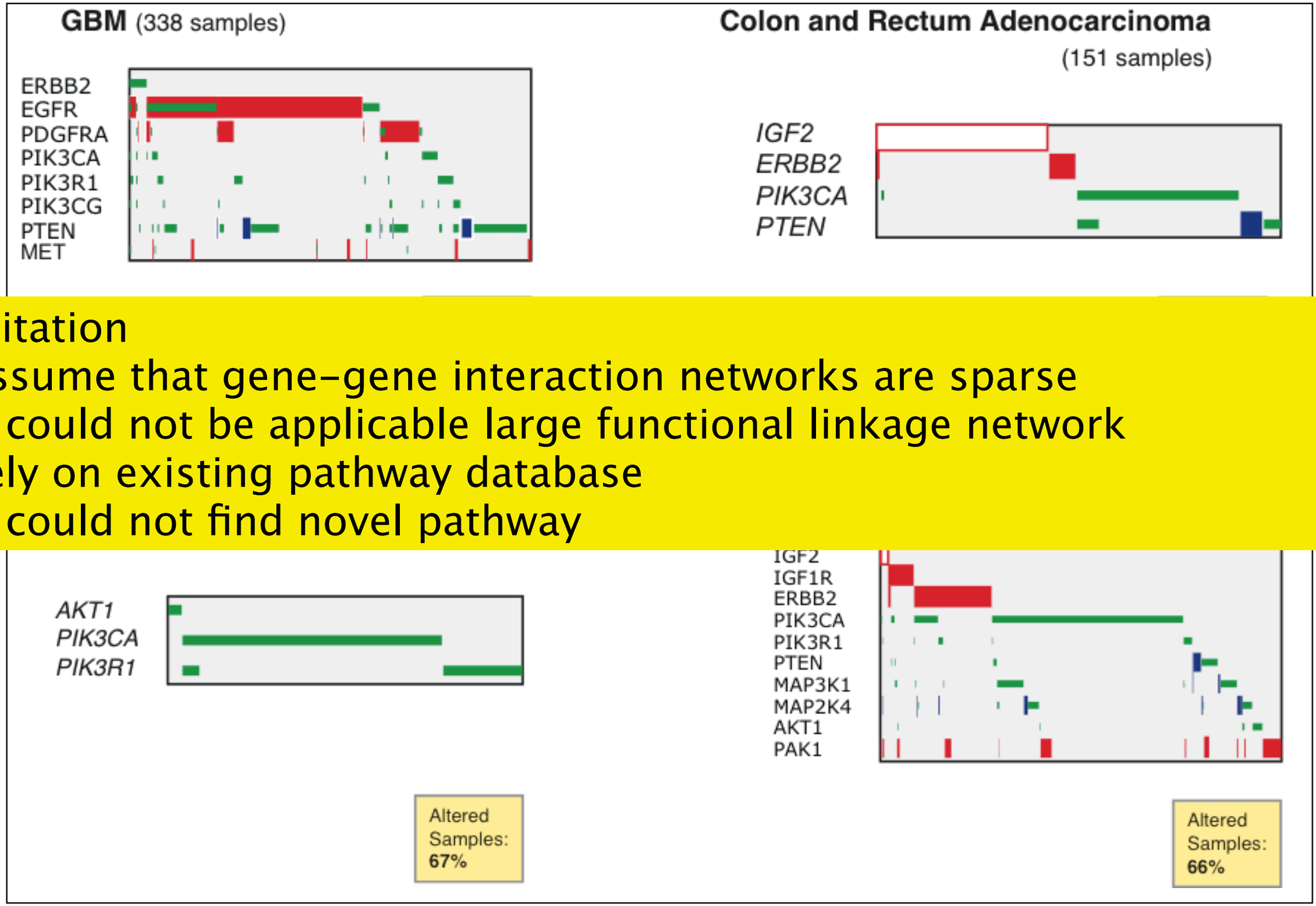
slide courtesy of Dr. Giovanni Ciriello

MEMo



slide courtesy of Dr. Giovanni Ciriello

MEMo



Limitation
- assume that gene-gene interaction networks are sparse
  - could not be applicable large functional linkage network
- rely on existing pathway database
  - could not find novel pathway

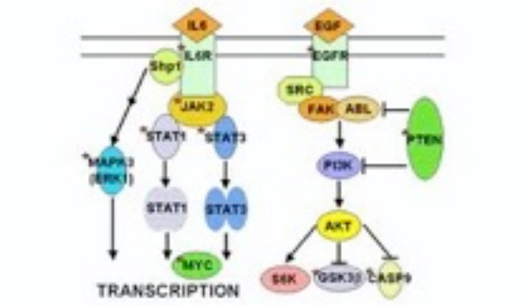
slide courtesy of Dr. Giovanni Ciriello

Patient stratification using genomic and pathway data integration

- **Develop novel computational methods to integrate genomic data with biological prior knowledge**

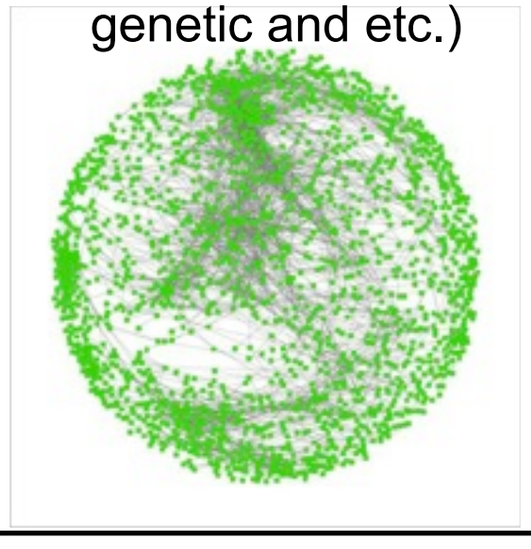
Input

Biological prior knowledge (e.g, pathway, literature, and etc)



Genomic data (e.g., gene expression, copy number, mutation, methylation and etc.)

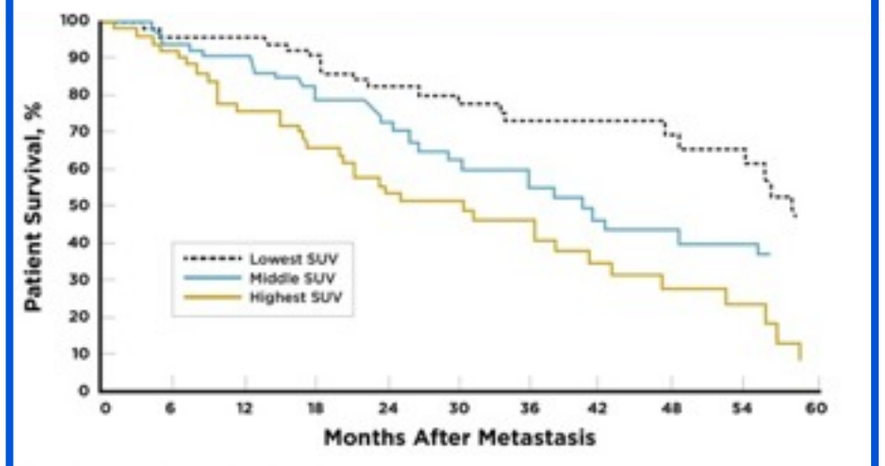
Gene-gene interaction (e.g., physical,coexpression, genetic and etc.)



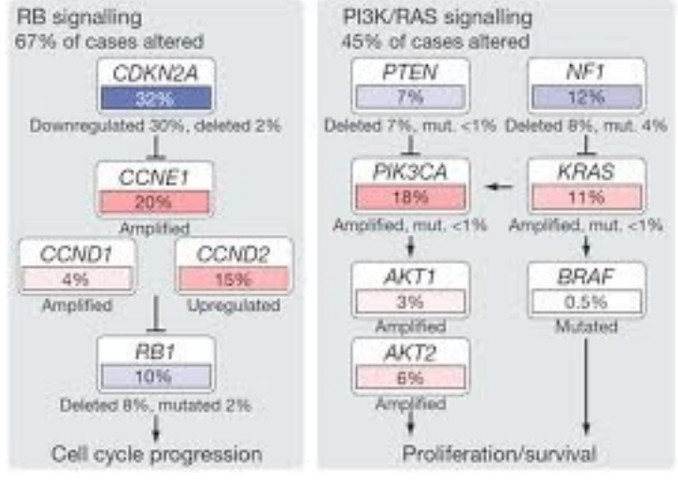
our integrated method

Output

Patient subgroup identification



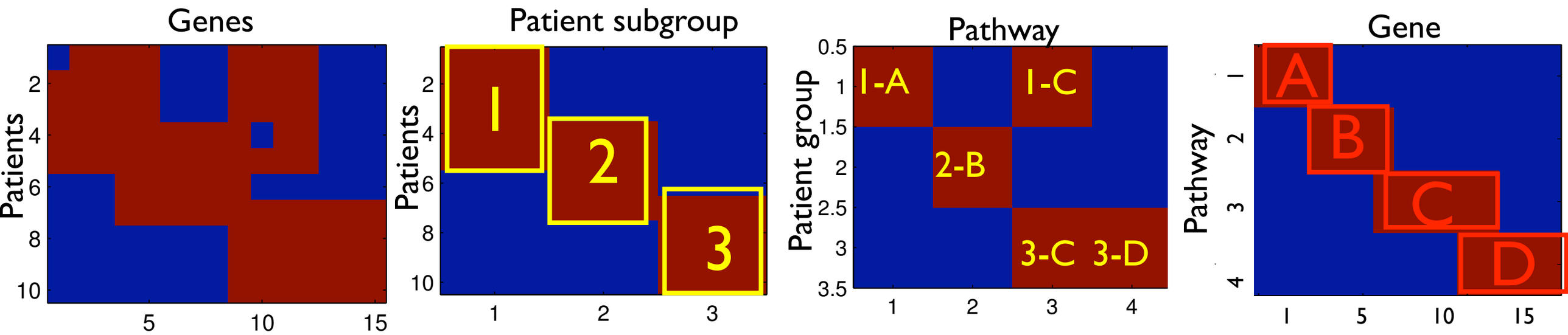
Pathways related with patient subgroup



SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., **Nucleic Acids Research** 2012

Matrix tri-factorization

- Given: Gene expression and pathway data
- Task : Identify patient subgroups and pathway activities related with patient subgroups



Gene expression | Patient subgroup | Patient subgroup-pathway association | Pathway

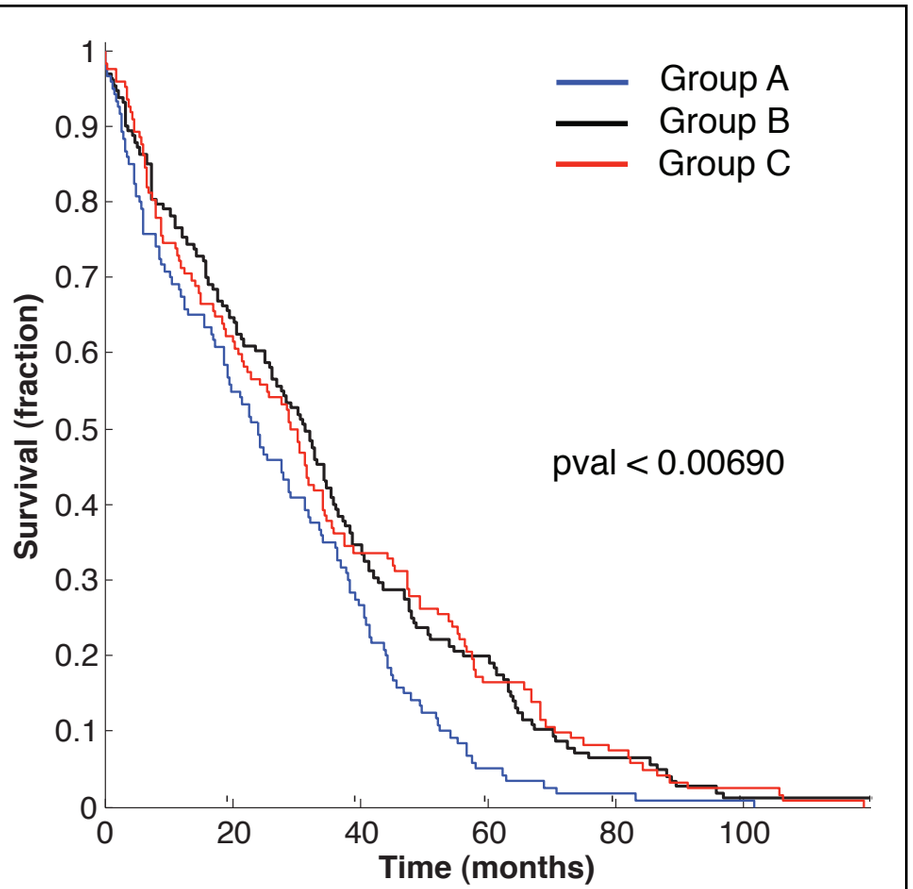
$$\min_{F,\,S} \quad \frac{1}{2}\|\mathbf{X}-\mathbf{F}\mathbf{S}\mathbf{G}^T\|_F^2 + \lambda_F\|\mathbf{F}\|_1^2 + \lambda_S\|\mathbf{S}\|_1^2$$

X: gene expression data
F: patient subgroups
S: patient subgroup–pathway association
G: pathway

SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., **Nucleic Acids Research** 2012

Experiments (TCGA)

- TCGA Ovarian Carcinoma: 377 patients with clinical data
- Gene Expression: 11,864 mRNA expression
- Pathway: KEGG pathway (186 pathways)



(a) microarray gene expression

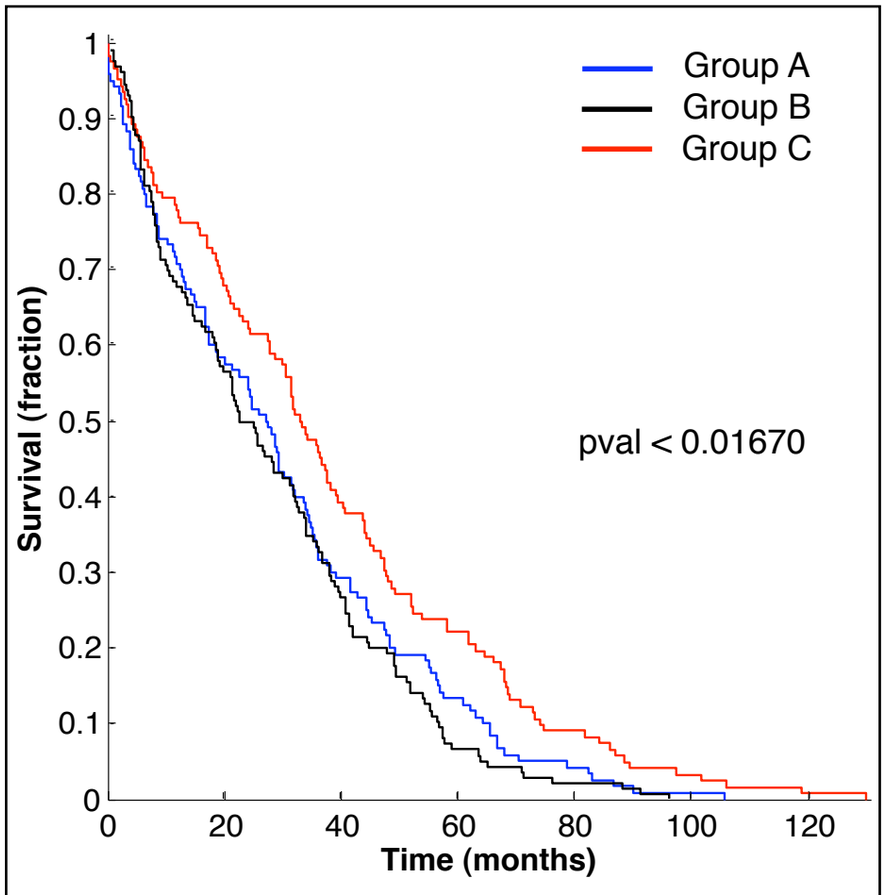
Ranking	Pathway (Microarray gene expression)
1	KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
2	KEGG COMPLEMENT AND COAGULATION CASCADES
3	KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION
4	KEGG CELL ADHESION MOLECULES CAMS
5	KEGG PATHWAYS IN CANCER
6	KEGG PURINE METABOLISM
7	KEGG CHEMOKINE SIGNALING PATHWAY
8	KEGG HEMATOPOIETIC CELL LINEAGE
9	KEGG MAPK SIGNALING PATHWAY
10	KEGG TGF BETA SIGNALING PATHWAY

✓ Matrix tri-factorization can accurately identify patient subgroups having different survival outcome and pathways associated with patient subgroups

SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., **Nucleic Acids Research** 2012

Experiments (TCGA)

- TCGA Ovarian Carcinoma: 377 patients with clinical data
- Copy Number Alteration: 11,864 copy number changes
- Pathway: KEGG pathway (186 pathways)



(B) Copy Number Alteration

Pahtway (Copy number alteration)
KEGG PATHWAYS IN CANCER
KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
KEGG RIBOSOME
KEGG CELL ADHESION MOLECULES CAMS
KEGG UBIQUITIN MEDIATED PROTEOLYSIS
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION
KEGG MAPK SIGNALING PATHWAY
KEGG WNT SIGNALING PATHWAY
KEGG HUNTINGTONS DISEASE
KEGG CHEMOKINE SIGNALING PATHWAY

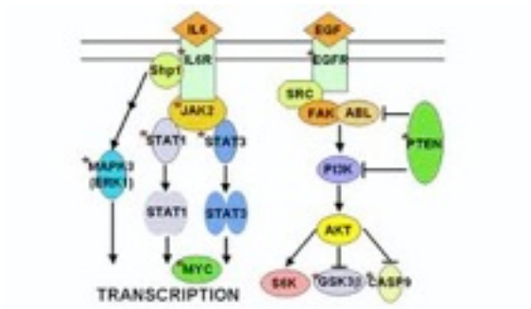
✓ Matrix tri-factorization can accurately identify patient subgroups having different survival outcome and pathways associated with patient subgroups

Patient stratification using genomic and pathway data integration using animal model

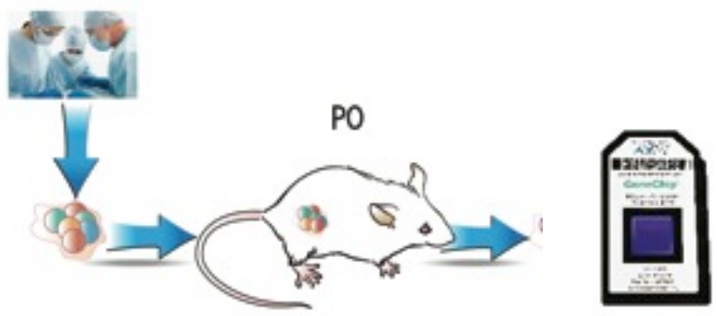
- **Develop novel computational methods to integrate cross-species genomic data for translational research**

Input

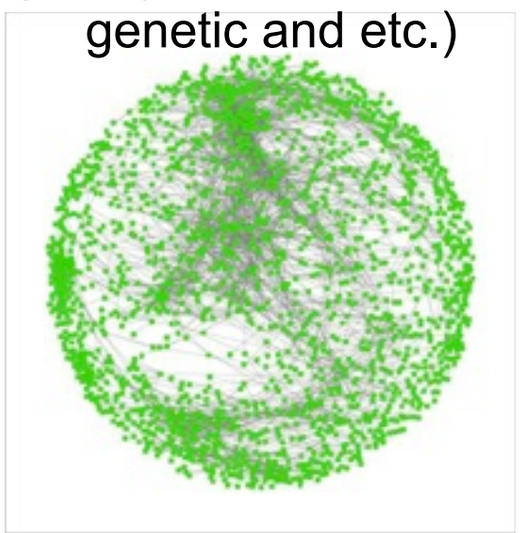
Biological prior knowledge (e.g, pathway, literature, and etc)



Genomic data (e.g., gene expression, copy number, mutation, methylation and etc.)

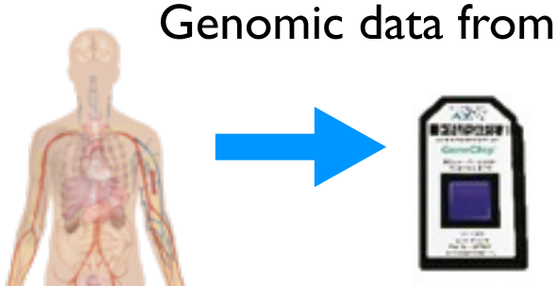


Gene-gene interaction (e.g., physical,coexpression, genetic and etc.)



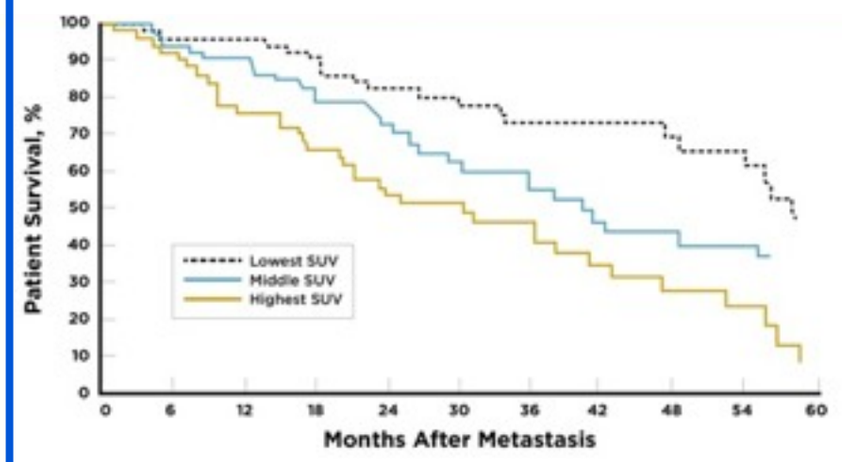
our integrated method

Genomic data from human

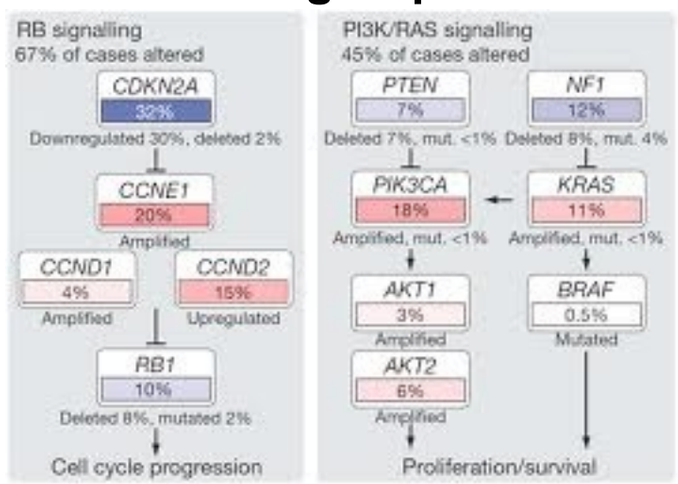


Output

Patient subgroup identification



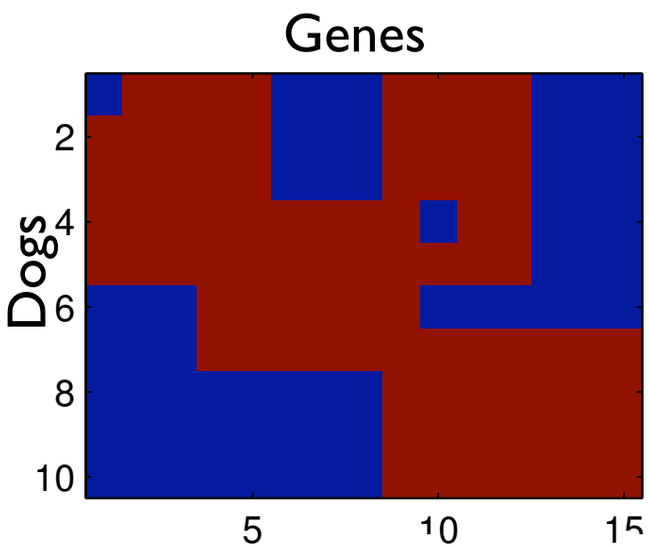
Pathways related with patient subgroup

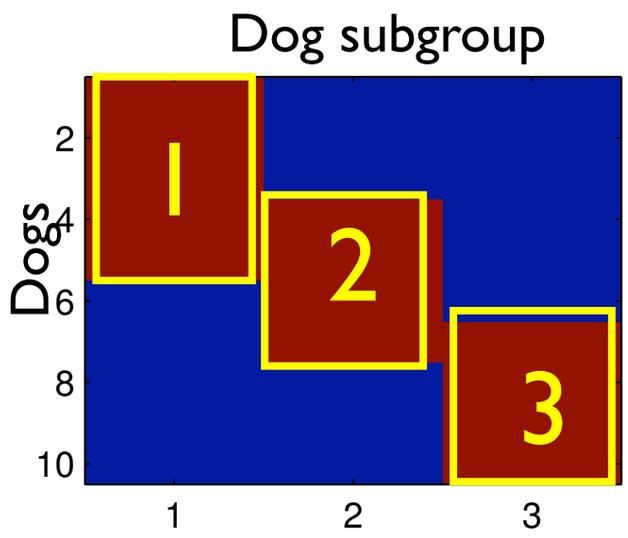


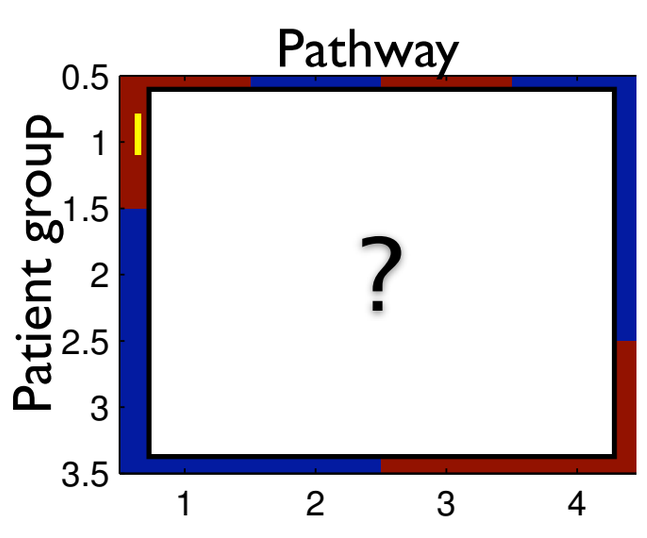
SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., **Nucleic Acids Research** 2012

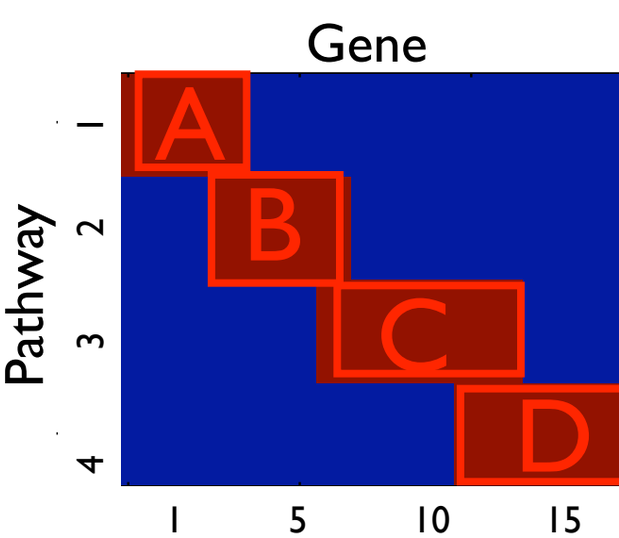
Cross-species Matrix tri-factorization

- Given: Dog and human gene expression, pathway data, and dog sugbgroup
- Task : Identify patient subgroups and pathway activities related with patient subgroups in human

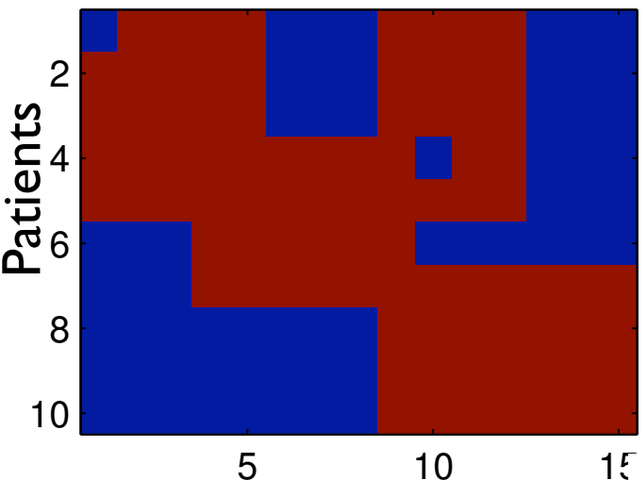


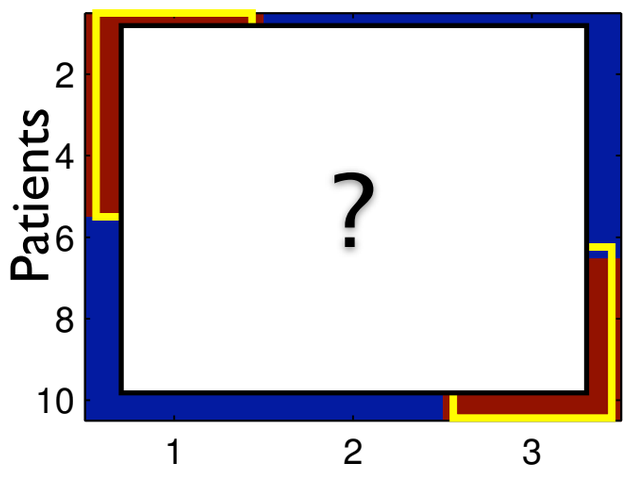


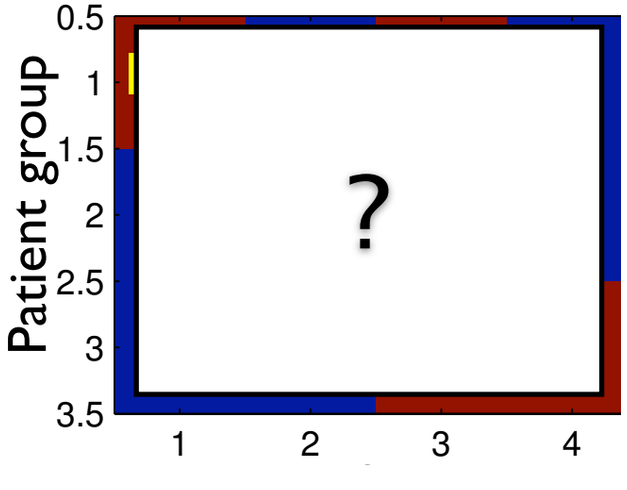


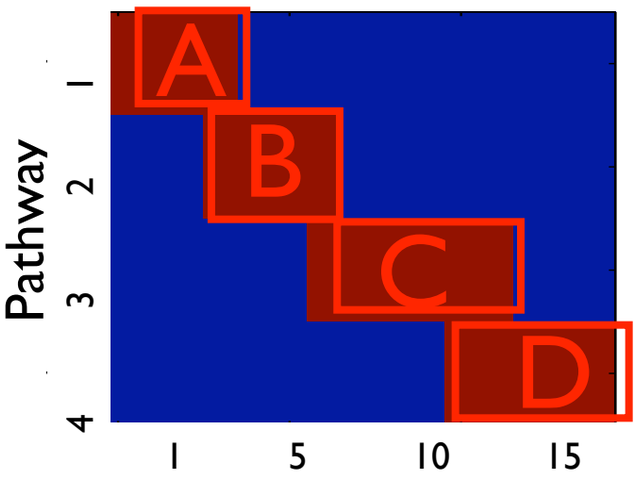


$X_{dog} \approx F_{dog} \times S_{dog} \times G^T$









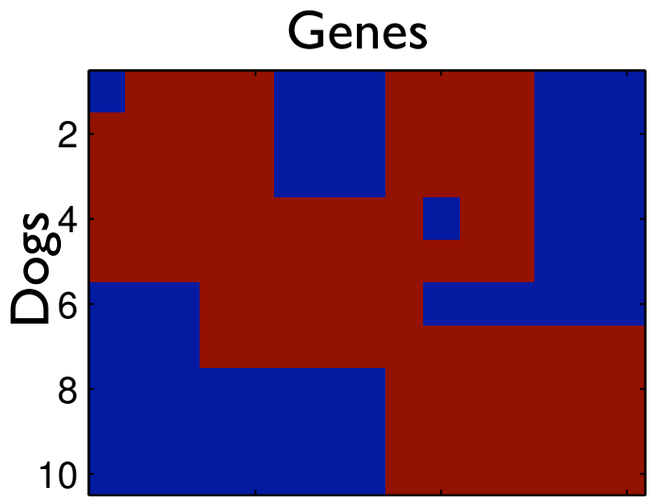
$X_{human} \approx F_{human} \times S_{human} \times G^T$

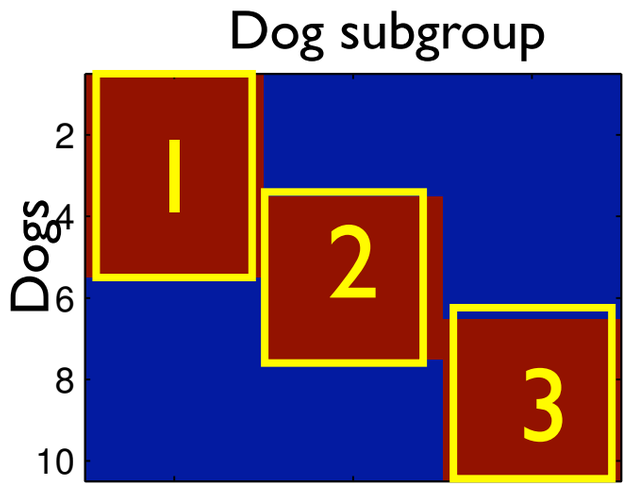
$$\min_{F_{human},\, S_{dog}} \frac{1}{2}\left(\|\mathbf{X}_{dog} - \mathbf{F}_{dog}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2 + \|\mathbf{X}_{human} - \mathbf{F}_{human}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2\right)$$

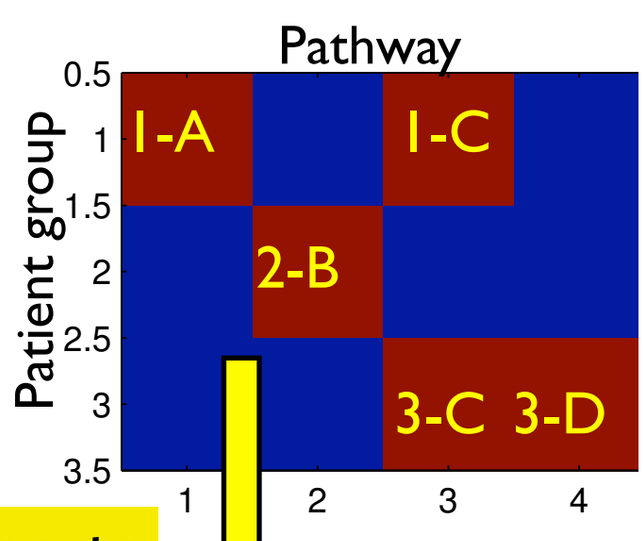
SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., **Nucleic Acids Research** 2012

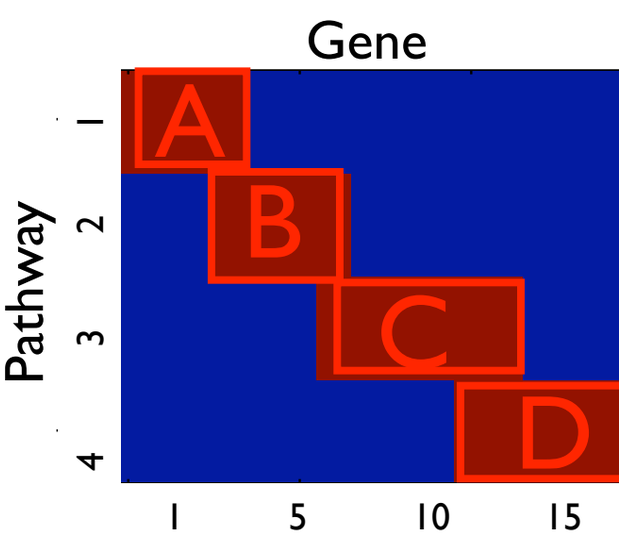
Cross-species Matrix tri-factorization

- Given: Dog and human gene expression, pathway data, and dog sugbgroup
- Task : Identify patient subgroups and pathway activities related with patient subgroups in human



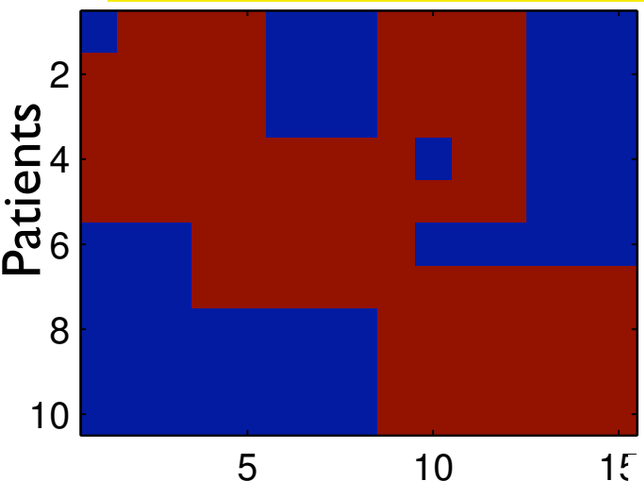


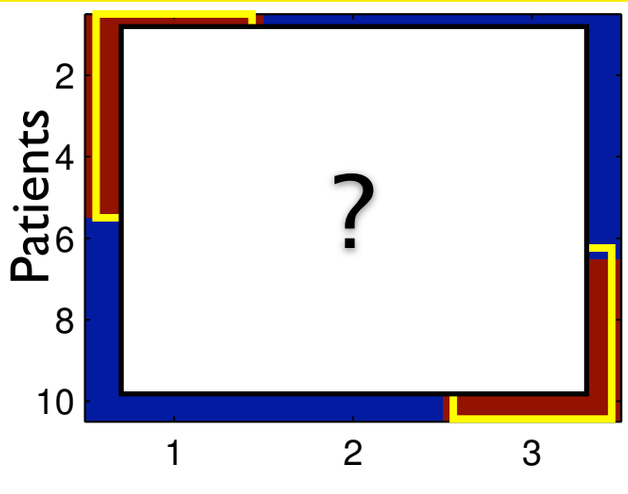


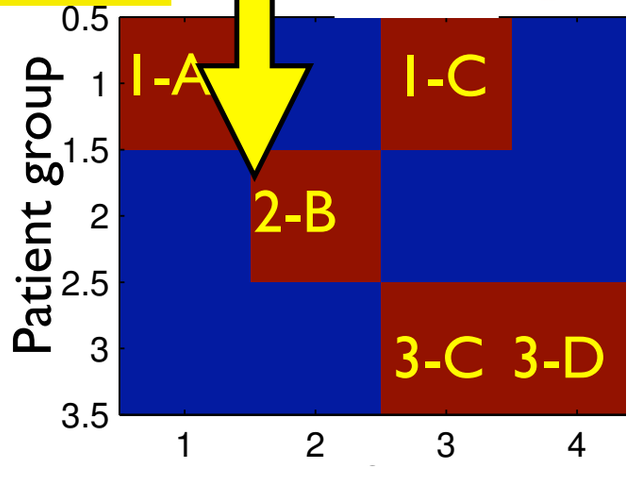


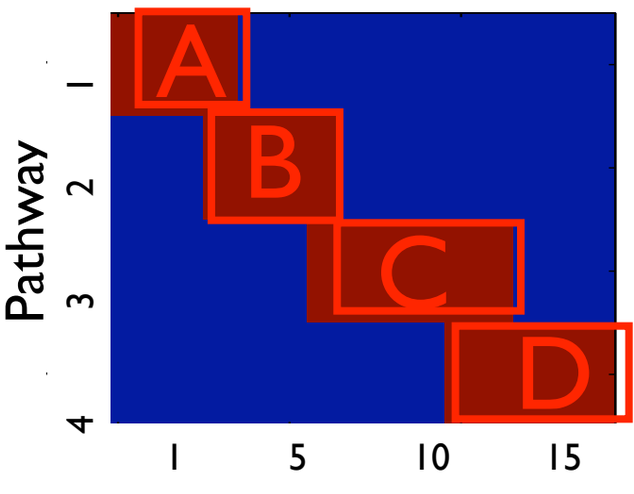
Use inferred pathway activities from dog to human cancer

S_{dog} × G^T









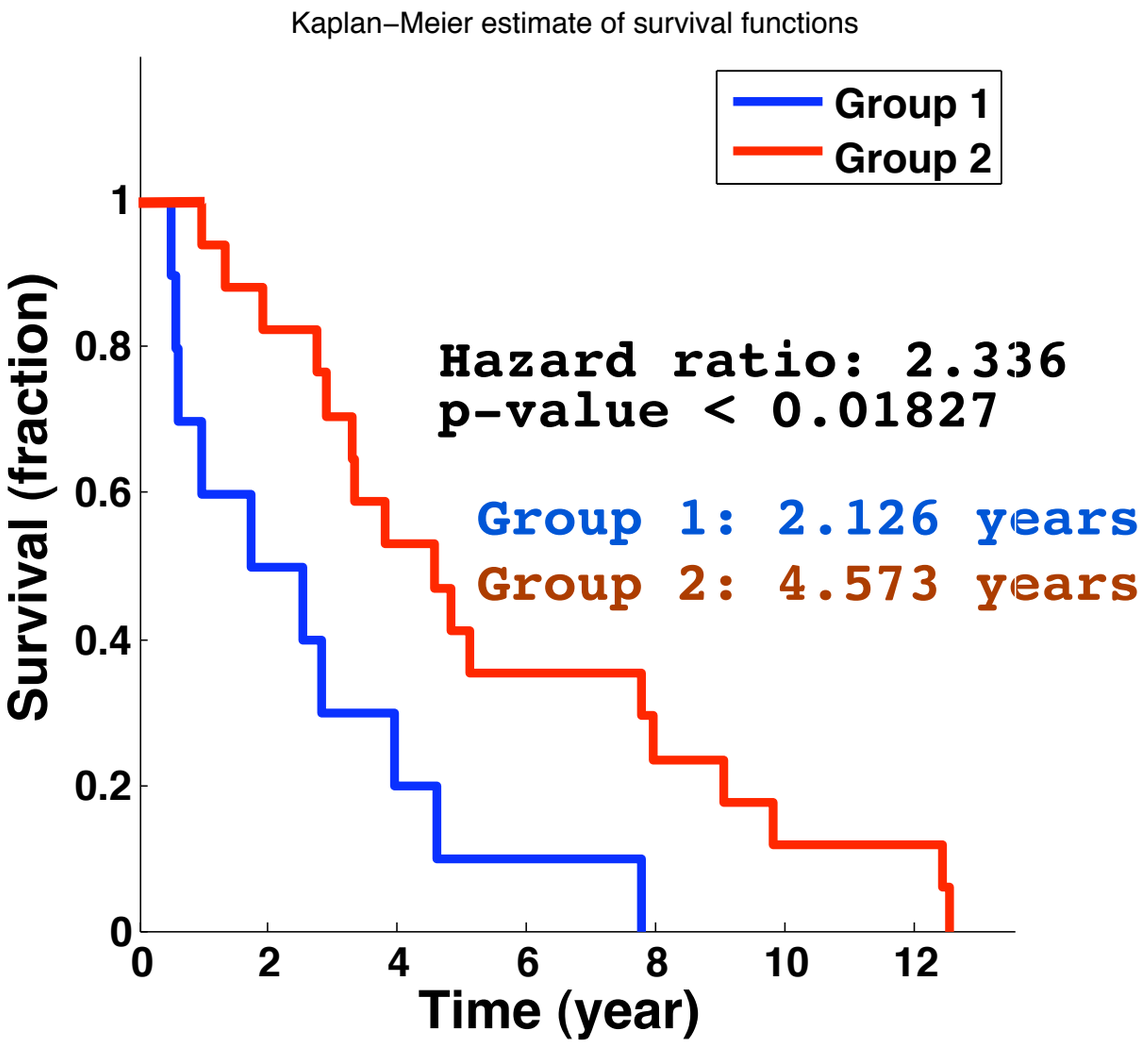
$$X_{human} \approx F_{human} \times S_{dog} \times G^T$$

$$\min_{F_{human},\, S_{dog}} \frac{1}{2}\left(\|\mathbf{X}_{dog} - \mathbf{F}_{dog}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2 + \|\mathbf{X}_{human} - \mathbf{F}_{human}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2\right)$$

SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., **Nucleic Acids Research** 2012

Experiments (Osteosarcoma)

- Osteosarcoma: 34 dogs (GSE27217) and 34 patients (GSE16091) with clinical data
- Pathway: Reactome pathway (430 pathways)
- 5 (short) vs 12 (long) months for dog subgroup



Ranking	Pathway
1	INFLUENZA LIFE CYCLE
2	**CELL CYCLE CHECKPOINTS**
3	**STABILIZATION OF P53**
4	S PHASE
5	DNA STRAND ELONGATION
6	SCF SKP2 MEDIATED DEGRADATION OF P27 P21
7	CYCLIN E ASSOCIATED EVENTS DURING G1 S TRANSITION
8	**SIGNALING BY NGF**
9	REGULATION OF INSULIN SECRETION BY GLUCAGON LIKE PEPTIDE 1
10	NEURORANSMITTER RECEPTOR BINDING
11	SYNTHESIS OF DNA
12	OPIOID SIGNALLING
13	**SIGNALING BY WNT**
14	ACTIVATION OF NMDA RECEPTOR UPON GLUTAMATE BINDING
15	VIF MEDIATED DEGRADATION OF APOBEC3G

SJ Kim**, T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., **Nucleic Acids Research** 2012

Take home message

- Integrating genomic data with pathway database can help to improve an ability for patient stratification and pathway discovery
- Leveraging knowledge (i.e., pathway activities) from dog cancer can help to study human cancer
- Our proposed method is a general method, and applicable to other problems
  - Inner-species analysis: infer pathway activities from one data, and use them to study another data
  - Tissue or cancer type specific dysregulated pathway activity analysis

•SJ Kim**, T. Hwang***, Georgios B. Giannakis, “Sparse Robust Matrix Tri-factorization with Application to Cancer Genomics”, International Workshop on Cognitive Information Processing, **CIP 2012**

•**T. Hwang**, Maoqiang Xie, Gowtham Atluri, Sanjoy Dey, Vipin Kumar, Changjin Hong and Rui Kuang. “Co-clustering Phenome-genome for Phenotype Classification and Disease Gene Discovery”, **Nucleic Acids Research 2012**

Large-scale network-based integrative analysis identifies common pathways disrupted by copy number alterations across cancers

TaeHyun Hwang[†1], Gowtham Atluri[2], Rui Kuang[2], Vipin Kumar[2,] Timothy Starr[1], Peter M Haverty[3], Zemin Zhang[3], Jinfeng Liu[†3]

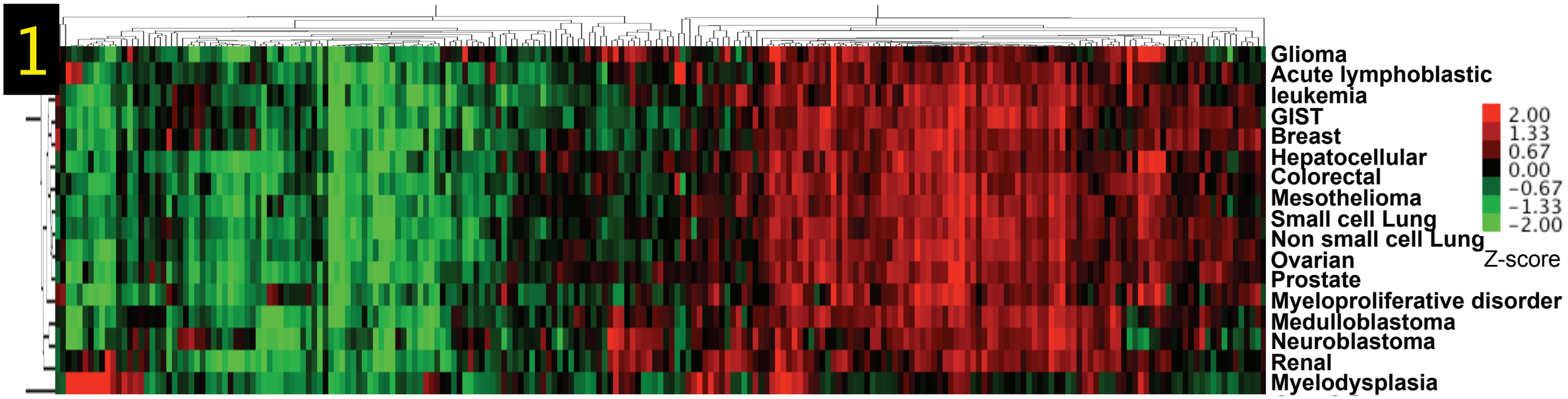
[1]Masonic Cancer Center, [2]Department of Computer Science and Engineering, University of Minnesota - Twin Cities; [3]Department of Bioinformatics and Computational Biology, Genentech Inc.

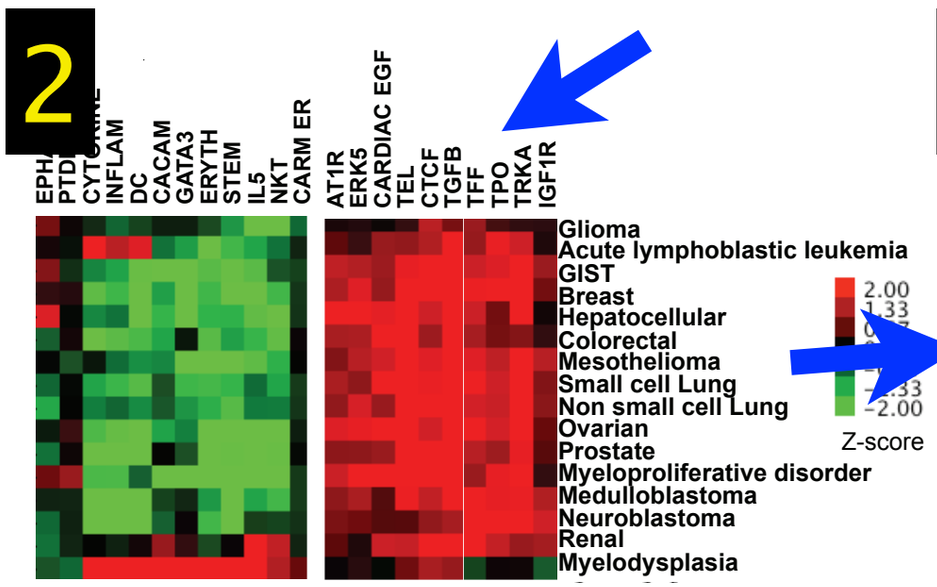
***Joint work with Genentech**

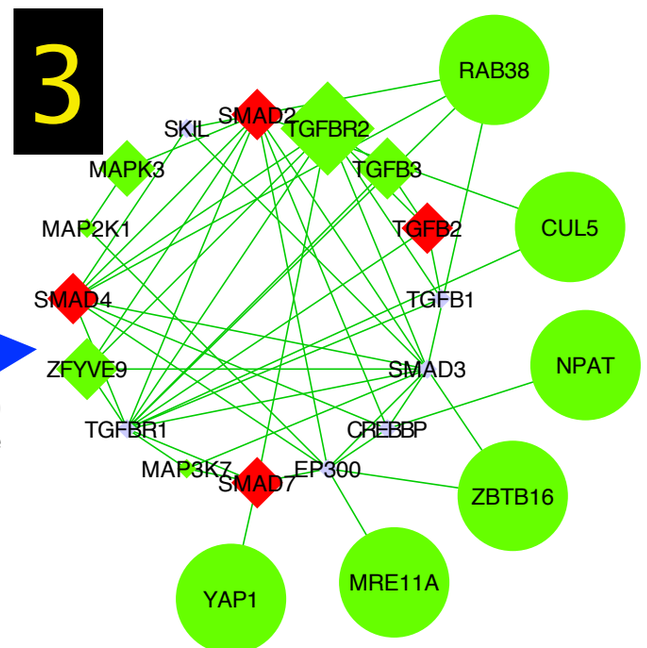
T. Hwang et. al, under review

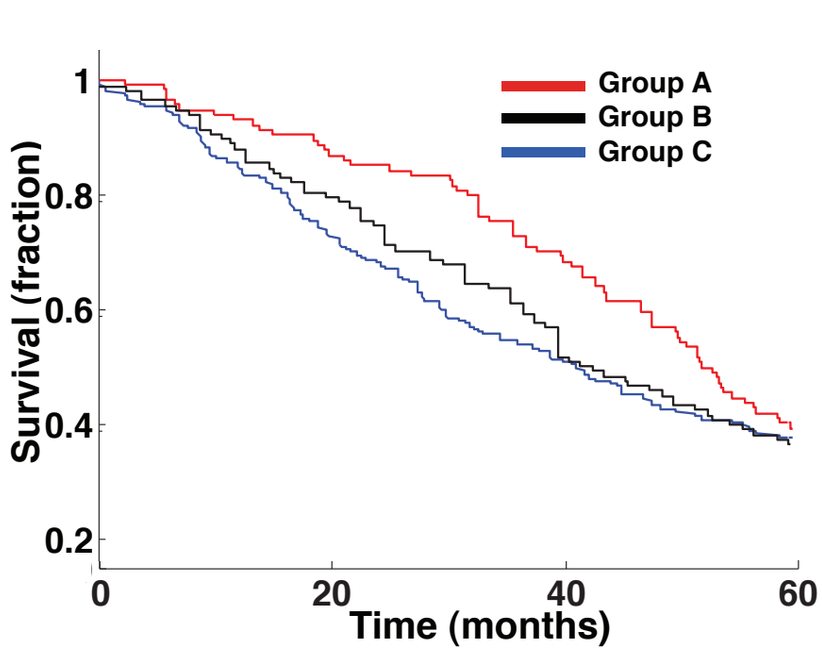
Motivation

1. Comprehensive pathway activity map across 16 types of cancers
2. Common and cancer-type specific disrupted pathway
3. Network view how copy number alterations can affect pathway
4. Pathway signatures to identify patient subgroups



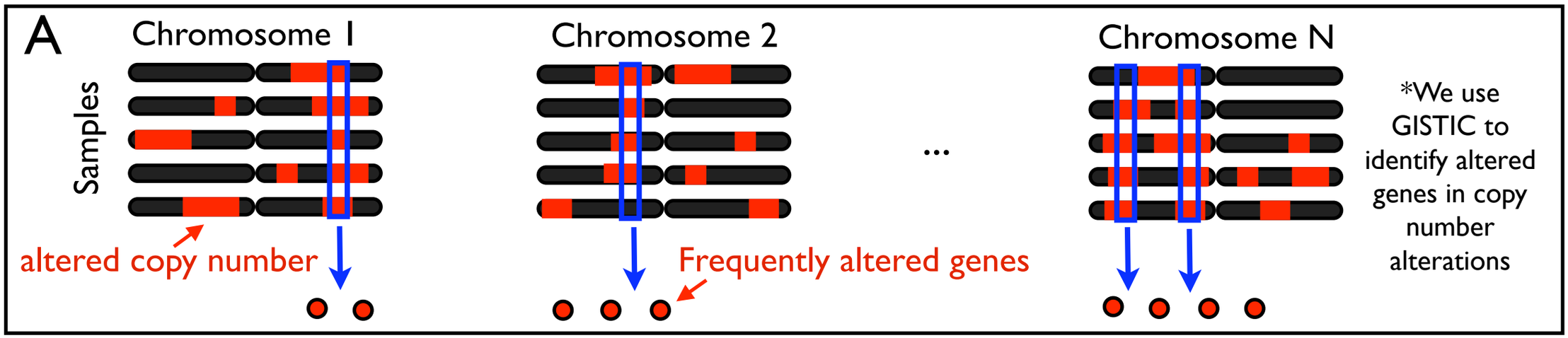






T. Hwang et. al, under review

Overview

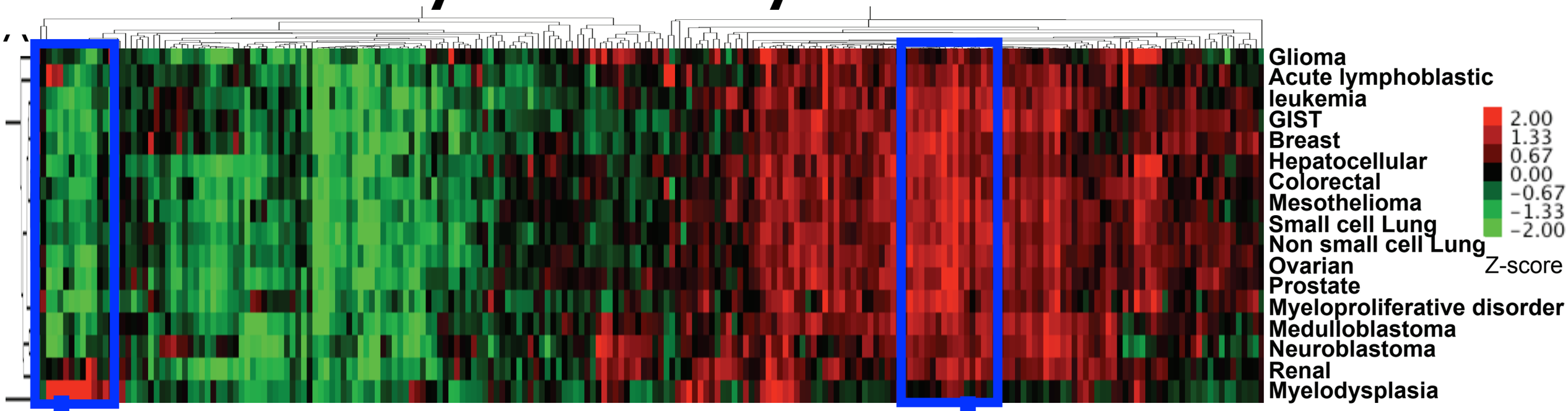


Experiments

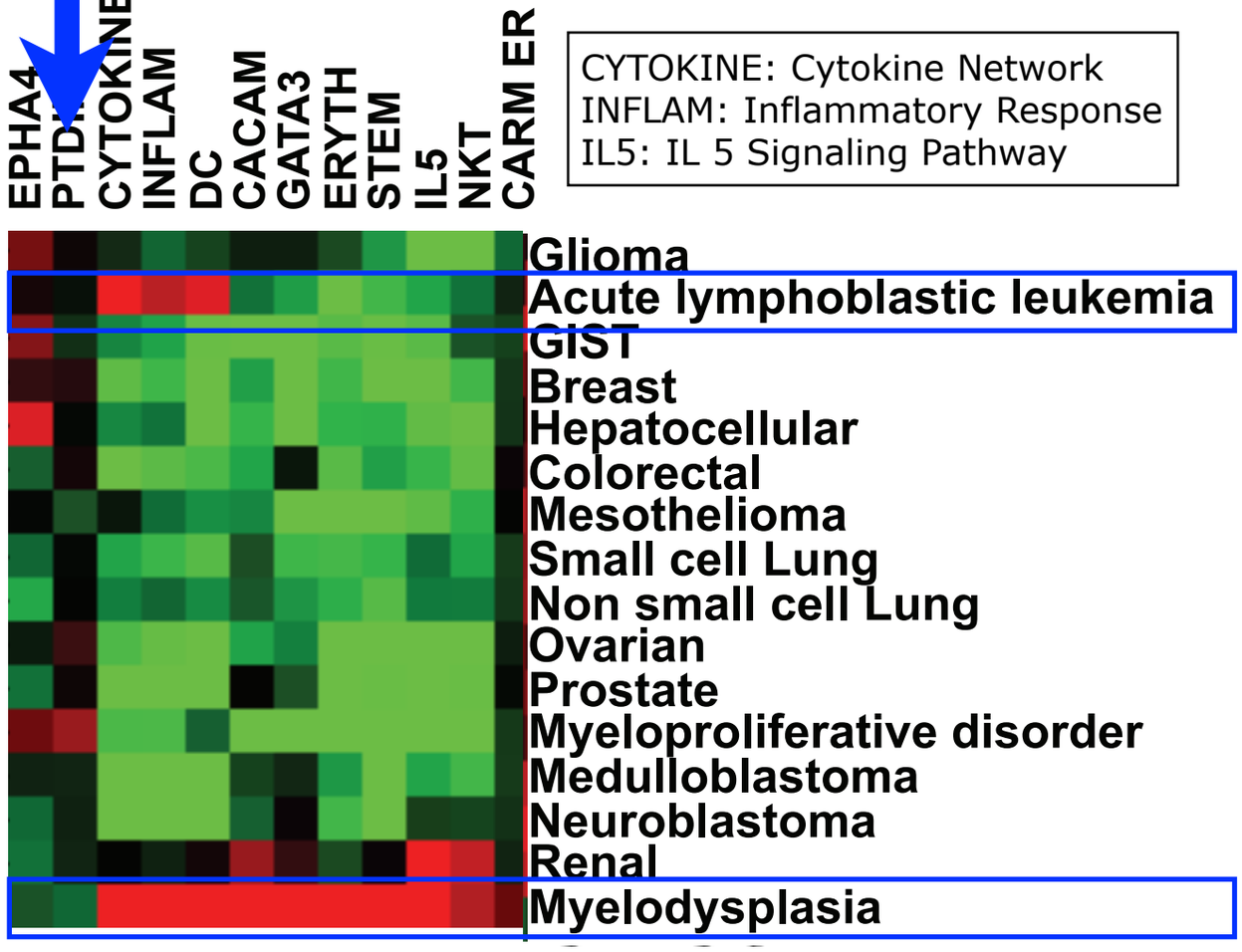
- Data
  - 2172 patients from 16 different types of cancers using Affymetrix 250k sty SNPs array data [Beroukhim et al., Nature 2010]
    - Use pennCNV to measure CNA, and use GLAD to segmentation
    - Use GISTIC to find significantly altered copy number region
  - Human protein-protein interaction network from HPRD database (May 2010)
    - 9674 proteins and 34,998 protein interactions
  - Pathway database
    - KEGG, Biocarta, and Reactome from MSigD, and conserved subnetworks cross species

T. **Hwang** et. al, under review

Pathway activity view of cancers

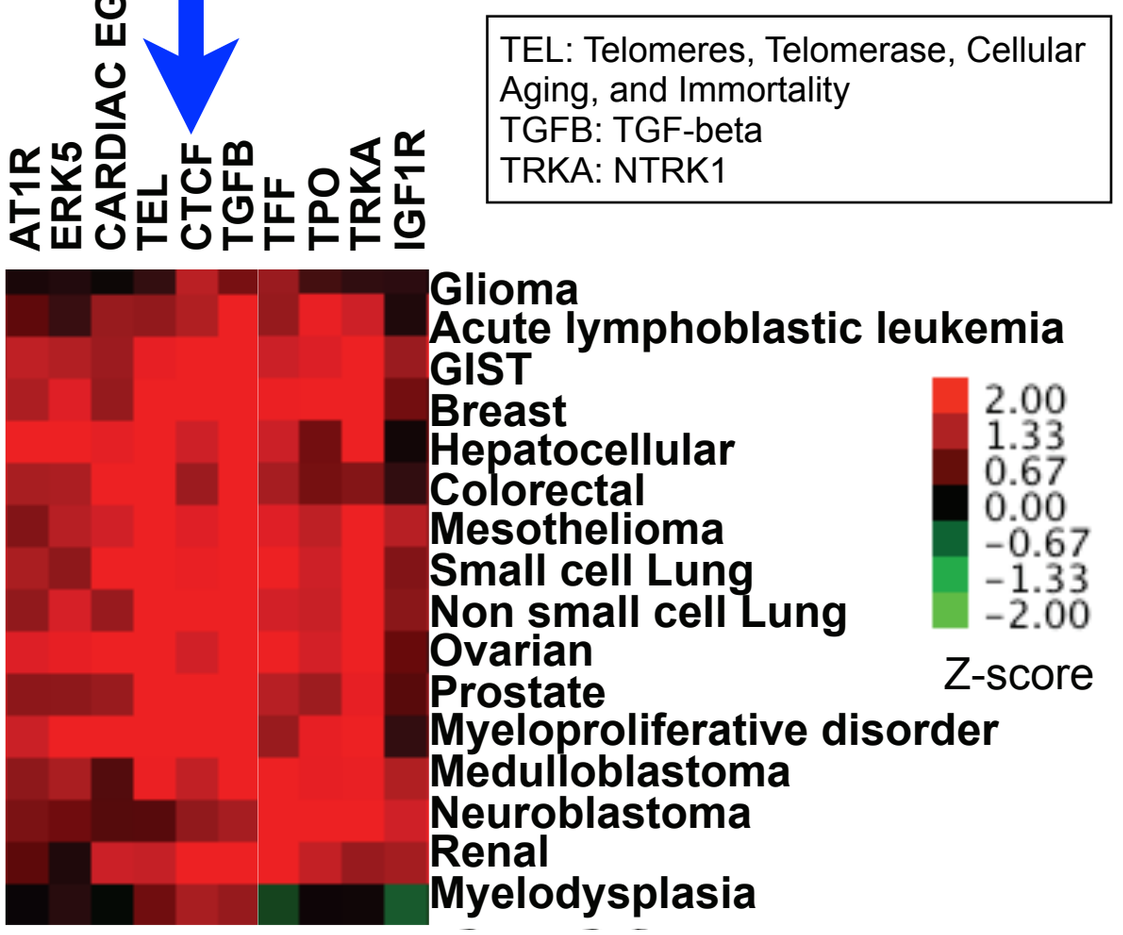


Cancer type specific disrupted pathway



CYTOKINE: Cytokine Network
INFLAM: Inflammatory Response
IL5: IL 5 Signaling Pathway

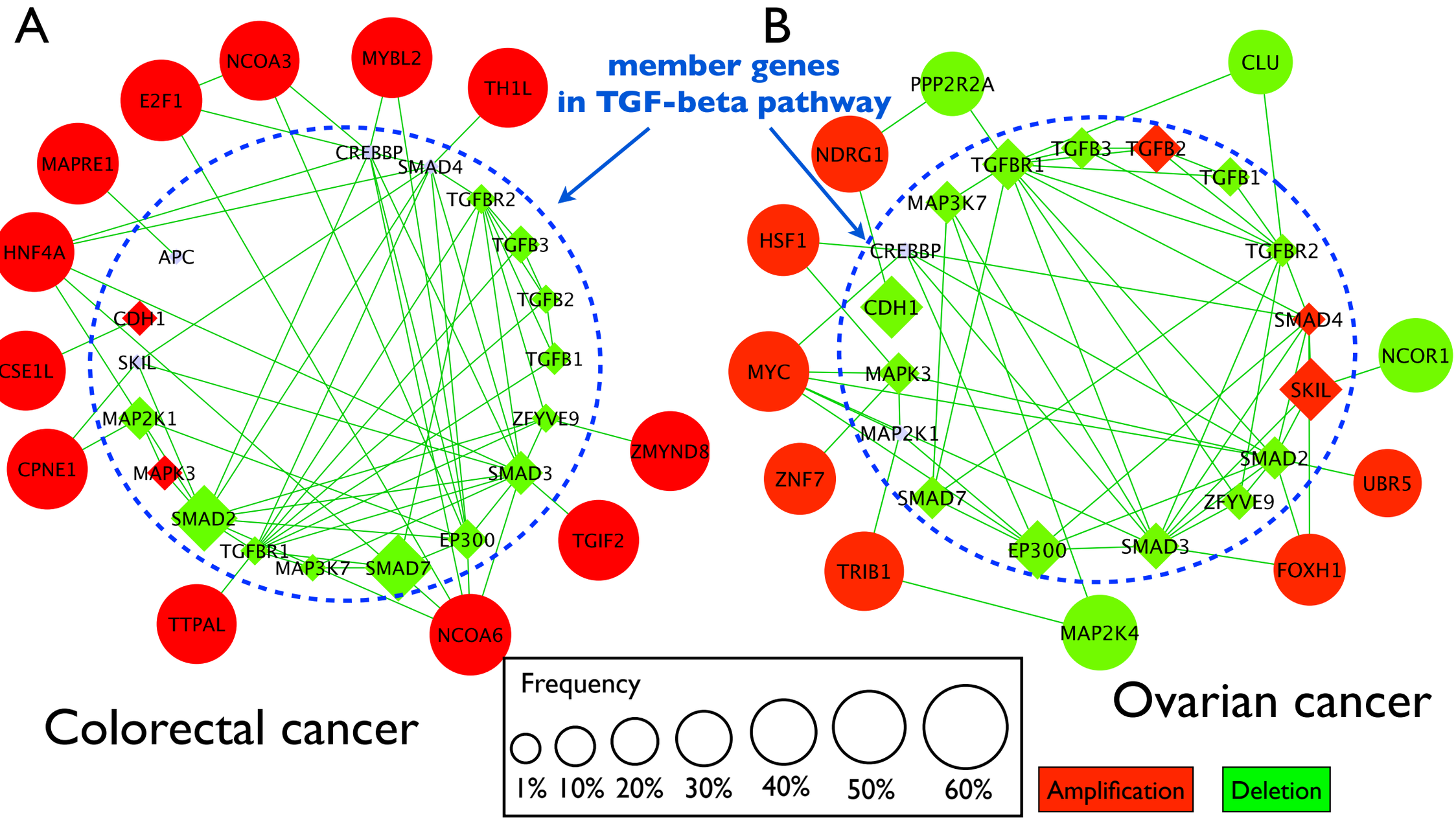
Commonly disrupted pathway



TEL: Telomeres, Telomerase, Cellular Aging, and Immortality
TGFB: TGF-beta
TRKA: NTRK1

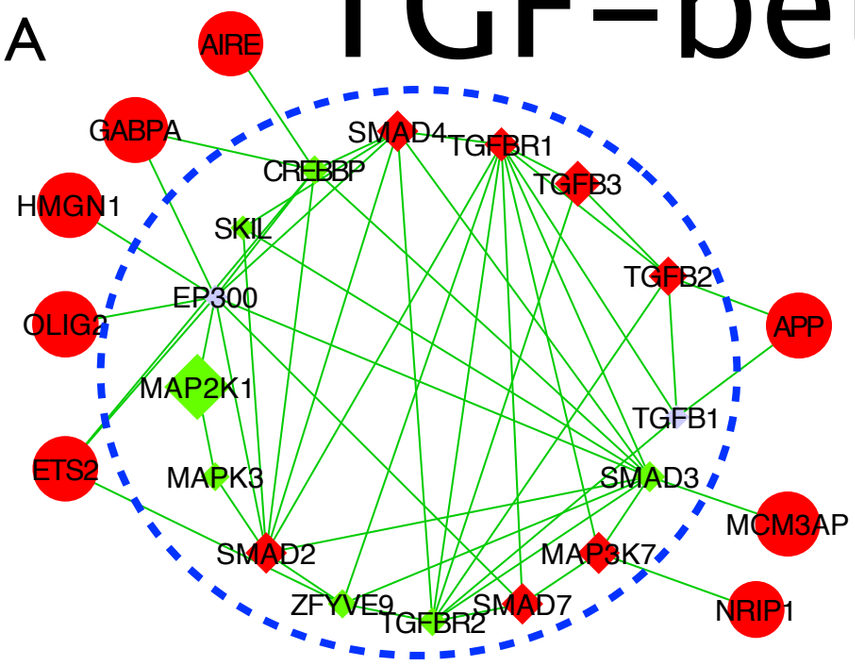
T. Hwang et. al, under review

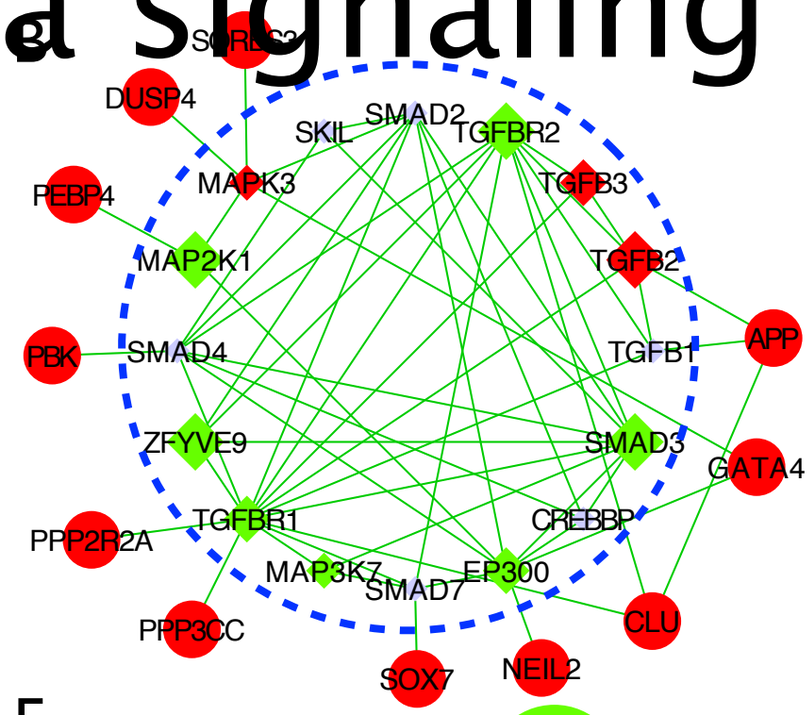
TGF-beta signaling pathway

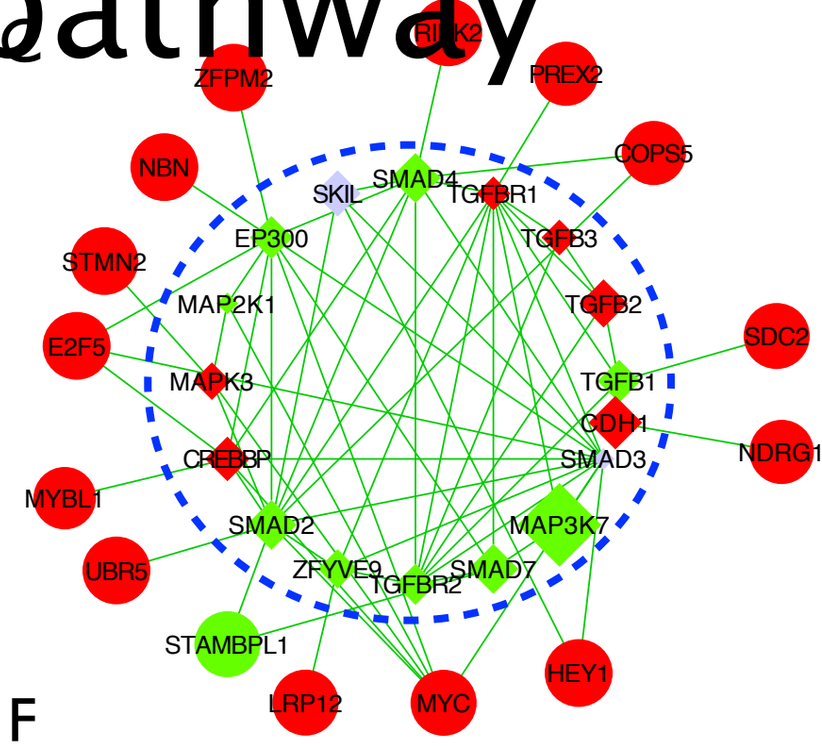


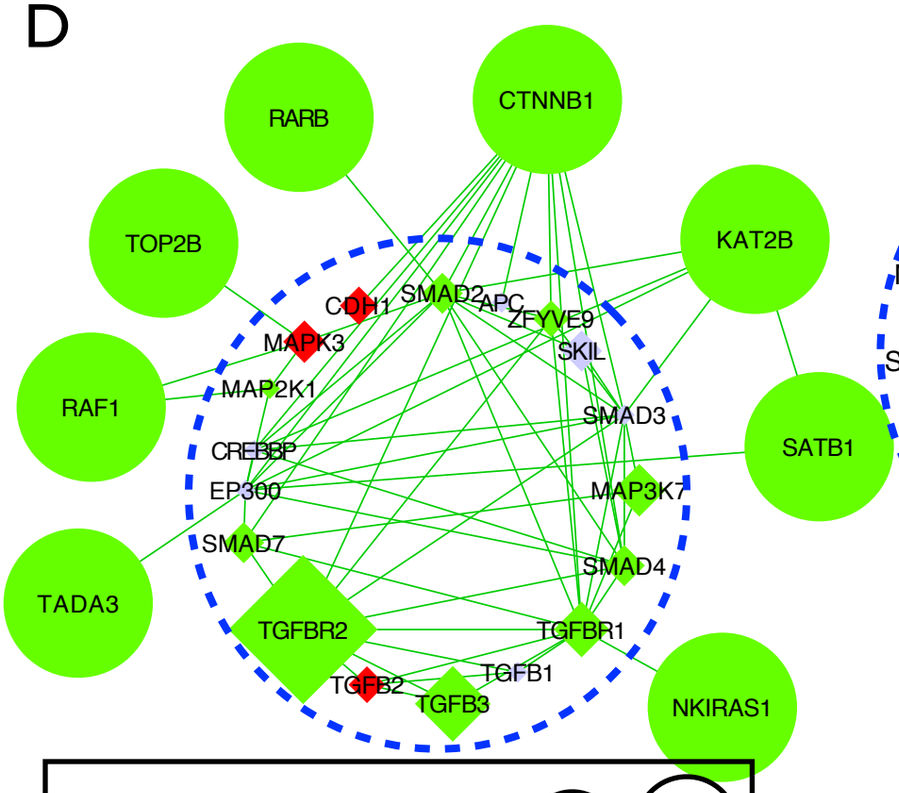
Member genes in the pathway have **low** frequency!
but...

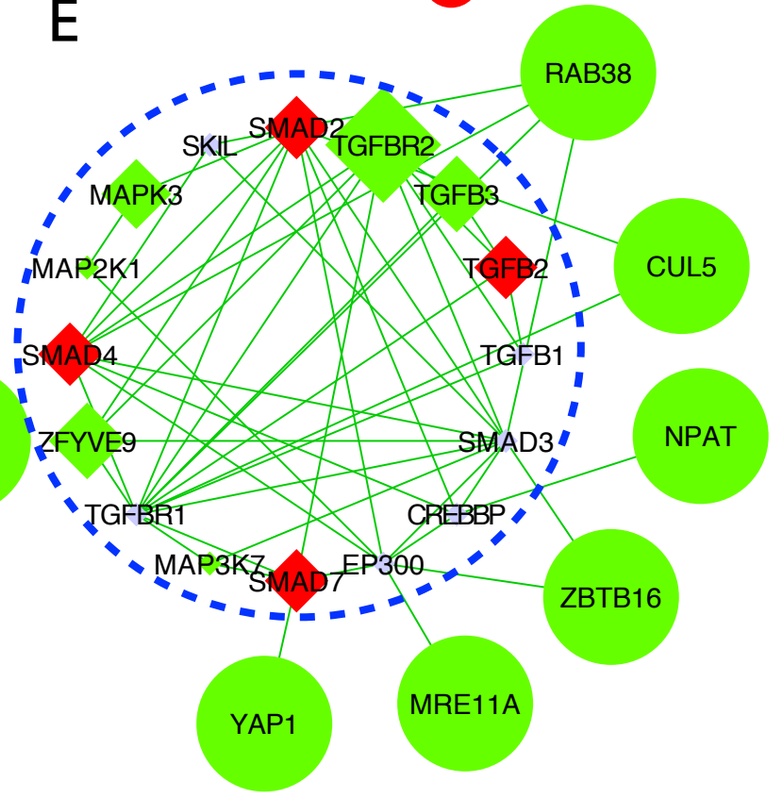
TGF-beta signaling pathway

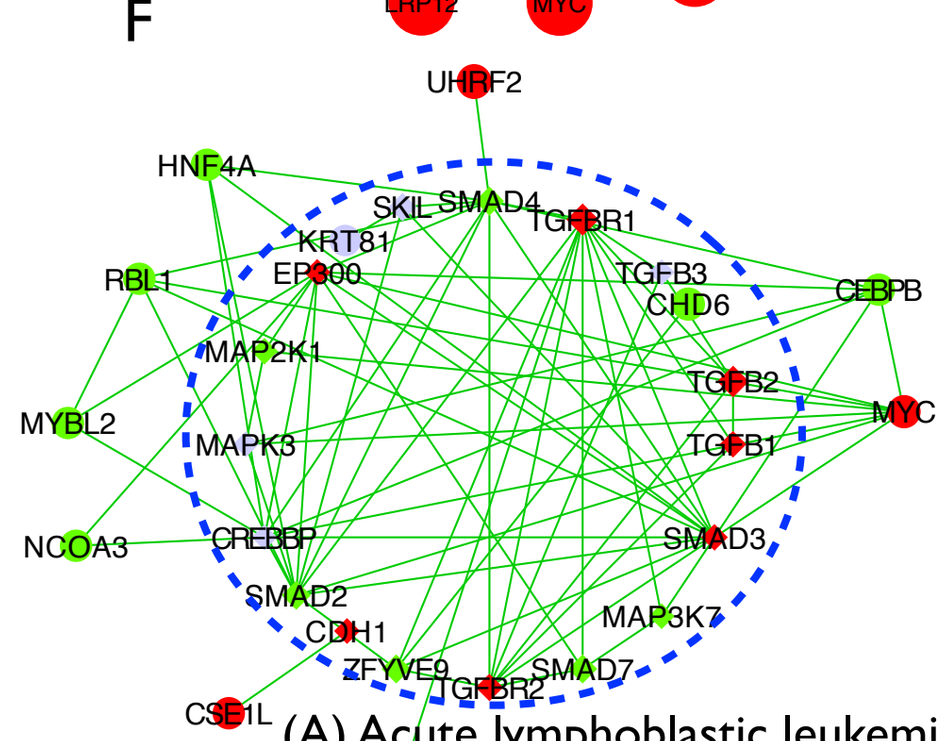






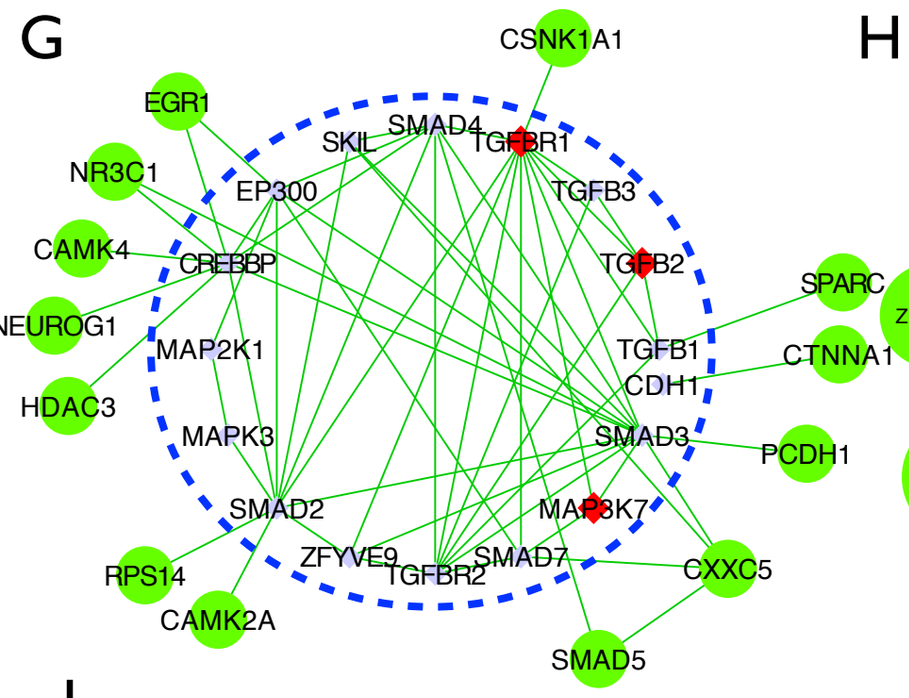


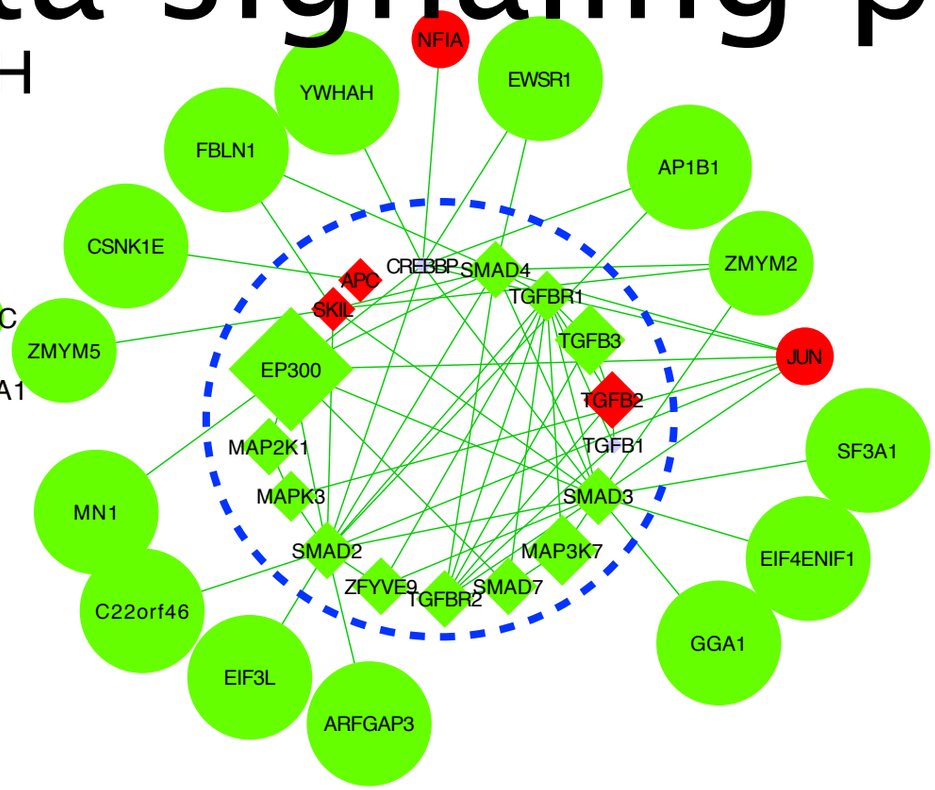


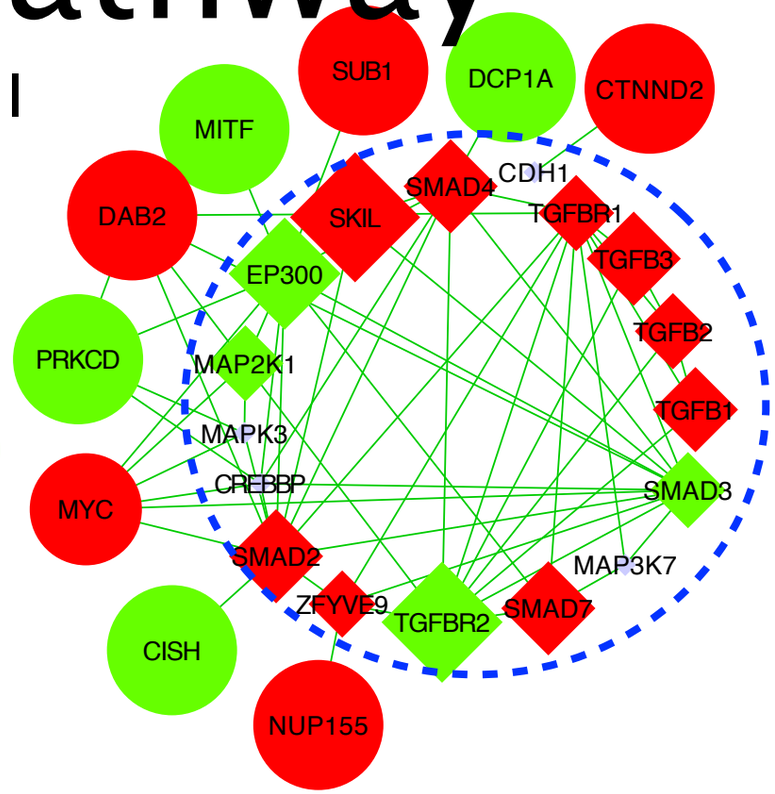


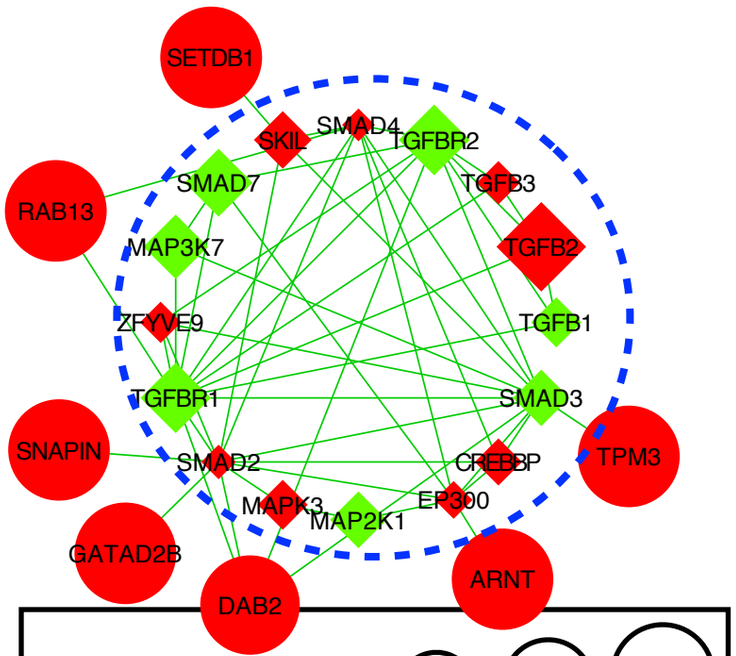
(A) Acute lymphoblastic leukemi
(B) Gastrointestinal stromal tum
(C) Prostate cancer
(D) Renal cancer
(E) Neuroblastoma
(F) Myeloproliferative disorder

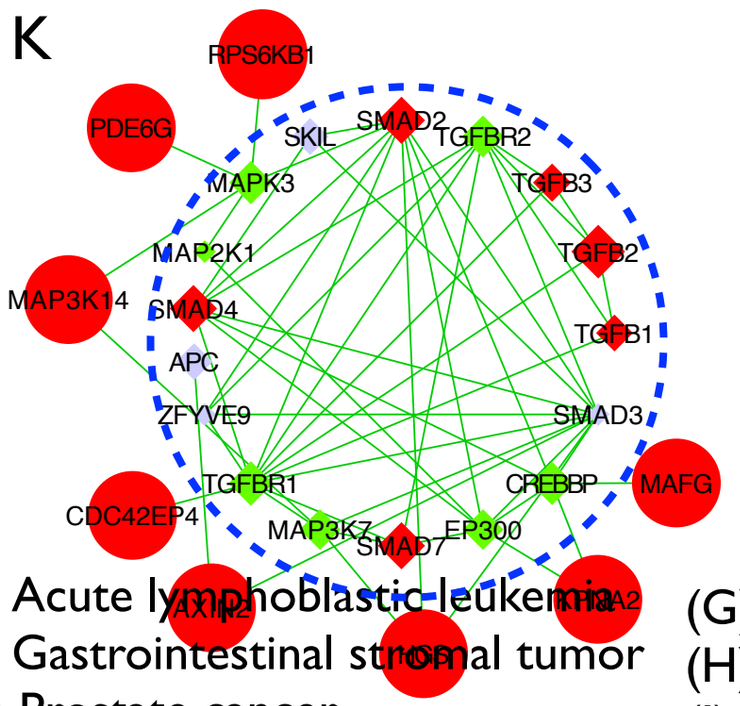
TGF-beta signaling pathway

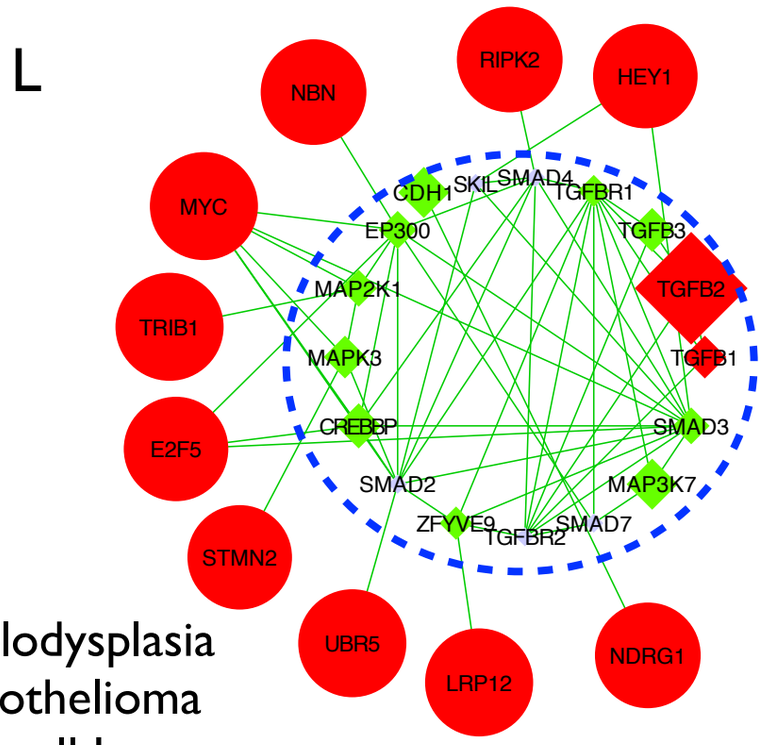










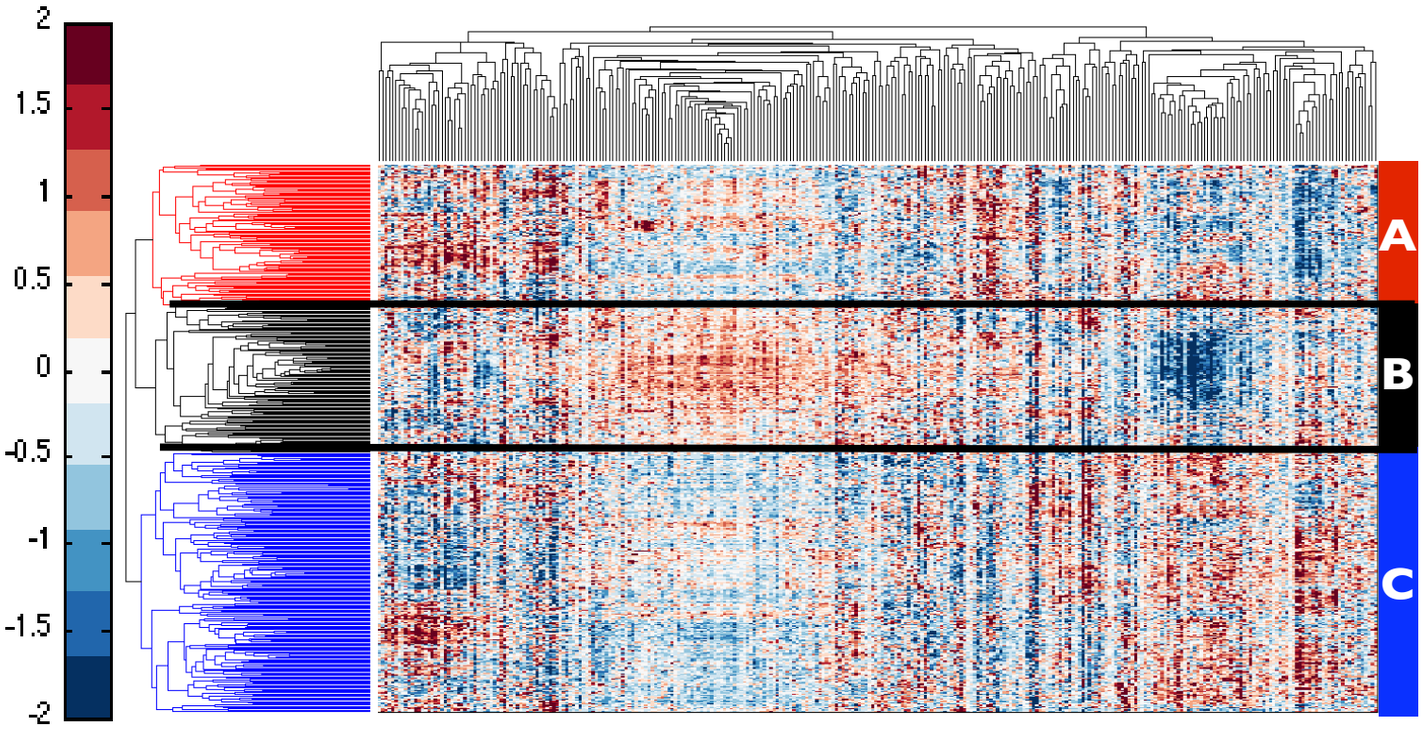


(A) Acute lymphoblastic leukemia
(B) Gastrointestinal stromal tumor
(C) Prostate cancer
(D) Renal cancer
(E) Neuroblastoma
(F) Myeloproliferative disorder
(G) Myelodysplasia
(H) Mesothelioma
(I) Small cell Lung cancer
(J) Non small cell Lung cancer
(K) Medulloblastoma
(L) Hepatocellular

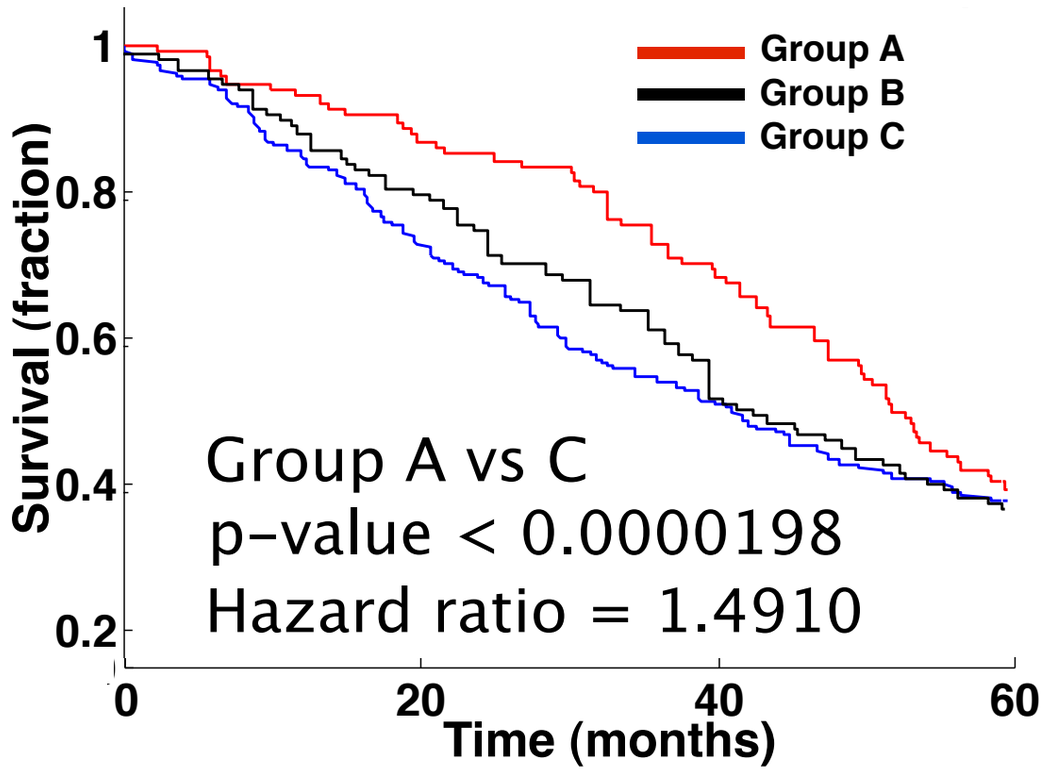
T. Hwang et. al, under review

Commonly disrupted pathways across cancers correlate with clinical outcomes

A



B

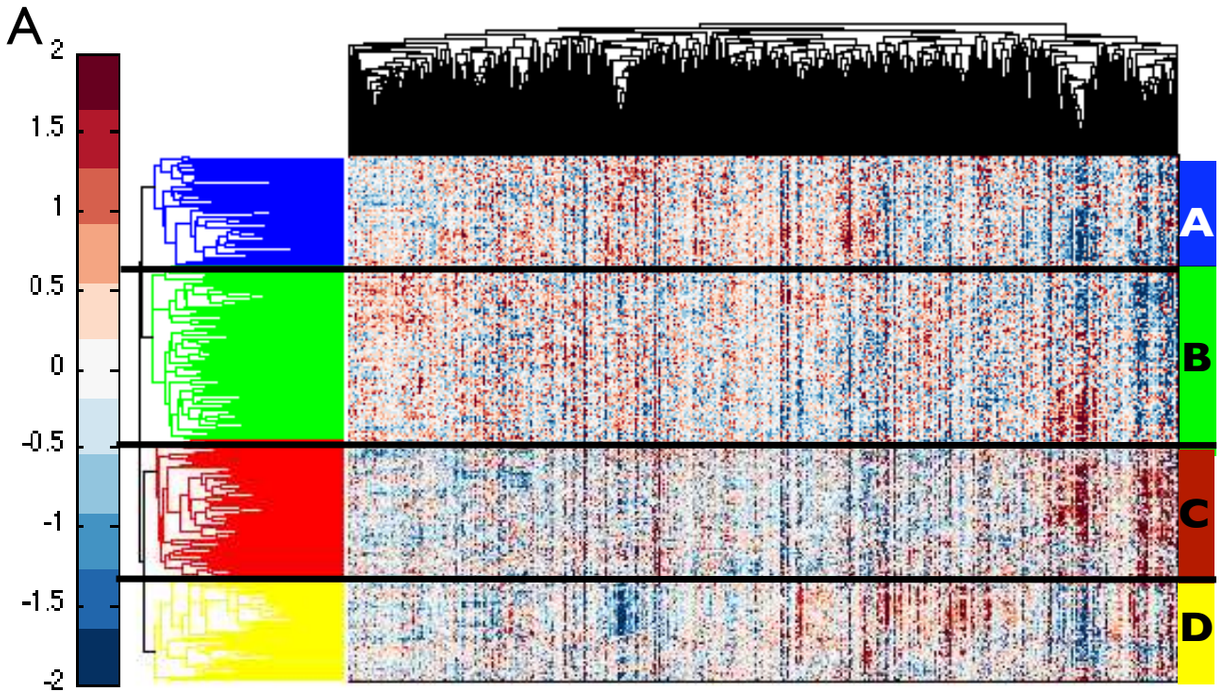


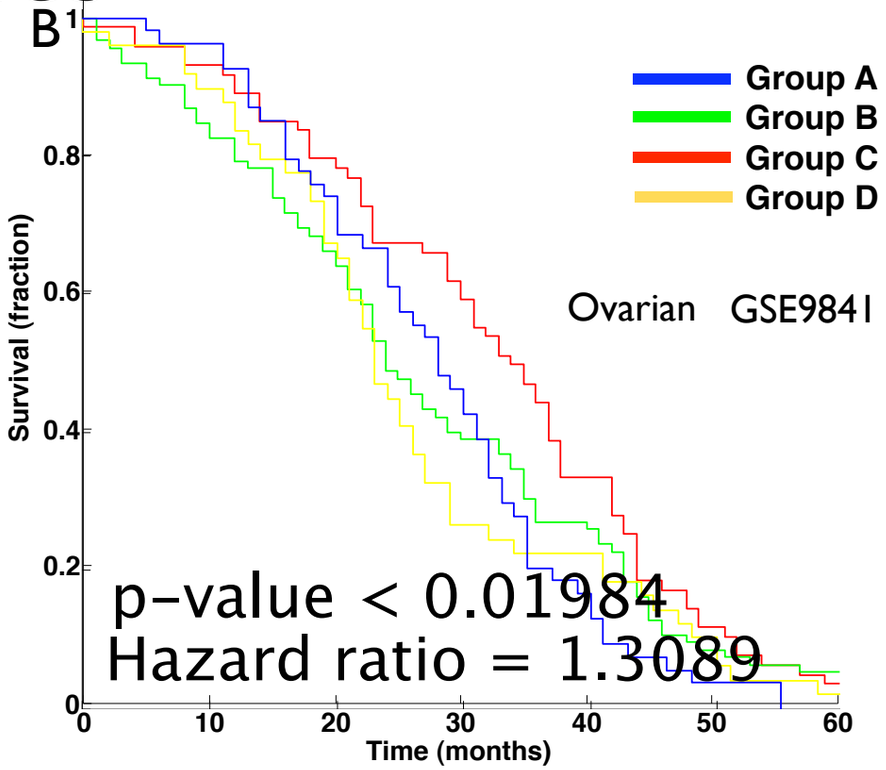
Commonly disrupted pathways may allow stratification of cancers at the pathway level, which could lead to the development of more targeted therapeutic!

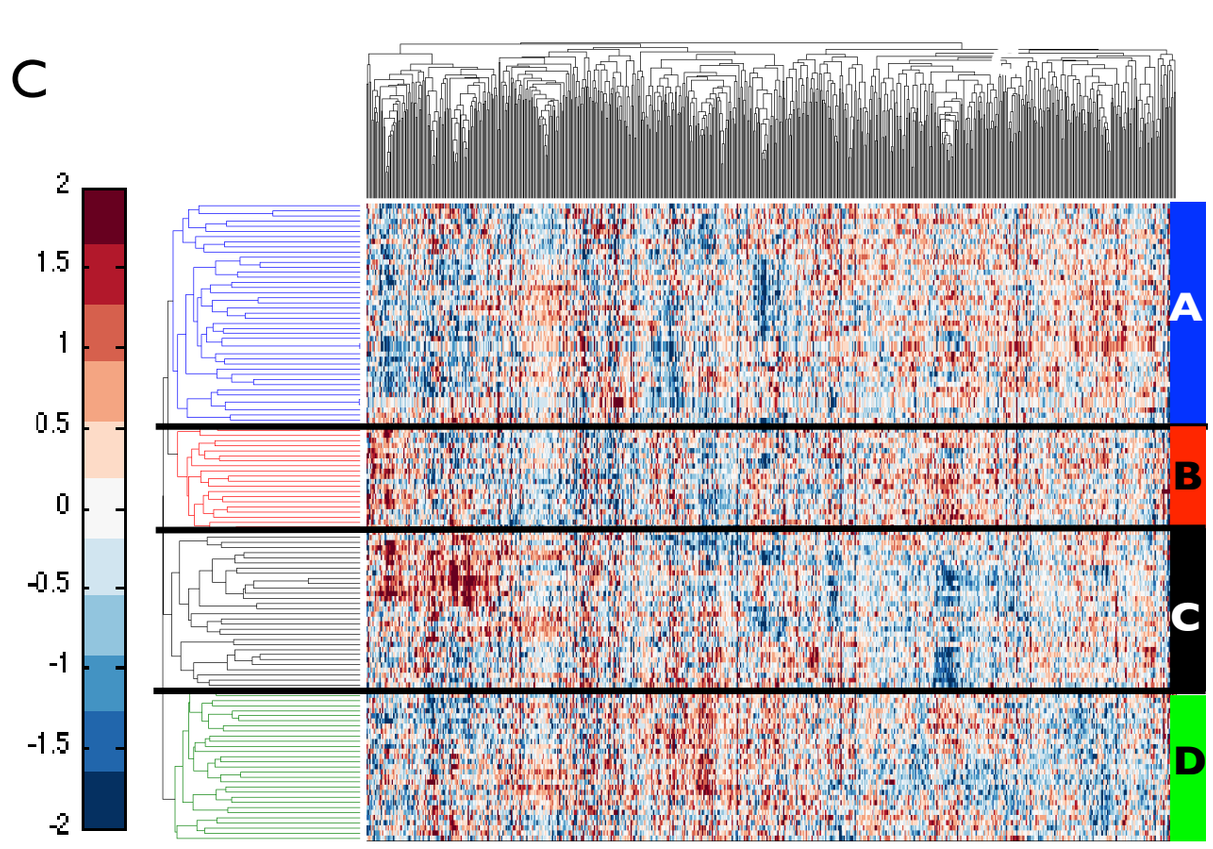
Shedden K, Taylor JMG, Enkemann SA, Tsao MS, Yeatman TJ, et al. (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nature medicine 14: 822-827

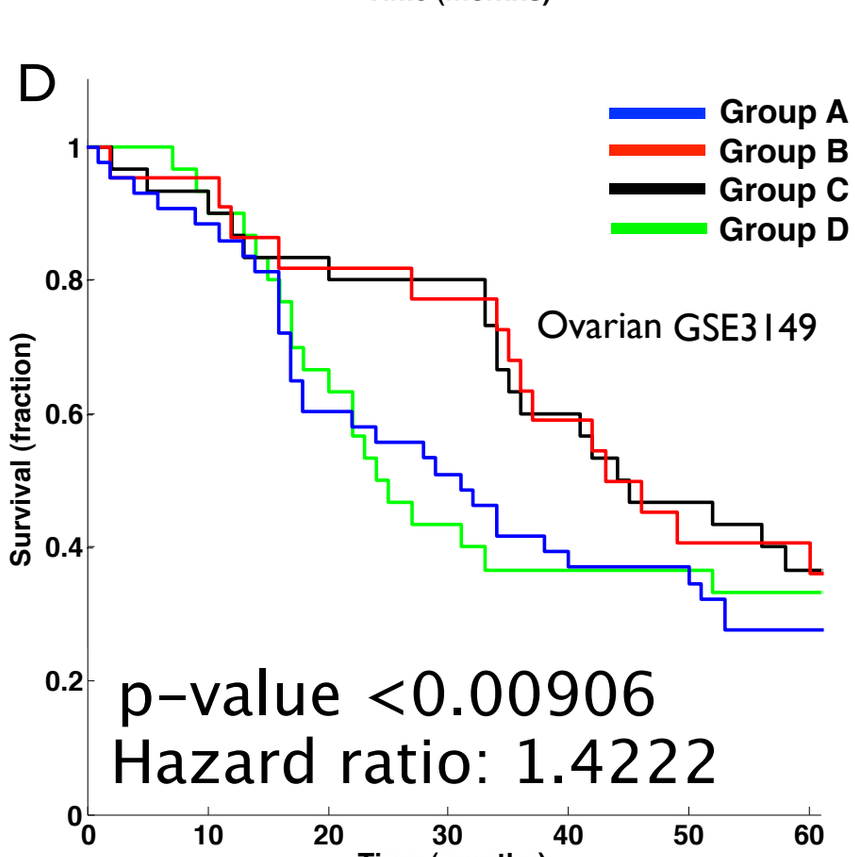
T. Hwang et. al, under review

Commonly disrupted pathways across cancers correlate with clinical outcomes





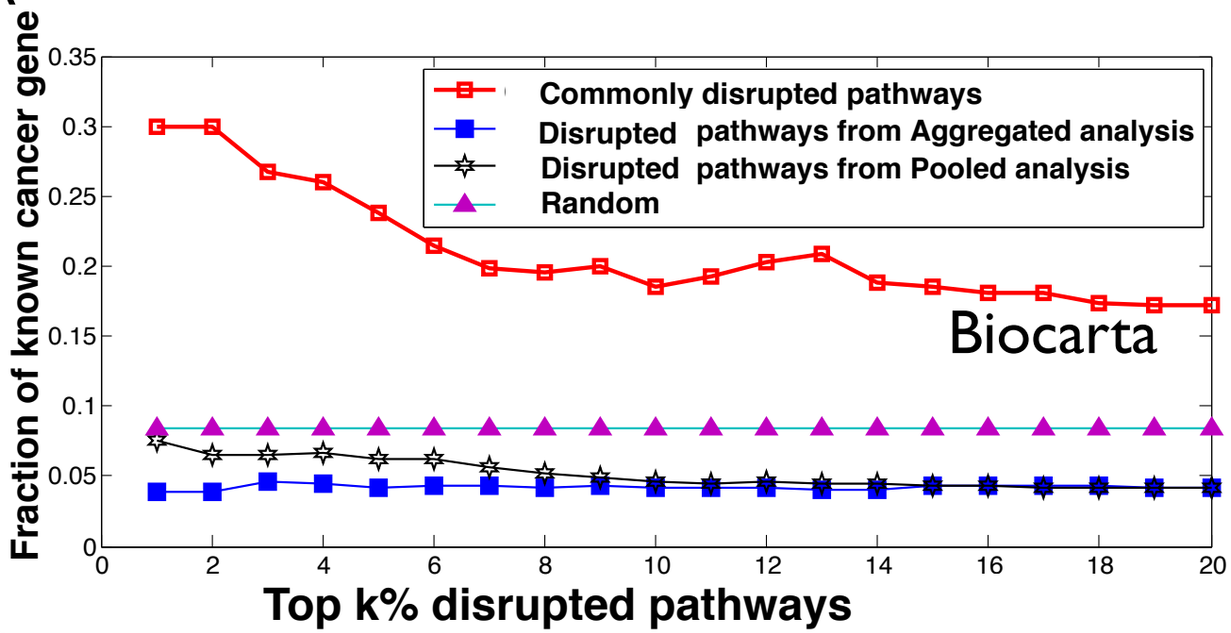




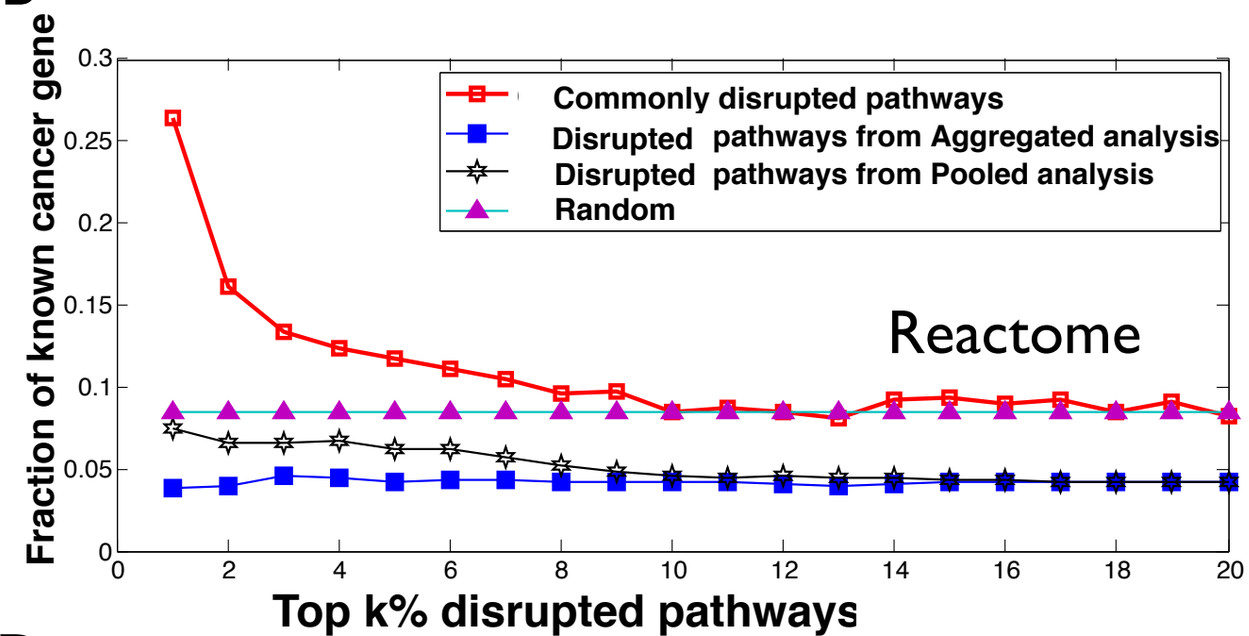
T. Hwang et. al, under review

Evaluation (Cancer gene enrichment)

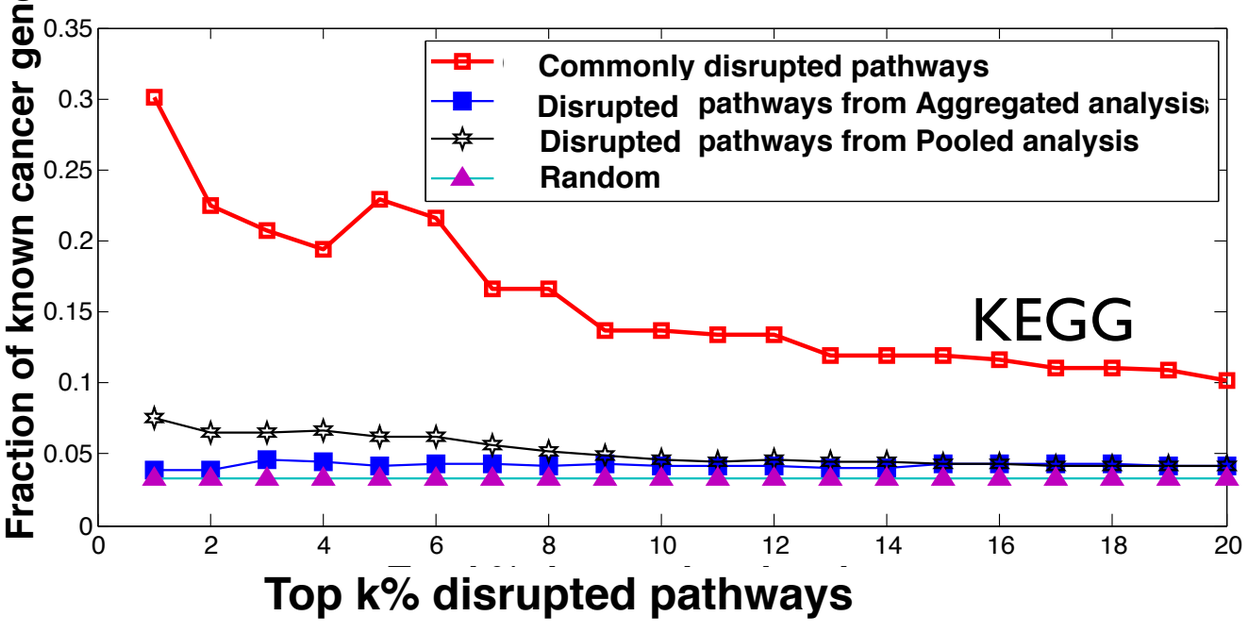
A



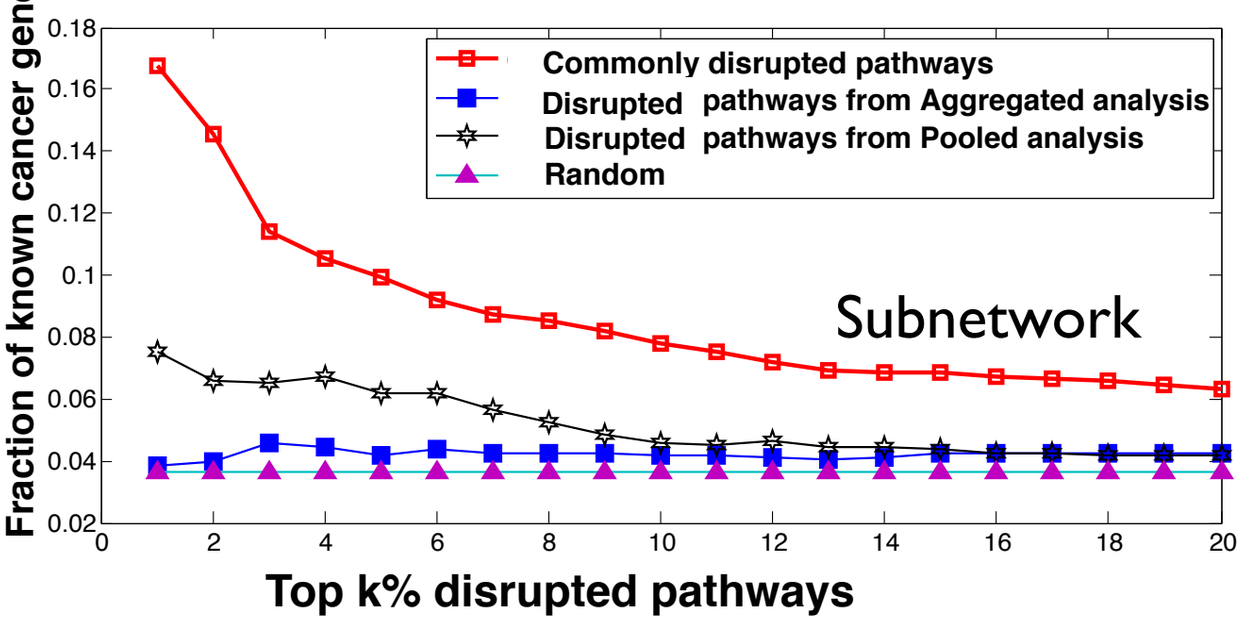
B



C



D



Oncogene (2012) 1 – 9

www.nature.com/Oncogene

ORIGINAL ARTICLE

AR intragenic deletions linked to androgen receptor splice variant expression and activity in models of prostate cancer progression

Y Li[1,10], TH Hwang[1,2,10], LA Oseth[3], A Hauge[4], RL Vessella[5,6], SC Schmechel[7,8], B Hirsch[3,7,9], KB Beckman[4], KA Silverstein[1,2] and SM Dehm[1,7]

1: Joint first author

✓ < 6 months for publication

- 2 months for data generation
- < 1 month for data analysis and validation

Y LI, **T. Hwang** et. al, **Oncogene** 2012

Motivation

• Reactivation of the androgen receptor (AR) during androgen depletion therapy (ADT) underlies castration-resistant prostate cancer (CRPCa).

• Alternative splicing of the AR gene and truncated AR variants lacking the AR ligand binding domain has emerged as an important mechanism of ADT-resistance in CRPCa.

• Truncated AR variants proteins were originally discovered and functionally characterized in the CRPCa 22Rv1 and CWR-R1 cell lines, and the LuCaP 86.2 PCa xenograft

•In a previous study, we demonstrated that altered AR splicing in CRPCa 22Rv1 cells was linked to a 35 kb intragenic tandem duplication of AR exon 3 and flanking sequences

✓ In this study, we wanted to investigate the link between AR gene structure alterations and enhanced synthesis of truncated AR variants in CRPCa CWR-R1 cell lines using paired-end sequencing data

Y LI, **T. Hwang** et. al, **Oncogene** 2012

Data preparation

- 2x76bp paired-end sequencing data using GAIIX illumina with SureSelect
  - 2x50bp paired-end seq using HiSeq
  - 2x76bp paired-end seq using Matepair
  - 2x150bp paired-end seq using MiSeq

- 6000X coverage

Select genomic regions that are interested in

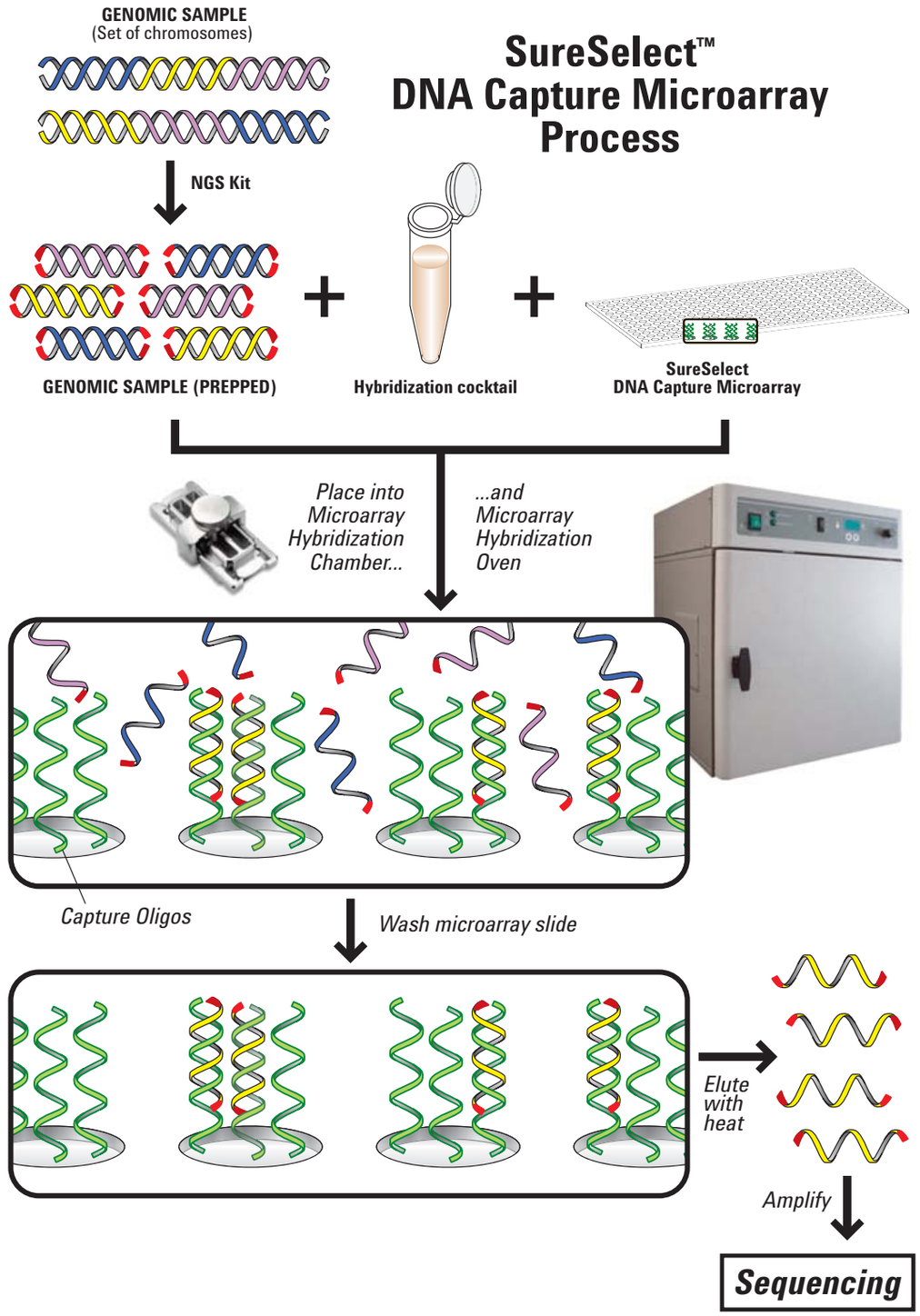
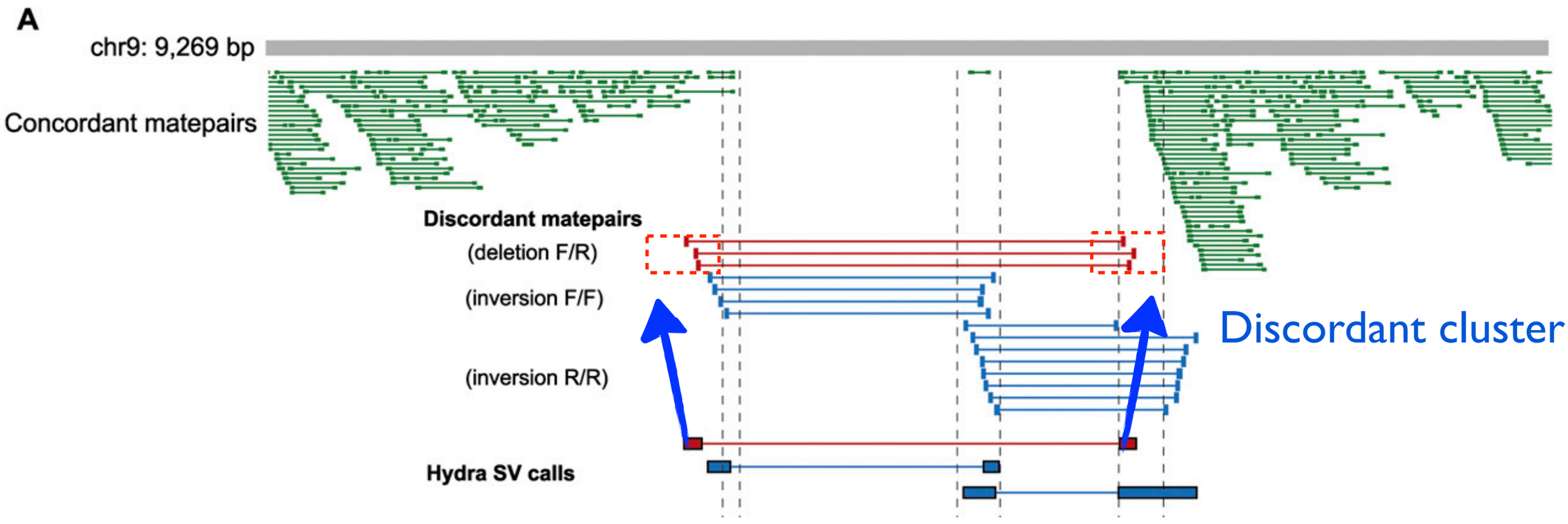


Figure 2 SureSelect Target Enrichment System Capture Process

Y LI, **T. Hwang** et. al, **Oncogene** 2012

Structural Variation Call w/ Hydra



Aaron et al, Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome, Genome biology 2010

Y LI, **T. Hwang** et. al, **Oncogene** 2012

Our pipeline

CRW-R1

1. raw sequences from illumina ← Paired-end reads (N = 11,267,612)

↓

2. Filter raw sequences (Remove poor quality reads) (N = 8,105,919)

↓

3. Convert raw sequences (qseq format) to fastq data

↓

4. Run fastq quality control (fastQC)

↓

Hydra pipeline

5. Align filtered pairs with BWA (N = 7,793,299) → Concordant w/ hg19 7,480,686

↓

6. Collect discordant or unaligned paired reads by BWA (N = 312,613)

↓

7. Re-align discordant or unaligned paired reads by BWA with Novoalign → Concordant w/ hg19 47,472

↓

8. Collect discordant or unaligned paired reads by Novoalign (N = 265,141) → Concordant w/ hg19 785

↓

9. Re-align discordant or unaligned paired reads by Novoalign

↓

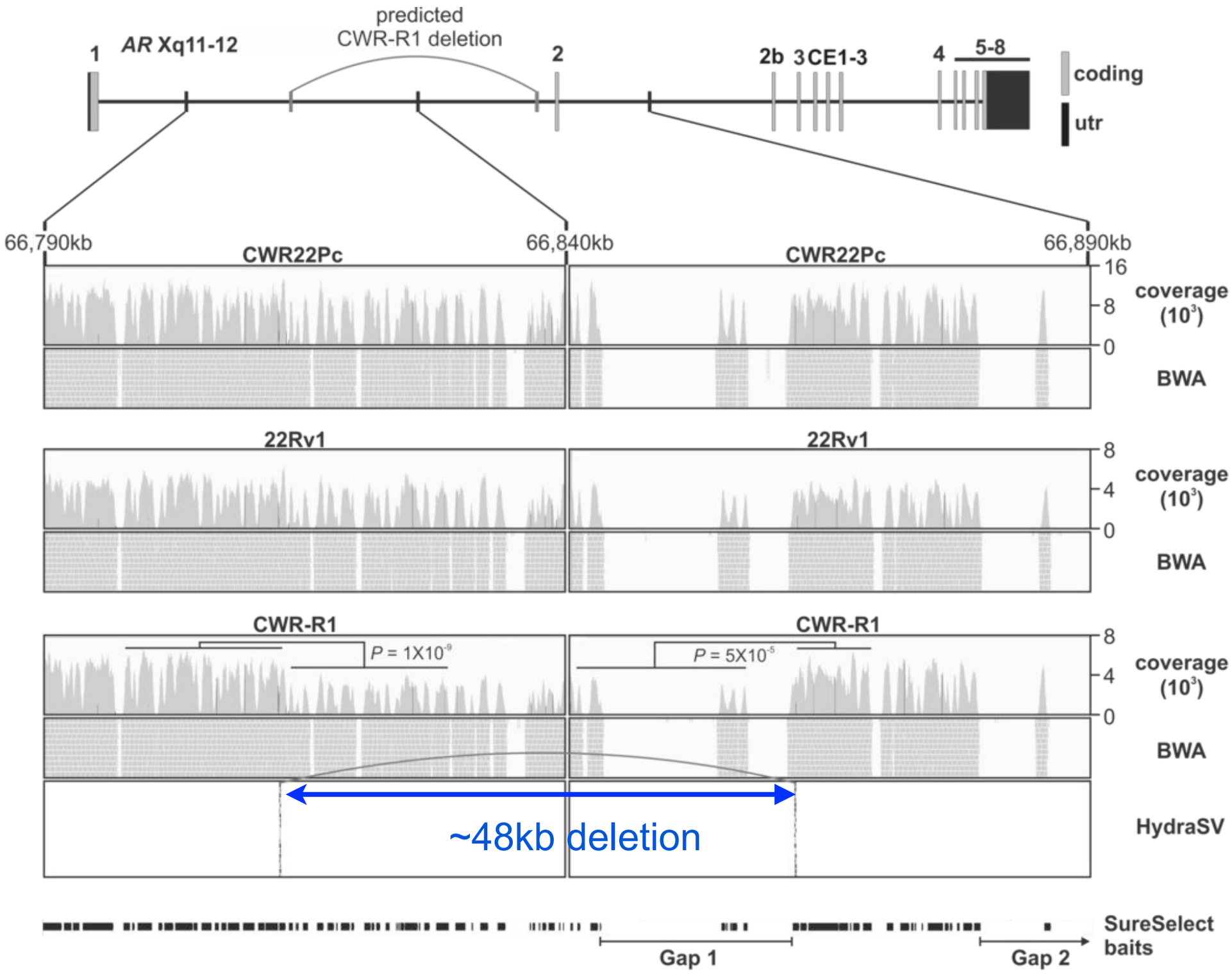
10. Identify Structural Variation (SV) from discordant paired reads using Hydra → 264,356 Discordant pairs yielding 566 discordant mappings

↓

Screen SV calls => 36 final SV calls

Y LI, **T. Hwang** et. al, **Oncogene** 2012

Structural Variation Discovery Visualization



✓ Hydra discovered ~48 kb deletion in AR intron 1 in CWR-R1 cell line

Y LI, **T. Hwang** et. al, **Oncogene** 2012

Validation

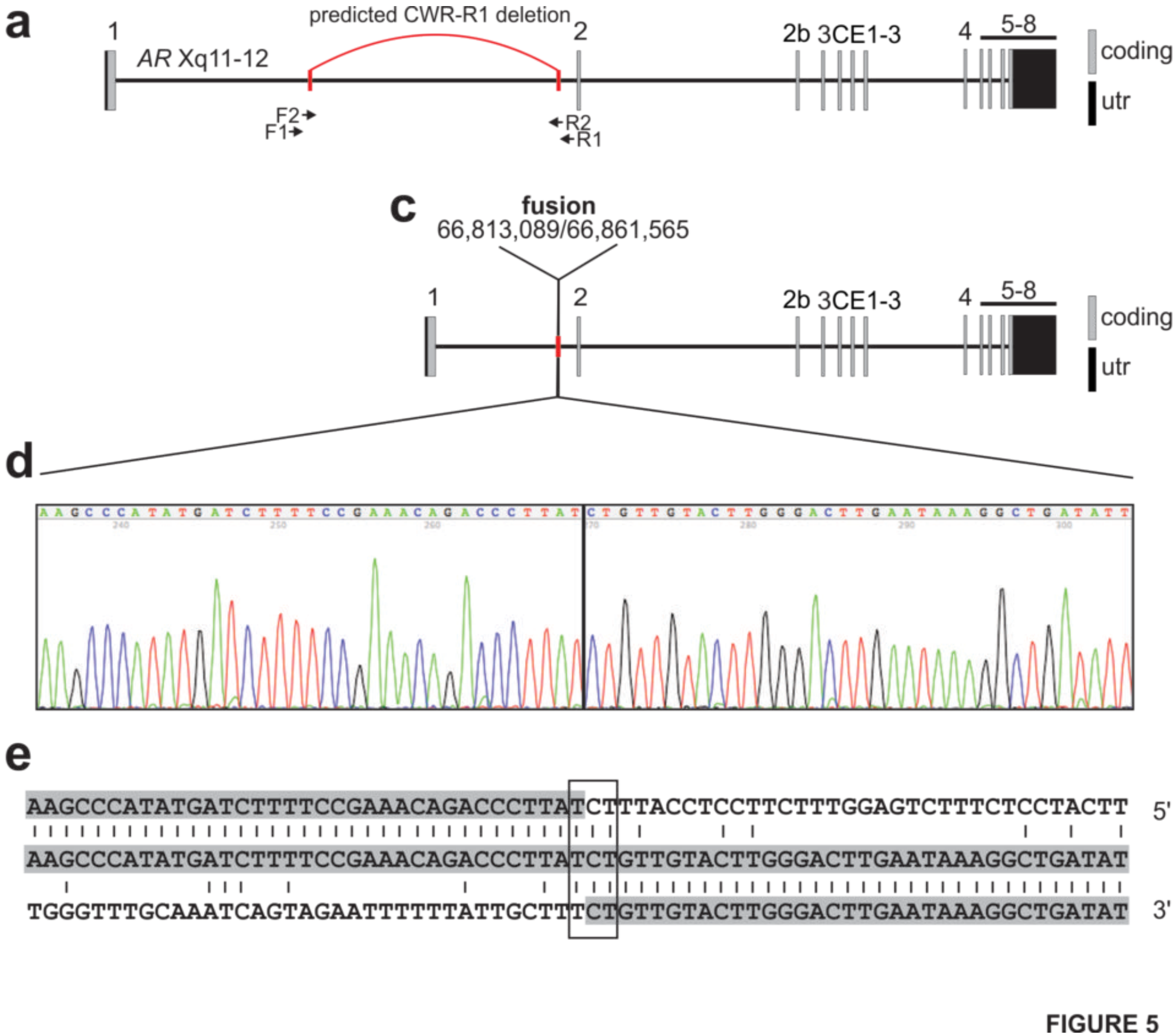
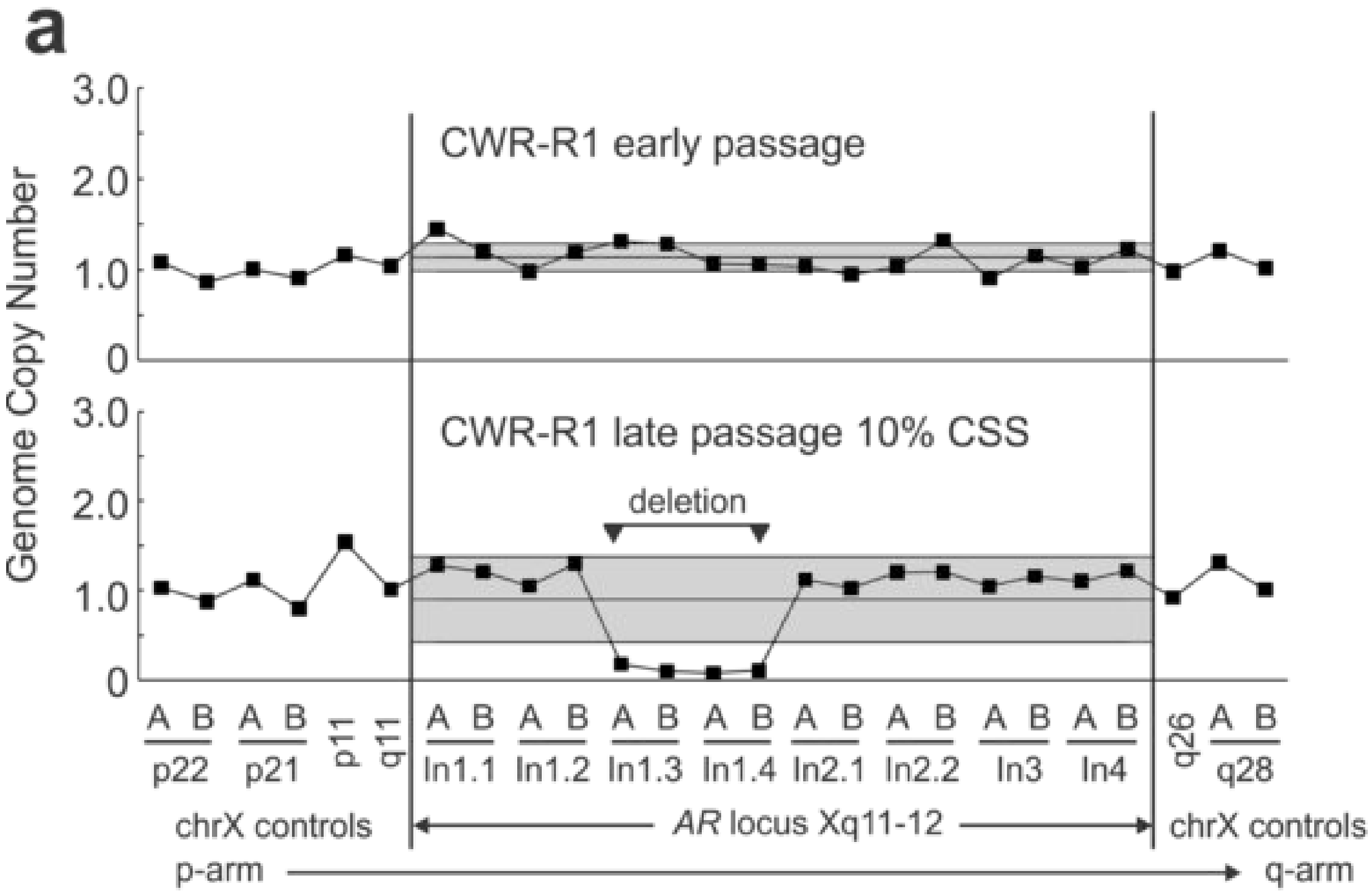


FIGURE 5

Y LI, **T. Hwang** et. al, **Oncogene** 2012

Validation



Y LI, **T. Hwang** et. al, **Oncogene** 2012

Take home message

- Design experiments with both biologists and computational biologists from the beginning (should know which tools will be used)
  - CREST (longer sequences) vs Hydra (more depth coverage)
    - GAIIX, HiSeq, MiSeq, or Mate-pair (sequence length, insertion size)
    - Depth coverage (10X, 100X, or 1500X)
- Start with a small number of genes with higher depth coverage (due to the heterogeneity of cell population)
- Should understand existing tools (e.g., how it works, and what are limitations)
- Quality control!!!!!

•Yingming Li*, **TaeHyun Hwang*,** LeAnn Oseth, Betsy Hirsch, Robert Vessella, Kenny Beckman, Kevin Silverstein, and Scott Dehm, “AR intragenic deletions linked to androgen receptor splice variant expression and activity in models of prostate cancer progression”, **Oncogene 2012**

***Joint first author**

Acknowledgement

University of Minnesota

Rui Kuang, Ph.D
Vipin Kumar, Ph.D
Chad L. Myers, Ph.D
Gotham Alturi
Sean Landman
Ze Tian
Wei Zhang
Maoqiang Xie, Ph.D
Gang Fang

Masonic Cancer Center @ University of Minnesota

Scott Dehm, Ph.D
Jaime Modiano, Ph.D
David Lagaespada, Ph.D
Kevin Silverstein,Ph.D

Boston University

Chang Jin Hong,Ph.D

Mount Sinai School of Medicine

Gaurav Pandey,Ph.D

Mayo Clinic

Dennis Wigle, M.D., Ph.D.
Hugues Sicotte, Ph.D.
Jean-Pierre Kocher, Ph.D.

Dept. of Bioinformatics and Computational Biology @ Genentech

Jinfeng Liu, Ph.D.
Peter Havert, Ph.D.
Zhaoshi Jiang, Ph.D.
Zemin Zhang, Ph. D.