Computational Problems in Cancer Genomics

Eli Upfal

With Fabio Vandin and Ben Raphael

BROWN

June 26, 2000 - Milestone for Humanity

Announcing a “Milestone for Humanity--Decoding the Book of Life” at the White House Ceremony for the Completion of the Human Genome Project

A Milestone for Humanity?

The New York Times

June 12, 2010

“A Decade Later, Genetic Map Yields Few New Cures”

“Ten years after President Bill Clinton announced that the first draft of the human genome was complete, medicine has yet to see any large part of the promised result.”

WHY?

Functional Driven Sequencing - The Cancer Genome Atlas (TCGA)

Compare DNA of cancer and healthy tissue from the same patient - somatic mutation

THE CANCER GENOME ATLAS

The Cancer Genome Atlas (TCGA) Pilot Project

Charting a new course for prevention, diagnosis, and treatment of cancer

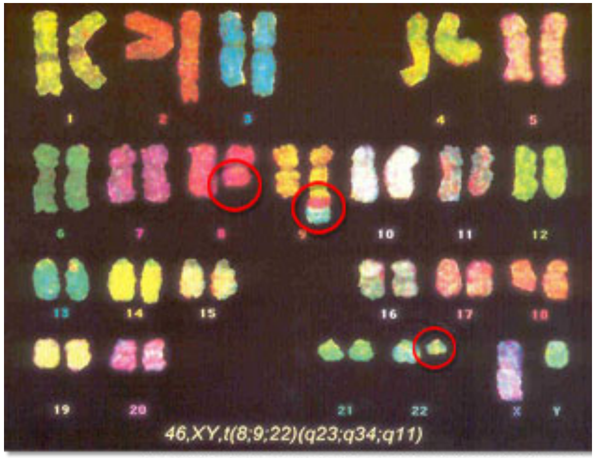
Mutations and other genomic measurements

- Hundreds of cancer samples
- Dozens of cancer types

Statistical approach: Find statistically significant *recurrent mutations*

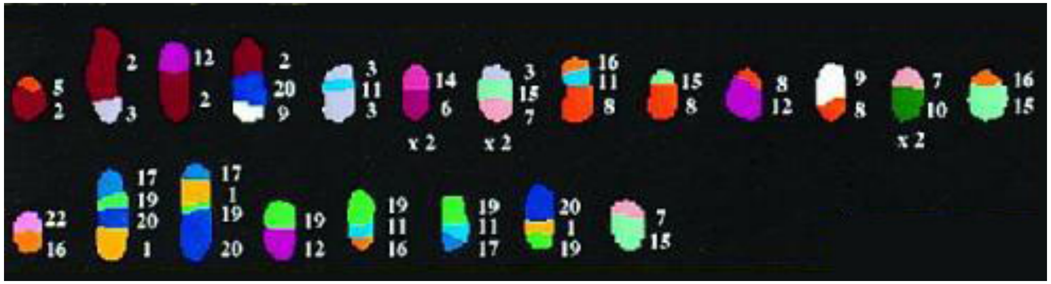
Cancer Genomes - Cancer is a disease of genome alterations

- Many mutations of various types
- Extensive diversity of mutations in tumors
    - Two tumors rarely (never?) have precisely the same set of somatic mutations



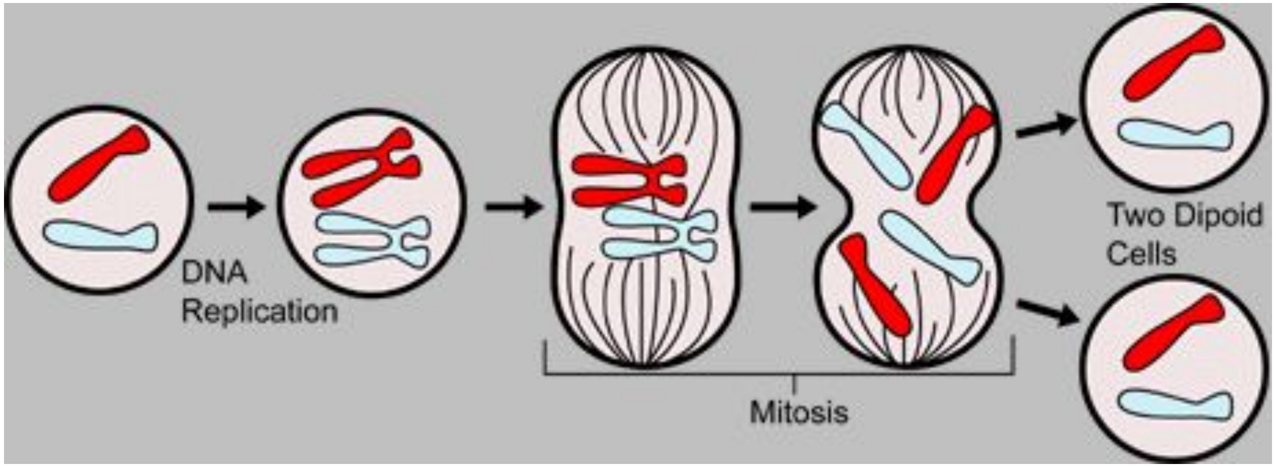
© Copyright 2002, Unistel Medical Laboratories,
Unistel Group Holdings (Pty) Ltd

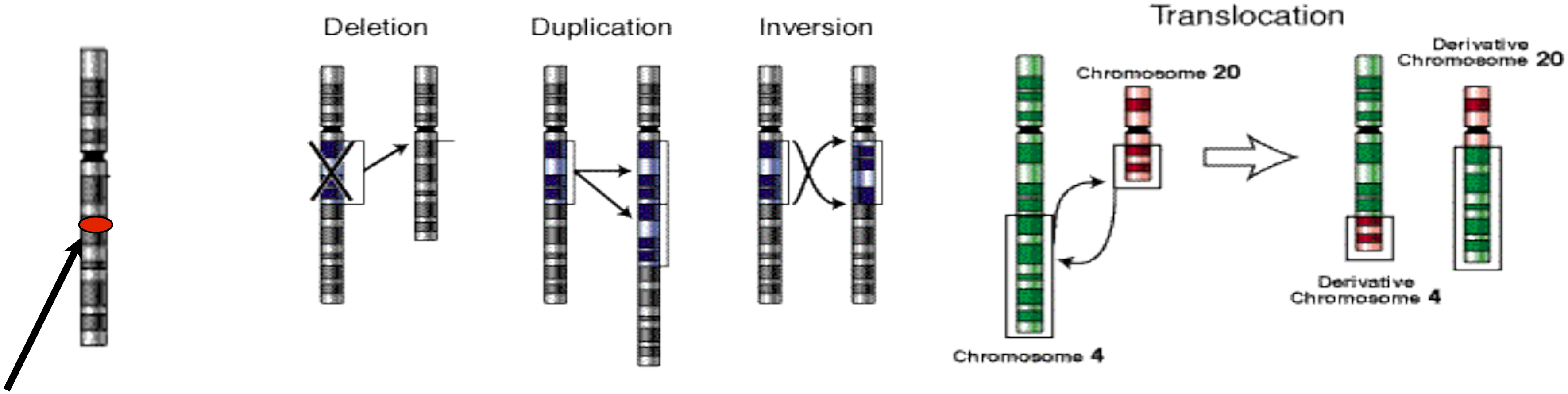
Leukemia



Breast

DNA Replication and Mutation



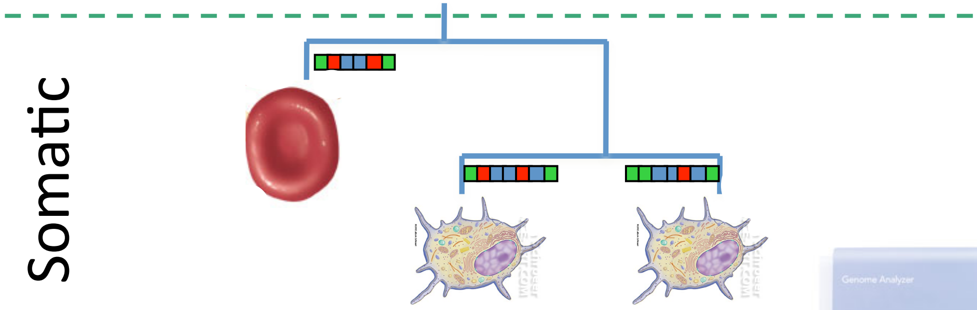


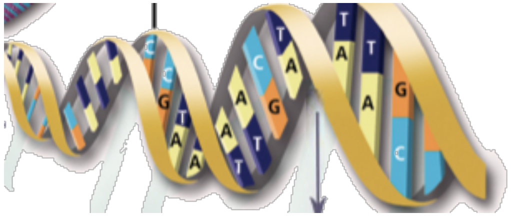
Single Nucleotide variant

Copy number variants

Structural variants

Challenges in Cancer Genomics





Human genome: ~3 billion letters

Reads of 30-1000 letters

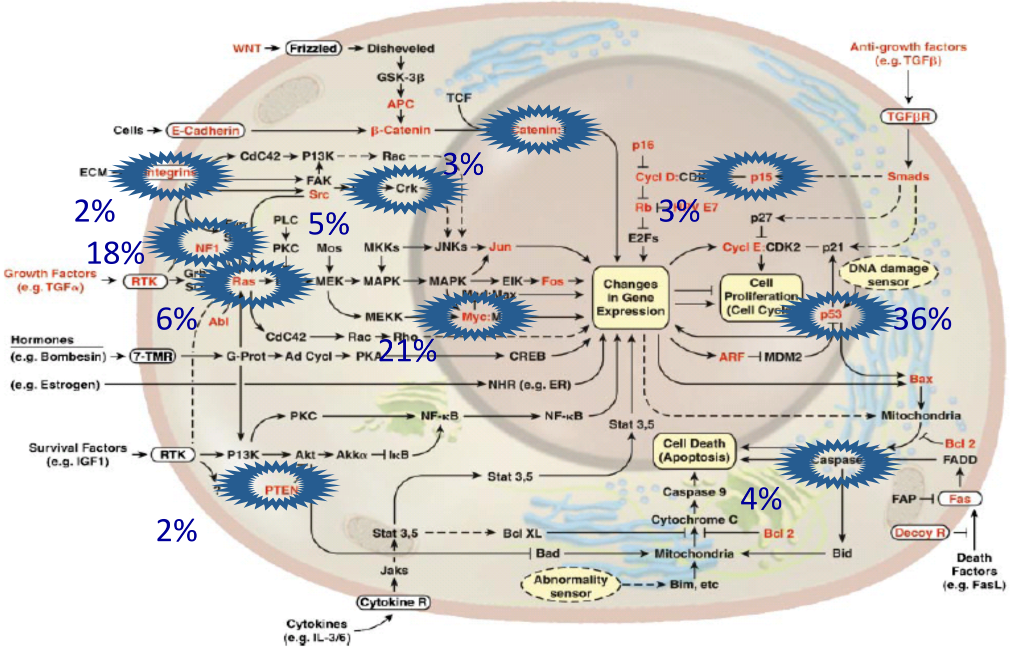
1. **Measurement** of all somatic mutations
2. **Identify** functionally significant mutations

Types of Mutations

- Driver mutations - functionally significant mutations (cause of the cancer)
- Passenger mutations – by product of the cancer process (faulty repair mechanism)

- Goal: identify the the driver mutations
- Problem: There is no small set of mutations that covers all patients

Cancer is a disease of “pathways”



[Hanahan and Weinberg, Cell 2000]

What pathways are altered/mutated?

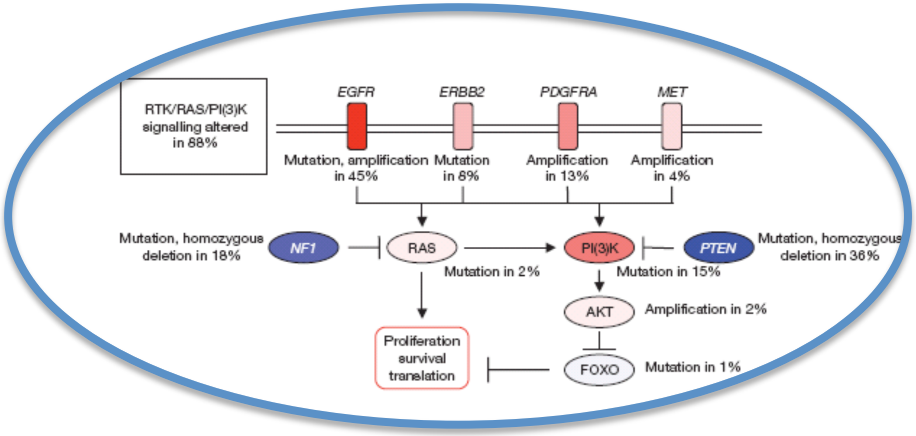
Mutations data

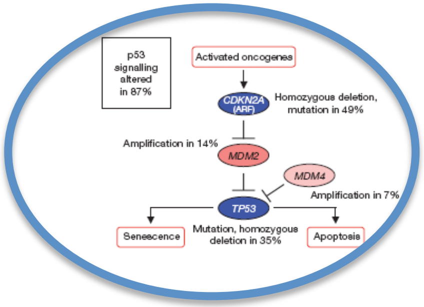
- The *driver* mutations are found in pathways - sets of genes responsible for functions associated with cancer.

- *Passenger mutations are* random mutations that were not repaired because the repair mechanism in cancer cell is broken

Finding Mutated Pathways

Standard practice: assess enrichment of mutations on known pathways

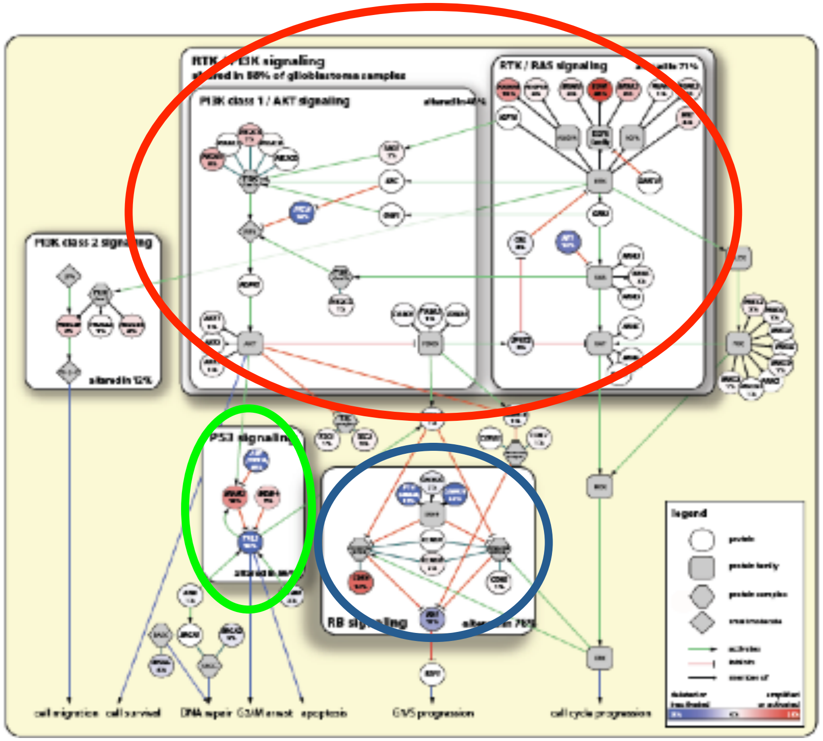

RTK/RAS/PI(3)K signalling altered in 88%
EGFR
ERBB2
PDGFRA
MET
Mutation, amplification in 45%
Mutation in 8%
Amplification in 13%
Amplification in 4%
Mutation, homozygous deletion in 18%
NF1
RAS
Mutation in 2%
PI(3)K
Mutation in 15%
PTEN
Mutation, homozygous deletion in 36%
AKT
Amplification in 2%
Proliferation survival translation
FOXO
Mutation in 1%


p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Mutation, homozygous deletion in 35%
Senescence
Apoptosis

Only known pathways are tested!

Finding Mutated Pathways

Manually constructed small network of interactions

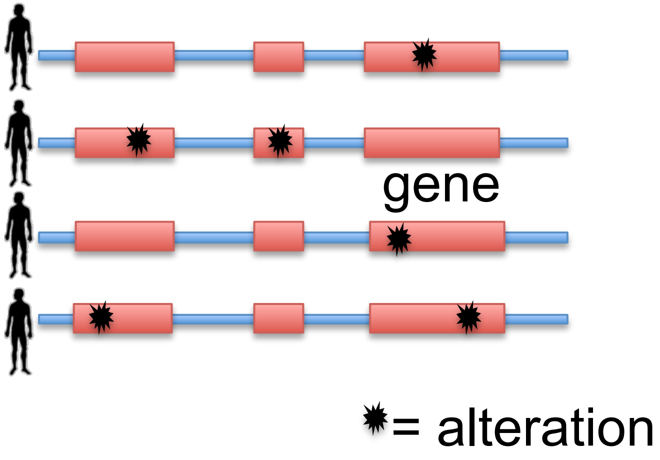


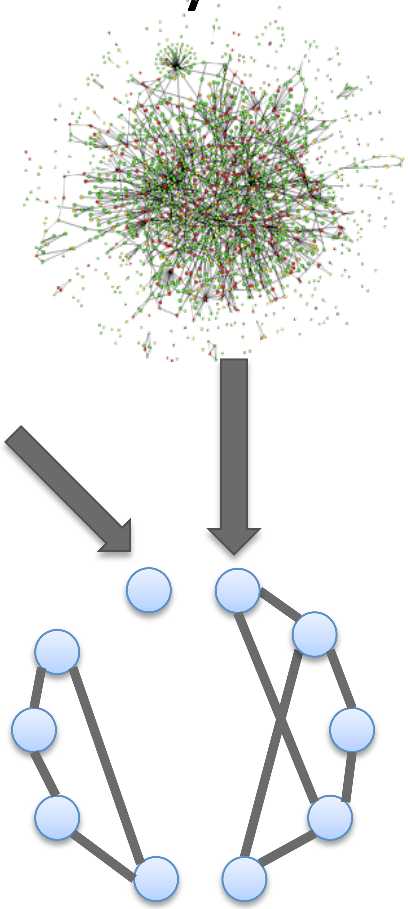
[TCGA, Nature 2008]

Many genes not included!

Network Methods

Use large interaction network to identify mutated subnewtorks





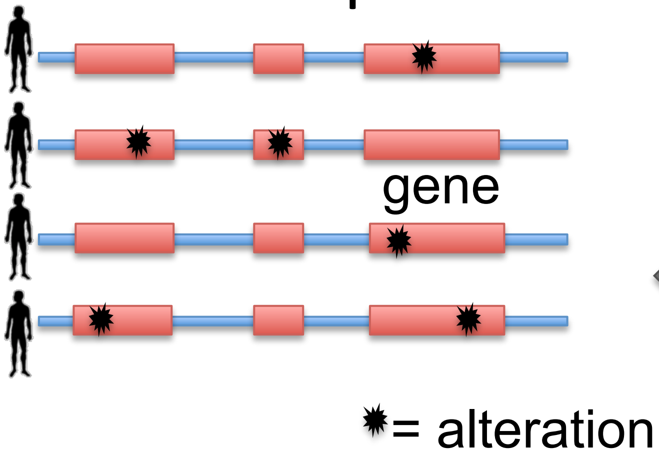
Networks are noisy!

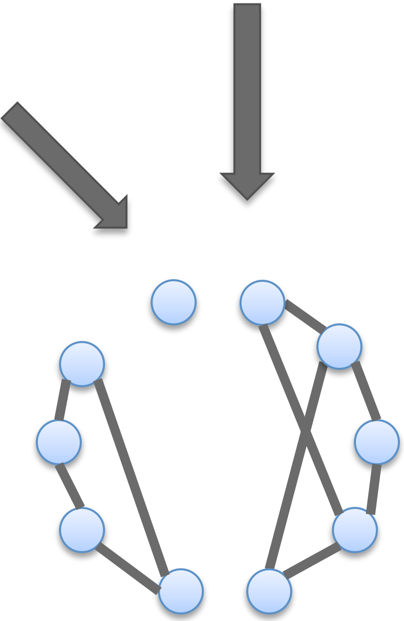
Can we get reliable information?

Problem

Given:

1. Large-scale interaction network
2. Mutation data from multiple cancer samples





Find: Subnetworks mutated in a significant number of samples

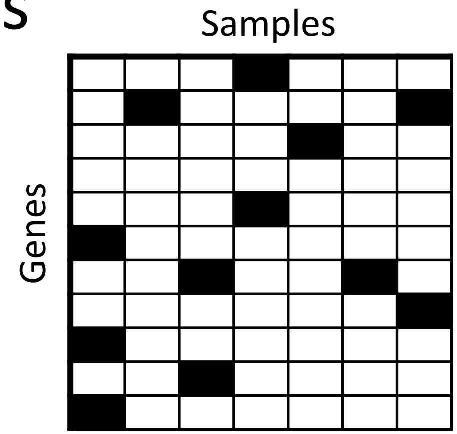
Problem Definition

Given:

1. Interaction network *G* = (*V*, *E*)

   V = genes. *E* = interactions b/w genes

2. Binary alteration matrix



			■			
	■					■
				■		
			■			
■						
		■			■	
						■
■						
		■				
■						

Find: Subnetworks mutated in a significant number of samples

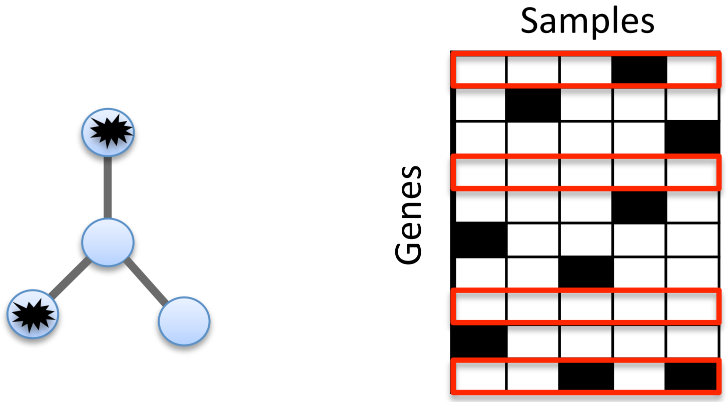
- subnetwork = connected subgraph
- subnetwork *mutated* in sample if ≥ 1 gene mutated in sample

Computational Formulation

For subnetwork *S*:

N_S = number of samples in which *S* is mutated with random alterations

m = number of *observed* samples in which S mutated

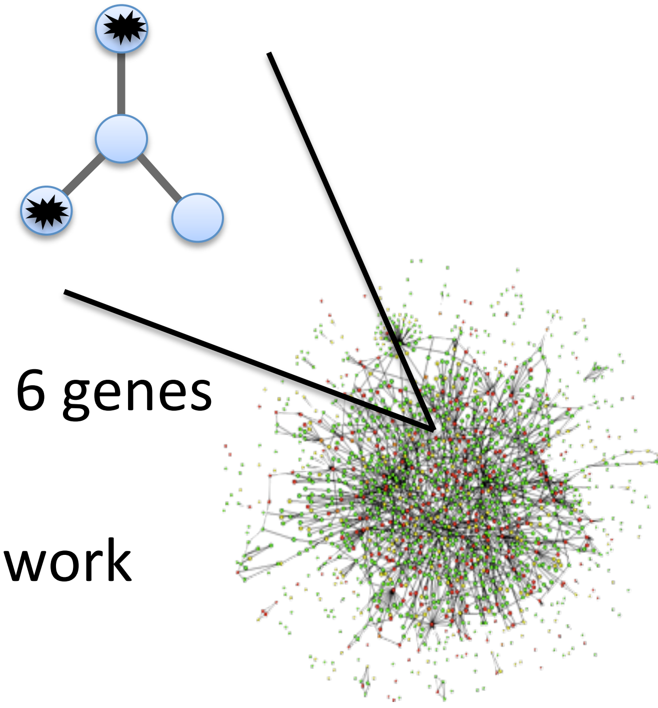


Goal: Find *S* such that Pr[$N_S \geq m$] < ε under suitable *null distribution*

Mutated subnetworks: Naïve Method

Find: S such that $\Pr[N_S \geq m] < \varepsilon$
under suitable null distribution

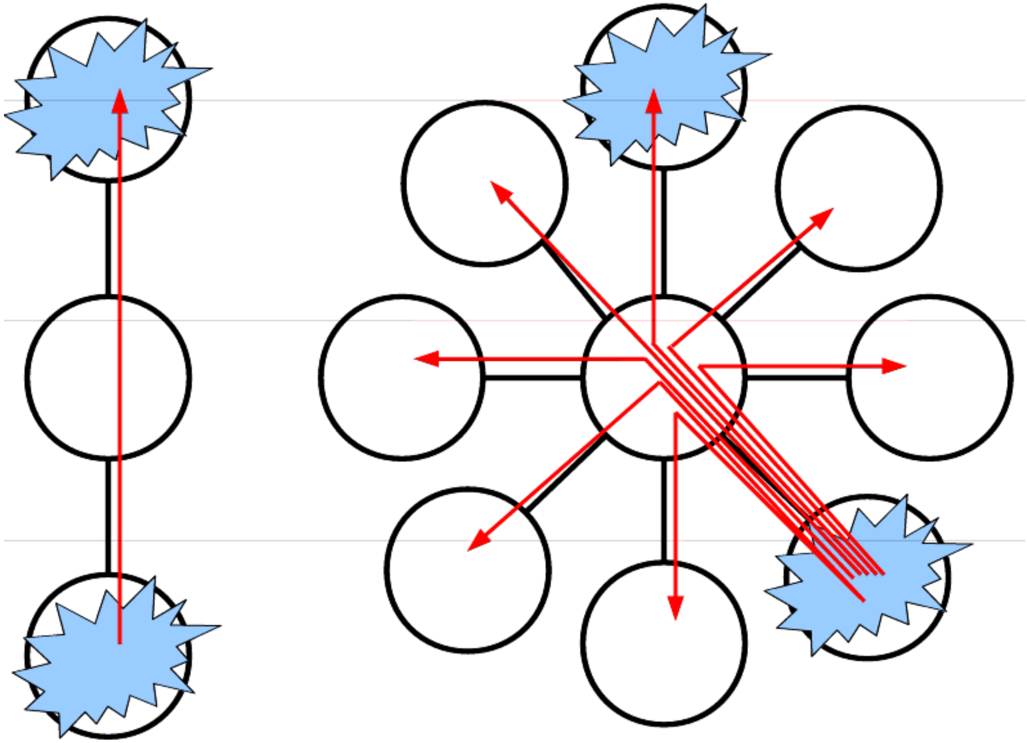
Naïve Method: Test each *S*



Problems

1. Multiple hypothesis testing:
   $> 10^{20}$ candidate subnetworks with < 6 genes
2. Network topology :
   TP53 has 238 neighbors in HPRD network

(Local) Topology Matters

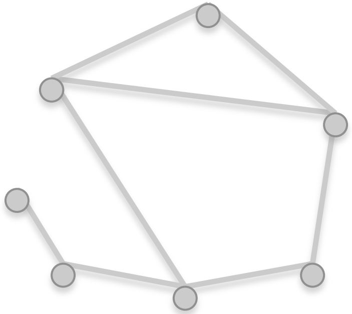


Single path between mutated genes

Path between mutated genes is one of many through node.

Our Contribution

1. *Methods* for *de novo* discovery of mutated subnetworks
   I. Combinatorial model
   II. Enhanced influence model

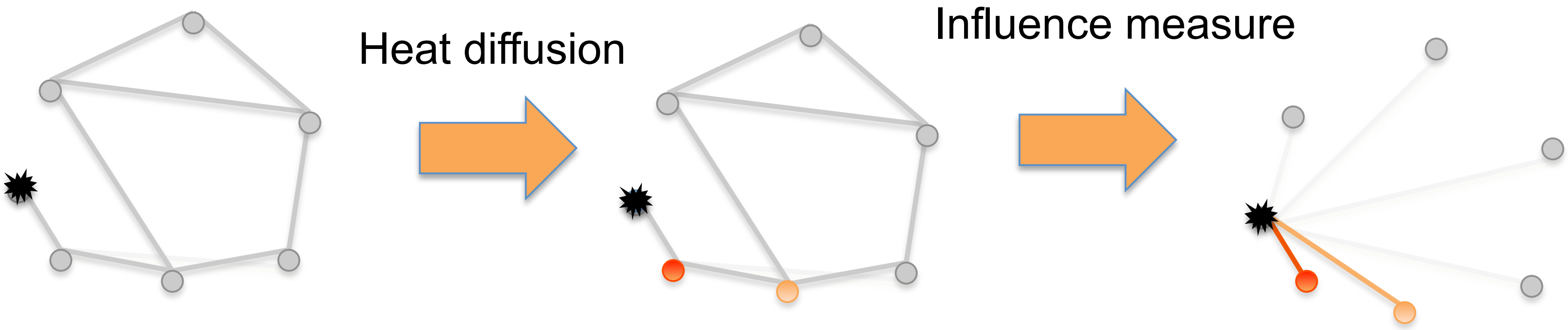


2. Definition of *Influence Graph*:
   Identify subnetworks using both frequency of alteration *and* network topology

3. *Statistical tests* to assess the significance

Influence Graph

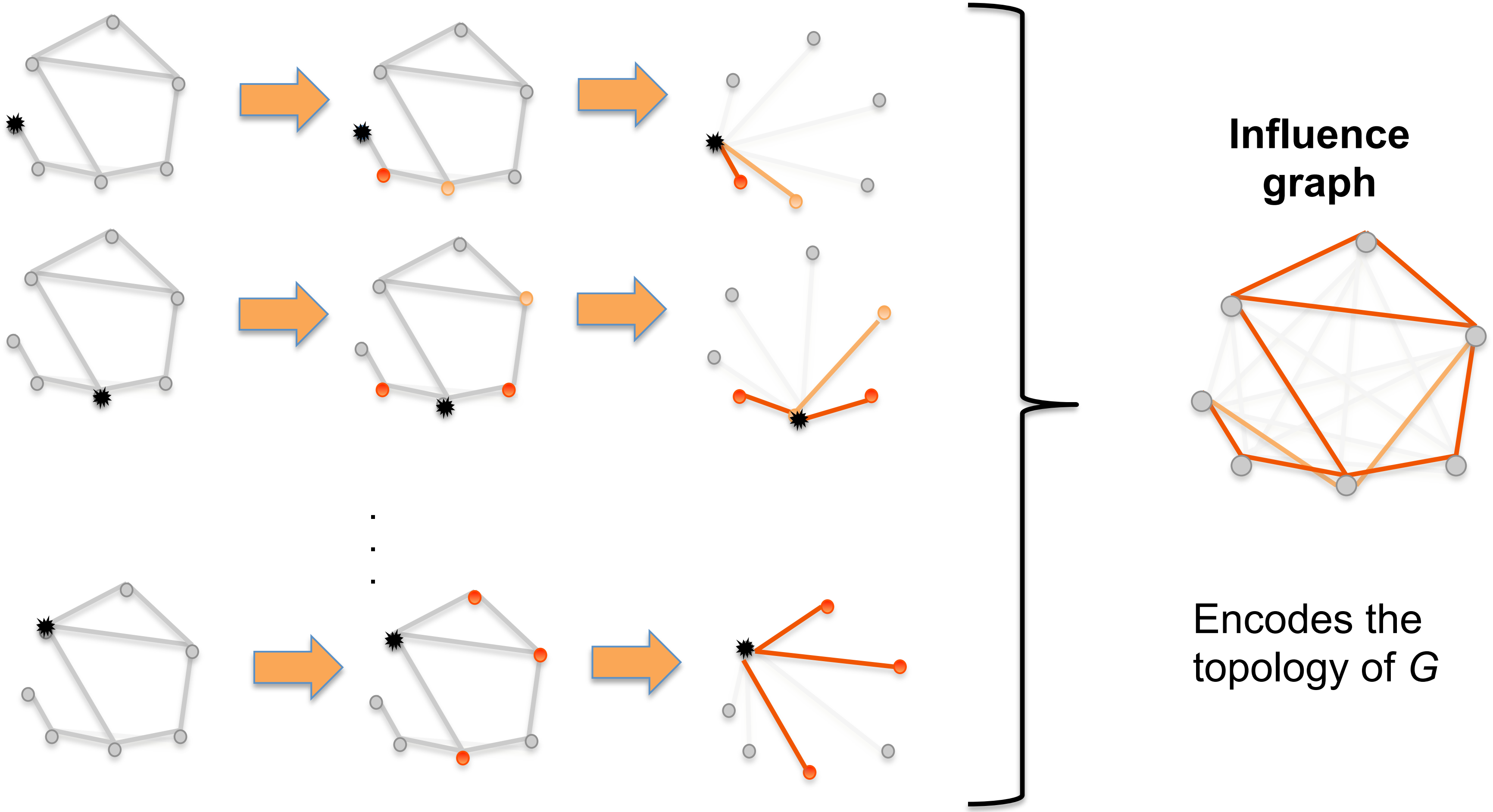
✸ alteration = unit source of heat



Easily derived from *Laplacian matrix* of *G*

Influence Graph

✸ alteration = unit source of heat



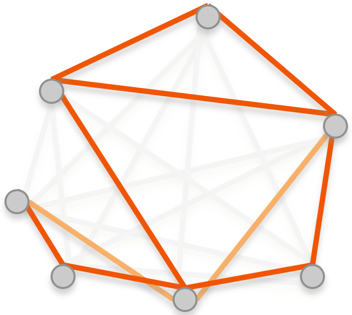
Influence graph

Encodes the topology of *G*

Heat equation

$f(t) = (f_1(t), \ldots, f_n(t))^T$

heat on vertices at time *t*.



$$\frac{df_i}{dt} = \sum_j a_{ij}(f_j(t) - f_i(t))$$

$df/dt = (A - D)\, f(t)$ $A = [a_{ij}]$ = **adjacency matrix** of *G*.

$f(t) = e^{-L\,t} f(0)$ $L = D - A$ **= Laplacian matrix** of *G*.

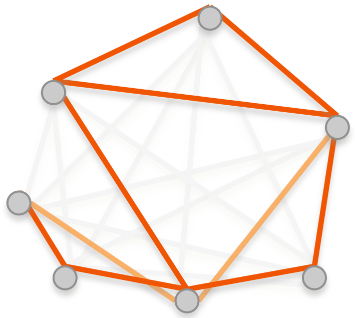
$e^{-L\,t}$ is **heat kernel** of *G*

Discovering Significant Subnetworks

Two approaches:

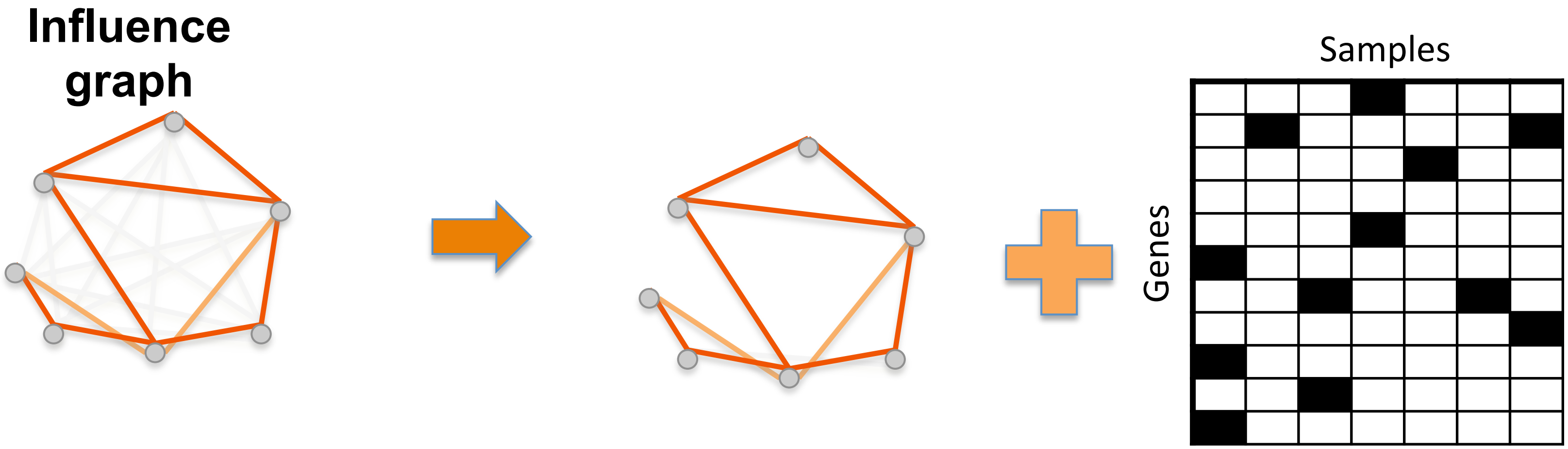
1. Combinatorial Model
2. Enhanced Influence Model

Based on Influence Graph



Statistical tests to assess significance

Combinatorial Model



Fix *K*: find the subnetwork with *K* genes mutated in the maximum number of samples

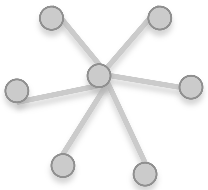
→ **Connected maximum coverage** problem

("graph version" of maximum coverage problem – NP-Hard)

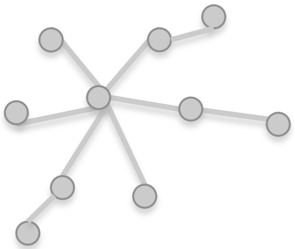
Connected maximum coverage problem

1. Thm. NP-Hard for general graphs.

2. Thm. NP-Hard for star graphs.



3. Thm. 1 – 1/e approx. alg. for spider graphs



4. Thm. 1/(cr) approx. alg. for general graphs
   - c=(2e-1)/(e-1)
   - r= radius of the optimal solution in G

Combinatorial Model: Statistical Test

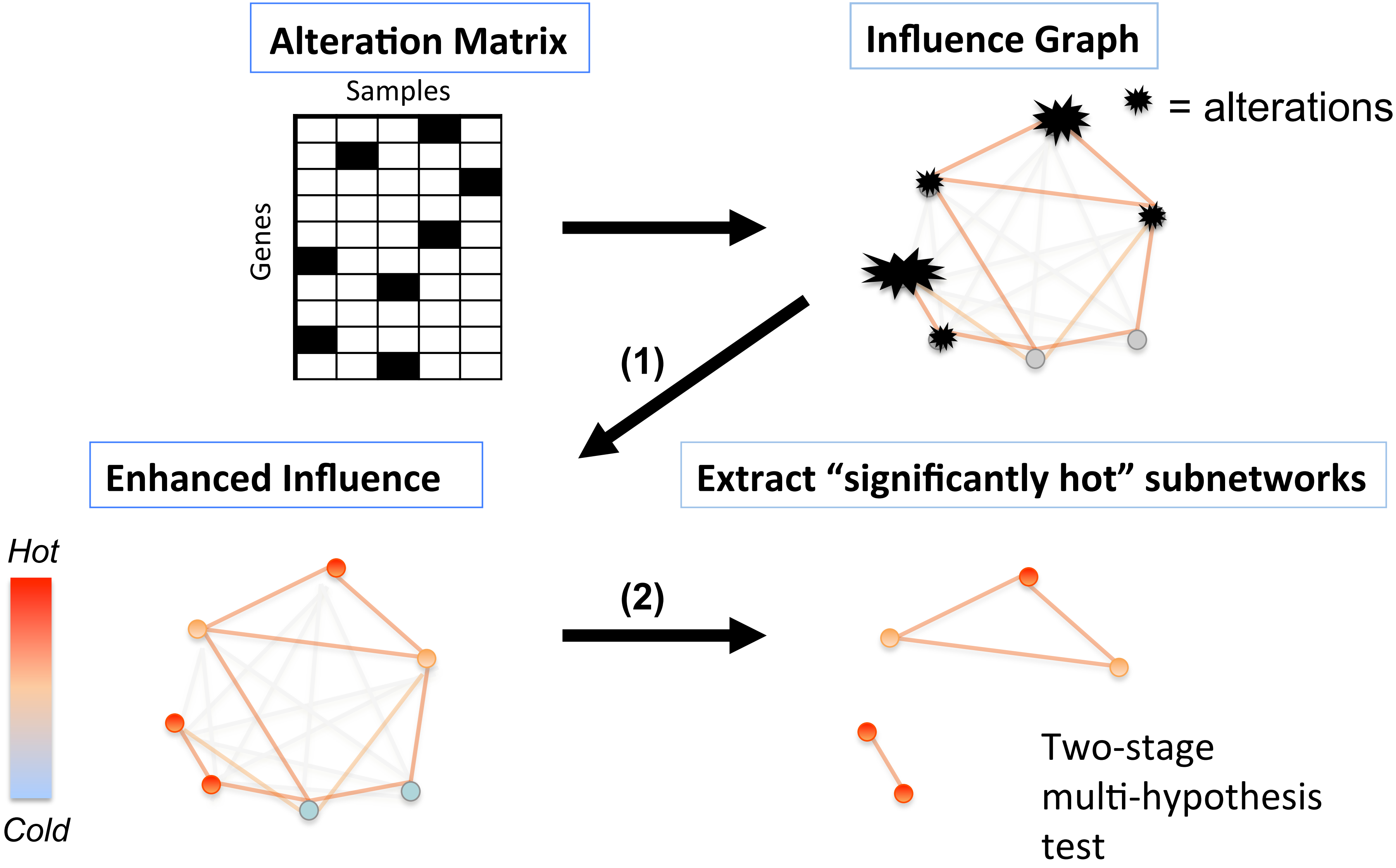
Fix *K*: find the subnetwork with *K* genes mutated in the maximum number of samples

testing the number of altered samples

only 1 hypothesis → no multiple correction!

Limitation: inadequate representation of *heterogeneity* of cancer alterations

Enhance Influence Model (EIM)



Alteration Matrix

Influence Graph

Enhanced Influence

Extract “significantly hot” subnetworks

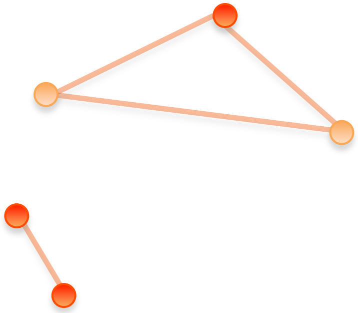
			■	
	■			
				■
			■	
■				
		■		
■				
		■		

Two-stage multi-hypothesis test

EIM: Statistical test

X_s = ***number*** of subnetworks with $\geq s$ genes using "random" alteration matrix.

$H_0^s : X_s \geq \eta_s$, $s = 1, \ldots, N$ = # genes.



2 subnetworks with 2 or more genes

Two-stage multi-hypothesis test

1. Let s^* = smallest s where H_0^s is rejected.

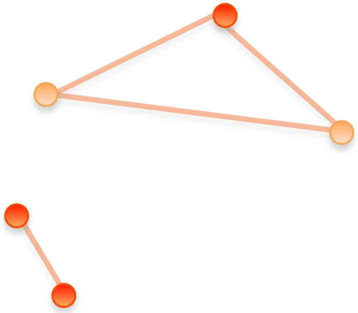
$$\Pr [X_s \geq \eta_s] < \alpha \;/\; N \quad \text{(Bonferroni correction)}$$

# hypotheses = #$s \leq$ # measured genes.

EIM: Statistical test

X_s = ***number*** of subnetworks with $\geq s$ genes using "random" alteration matrix.

$H_0^s : X_s \geq \eta_s$, $s = 1, \ldots, N$ = # genes.



2 subnetworks with 2 or more genes

Two-stage multi-hypothesis test

2. Bound false discovery rate (FDR) for ***list of identified*** subnetworks.

Thm. Fix $\beta_1, \ldots, \beta_N$ such that $\Sigma_i \beta_i \leq \beta$. Let s^* be smallest s such that $\eta_s \geq E[X_s] / \beta_s$. If return all subnetworks of size $\geq s^*$ as significant, then FDR $\leq \beta$.

Two Stage Statistical Test

- Instead of testing the significance of the support of individual itemsets we test the significance of the number of itemsets with a given support
- The null hypothesis distribution is specified by the Poisson approximation result
- Reduces the number of simultaneous tests
- More powerful test – less false negatives

[JACM 2012 - Kirsch, Mitzenmacher, Pietracaprina, Pucci, U, Vandin]

Test I

- Define $\alpha_{1,}\ \alpha_{2,}\ \alpha_{3,}$ **…** such that $\sum \alpha_i \leq \alpha$
- For **i=0,…,log** $(s_{max} - s_{min}) + 1$
  - $s_i = s_{min} + 2^i$
  - $Q(k, s_i)$ = observed number of itemsets of size **k** and support $\geq s_i$
  - $H_0(k,s_i)$ = "$Q(k,s_i)$ conforms with **Poisson**(λ_i)"

  - Reject $H_0(k,s_i)$ if **p-value** $< \alpha_i$

Test I

- Let **s*** be the smallest **s** such that **H_0 (k,s)** rejected by Test I
- With confidence level α the number of itemsets with support ≥ **s*** is significant
- Some itemsets with support ≥ **s*** could still be false positive

Test II

- Define $\beta_1, \beta_2, \beta_3, \ldots$ such that $\sum \beta_i \leq \beta$
- Reject $\mathbf{H_0 (k,s_i)}$ if:

  $\mathbf{p\text{-}value} < \alpha_i$ and $\mathbf{Q(k,s_i)} \geq \lambda_i / \beta_i$

- Let **s*** be the minimum **s** such that $\mathbf{H_0(k,s)}$ was rejected
- If we flag all itemsets with support **≥ s*** as significant, **FDR ≤** β

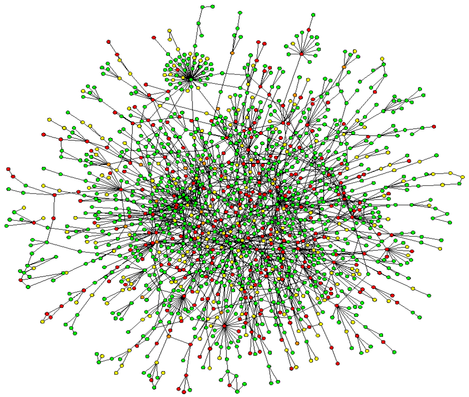
Proof

- $\mathbf{V_i}$ = false discoveries if $\mathbf{H_0(k,s_i)}$ first rejected
- $\mathbf{E_i}$ = "$\mathbf{H_0(k,s_i)}$ rejected"

$$\begin{aligned} FDR &= \sum_{i=0}^{h-1} E\left[\frac{V_i}{Q_{k,s_i}}\right] \mathbf{Pr}(E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0) \\ &\leq \sum_{i=0}^{h-1} \frac{E[X_i \mid E_i \bar{E}_{i-1}, \ldots, \bar{E}_0]}{\lambda_i/\beta_i} \mathbf{Pr}(E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0) \\ &= \sum_{i=0}^{h-1} \frac{\sum_j j \mathbf{Pr}(X_i = j, E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0)}{\lambda_i/\beta_i} \\ &\leq \sum_{i=0}^{h-1} \frac{\beta_i \lambda_i}{\lambda_i} \leq \sum_{i=0}^{h-1} \beta_i \leq \beta. \end{aligned}$$

□

Experimental Results



Interaction network

HPRD: 18796 nodes, 37107 edges

Datasets

1. **Glioblastoma Multiforme (GBM)** [TCGA, *Nature*, 2008]

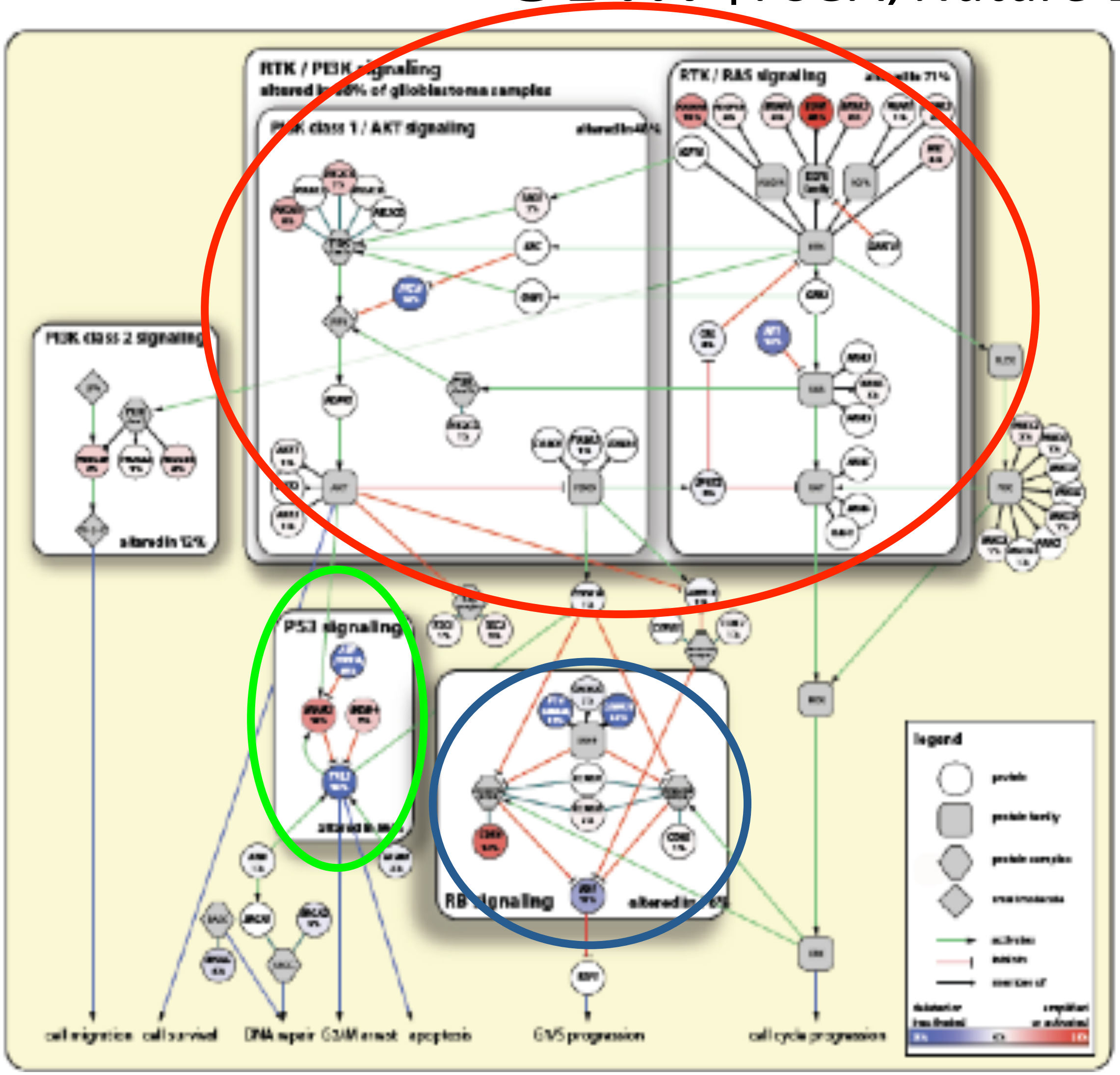
   601 sequenced genes in 91 samples

   Array copy number data on *all* genes

2. **Lung Adenocarcinoma** [Ding et al., *Nature*, 2008]

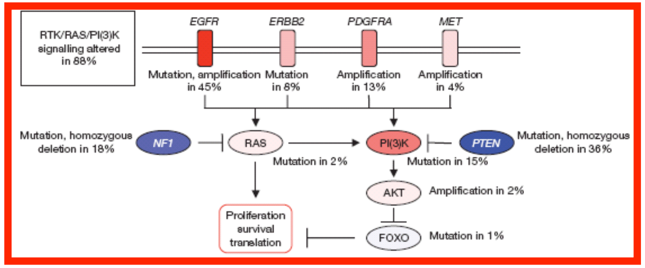
   623 sequenced genes in 188 samples

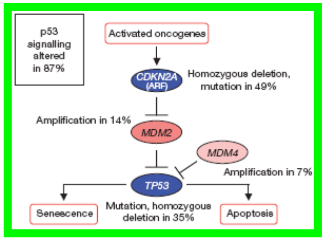
GBM [TCGA, *Nature* 2008]



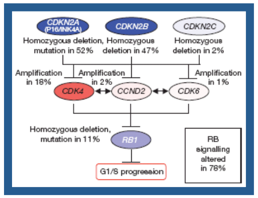
Manually created

RTK/RAS/PI(3)K


RTK/RAS/PI(3)K signalling altered in 88%
EGFR
ERBB2
PDGFRA
MET
Mutation, amplification in 45%
Mutation in 8%
Amplification in 13%
Amplification in 4%
Mutation, homozygous deletion in 18%
NF1
RAS
Mutation in 2%
PI(3)K
PTEN
Mutation, homozygous deletion in 36%
Mutation in 15%
AKT
Amplification in 2%
Proliferation survival translation
FOXO
Mutation in 1%


p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Senescence
Mutation, homozygous deletion in 35%
Apoptosis

p53


CDKN2A (P16/INK4A)
CDKN2B
CDKN2C
Homozygous deletion, mutation in 52%
Homozygous deletion in 47%
Homozygous deletion in 2%
Amplification in 18%
Amplification in 2%
Amplification in 1%
CDK4
CCND2
CDK6
Homozygous deletion, mutation in 11%
RB1
RB signalling altered in 78%
G1/S progression

RB1

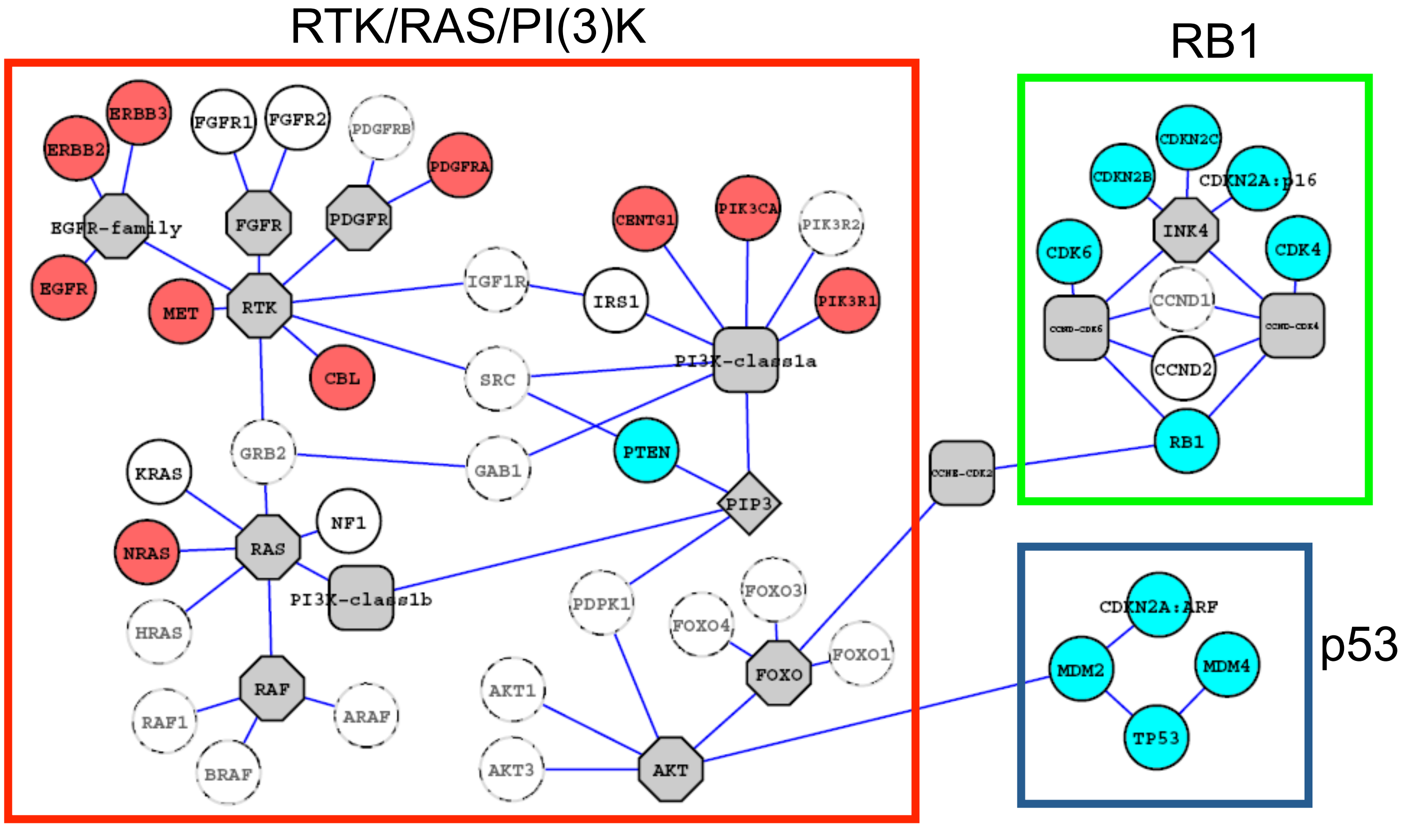
Significant?

GBM: Mutations + Copy number

			Enrichment *p*-val		
s	#net ≥ *s*	*p*-val	RTK/RAS/PI(3)K	P53	RB1
20	2	$<10^{-2}$	0.69	$2\text{x}10^{-6}$	$4\text{x}10^{-8}$
26	1	$5\text{x}10^{-2}$	10^{-8}	-	-

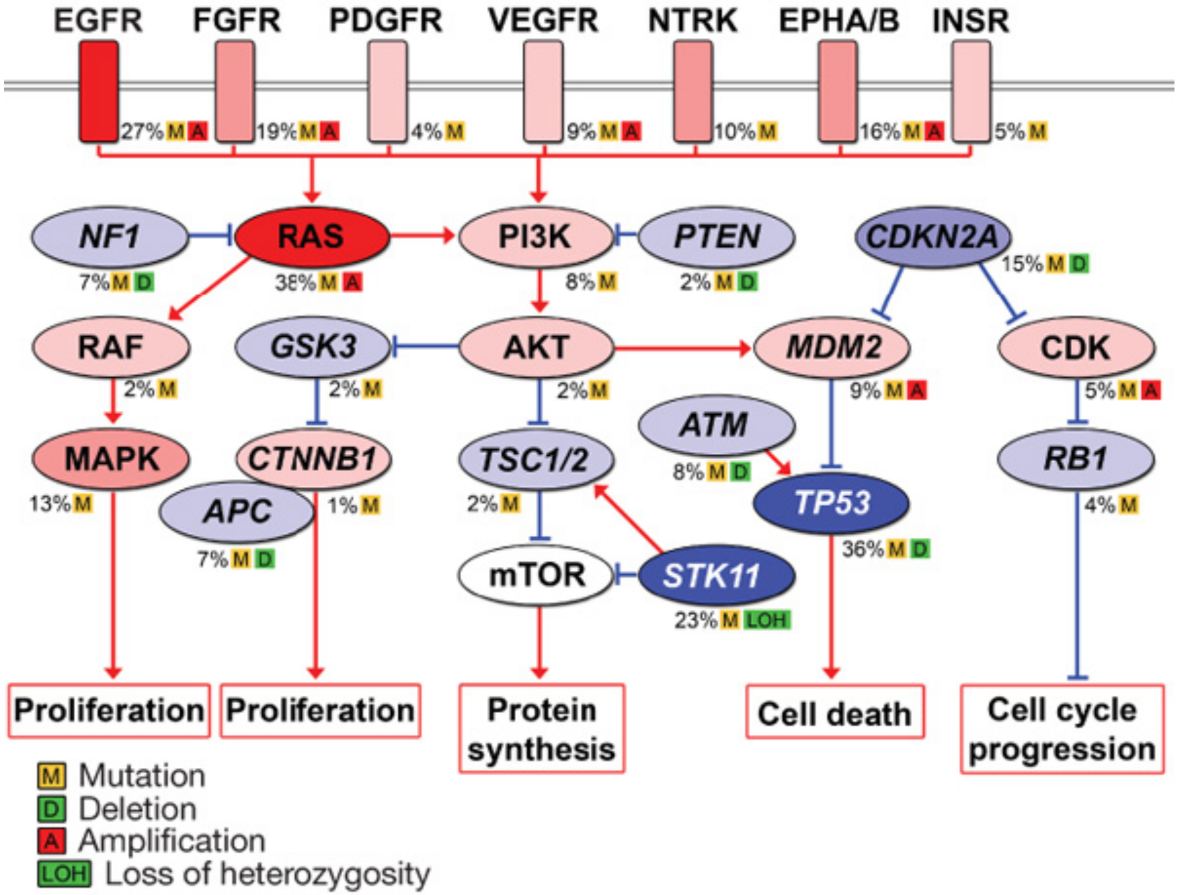
FDR <0.1

total enrichment for **s ≥** 20 : *p* < 10^{-2}



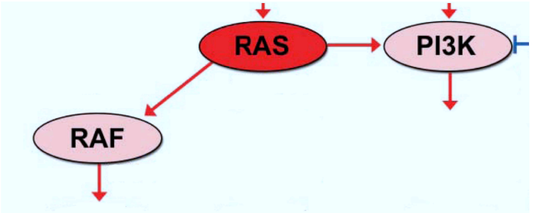
Lung Adenocarcinoma

[Ding et al., *Nature* 2008]



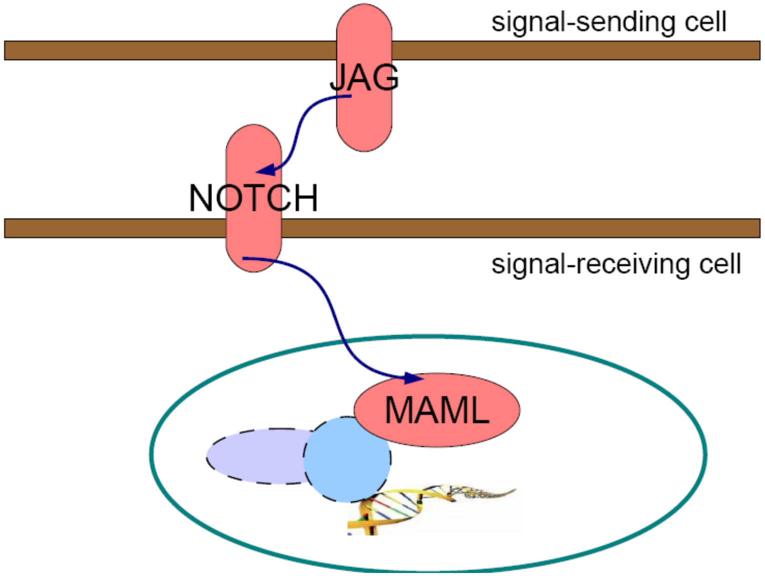
Results: Lung Adenocarcinoma

			enrichment
s	**#net ≥ *s***	***p*-val**	**KEGG pathway/*p*-val**
6	3	$<10^{-2}$	Notch signaling/$2\text{x}10^{-9}$
8	2	$<10^{-2}$	MAPK signaling/$3\text{x}10^{-2}$
48	1	$<10^{-2}$	p53 signaling/$7\text{x}10^{-4}$



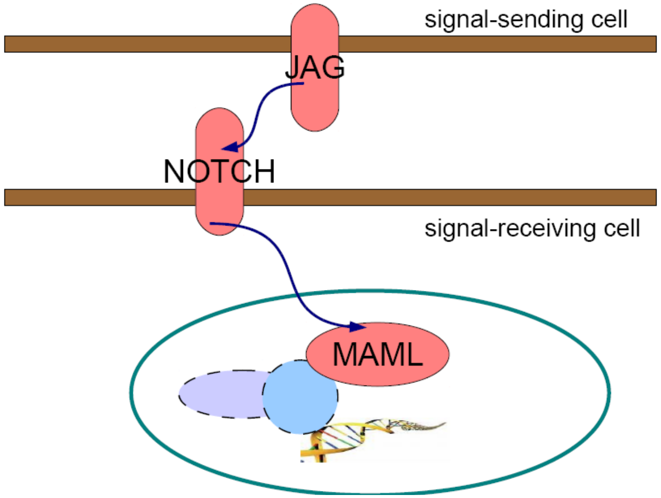
FDR <0.07

total enrichment for **s ≥** 6 : $p < 7\text{x}10^{-9}$



Notch signaling

Lung Adenocarcinoma: Notch



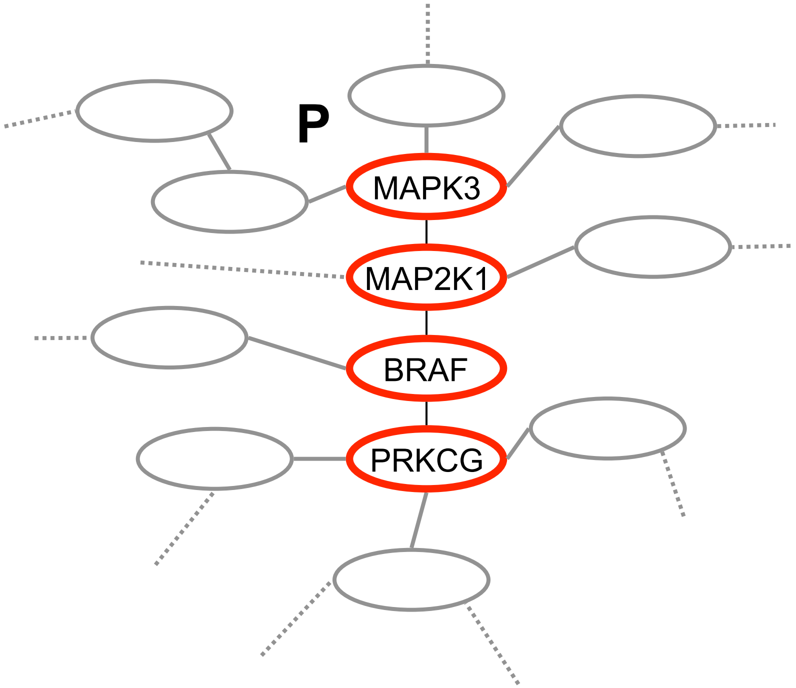
Implicated in a variety of cancers including lung

[Axelson, *Sem. Cancer Biol*. 2004, Collins et al., *Sem. Cancer Biol.* 2004]

Gene	# samples
JAG2	3
NOTCH2	1
NOTCH3	2
NOTCH4	3
MAML1	3
MAML2	1

Not reported in Ding et al. [*Nature* 2008]

Simulated data

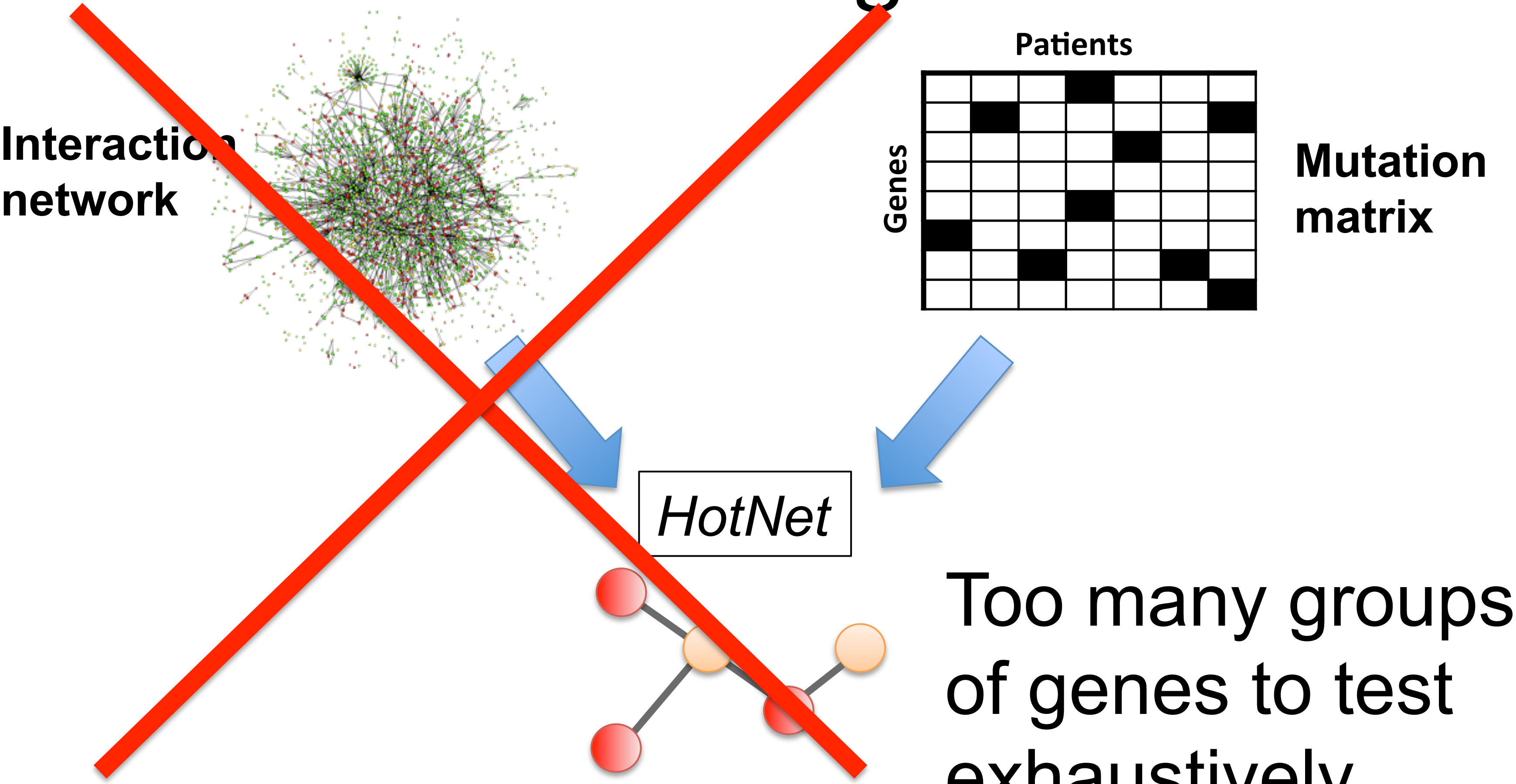


- **Graph**: KEGG pathway + random interactions
    - 258 genes
    - 1762 “real” edges
    - 440 random edges
- **Alteration Matrix**
    - 30 tested genes including **P**
    - Random mutations (parameters from real data)
    - Mutations in **P** (17% of samples)

s	#c.c.≥s	FDR	*p*-val
4	1	$<10^{-2}$	$<10^{-2}$

- removing mutations in **P**: nothing significant
- making BRAF hub: nothing significant

Dendrix: Removing the Network



Interaction network

Patients

Genes

Mutation matrix

HotNet

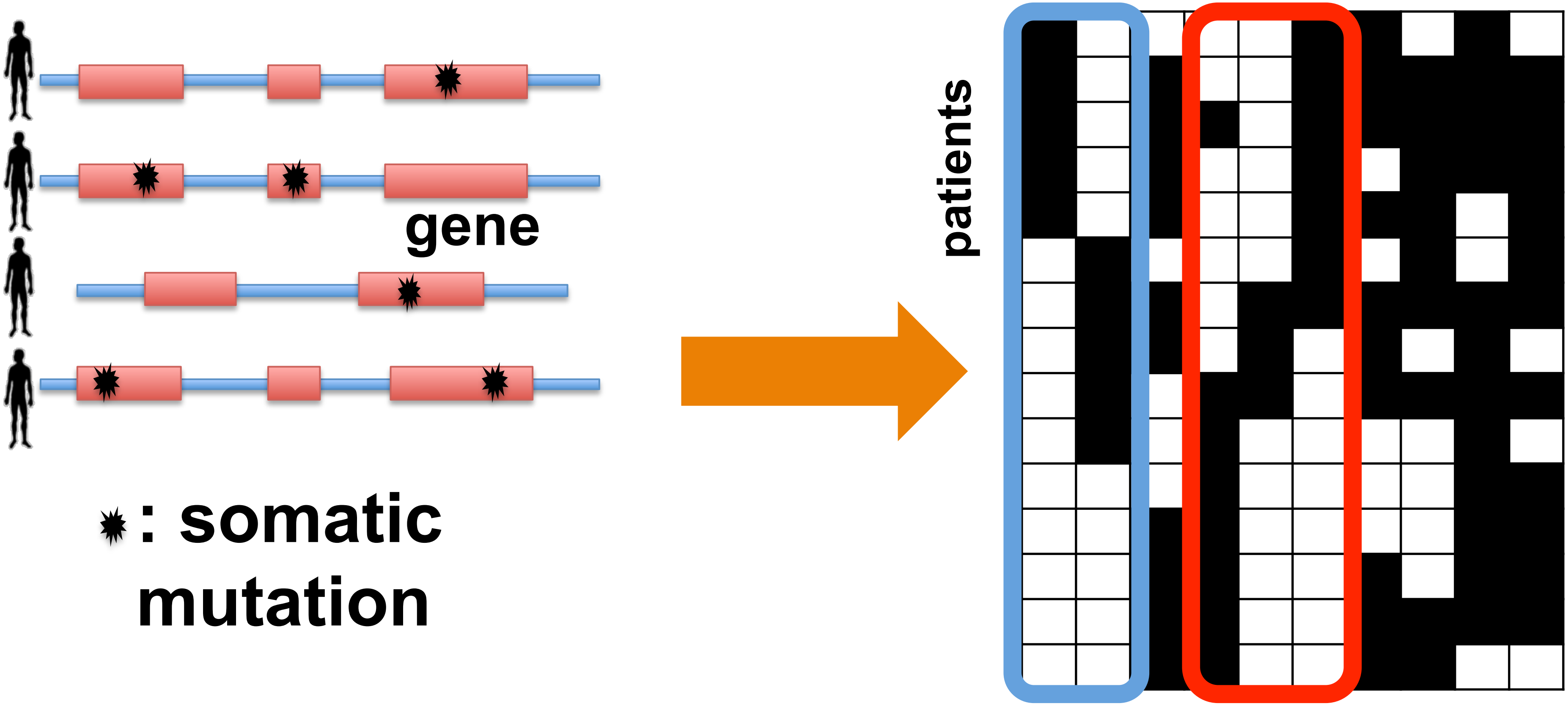
Too many groups of genes to test exhaustively.

Networks are noisy.
Do we need them?

Genomes

Mutation Matrix

genes



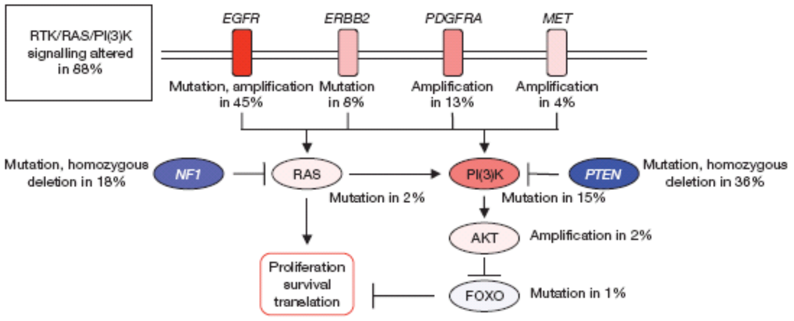
Naïve: Test groups of genes
Too many hypotheses
Network reduced hypotheses. Other information?

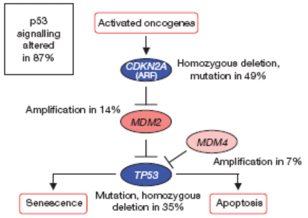
Pathways and Mutational Signatures

Driver mutations are rare.

→Cancer pathway has ***exactly one*** driver mutation (gene) per patient [REFs]

[**Exclusivity**]


RTK/RAS/PI(3)K signalling altered in 88%
EGFR
ERBB2
PDGFRA
MET
Mutation, amplification in 45%
Mutation in 8%
Amplification in 13%
Amplification in 4%
Mutation, homozygous deletion in 18%
NF1
RAS
Mutation in 2%
PI(3)K
Mutation in 15%
PTEN
Mutation, homozygous deletion in 36%
AKT
Amplification in 2%
Proliferation survival translation
FOXO
Mutation in 1%


p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Senescence
Mutation, homozygous deletion in 35%
Apoptosis

Most patients have mutation in pathway [**Coverage**]

Properties of *driver* mutations

- ***M*** = pathway (set of genes)
- ***n*** = number of tested genes

- From current understanding of mutational process of cancer:

  - **Coverage**: Most samples have at least one mutation in ***M***

  - **Exclusivity:** Most samples have no more than one mutation in ***M***

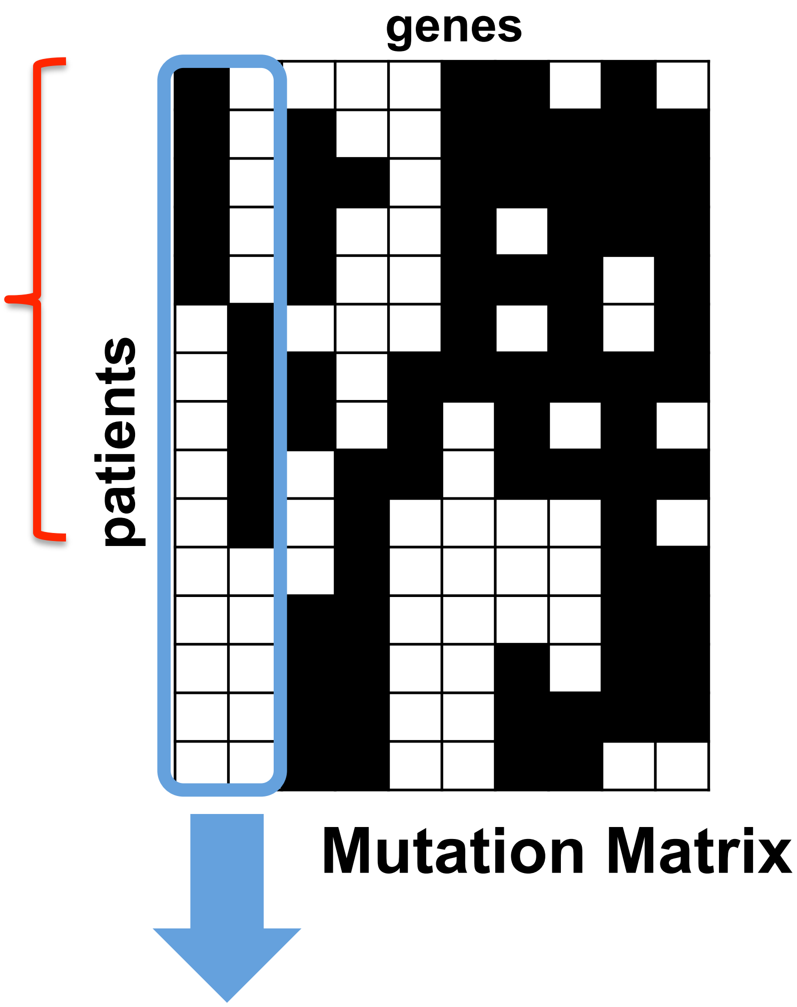
Mutation Exclusivity and Coverage

Coverage:

$\boldsymbol{\Gamma(g)}$ = {patients in which gene $\boldsymbol{g}$ mutated}

$\boldsymbol{\Gamma(M) = \cup_i \Gamma(g_i)}$ =

{patients in which ≥ 1 of $\boldsymbol{\{g_1, g_2, \ldots, g_k\}}$ is mutated}



Exclusive **(Column) Submatrix**

Mutual exclusivity and coverage

Coverage:

$\Gamma(g)$ = {patients in which gene ***g*** mutated}

$\Gamma(M) = U_i\, \Gamma(g_i)$ = {patients in which ≥ 1 of $\{g_1,g_2,\ldots,g_k\}$ is mutated}

Maximum Coverage Exclusive Submatrix Problem*:* Given ***k>0***, find the exclusive set ***M*** of ***k*** genes that maximizes $|\Gamma(M)|$

Theorem **Maximum Covering Exclusive Submatrix Problem** is NP-Hard.

Relaxing Constraints

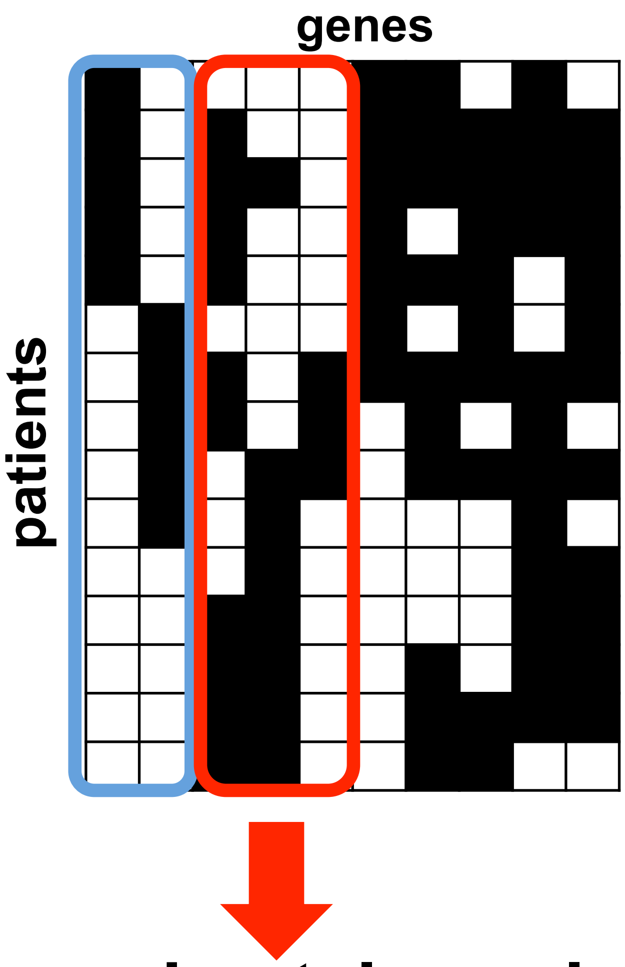
For set $\boldsymbol{M}$ of genes:

Coverage overlap:

$\gamma(\boldsymbol{M}) = \Sigma_i|\Gamma(g_i)| - |\Gamma(\boldsymbol{M})|$

$\gamma(\boldsymbol{M}) = 0$ if and only if $\boldsymbol{M}$ is exclusive.

Goal: $|\Gamma(\boldsymbol{M})|$ large and $\gamma(\boldsymbol{M})$ small.



"Approximately exclusive",
high coverage submatrix

Approximate Exclusivity

Goal: $\Gamma(M)$ large and $\gamma(M)$ small.

Weight: $W(M) = |\Gamma(M)| - \gamma(M) = 2\,|\Gamma(M)| - \Sigma_i|\Gamma(g_i)|$

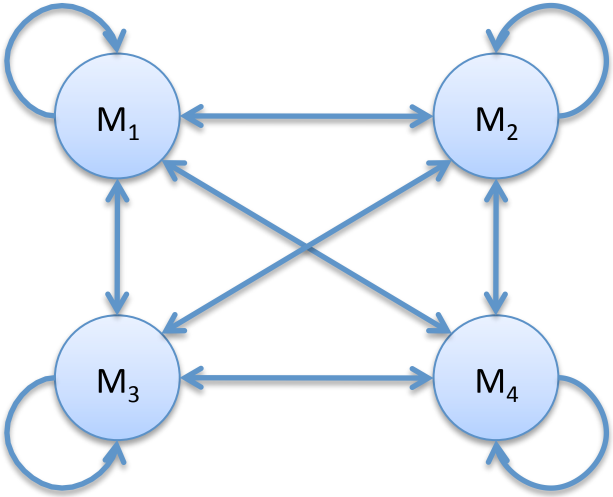
Maximum Weight Submatrix Problem*:* Given $k>0$, find the set M of k genes that maximizes $W(M)$

Thm. **Maximum Weight Submatrix Problem** is NP-Hard.

Markov Chain Monte Carlo

Sample gene sets $|\boldsymbol{M}| = \boldsymbol{k}$ according to $\boldsymbol{W(M)}$

Markov chain:
States = sets
M



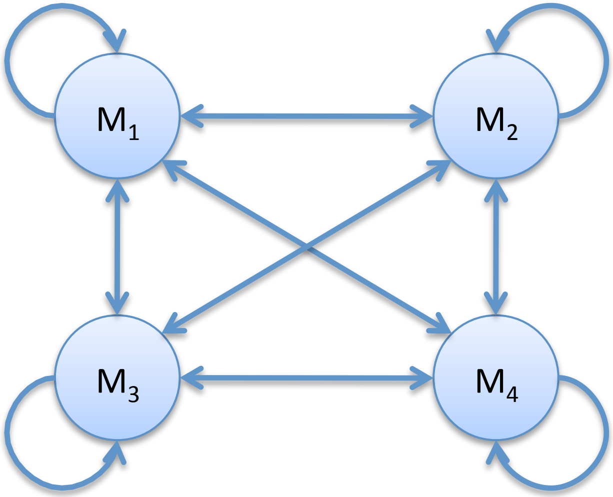
Generate sequence of states: $M^{(1)}, M^{(2)}, M^{(3)}, \ldots$

Markov Chain Convergence Thm: $M^{(i)} \rightarrow \pi$

Metropolis-Hastings

Define transition probabilities of Markov Chain so
π = desired distribution.

Markov chain:
States = sets
M



Distribution on gene sets: Pr[***M***] ~ $\boldsymbol{e}^{c\,W(M)}$

In general: no guarantees on rate of convergence

MCMC approach

Thm. Markov Chain is rapidly mixing.

Returns a distribution on sets, not just optimal [max W(**M**)] set

No assumptions on distribution of mutations

- i.e. independence not necessary
- can handle various mutation types

*Vandin, Upfal, Raphael. (submitted)

Experimental Results

- Simulated data
- Cancer data

1. **Brain cancer (GBM)** [TCGA, *Nature* (2008)]

   601 sequenced genes in 84 samples

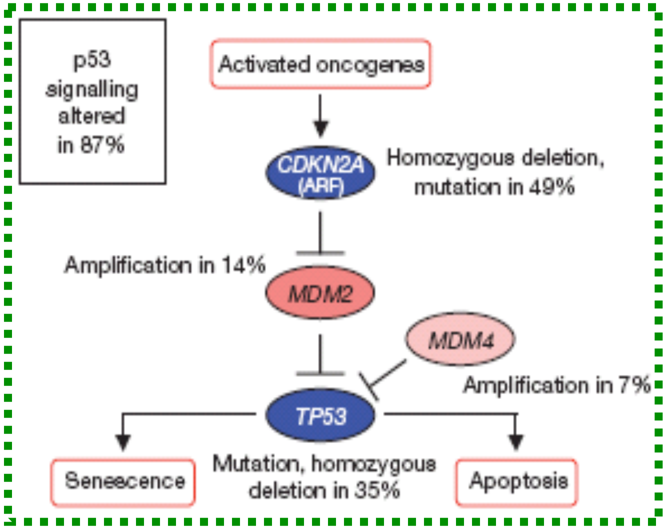
   Array copy number data on *all* genes

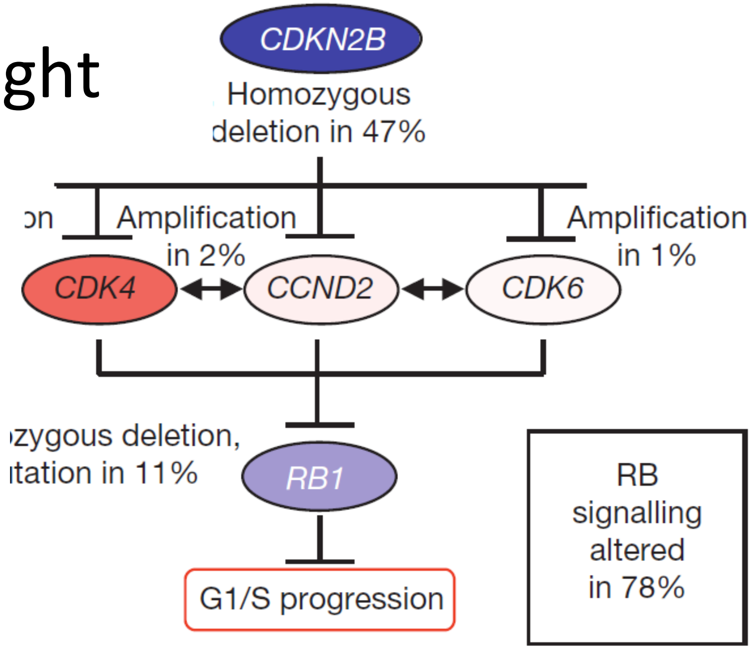
2. **Lung Adenocarcinoma** [Ding et al., *Nature* (2008)]

   623 sequenced genes in 188 samples

Brain Cancer (GBM)

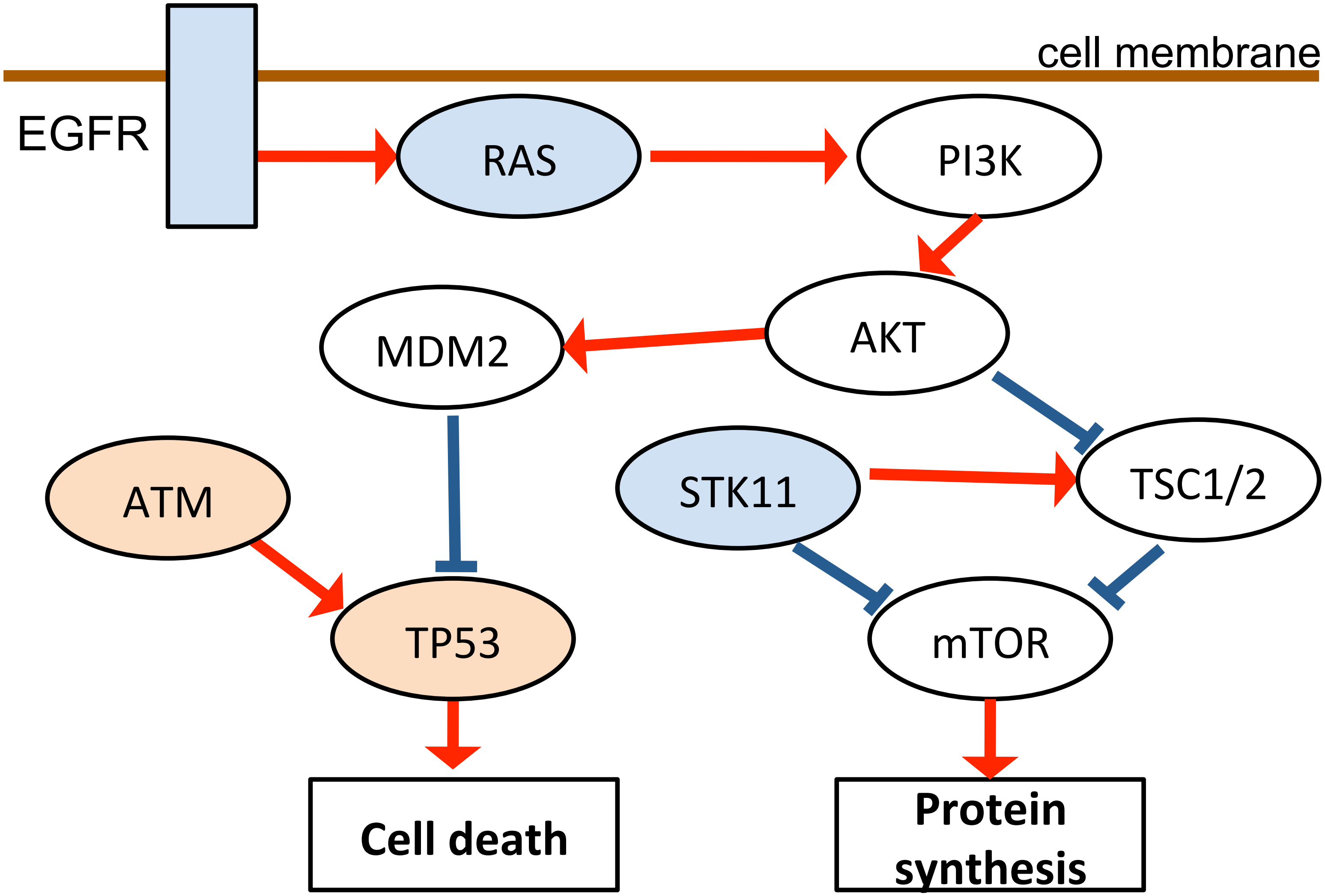
- **M** = {CDKN2B, RB1, CDK4}
  - not the set with highest weight
- **M** = {TP53, CDKN2A}
  - p53 signaling pathway


p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Senescence
Mutation, homozygous deletion in 35%
Apoptosis



From [TCGA, *Nature*, 2008]

Lung Adenocarcinoma


cell membrane
EGFR
RAS
PI3K
AKT
MDM2
ATM
STK11
TSC1/2
TP53
mTOR
Cell death
Protein synthesis