

Algorithms for Automated Discovery of Mutated Pathways in Cancer

Ben Raphael



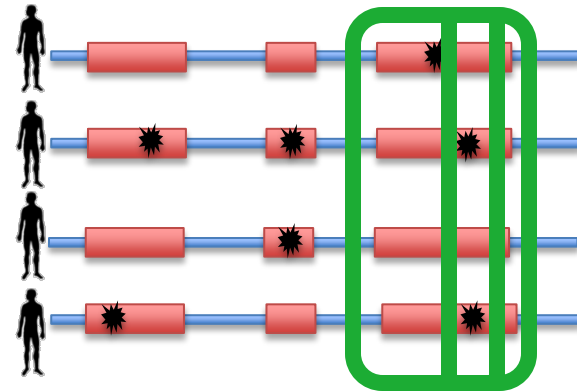
BROWN

Department of Computer Science
Center for Computational Molecular Biology

Driver Mutations



Next-generation
DNA sequencing

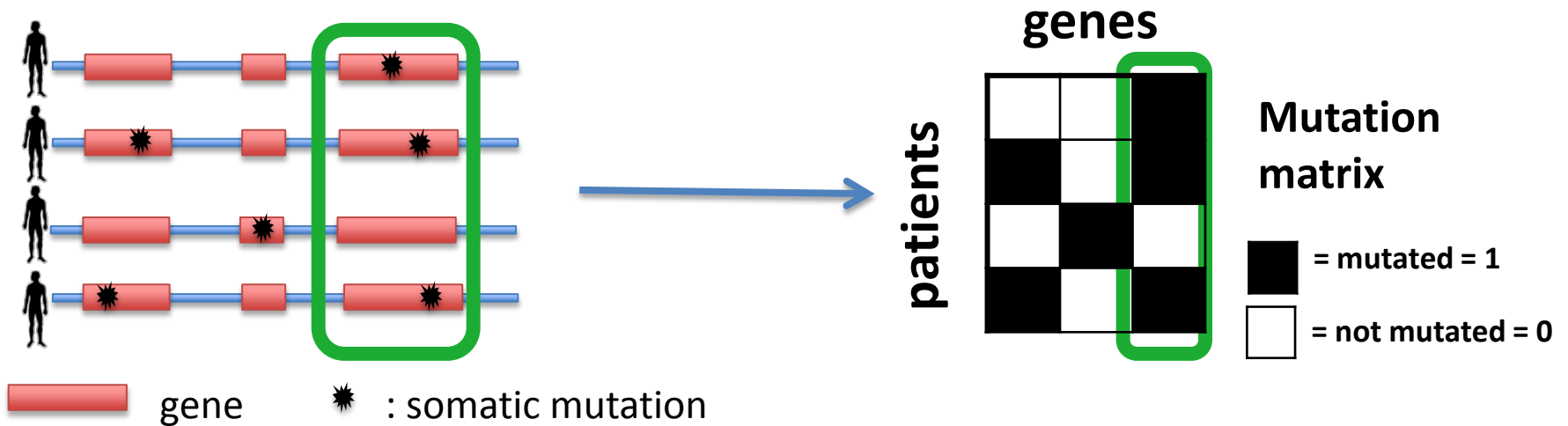


gene * : somatic mutation

Distinguish functional (*driver*) mutations from background (*passenger*) mutations.

Recurrent mutations/mutated genes → *driver* mutations

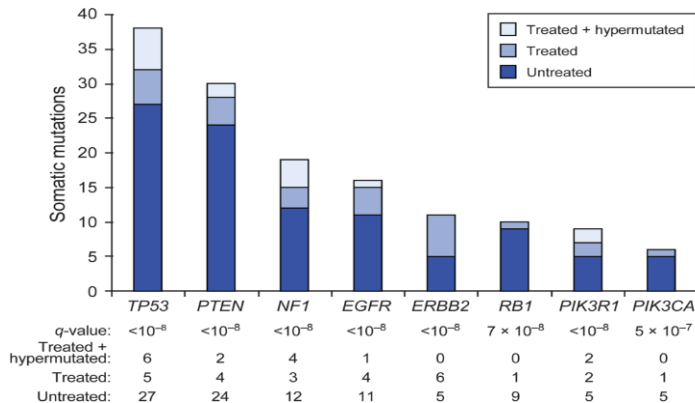
Recurrent Mutations



Mutated more than expected by chance?

Single-gene test → Multiple hypotheses correction

91 GBM samples. TCGA, *Nature* (2008)



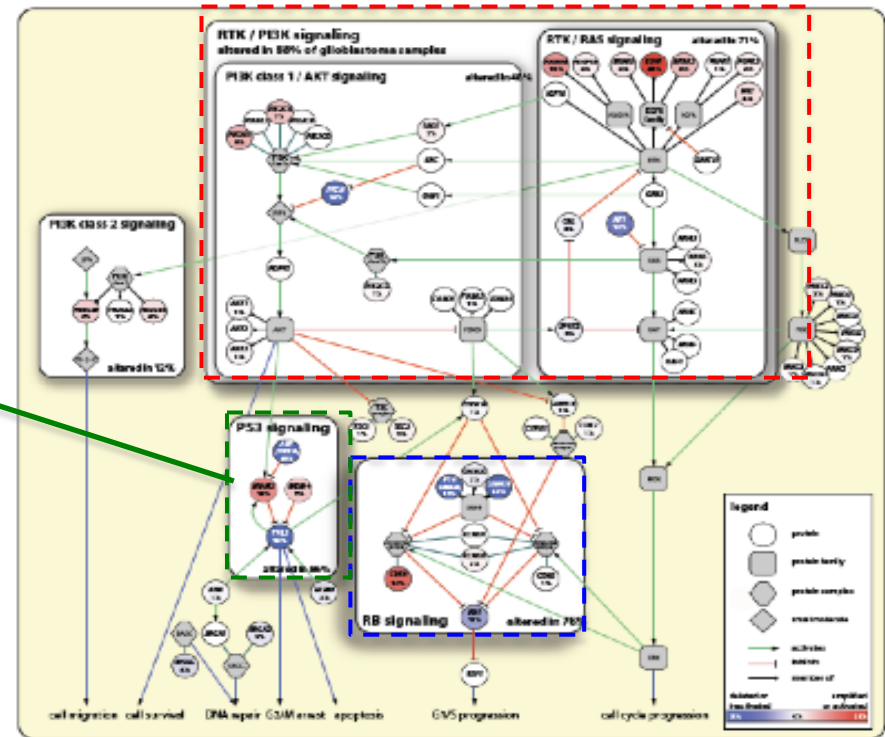
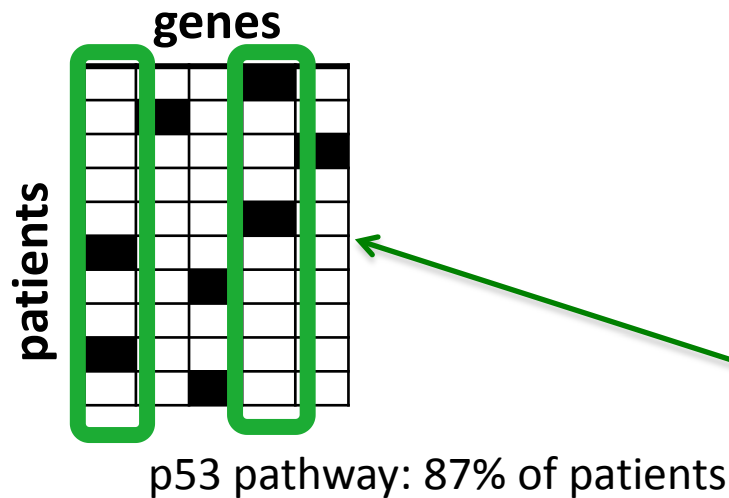
316 OV samples. TCGA, *Nature* (2011)

Table 2 | Significantly mutated genes in HGS-OvCa

Gene	No. of mutations	No. validated	No. unvalidated
TP53	302	294	8
BRCA1	11	10	1
CSMD3	19	19	0
NF1	13	13	0
CDK12	9	9	0
FAT3	19	18	1
GABRA6	6	6	0
BRCA2	10	10	0
RB1	6	6	0

Mutated Pathways

Standard approach: enrichment of mutations on *known pathways*.



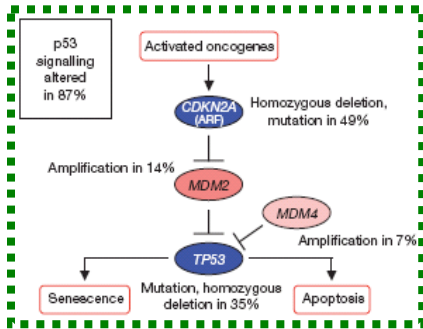
Limitations

- Only existing pathways are tested.
- Topology of pathways ignored.
- Pathways are interconnected (*crosstalk*).

TCGA GBM. *Nature*, 2008.

Advantage of Large Datasets?

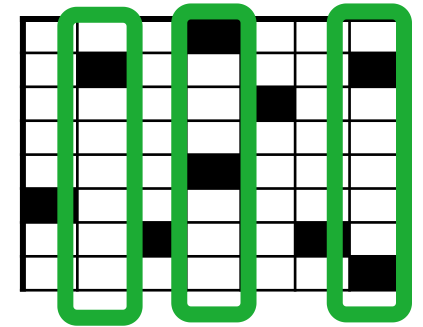
Prior knowledge of groups of genes



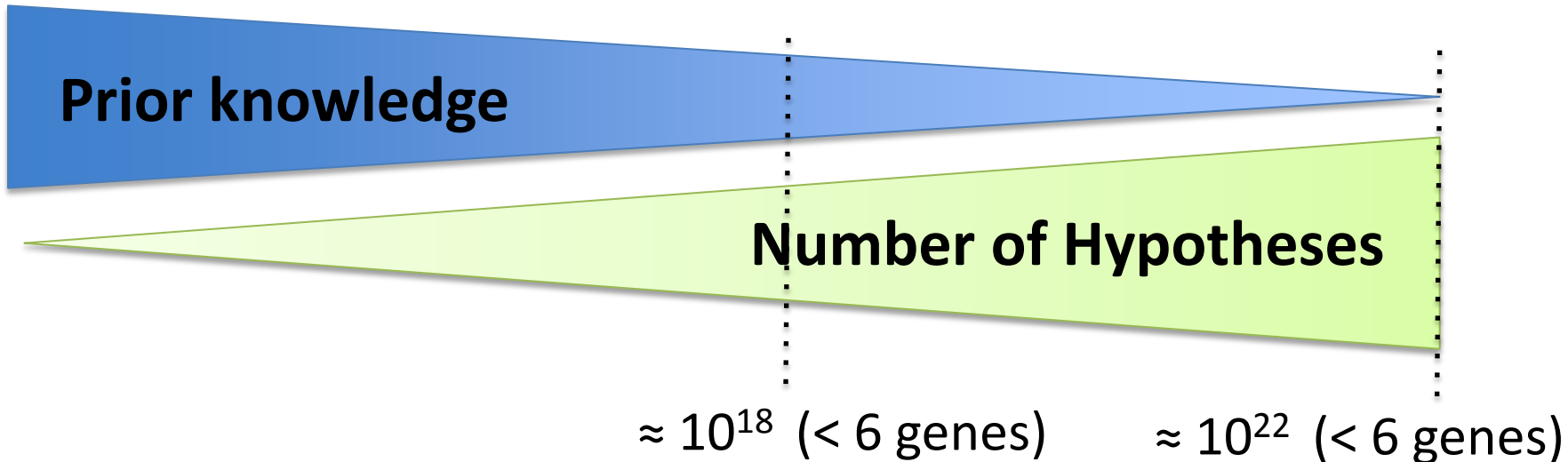
Known pathways



Interaction Network

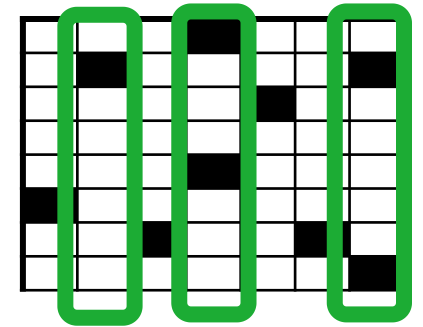
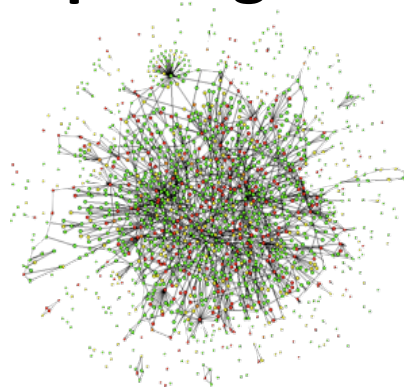
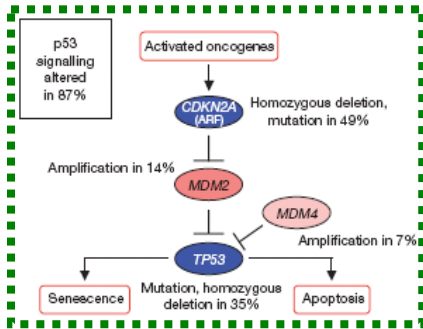


None



Two Algorithms

Prior knowledge of groups of genes



Known pathways

Interaction Network

None

Prior knowledge

Number of Hypotheses

HotNet

Dendrix

subnetworks of
interaction network

Exclusive gene sets

HotNet: Problem Definition

Given:

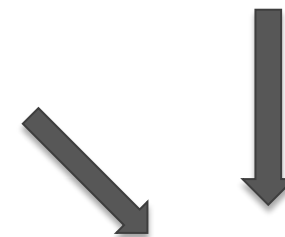
1. Network $G = (V, E)$

V = genes. E = interactions b/w genes

2. Binary mutation matrix

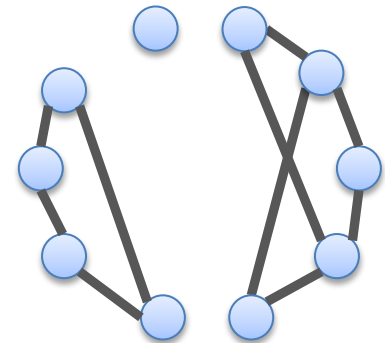
■ = mutated
□ = not mutated

	Genes							
Patients				■				
		■						■
						■		
					■			
	■							
			■				■	
	■							■
				■				

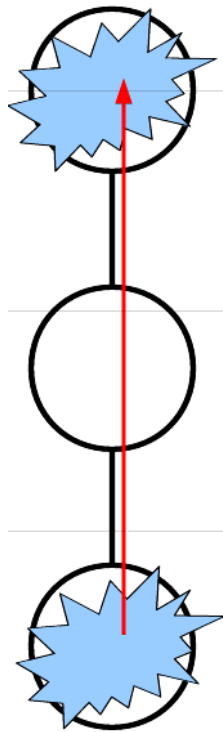


Find: *Connected subnetworks* mutated in a significant number of patients

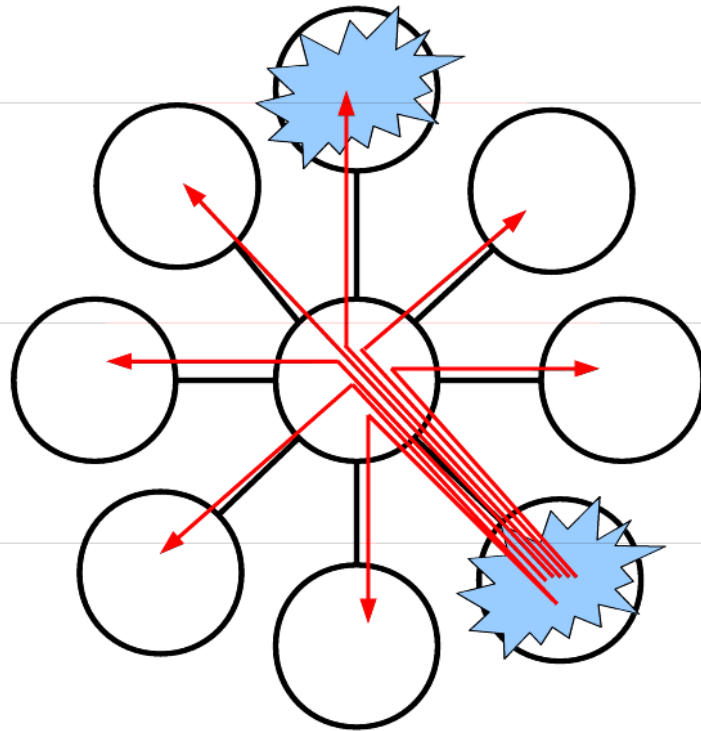
— *mutated* in patient if ≥ 1 gene mutated in patient



(Local) Topology Matters



Single path
between mutated
genes



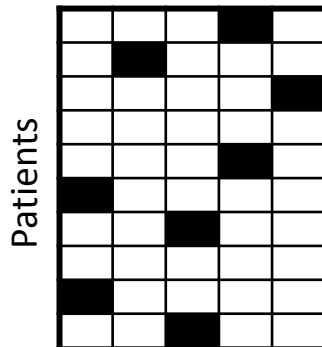
Path between mutated
genes is one of many
through node.

Example: TP53 has 238 neighbors in HPRD network

Mutated subnetworks: *HotNet**

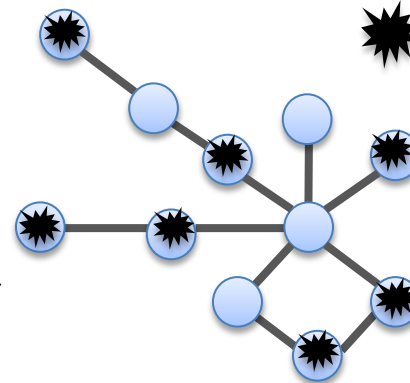
Mutation Matrix

Genes



Human Interaction Network

* = mutated genes



(1)

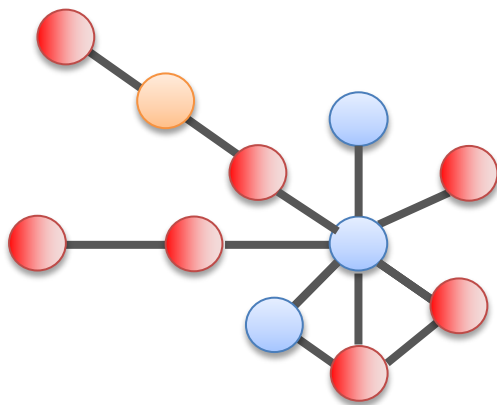
Mutation \rightarrow *heat diffusion*

Extract "significantly hot" subnetworks

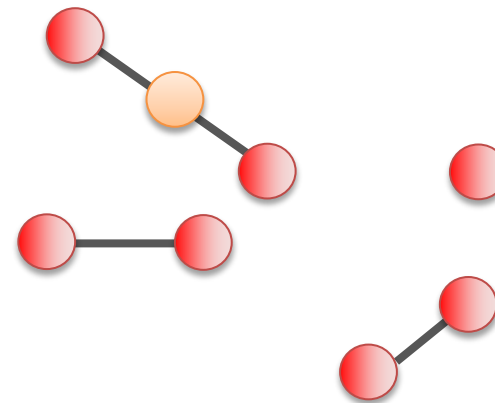
Hot



Cold

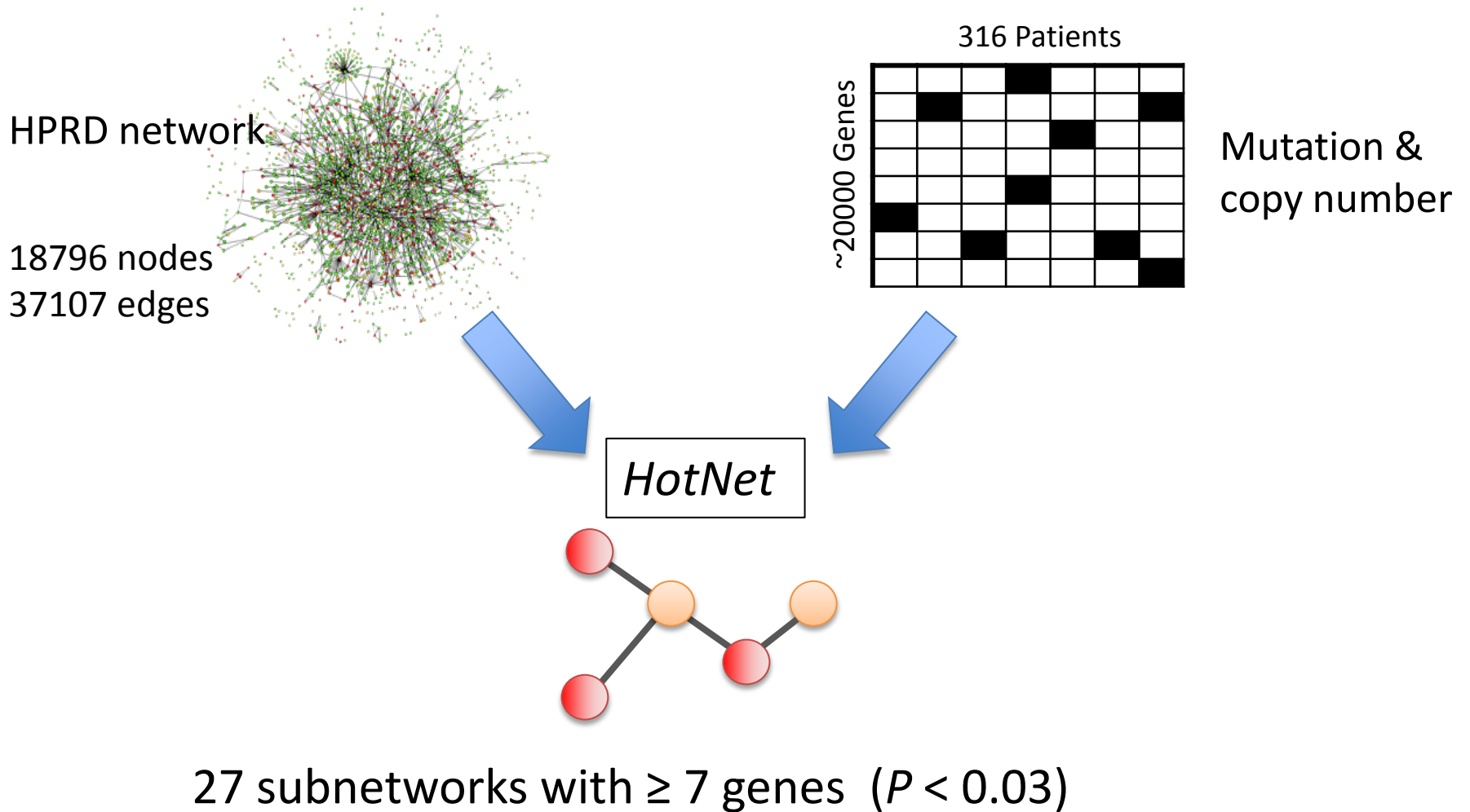


(2)

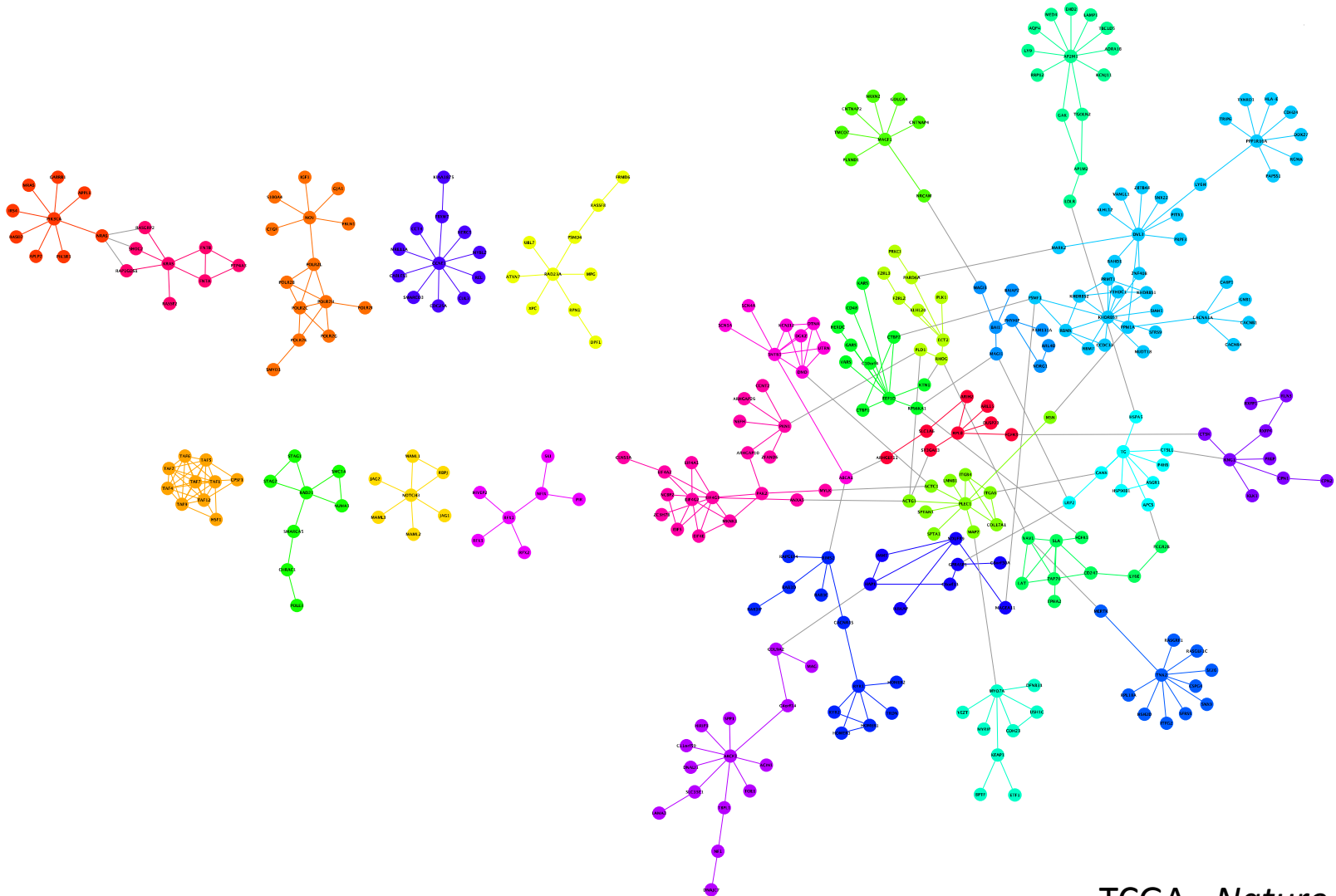


*F. Vandin, E. Upfal, and B. J. Raphael. J. Comp.Biol. (2011). Also *RECOMB* (2010).

Results: TCGA Ovarian



Ovarian Subnetworks



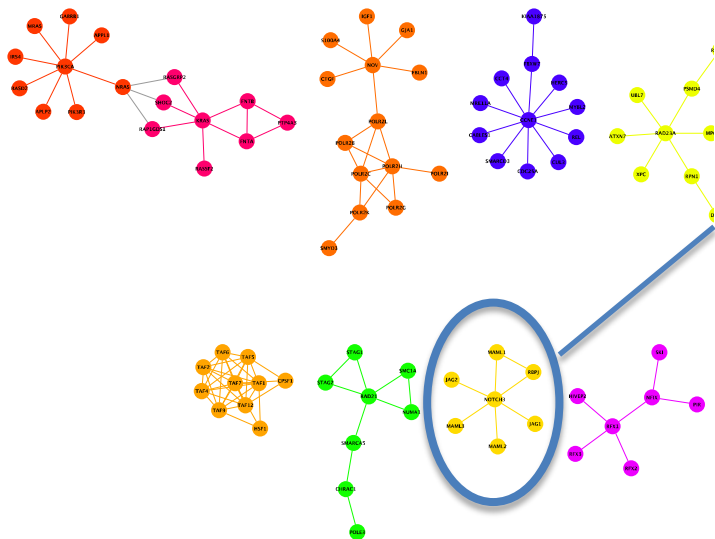
Ovarian Subnetworks

ARTICLE

doi:10.1038/nature10166

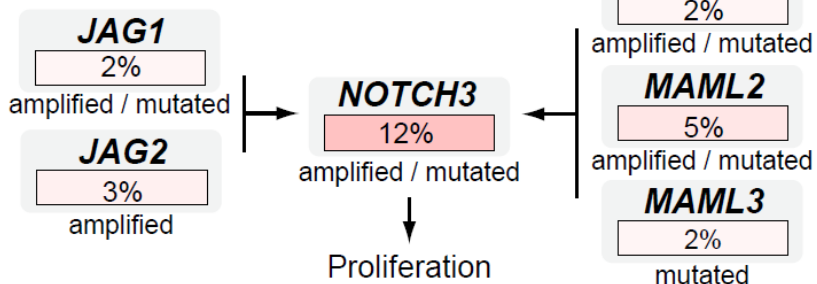
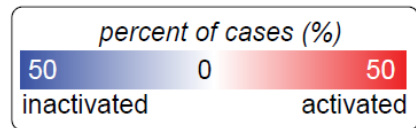
Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*



B. NOTCH Signaling

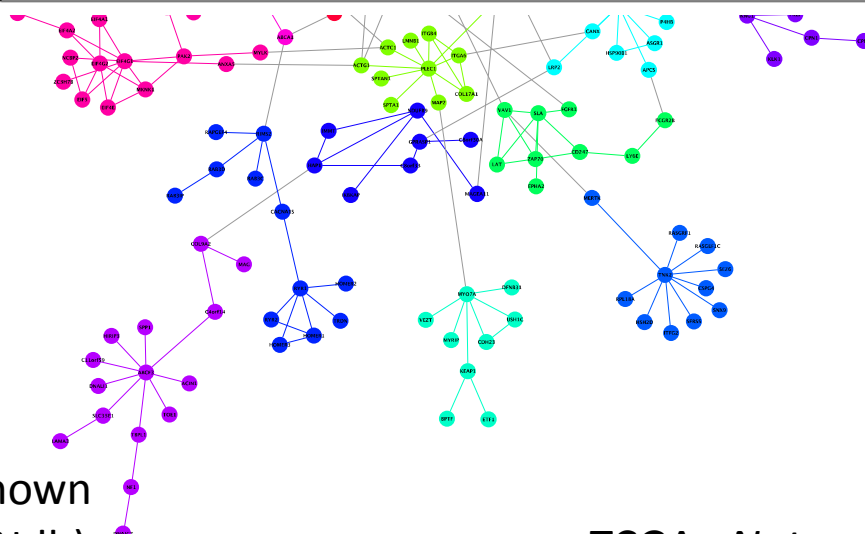
23% of cases altered



Kegg Pathway

Notch signaling ($p < 6 \times 10^{-7}$)

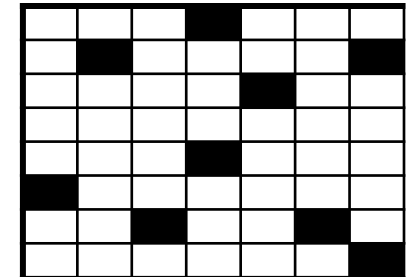
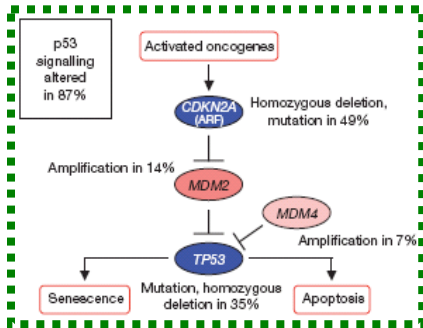
12/27 subnetworks significantly overlap known pathways (KEGG) or protein complexes (PINdb)



TCGA. *Nature* (2011)

Two Algorithms

Prior knowledge of groups of genes



Known pathways

Interaction Network

None

Prior knowledge

Number of Hypotheses

HotNet

Dendrix

subnetworks of
interaction network

Exclusive gene sets

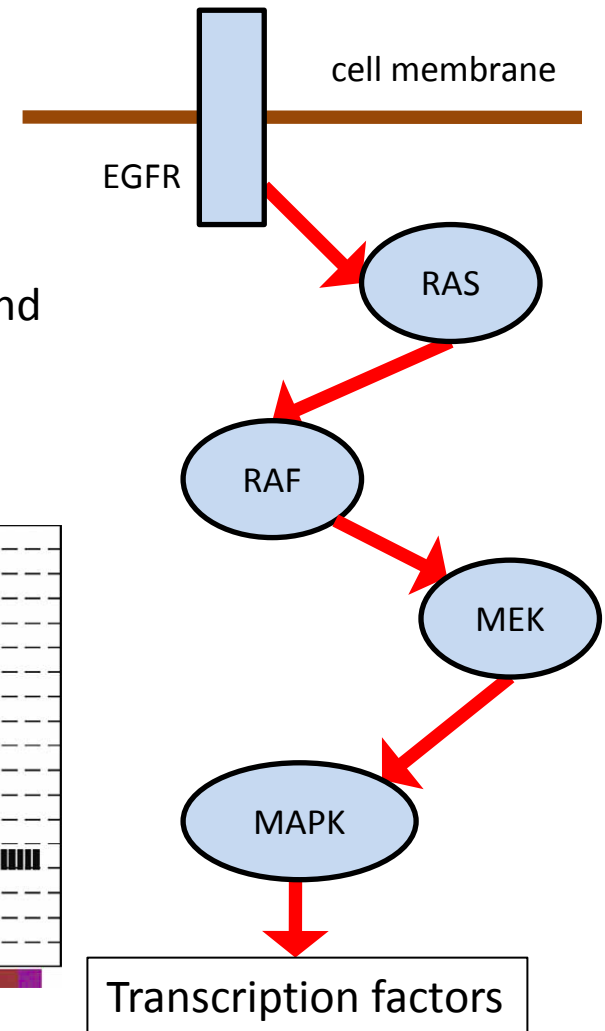
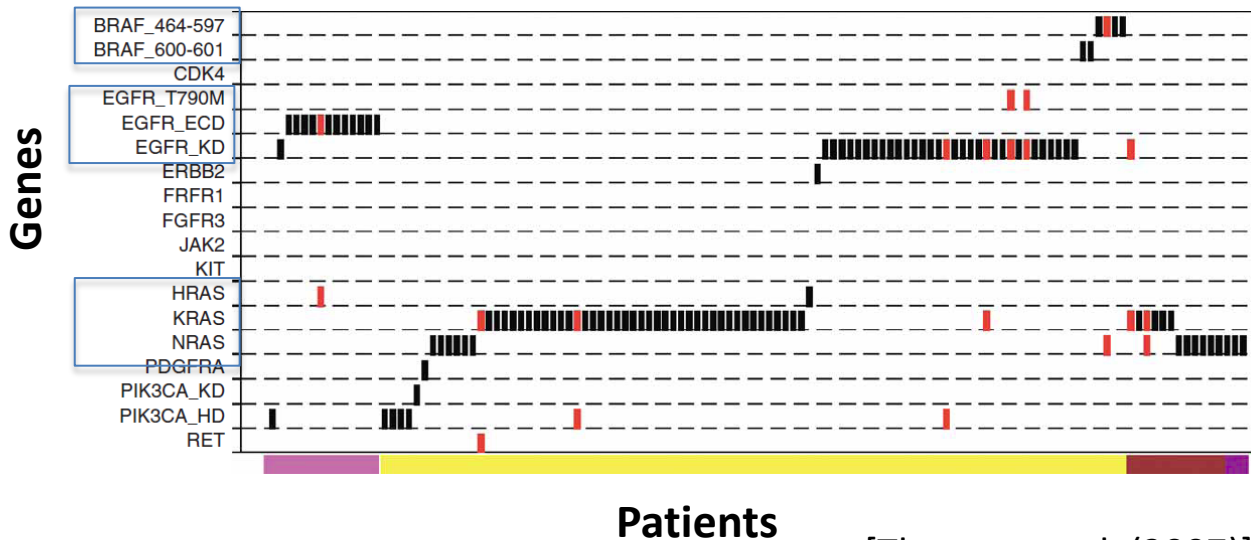
Pathways and Mutations

Driver mutations are rare.

➔ Cancer pathway has **one** driver mutation (gene) per patient

[Vogelstein and K. W. Kinzler (2004), Yeang, McCormick, and Levine (2008)]

1. Exclusivity



Pathways and Mutations

Driver mutations are rare.

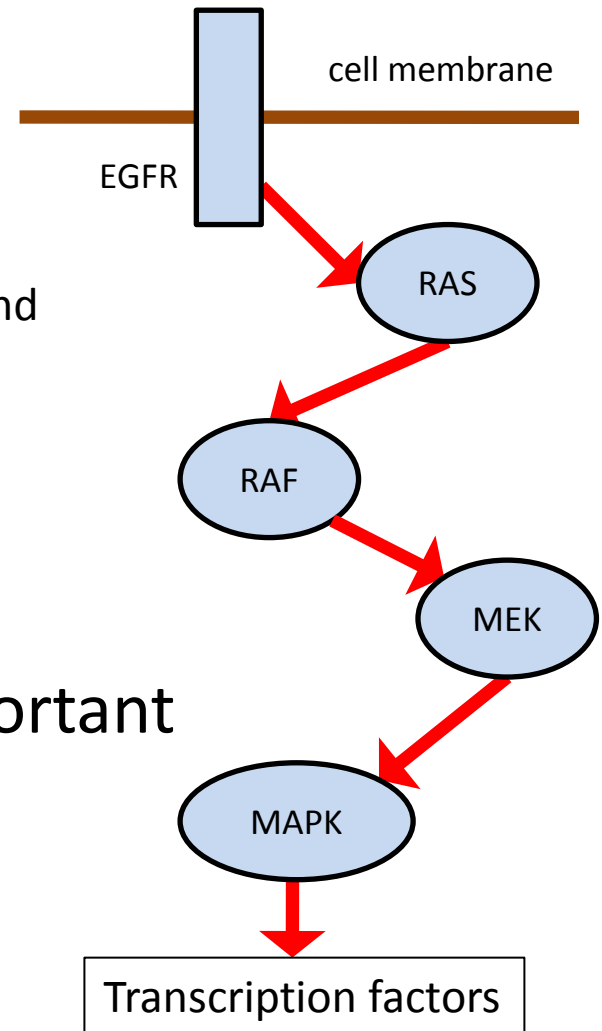
➔ Cancer pathway has ***one*** driver mutation (gene) per patient

[Vogelstein and K. W. Kinzler (2004), Yeang, McCormick, and Levine (2008)]

1. Exclusivity

Many patients have mutation in important cancer pathway.

2. Coverage



De novo driver exclusivity (Dendrix*)

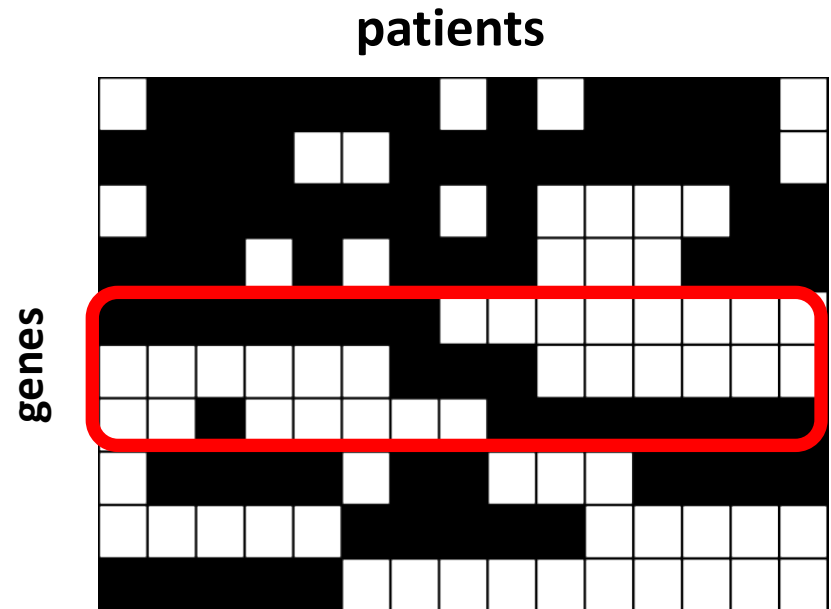
Given:

Binary mutation matrix A

Find:

Set M of genes with:

- **High Coverage:** many patients have a least one mutation in M
- **Approximate Exclusivity:** most patients have no more than one mutation in M



Finding largest M is difficult!

Greedy algorithm and *MCMC* algorithm

- Theoretical results on convergence and optimality.

Dendrix++: extension with scoring based on probabilistic model

*Vandin, Upfal, & Raphael. *Genome Res. (Advance online)* Also *RECOMB 2011*.

Acute Myeloid Leukemia (AML)

Data: Somatic mutations (from whole-exome sequencing) and fusion genes (RNA-seq) of 199 patients.

Dendrix++: Several “approximately exclusive” sets (each $p < 0.001$).

HotNet: 5 subnetworks containing ≥ 5 genes ($p < 0.01$).

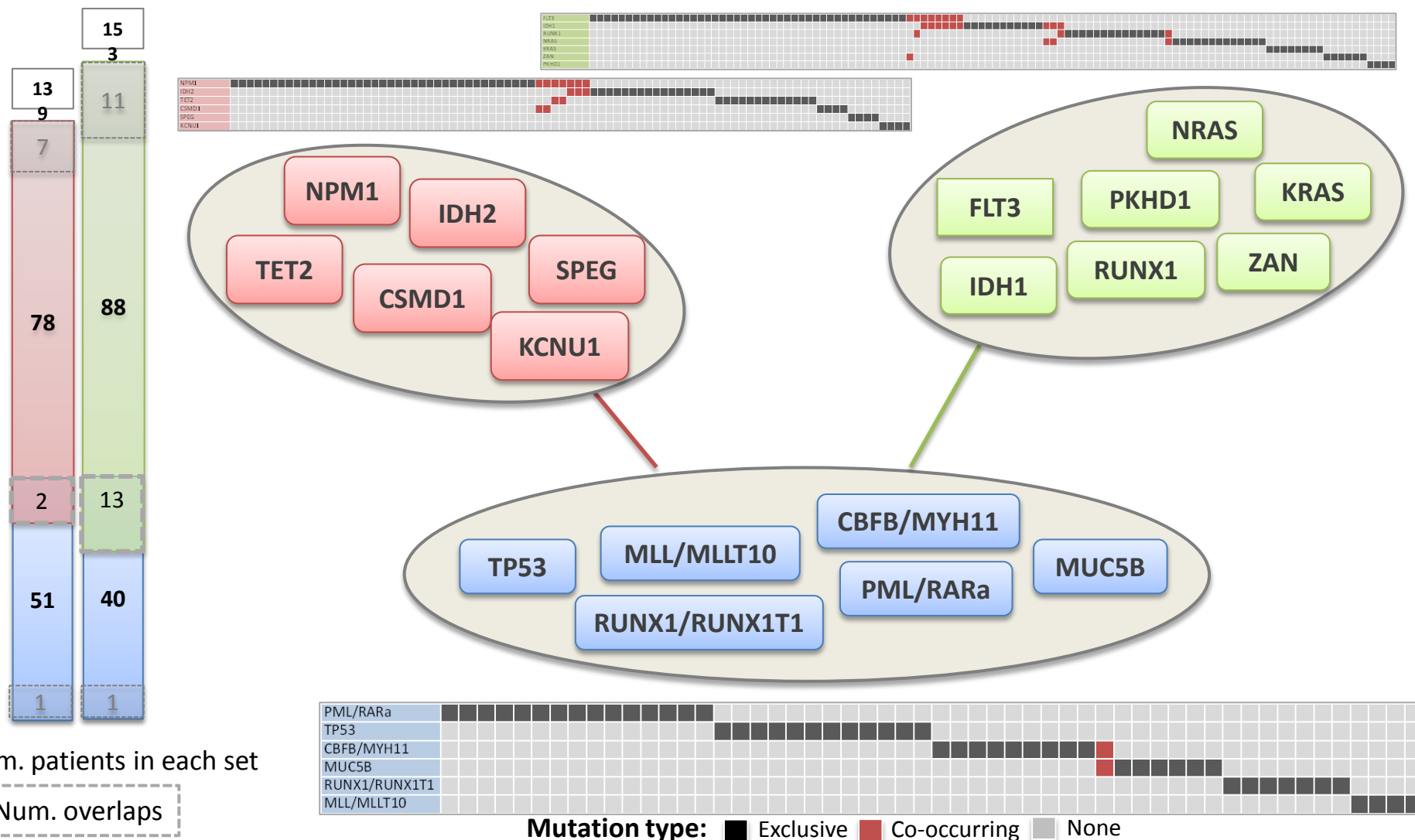


CAUTION

Mutations are not (yet) validated.
Any conclusions drawn from this data are preliminary and subject to change.

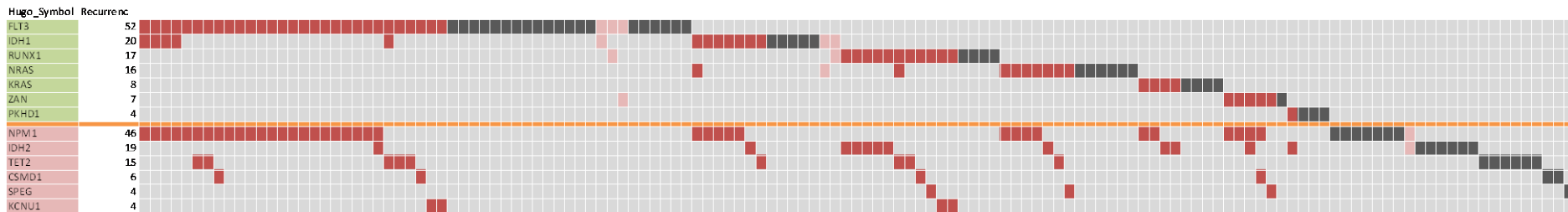
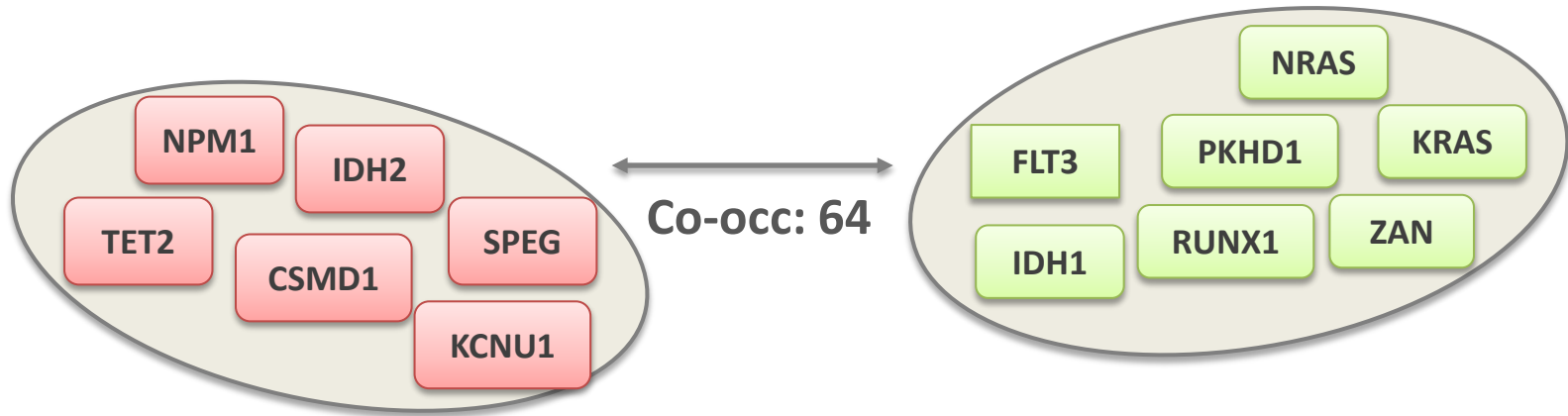
Dendrix++: AML

Top 2 scoring sets ($p < 10^{-3}$ for each)



Dendrix++: AML

Top 2 scoring sets ($p < 10^{-3}$ for each)

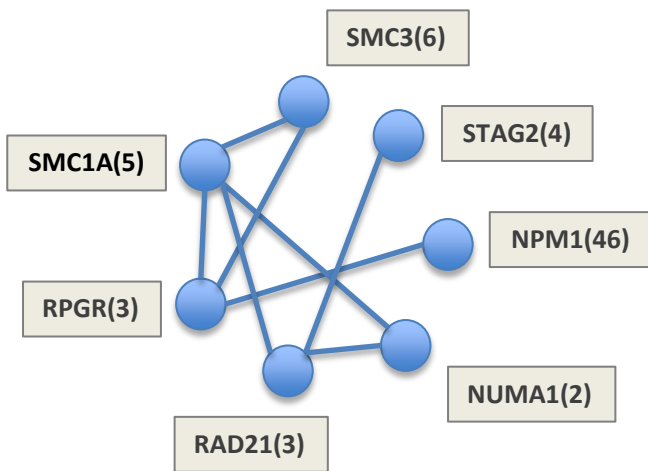


■ Co-occurrence between two groups
■ Co-occurrence inside the group

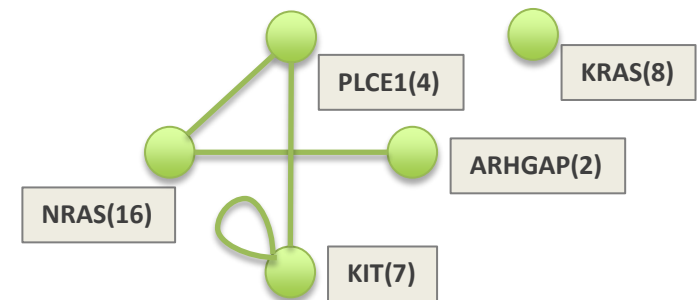
HotNet: AML

Data: Somatic mutations (from whole-exome sequencing) and fusion genes (RNA-seq) of 199 samples.

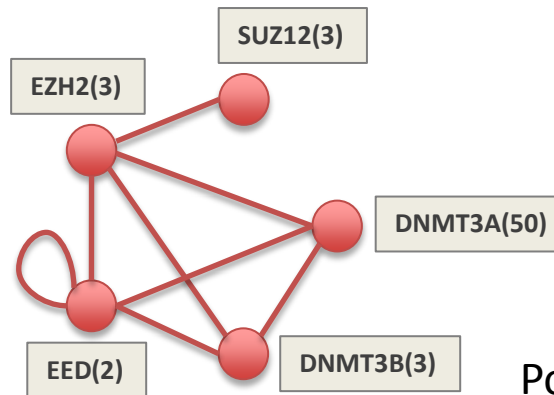
HotNet: 5 subnetworks containing ≥ 5 genes ($p < 0.01$).



Cohesin complex ($p = 10^{-5}$)



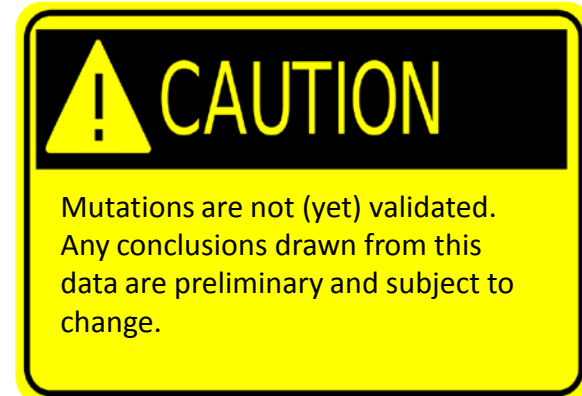
KEGG: AML ($p = 10^{-3}$)



Polycomb complex ($p = 10^{-3}$)

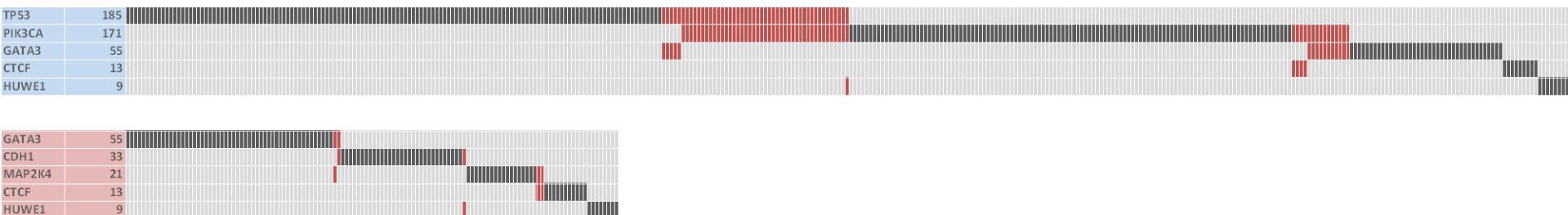
Results: TCGA Breast

Data: Somatic mutations (whole-exome sequencing) of 514 samples with copy number aberrations for 438 of these.



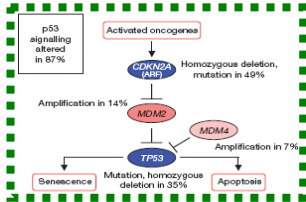
Dendrix++: Multiple approximately exclusive sets (each $p < 0.001$).

HotNet: 13 subnetworks containing ≥ 8 genes ($p < 0.01$). 4 enriched for known pathways (KEGG) and protein complexes (PINDb)

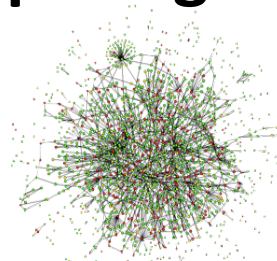


Summary

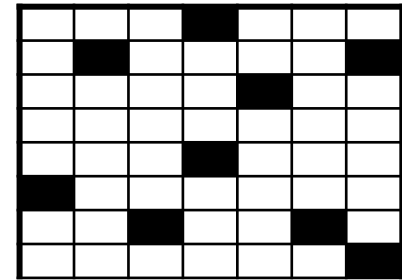
Prior knowledge of groups of genes



Known pathways



Interaction Network



None

Prior knowledge

Number of Hypotheses

HotNet

subnetworks of
interaction network

Dendrix

Exclusive gene sets

Future: Incorporate more data types (methylation, gene expression).
Perform pre/post filtering of predictions.

Acknowledgements

Fabio Vandin
Eli Upfal
Hsin-Ta Wu



Tim Ley
Li Ding
Elaine Mardis
Rick Wilson
and others...



Andy Mungall
and others..



Funding

