BROWN

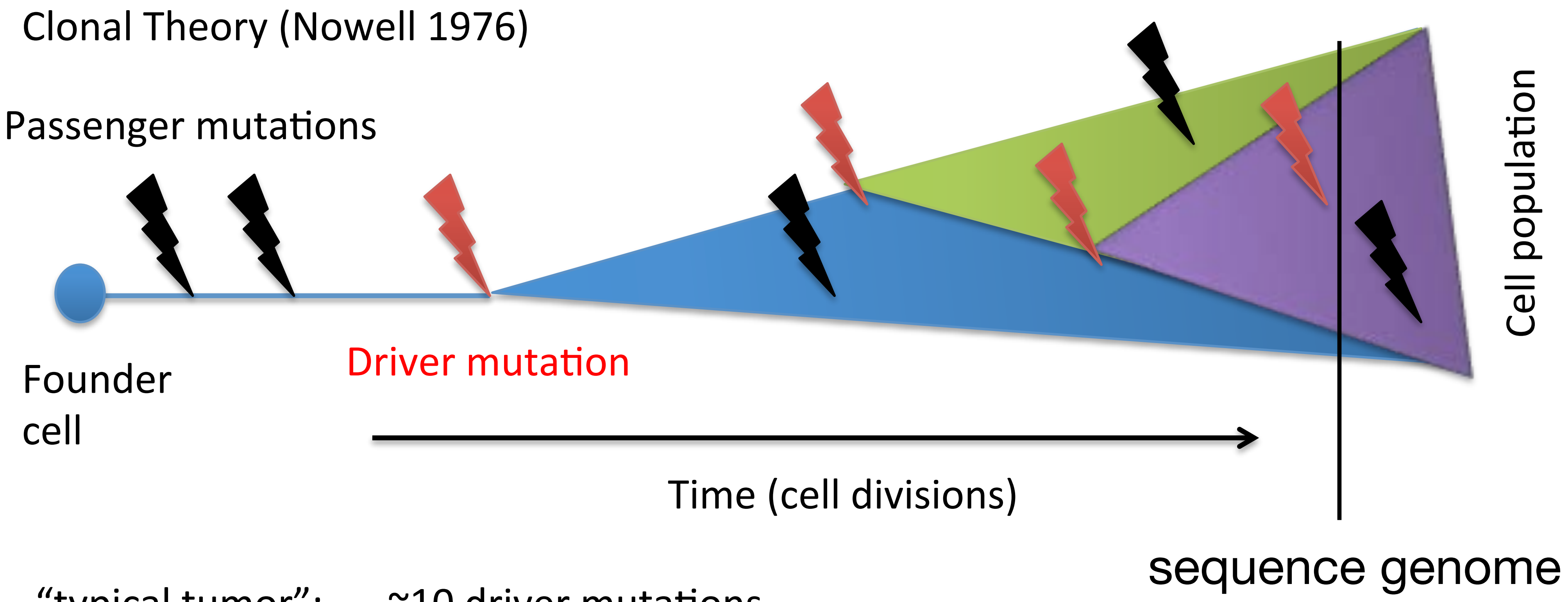
Simultaneous Identification of Multiple Driver Pathways in Cancer

Max Leiserson
ISMB 2014

Joint work with Benjamin Raphael, Roded Sharan and Dima Blokh.

July 13th, 2014

Clonal theory of cancer



Clonal Theory (Nowell 1976)

Passenger mutations

Founder cell

Driver mutation

Time (cell divisions)

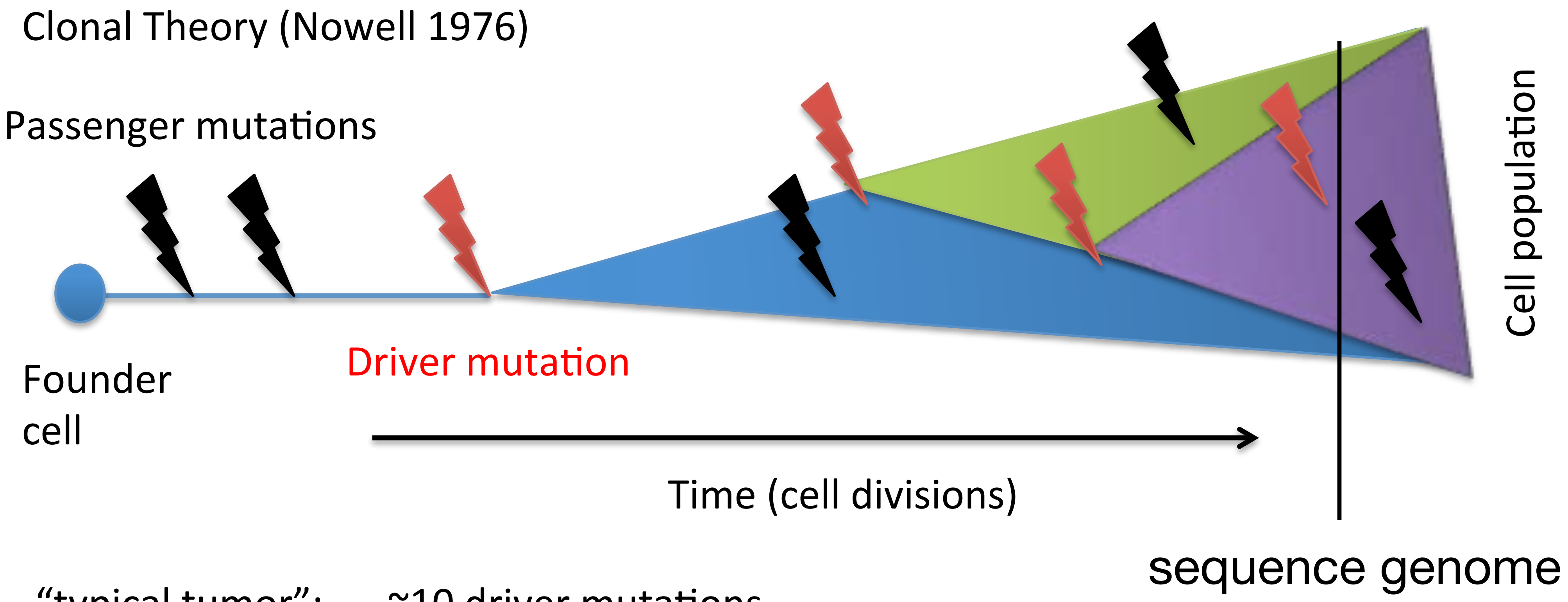
Cell population

sequence genome

“typical tumor”: ~10 driver mutations

100’s – 1000’s of passenger mutations

Clonal theory of cancer



“typical tumor”: ~10 driver mutations

100’s – 1000’s of passenger mutations

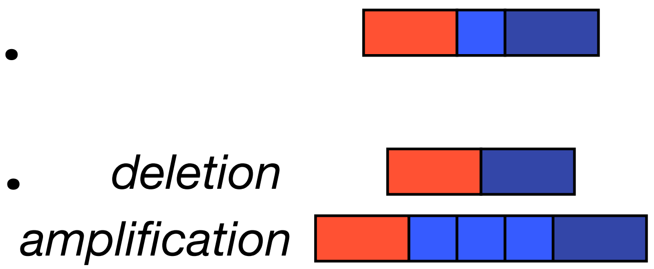
Types of Variation in Tumor Genomes

Single Nucleotide Variants

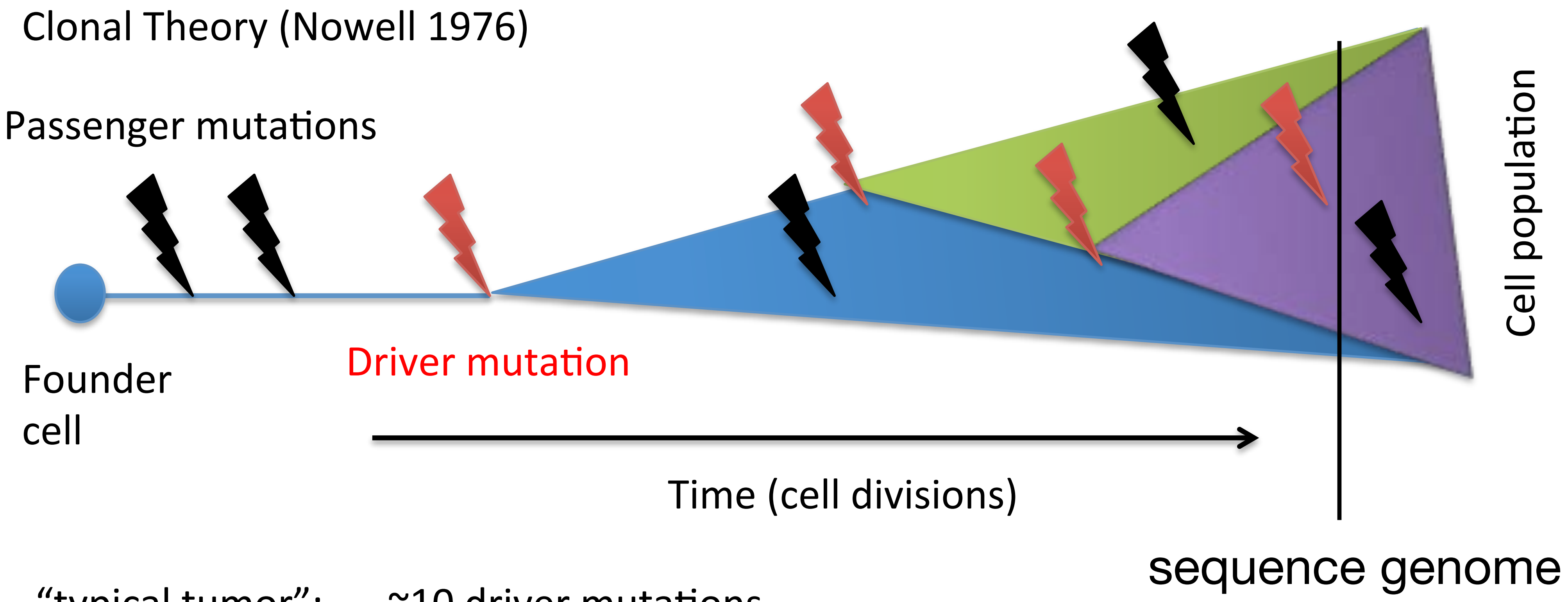
Copy Number Variants

Healthy ..ACGTCA**T**CATGA..

Tumor ..ACGTCA**G**CATGA..



Clonal theory of cancer



“typical tumor”: ~10 driver mutations

100’s – 1000’s of passenger mutations

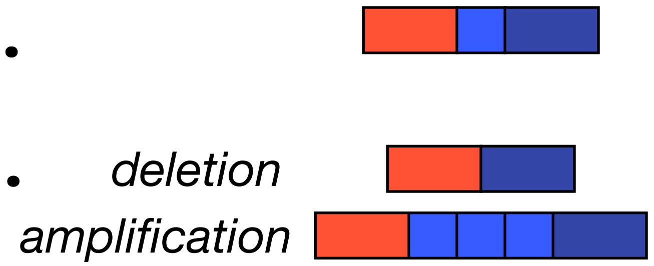
Types of Variation in Tumor Genomes

Single Nucleotide Variants

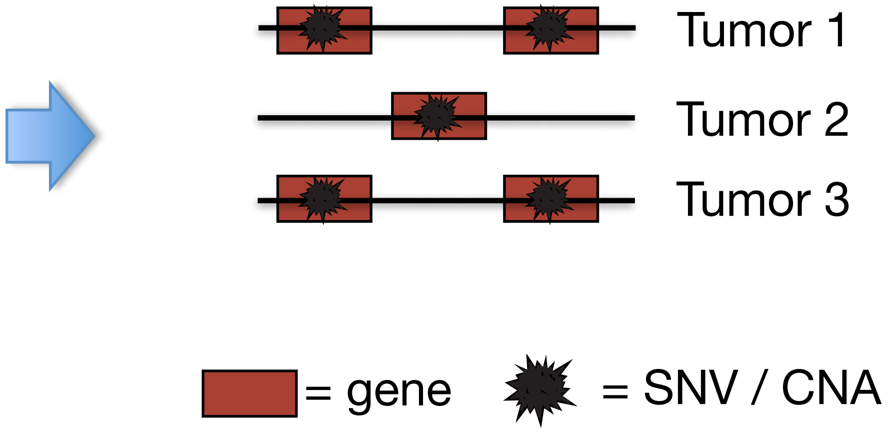
Healthy ..ACGTCA**T**CATGA..

Tumor ..ACGTCA**G**CATGA..

Copy Number Variants



Compare variation across tumors

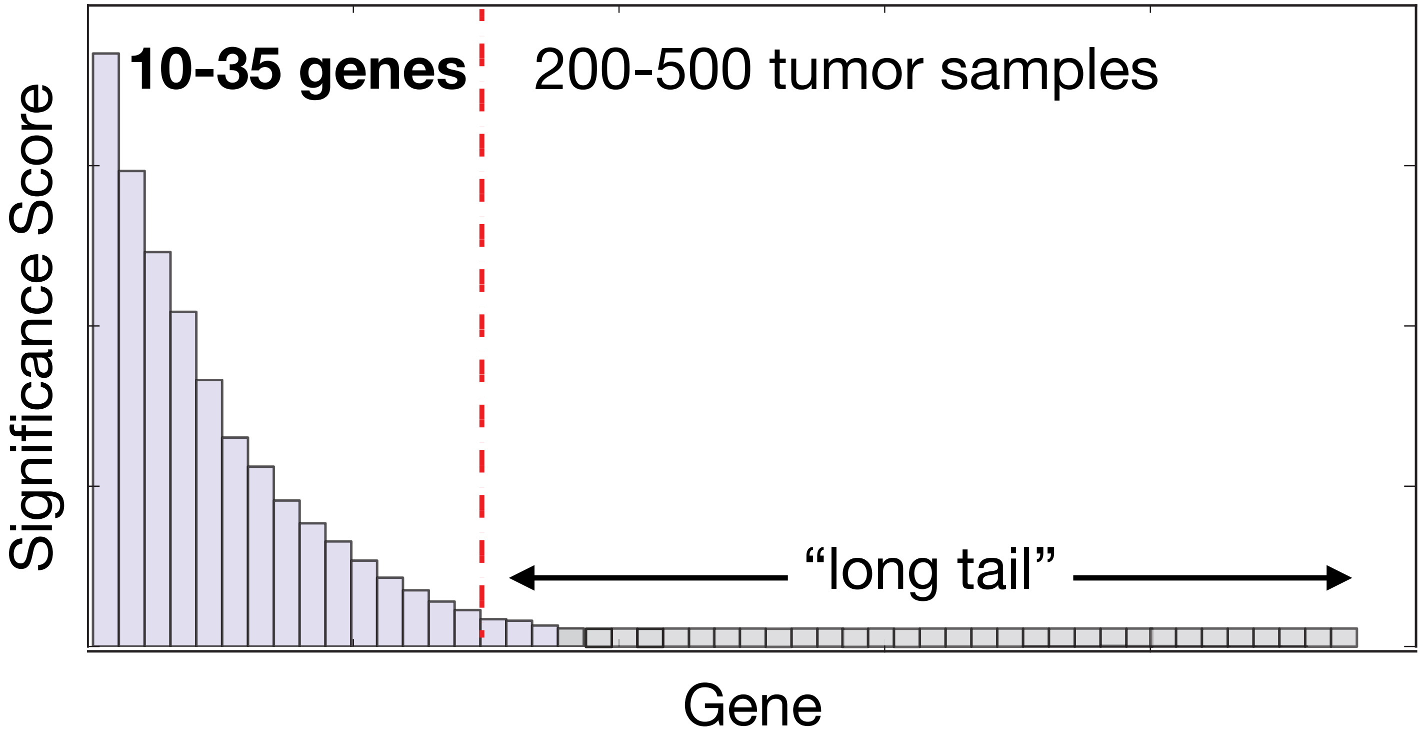


Significantly mutated genes in cancer

Significance Score

Mutations weighted by:

- Gene length
- Mutation context
- Expression level
- Replication timing
- …

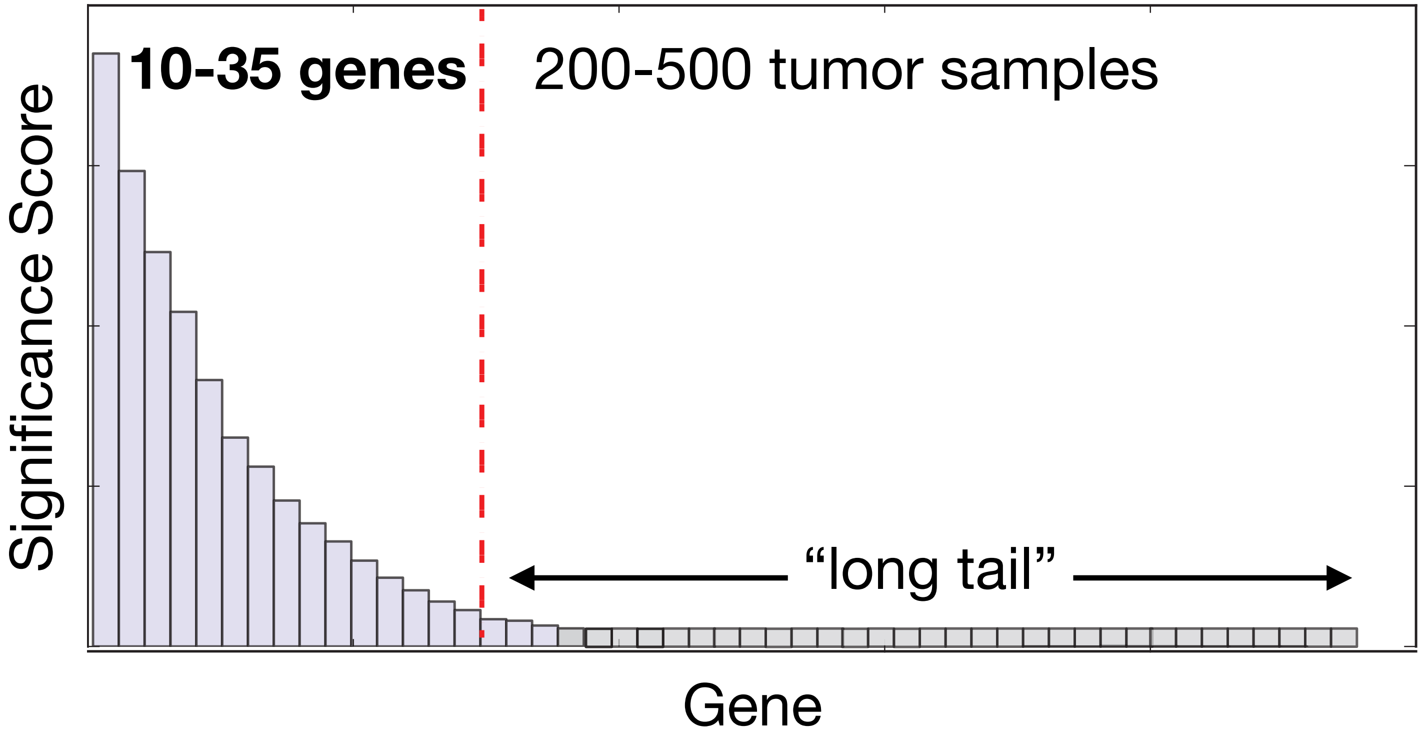


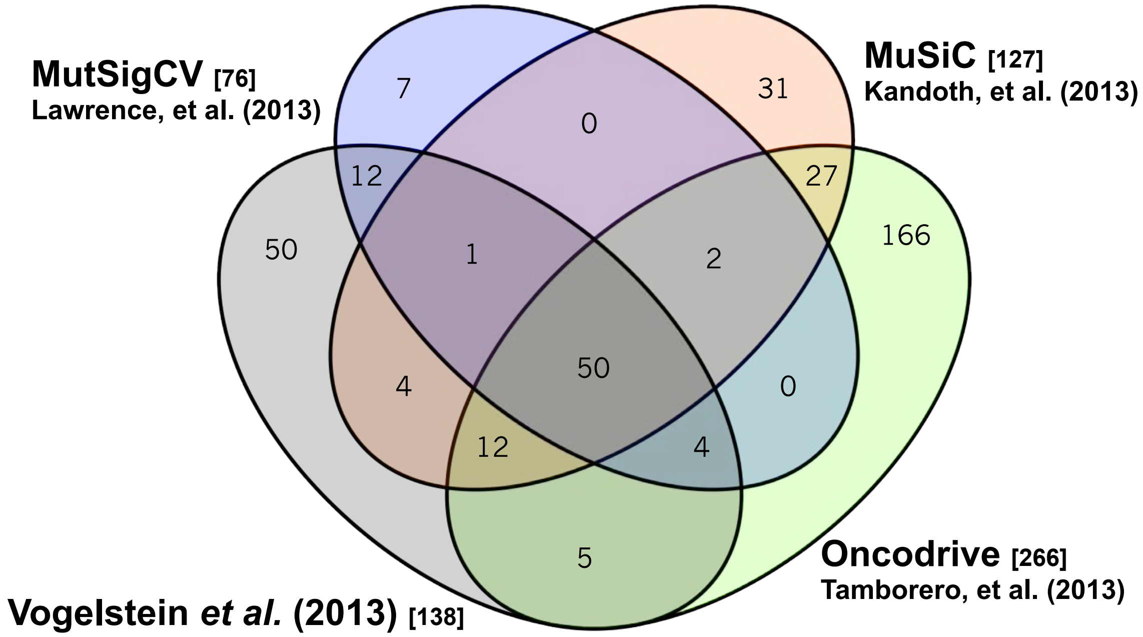
“Long tail” of mutated genes complicates finding driver mutations

Significance Score

Mutations weighted by:

- Gene length
- Mutation context
- Expression level
- Replication timing
- …





Comparison of significantly mutated genes

TCGA Pan-Cancer Dataset [TCGA Research Network, 2013]

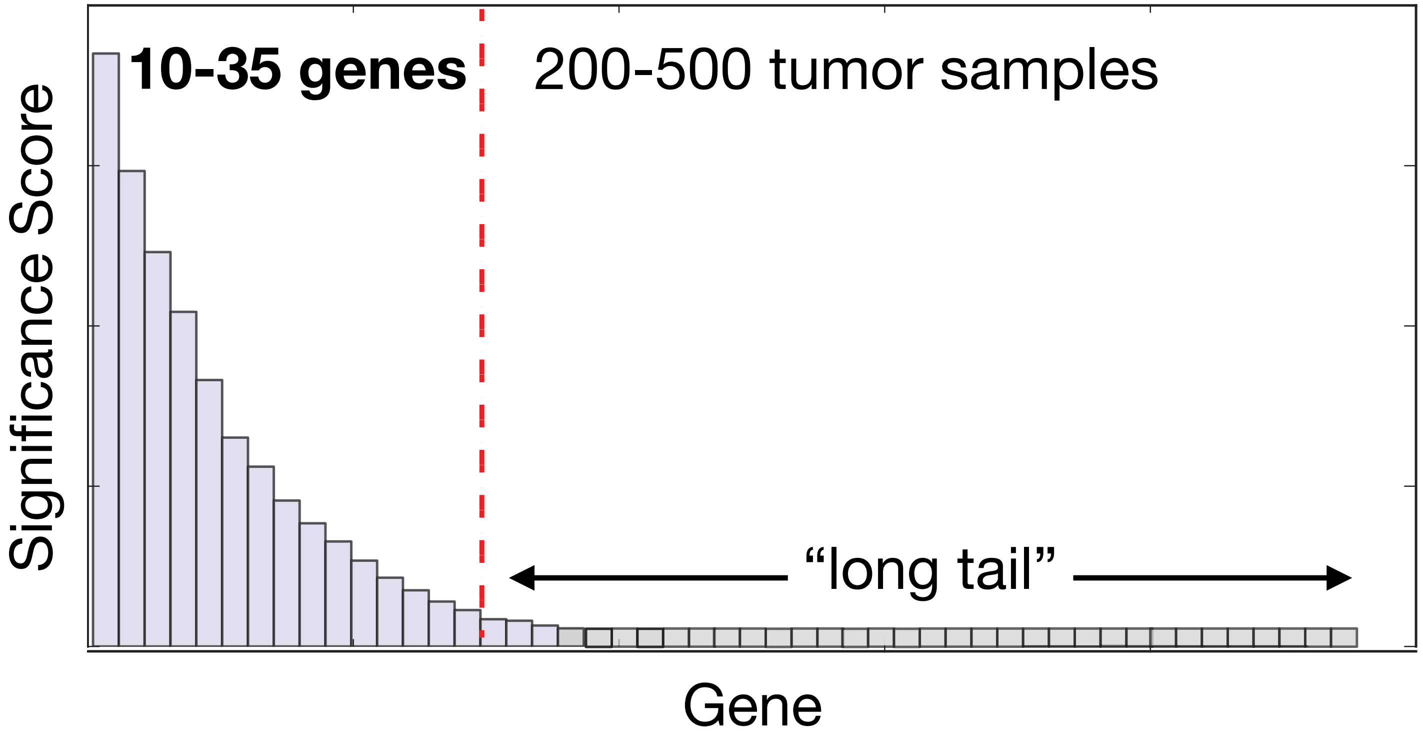
>3000 tumor samples of twelve cancer types

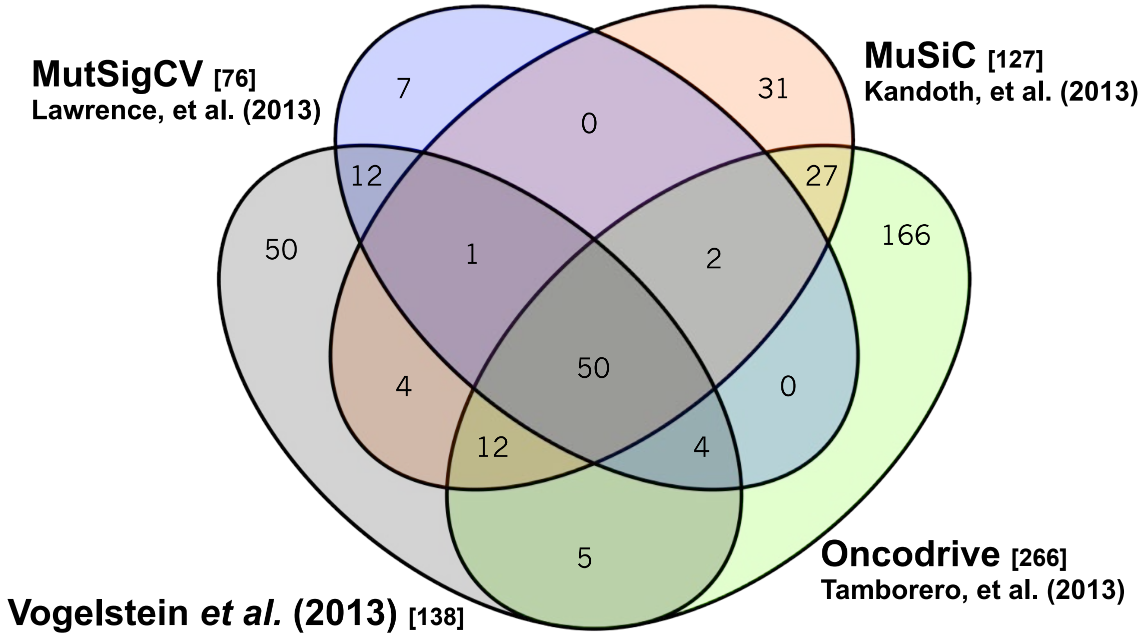
"Long tail" of mutated genes complicates finding driver mutations

Significance Score

Mutations weighted by:

- Gene length
- Mutation context
- Expression level
- Replication timing
- …





Comparison of significantly mutated genes

TCGA Pan-Cancer Dataset [TCGA Research Network, 2013]

>3000 tumor samples of twelve cancer types

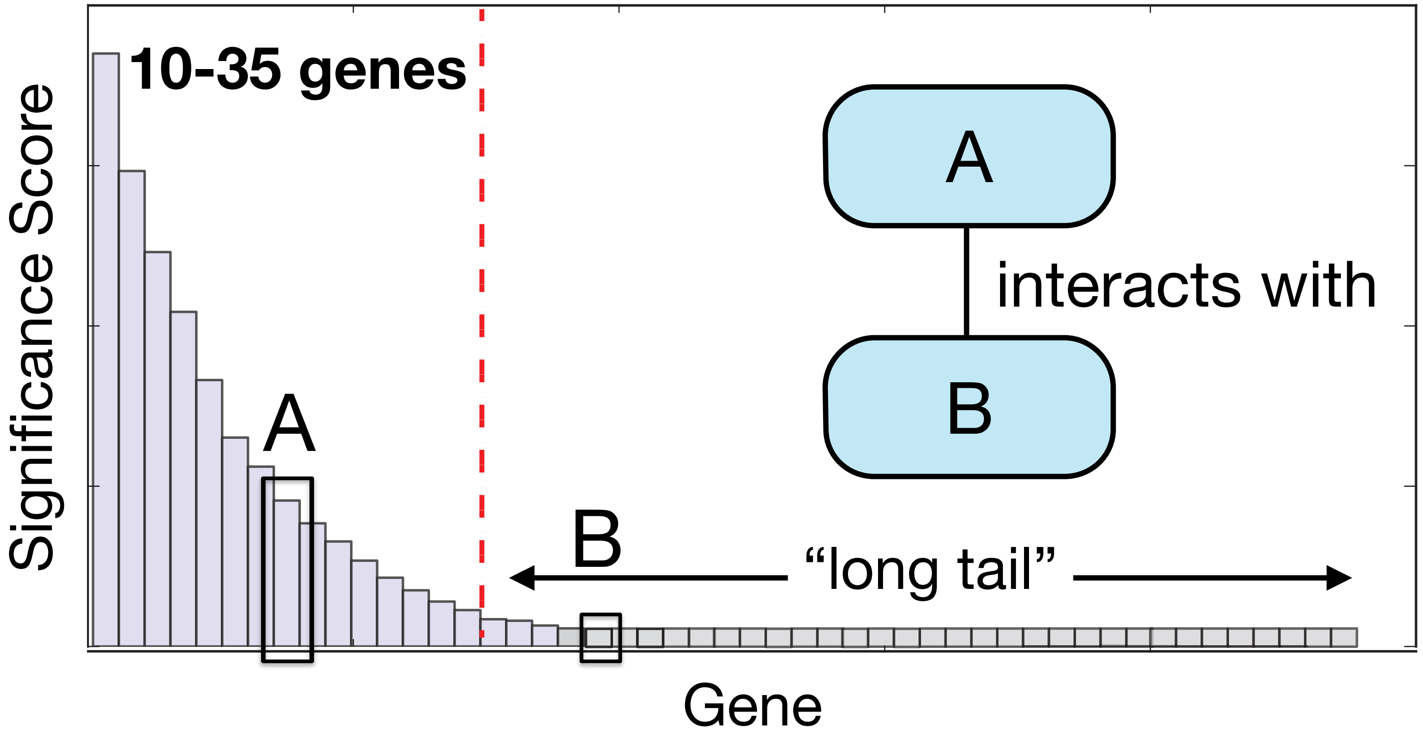
50 genes agreed upon by all methods
Hundreds found by only one method

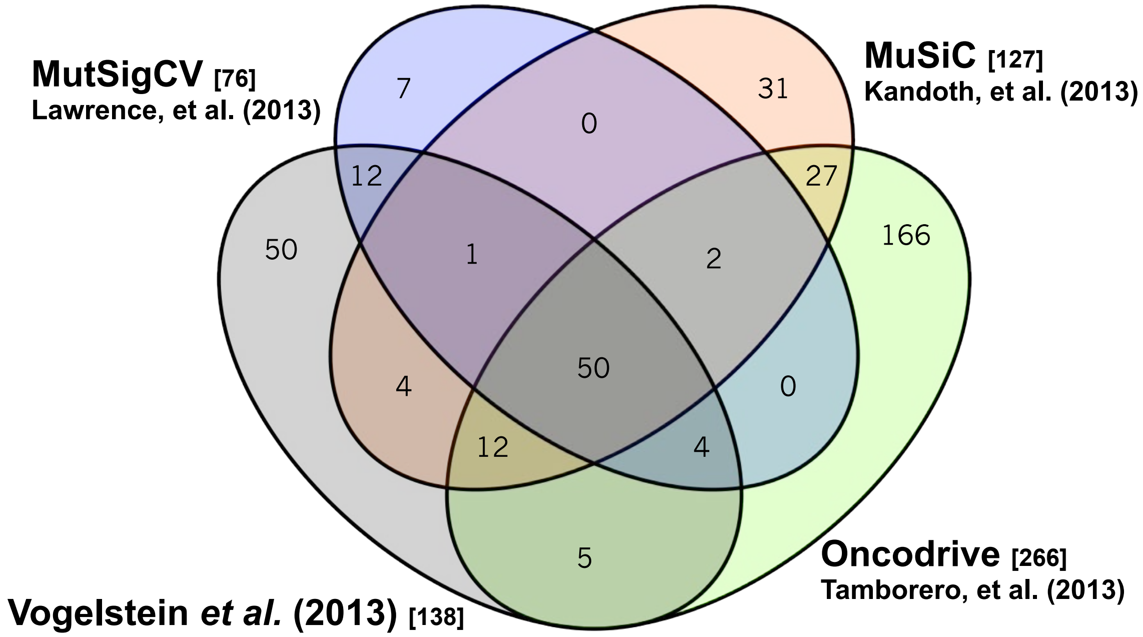
Mutations target pathways

Significance Score

Mutations weighted by:

- Gene length
- Mutation context
- Expression level
- Replication timing
- …





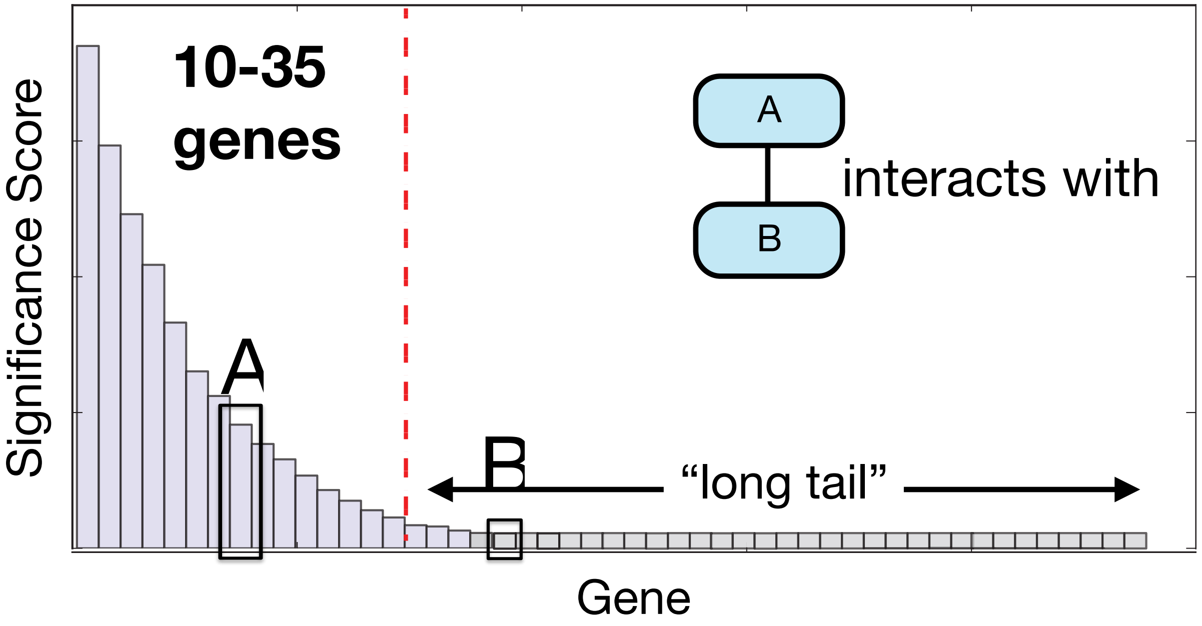
Comparison of significantly mutated genes

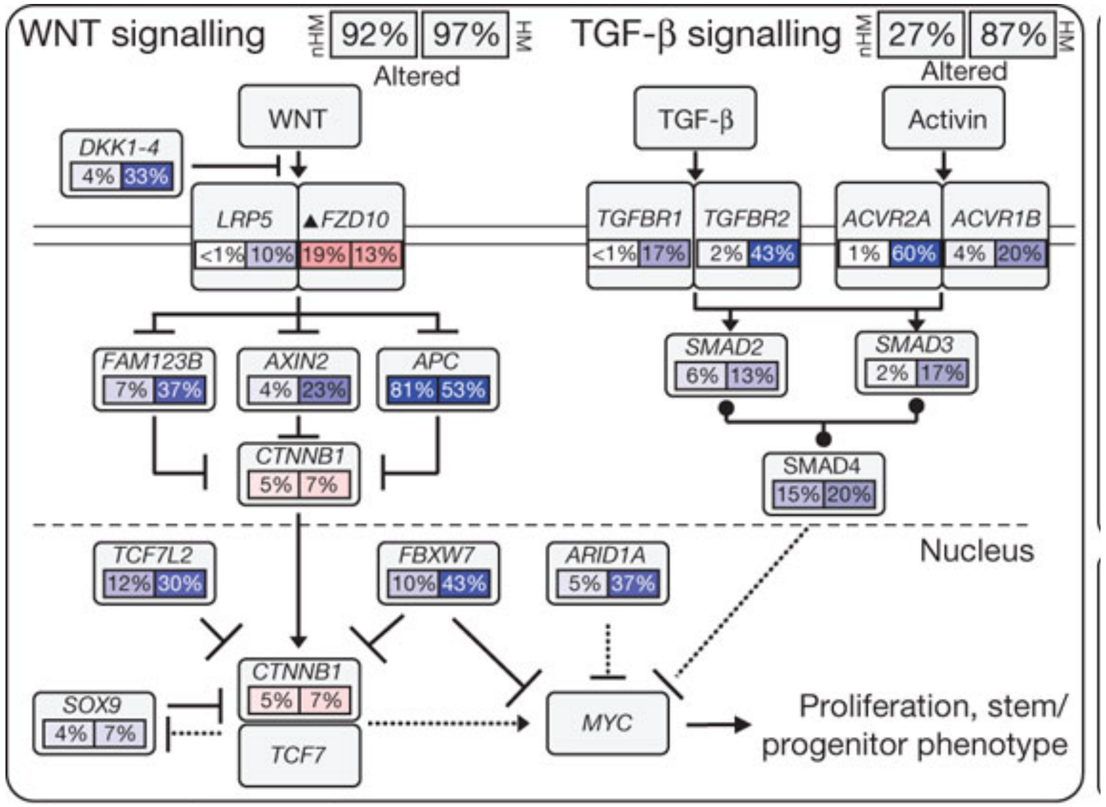
TCGA Pan-Cancer Dataset *[TCGA Research Network, 2013]*

>3000 tumor samples of twelve cancer types

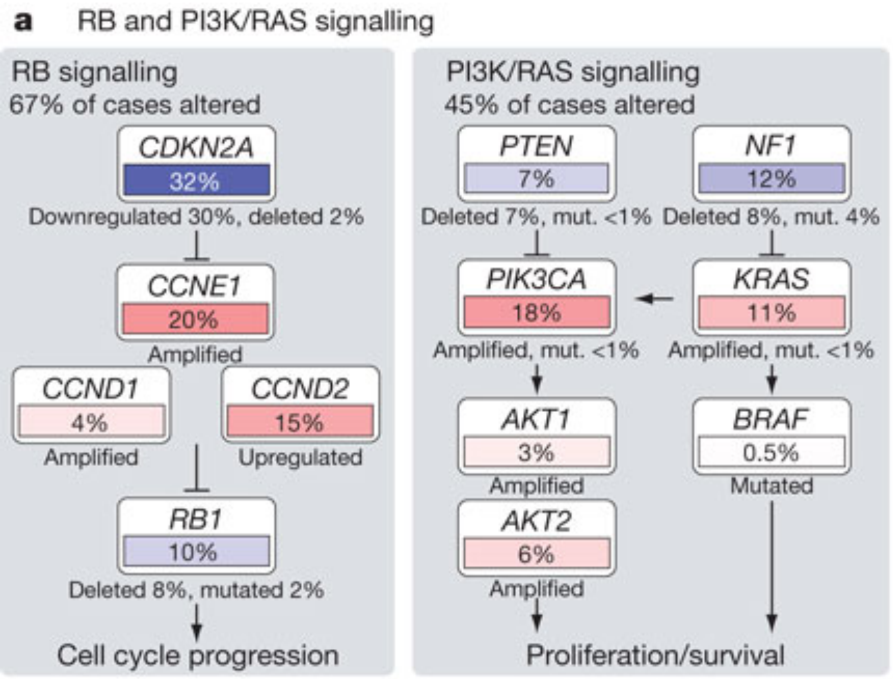
50 genes agreed upon by all methods
Hundreds found by only one method

Identifying significantly mutated pathways



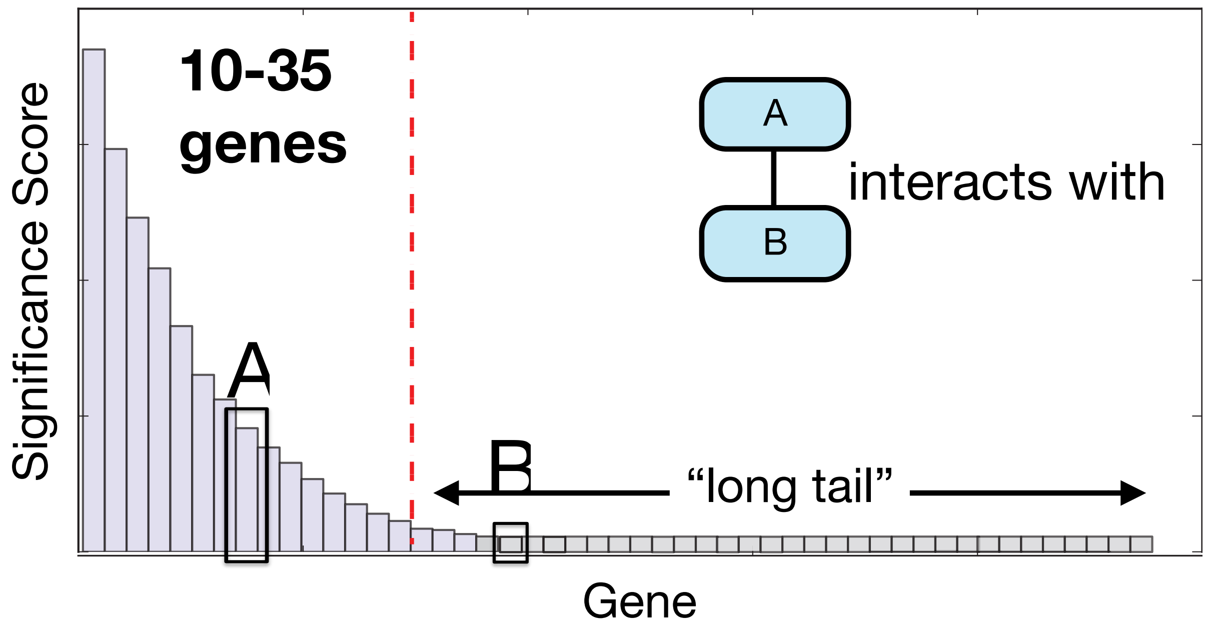


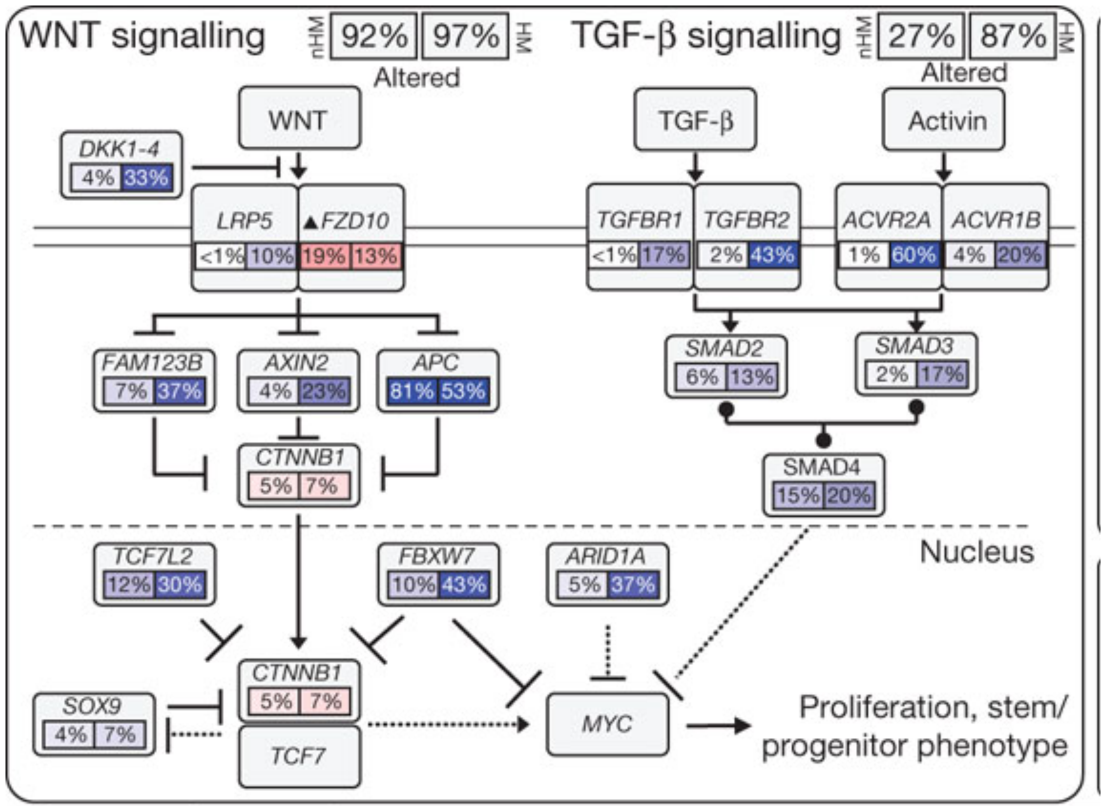
TCGA Colorectal (*Nature* 2012)



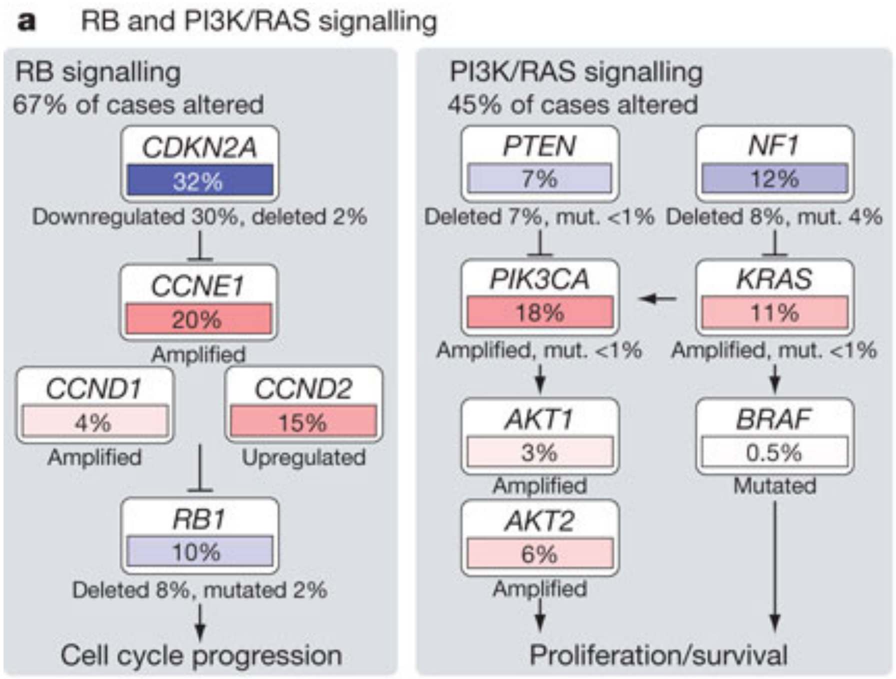
TCGA Ovarian (*Nature* 2011)

Identifying significantly mutated pathways





TCGA Colorectal (*Nature* 2012)



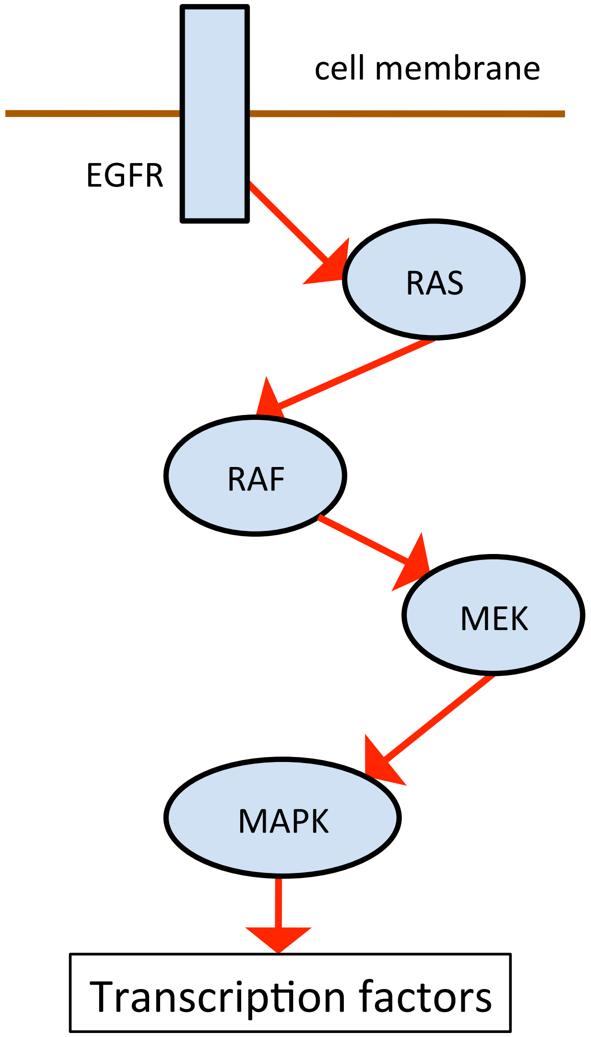
TCGA Ovarian (*Nature* 2011)

- Novel pathways
- *Crosstalk*
- Origin specific

Cancer pathways often harbor *mutually exclusive* mutations

Few driver mutations distributed across *multiple* pathways

→ Approximately one driver mutation per pathway per patient

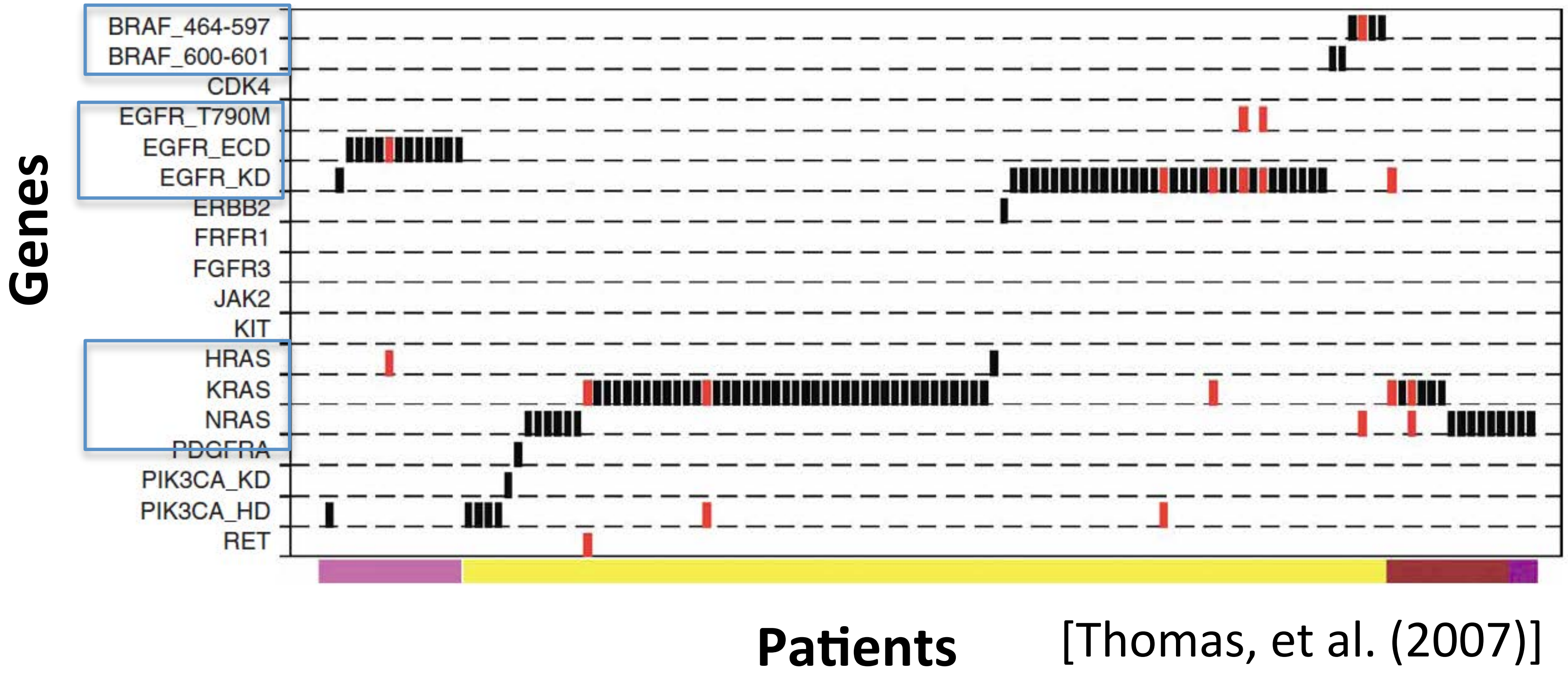


Cancer pathways often harbor *mutually exclusive* mutations

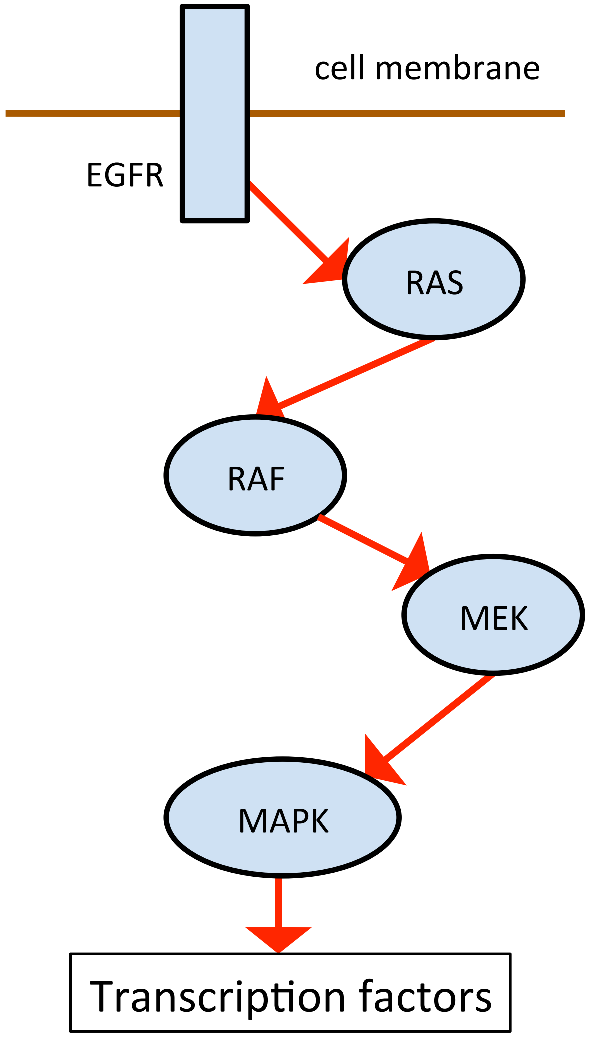
Few driver mutations distributed across *multiple* pathways

→ Approximately one driver mutation per pathway per patient

1. Exclusivity



[Thomas, et al. (2007)]



Cancer pathways often harbor *mutually exclusive* mutations

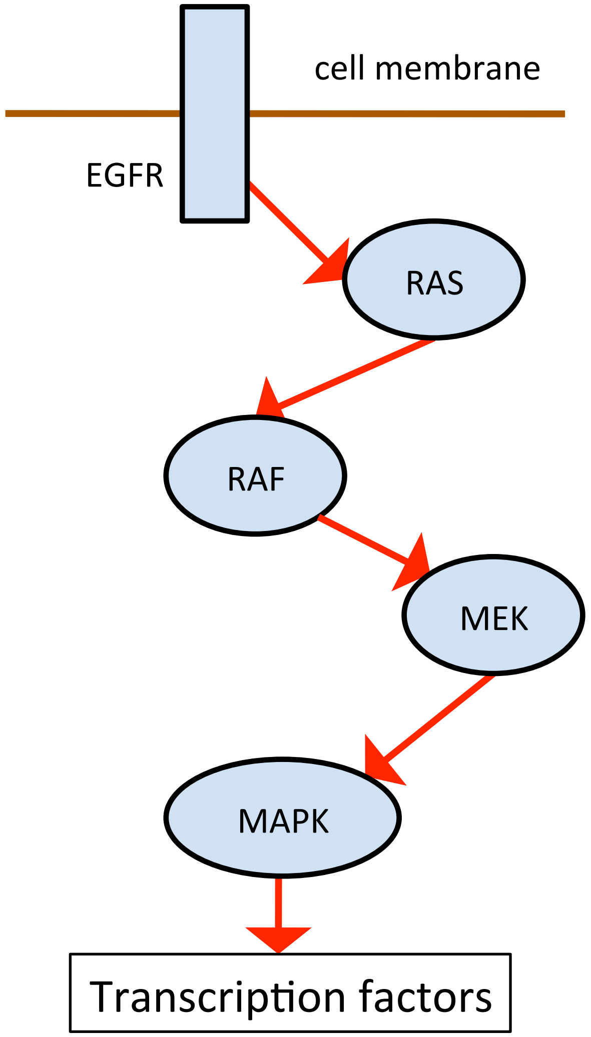
Few driver mutations distributed across *multiple* pathways

→ Approximately one driver mutation per pathway per patient

1. Exclusivity

→ Many patients have a mutation in important cancer pathway

2. High Coverage



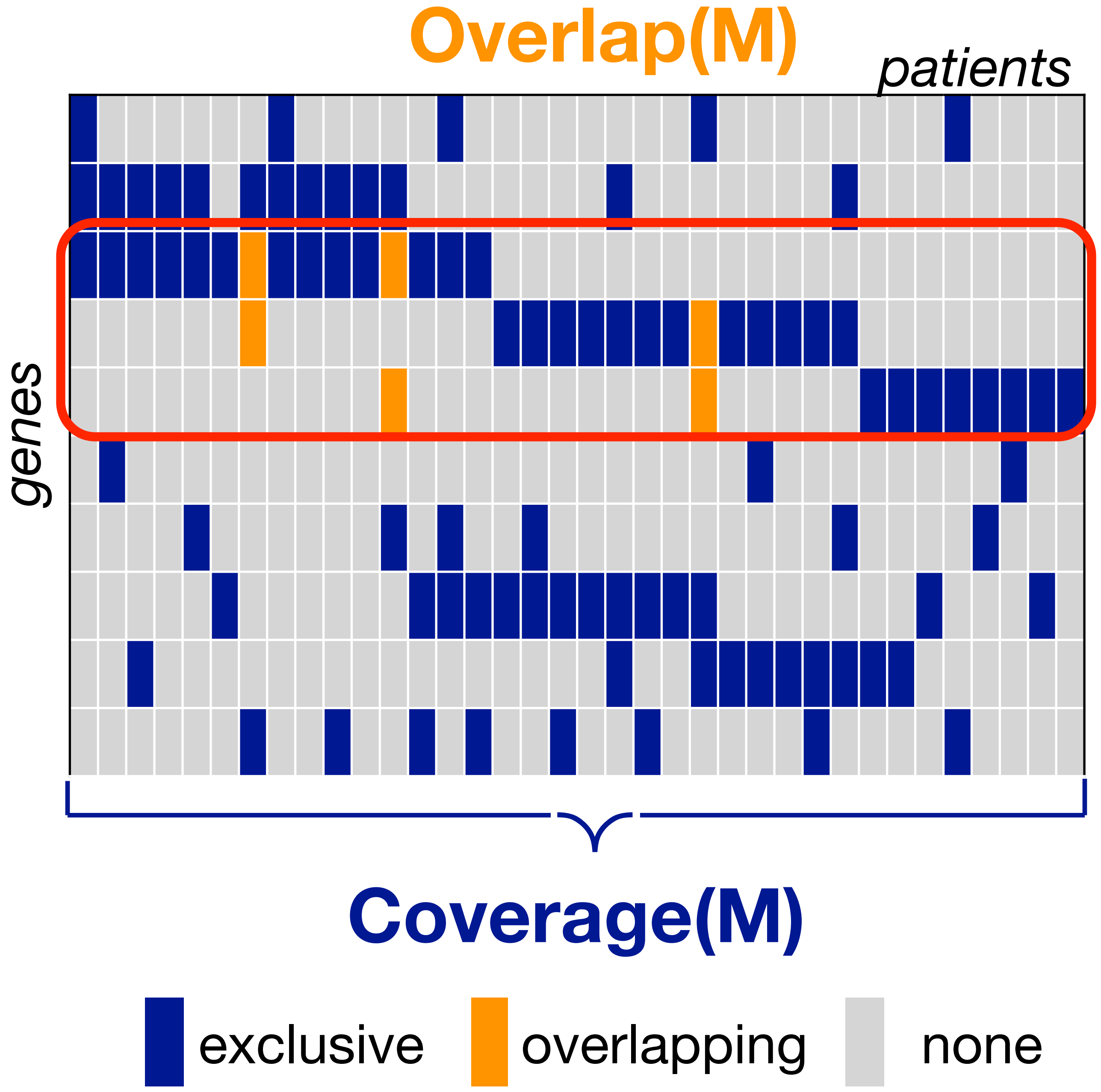
De novo driver exclusivity (Dendrix)

de novo: ***without prior biological information (pathways, interactions, etc.)***

Goal: Find sets **M** of genes with:

- *High coverage*: many patients with ≥ 1 mutation in **M**
- *Approximate exclusivity:* most patients have ≤ 1 mutation in **M**

Weight W(M)



- Finding optimal set is NP-Hard.
- MCMC algorithm samples sets in proportion to weight.

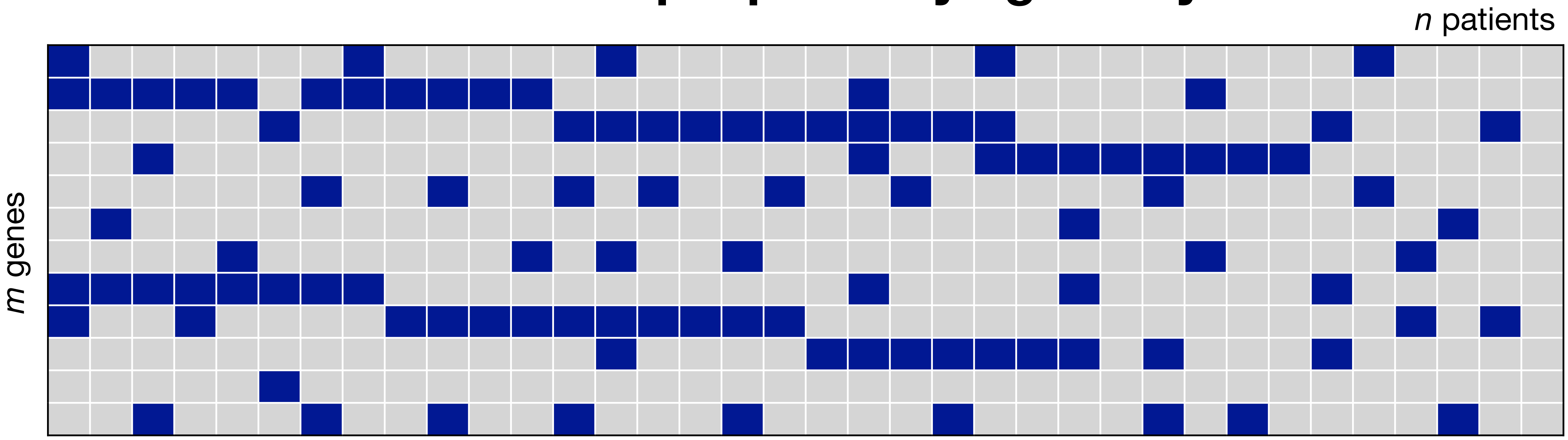
Vandin *et al*. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22:375-85.

Iter-Dendrix: multiple pathways

Cancer requires mutations to more than one pathway

⇩

Find multiple pathways greedily.



🟦 = gene *i* is mutated in patient *j*,

⬜ = otherwise.

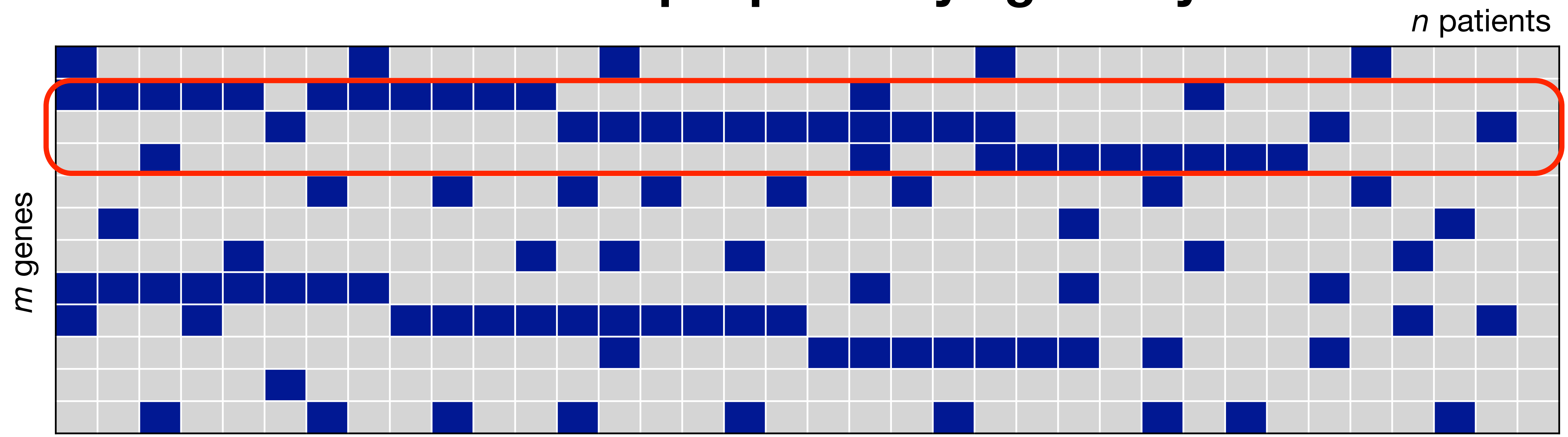
Vandin *et al*. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22:375-85.

Iter-Dendrix: multiple pathways

Cancer requires mutations to more than one pathway

⇩

Find multiple pathways greedily.



■ = gene *i* is mutated in patient *j*,

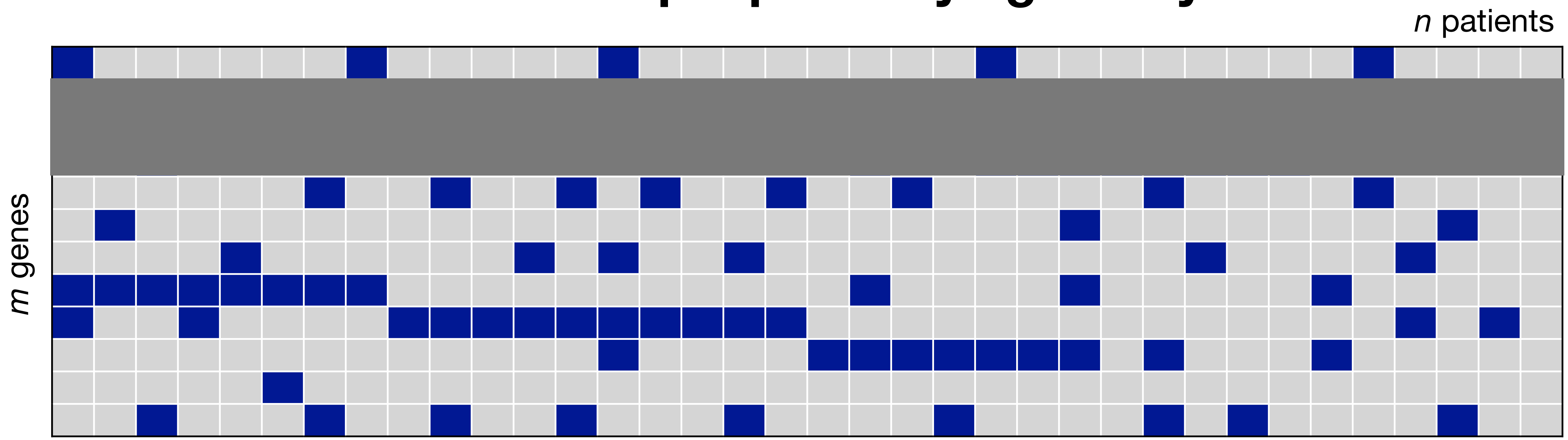
□ = otherwise.

Vandin *et al*. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22:375-85.

Iter-Dendrix: multiple pathways

Cancer requires mutations to more than one pathway

Find multiple pathways greedily.



🟦 = gene *i* is mutated in patient *j*,

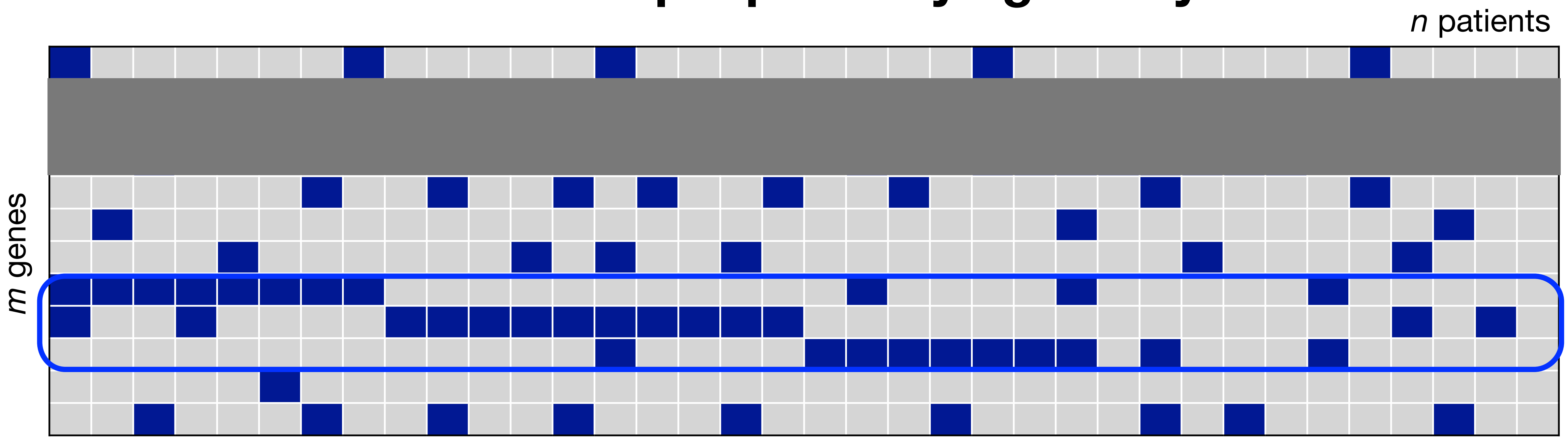
⬜ = otherwise.

Vandin *et al*. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22:375-85.

Iter-Dendrix: multiple pathways

Cancer requires mutations to more than one pathway

Find multiple pathways greedily.



🟦 = gene *i* is mutated in patient *j*,

⬜ = otherwise.

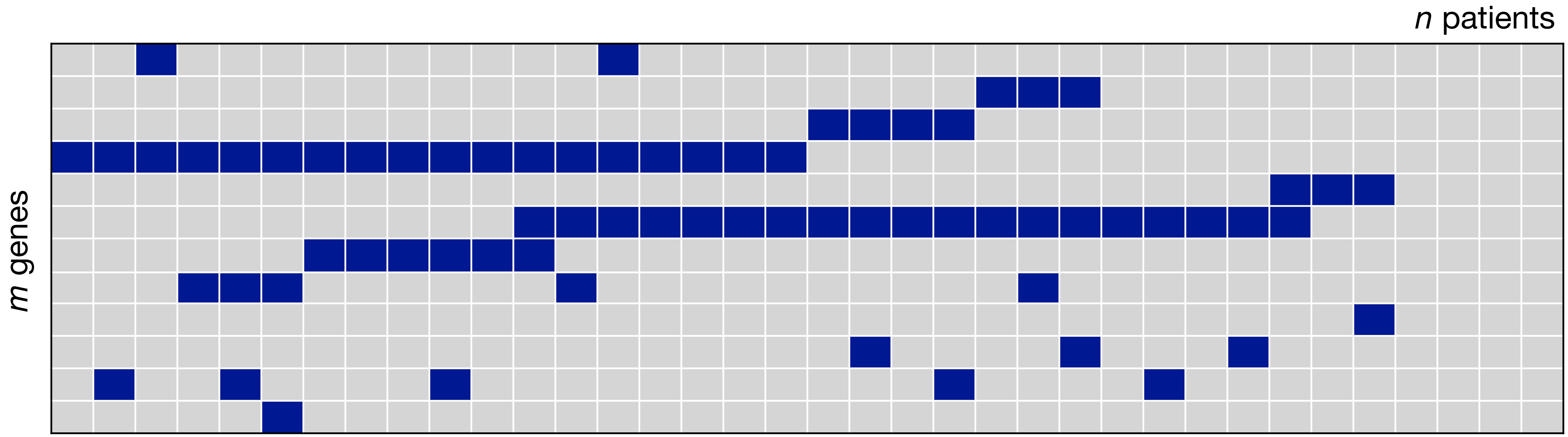
Vandin *et al*. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22:375-85.

Greedy can fail

Cancer requires mutations to more than one pathway

Find multiple pathways greedily.

Groups most frequently mutated genes even without exclusivity.



🟦 = gene *i* is mutated in patient *j*,
⬜ = otherwise.

Greedy can fail

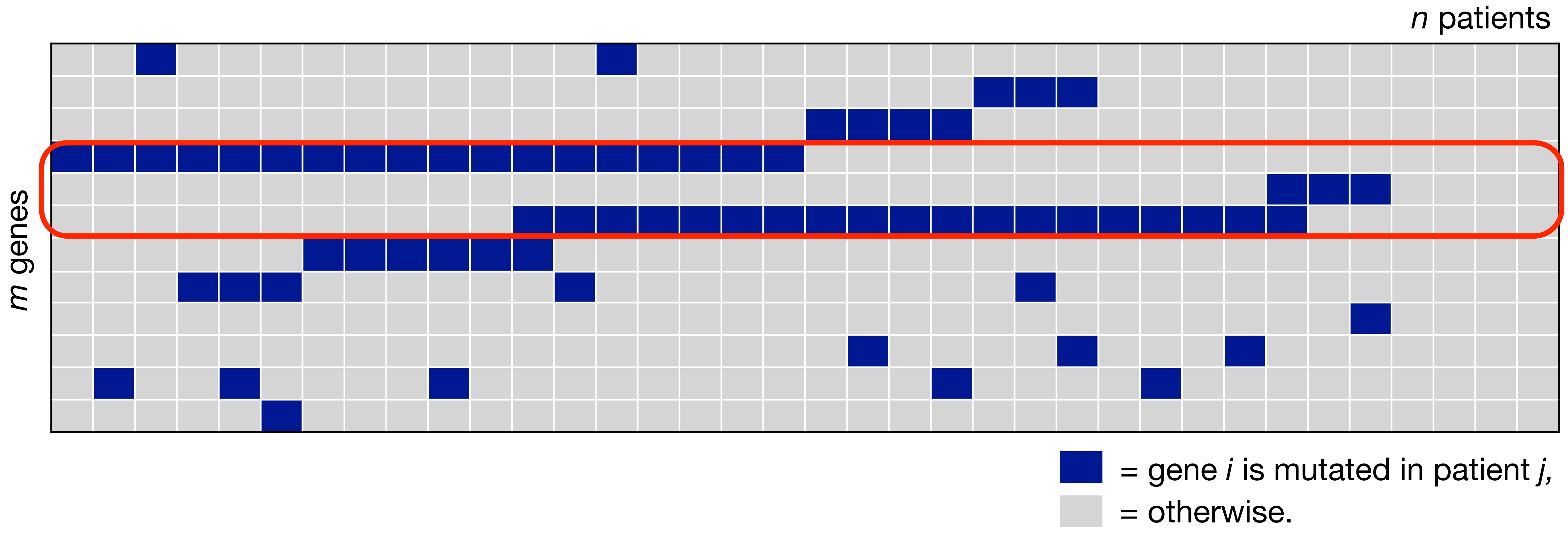
Cancer requires mutations to more than one pathway

↓

Find multiple pathways greedily.

↓

Groups most frequently mutated genes even without exclusivity.

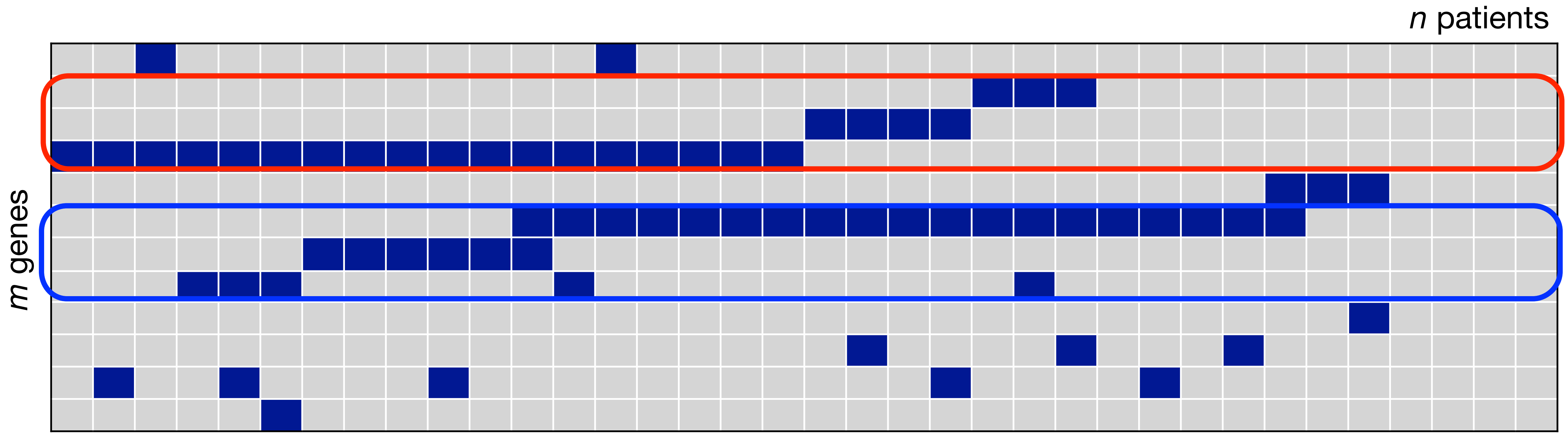


Greedy can fail

Cancer requires mutations to more than one pathway

Find multiple pathways greedily.

Groups most frequently mutated genes even without exclusivity.



🟦 = gene *i* is mutated in patient *j*,

⬜ = otherwise.

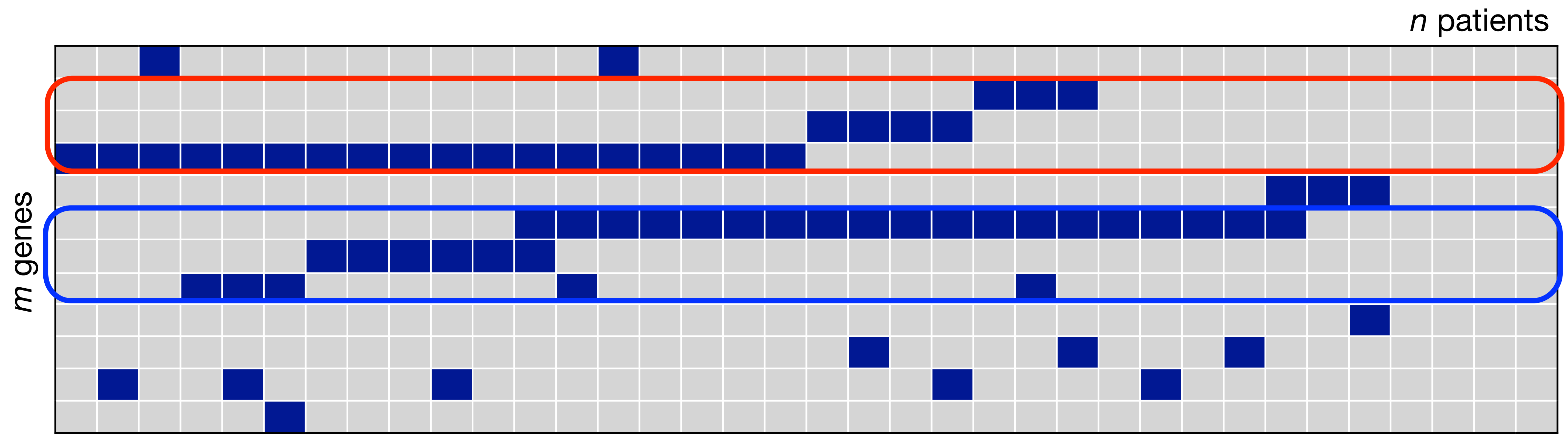
Multi-Dendrix

Cancer requires mutations to more than one pathway

⇩

Find <u>multiple</u> pathways simultaneously

- ILP rapidly finds optimal solution
- Searches wide range of parameters to find stable collections of gene sets



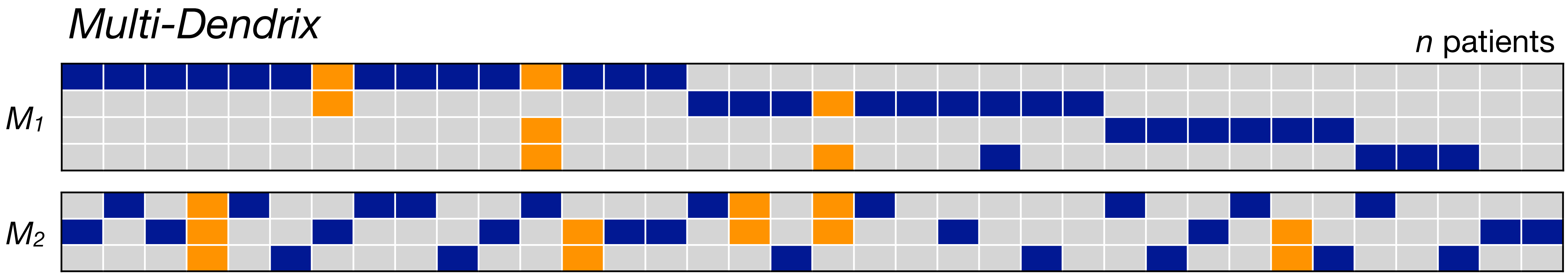
■ = gene *i* is mutated in patient *j*,

□ = otherwise.

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix

Cancer requires mutations to more than one pathway



Most samples have approximately one mutation in *each* of *t* pathways.

maximize $W'_\alpha(\mathbf{M}) = \sum_{M\in\mathbf{M}} |\Gamma(M)| - \alpha\omega(M)$

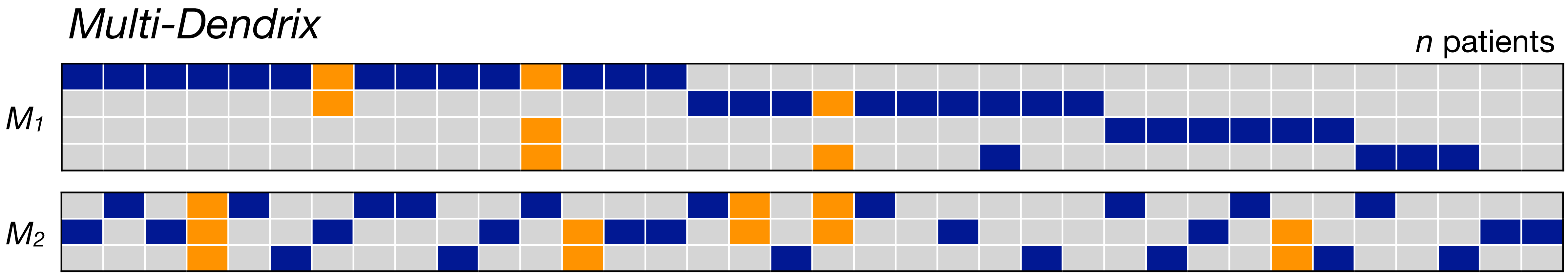
coverage overlap

Integer Linear Program

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix

Cancer requires mutations to more than one pathway



Most samples have approximately one mutation in *each* of *t* pathways.

$$\textbf{maximize } W'_{\alpha}(\mathbf{M}) = \sum_{M \in \mathbf{M}} |\Gamma(M)| - \alpha\omega(M)$$

parameter to control “weight” of exclusivity

coverage

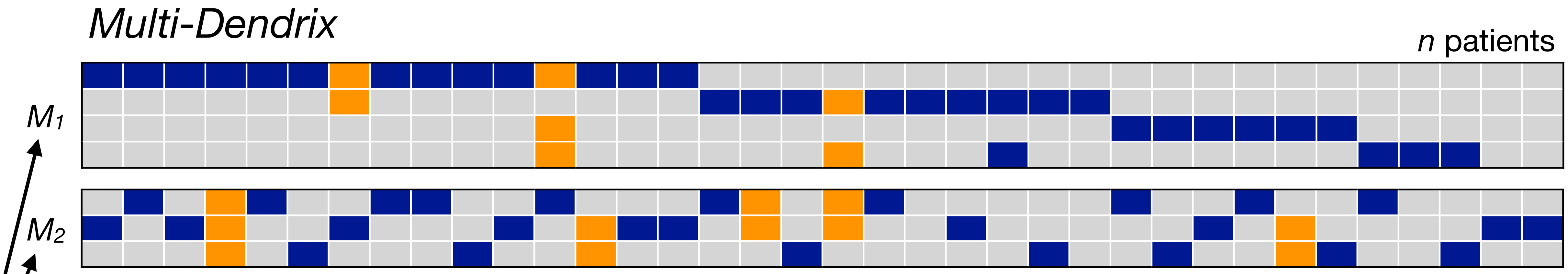
overlap

Integer Linear Program

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix

Cancer requires mutations to more than one pathway



Most samples have approximately one mutation in *each* of *t* pathways.

$|M_i| \in [k_{\min}, k_{\max}]$

$$\textbf{maximize } W'_\alpha(\mathbf{M}) = \sum_{M \in \mathbf{M}} |\Gamma(M)| - \alpha\omega(M)$$

parameter to control “weight” of exclusivity

coverage

overlap

Integer Linear Program

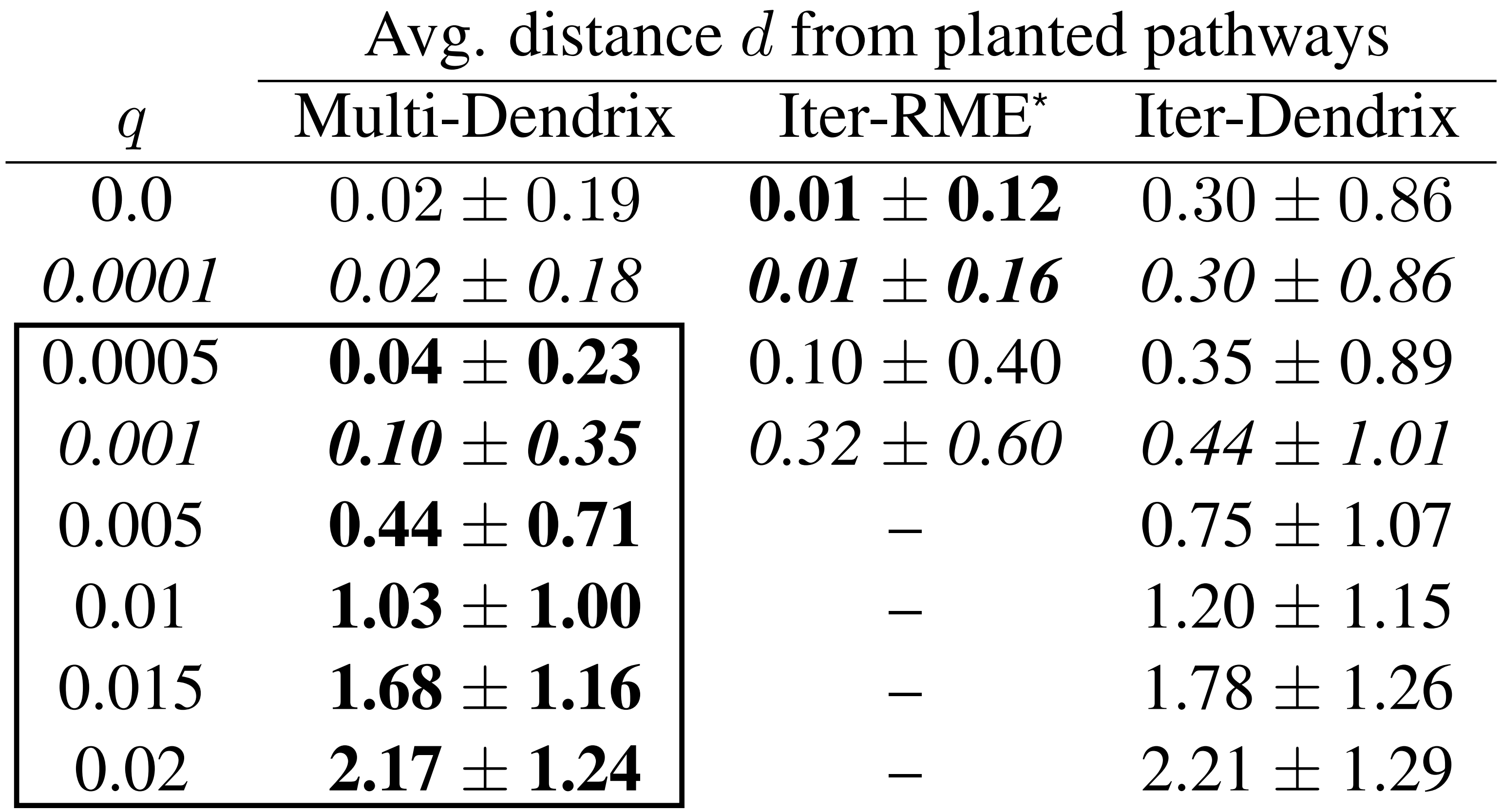
Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Contributions

A new algorithm, **Multi-Dendrix**, for identifying driver pathways *de novo:*

1. **Outperforms previous methods on simulated data in speed and accuracy**
2. Identifies gene sets that overlap known pathways in TCGA datasets
3. Ongoing work extending Multi-Dendrix to large datasets and overlapping pathways

Multi-Dendrix is significantly better on simulated data

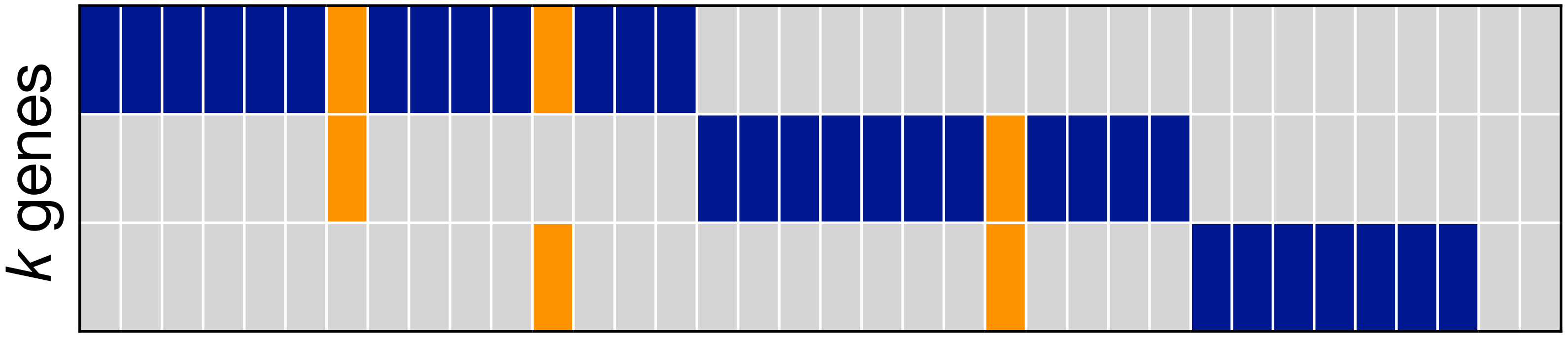


	Avg. distance d from planted pathways		
q	Multi-Dendrix	Iter-RME*	Iter-Dendrix
0.0	0.02 ± 0.19	**0.01 ± 0.12**	0.30 ± 0.86
0.0001	*0.02 ± 0.18*	***0.01 ± 0.16***	*0.30 ± 0.86*
0.0005	**0.04 ± 0.23**	0.10 ± 0.40	0.35 ± 0.89
0.001	***0.10 ± 0.35***	*0.32 ± 0.60*	*0.44 ± 1.01*
0.005	**0.44 ± 0.71**	–	0.75 ± 1.07
0.01	**1.03 ± 1.00**	–	1.20 ± 1.15
0.015	**1.68 ± 1.16**	–	1.78 ± 1.26
0.02	**2.17 ± 1.24**	–	2.21 ± 1.29

$P < 0.01$

360 mutated genes in 160 patients

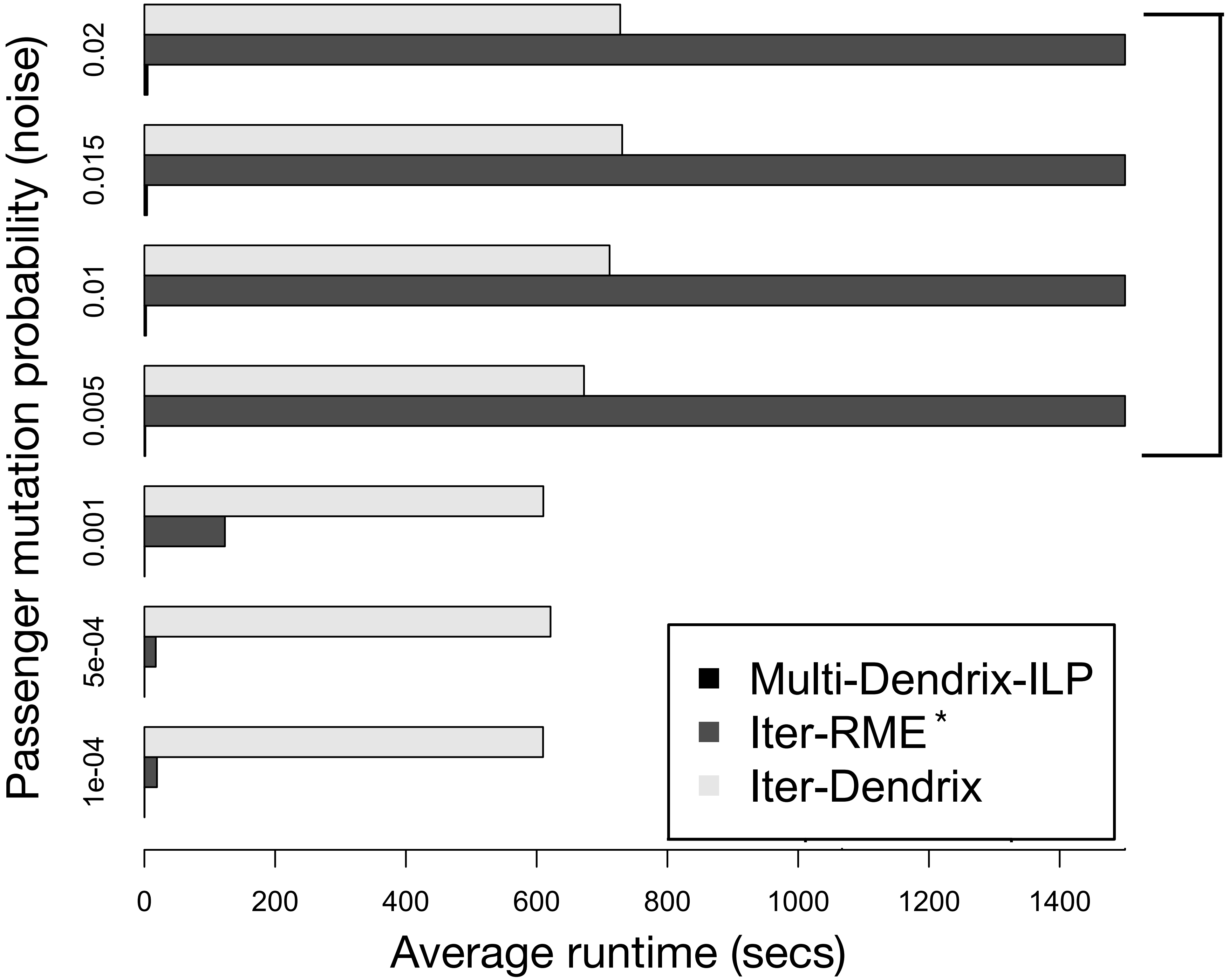
Example planted pathway



- High coverage and mutually exclusive
- Add passenger mutations (noise) to all genes

*Miller *et al*. (2011). *BMC medical genomics*, **4**:34.

Multi-Dendrix is significantly **faster** on simulated data



Multi-Dendrix >2 orders of magnitude faster

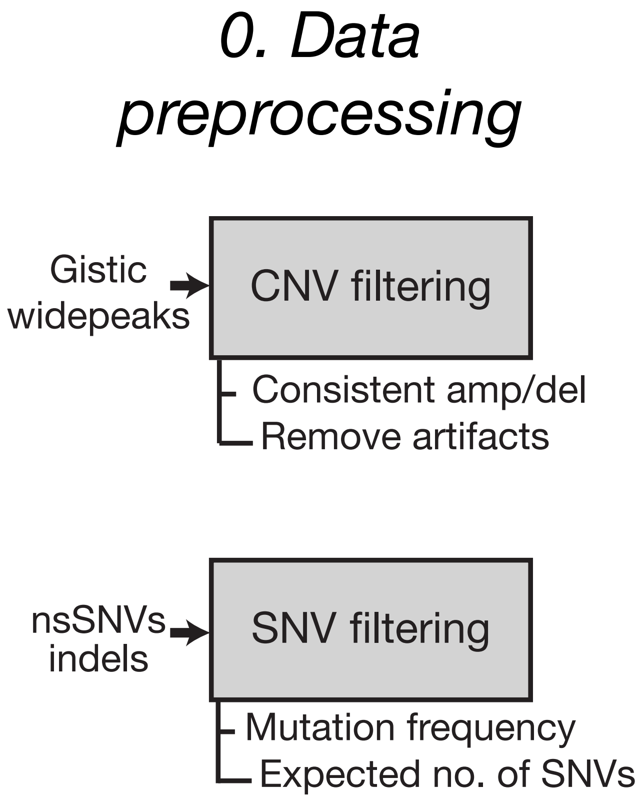
Multi-Dendrix completes in **fewer than 5 seconds** for each dataset.

*Miller *et al*. (2011). *BMC medical genomics*, **4**:34.

Multi-Dendrix pipeline for identifying mutated cancer pathways

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix pipeline for identifying mutated cancer pathways



0. Data preprocessing

Gistic widepeaks

CNV filtering

Consistent amp/del

Remove artifacts

nsSNVs indels

SNV filtering

Mutation frequency

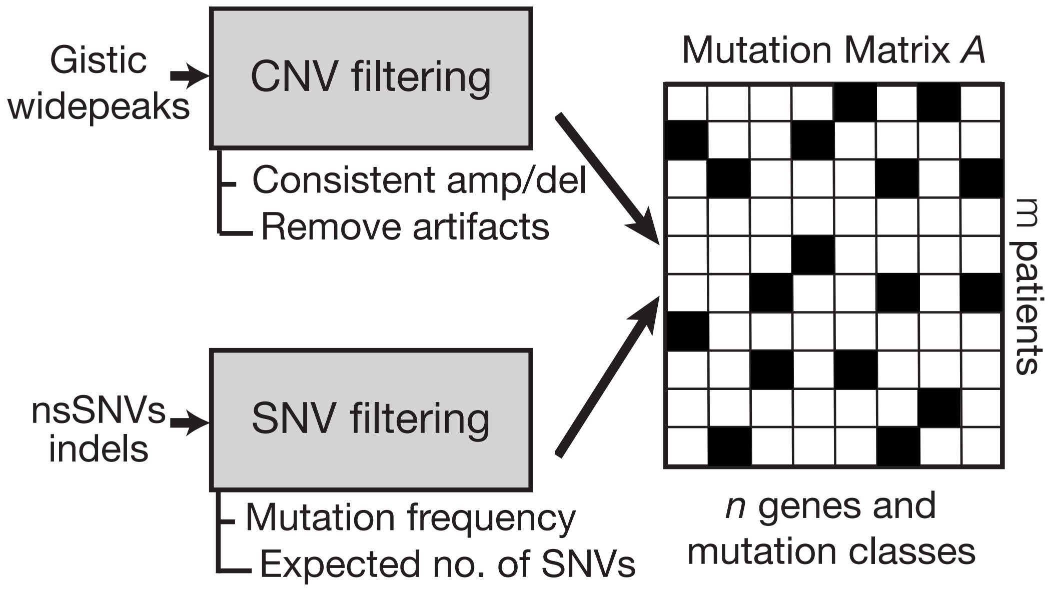
Expected no. of SNVs

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix pipeline for identifying mutated cancer pathways

0. Data preprocessing

1. Input mutation matrix



Gistic widepeaks

CNV filtering

- Consistent amp/del
- Remove artifacts

nsSNVs indels

SNV filtering

- Mutation frequency
- Expected no. of SNVs

Mutation Matrix *A*

m patients

n genes and mutation classes

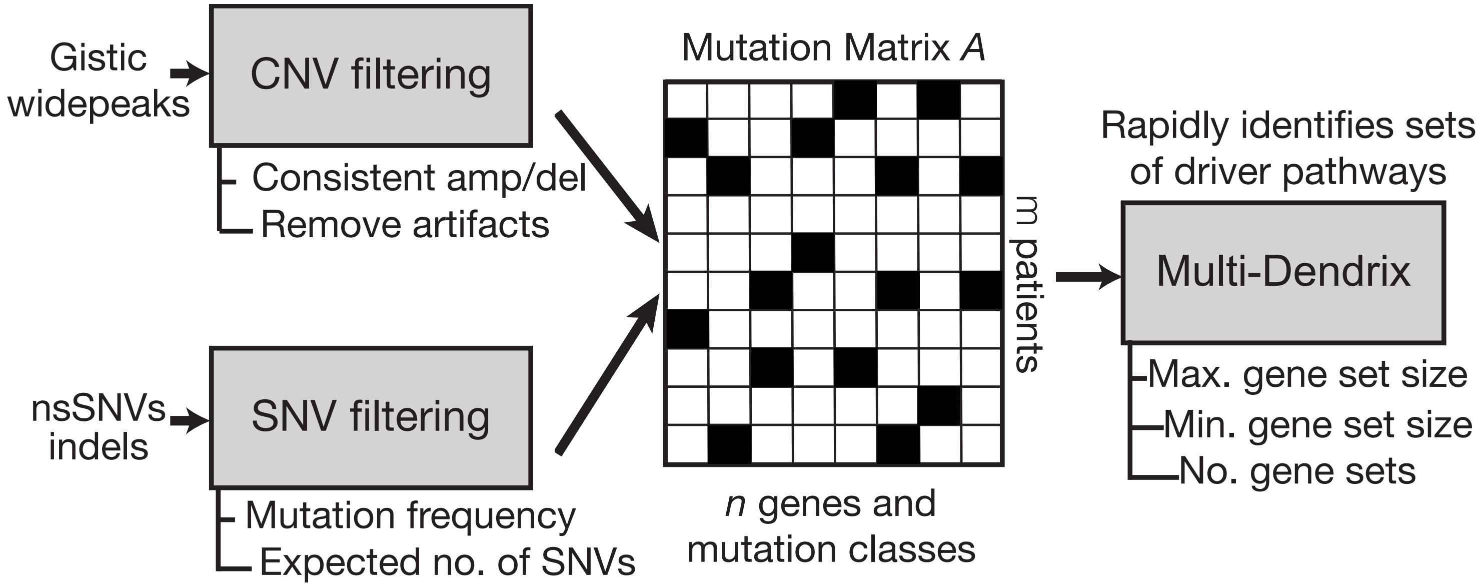
Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix pipeline for identifying mutated cancer pathways

0. Data preprocessing

1. Input mutation matrix

2. Analysis and annotation



Gistic widepeaks

CNV filtering

- Consistent amp/del
- Remove artifacts

nsSNVs indels

SNV filtering

- Mutation frequency
- Expected no. of SNVs

Mutation Matrix *A*

m patients

n genes and mutation classes

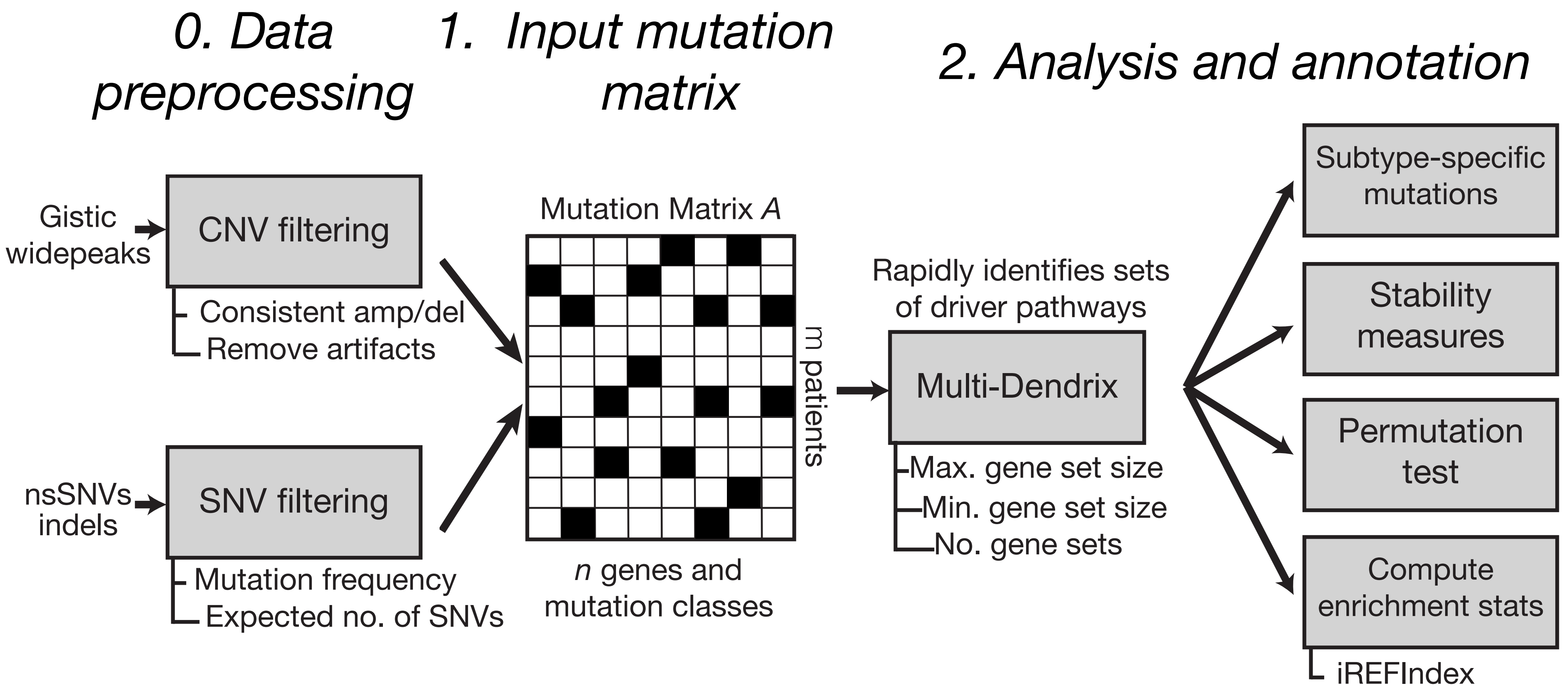
Rapidly identifies sets of driver pathways

Multi-Dendrix

- Max. gene set size
- Min. gene set size
- No. gene sets

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix pipeline for identifying mutated cancer pathways



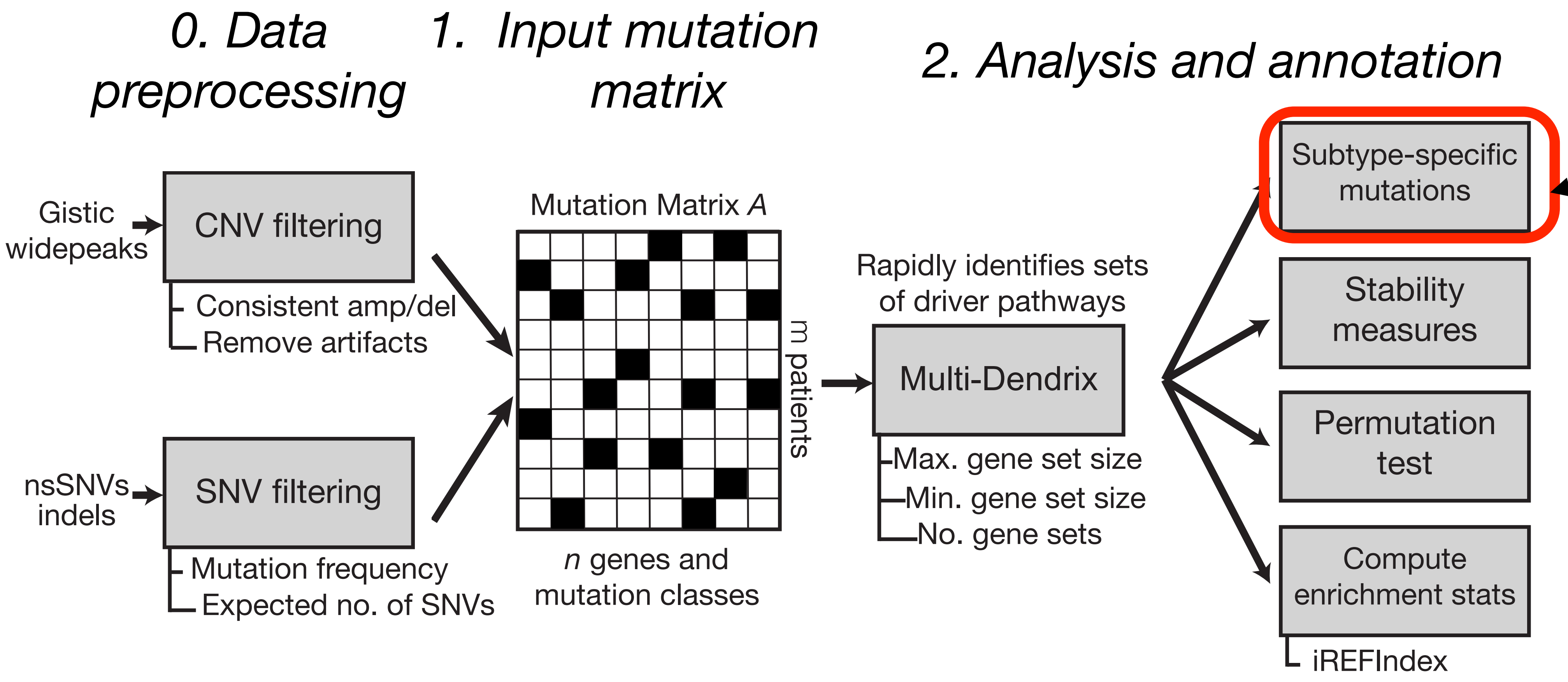
0. Data preprocessing

1. Input mutation matrix

2. Analysis and annotation

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

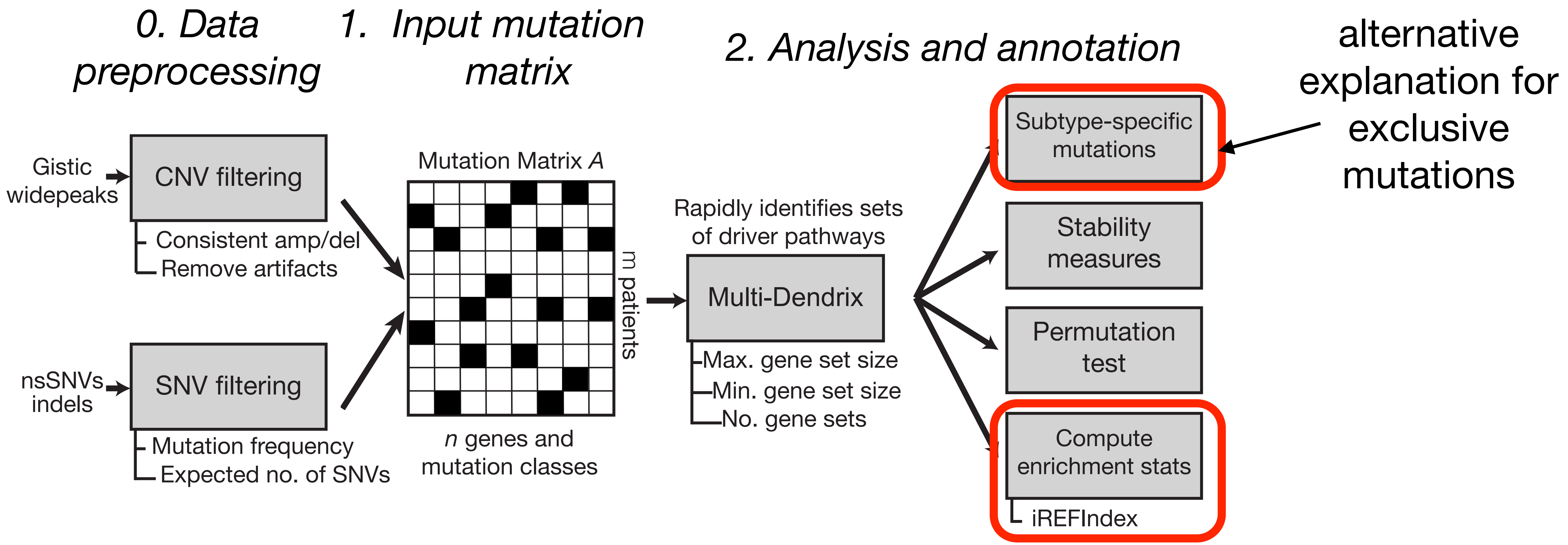
Multi-Dendrix pipeline for identifying mutated cancer pathways



alternative explanation for exclusive mutations

Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix pipeline for identifying mutated cancer pathways



0. Data preprocessing

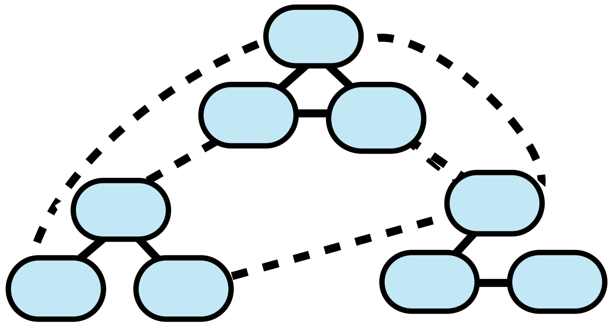
1. Input mutation matrix

2. Analysis and annotation

alternative explanation for exclusive mutations

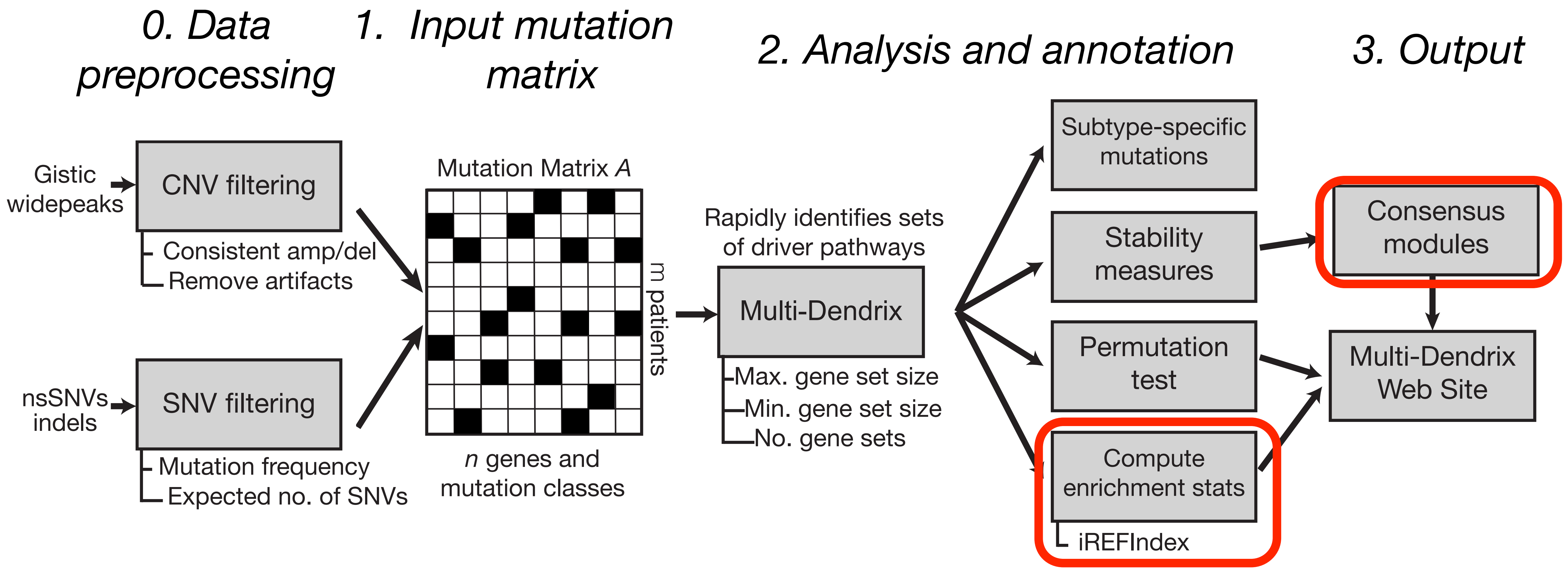
Direct Interactions Test

- Measures enrichment of PPI interactions within individual gene sets or in a collection of gene sets



Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Multi-Dendrix pipeline for identifying mutated cancer pathways



0. Data preprocessing

1. Input mutation matrix

2. Analysis and annotation

3. Output

Gistic widepeaks

CNV filtering

- Consistent amp/del
- Remove artifacts

nsSNVs indels

SNV filtering

- Mutation frequency
- Expected no. of SNVs

Mutation Matrix *A*

m patients

n genes and mutation classes

Rapidly identifies sets of driver pathways

Multi-Dendrix

- Max. gene set size
- Min. gene set size
- No. gene sets

Subtype-specific mutations

Stability measures

Permutation test

Compute enrichment stats

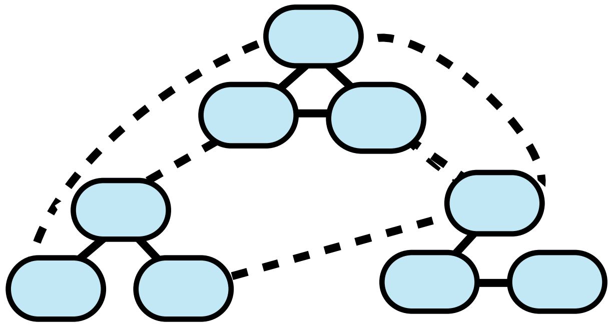
- iREFIndex

Consensus modules

Multi-Dendrix Web Site

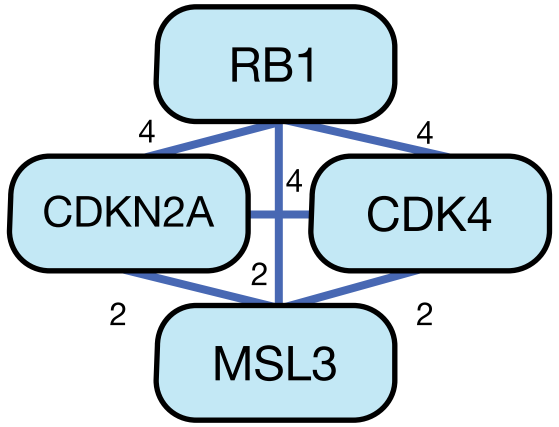
Direct Interactions Test

- Measures enrichment of PPI interactions within individual gene sets or in a collection of gene sets



Consensus modules

- Run Multi-Dendrix across a range of parameters.
- Identify the stable “modules” of genes that appear together multiple times.



Leiserson *et al*. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* 9(**5**): e1003054.

Contributions

A new algorithm, **Multi-Dendrix**, for identifying driver pathways *de novo:*

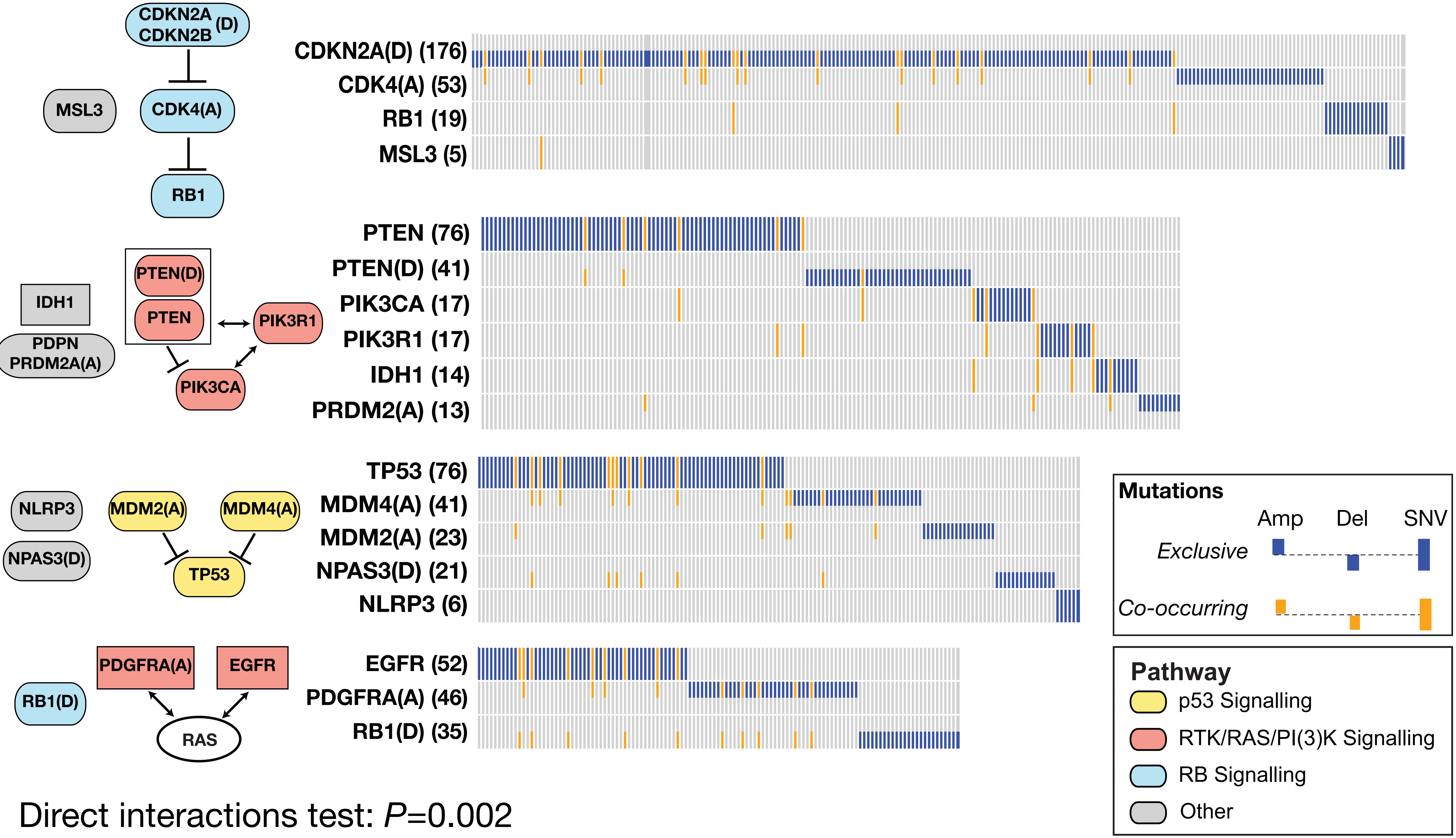
1. Outperforms previous methods on simulated data in speed and accuracy

2. **Identifies gene sets that overlap known pathways in TCGA datasets**

3. Ongoing work extending Multi-Dendrix to large datasets and overlapping pathways

Results: Glioblastoma

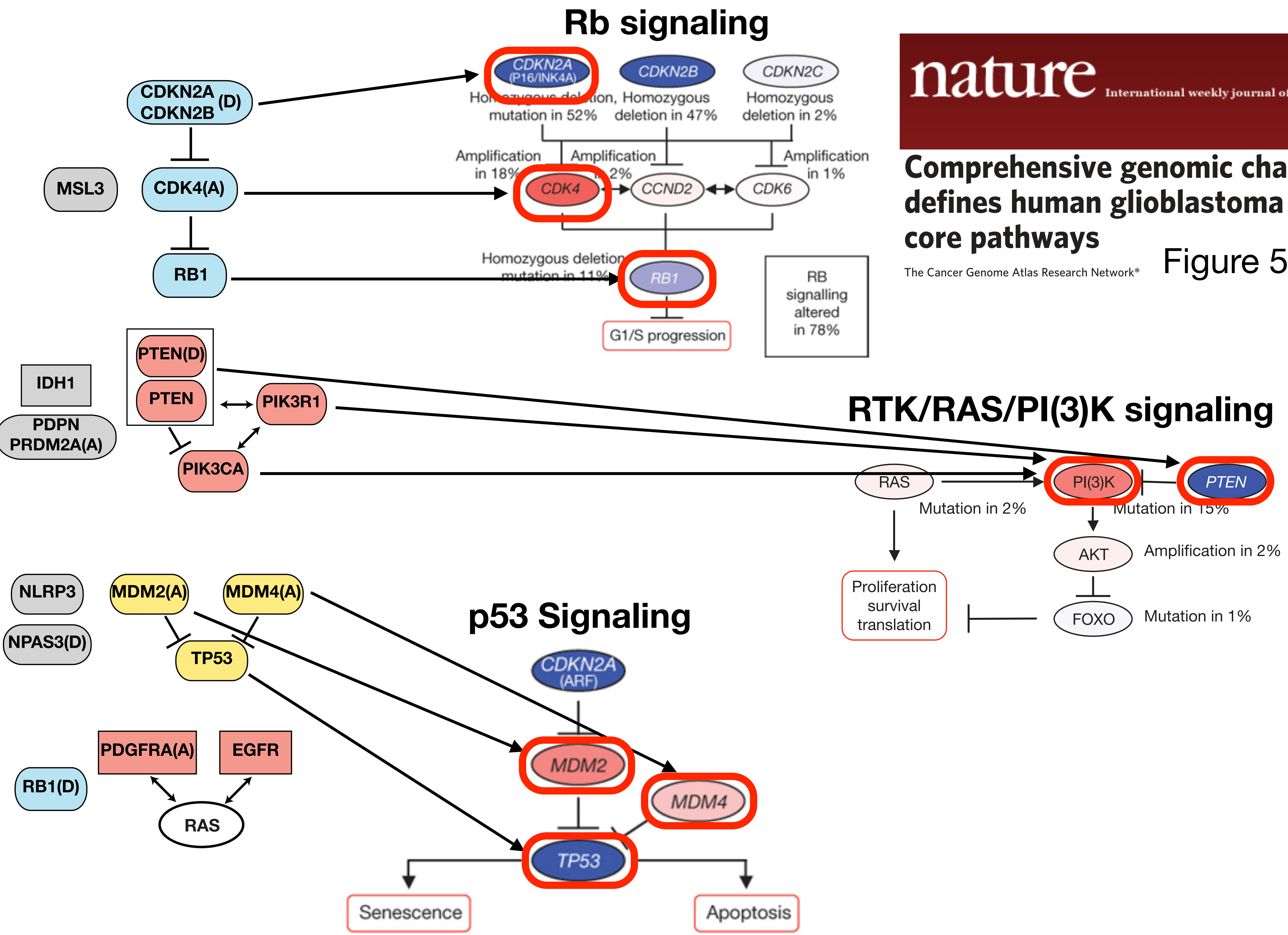
**Mutation data:* 398 genes (events) in 261 patients



Direct interactions test: *P*=0.002

*TCGA (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature **455**:1061-1088.*

Results: Glioblastoma



nature International weekly journal of science

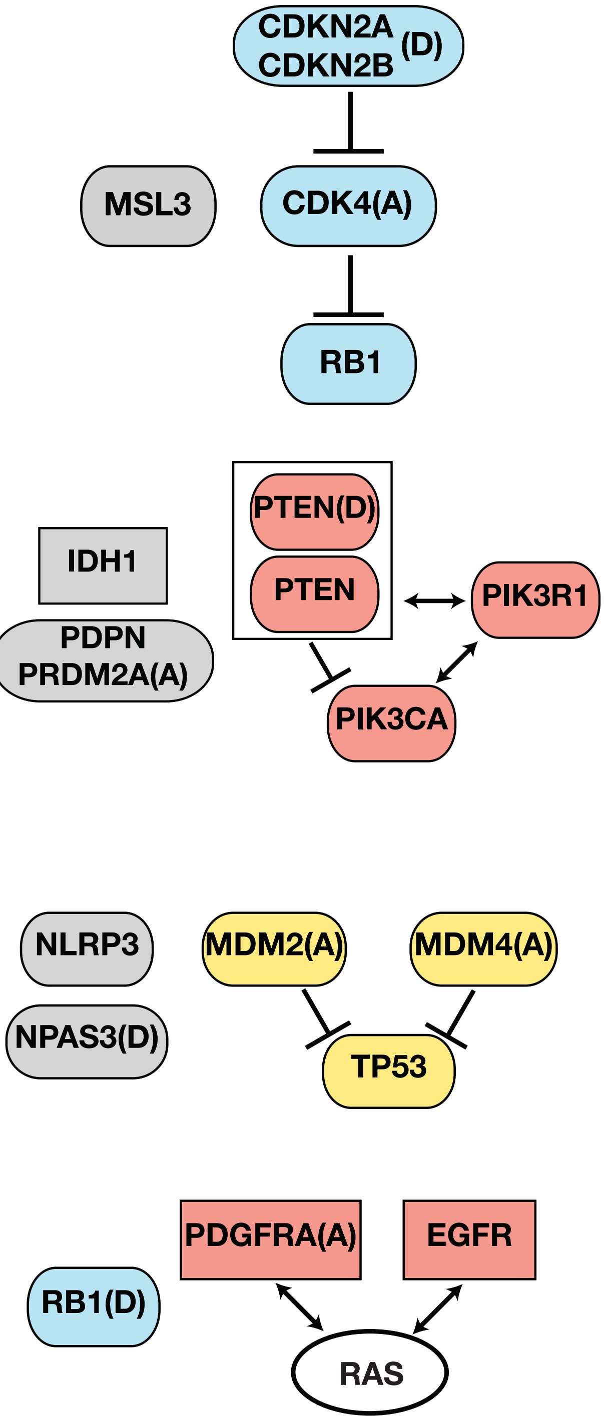
Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

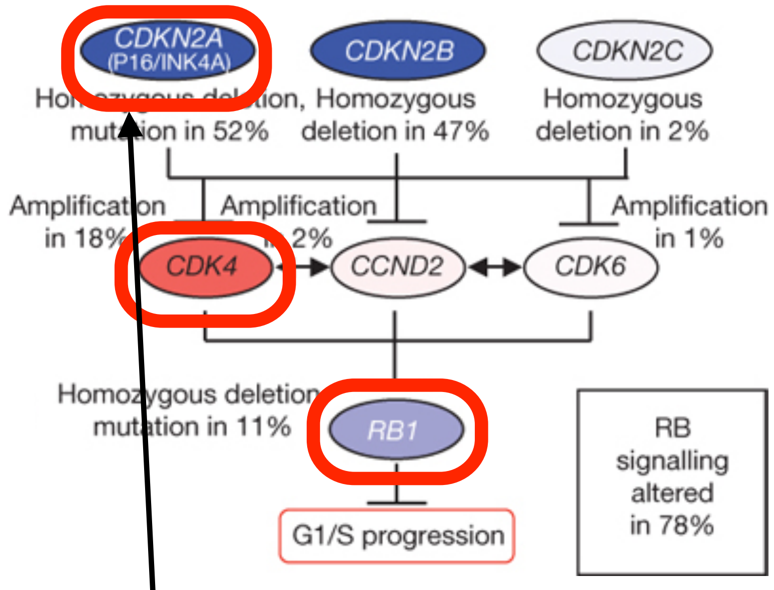
Figure 5

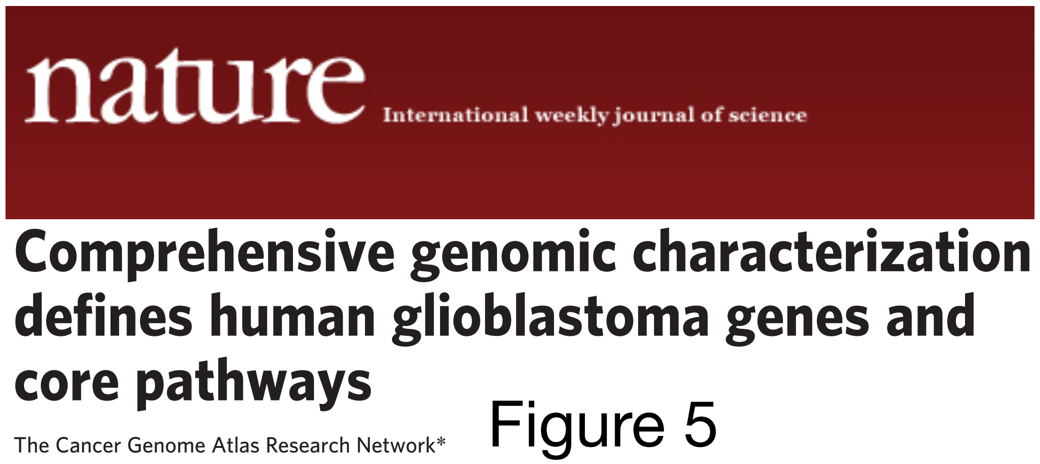
*TCGA (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* ***455****:1061-1088.*

Results: Glioblastoma



Rb signaling


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
G1/S progression
RB signalling altered in 78%



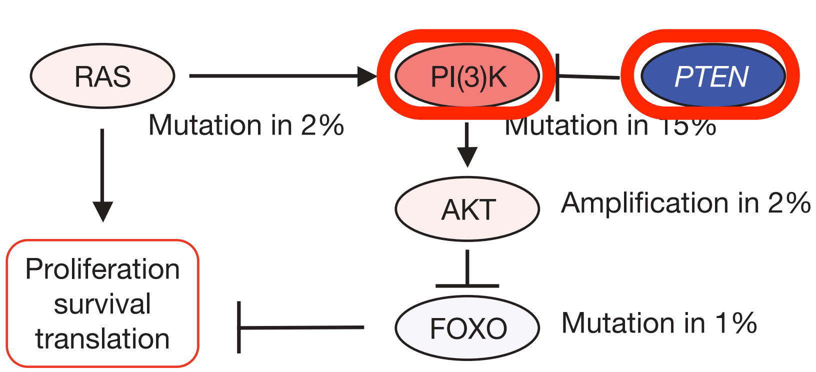
nature International weekly journal of science

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

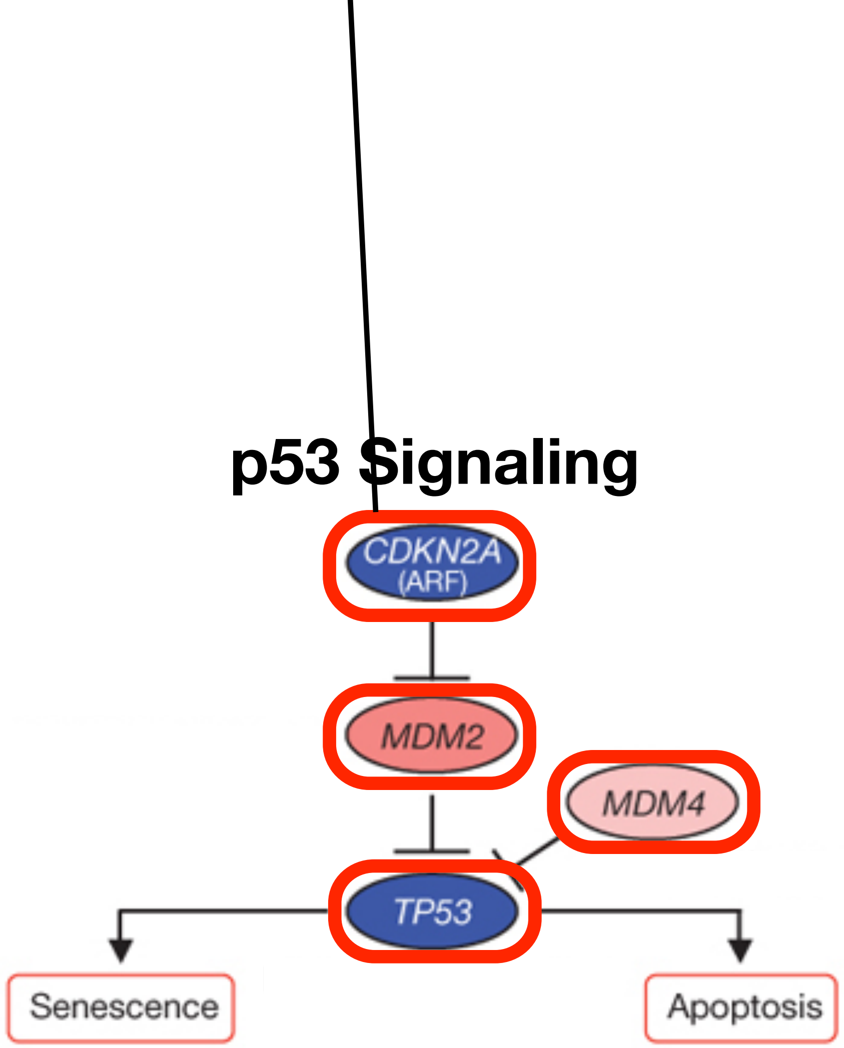
The Cancer Genome Atlas Research Network*

Figure 5

RTK/RAS/PI(3)K signaling



p53 Signaling



*TCGA (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* ***455****:1061-1088.*

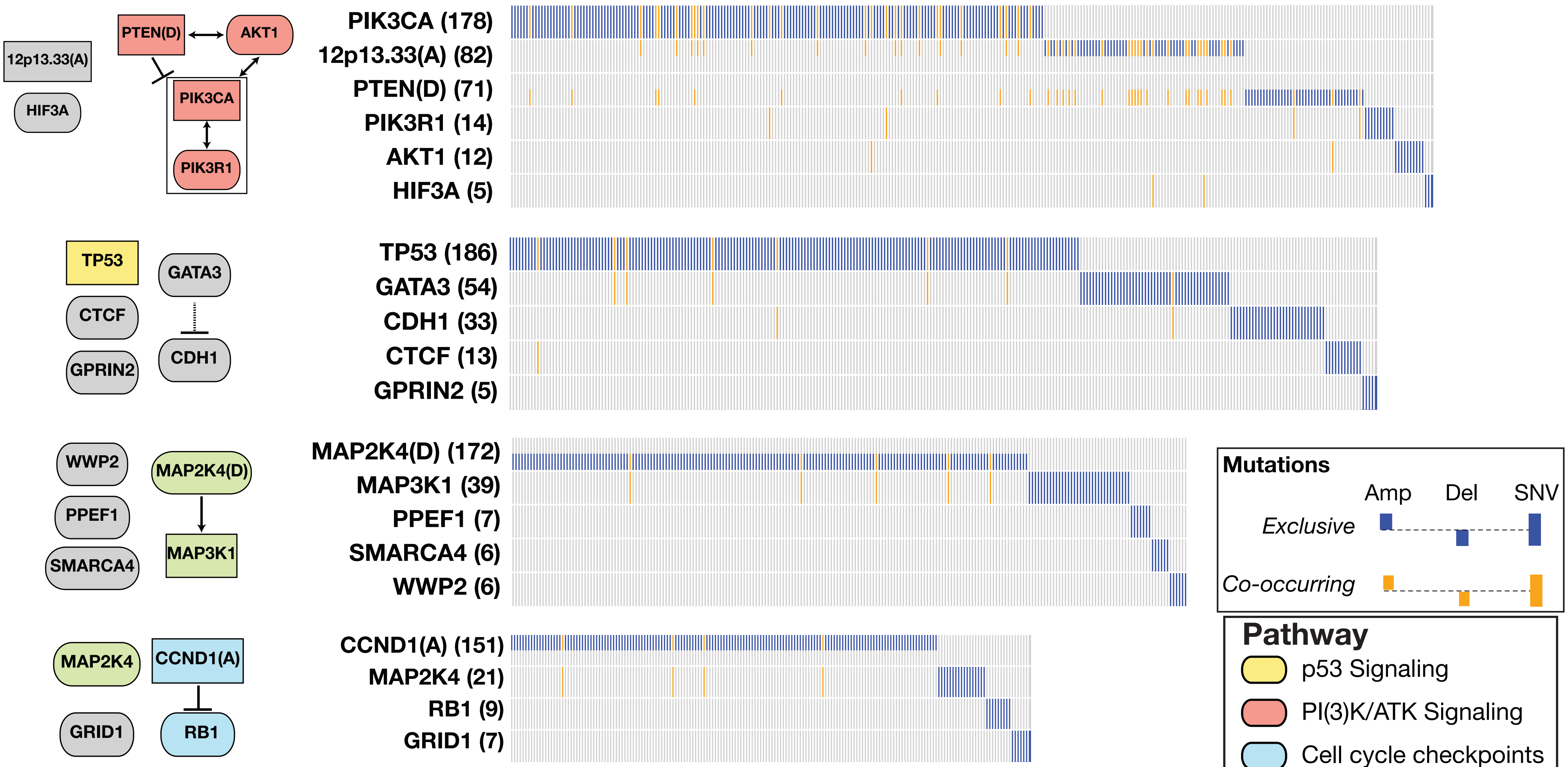
Results: TCGA breast cancer

$$W'_\alpha(\mathbf{M}) = \sum_{M \in \mathbf{M}'} \Gamma(M) - \alpha\omega(M)$$

higher $\alpha \Rightarrow$ more exclusivity

Multi-Dendrix (α=2.5)

**Mutation data:* 375 genes (events) in 507 patients



Pathway

- 🟨 p53 Signaling
- 🟥 PI(3)K/ATK Signaling
- 🟦 Cell cycle checkpoints
- 🟩 p38-JNK1
- ⬜ Other

Direct interactions test: $P < 0.001$

*TCGA. (2012) Comprehensive molecular portraits of human breast tumors. *Nature*, **490**:61-70.

Contributions

A new algorithm, **Multi-Dendrix**, for identifying driver pathways *de novo:*

1. Outperforms previous methods on simulated data in speed and accuracy

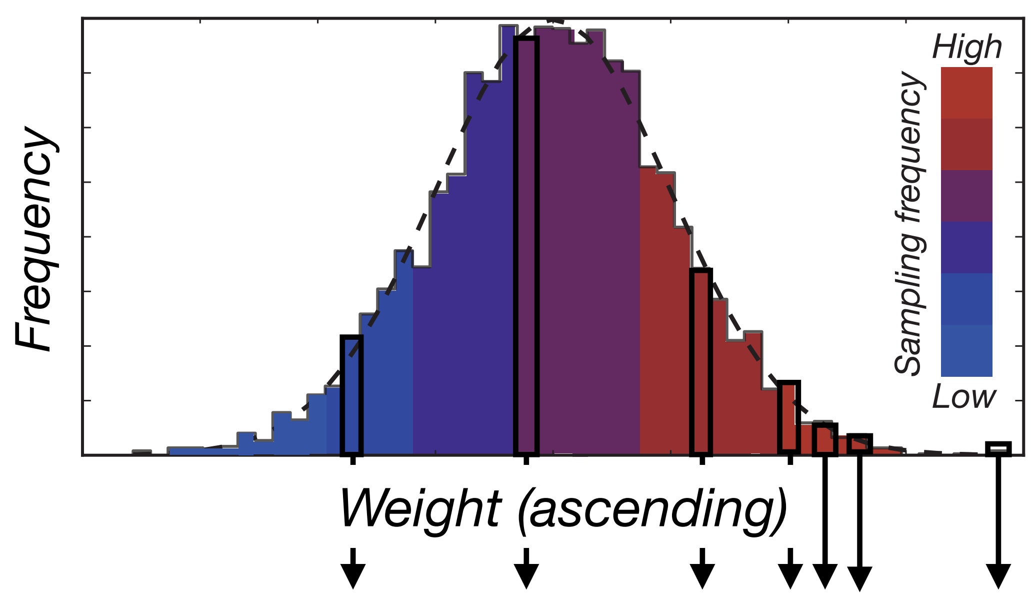
2. Identifies gene sets that overlap known pathways in TCGA datasets

3. Ongoing work extending Multi-Dendrix to large datasets and overlapping pathways

Multi-Dendrix MCMC

Sample collections of gene sets in proportion to their weight

Distribution of collections of gene sets



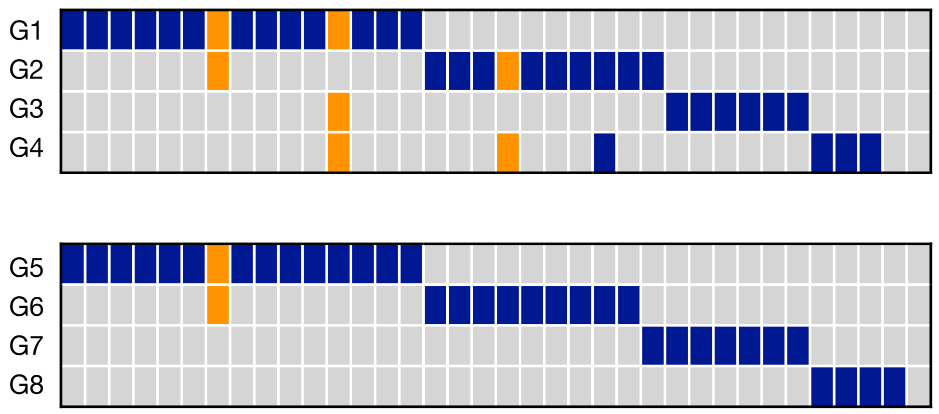
- Finds a distribution — optimal and suboptimal — of solutions

- Newest version on GitHub, requirements are all open-source

Sampling table

Geneset	Weight	Sampling Frequency
G1, G2, G3, G4 G5, G6, G7, G8	200	60
G9, G10, G11, G12 G13, G14, G15, G16	170	30
G1, G2, G3, G4 G5, G7, G15, G16	130	9
G17, G18, G19, G20 G13, G14, G15, G16	80	1

Top collection



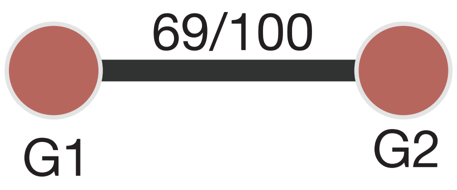
Weight: 200

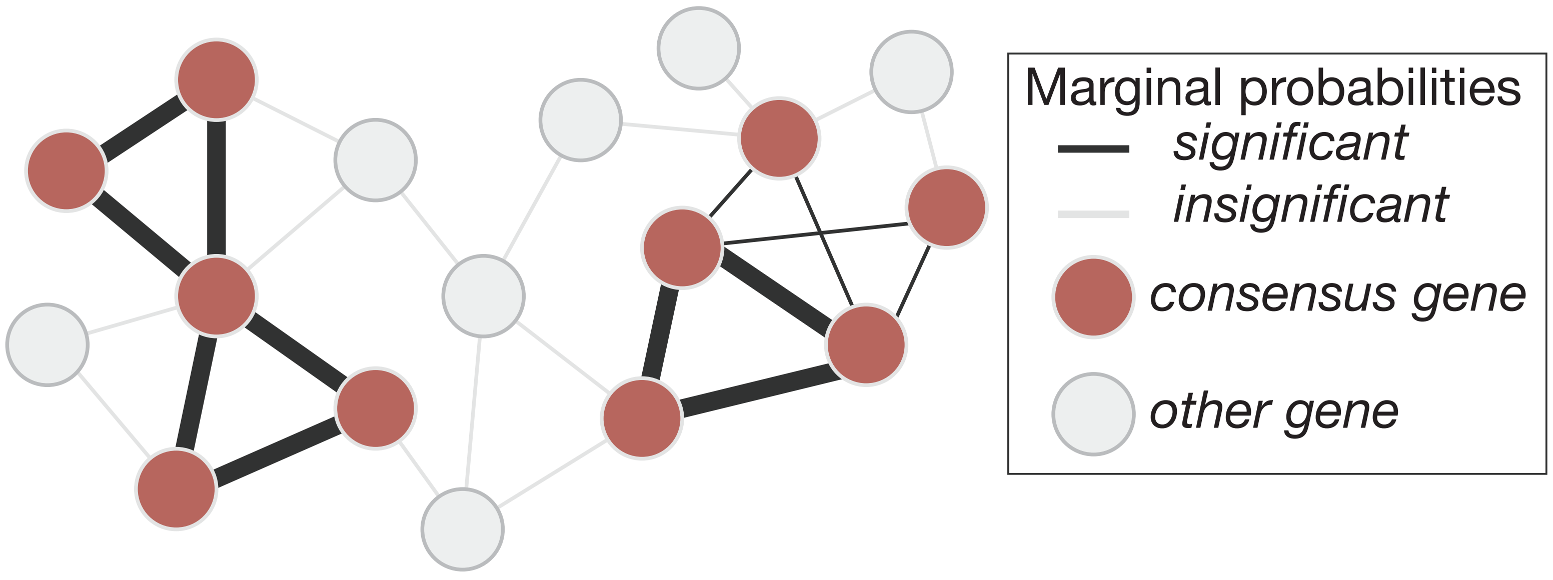
http://github.com/raphael-group/multi-dendrix

Marginal probability graph

- Marginal probability graph defines consensus subnetworks
- Edges (*u*, *v*) are weighted by how often gene *u* is sampled in the same gene set as gene *v*

Geneset	Weight	Sampling Frequency
G1, **G2**, G3, G4 G5, G6, G7, G8	200	60
G1, **G2**, G3, G4 G5, G7, G15, G16	130	9





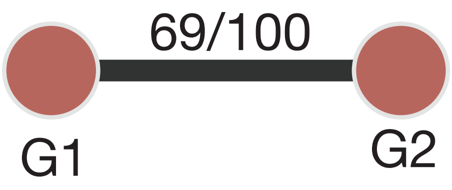
Complete, weighted marginal probability graph

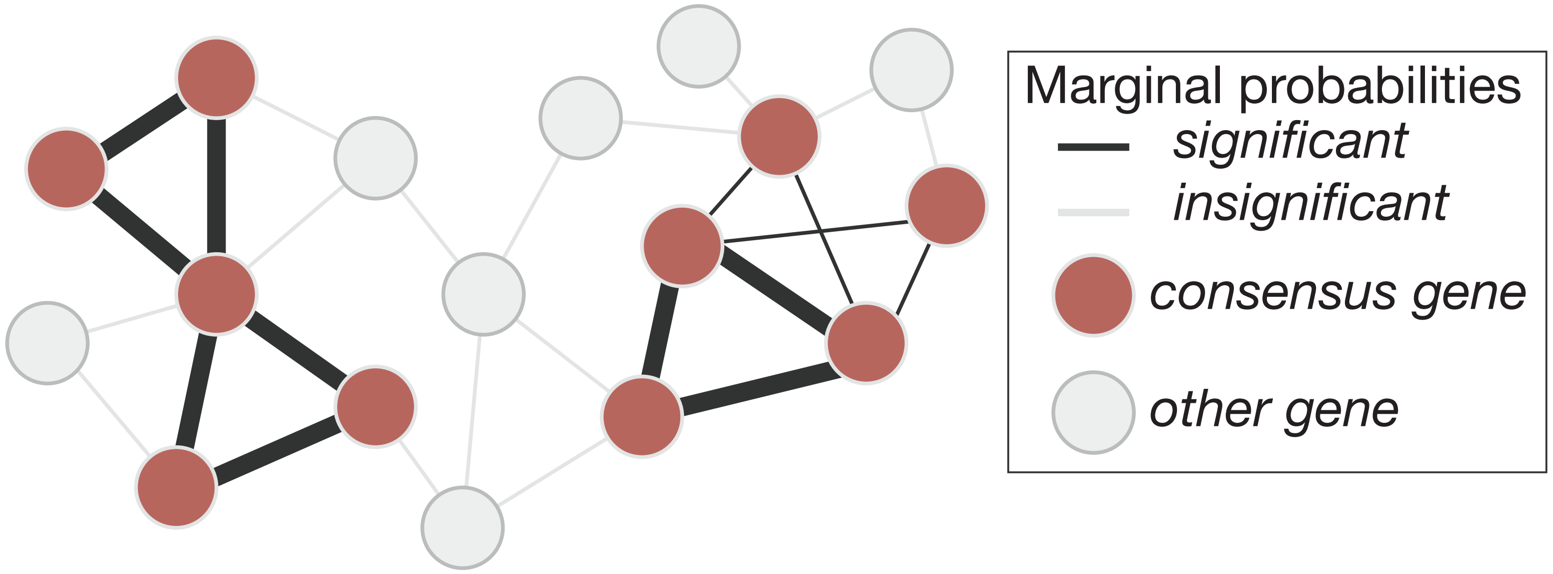
→ unconstrained size and number of gene sets
→ gene sets can overlap

Marginal probability graph

- Marginal probability graph defines consensus subnetworks
- Edges (*u*, *v*) are weighted by how often gene *u* is sampled in the same gene set as gene *v*

Geneset	Weight	Sampling Frequency
G1, **G2**, G3, G4 G5, G6, G7, G8	200	60
G1, **G2**, G3, G4 G5, G7, G15, G16	130	9





Complete, weighted marginal probability graph

Hsin-Ta Wu

→ unconstrained size and number of gene sets
→ gene sets can overlap

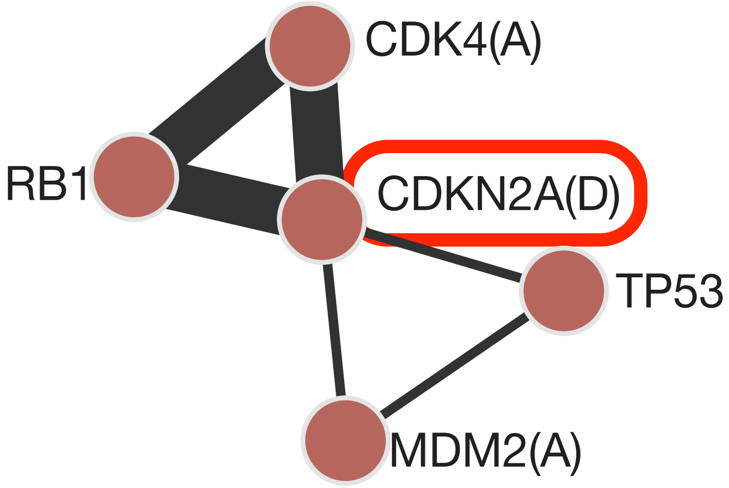
Leiserson, Wu *et al*. Dendrix++. (In preparation).

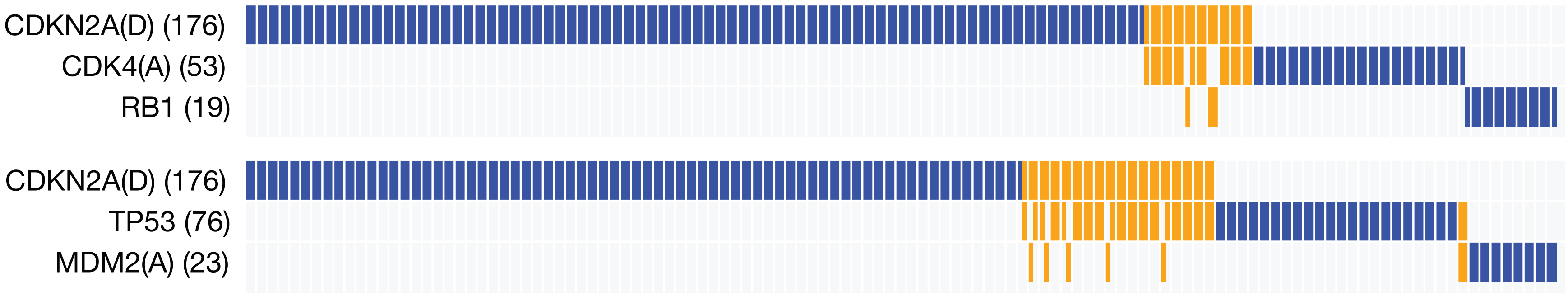
Dendrix++ results: GBM

nature International weekly journal of science

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*





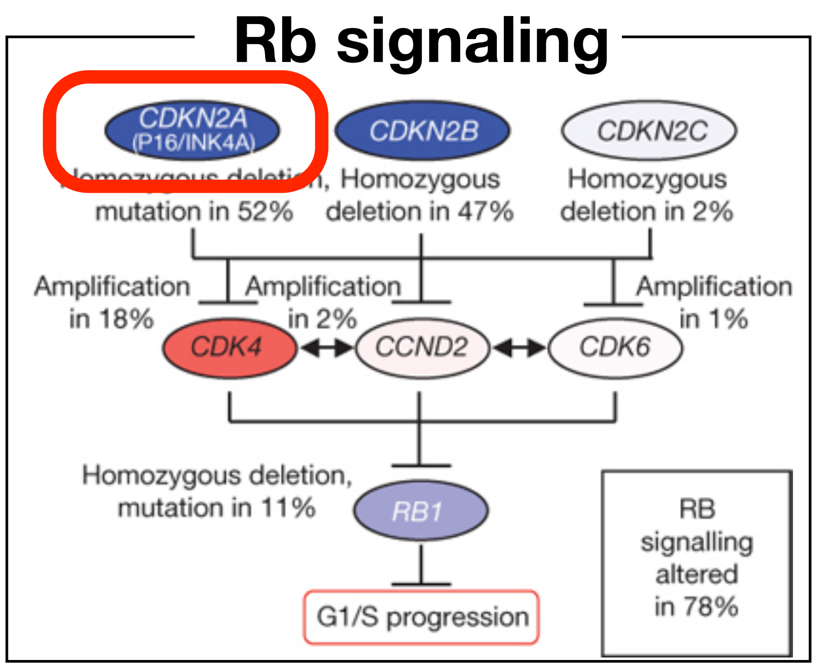
Leiserson, Wu *et al*. Dendrix++. (In preparation).

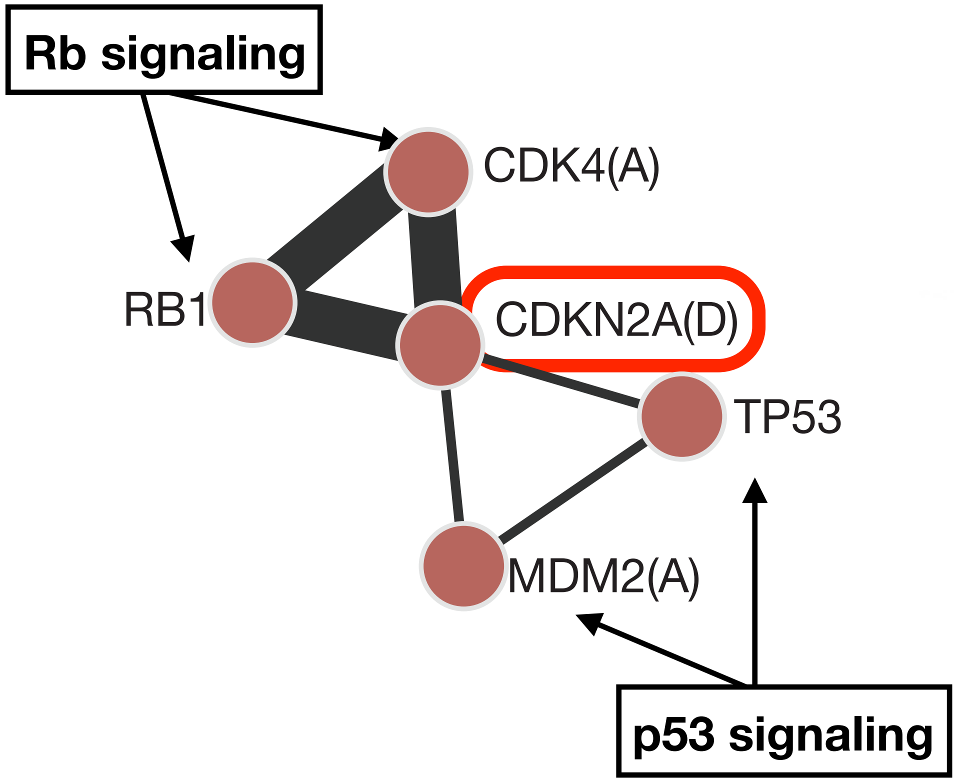
Dendrix++ results: GBM

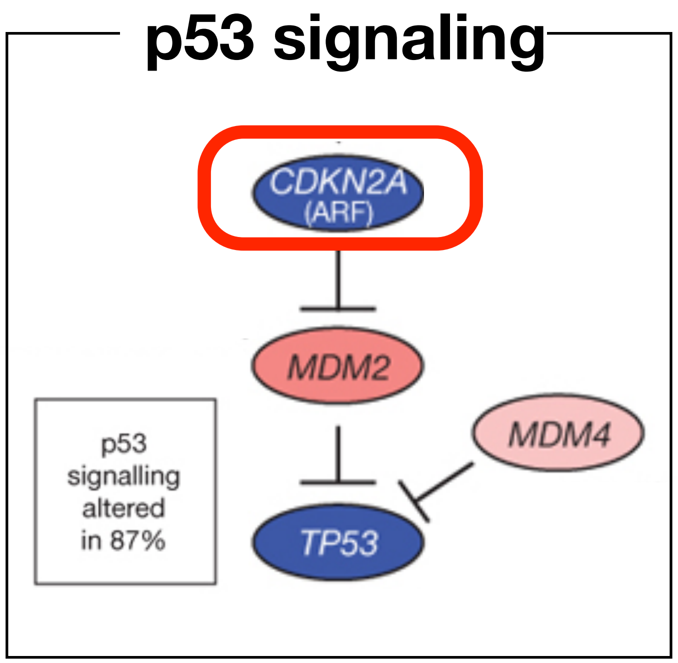
nature International weekly journal of science

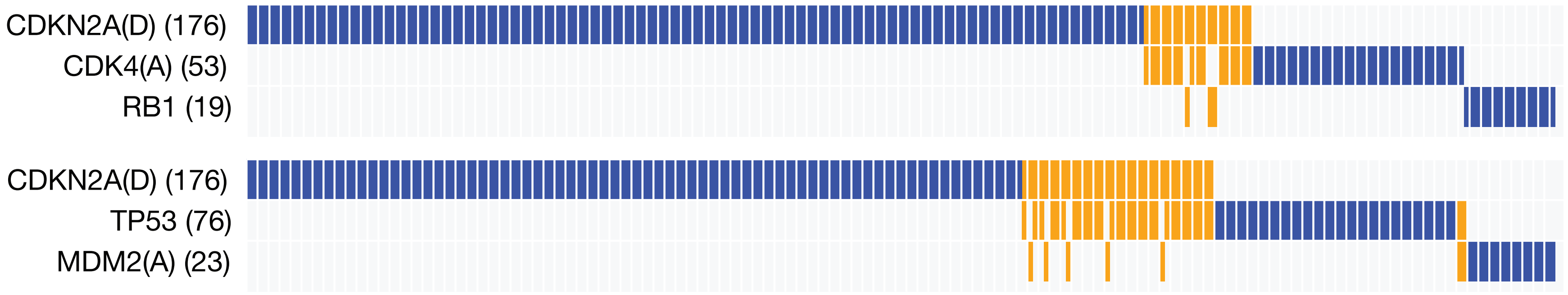
Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*


Rb signaling
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
G1/S progression
RB signalling altered in 78%







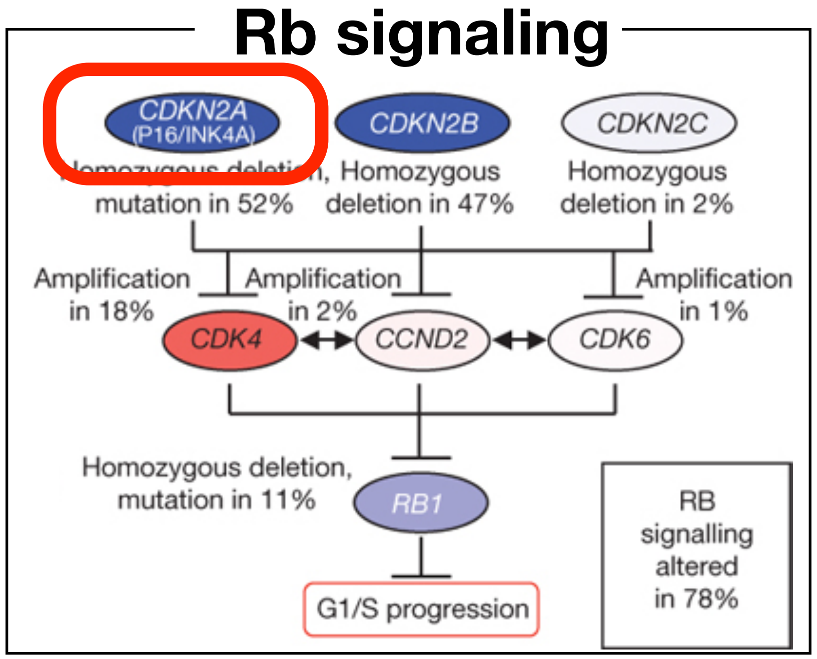
Leiserson, Wu *et al*. Dendrix++. (In preparation).

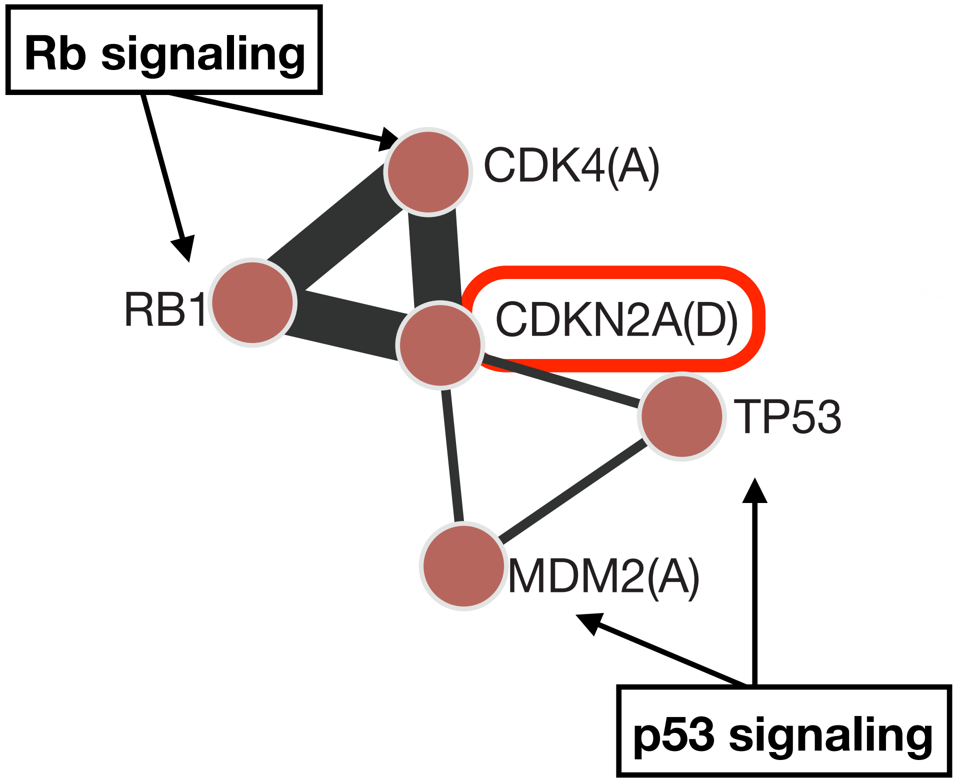
Dendrix++ results: GBM

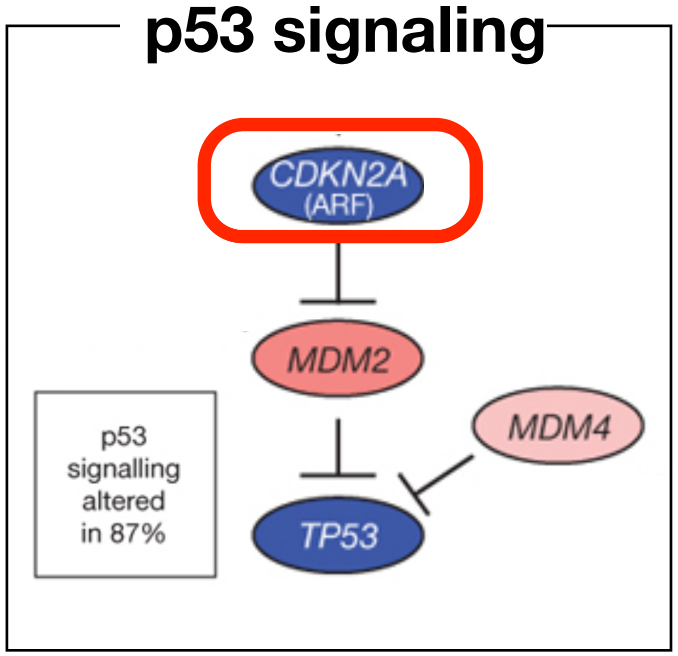
nature International weekly journal of science

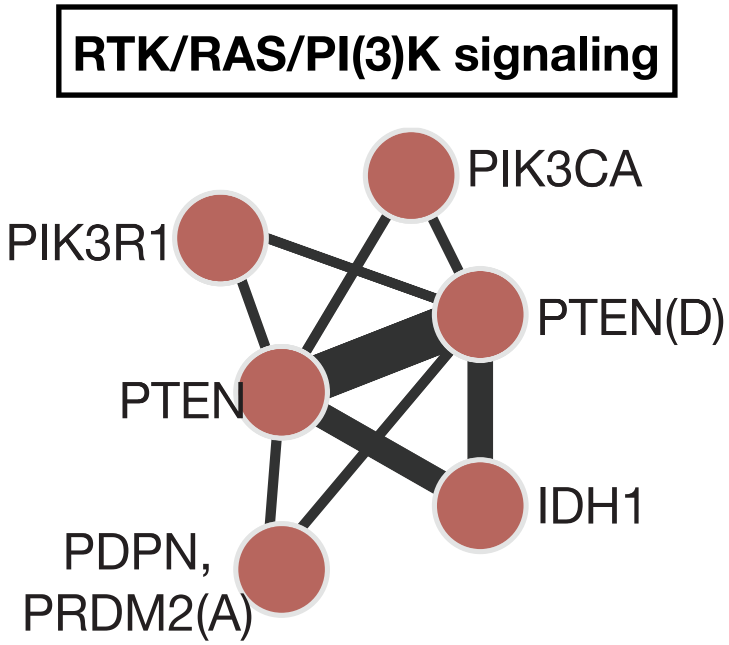
Comprehensive genomic characterization defines human glioblastoma genes and core pathways

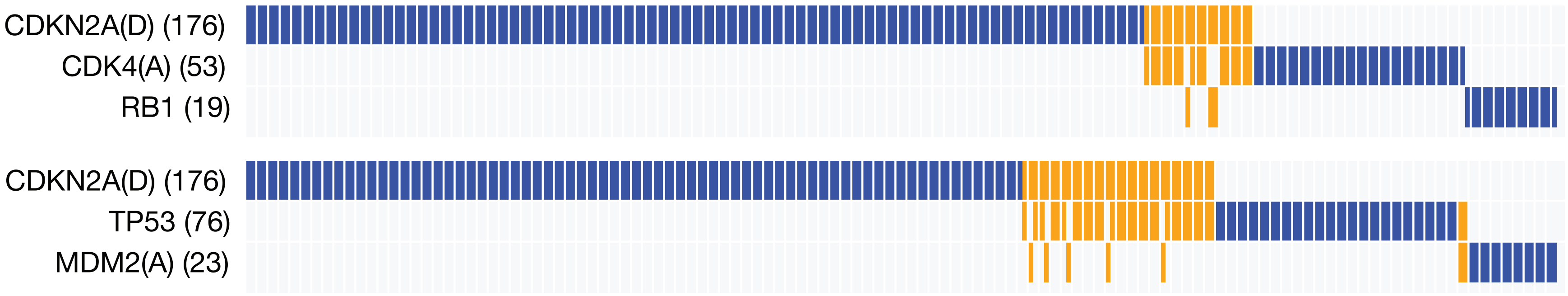
The Cancer Genome Atlas Research Network*


Rb signaling
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
G1/S progression
RB signalling altered in 78%









Leiserson, Wu *et al*. Dendrix++. (In preparation).

Summary

- Multi-Dendrix: Fast, exact ILP for identifying collections of gene sets with exclusive mutations

- Identifies modules that overlap multiple cancer pathways in glioblastoma and breast cancer

- Dendrix++: New algorithm that can identify more complex pathways (in preparation)

Acknowledgements

Research group

Ben Raphael

Jason Dobson

Iman Hajirasouliha

Fabio Vandin

Layla Oesper

Hsin-Ta Wu

BROWN

Collaborators

Roded Sharan

Dima Blokh

אוניברסיטת תל-אביב
TEL AVIV UNIVERSITY

Funding + Data

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

The Cancer Genome Atlas

NATIONAL INSTITUTES OF HEALTH