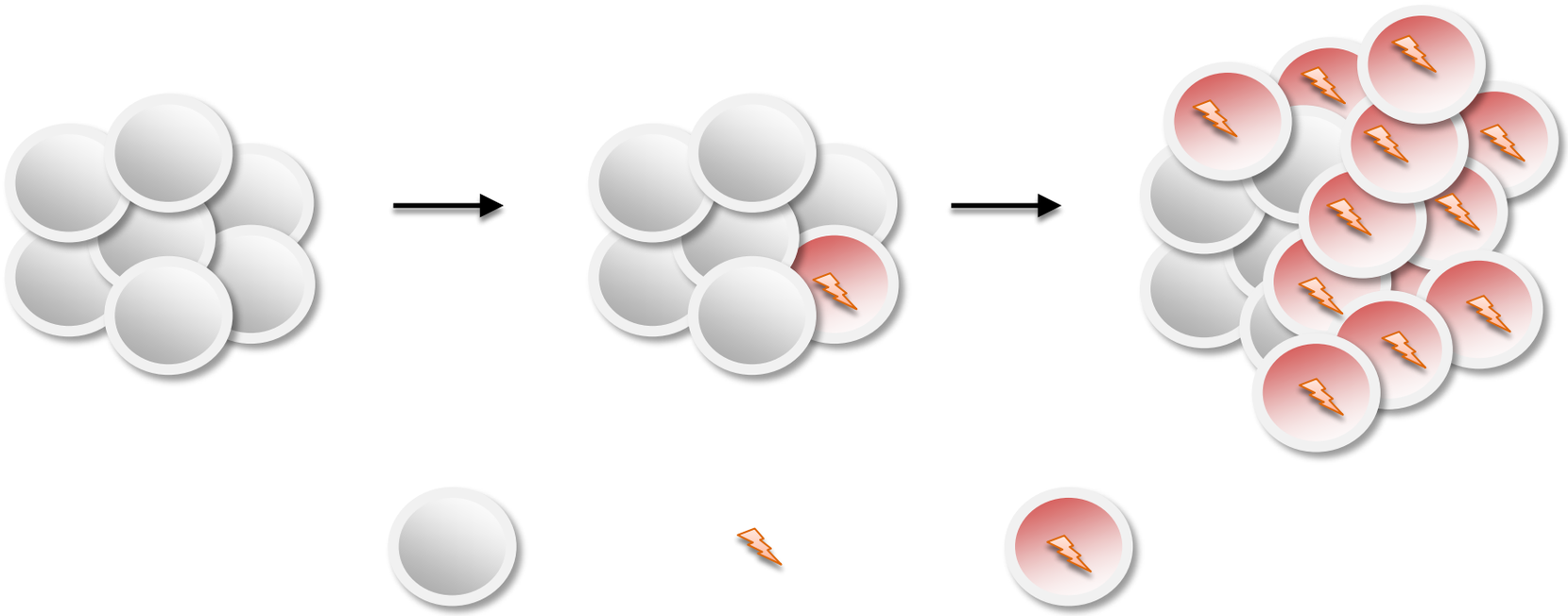
CoMEt:
A Statistical Approach to Identify Combinations of Mutually Exclusive Alterations in Cancer

Hsin-Ta Wu*,
Max Leiserson*, Fabio Vandin, Ben Raphael

TCGA 4th Annual Scientific Symposium

May 11th, 2015

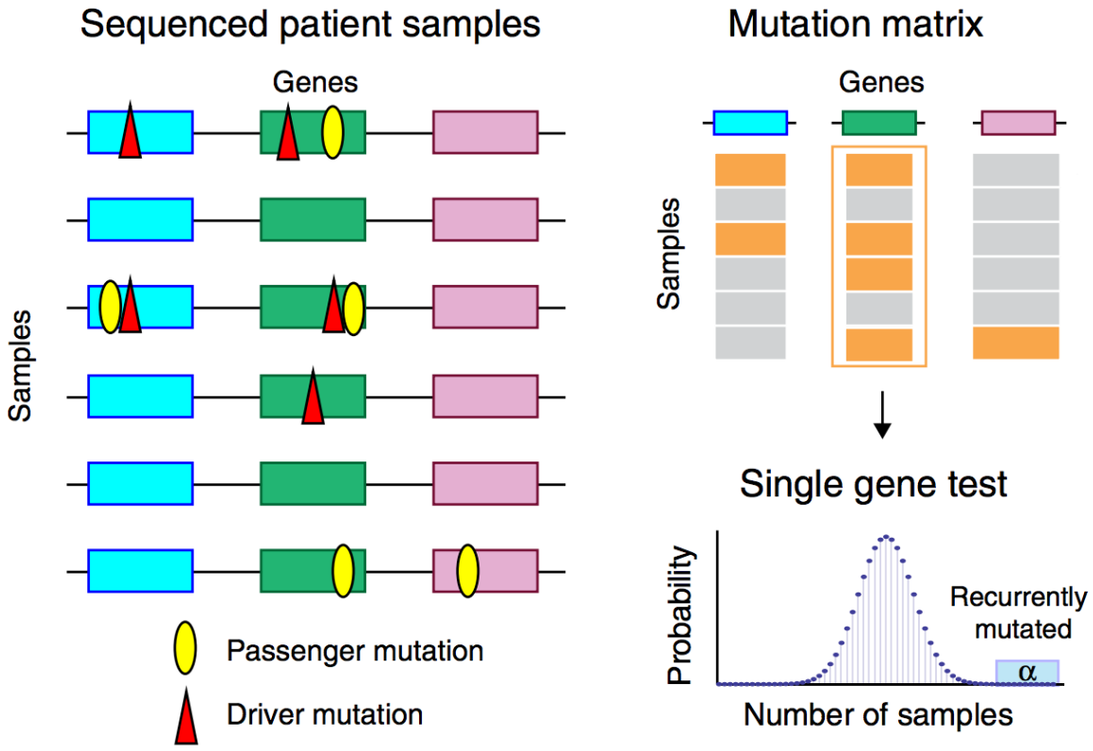
Distinguish driver mutations from passenger mutations



typical tumor: ~10 driver mutations, 100’s~1000’s of passenger mutations

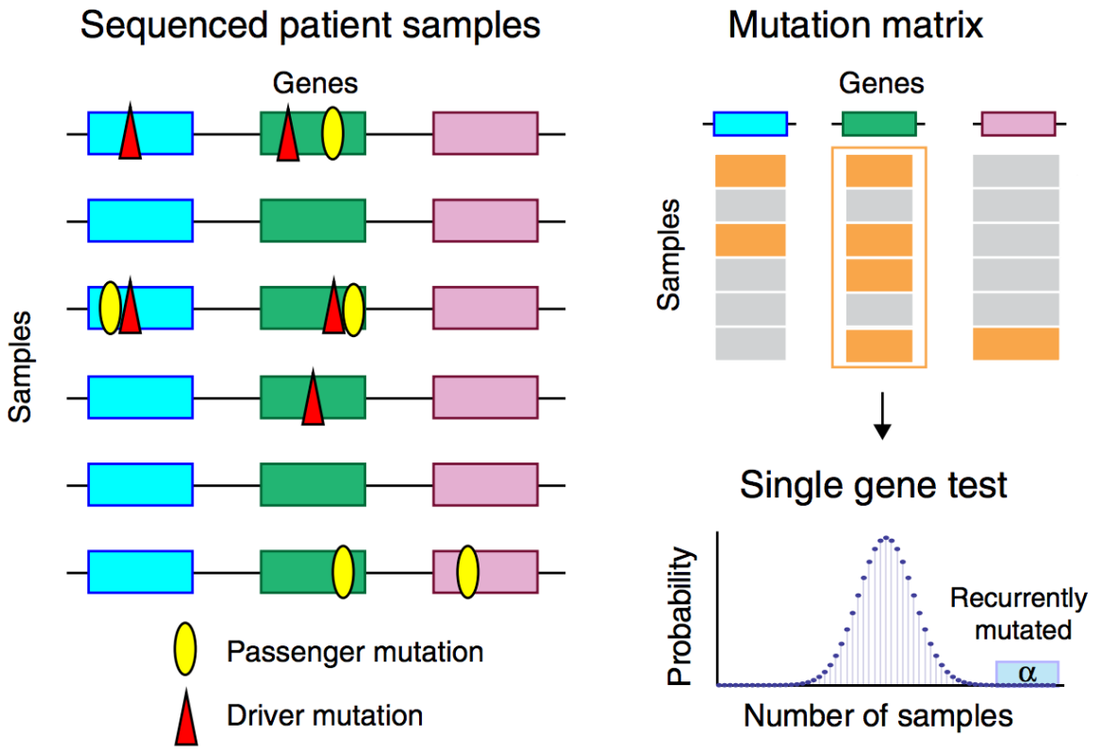
How to distinguish driver from passenger mutations?

Finding recurrence by comparing mutations across tumors

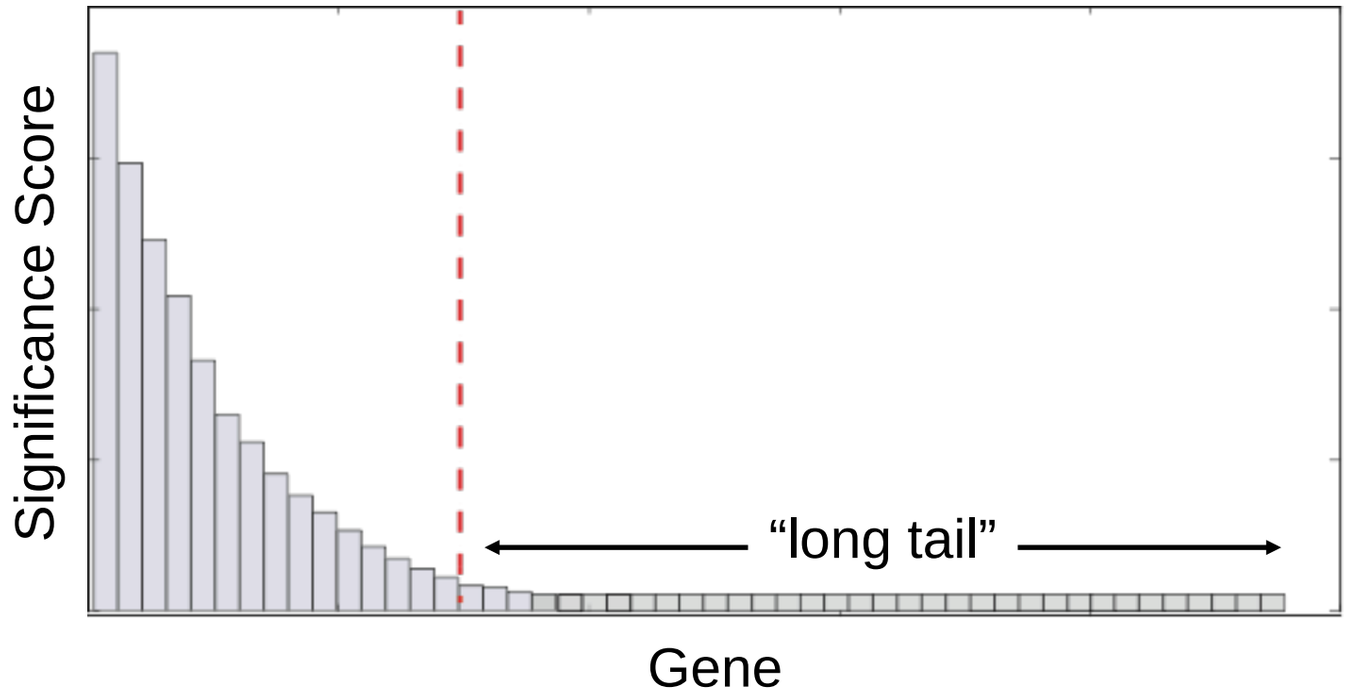


Raphael et al. *Genome Medicine.* 2014

Finding recurrence by comparing mutations across tumors



Raphael et al. *Genome Medicine.* 2014

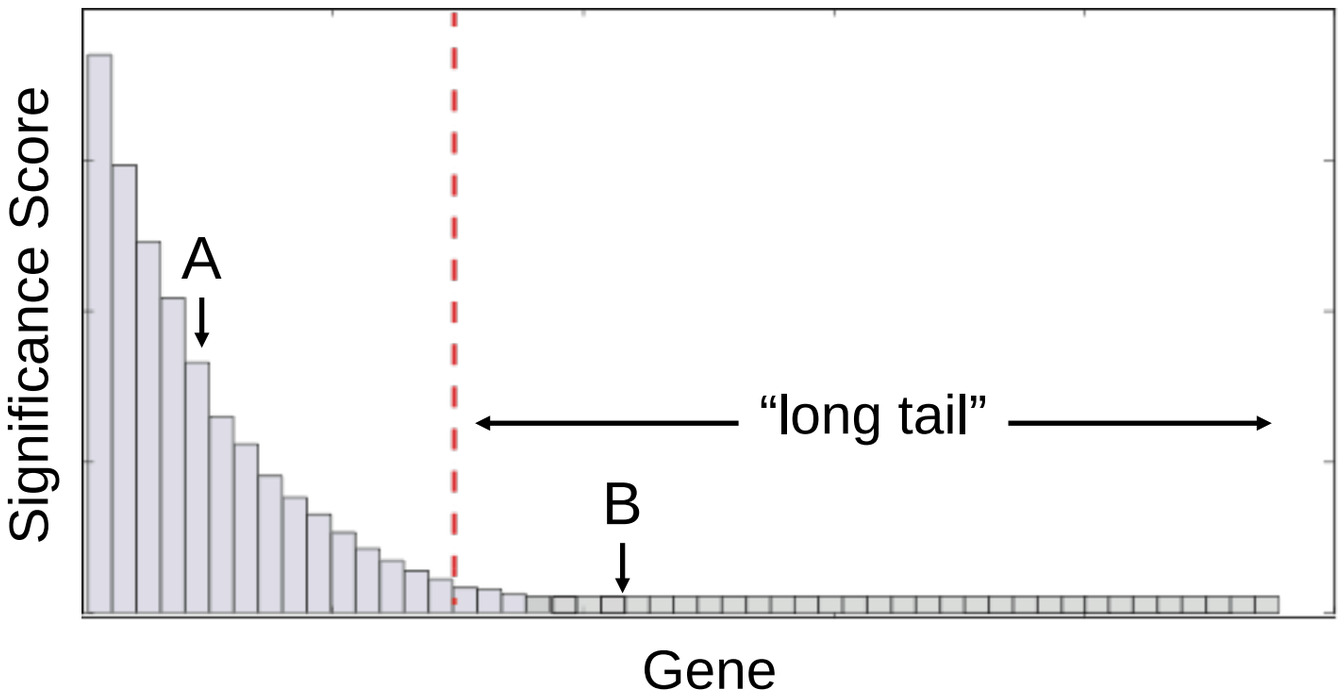


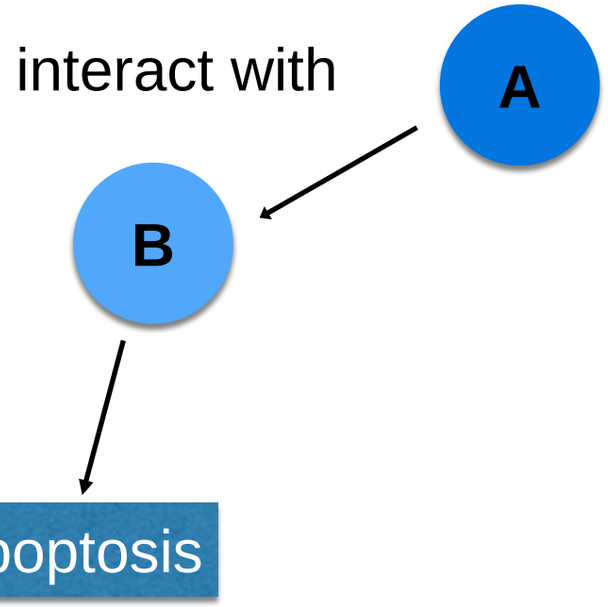
Significance Score
Mutations weighted by:

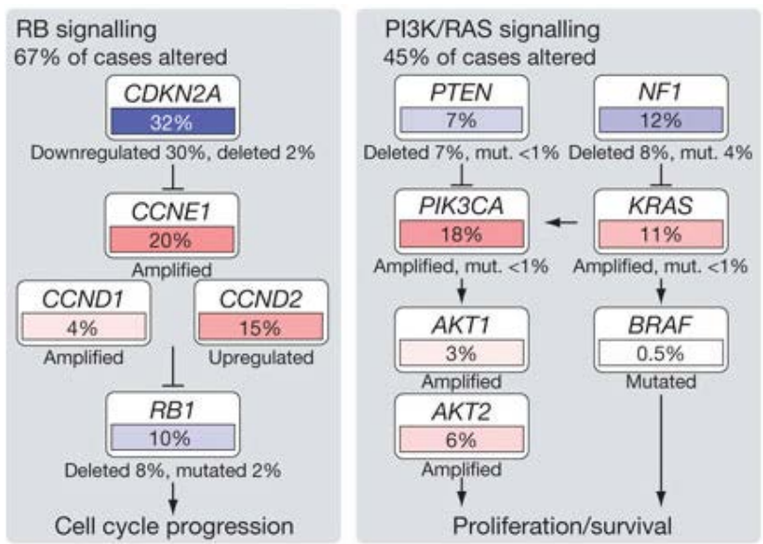
- Gene length
- Mutation context
- Expression level
- Replication timing
- …

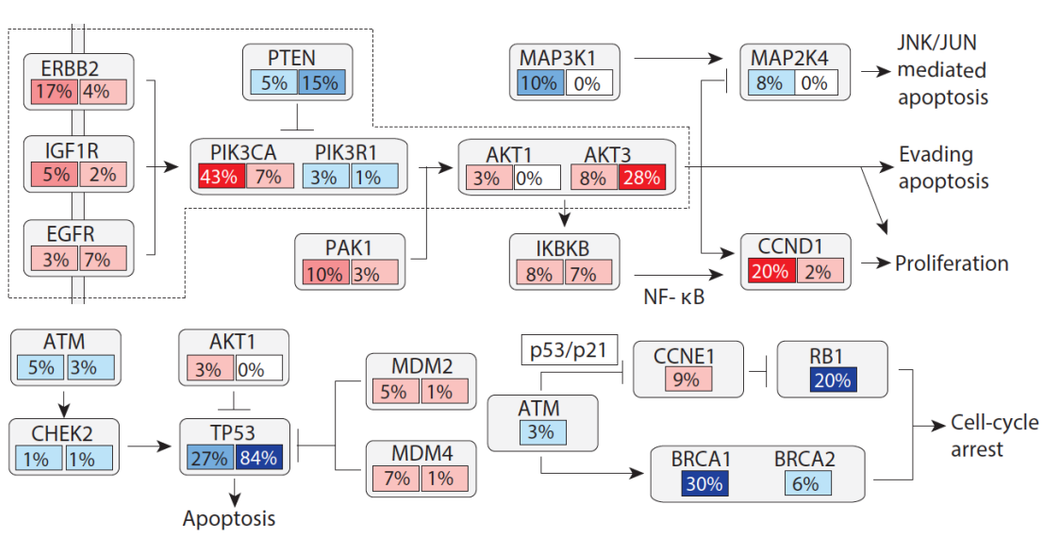
Lawrence, et al. 2013
Tamborero, et al. 2013
Kandoth, et al. 2014

Driver mutations target pathways rather than individual genes



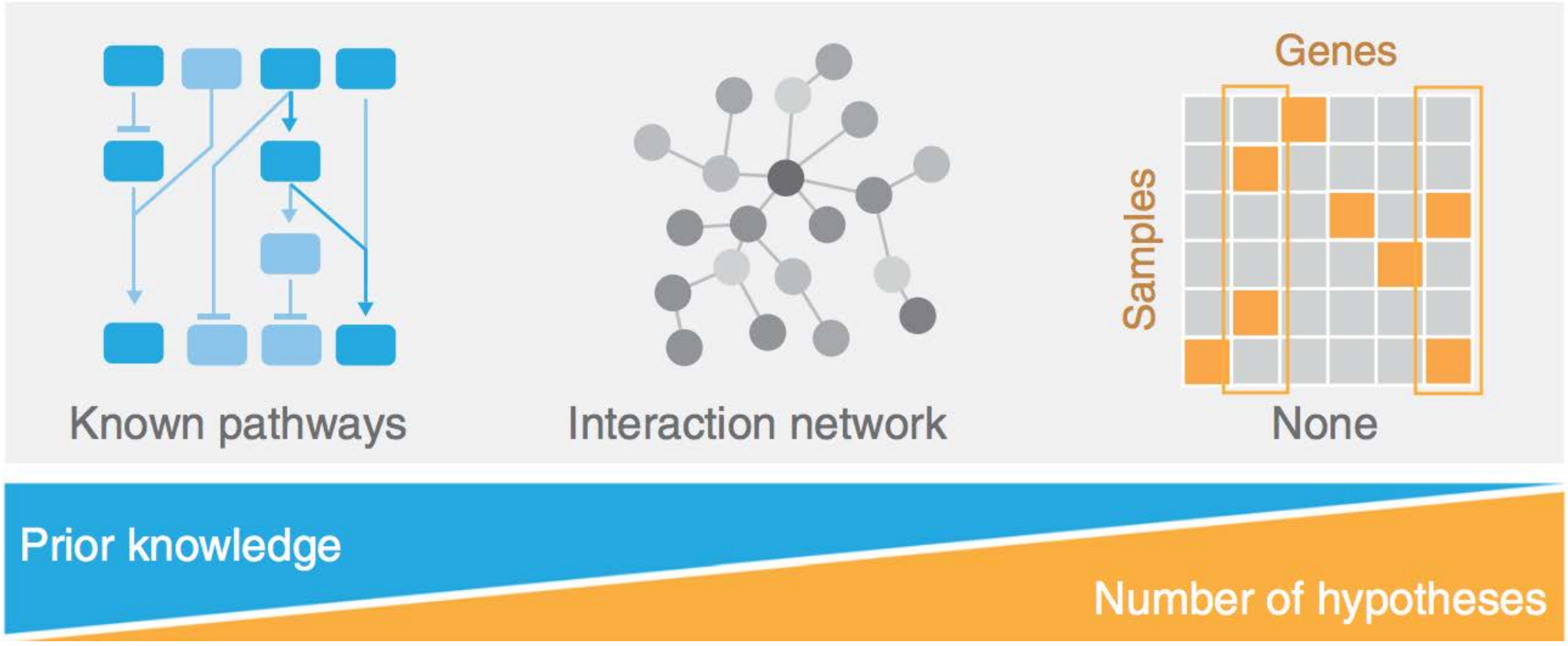






TCGA Ovarian (2011), Breast (2012)

Combinations of mutations

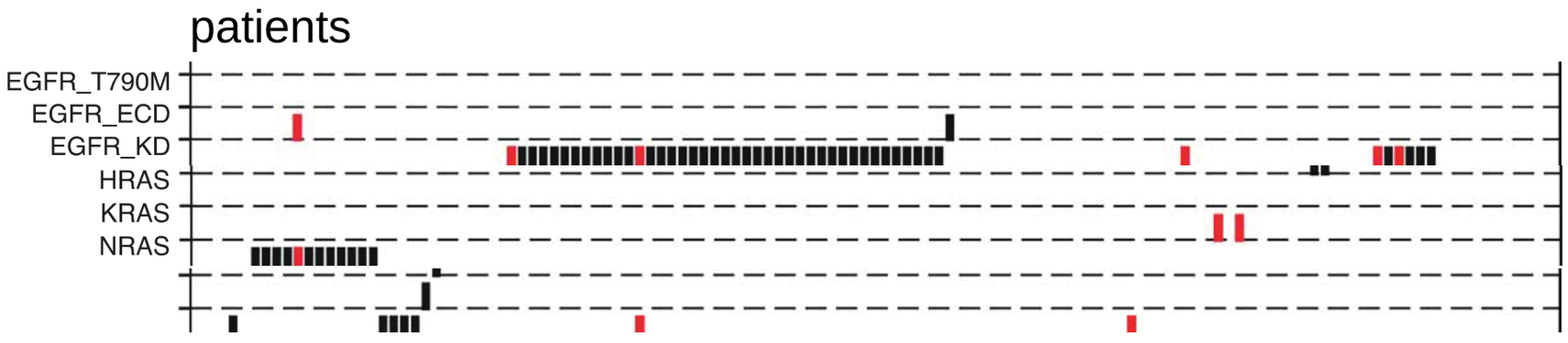


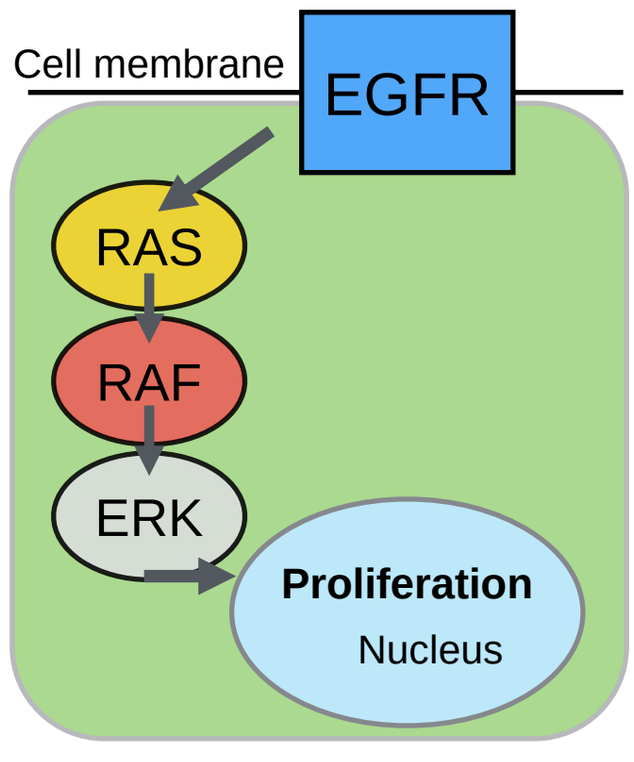
Pathways and interaction networks are incomplete
→ Difficult to detect novel pathways

Raphael et al. *Genome Medicine.* 2014

Cancer pathways harbor mutually exclusive mutations

Exclusivity: most patients in the cancer pathway have only one mutation.

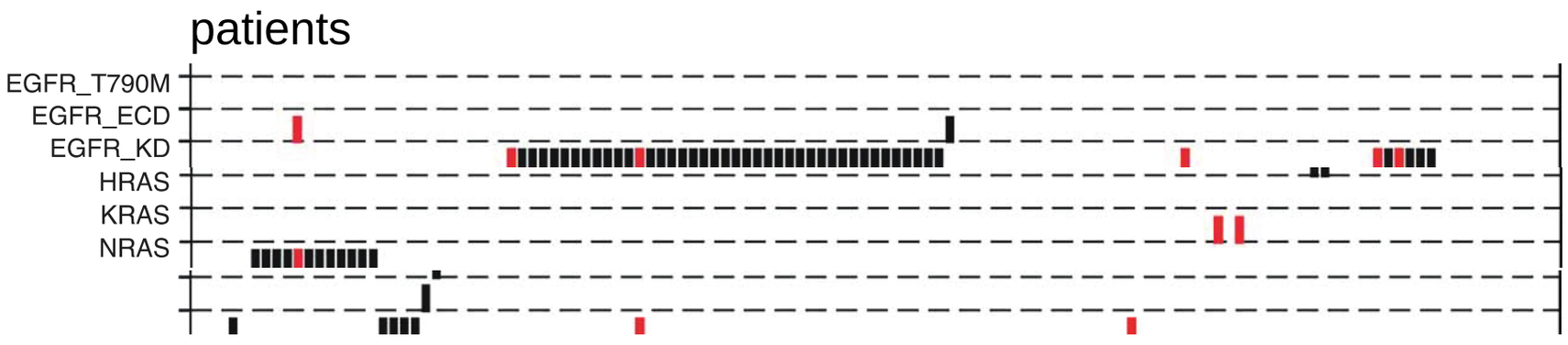




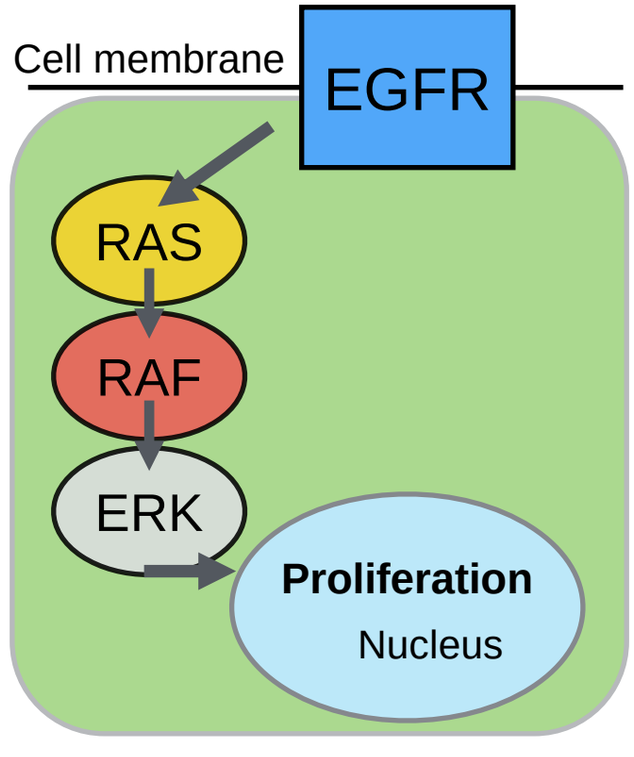
Thomas, et al. 2007

Cancer pathways harbor mutually exclusive mutations

Exclusivity: most patients in the cancer pathway have only one mutation.



Dendrix	Vandin et al. ***RECOMB****.* 2011
RME	Miller et al. ***BMC Med Genomics****.* 2011
MEMo	Ciriello et al. ***Genome Res****.* 2012
Dendrix++	TCGA. ***NEJM****.* 2013
Multi-Dendrix	Leiserson et al. ***PLoS Comp. Bio***. 2013
muex	Szczurek et al. ***RECOMB****.* 2014
mutex	Babur et al. ***Genome Biology.*** *2015*



Thomas, et al. 2007

How do current methods score exclusivity ?

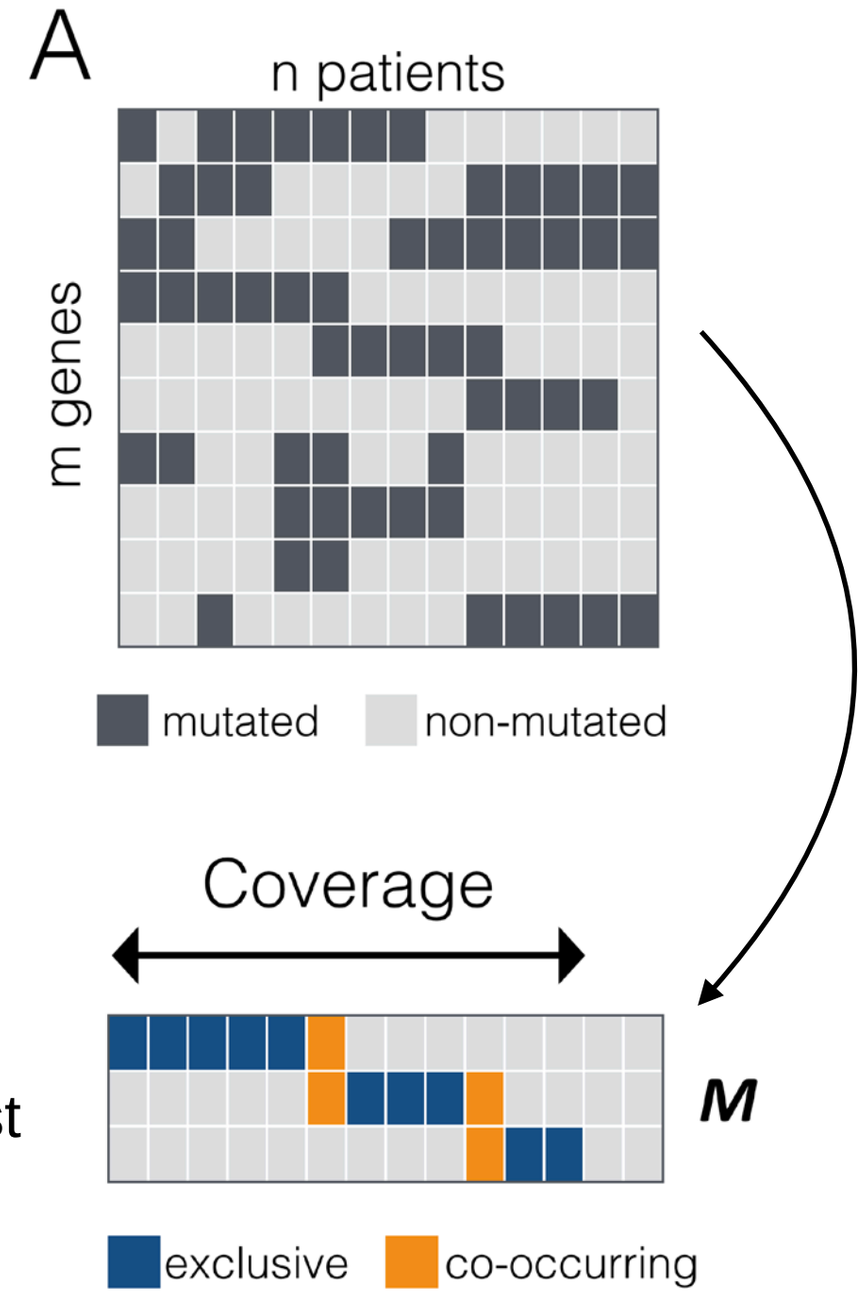
Vandin et al. 2011; Miller et al. 2011;
Ciriello et al. 2012; Szczurek et al. 2014

Given:

Binary mutation matrix **A**

Find: A combination ***M*** of genes

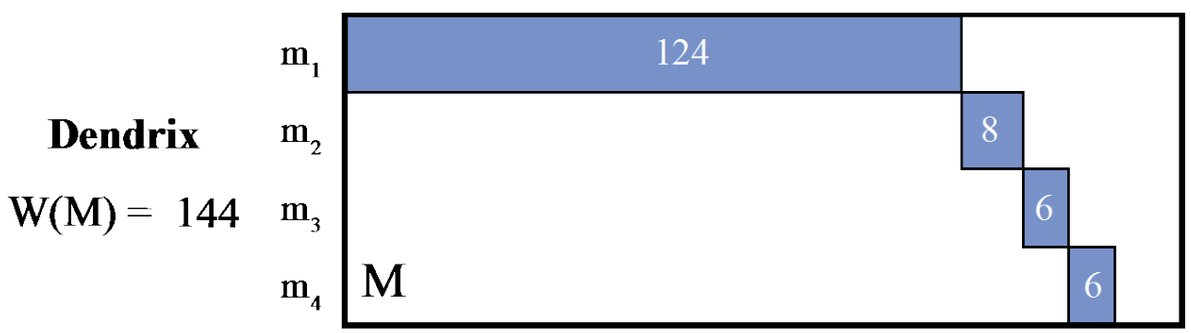
- **Dendrix*, Multi-Dendrix**
- **RME**
  - ***Exclusivity*** and ***Coverage.***
- **muex**
  - Generative model of ***Exclusivity.***
- **MEMo+**
  - Only consider ***M*** in **interaction network**.
  - Permutation test with ***Coverage*** as the test statistic.

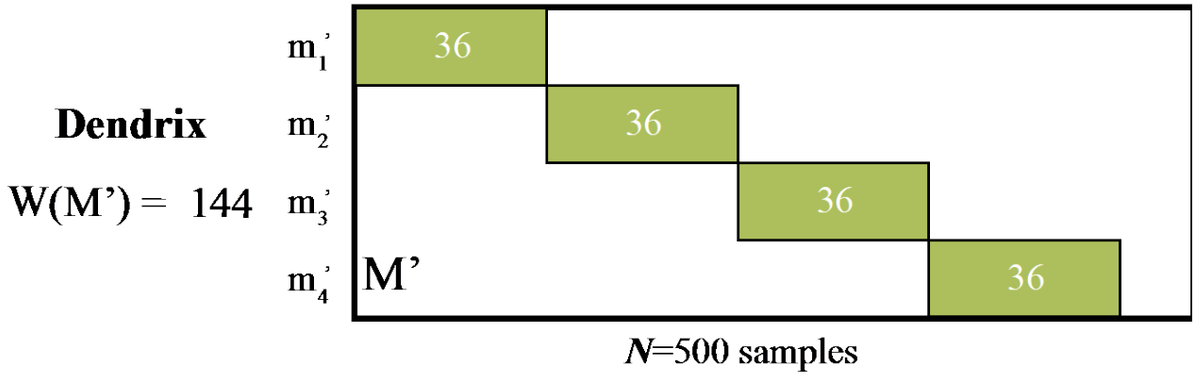


* Kandoth, et al. *Nature* (2013). Mutational landscape and significance across 12 major cancer types.

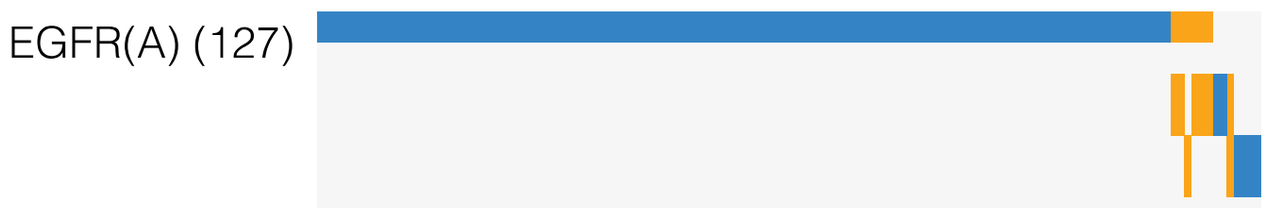
+ TCGA BRCA COAD UCEC studies

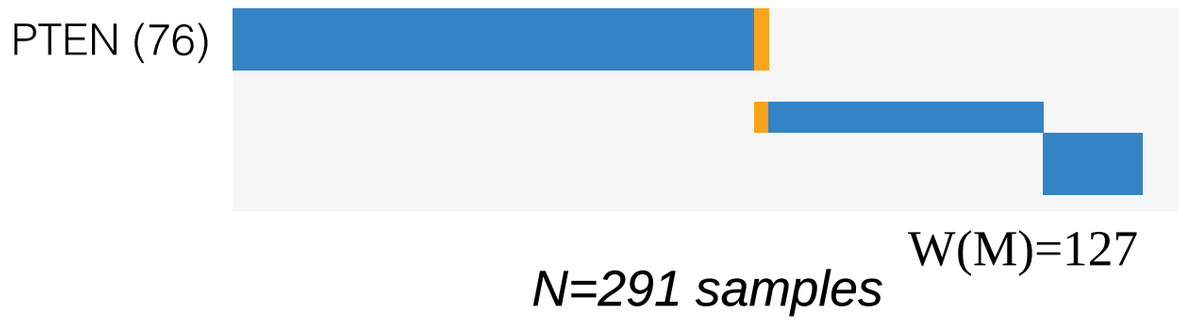
Dendrix favors highly mutated genes





Glioblastoma (GBM) mutation dataset

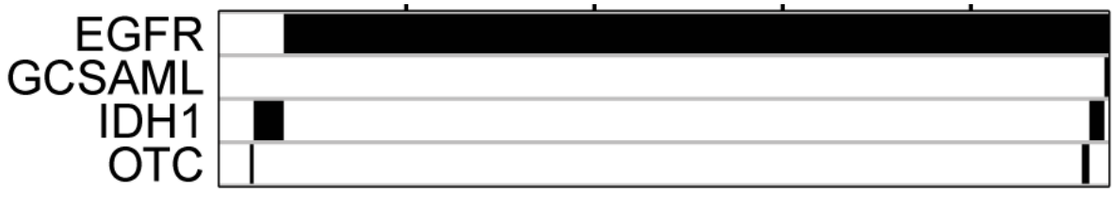




muex

Szczurek et al. 2014

the largest weight →

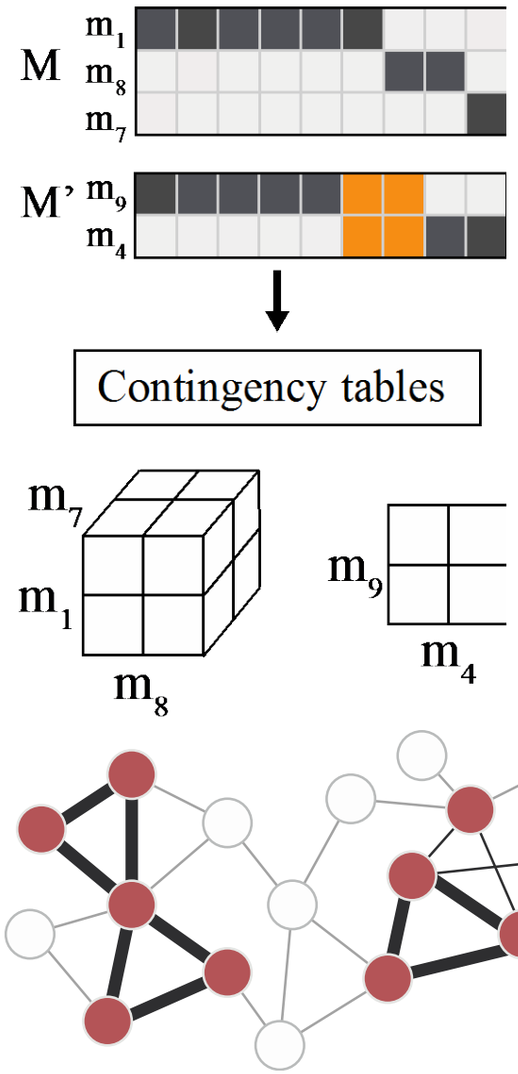


Genes with high mutation frequencies can dominate the mutual exclusivity signal.

Contributions

A new algorithm, **CoMEt**, for identifying driver pathways *de novo:*

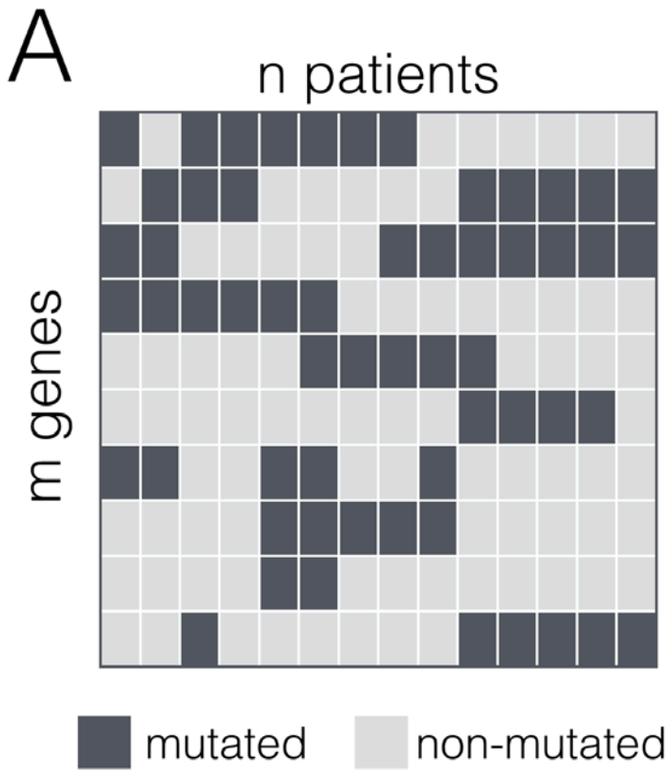
- ***Statistical*** score for exclusivity.
- Simultaneous analysis of ***multiple*** combinations.
- Summarize mutual exclusivity over high-scoring collections.
- Outperform other methods on simulated and real data.



Leiserson, Wu, et al. (2015)

CoMEt

Score a combination of mutations M

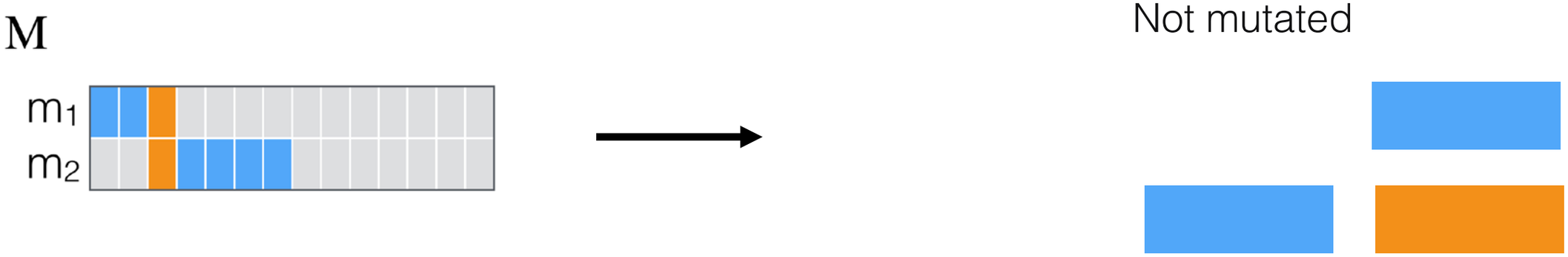


Given a binary mutation matrix **A**,

Find a combination **M** of ***k*** genes with significant mutually exclusive mutations, **conditional on the the number of mutations in each gene**

Score a combination of two genes

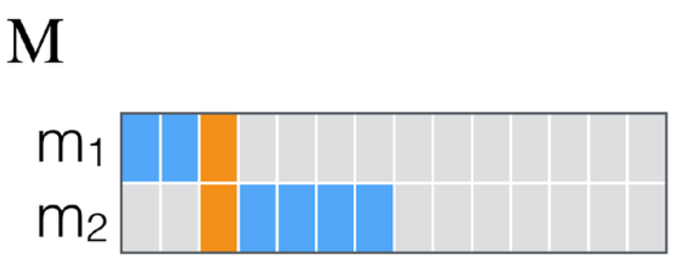
2x2 contingency Table $\mathbf{X}_M$

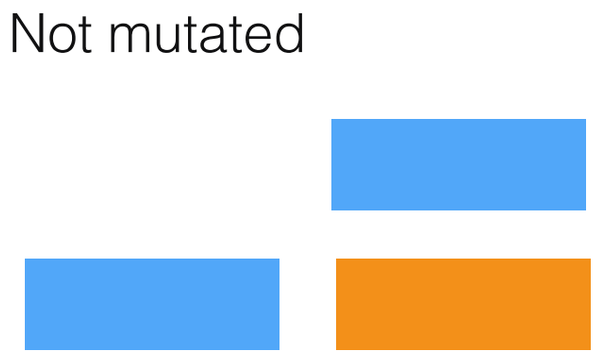


Compute significance of observed mutual exclusivity?

Score a combination of two genes

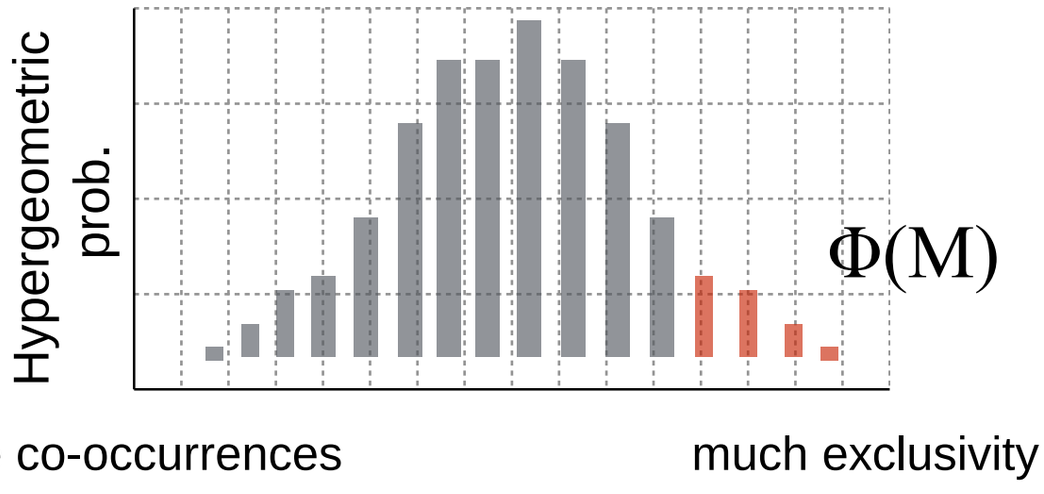
2x2 contingency Table $\mathbf{X}_M$





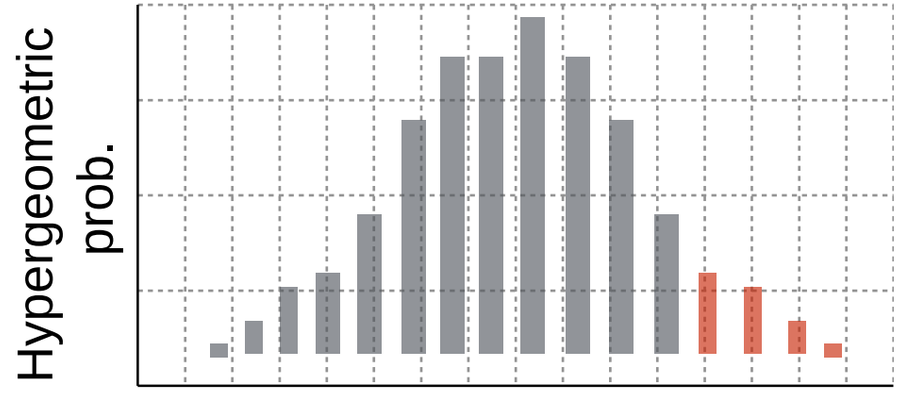
Compute significance of observed mutual exclusivity?

One-sided Fisher's exact test for independence



Score a combination of two genes

One-sided Fisher's exact test for independence



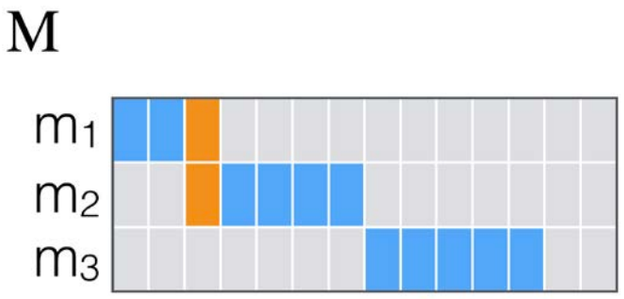
Yeang *et al*. ***FASEB J*** 2008

TCGA Acute Myeloid Leukemia. ***NEJM*** 2013

TCGA Papillary Thyroid Carcinoma. ***Cell*** 2013

Babur *et al.* ***Genome Biology*** 2015 - *mutex*

Score a combination of three genes
(k=3)



2x2x2 contingency table $\mathbf{X}_M$

One-sided Fisher's exact test for independence?

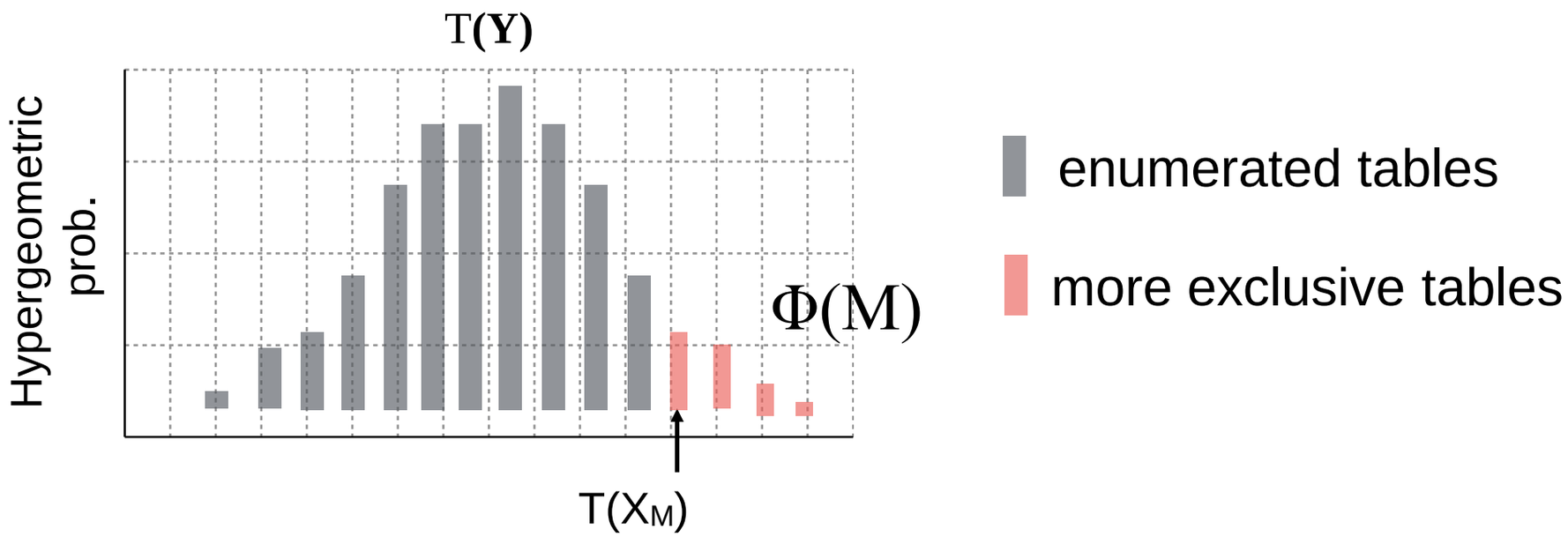
Degrees of freedom : 2^k-k-1

Score a combination of three genes
(k=3)

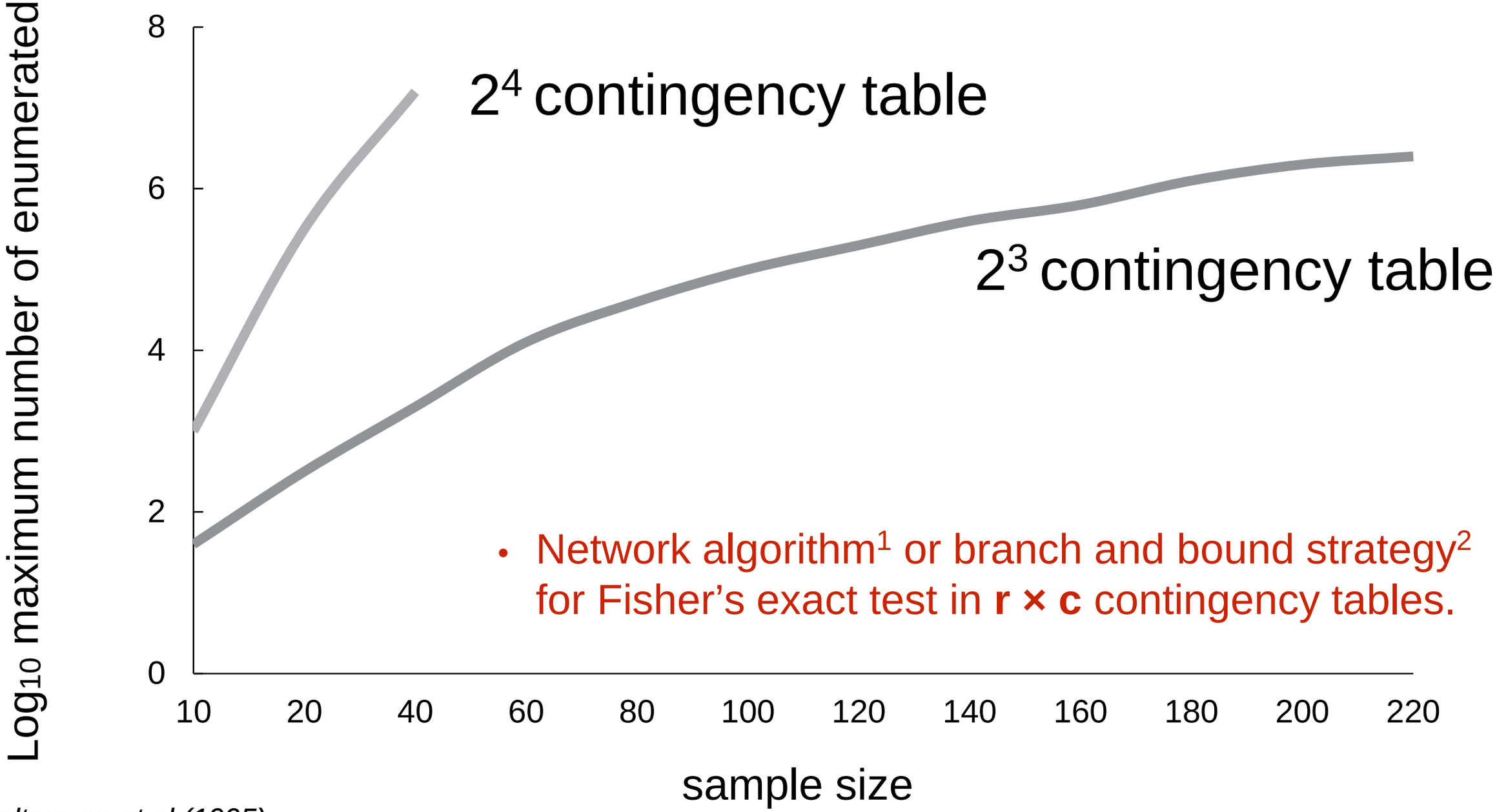
M

2x2x2 contingency table $\mathbf{X}_M$

Test statistic: $\mathrm{T}(\mathbf{X}_M)$: the sum of **exclusive entries** in $\boldsymbol{X_M}$



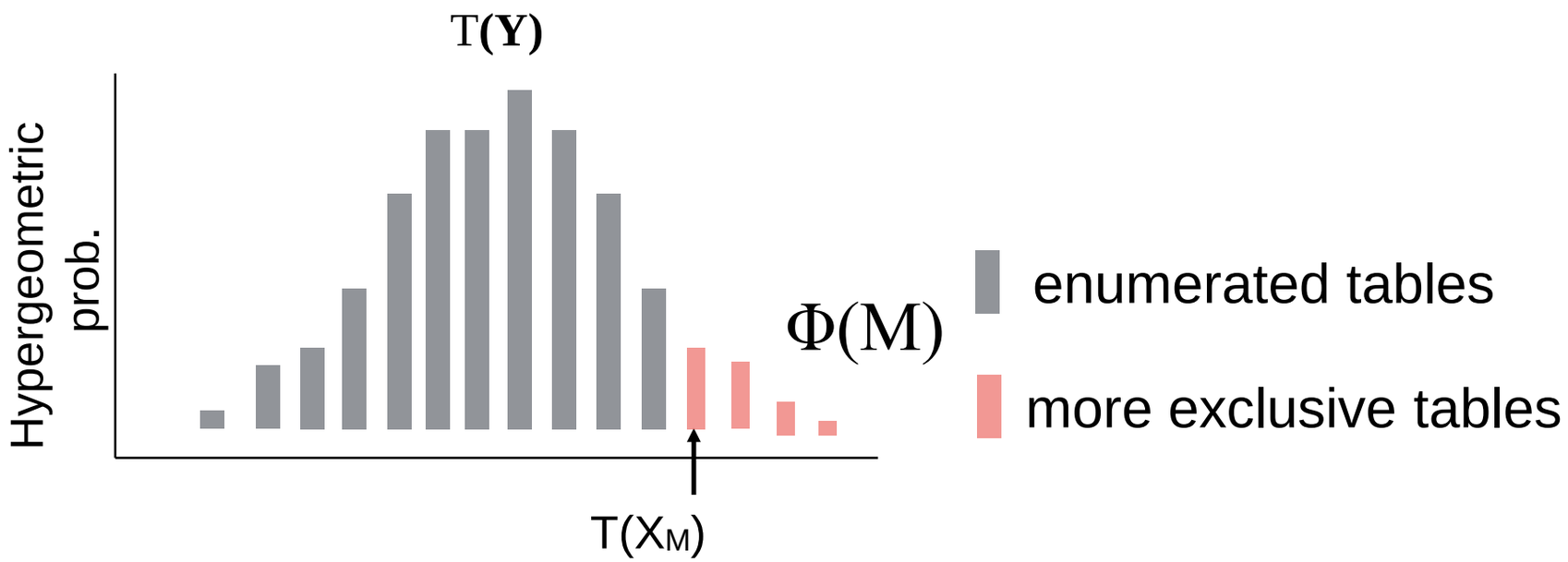
Exponential growth of table enumeration for exact distribution in higher k



- Network algorithm[1] or branch and bound strategy[2] for Fisher's exact test in **r × c** contingency tables.

Zelterman et al.(1995)

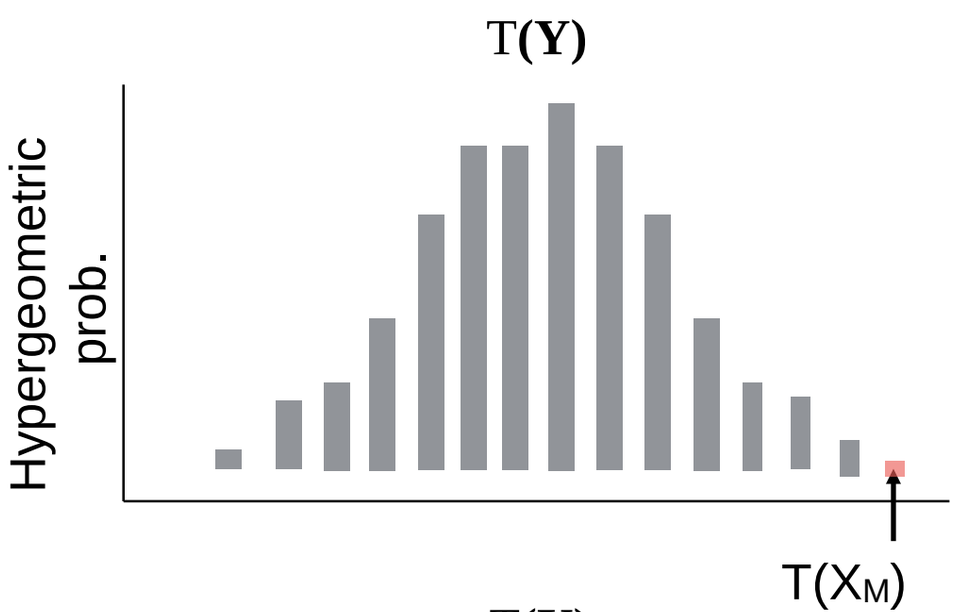
Efficient algorithm for enumerating 2^k table

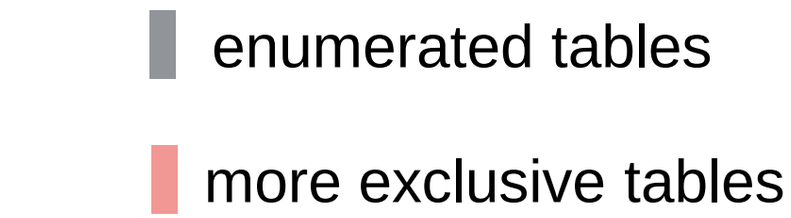


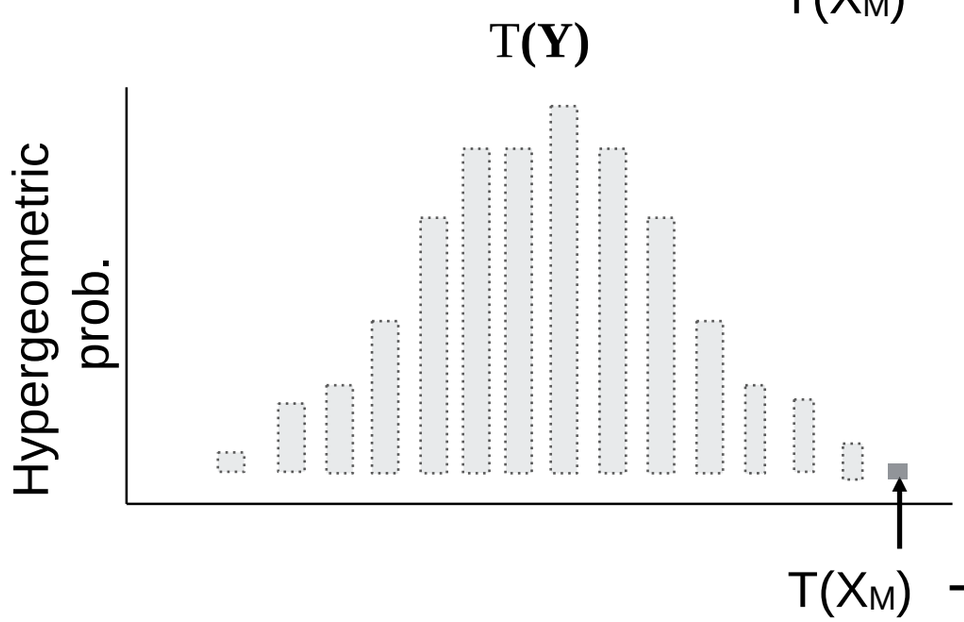
Do we need to enumerate all tables?

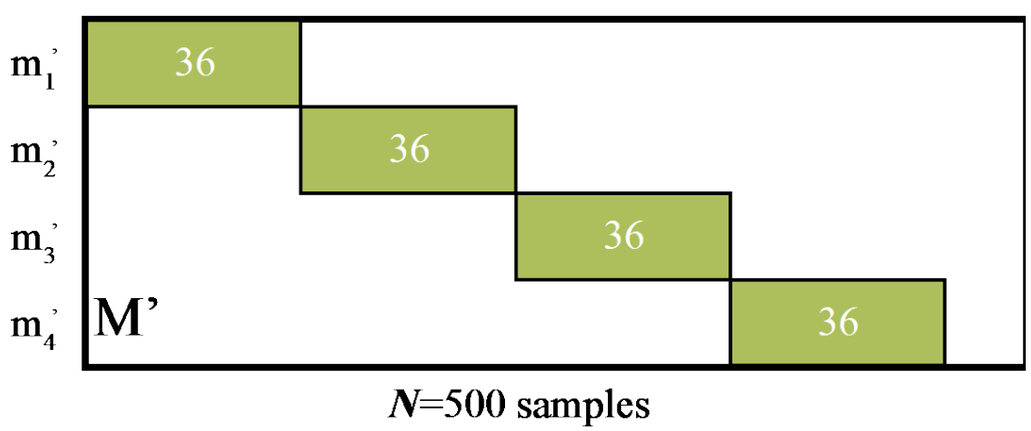
Efficient algorithm for enumerating 2^k table

Perfectly exclusive case



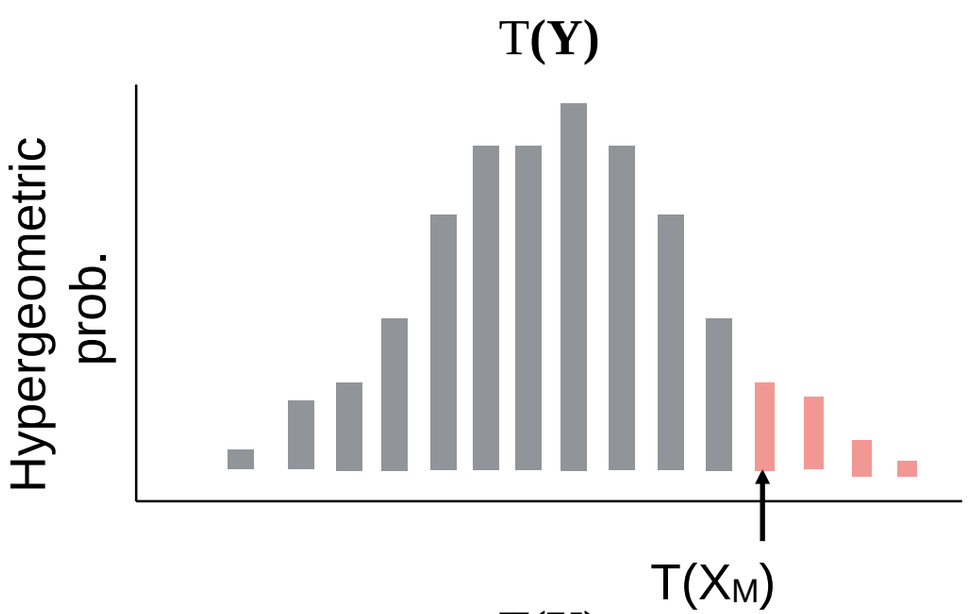


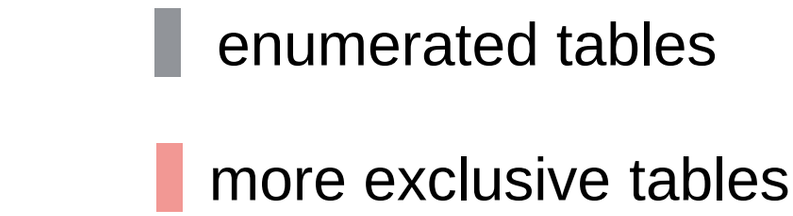


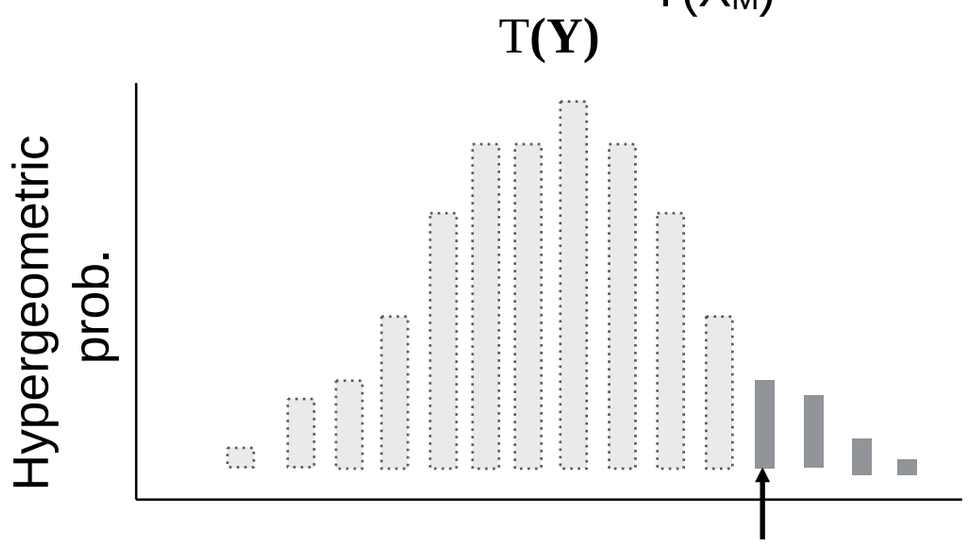


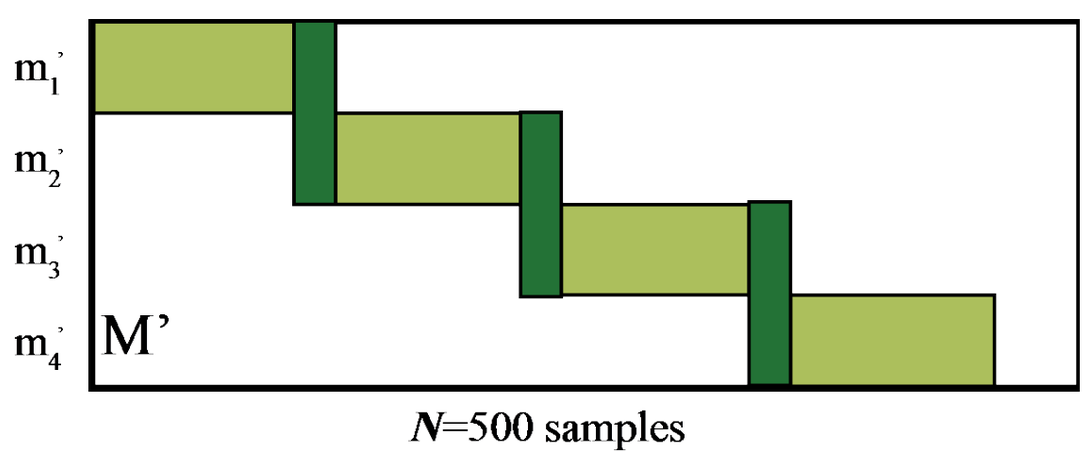
Efficient algorithm for enumerating 2^k table

CoMEt tail enumeration procedure





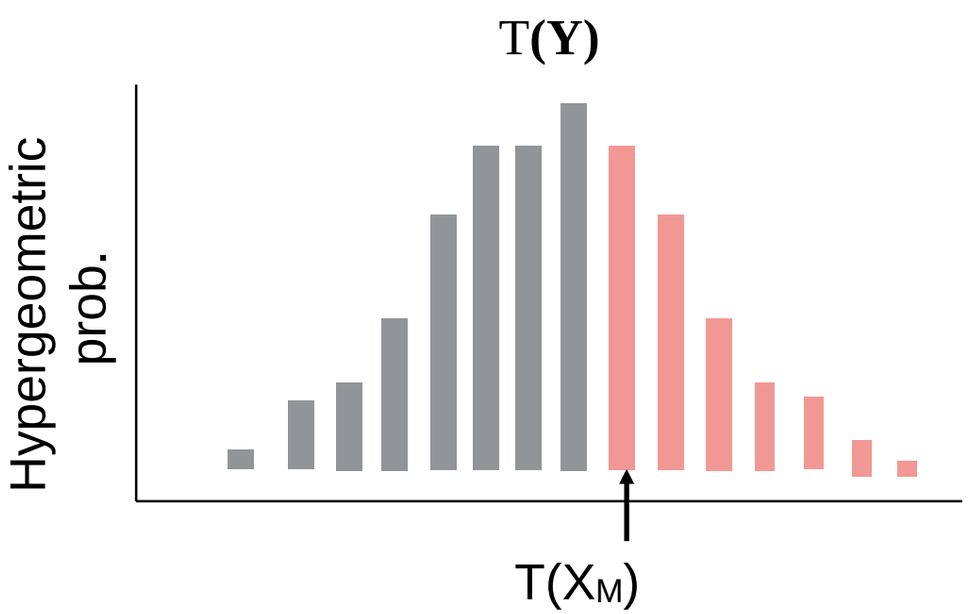


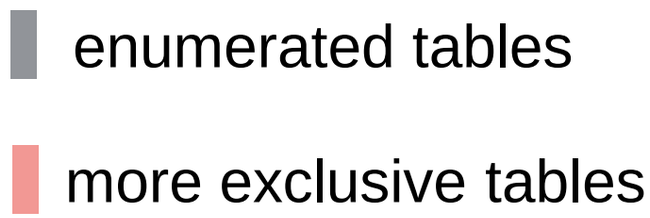


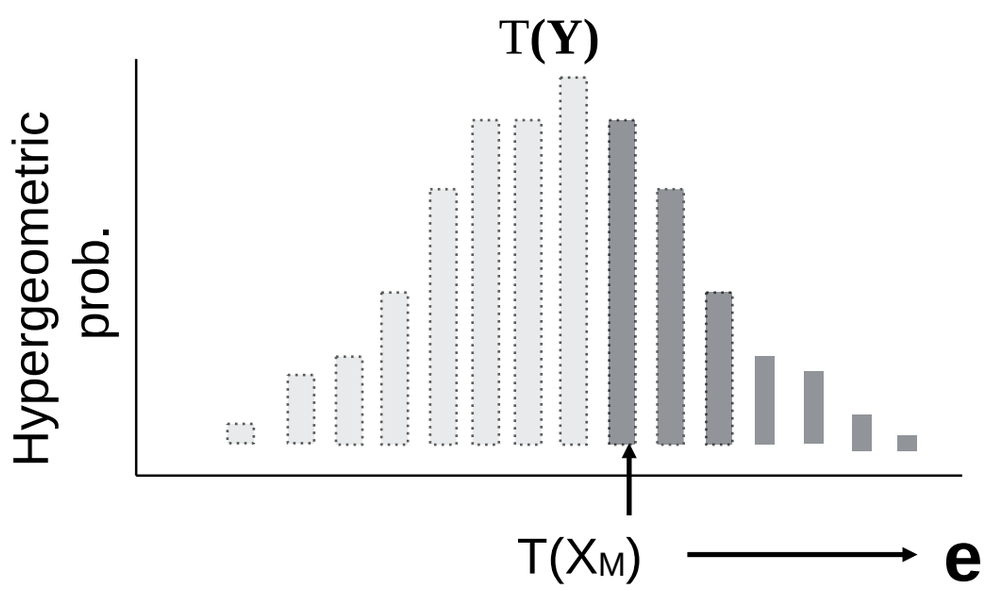
T(X_M) → **enumeration** → $\Phi(M)$

Efficient algorithm for enumerating 2^k table

More co-occurring case







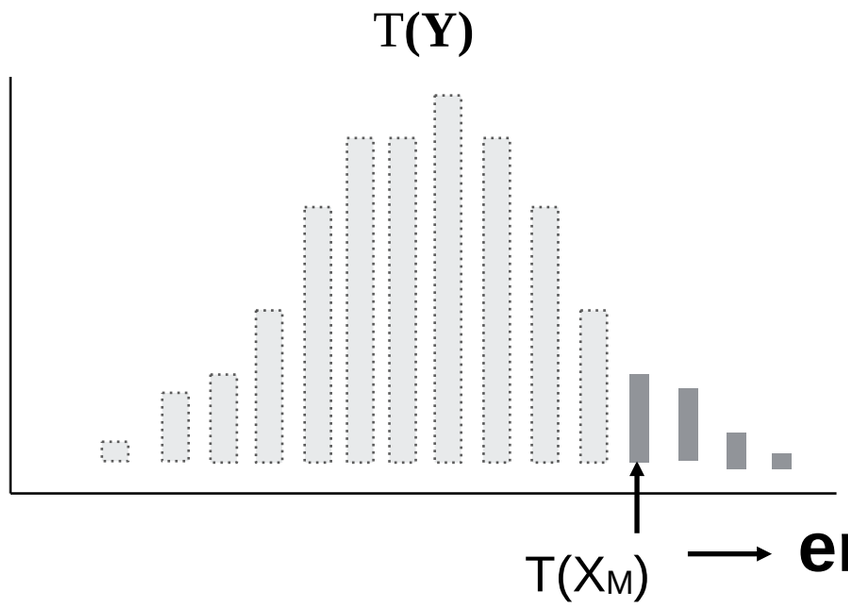
Approximation:

- Permutation approximation
- Binomial approximation

→ **enumeration**

Efficient algorithm for enumerating 2^k table

Exclusive case

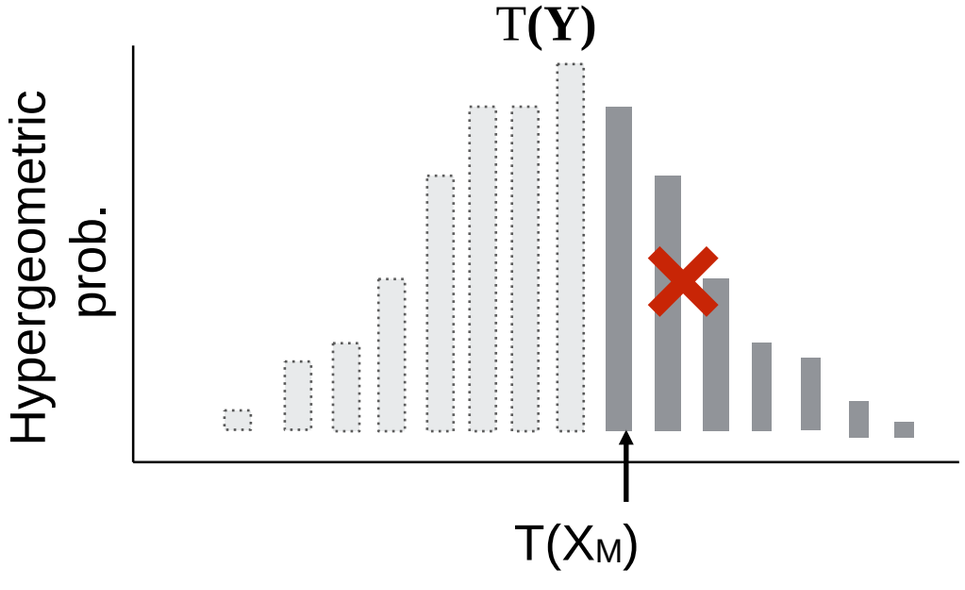


Exact 2^k:

- Tail enumeration procedure

→ **enumeration** → $\Phi(M)$

Co-occurring case



Approximation:

- Permutation approximation
- Binomial approximation

↓

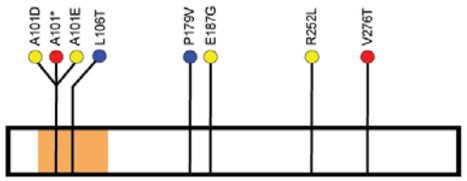
$\Phi(M)$

CoMEt

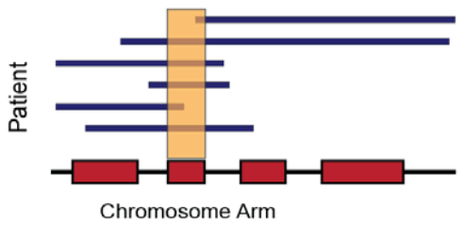
Simultaneous analysis of multiple combinations

Alteration data

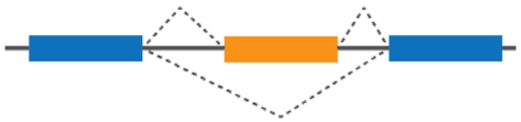
SNV/Indels



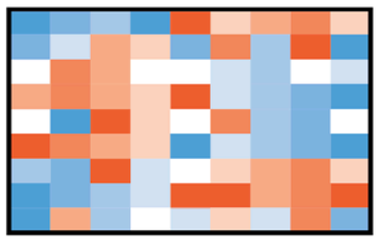
CNA



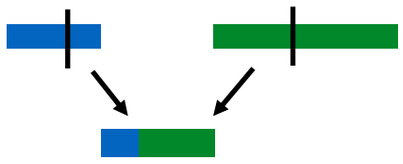
Splicing



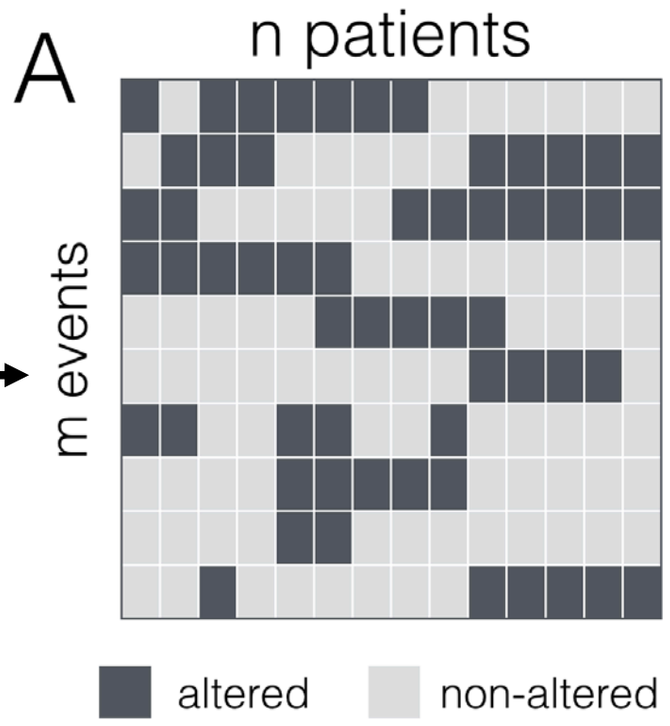
Expression



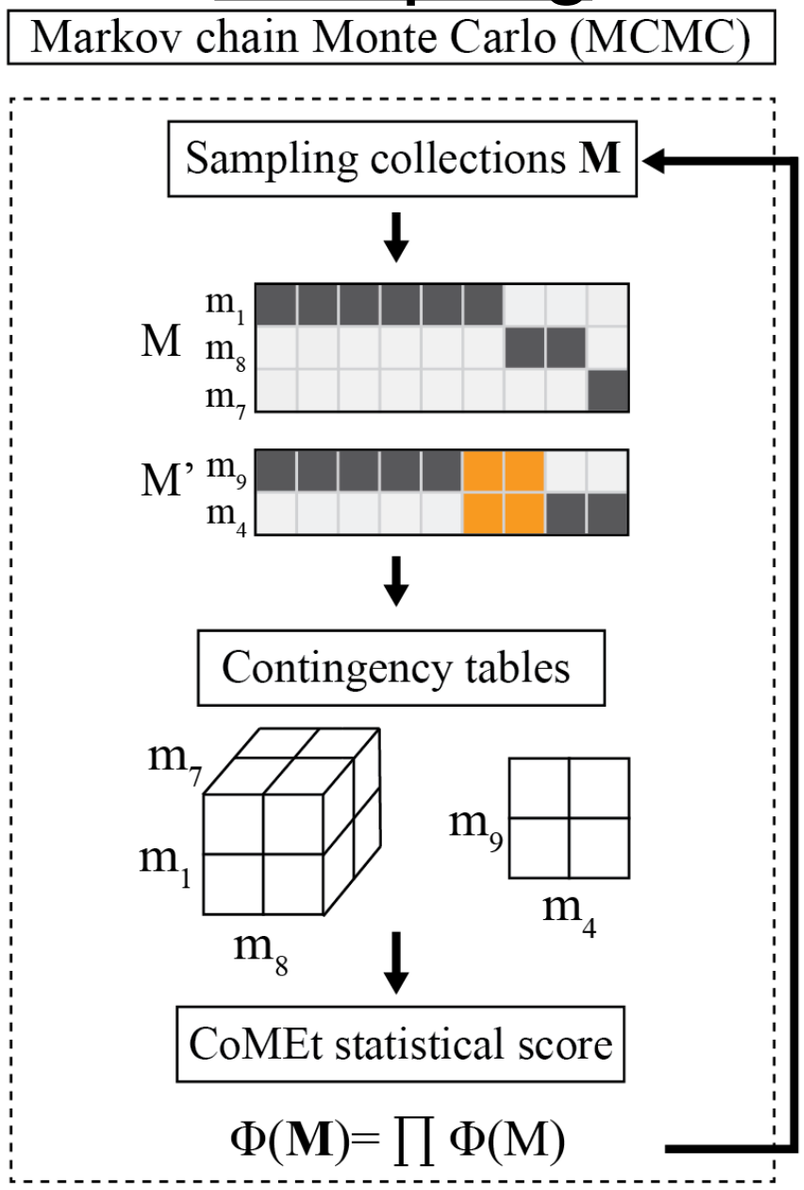
Fusion gene



Binary matrix



Sampling

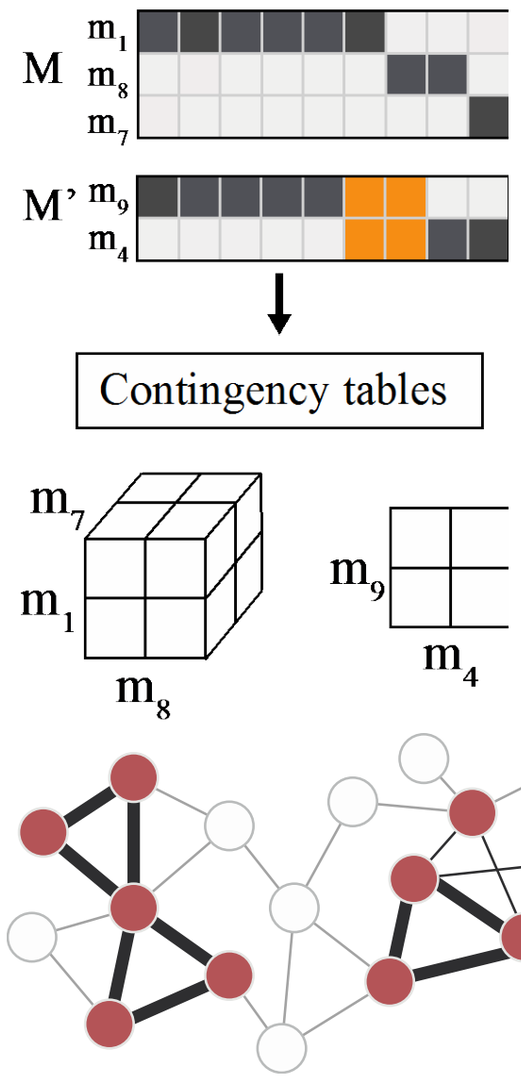


Leiserson, Wu, et al. (2015). *19th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2015).*

Contributions

A new algorithm, **CoMEt**, for identifying driver pathways *de novo:*

- ***Statistical*** score for exclusivity.
- Simultaneous analysis of ***multiple*** combinations.
- Summarize mutual exclusivity over high-scoring collections.
- Outperform other methods on simulated and real data.



Leiserson, Wu, et al. (2015)

High scoring collections

t=2, k=4

Combinations of genes (set1;set2)	$\Phi^{-1}(M)$	Sampling frequency
m1,m2,m3,m4; m5,m6,m7,m8	210	2106
m1,m2,m3,m4; m5,m9,m10,m12	160	1599
m1,m2,m4,m16; m5,m6,m7,m13	150	1511
m1,m2,m3,m16; m5,m9,m10,m14	130	1302
m1,m2,m15,m16; m5,m6,m7,m11	110	1098
m9,m10,m17,m18; m20,m21,m22,m23	100	1000
m5,m9,m10,m12; m13,m14,m19,m20	80	789
m3,m4,m15,m16; m5,m6,m7,m8	50	501
m1,m2,m16,m18; m5,m6,m7,m8	10	94

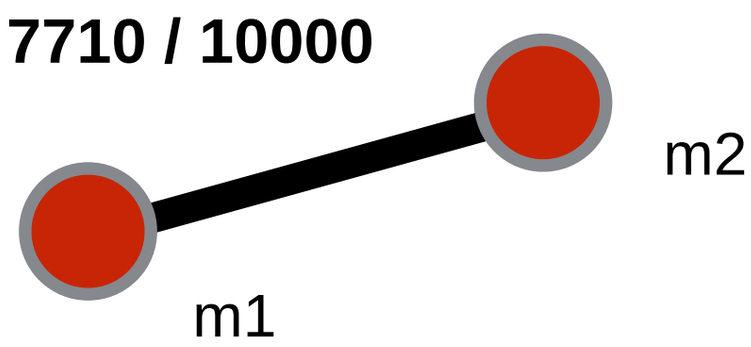
Summarize sampling results by marginal prob. graph

t=2, k=4

Marginal probability graph

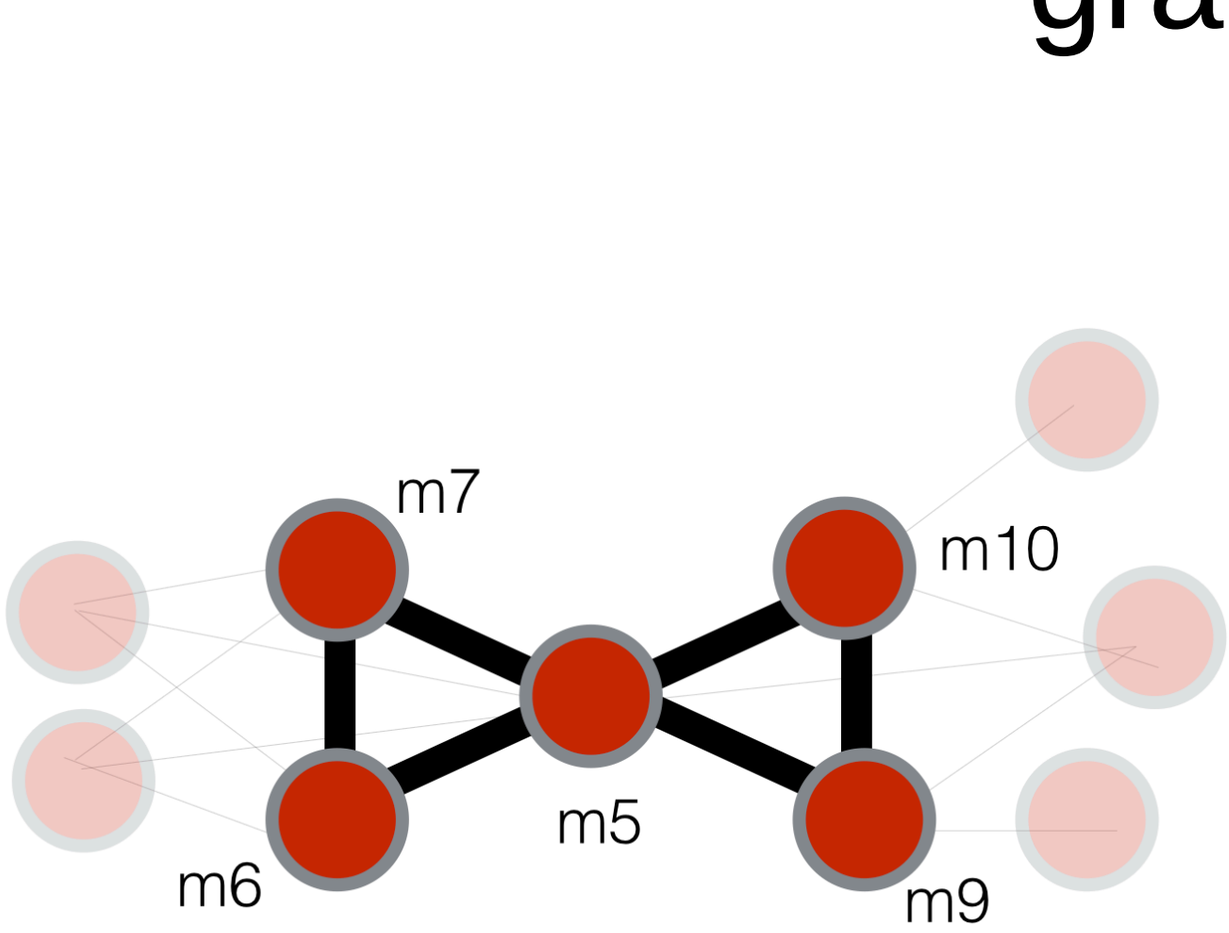
- Complete graph with weighted edges
- Reveal consensus subgraphs with high sampling freq.

Edges (m1, m2) are weighted by how often gene m1 is sampled in the same combination as gene m2

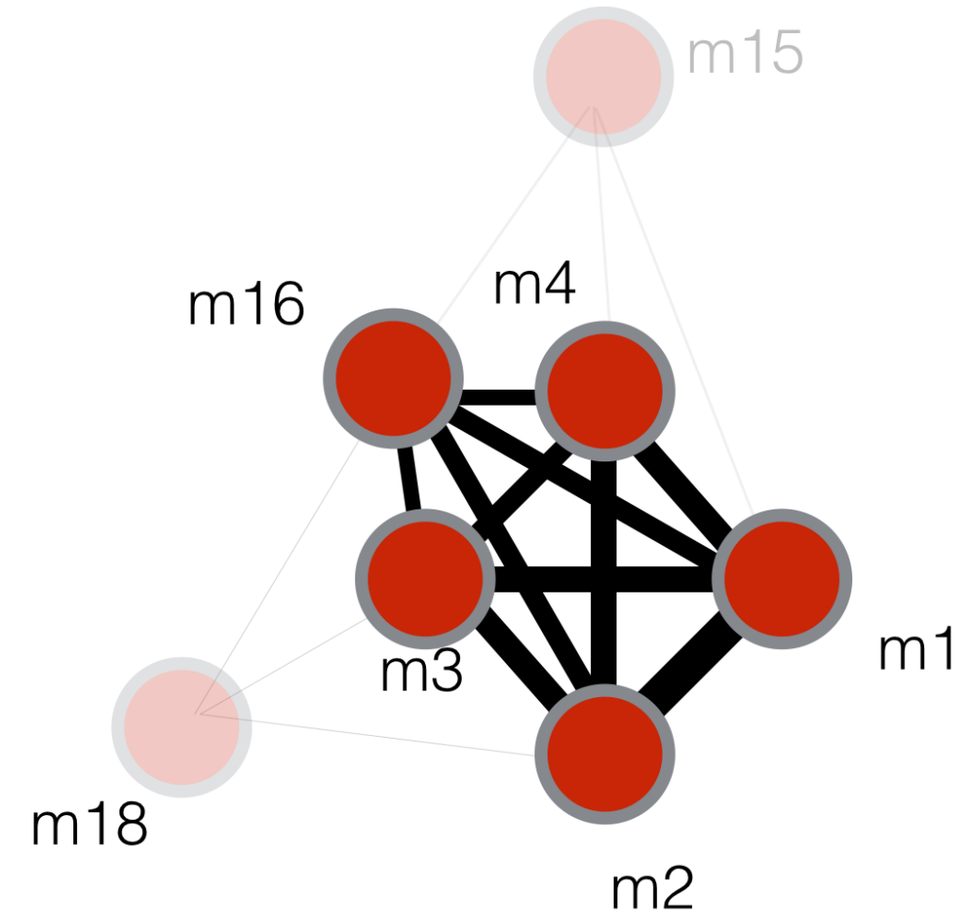


Combinations of genes (set1, set2)	$\Phi^{-1}(M)$	Sampling frequency
m1,m2,m3,m4; m5,m6,m7,m8	210	**2106**
m1,m2,m3,m4**;** m5,m9,m10,m12	160	**1599**
m1,m2,m4,m16**;** m5,m6,m7,m13	150	**1511**
m1,m2,m3,m16**;** m5,m9,m10,m14	130	**1302**
m1,m2,m15,m16**;** m5,m6,m7,m11	110	**1098**
m9,m10,m17,m18**;** m20,m21,m22,m23	100	1000
m5,m9,m10,m12**;** m13,m14,m19,m20	80	789
m3,m4,m15,m16**;** m5,m6,m7,m8	50	501
m1,m2,m16,m18**;** m5,m6,m7,m8	10	**94**

CoMEt modules from marginal prob. graph



t=2, k=4



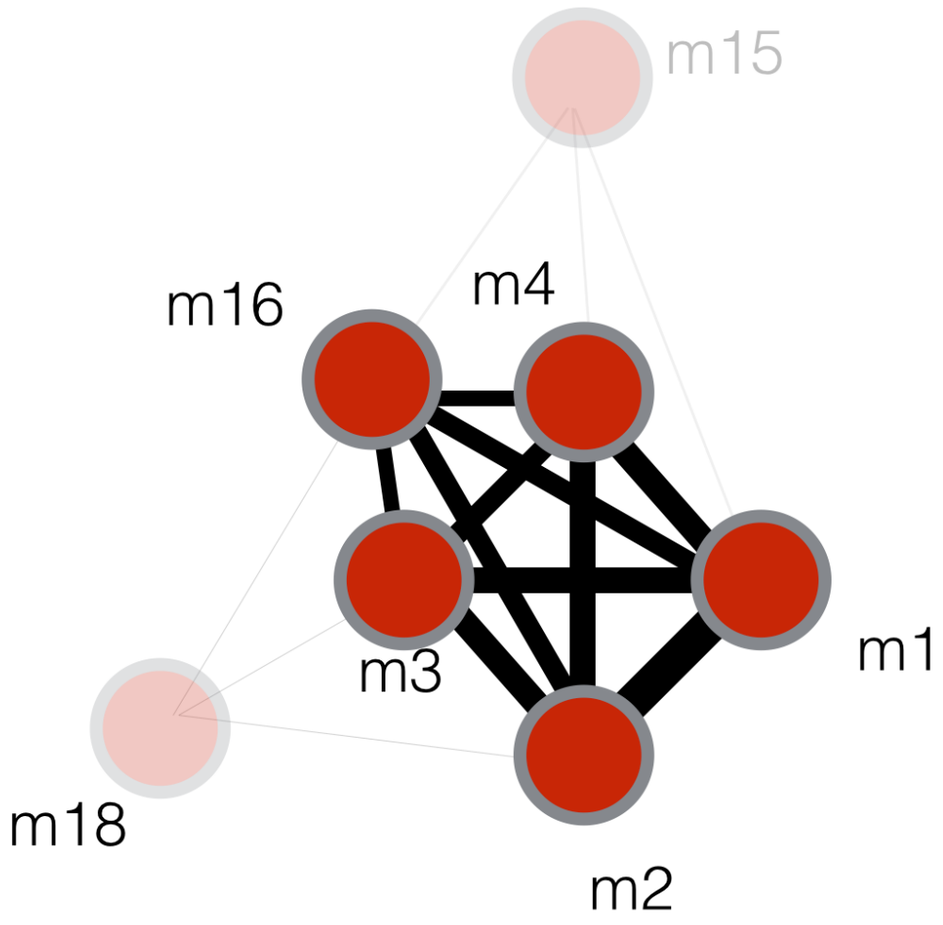
Advantages:

1. Discover complex relationship, e.g. **overlapping pathways**

2. Unconstrained size **k** and number **t** of mutually exclusive sets

Identify modules that with different sizes specifying in the parameters

t=2, k=4



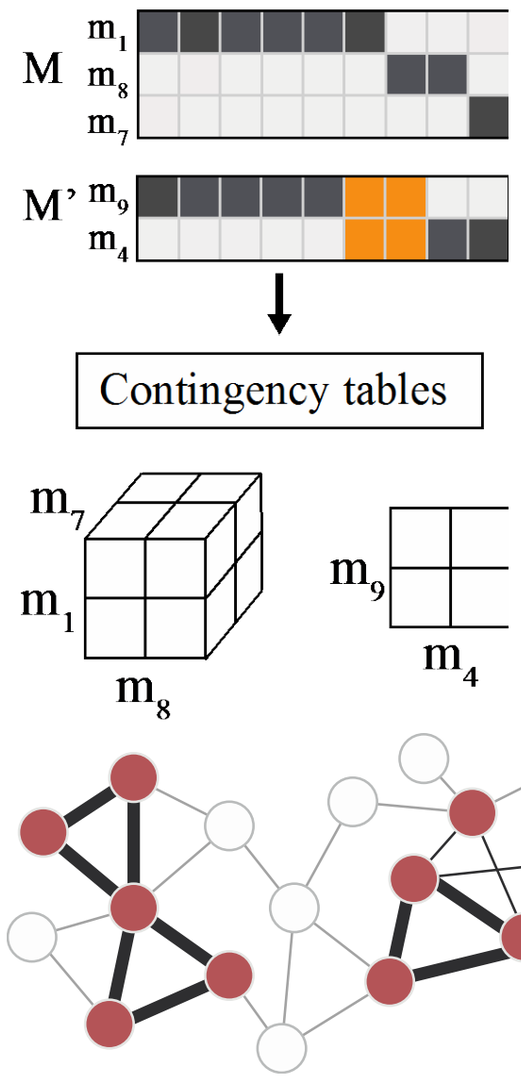
Combinations of genes	$\Phi^{-1}(M)$	Sampling frequency
m1,m2,m3,m4; m5,m6,m7,m8	210	2106
m1,m2,m3,m4; m5,m9,m10,m12	160	1599
m1,m2,m4,m16; m5,m6,m7,m13	150	1511
m1,m2,m3,m16; m5,m9,m10,m14	130	1302
m1,m2,m3,m16; m5,m6,m7,m11	110	1098
m9,m10,m17,m18; m20,m21,m22,m23	100	1000
m5,m9,m10,m12; m13,m14,m19,m20	80	789
m3,m4,m15,**m16**; m5,m6,m7,m8	50	501
m1,m2,m16,m18; m5,m6,m7,m8	10	94

2. Unconstrained size **k** and number **t** of mutually exclusive sets

Contributions

A new algorithm, **CoMEt**, for identifying driver pathways *de novo:*

- ***Statistical*** score for exclusivity.
- Simultaneous analysis of ***multiple*** combinations.
- Summarize mutual exclusivity over high-scoring collections.
- Outperform other methods on simulated and real data.

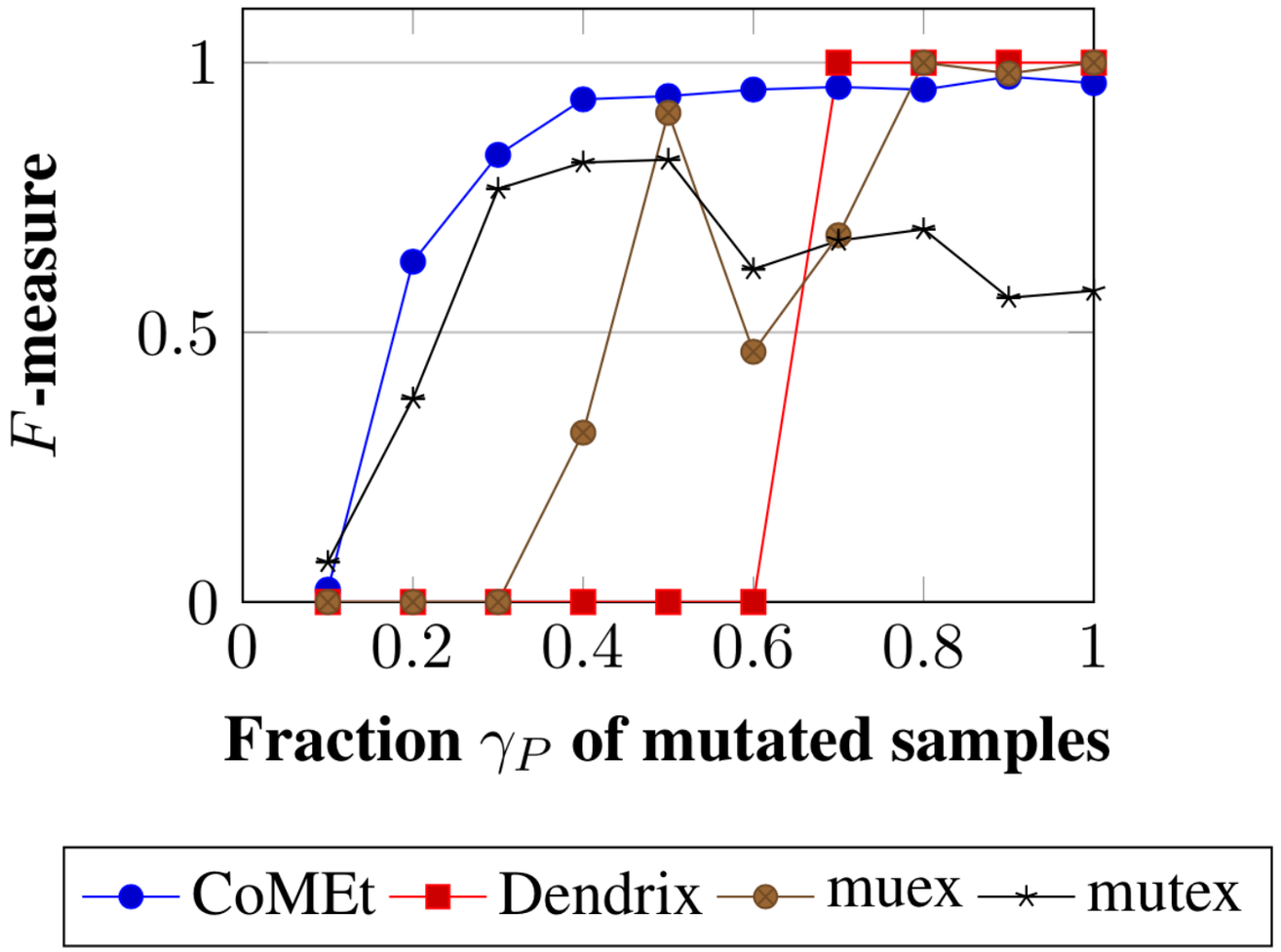


Comparison to other methods in simulated data

Run each **algorithm** on 25 simulated data sets for each coverage of the implanted pathway γ_p

Examine true positive and false positive between implanted pathway ***P*** and predicted gene sets

F-measure:
2* precision * recall / (precision+recall)

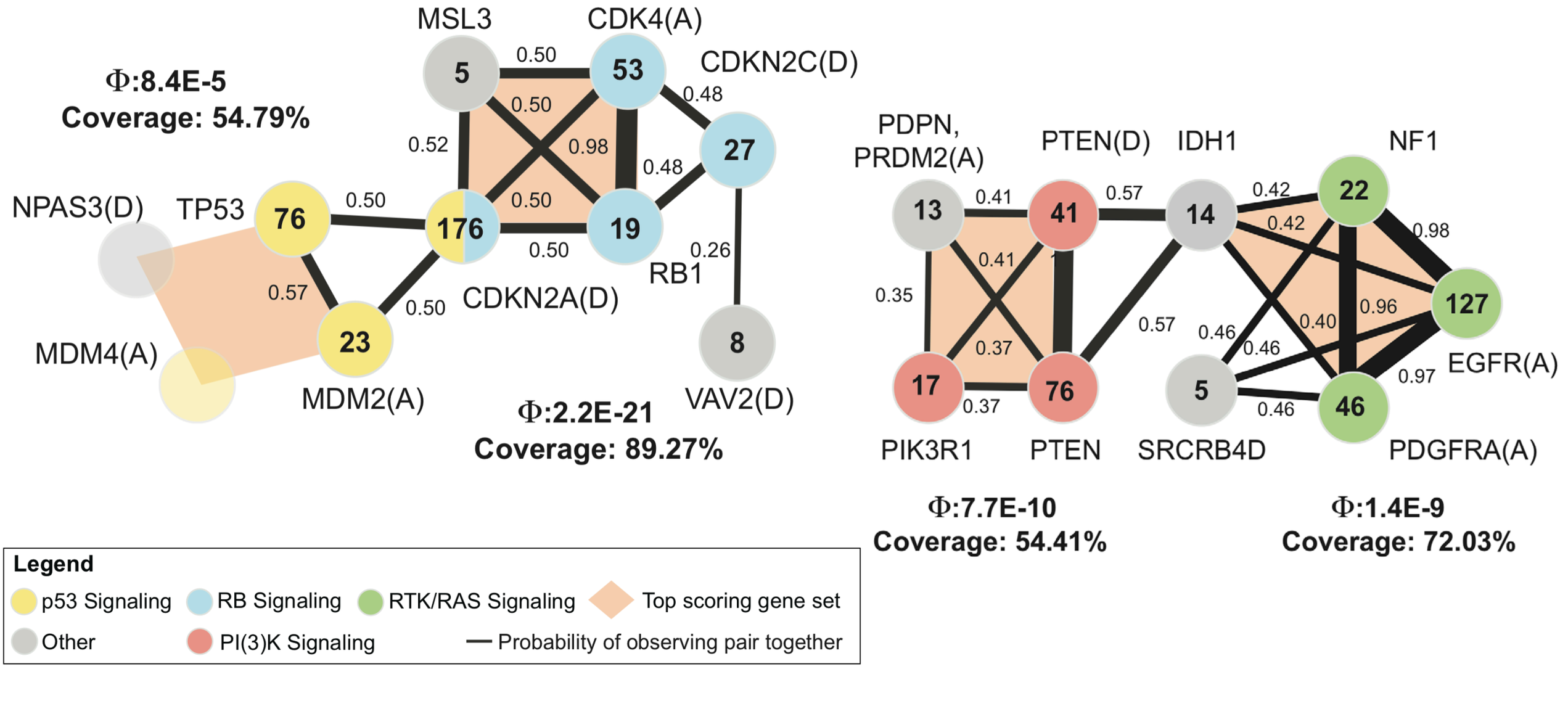


Vandin, et al. 2011; Szczurek, et al. 2014; Babur, et al. *2015*

TCGA Glioblastoma (GBM)

261 patients and 398 genes, t=4, k=4

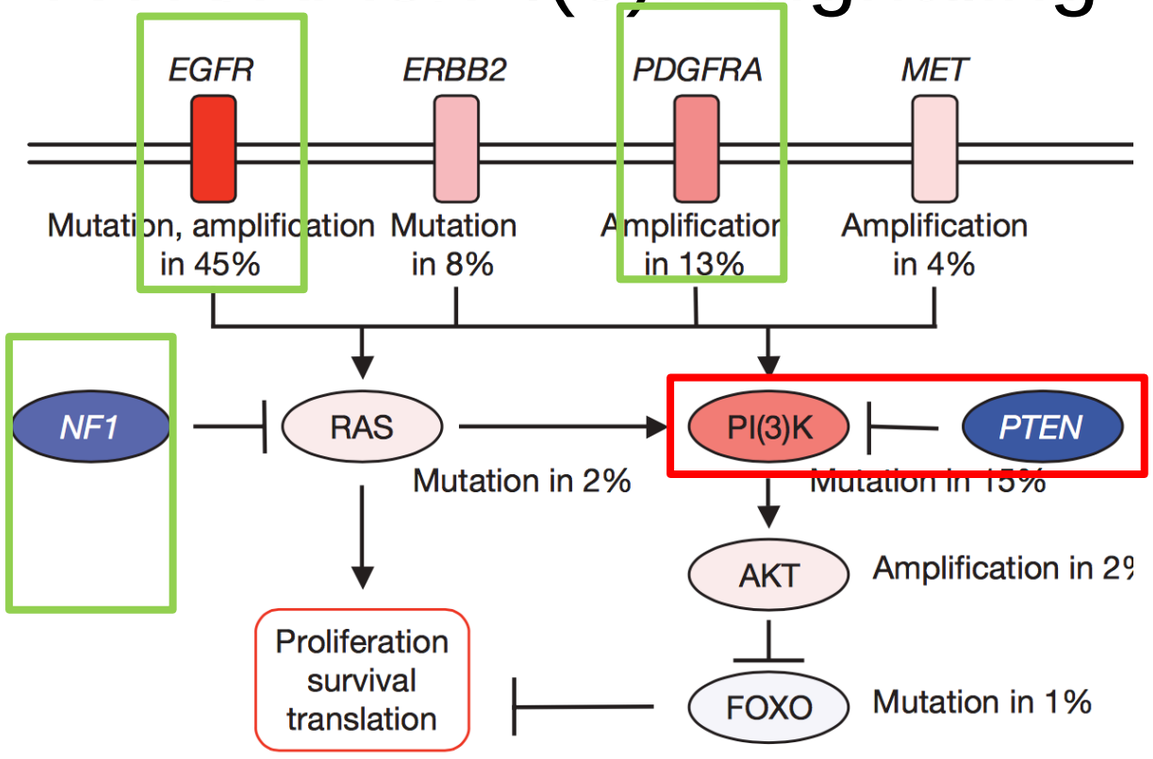
CoMEt modules



TCGA Glioblastoma (GBM)

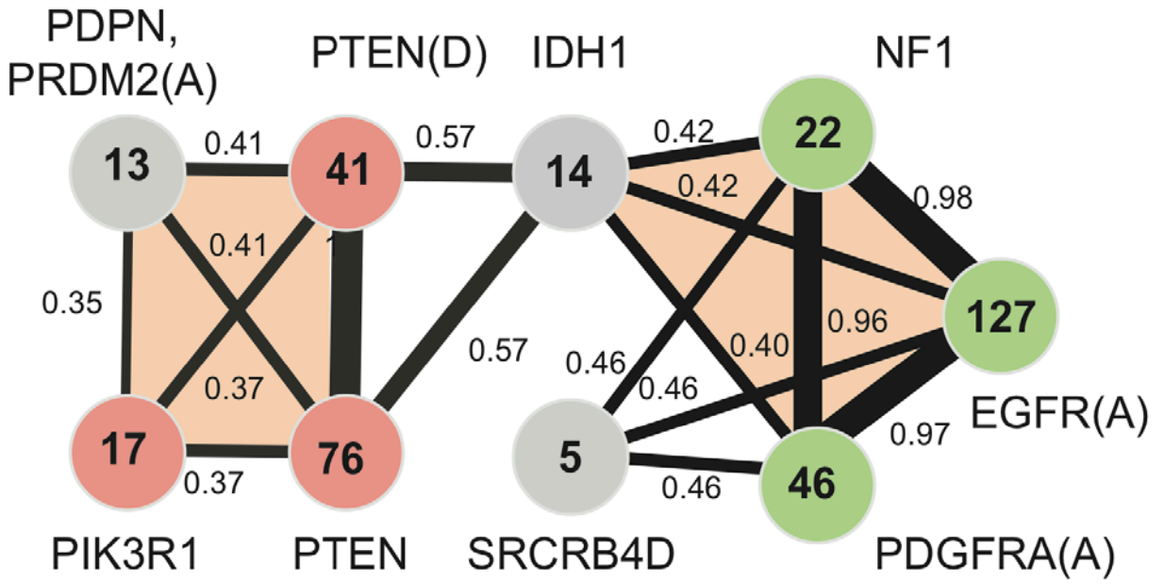
261 patients and 398 genes, t=4, k=4

RTK/RAS/PI(3)K signaling



TCGA 2008

CoMEt modules

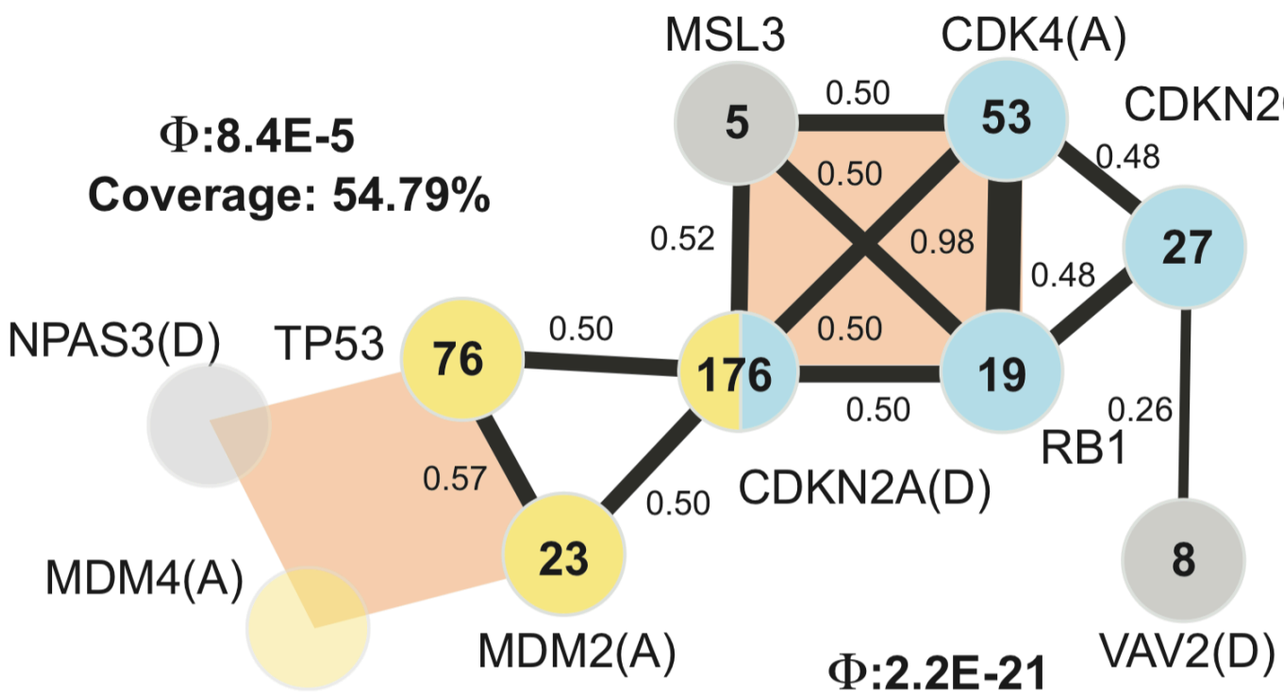


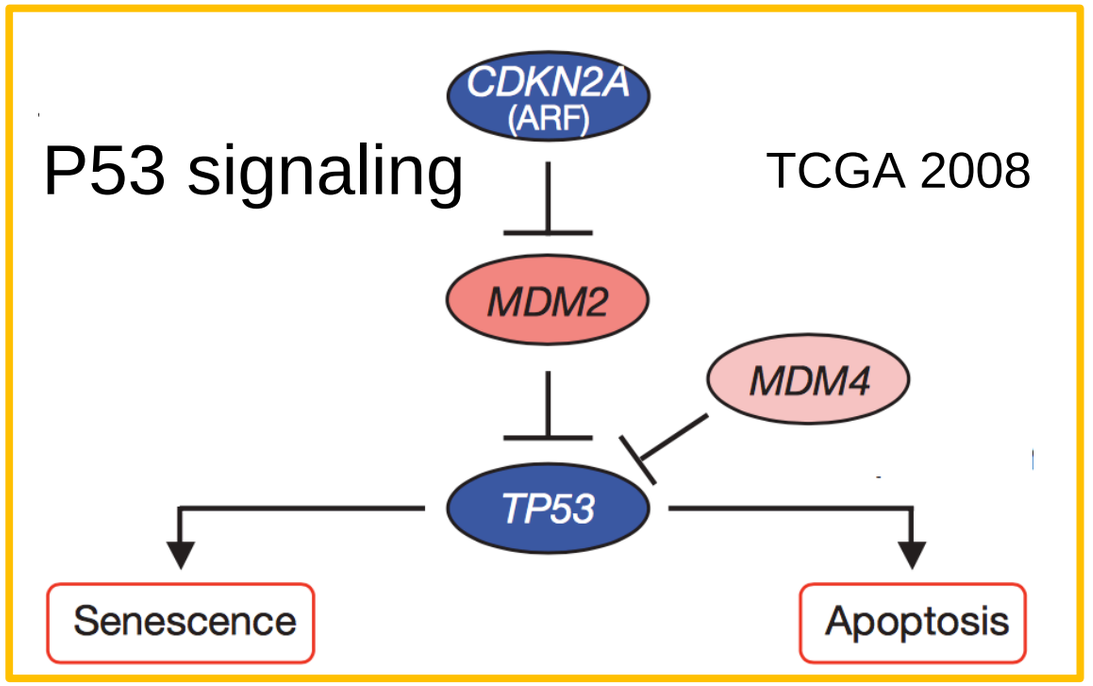
Φ:7.7E-10
Coverage: 54.41%

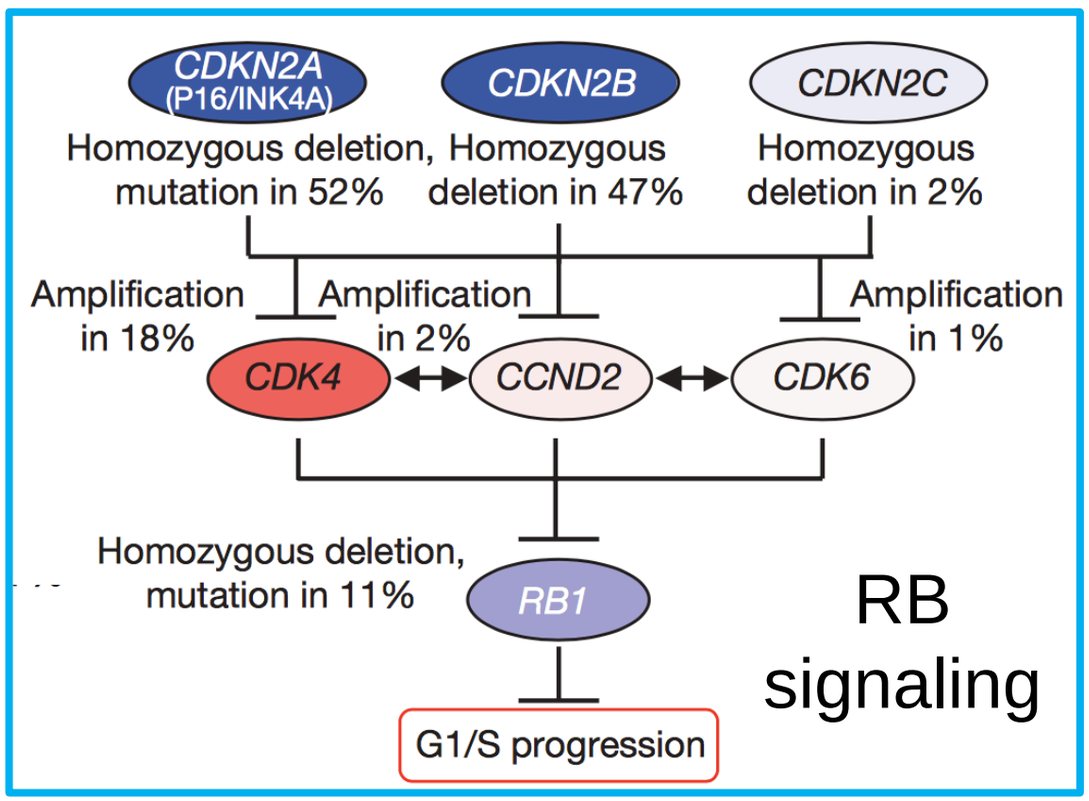
Φ:1.4E-9
Coverage: 72.03%

TCGA Glioblastoma (GBM)

261 patients and 398 genes, t=4, k=4

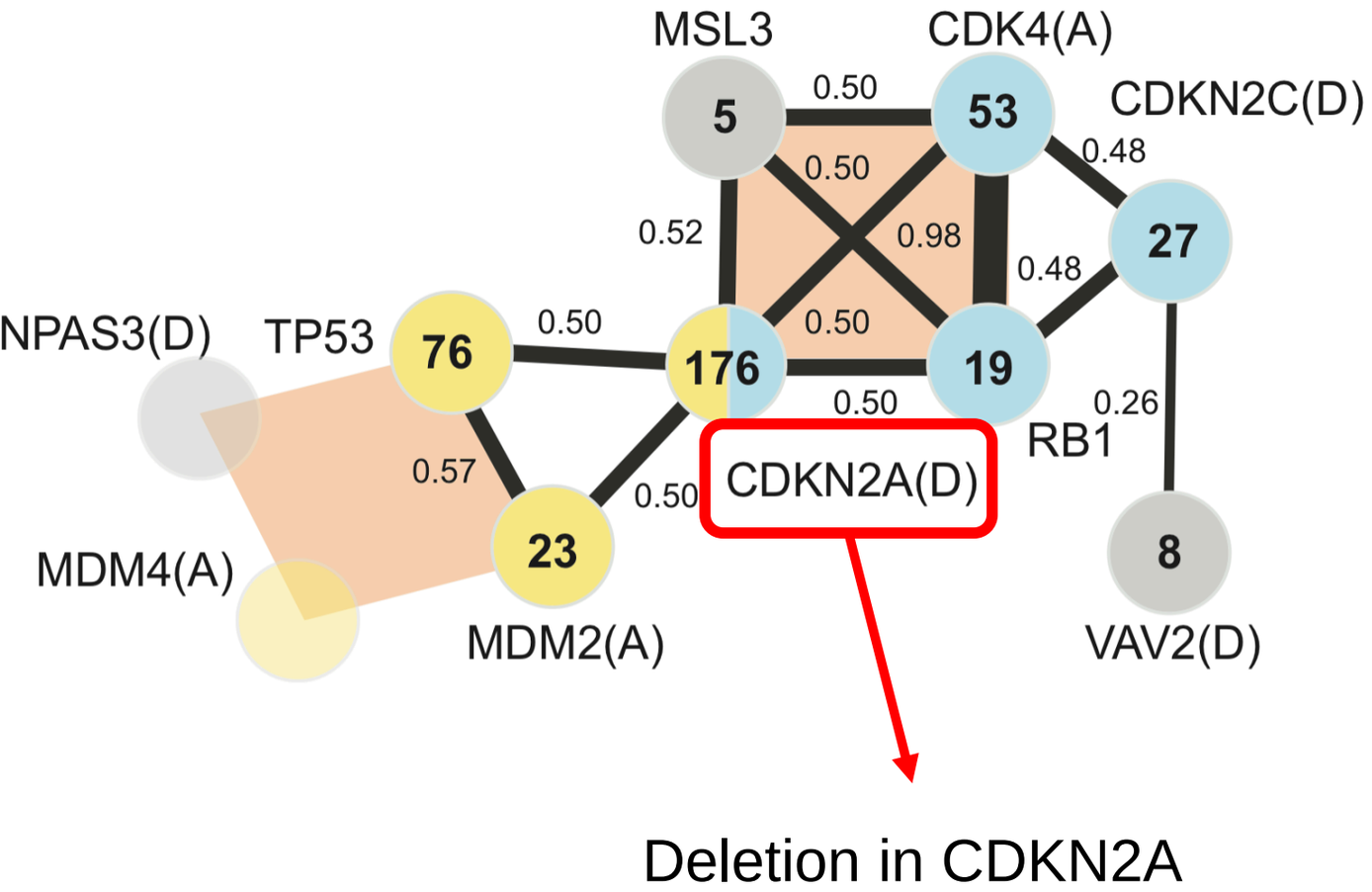






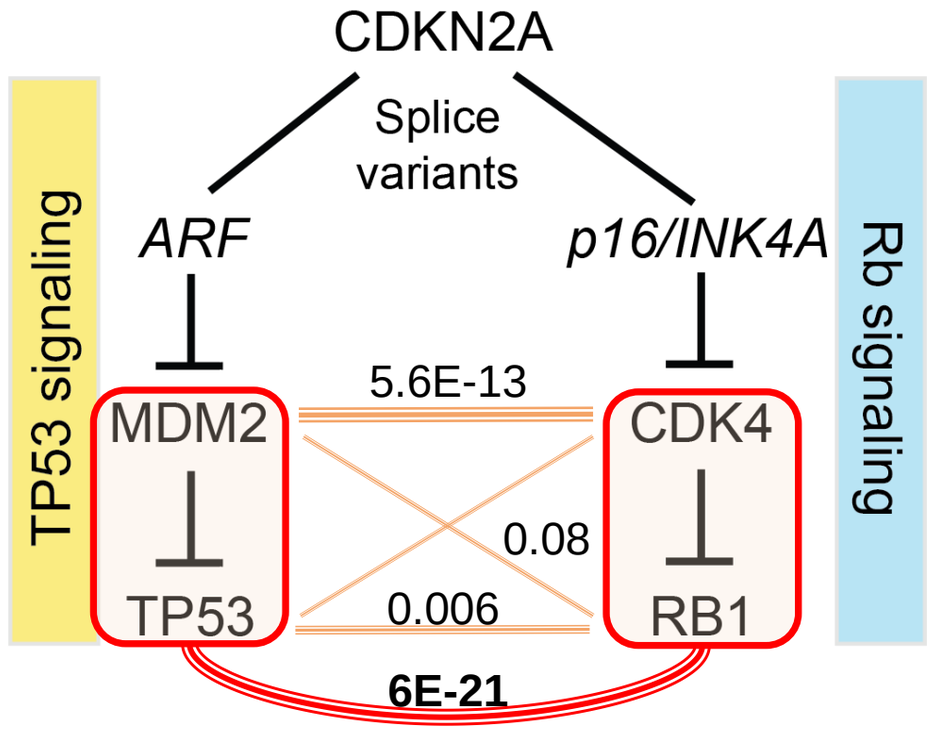
Overlapping pathways in GBM

CoMEt modules



Deletion in CDKN2A

Different isoforms of the CDKN2A are involved in the Rb and p53 signaling pathways

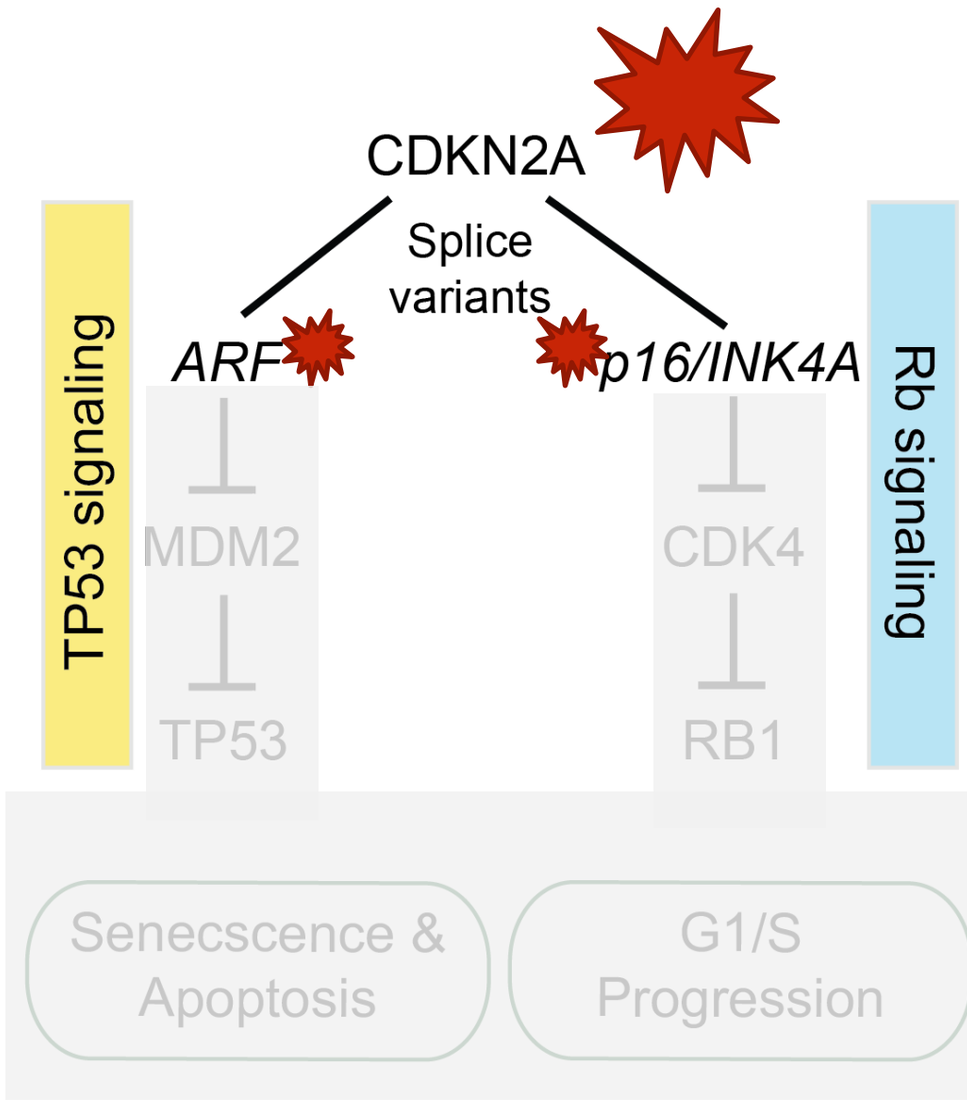


co-occurs between paris

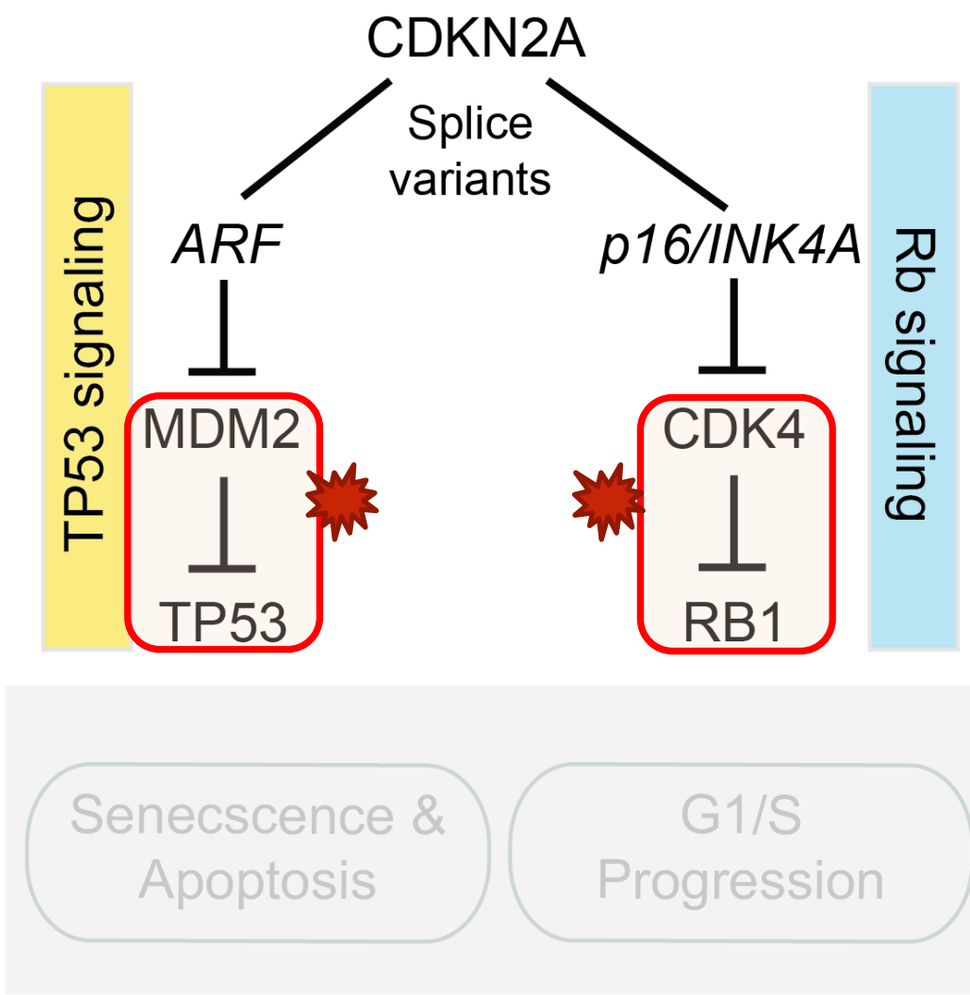
co-occurs between genes

Overlapping pathways in GBM

Copy number deletion on CDKN2A affects both isoforms

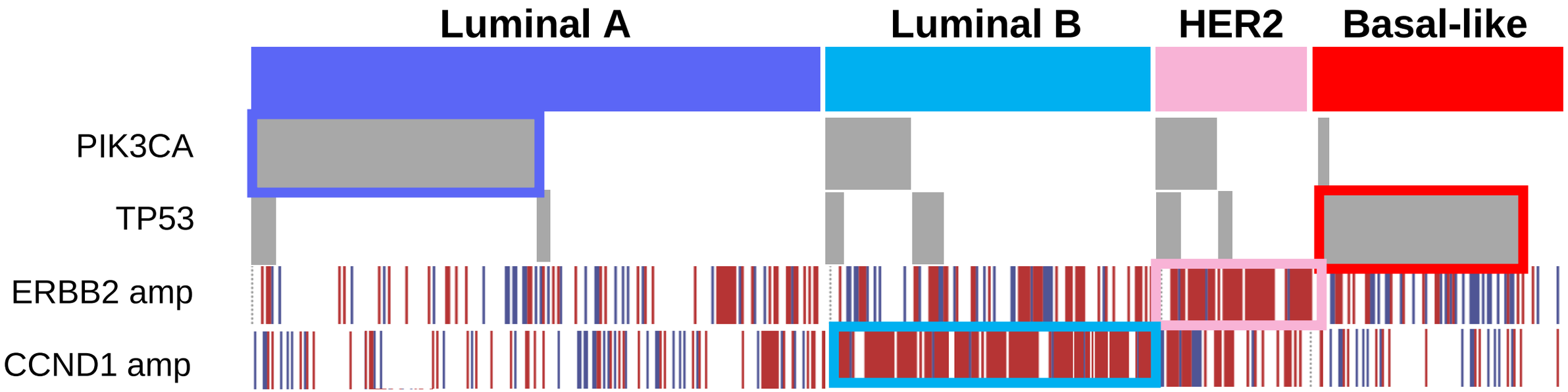


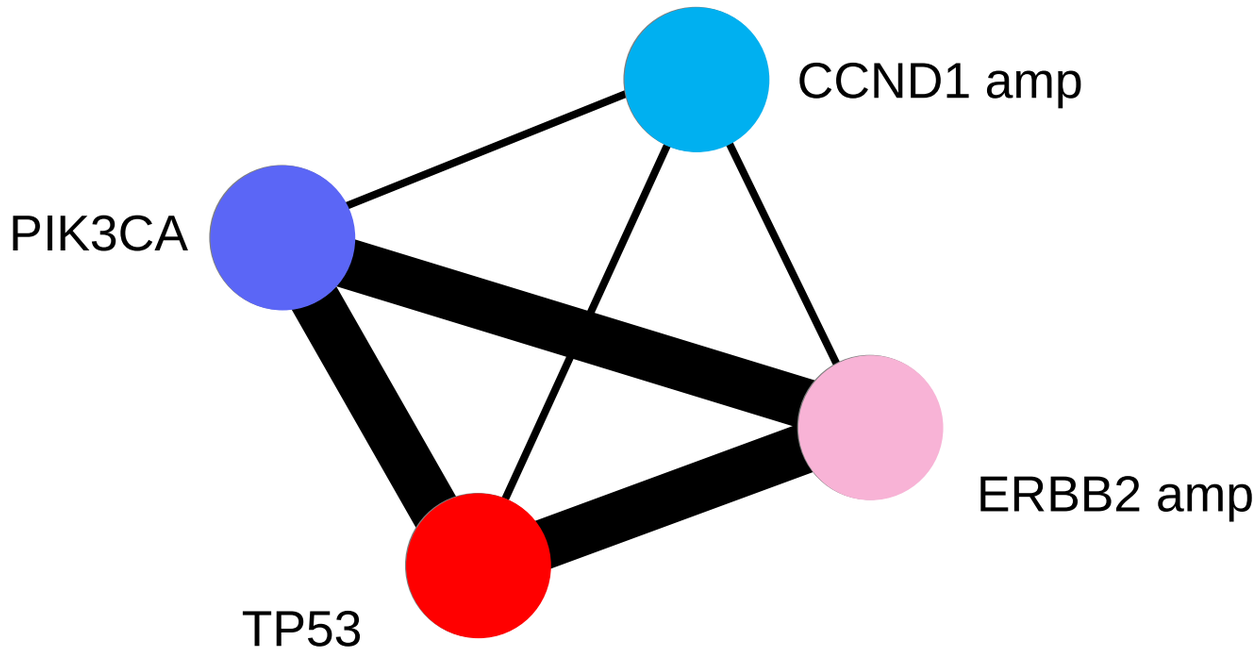
High co-occurring between pairs in Rb and P53 signaling pathways



mutations

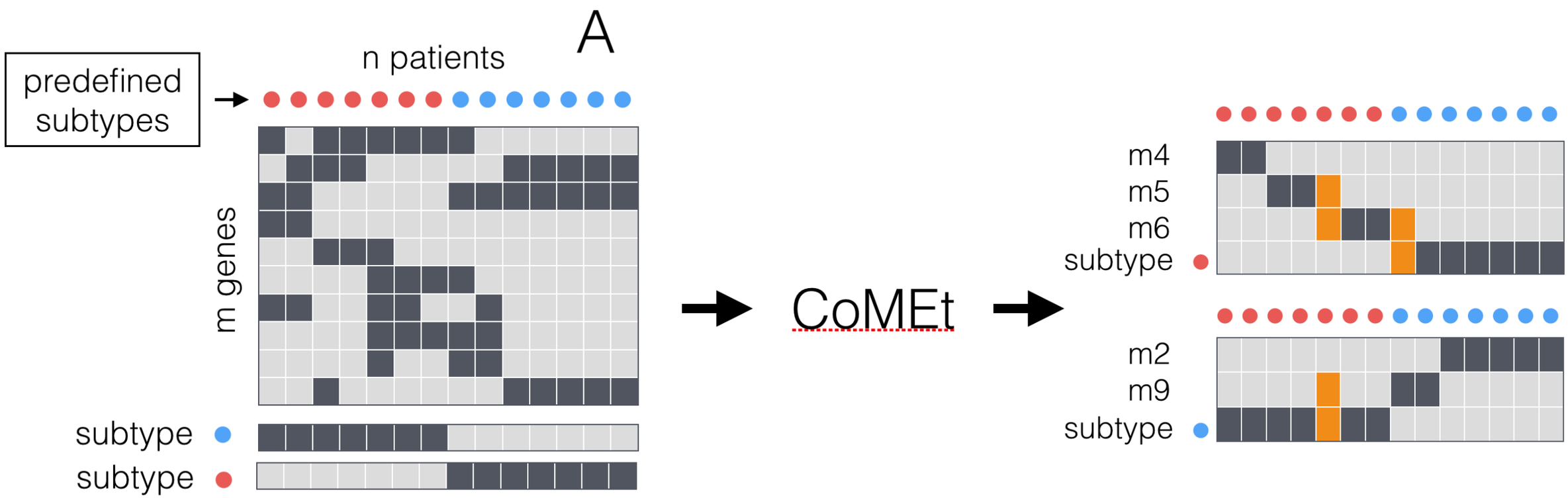
Mutual exclusivity between subtype-enriched mutations





TCGA BRCA 2012

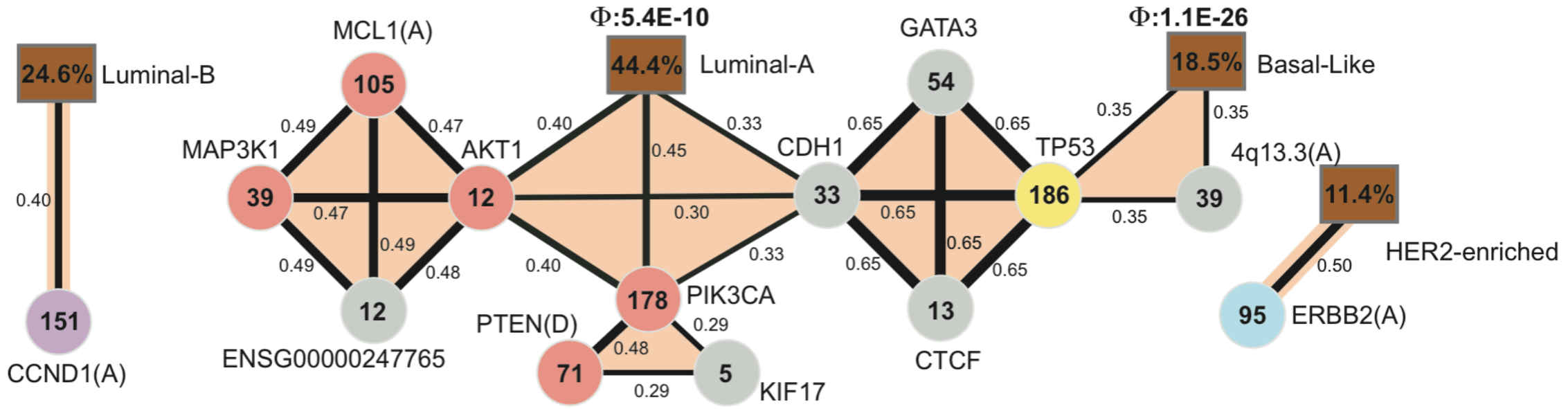
Simultaneous analysis of subtype-specific mutations/pathways



TCGA Breast cancer (BRCA)

507 patients and 375 genes + 4 molecular subtypes, t=4, k=4

CoMEt modules



Legend

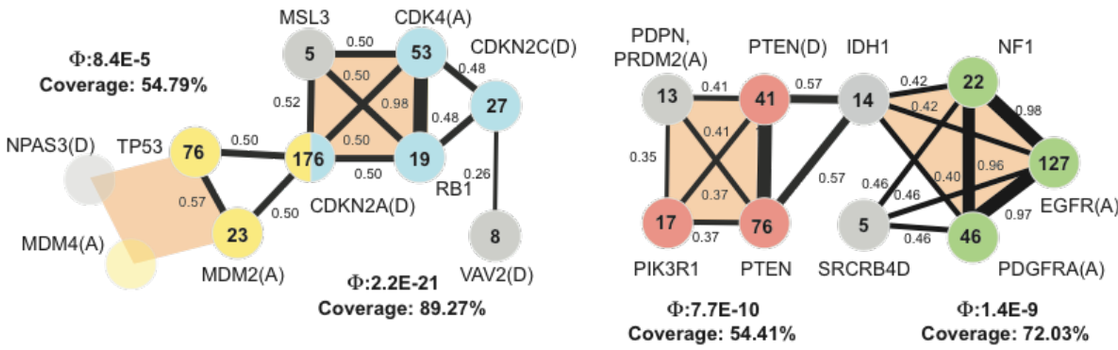
p53 Signaling · RTK/RAS Signaling · PI(3)K/Akt Signaling · Other · Probability of observing pair together

Cell cyle mediators · Genomically stable gastric cancer · c-MET Signaling · Subtype · Top scoring gene set

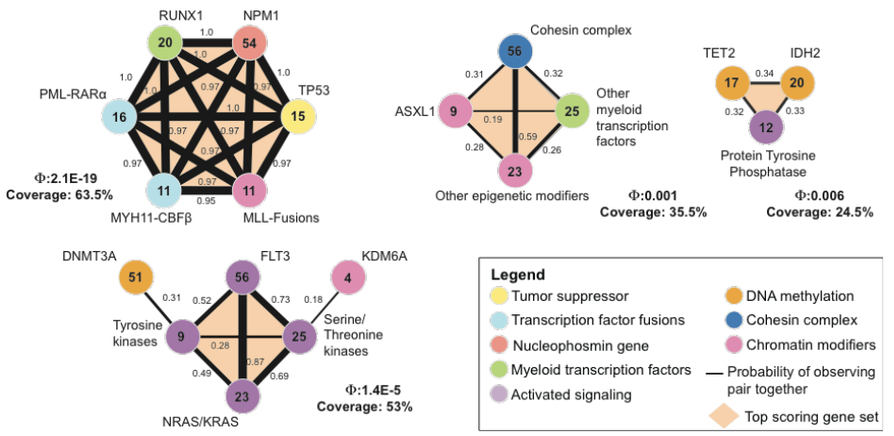
CoMEt

Simultaneous analysis of (sub)type and generic exclusivity in TCGA data

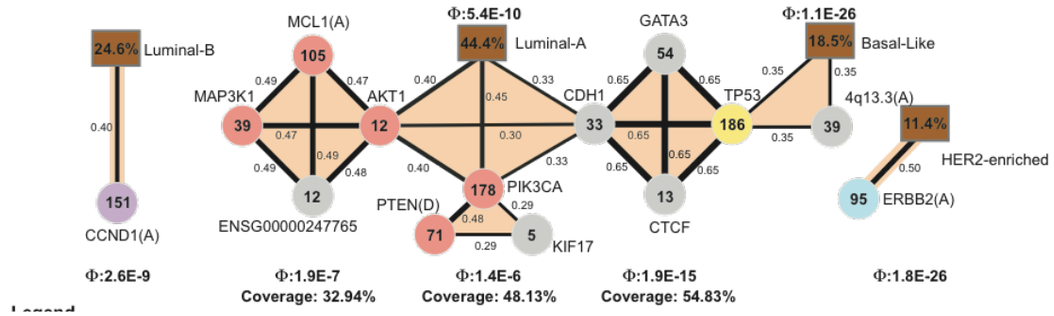
Glioblastoma (GBM)



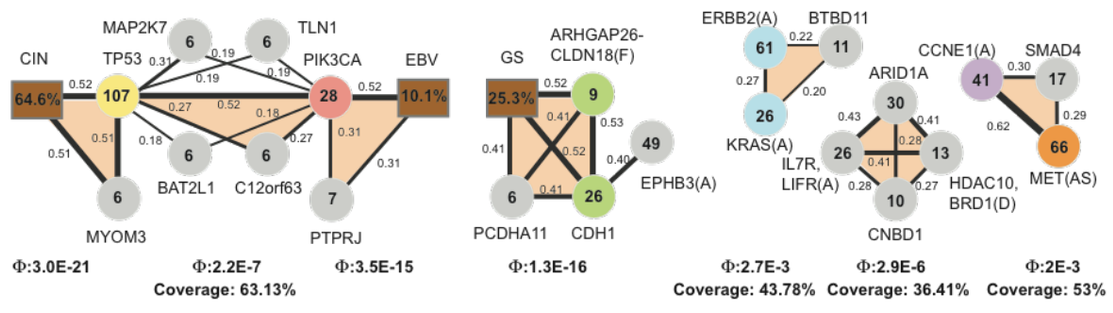
Acute myeloid leukemia (AML)



Breast cancer (BRCA) with subtypes



Gastric cancer (STAD) with subtypes

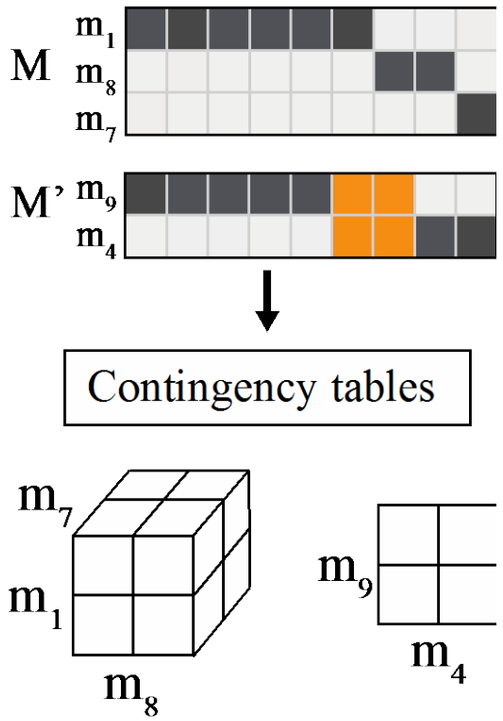


Poster #32

Summary

A new algorithm, **CoMEt**, for identifying driver pathways *de novo:*

- ***Statistical*** score for exclusivity.
- Simultaneous analysis of ***multiple*** combinations.
- Summarize mutual exclusivity over high-scoring collections.
- Outperform other methods on simulated and real data.



Paper is available to download at http://arxiv.org/abs/1503.08224
Software: http://compbio.cs.brown.edu/software/

Leiserson, Wu, et al. (2015)

Acknowledgements

RAPHAEL LAB

Ben Raphael
Max Leiserson*
Fabio Vandin
Mohammed El-Kebir
Connor Gramazio
Ahmad Mahmoody
Layla Oesper
Matthew Reyna
Gryte Satas

NSF

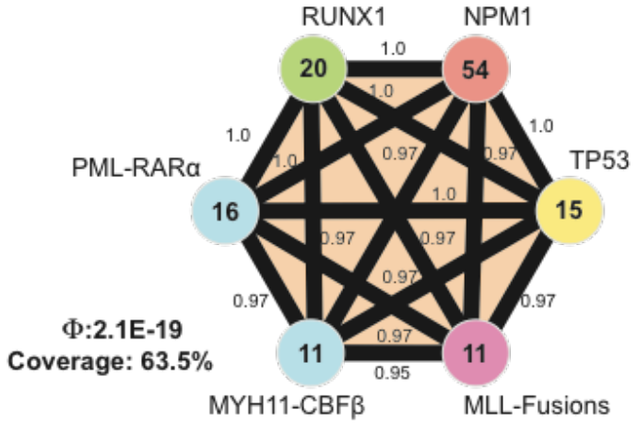
NATIONAL INSTITUTES OF HEALTH

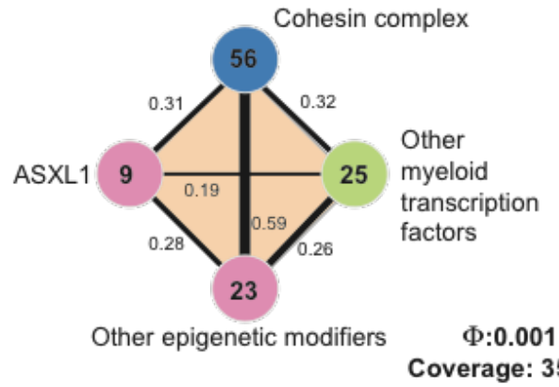
The Cancer Genome Atlas
Understanding genomics to improve cancer care

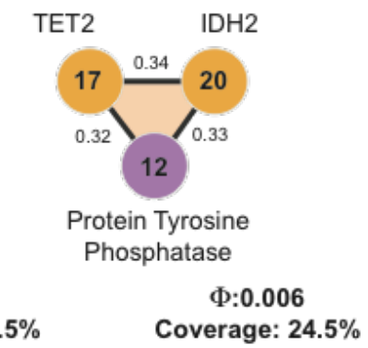
CCMB

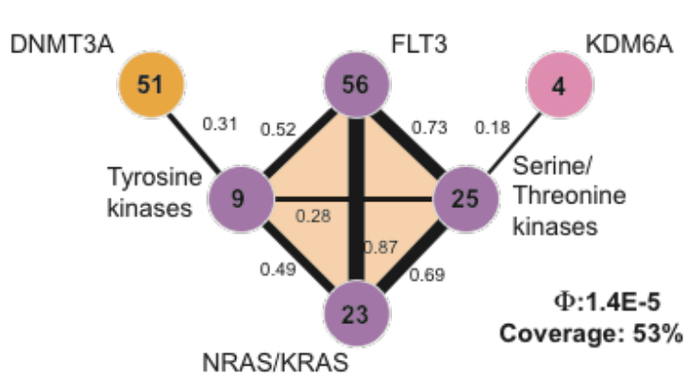
TCGA Acute myeloid leukemia (AML)

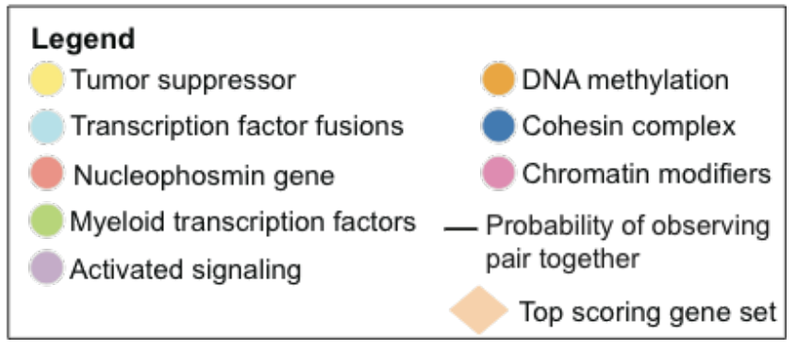
200 patients and 51 genes / categories, t=4, k=[6,4,4,3]

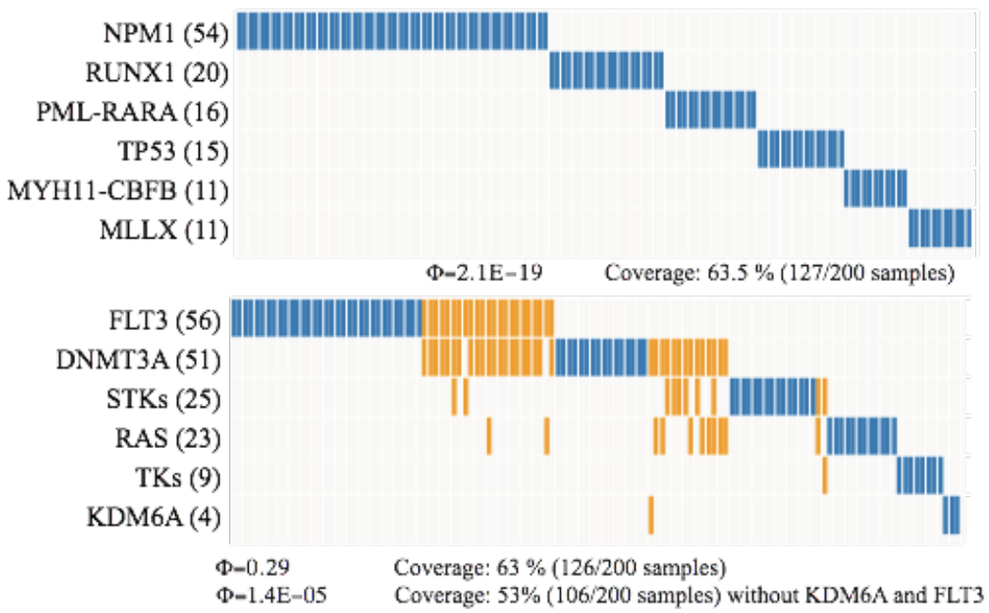


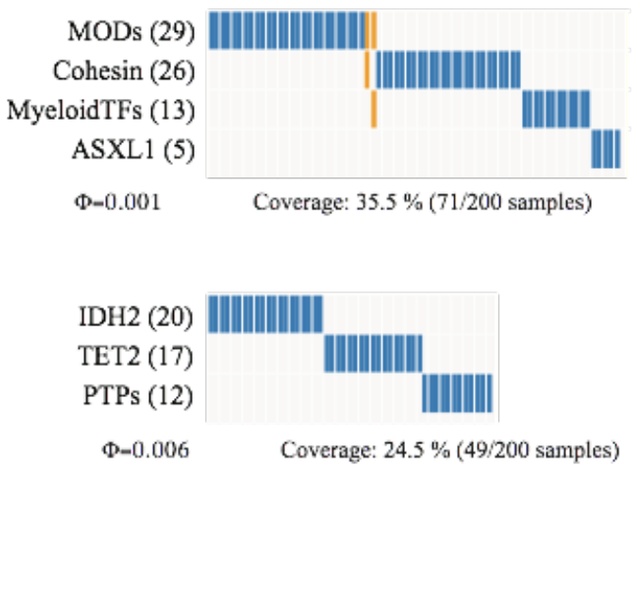


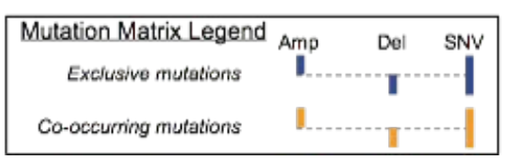






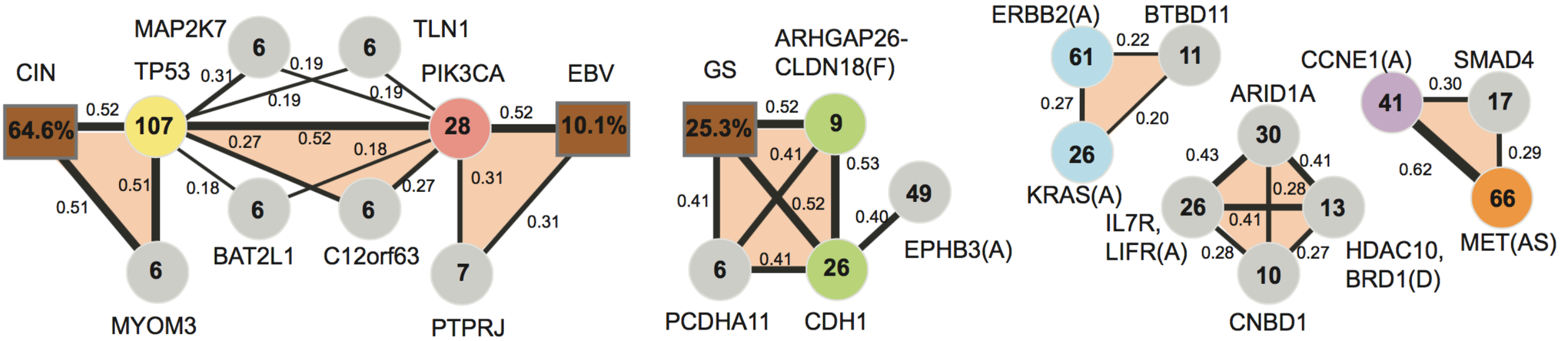






TCGA Gastric cancer with subtypes (STAD)

217 patients and 397 genes / categories, t=4, k=4

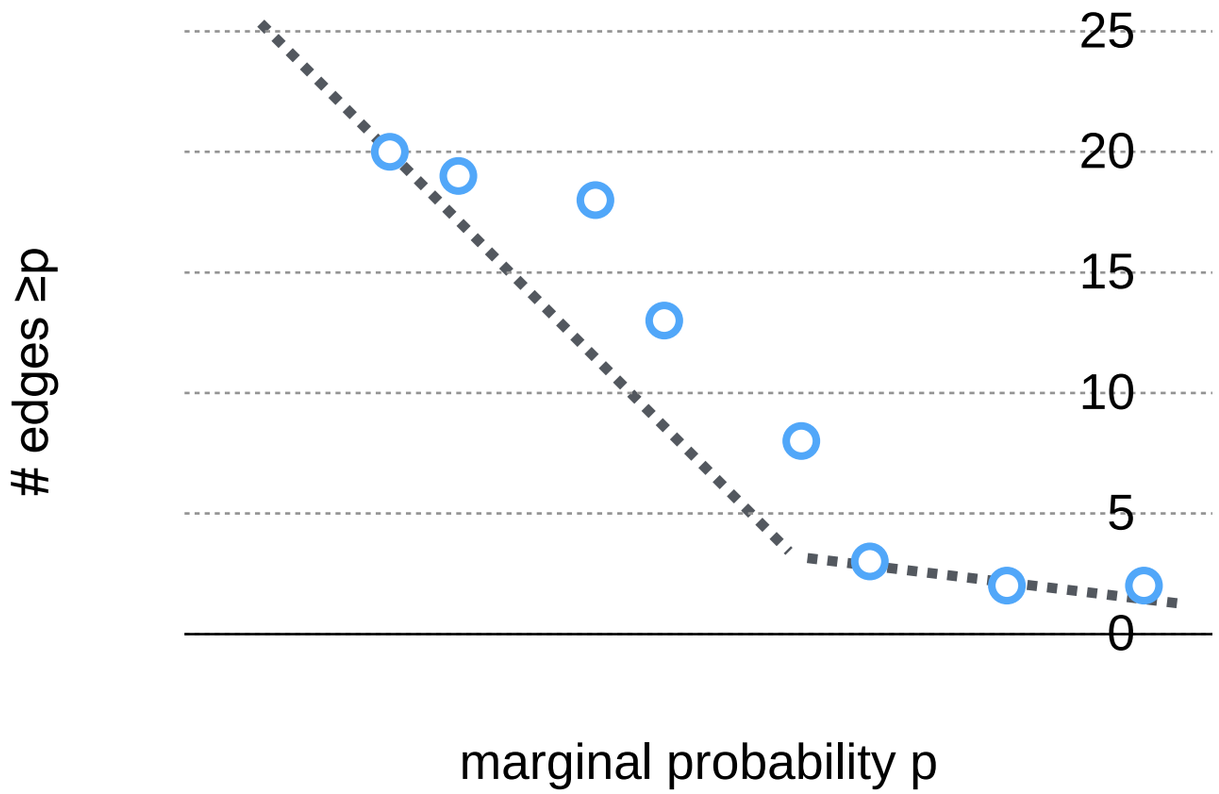


Legend

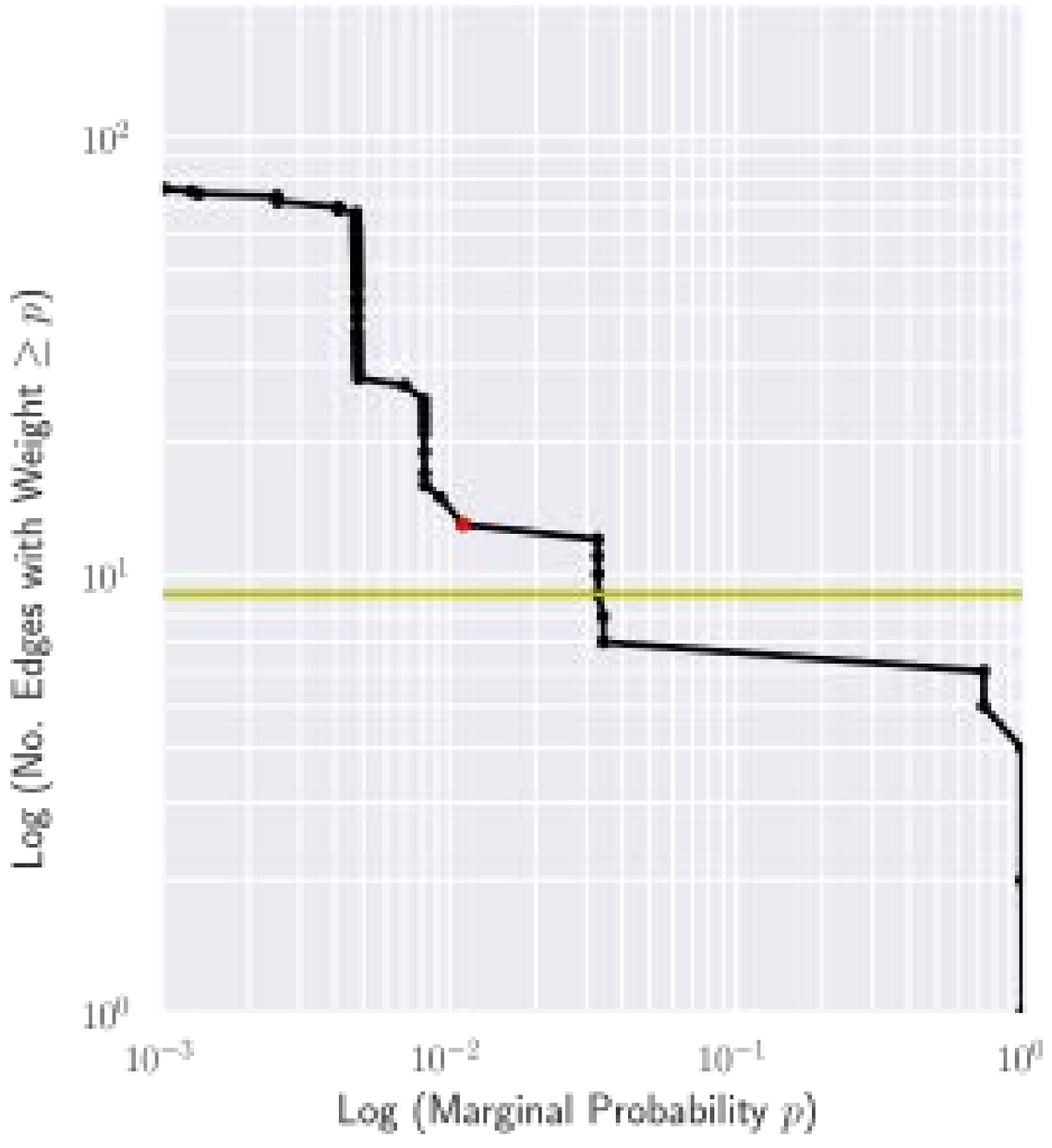
p53 Signaling · RTK/RAS Signaling · PI(3)K/Akt Signaling · Other · Probability of observing pair together

Cell cyle mediators · Genomically stable gastric cancer · c-MET Signaling · Subtype · Top scoring gene set

Delta selection - finding L-corner of edge distribution



Delta selection - finding L-corner of edge distribution



Comparison of **run times** among CoMEt, Dendrix and muex

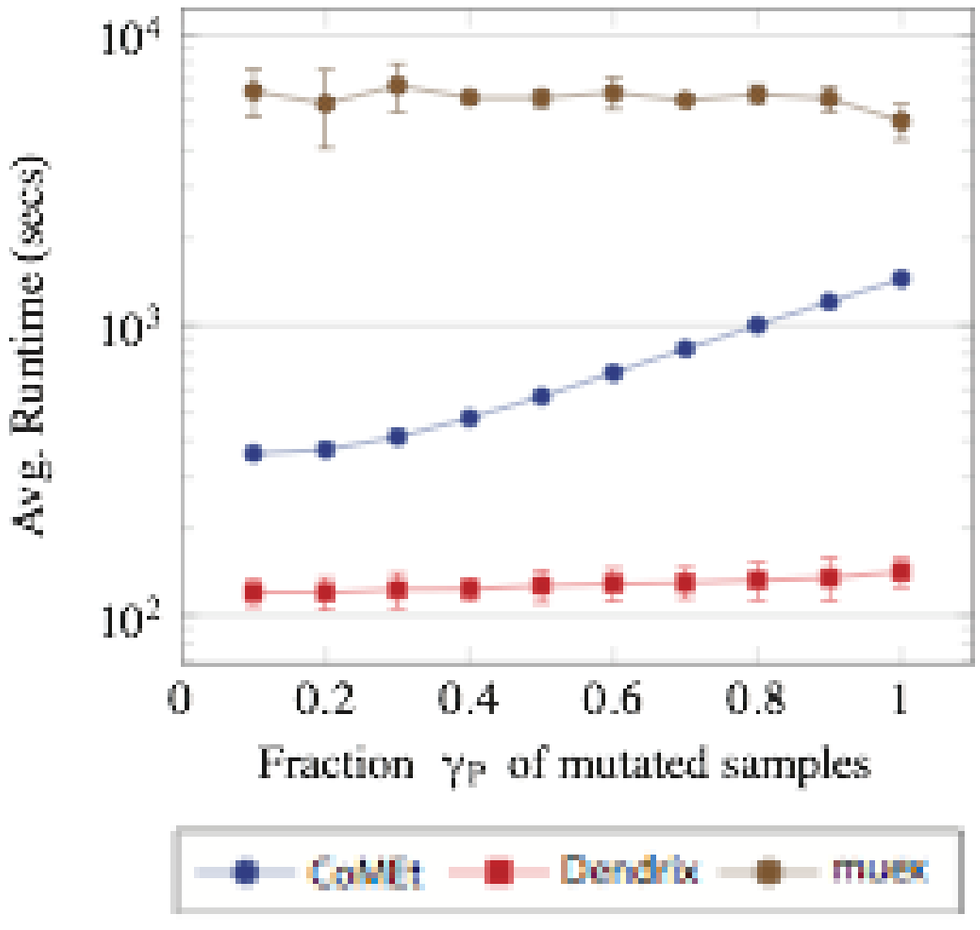
Simulated data:

- One implanted pathway ***P*** of three genes with coverage $n\gamma_p$, where the proportion of mutations in each gene in P is given by

  $\mu_P = (0.5, 0.35, 0.15)$.

- Implant **5 highly altered genes** by randomly selecting samples to be mutated.

- Introduces **NOISES** into simulated dataset by fixed probability.

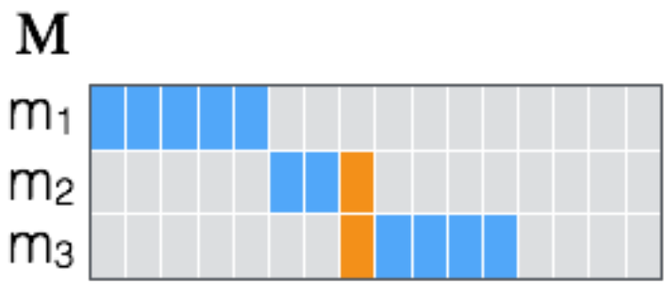


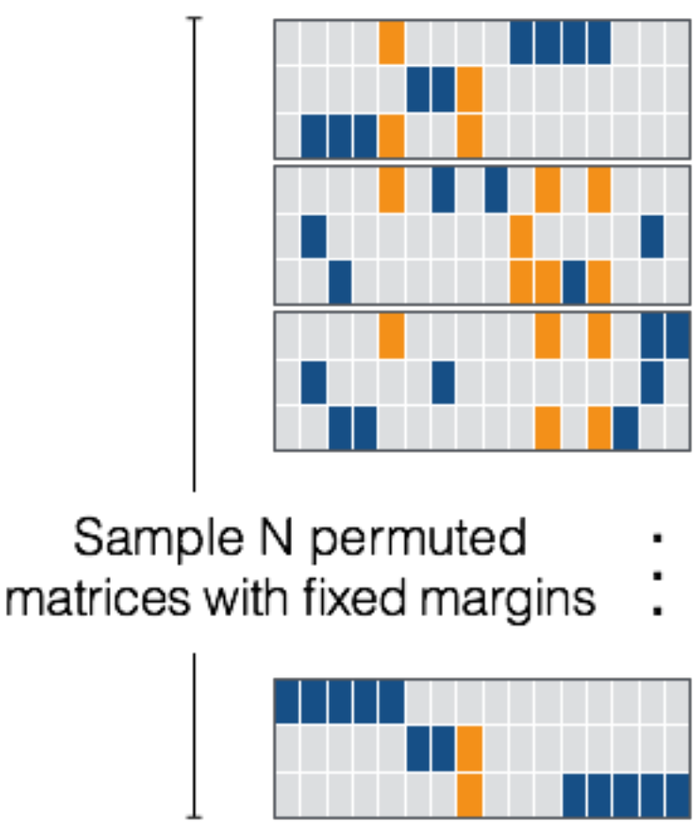
Dendrix: Vandin et al. ***RECOMB***. 2011

muex: Szczurek et al. ***RECOMB***. 2014

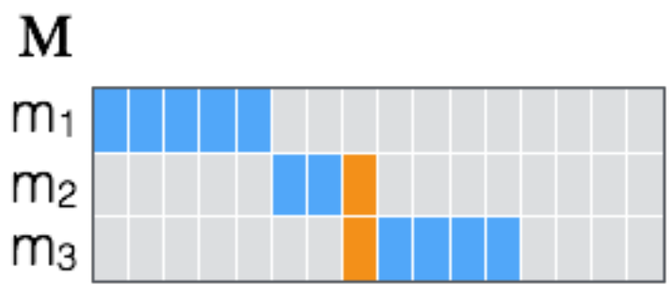
Approximations for higher k

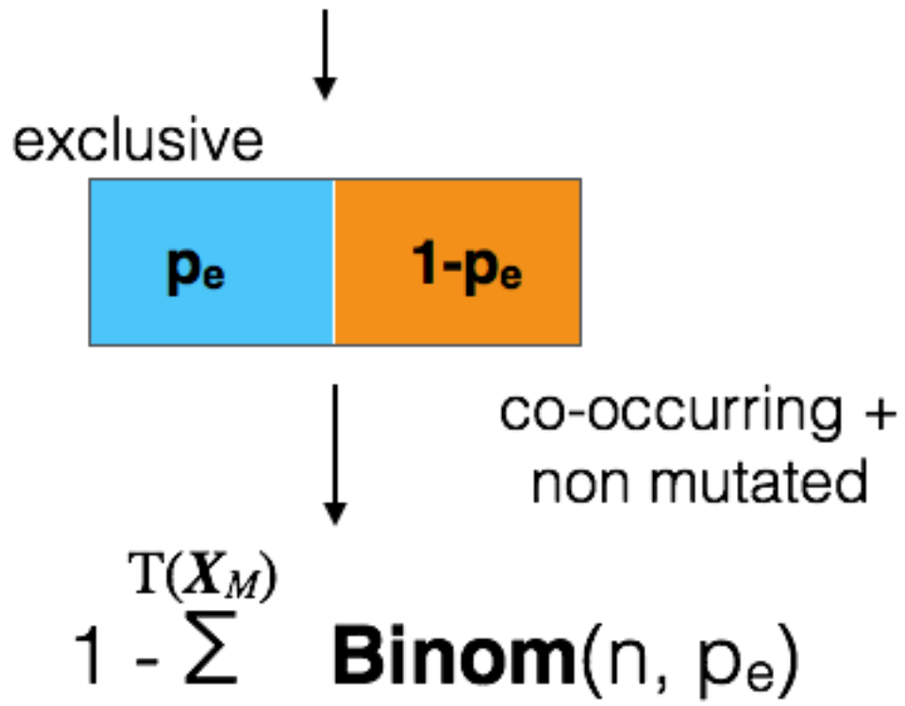
Get null prob. from permuted mutation data in M





Binomial approximation





De novo driver exclusivity (Dendrix)

Given:

Binary mutation matrix **A**

Find: A combination ***M*** of genes with

Approx. Exclusivity: most patients have ≤1 mutation in ***M***

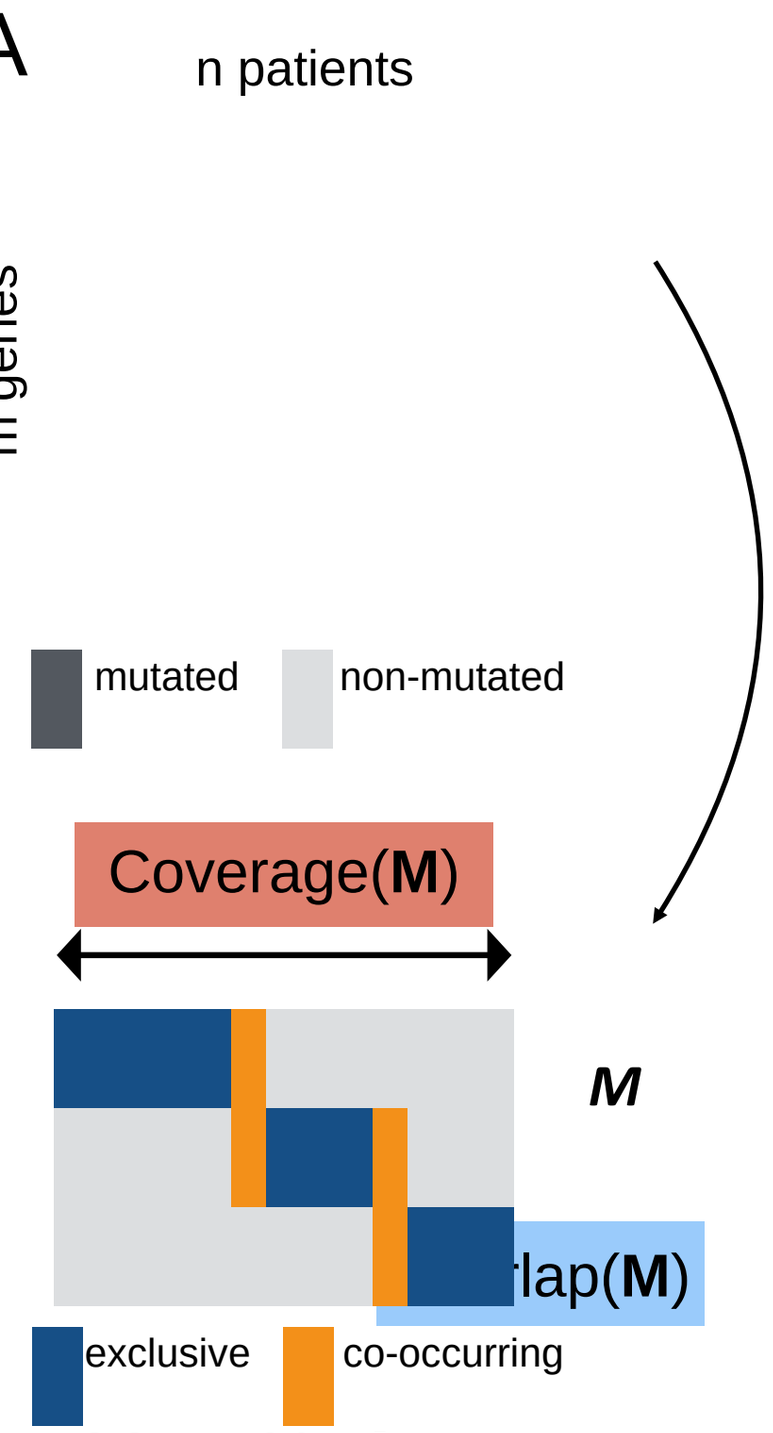
Coverage: most patients have ≥ 1 mutation in ***M***

W(**M**) = Coverage(**M**) - Overlap(**M**)

$$max_{\mathbf{M}}\ W(\mathbf{M})$$

Finding optimal combination is NP-Hard.

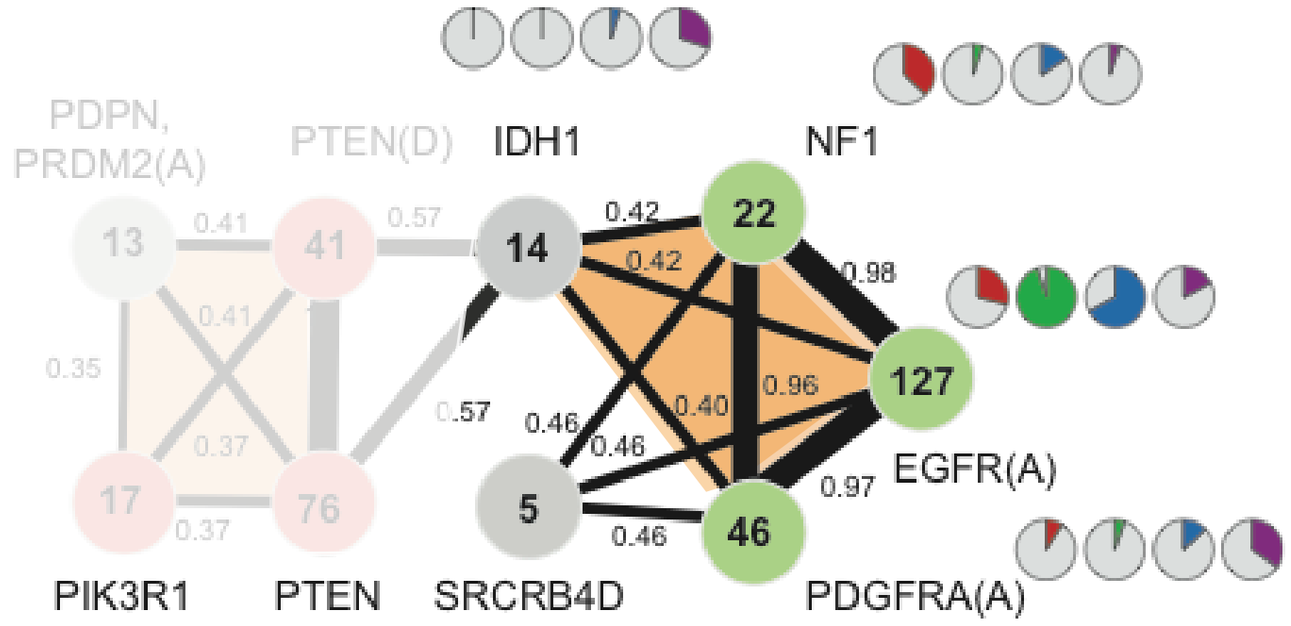
MCMC algorithm samples combinations in proportion to W(**M**)



Vandin, Upfal, & Raphael. *Genome Res*. (2012) [Also *RECOMB 2011*].
Kandoth, et al. *Nature* (2013). Pan-Cancer analysis.

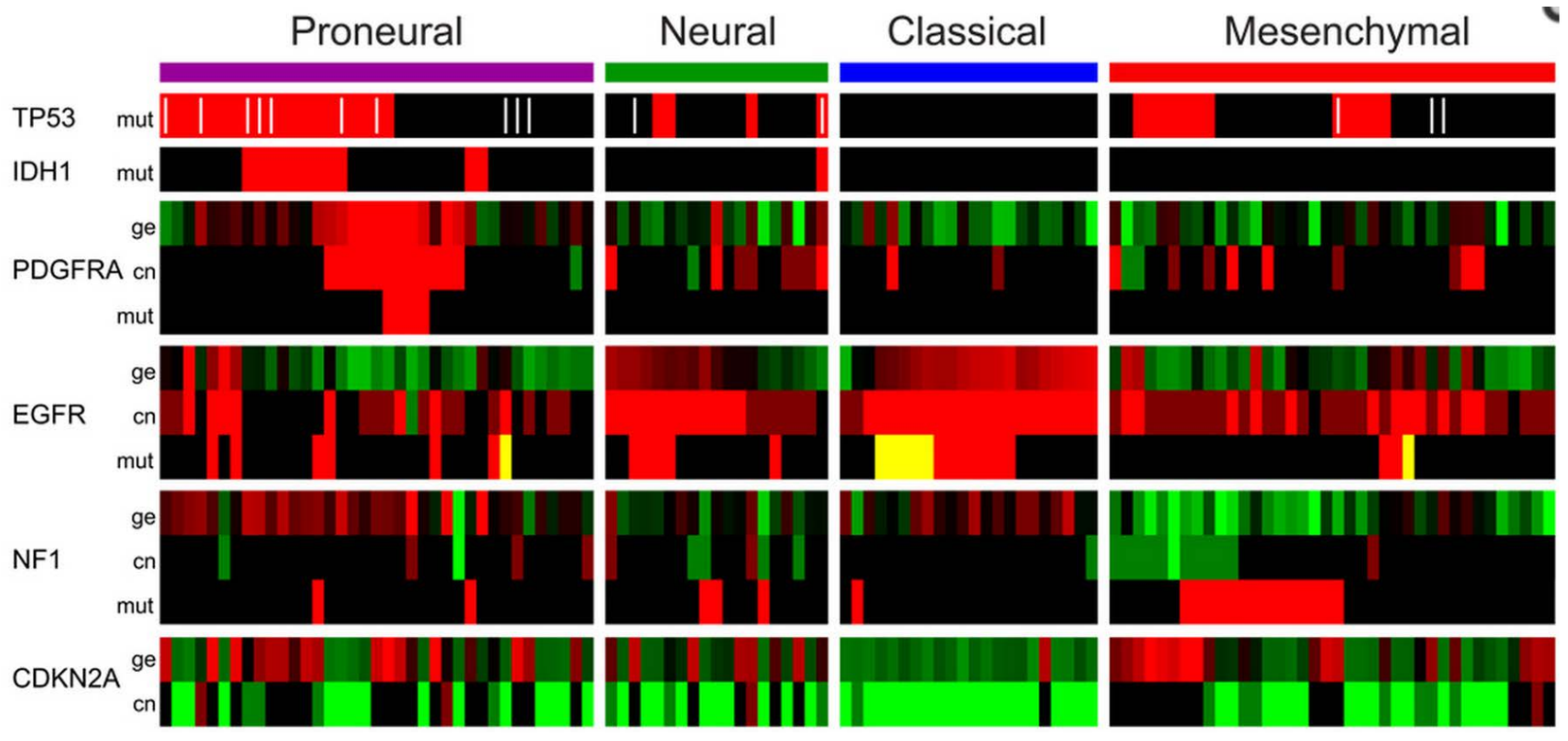
CoMEt module reveals subtype-specific mutations in GBM

CoMEt modules



GBM molecular subtypes

(Verhaak et al. 2010)



mutations	Luminal A	Luminal B	Basal-like	HER2-enriched
TP53	12%	32%	**84%**	75%
PIK3CA	**50%**	32%	7%	42%
ERBB2 amp	-	-	-	**71%**
CCND1 amp	29%	**58%**	-	38%