Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes

Nature Genetics **47**, 106-114 (2015) doi:10.1038/ng.3168

Max Leiserson

RECOMB 2015
April 14, 2015

Raphael Lab
Department of Computer Science &
Center for Computational Molecular Biology

Identifying cancer driver genes

Science
AAAS

Cancer Genome Landscapes
Vogelstein *et al*. (2013)

“>99.9% of mutations are passengers”

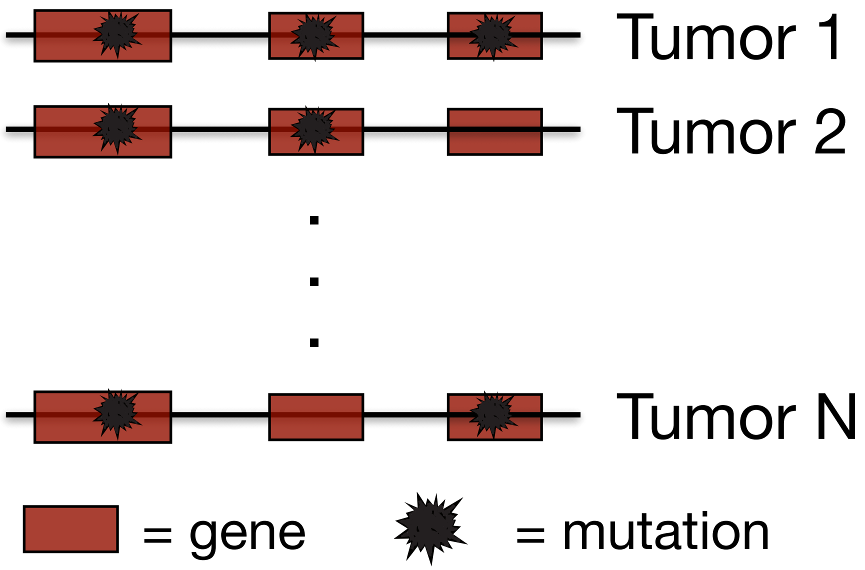
“3-8 drivers per tumor"

Identifying cancer driver genes

Cancer Genome Landscapes
Vogelstein *et al*. (2013)

“>99.9% of mutations are passengers”
“3-8 drivers per tumor"

Compare variation across tumors



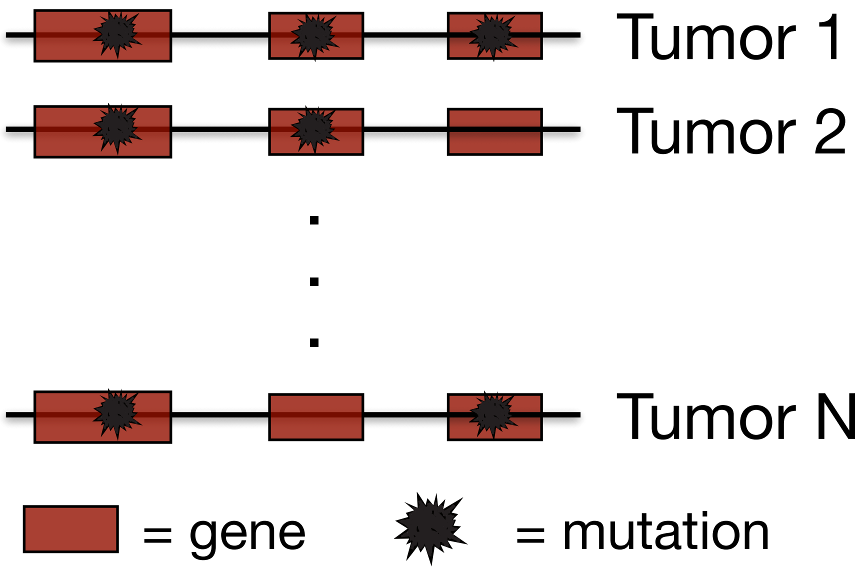
- Single nucleotide variants
- Copy number aberrations
- Gene expression
- ….

Identifying cancer driver genes

Cancer Genome Landscapes
Vogelstein *et al*. (2013)

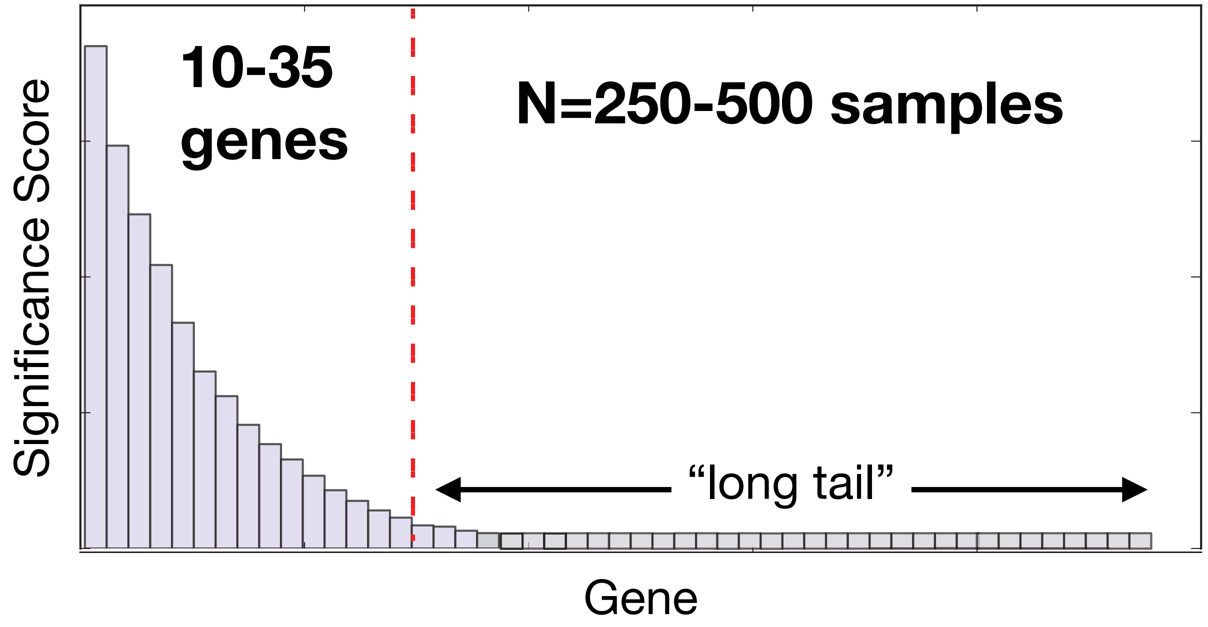
“>99.9% of mutations are passengers”
“3-8 drivers per tumor"

Compare variation across tumors



- Single nucleotide variants
- Copy number aberrations
- Gene expression
- ….

Identify cancer driver genes



Mutations weighted by:

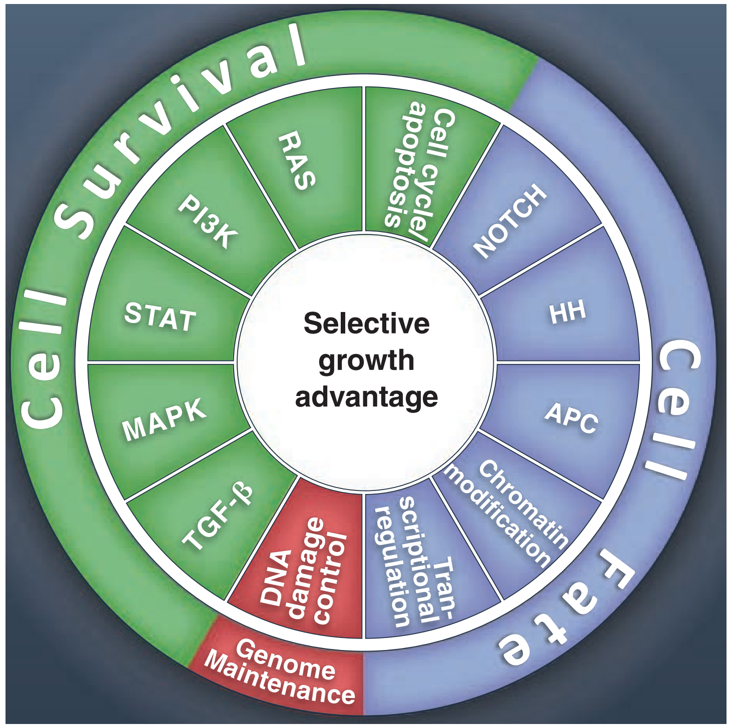
- Recurrence
- Gene length
- Mutation context
- Expression level
- Replication timing
- …

Cancer driver mutations target *pathways*

Driver mutations confer a growth advantage to the tumor → driver genes are members of cancer signaling pathways

Cancer driver mutations target *pathways*

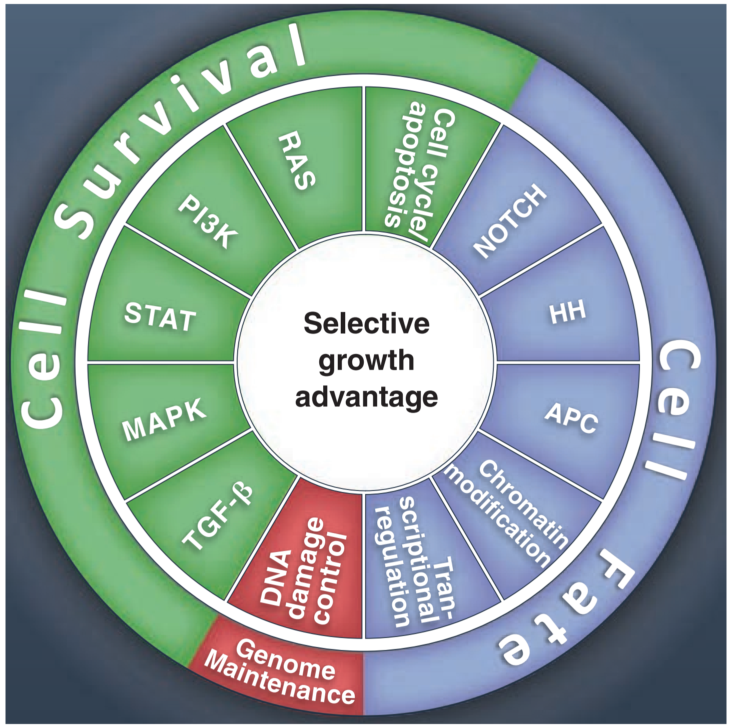
Driver mutations confer a growth advantage to the tumor → driver genes are members of cancer signaling pathways



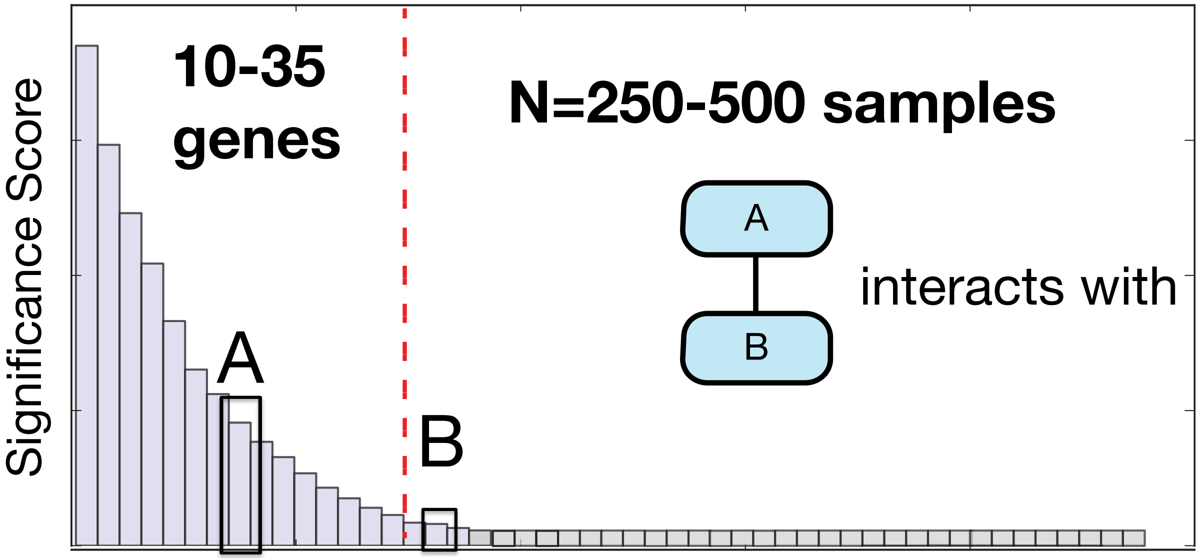
Vogelstein et al. (*Science* 2013)

Cancer driver mutations target *pathways*

Driver mutations confer a growth advantage to the tumor → driver genes are members of cancer signaling pathways

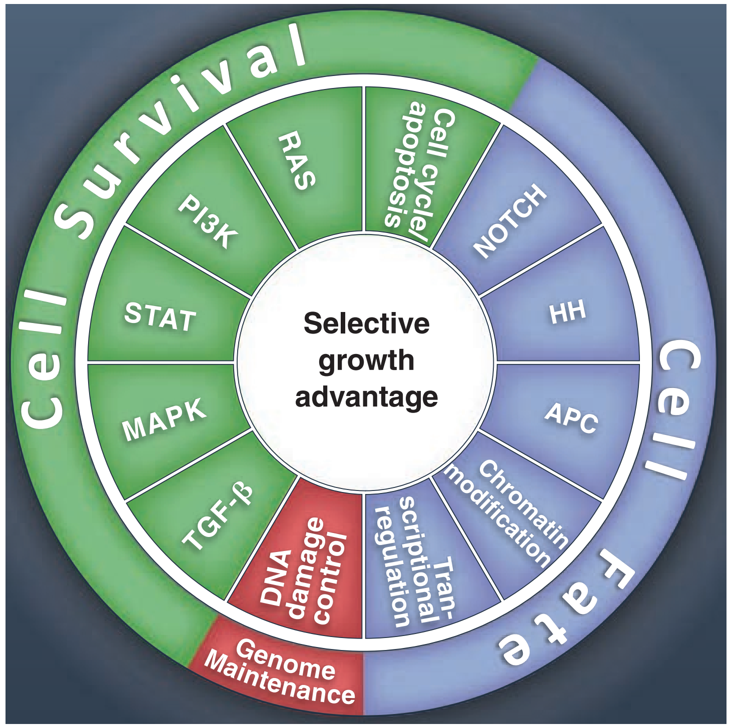


Vogelstein et al. (*Science* 2013)

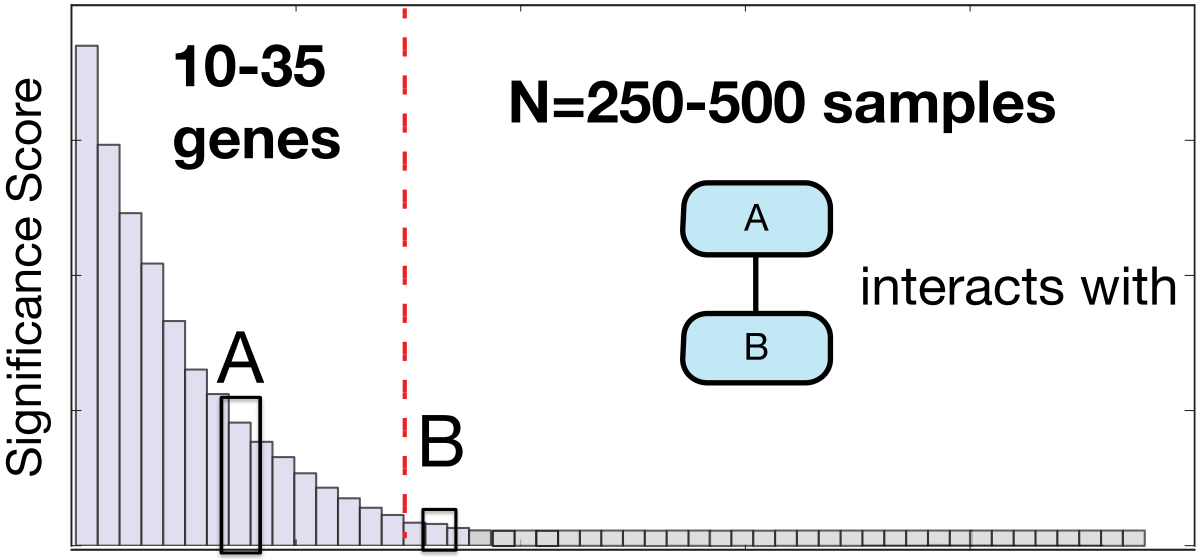


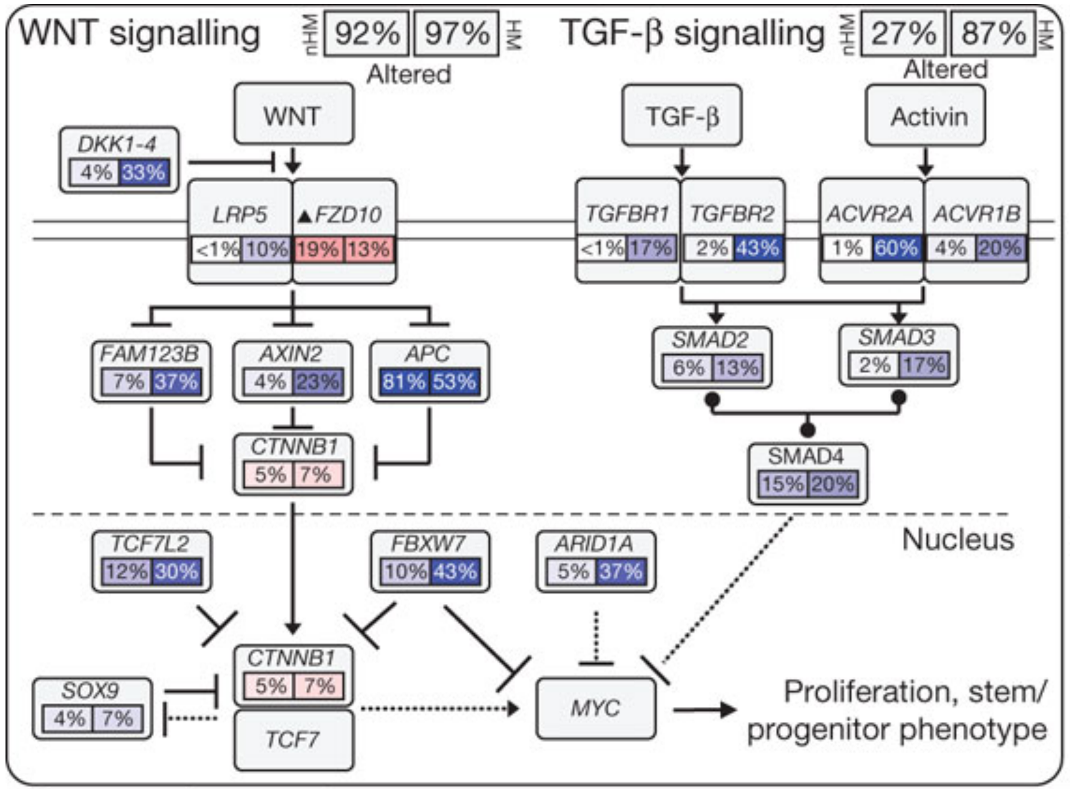
Cancer driver mutations target *pathways*

Driver mutations confer a growth advantage to the tumor → driver genes are members of cancer signaling pathways



Vogelstein et al. (*Science* 2013)



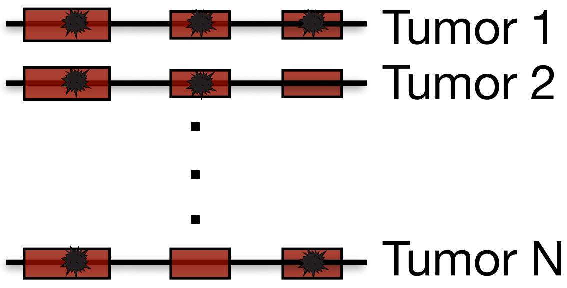


TCGA Ovarian (*Nature* 2011)

Testing known gene sets and pathways

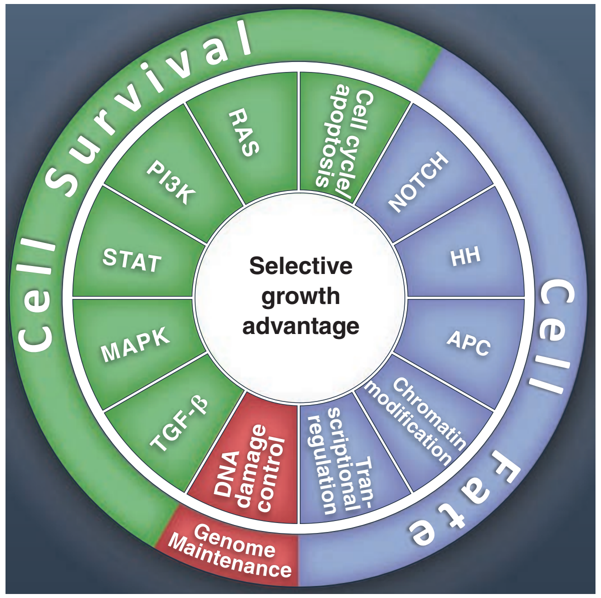
Input data

Mutation data



(e.g. most mutated genes: EGFR, KRAS, BRAF)

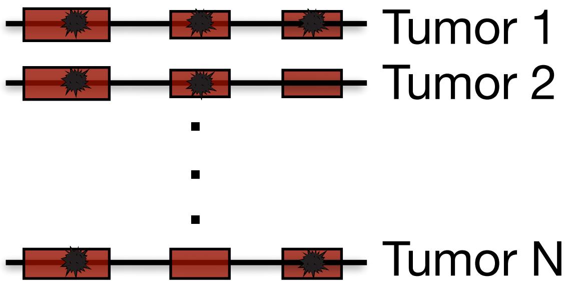
Gene set database



Testing known gene sets and pathways

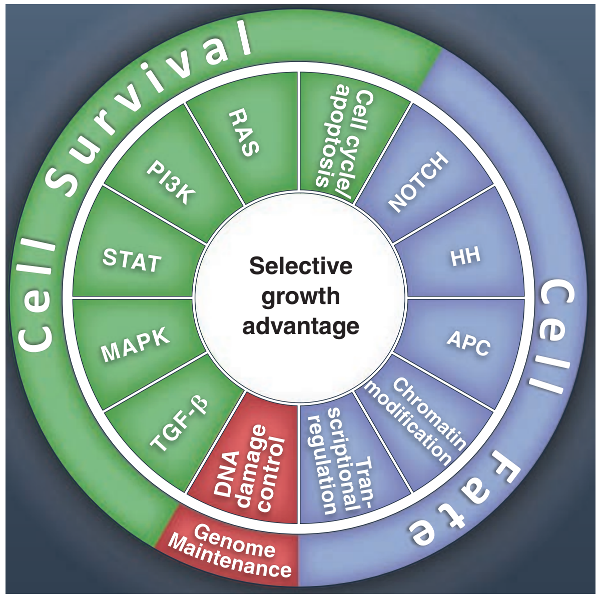
Input data

Mutation data



(e.g. most mutated genes: EGFR, KRAS, BRAF)

Gene set database



Enrichment tests

GSEA [1,2]

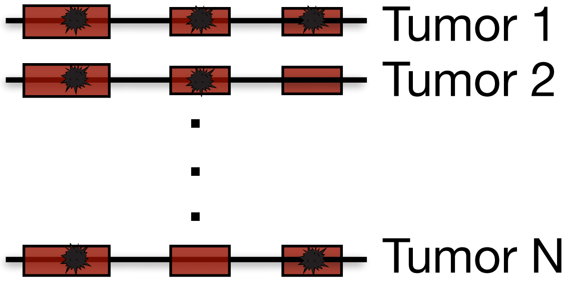
DAVID [3,4]

[1] Mootha *et al. Nat. Genet.* (2003).
[2] Subramanian *et al. PNAS* (2005).
[3] Huang *et al. Nat. Protoc.* (2009).
[4] Huang *et al. Nucleic Acids Res.* (2009)

Testing known gene sets and pathways

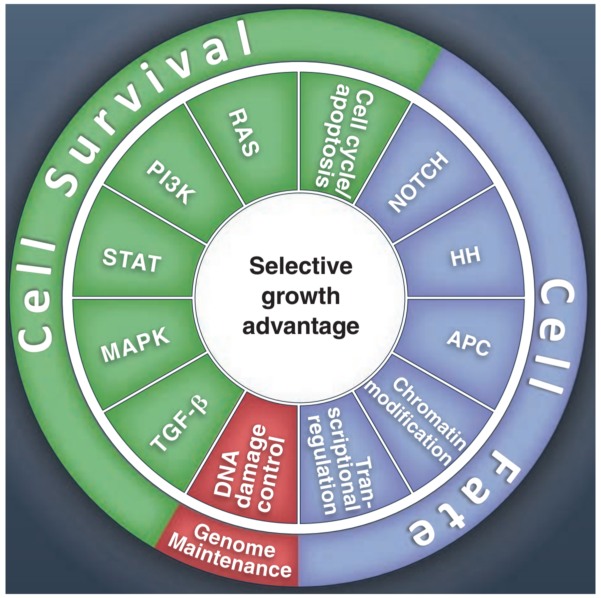
Input data

Mutation data



(e.g. most mutated genes:
EGFR, KRAS, BRAF)

Gene set database



Enrichment tests

GSEA [1,2]

DAVID [3,4]

Enriched gene sets

MYC ARID1A
FBXW7 APC
SOX9 CTNNB1

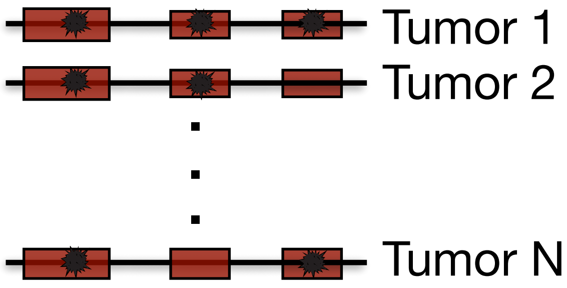
⋮

[1] Mootha *et al. Nat. Genet.* (2003).
[2] Subramanian *et al. PNAS* (2005).
[3] Huang *et al. Nat. Protoc.* (2009).
[4] Huang *et al. Nucleic Acids Res.* (2009)

Testing known gene sets and pathways

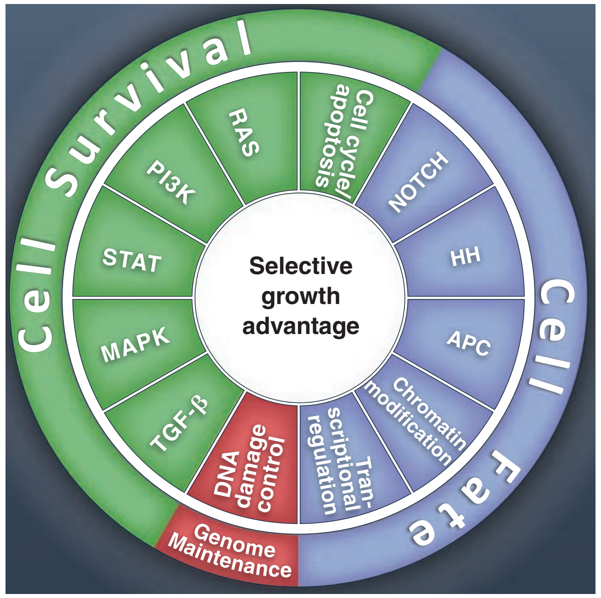
Input data

Mutation data



(e.g. most mutated genes: EGFR, KRAS, BRAF)

Gene set database



Enrichment tests

GSEA [1,2]

DAVID [3,4]

Enriched gene sets

MYC ARID1A

FBXW7 APC

SOX9 CTNNB1

⋮

Key drawbacks

- Novel pathways and *crosstalk*?
- Topology of interactions?
- Handling large and/or overlapping pathways?

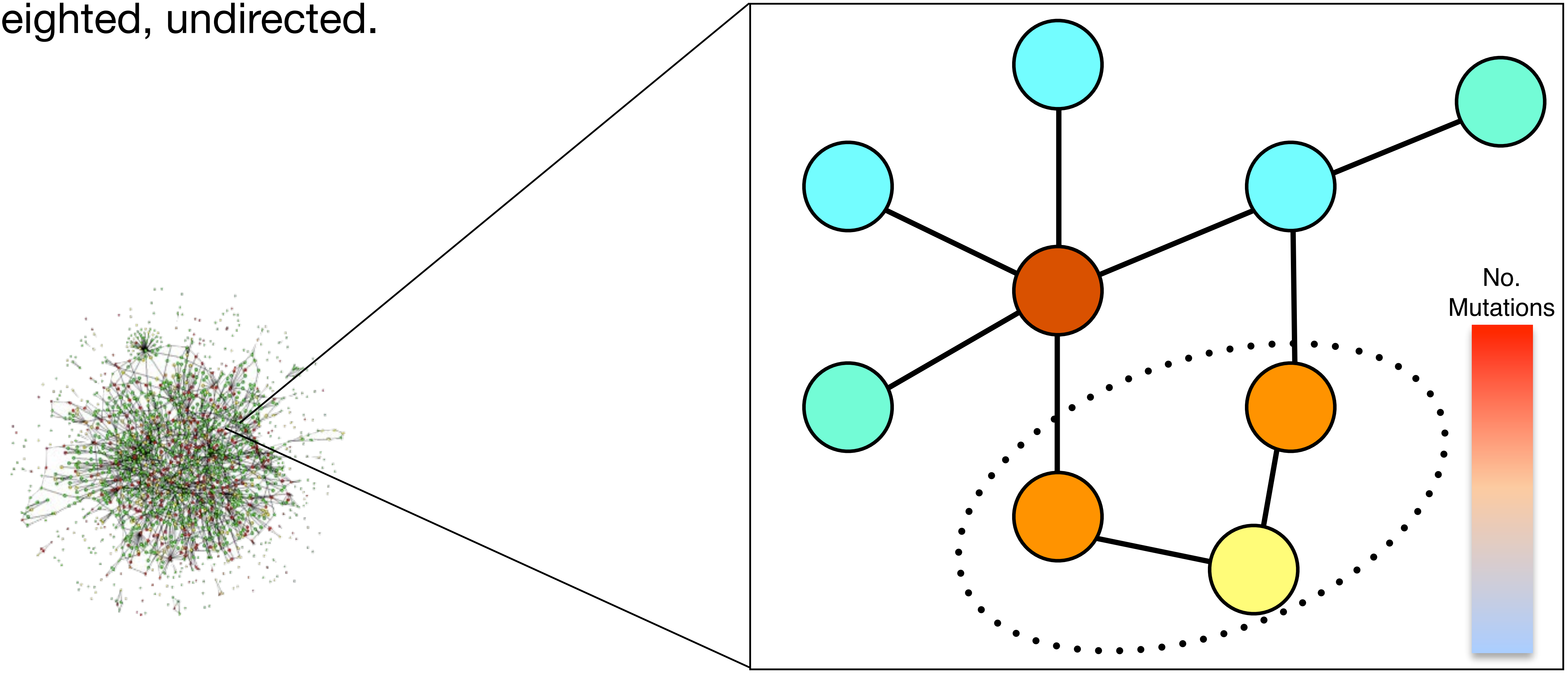
[1] Mootha *et al. Nat. Genet.* (2003).
[2] Subramanian *et al. PNAS* (2005).
[3] Huang *et al. Nat. Protoc.* (2009).
[4] Huang *et al. Nucleic Acids Res.* (2009)

Significantly mutated subnetworks of a protein-protein interaction network

Protein-protein interaction networks

- Nodes: genes/protein
- Edges: connect genes if the proteins they encode physically interact.
- Unweighted, undirected.

Goal: identify connected subnetworks with more mutations than expected by chance.

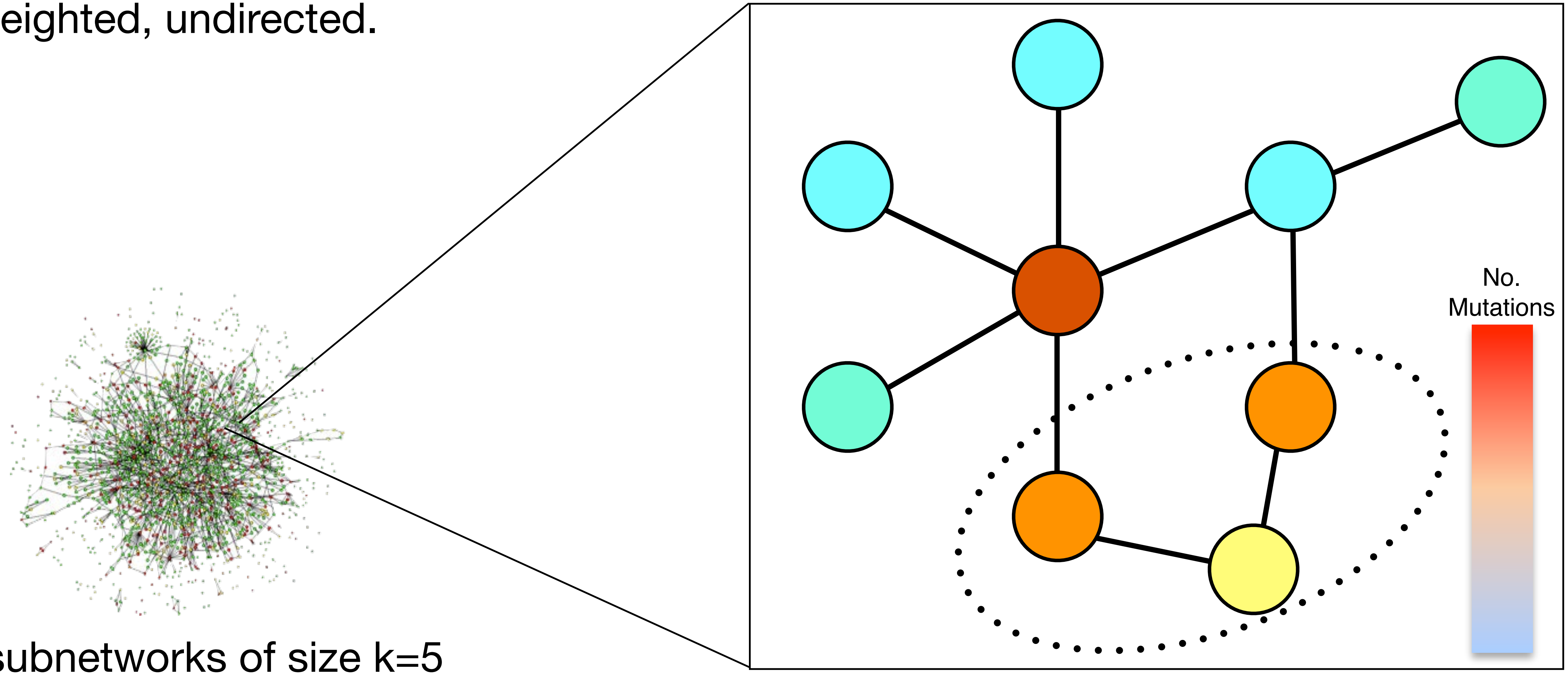


Significantly mutated subnetworks of a protein-protein interaction network

Protein-protein interaction networks

- Nodes: genes/protein
- Edges: connect genes if the proteins they encode physically interact.
- Unweighted, undirected.

Goal: identify connected subnetworks with more mutations than expected by chance.



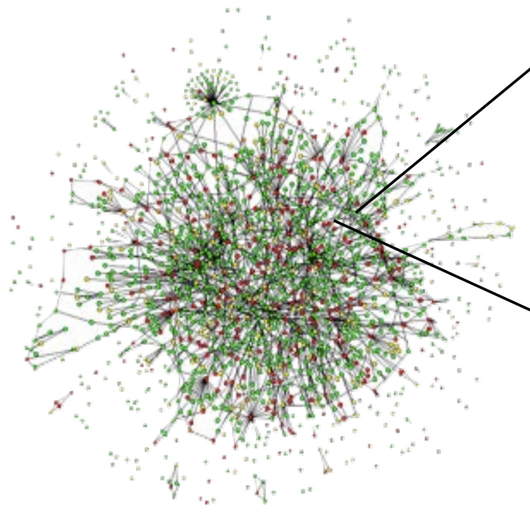
~10^{18} subnetworks of size k=5

Significantly mutated subnetworks of a protein-protein interaction network

Protein-protein interaction networks

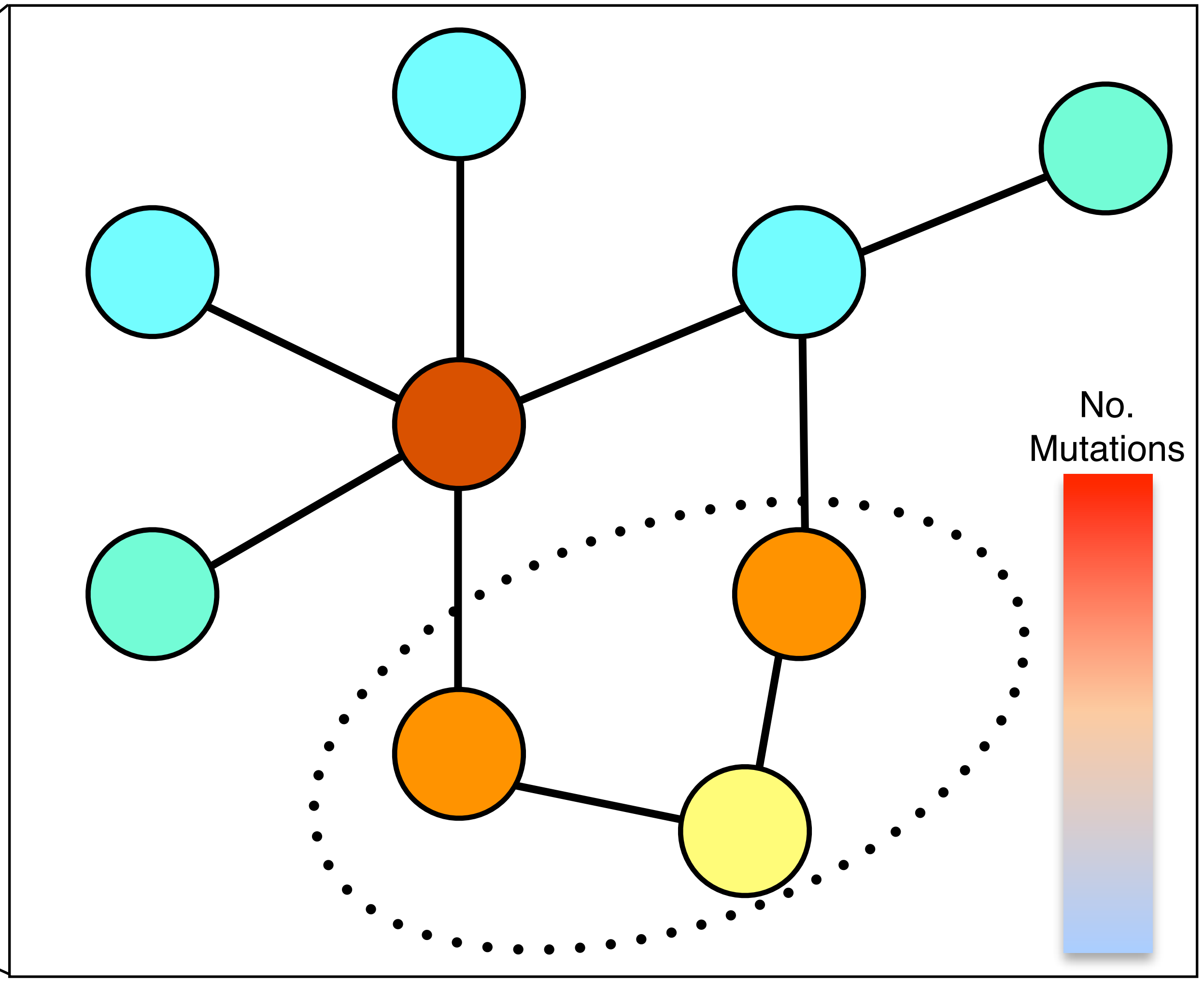
Network	Nodes	Edges	Diameter	ASP
HPRD	9,205	36,720	**14**	**4.22**
HINT+HI2012	9,859	40,705	**14**	**4.08**
iRefIndex	12,129	91,809	**12**	**3.64**
MultiNet	14,399	109,570	**9**	**3.39**

Low diameter → Most genes have a high-scoring neighbor



~10^{18} subnetworks of size k=5

Goal: identify connected subnetworks with more mutations than expected by chance.



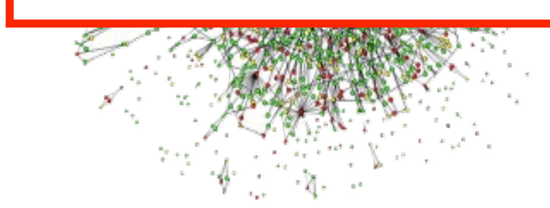
Significantly mutated subnetworks of a protein-protein interaction network

Protein-protein interaction networks

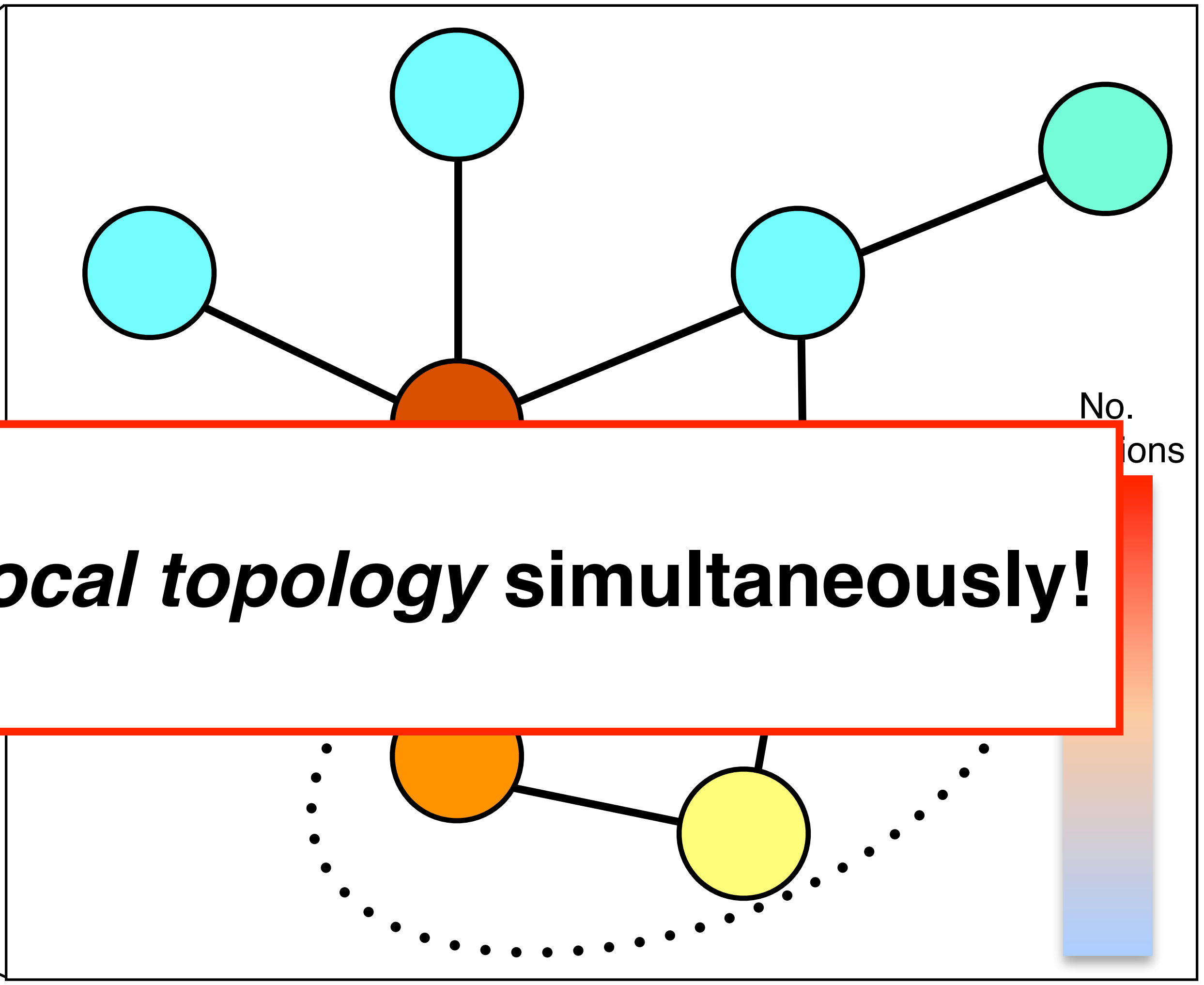
Network	Nodes	Edges	Diameter	ASP
HPRD	9,205	36,720	**14**	**4.22**
HINT+HI2012	9,859	40,705	**14**	**4.08**
iRefIndex	12,129	91,809	**12**	**3.64**
MultiNet	14,399	109,570	**9**	**3.39**

Low diameter → Most genes have a high-scoring neighbor

Goal: identify connected subnetworks with more mutations than expected by chance.



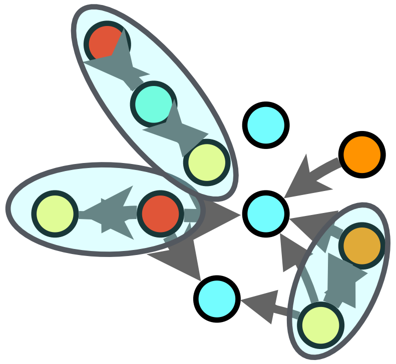
~10^{18} subnetworks of size k=5



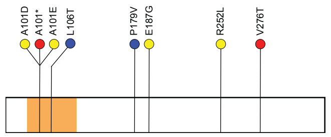
Must analyze mutations and *local topology* simultaneously!

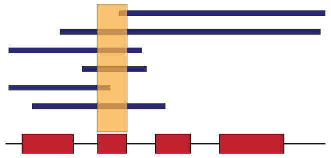
Outline

1. **A new algorithm, HotNet2.**



2. Application to TCGA Pan-Cancer data.

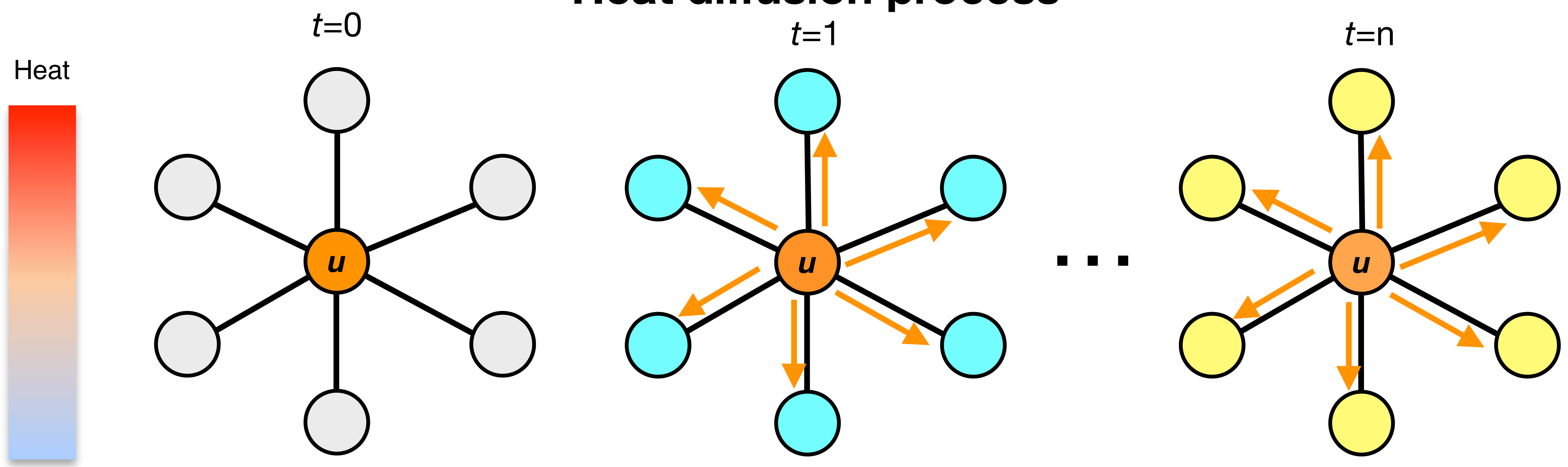




3. Comparison of HotNet2 to other methods.

Encoding mutations and graph topology with heat diffusion

Heat diffusion process

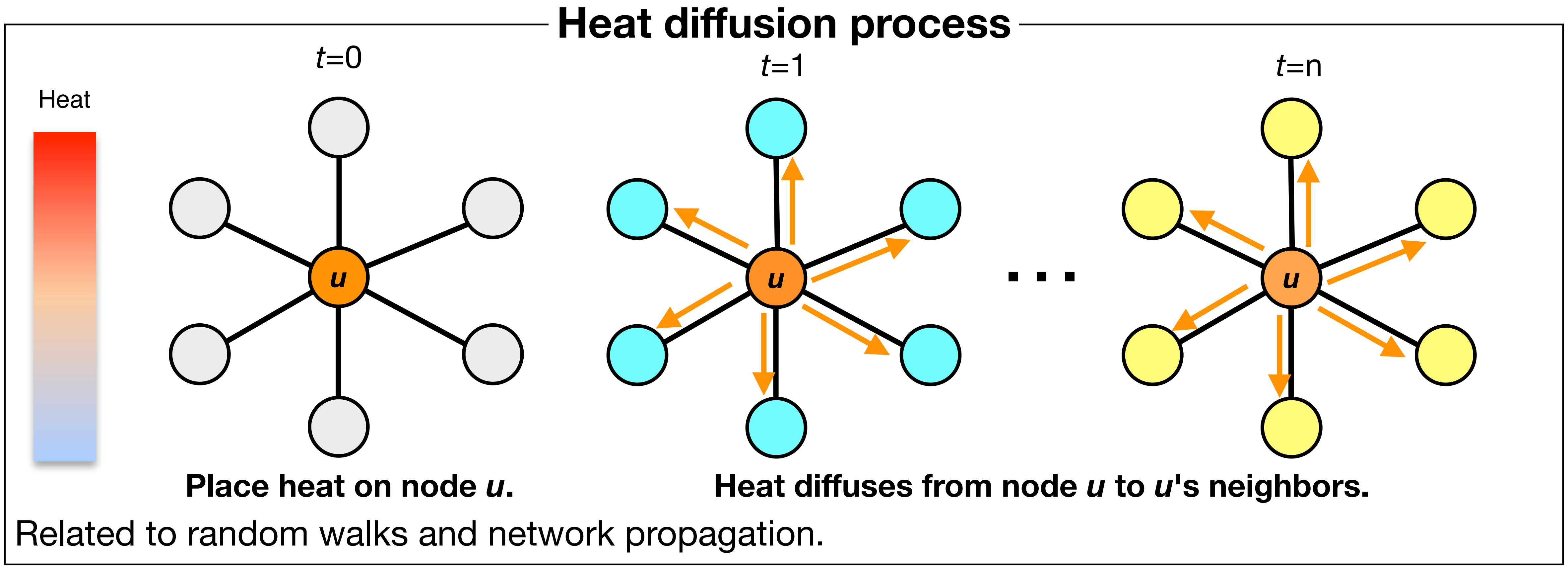


Place heat on node *u*.

Heat diffuses from node *u* to *u*'s neighbors.

Related to random walks and network propagation.

Encoding mutations and graph topology with heat diffusion



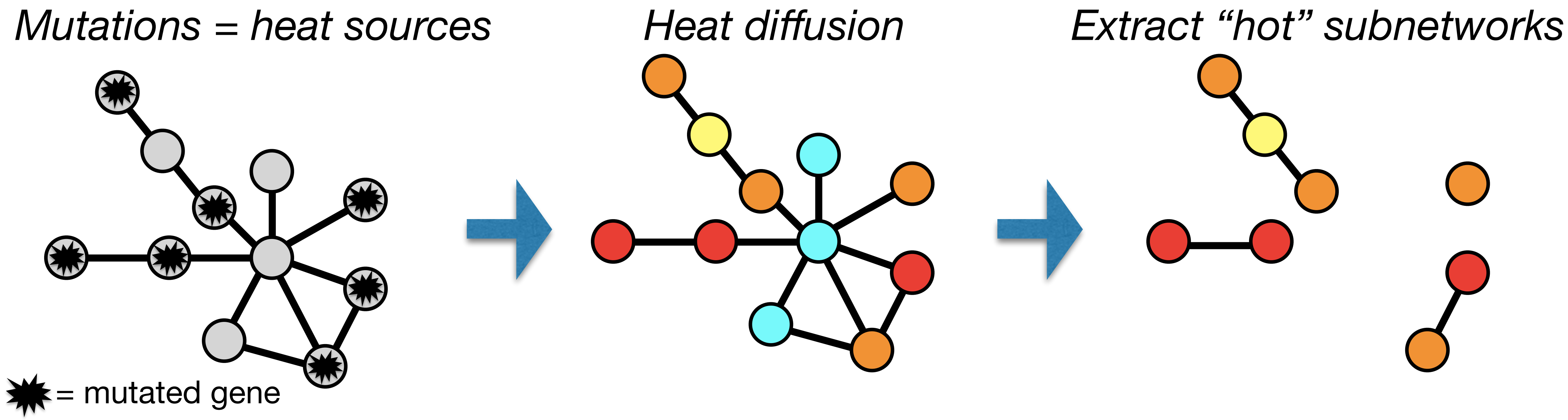
Heat diffusion process

Place heat on node *u*.

Heat diffuses from node *u* to *u*'s neighbors.

Related to random walks and network propagation.

HotNet (Vandin *et al*. JCB & *RECOMB* 2010)



Mutations = heat sources

Heat diffusion

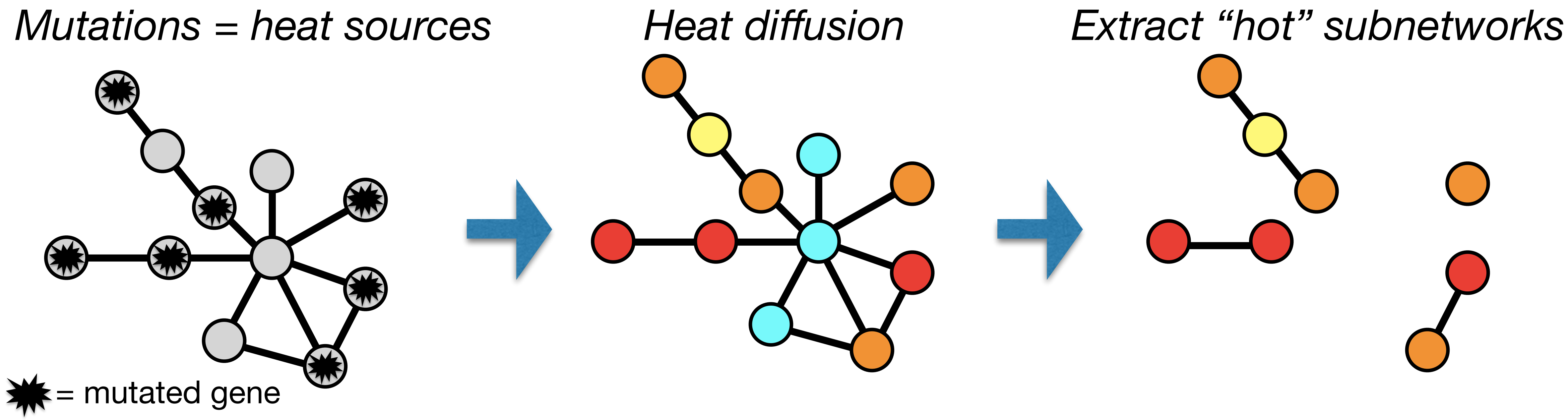
Extract “hot” subnetworks

HotNet applied to TCGA data

TCGA Papers
(~300 samples)

- Leukemia (*NEJM* 2013)
- Kidney (*Nature* 2011)
- Ovarian (*Nature* 2011)

HotNet (Vandin *et al*. JCB & *RECOMB* 2010)



Mutations = heat sources

Heat diffusion

Extract "hot" subnetworks

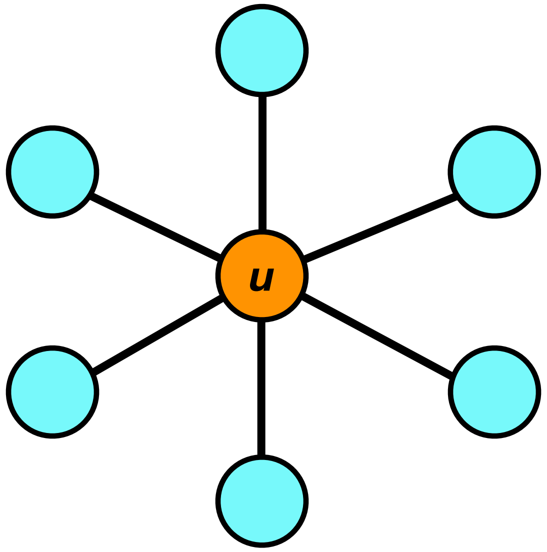
✸ = mutated gene

HotNet applied to TCGA data

TCGA Papers
(~300 samples)

- Leukemia (*NEJM* 2013)
- Kidney (*Nature* 2011)
- Ovarian (*Nature* 2011)

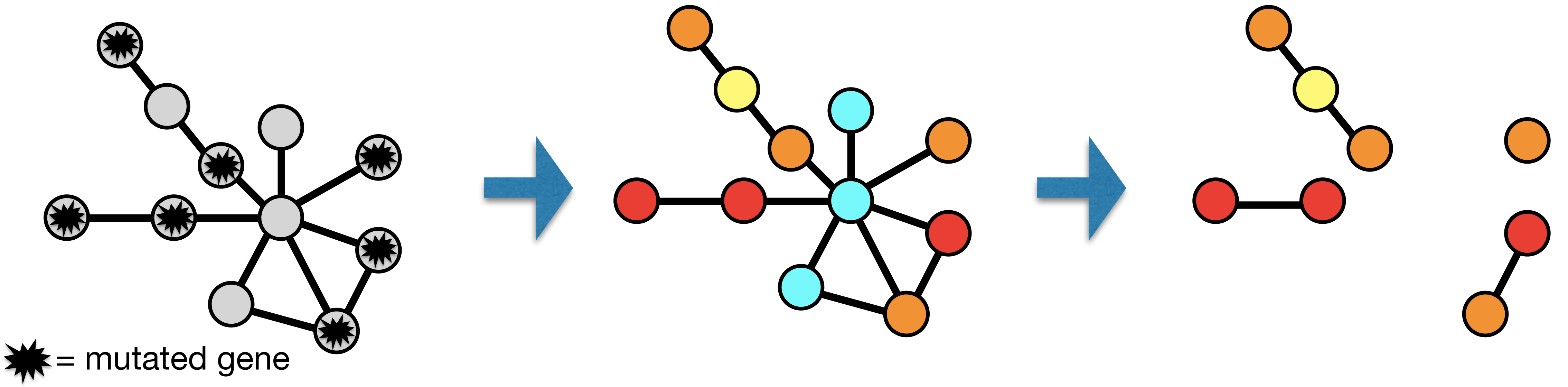
TCGA Pan-Cancer (>3000 samples)



HotNet finds many "star" subnetworks with one central, hot node

HotNet (Vandin *et al*. JCB & *RECOMB* 2010)

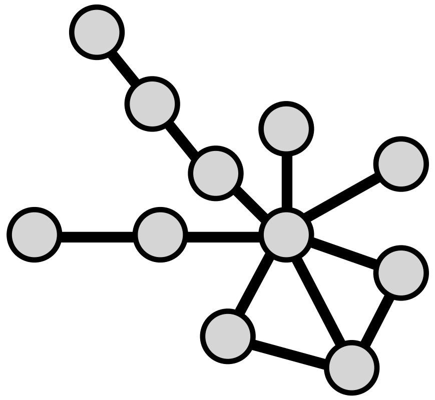
Mutations = heat sources *Heat diffusion* *Extract "hot" subnetworks*



= mutated gene

HotNet Algorithm

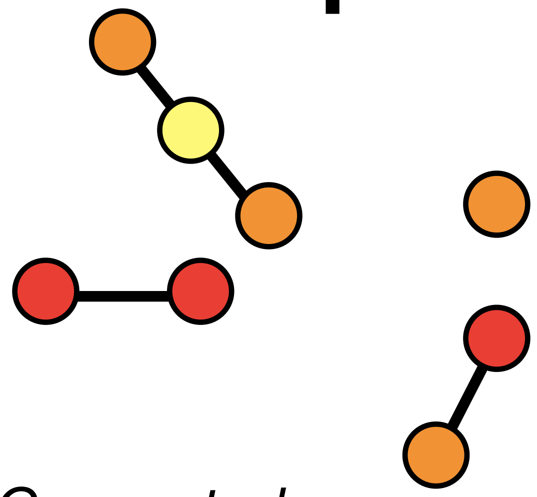
Input



$$\begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}$$

A = adjacency matrix **h** = gene scores

Output



Connected components

Threshold at δ

$$\begin{pmatrix} \text{Heat kernel} \\ f(A, t) \end{pmatrix} \begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_n \end{pmatrix} = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{pmatrix} \xrightarrow{\text{symmetrize}} \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix}$$

Diffusion matrix
*(**symmetric**)*

Similarity matrix
*(**asymmetric**)*

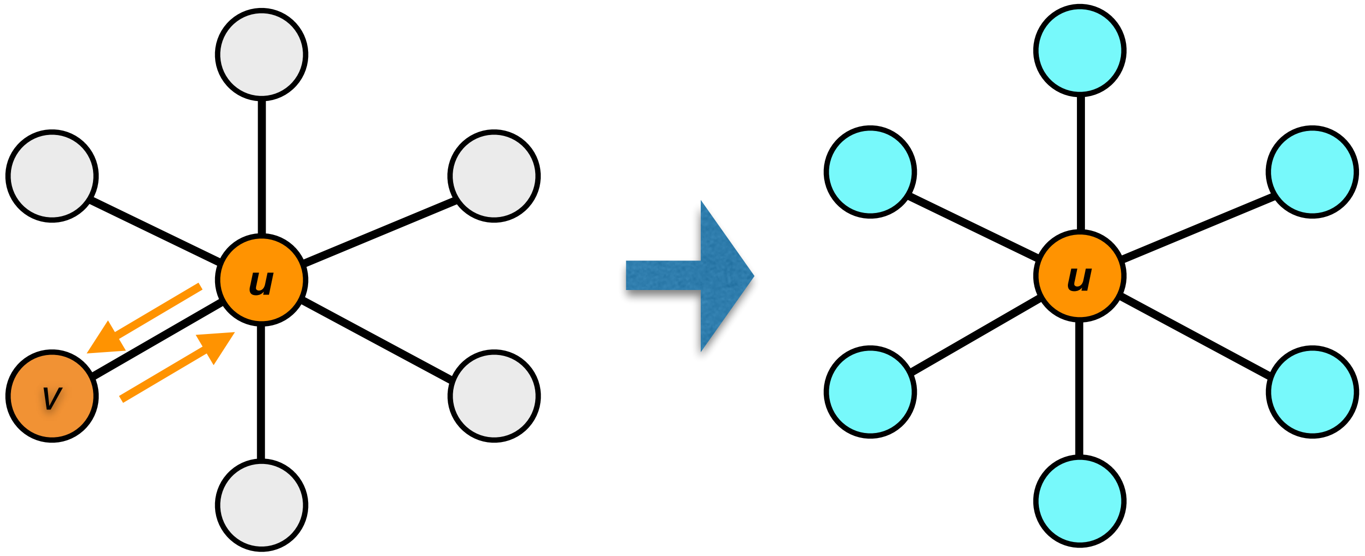
Time parameter t

s_{ij} = heat on vertex *i* at time *t* given initial heat h_j on vertex *j* at time 0.

Direction of heat is important →HotNet can fail

HotNet's heat is symmetric

u sends the same heat to *v* even though *u* has much higher degree



Star graph

Potential artifacts

Hot nodes with high degree often form large "star" subnetworks with many cold nodes

$$\begin{pmatrix} \text{Heat kernel} \\ f(A, t) \end{pmatrix} \begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_n \end{pmatrix} = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{pmatrix} \xrightarrow{\textbf{symmetrize}} \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix}$$

Diffusion matrix
*(**symmetric**)*

Similarity matrix S
*(**asymmetric**)*

s_{ij} = heat on vertex *i* at time *t* given initial heat h_j on vertex *j* at time 0.

HotNet2 algorithm

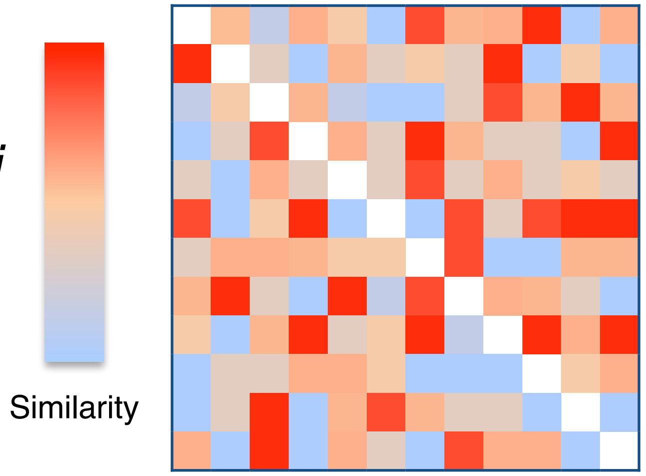
(HotNet diffusion oriented subnetworks)

Need to consider the source of heat

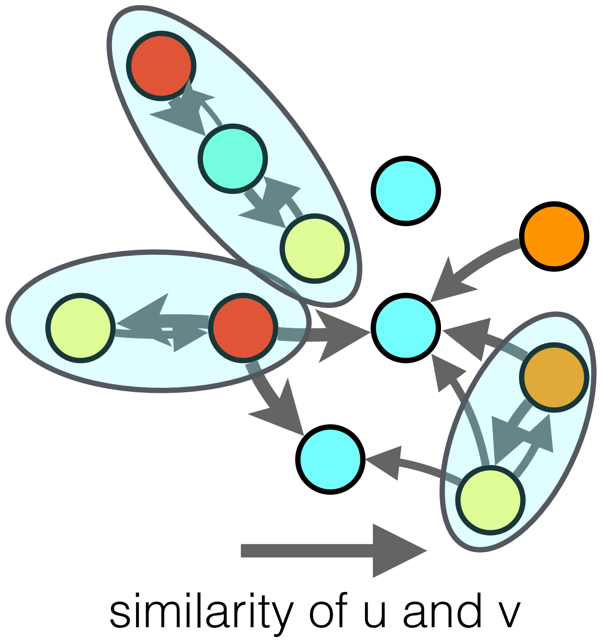
Encode directionality with asymmetric heat diffusion.

- Hot genes do not necessarily implicate their neighbors.
- Hot subnetworks have a directed path between each pair of nodes.

s_{ij} = heat on vertex *i* at *equilibrium* given initial heat h_j on vertex *j* at time 0.



Similarity matrix *S*
Asymmetric!



Identify *strongly* connected components

Leiserson, Vandin *et al. Nat. Genet.* (2015).
http://compbio.cs.brown.edu/projects/hotnet2

HotNet2 algorithm

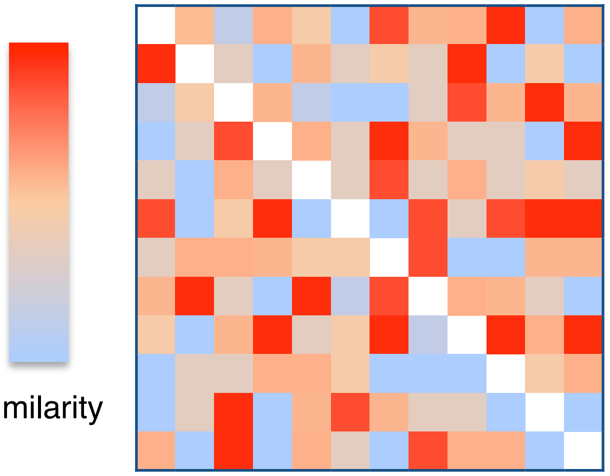
(HotNet diffusion oriented subnetworks)

Need to consider the source of heat

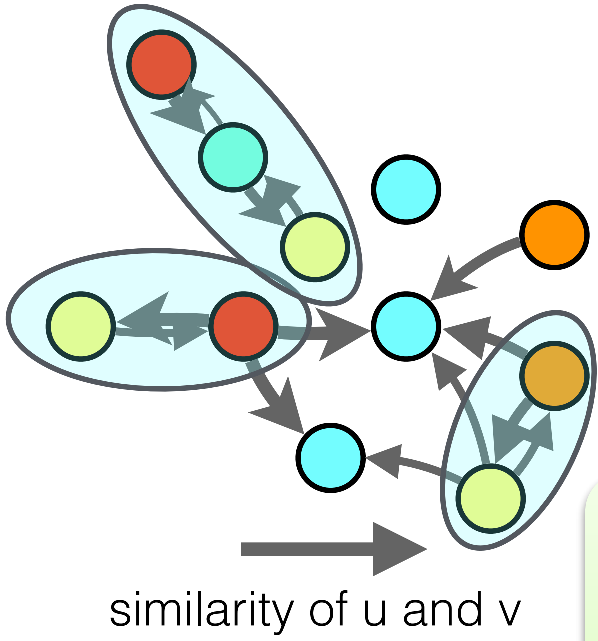
Encode directionality with asymmetric heat diffusion.

- Hot genes do not necessarily implicate their neighbors.
- Hot subnetworks have a directed path between each pair of nodes.

s_{ij} = heat on vertex *i* at *equilibrium* given initial heat h_j on vertex *j* at time 0.



Similarity matrix *S*
Asymmetric!



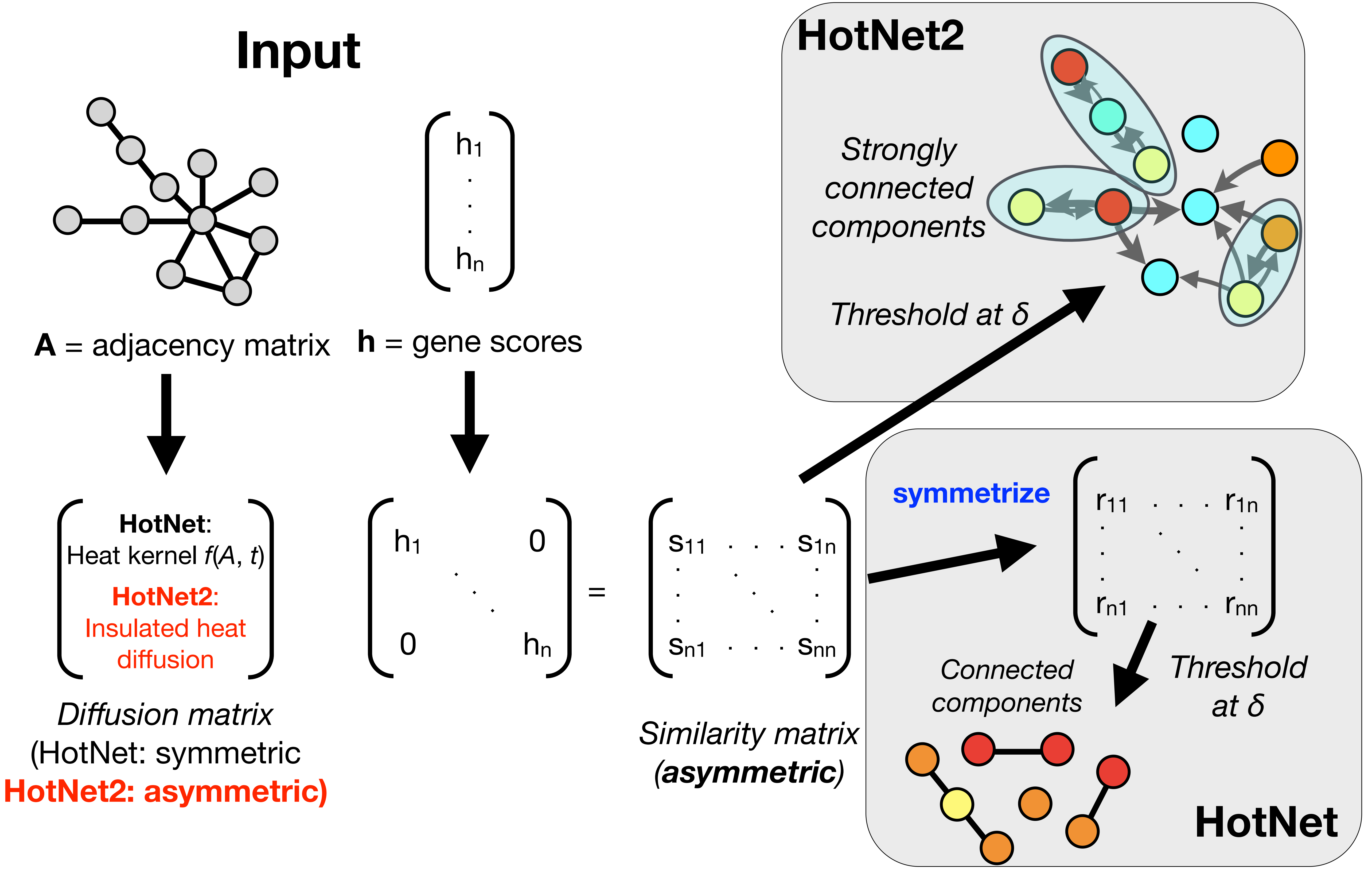
Identify *strongly* connected components

TCGA Papers

Gastric, *Nature* (2014)
Thyroid, *Cell* (2014)
…

Leiserson, Vandin *et al. Nat. Genet.* (2015).
http://compbio.cs.brown.edu/projects/hotnet2

HotNet2 vs. HotNet



Input

A = adjacency matrix

h = gene scores

$$\mathbf{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}$$

HotNet: Heat kernel *f*(*A*, *t*)

HotNet2: Insulated heat diffusion

$$\begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_n \end{pmatrix} = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{pmatrix}$$

Diffusion matrix
(HotNet: symmetric
HotNet2: asymmetric)

Similarity matrix
(***asymmetric***)

HotNet2

Strongly connected components

Threshold at δ

symmetrize

$$\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix}$$

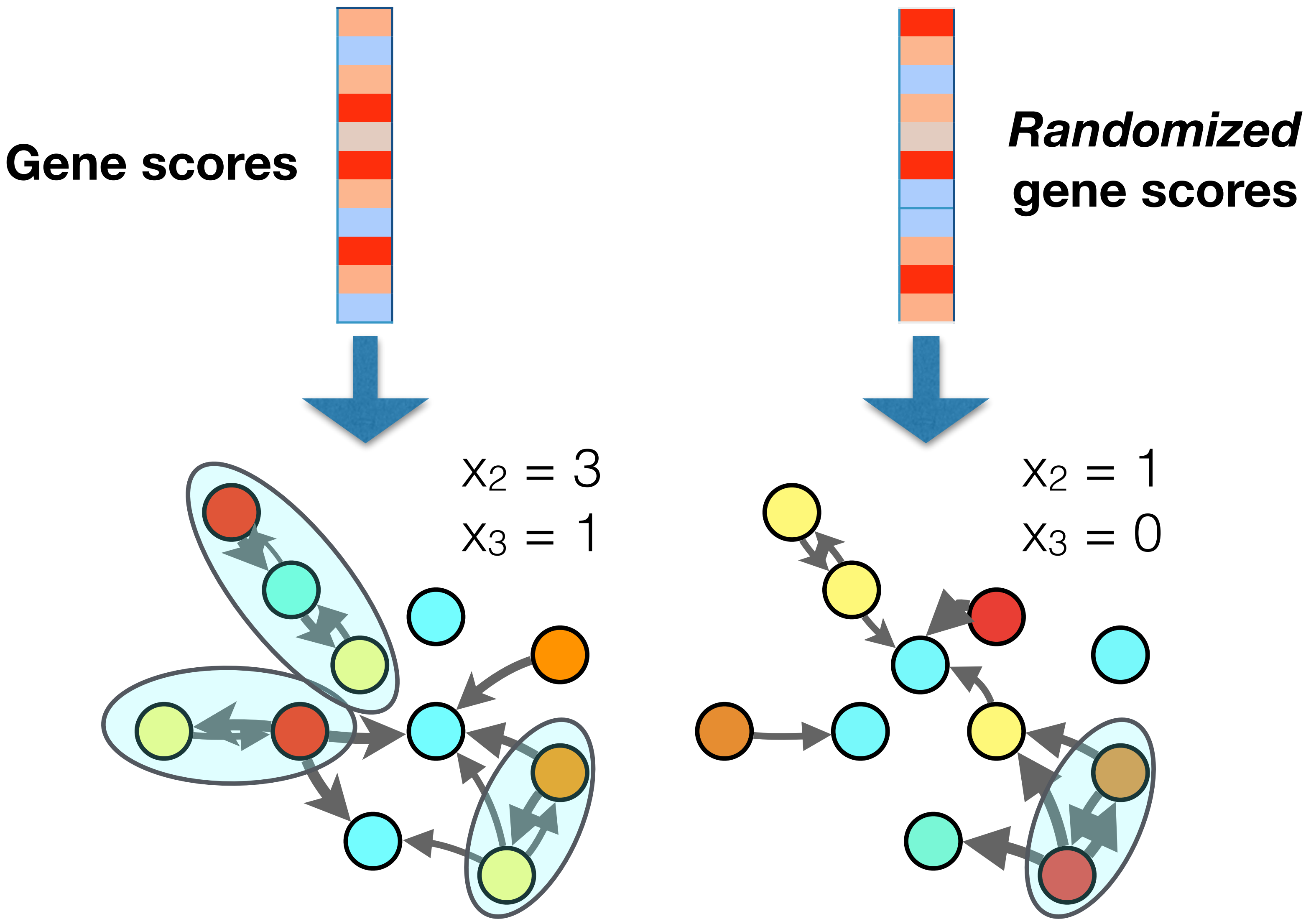
Threshold at δ

Connected components

HotNet

Statistical test

Evaluate graph partition with rigorously bounded False Discovery Rate (FDR)



Gene scores

Randomized **gene scores**

$x_2 = 3$
$x_3 = 1$

$x_2 = 1$
$x_3 = 0$

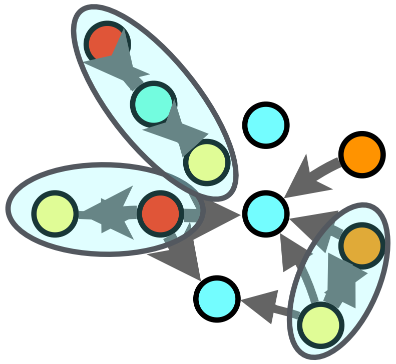
$\mathbf{X_k}$: number of subnetworks of size ≥ **k**

$\mathbf{Pr(X_k \geq x_k \mid h, \delta)}$

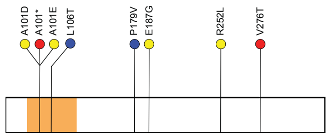
Leiserson, Vandin *et al. Nat. Genet.* (2015).

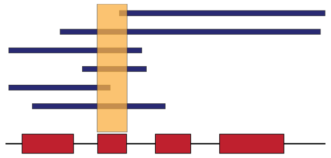
Outline

1. A new algorithm, HotNet2.



2. **Application to TCGA Pan-Cancer data.**





3. Comparison of HotNet2 to other methods.

TCGA Pan-Cancer

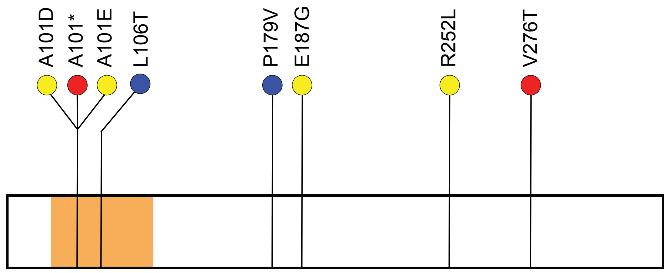
Tumor samples

Cancer	*Samples*	*Color*
BLCA	***99***	🟦
BRCA	***772***	🟦
COAD/READ	***224***	🟩
GBM	***291***	🟩
HNSC	***306***	🟥
KIRC	***417***	🟥
LAML	***196***	🟧
LUAD	***230***	🟧
LUSC	***178***	🟪
OV	***316***	🟪
UCEC	***248***	🟫

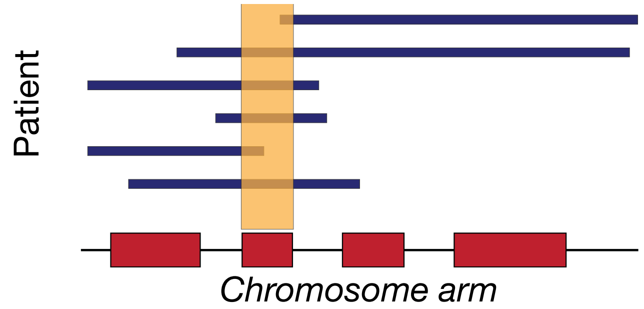
3,110 tumors of 12 cancer types

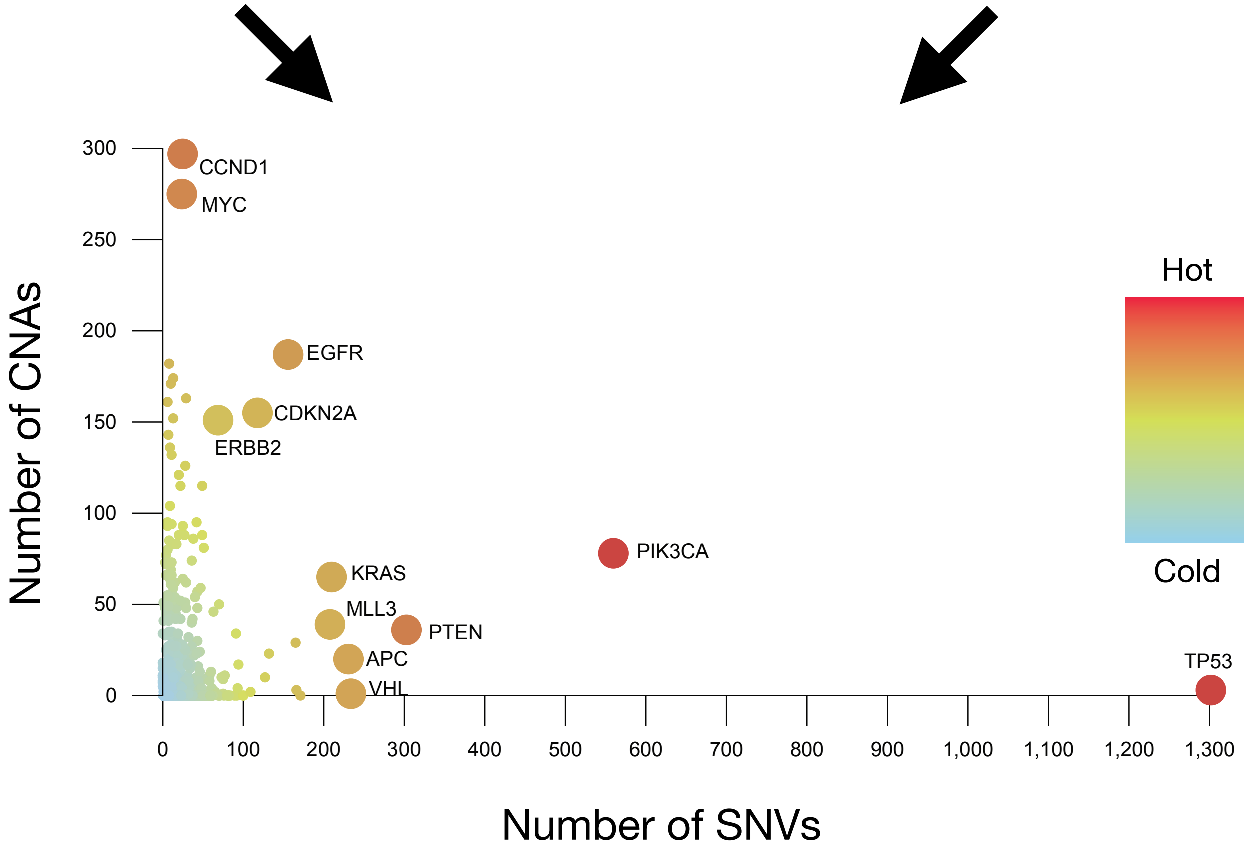
Mutations

Single nucleotide variants



Copy number aberrations

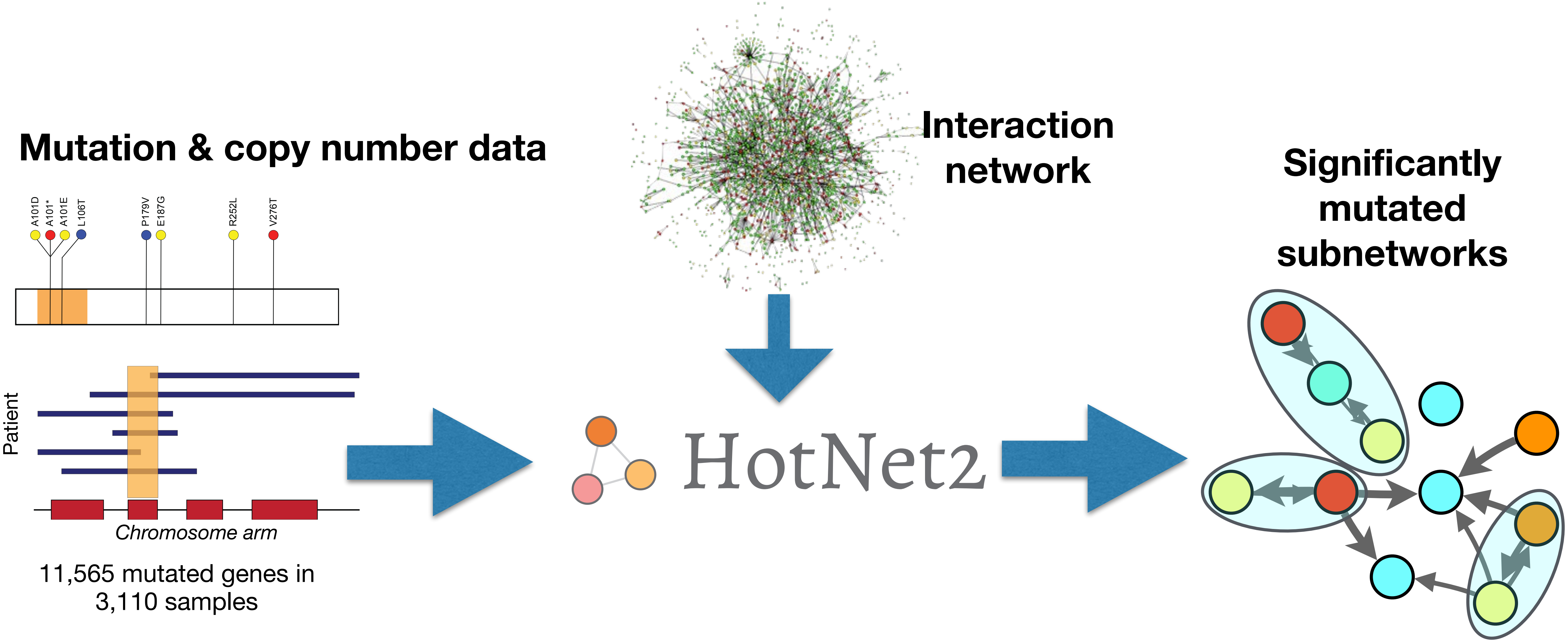




SNVs and CNAs in 3,110 samples among 11,565 *expressed* genes

Weinstein et al. *Nature Genetics* (2013)

HotNet2 runs on TCGA Pan-Cancer dataset



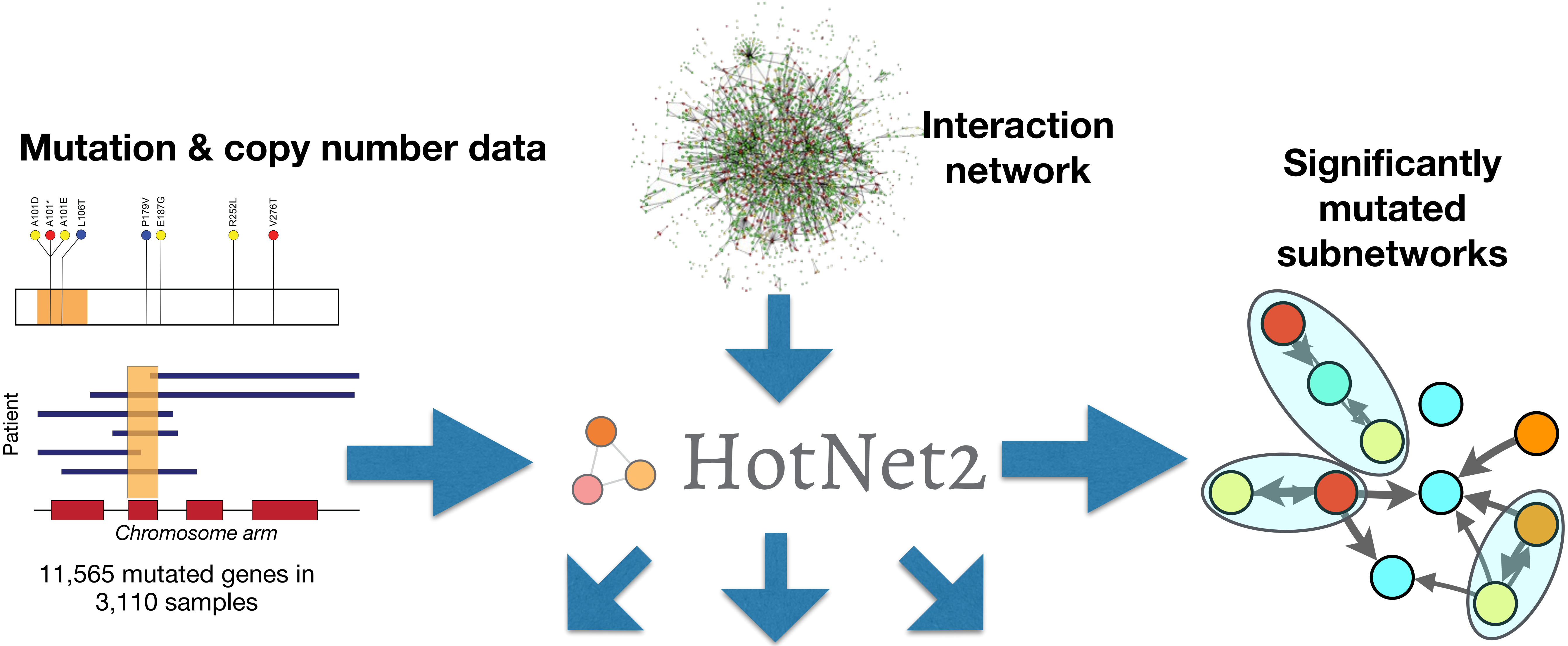
Mutation & copy number data

11,565 mutated genes in 3,110 samples

Interaction network

Significantly mutated subnetworks

HotNet2 runs on TCGA Pan-Cancer dataset


Mutation & copy number data

Interaction network

Significantly mutated subnetworks

11,565 mutated genes in 3,110 samples

HINT+HI2012 (P < 0.01)

40,704 interactions
9,858 proteins

iRefIndex 9.0 (P < 0.01)

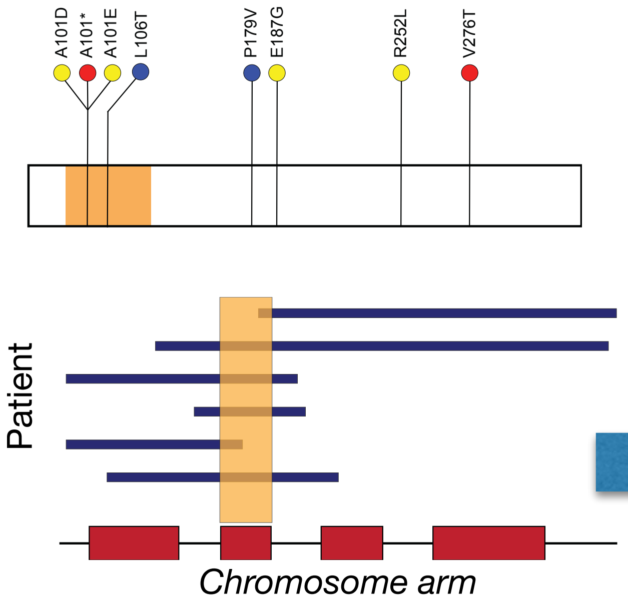
91,808 interactions
12,128 proteins

Multinet (P < 0.01)

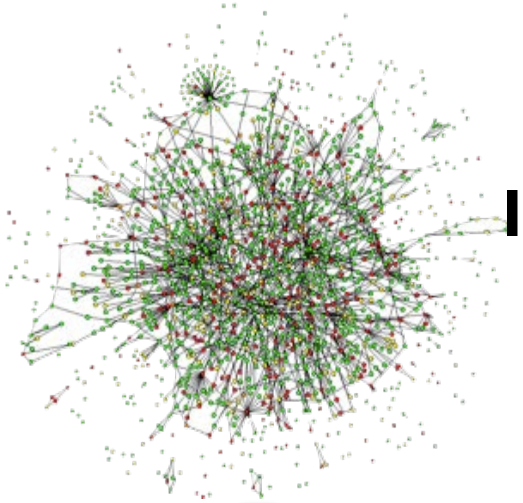
109,569 interactions
14,398 proteins

HotNet2 runs on TCGA Pan-Cancer dataset

Mutation & copy number data



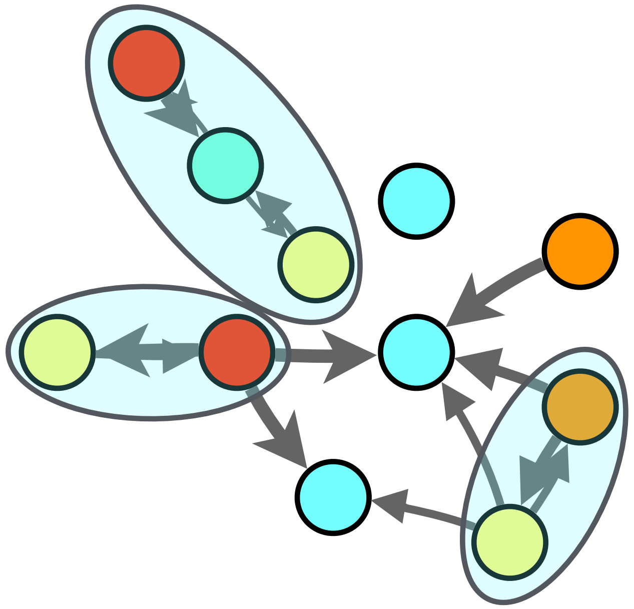
11,565 mutated genes in
3,110 samples



Interaction network

HotNet2

Significantly mutated subnetworks



HINT+HI2012 (P < 0.01)

40,704 interactions
9,858 proteins

iRefIndex 9.0 (P < 0.01)

91,808 interactions
12,128 proteins

Multinet (P < 0.01)

109,569 interactions
14,398 proteins

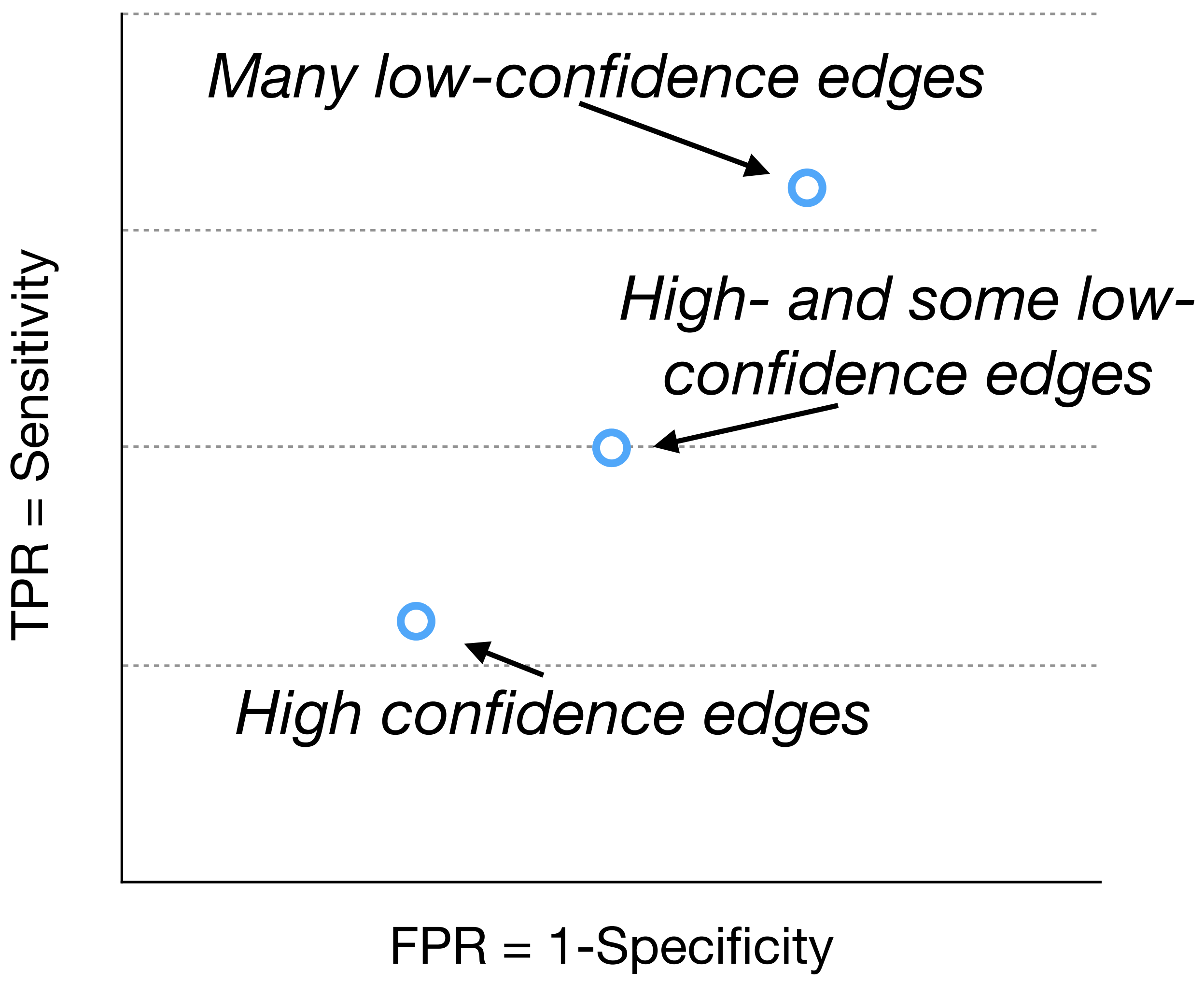
Consensus subnetworks

16 consensus subnetworks with ≥ 4 genes (*P*=0.004)
13 “linkers” between consensus subnetworks

HotNet2 Consensus

HotNet2 Runs
HINT+HI2012 ($P < 0.01$)
iRefIndex 9.0 ($P < 0.01$)
Multinet ($P < 0.01$)

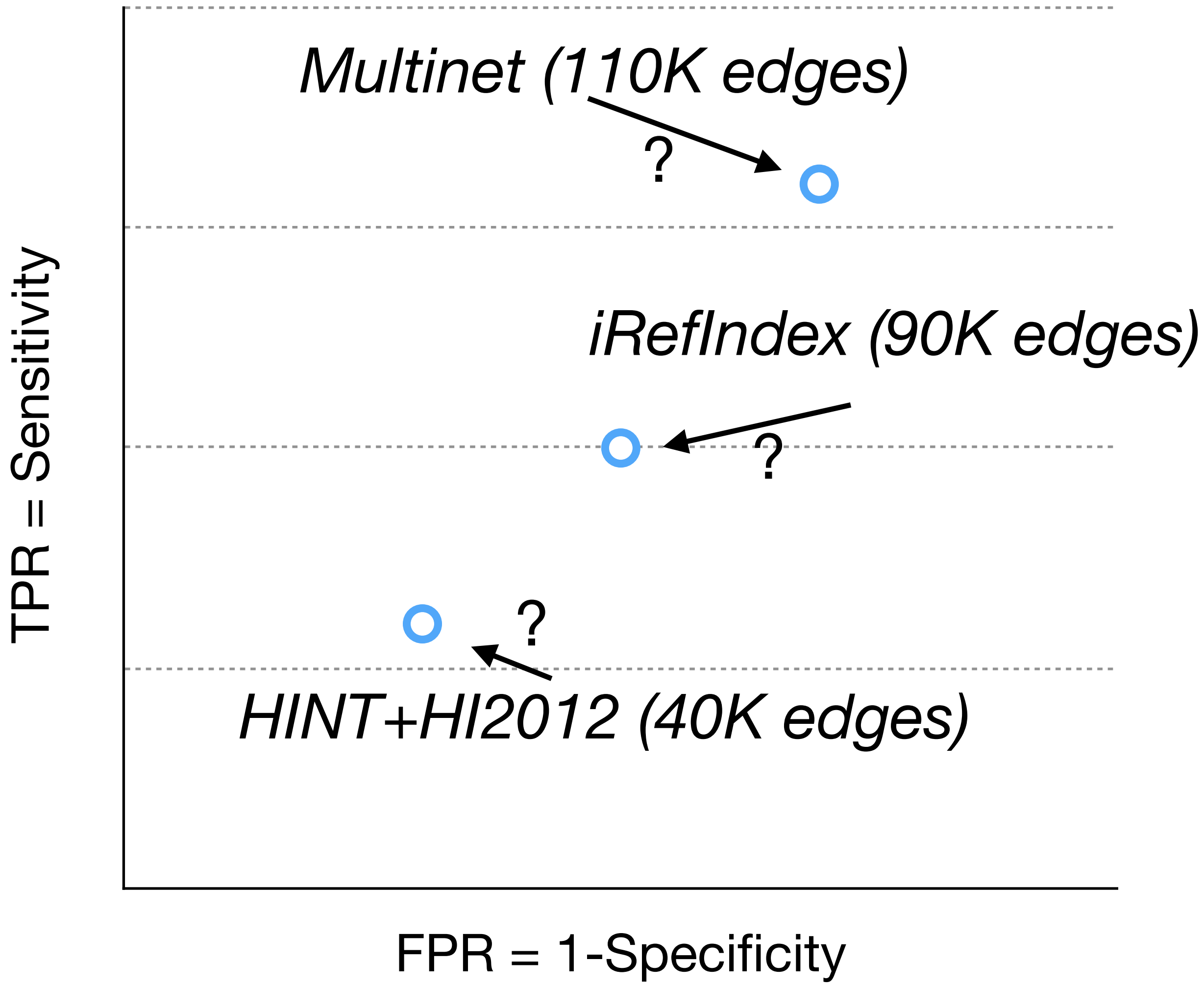
Interaction networks



HotNet2 Consensus

HotNet2 Runs
HINT+HI2012 (*P* < 0.01)
iRefIndex 9.0 (*P* < 0.01)
Multinet (*P* < 0.01)

Interaction networks

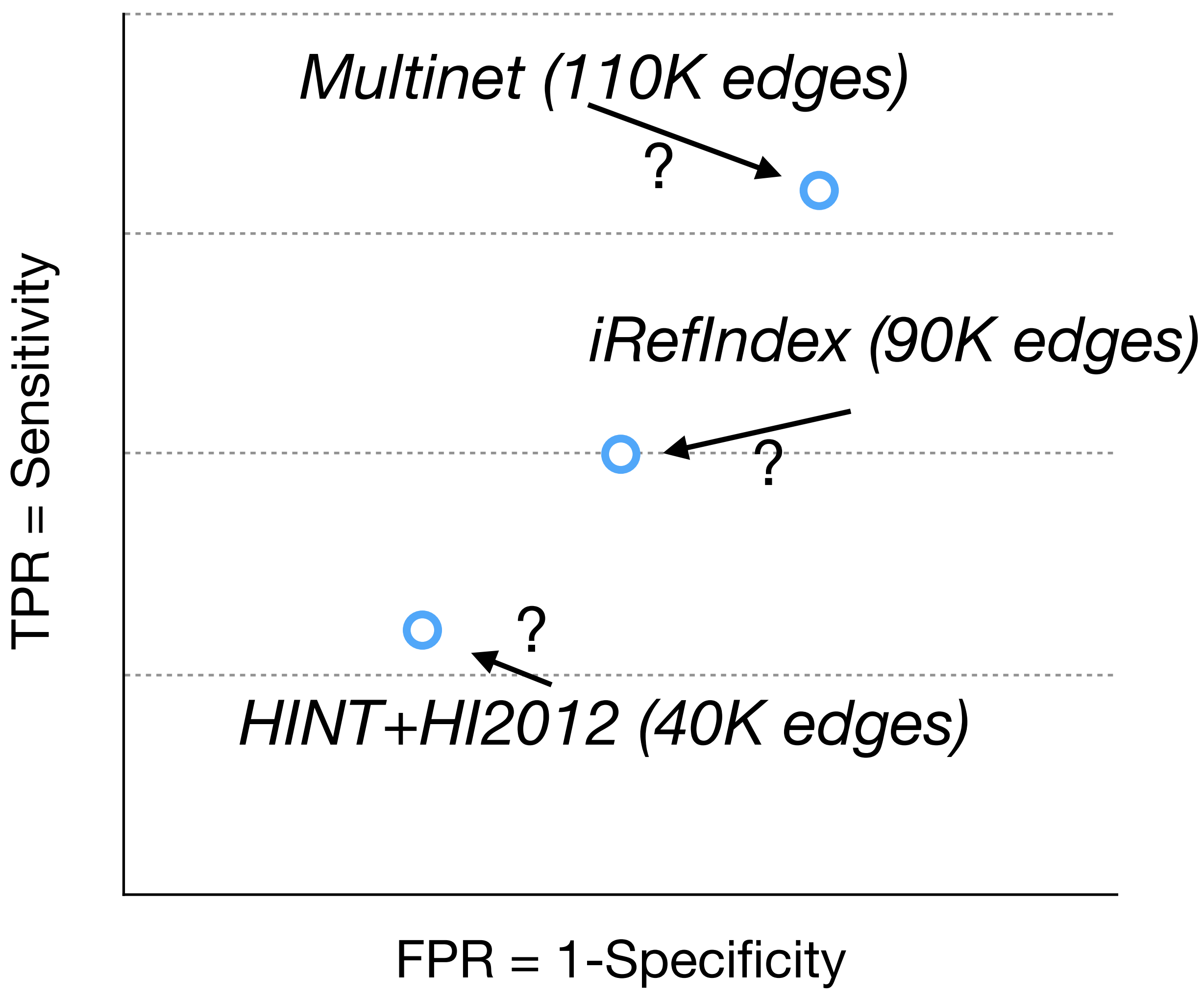


HotNet2 Consensus

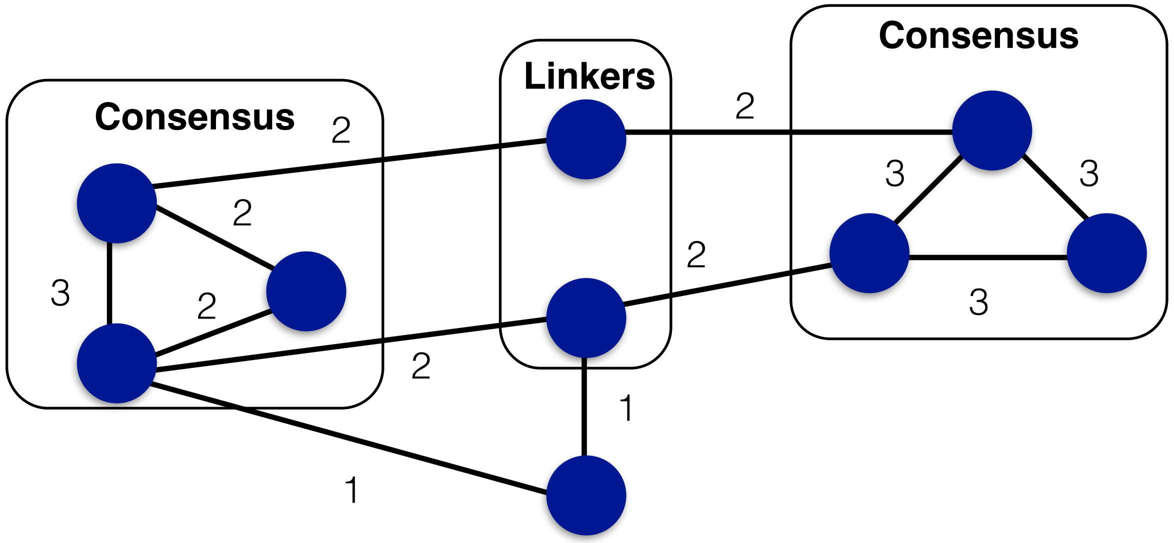
HotNet2 Runs
HINT+HI2012 ($P < 0.01$)
iRefIndex 9.0 ($P < 0.01$)
Multinet ($P < 0.01$)

Consensus
16 consensus subnetworks
13 “linkers” between subnetworks

Interaction networks

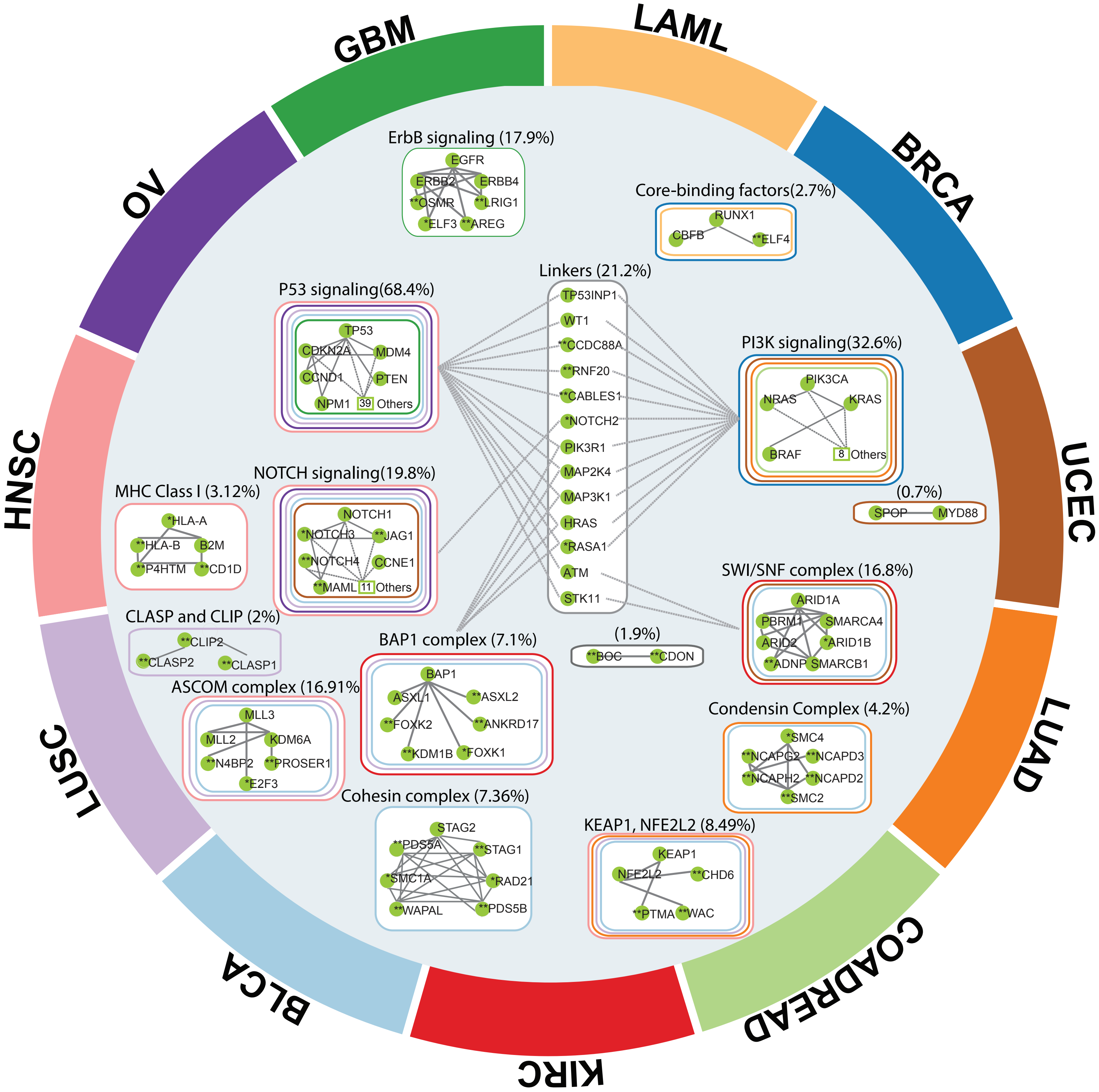


Main Idea: Incorporate low-confidence edges but give high-confidence edges more weight.



Consensus Graph
Edges connect genes identified by HotNet2 in the same subnetwork.

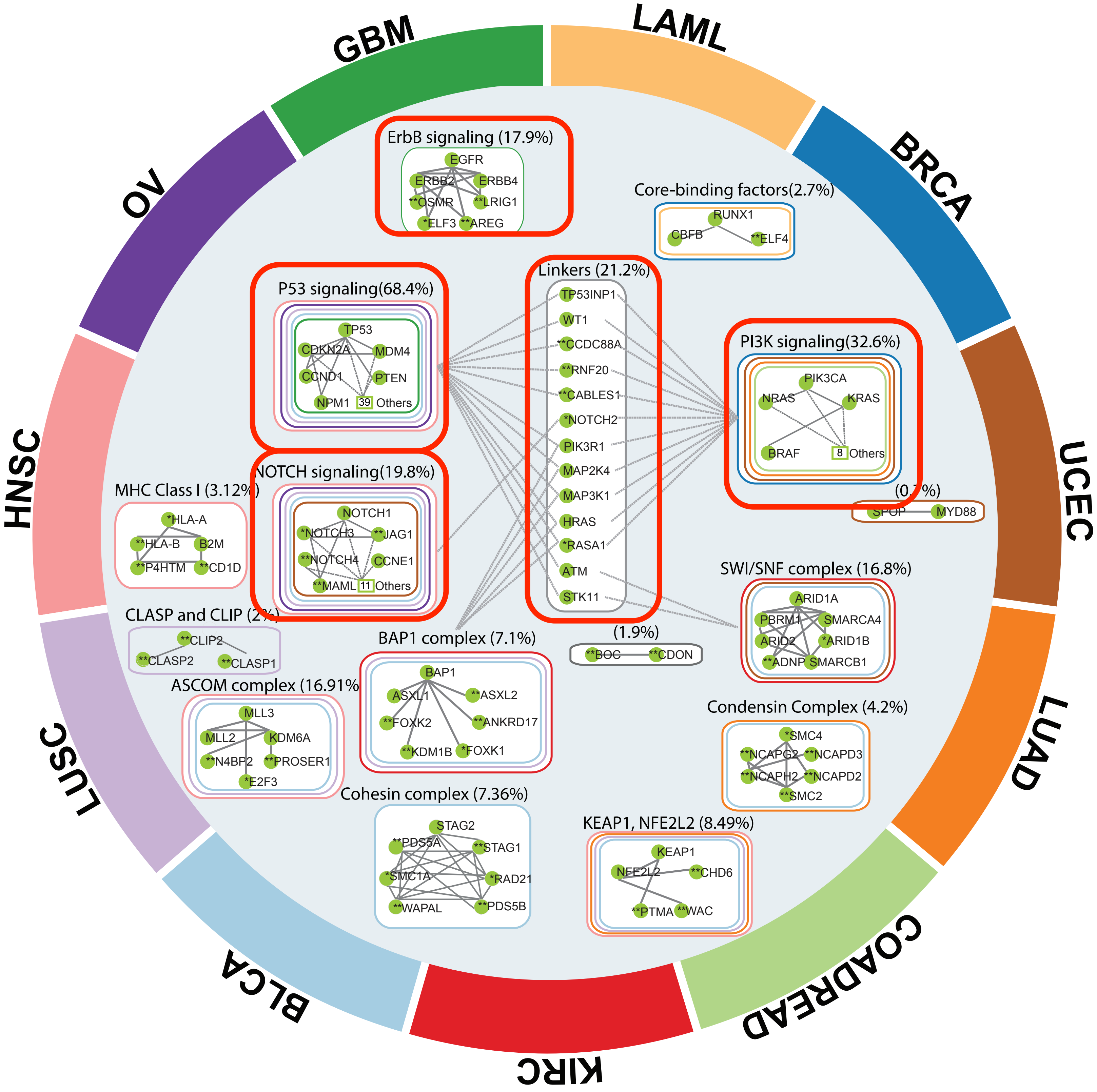
HotNet2 Consensus Subnetworks



Frequently and rarely mutated cancer genes

http://compbio.cs.brown.edu/pancancer/hotnet2

HotNet2 Consensus Subnetworks



Frequently and rarely mutated cancer genes

Well-known cancer pathways

- PI(3)K signaling
- p53 signaling
- NOTCH signaling
- ErbB signaling
- Linkers: *HRAS*, *STK11*, *ATM*

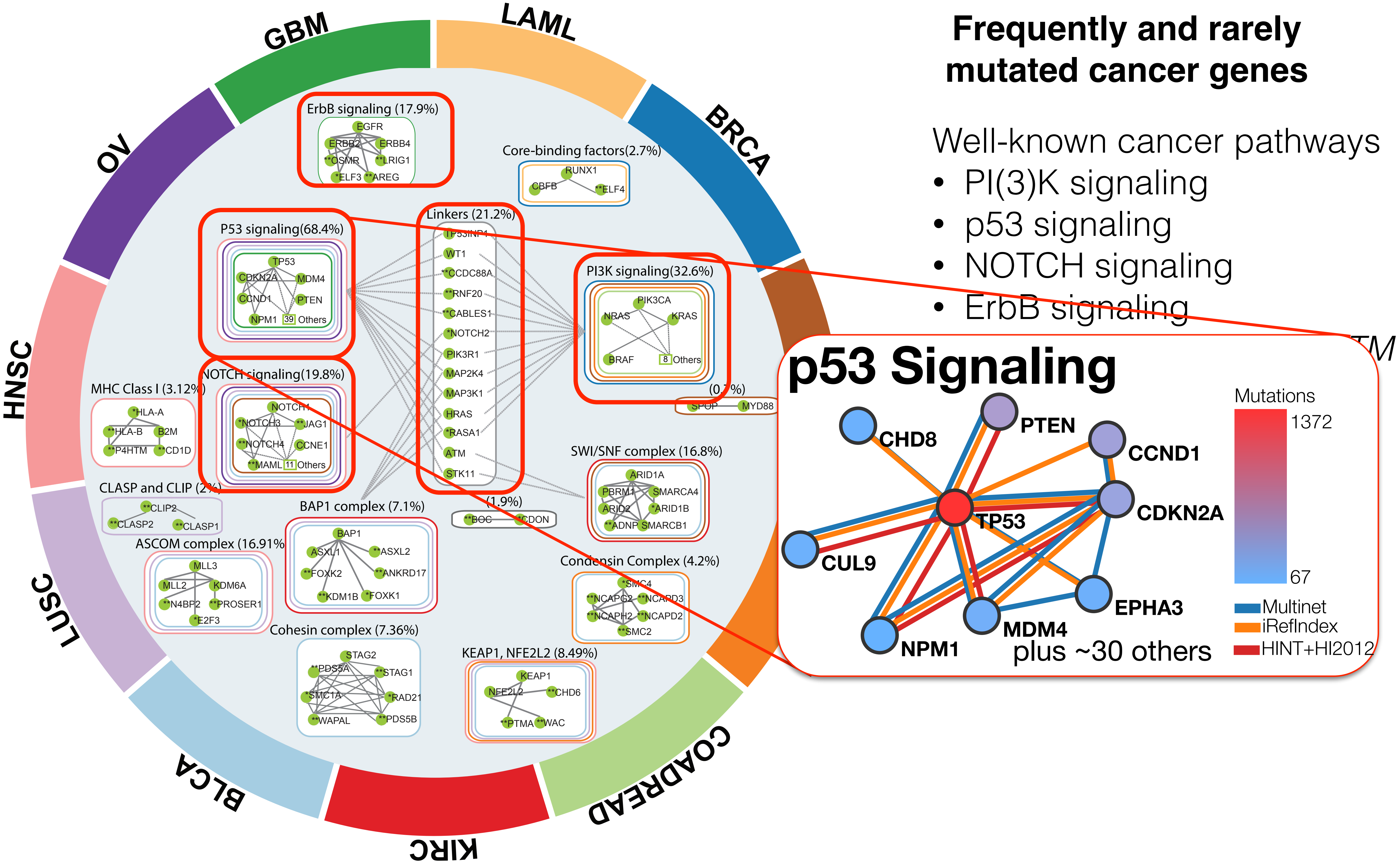
http://compbio.cs.brown.edu/pancancer/hotnet2

HotNet2 Consensus Subnetworks

Frequently and rarely mutated cancer genes

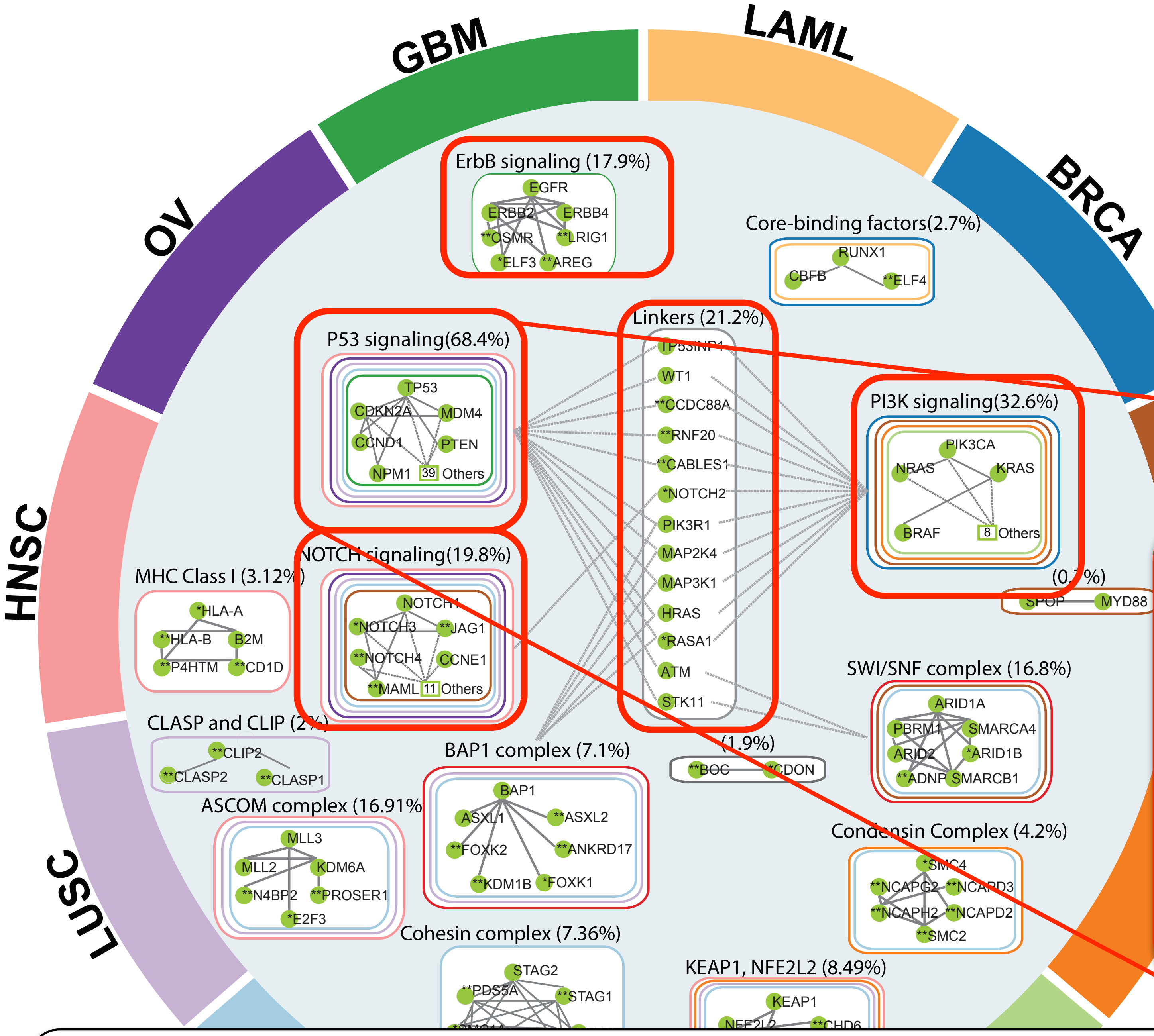
Well-known cancer pathways

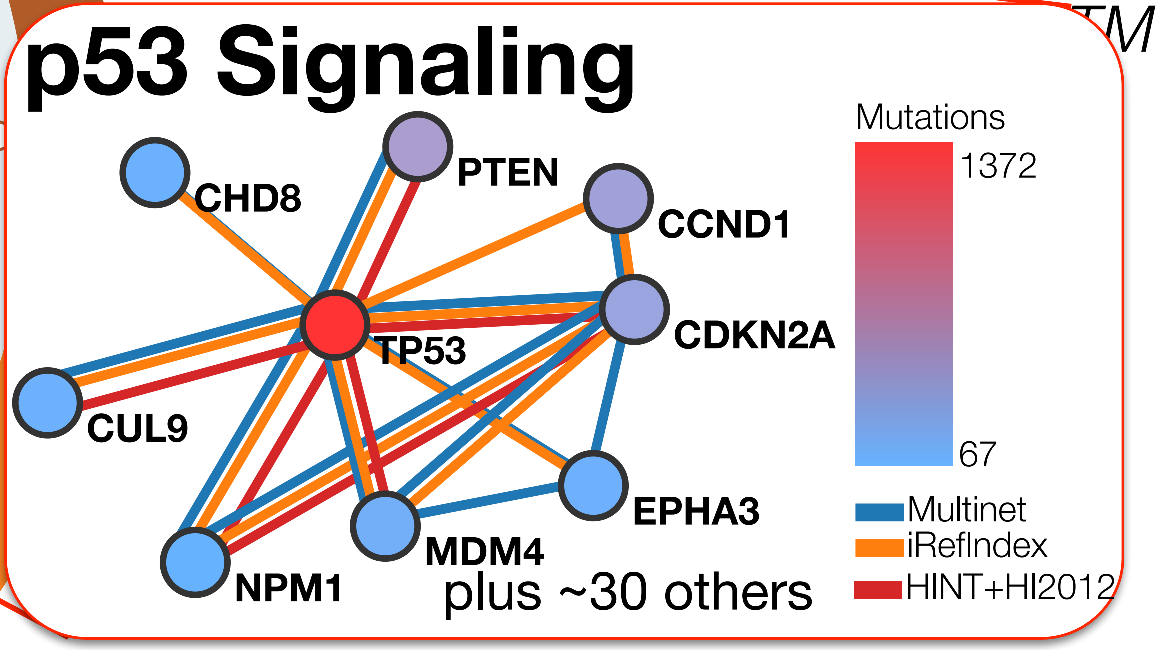
- PI(3)K signaling
- p53 signaling
- NOTCH signaling
- ErbB signaling



http://compbio.cs.brown.edu/pancancer/hotnet2

HotNet2 Consensus Subnetworks





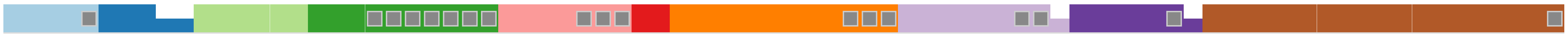
Frequently and rarely mutated cancer genes

Well-known cancer pathways

- PI(3)K signaling
- p53 signaling
- NOTCH signaling
- ErbB signaling

TM

CHD8 (mutated in 54 of 3110 samples)



Legend

Cancer types

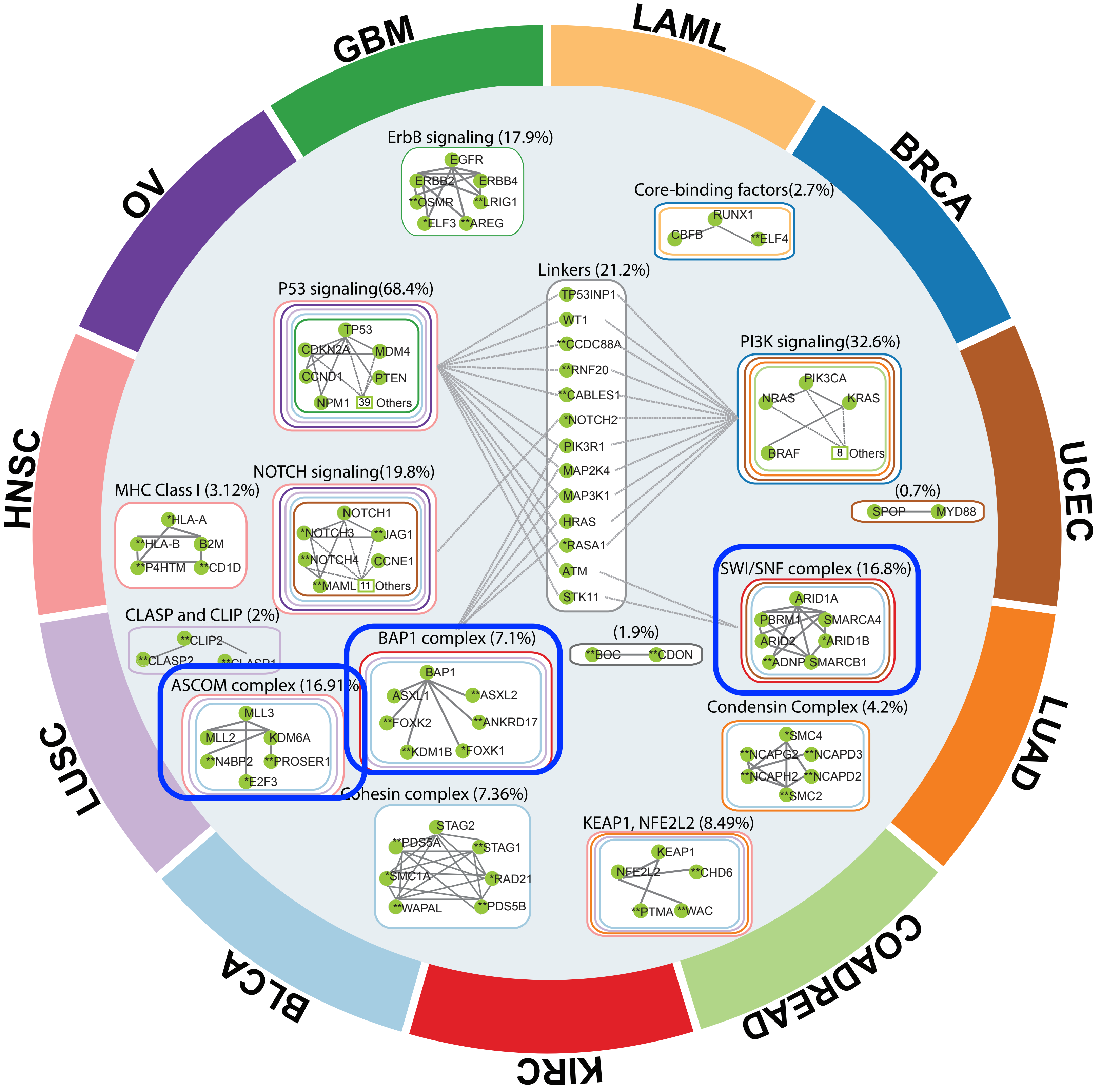
▮	BLCA	▮	BRCA	▮	COADREAD	▮	GBM	▮	HNSC
▮	KIRC	▮	LAML	▮	LUAD	▮	LUSC	▮	OV	▮	UCEC

Mutation types

▮	Inactivating SNV	▮	SNV
▮	Amplification	▮	Deletion

http://compbio.cs.brown.edu/pancancer/hotnet2

HotNet2 Consensus Subnetworks



Frequently and rarely mutated cancer genes

Well-known cancer pathways

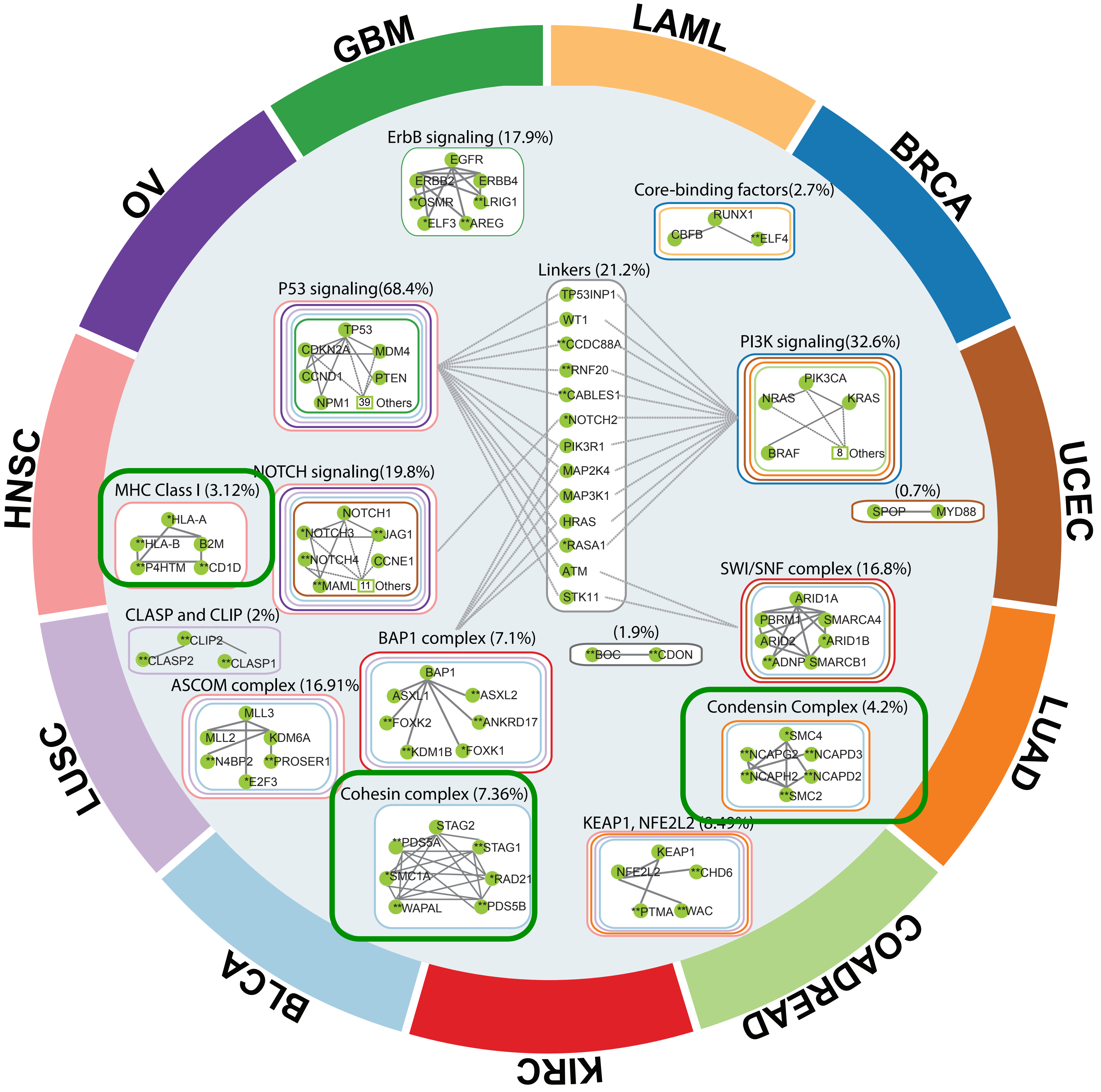
- PI(3)K signaling
- p53 signaling
- NOTCH signaling
- ErbB signaling
- Linkers: *HRAS*, *STK11*, *ATM*

Recently characterized complexes:

- SWI/SNF complex
- ASCOM complex
- BAP1 complex

http://compbio.cs.brown.edu/pancancer/hotnet2

HotNet2 Consensus Subnetworks



Frequently and rarely mutated cancer genes

Well-known cancer pathways

- PI(3)K signaling
- p53 signaling
- NOTCH signaling
- ErbB signaling
- Linkers: *HRAS*, *STK11*, *ATM*

Recently characterized complexes:

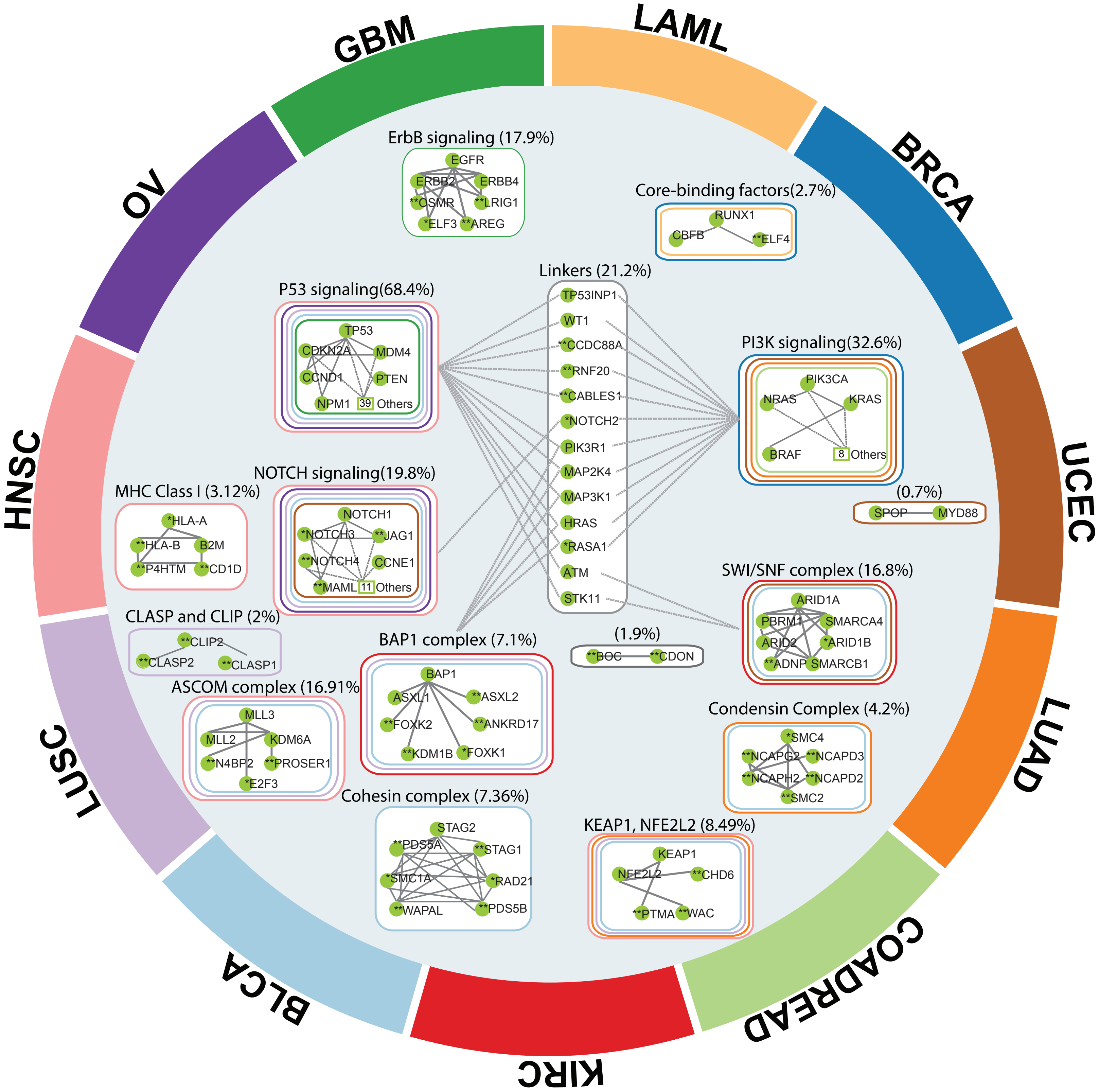
- SWI/SNF complex
- ASCOM complex
- BAP1 complex

Potentially novel complexes:

- Cohesin complex
- Condensin complex
- MHC Class I proteins

http://compbio.cs.brown.edu/pancancer/hotnet2

HotNet2 Consensus Subnetworks



Frequently and rarely mutated cancer genes

Well-known cancer pathways

- PI(3)K signaling
- p53 signaling
- NOTCH signaling
- ErbB signaling
- Linkers: *HRAS*, *STK11*, *ATM*

Recently characterized complexes:

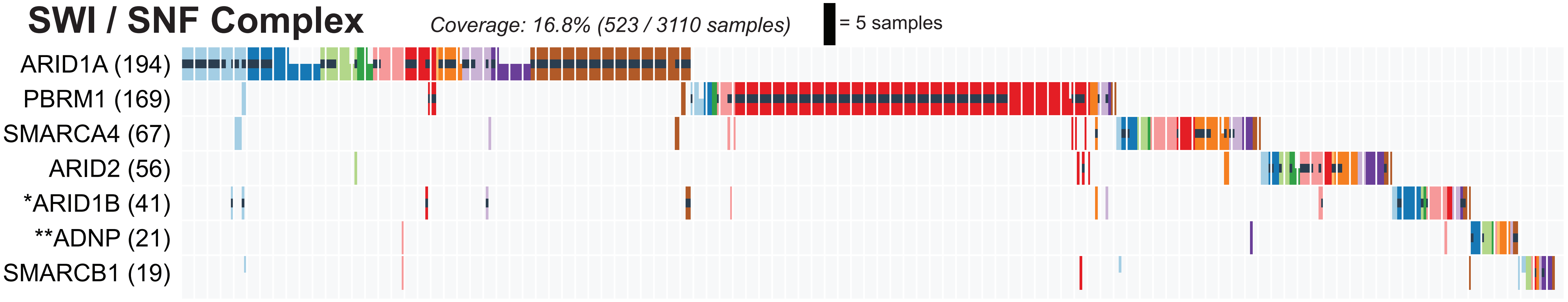
- SWI/SNF complex
- ASCOM complex
- BAP1 complex

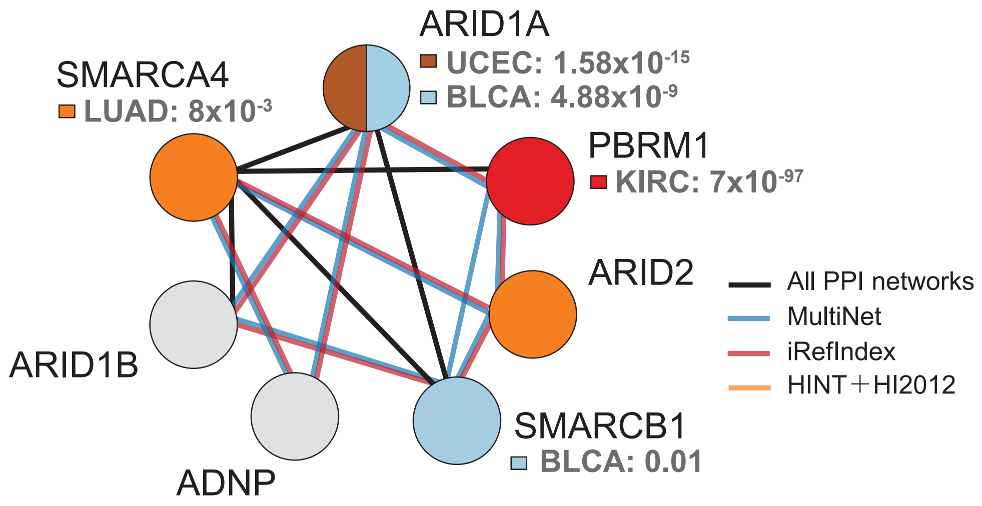
Potentially novel complexes:

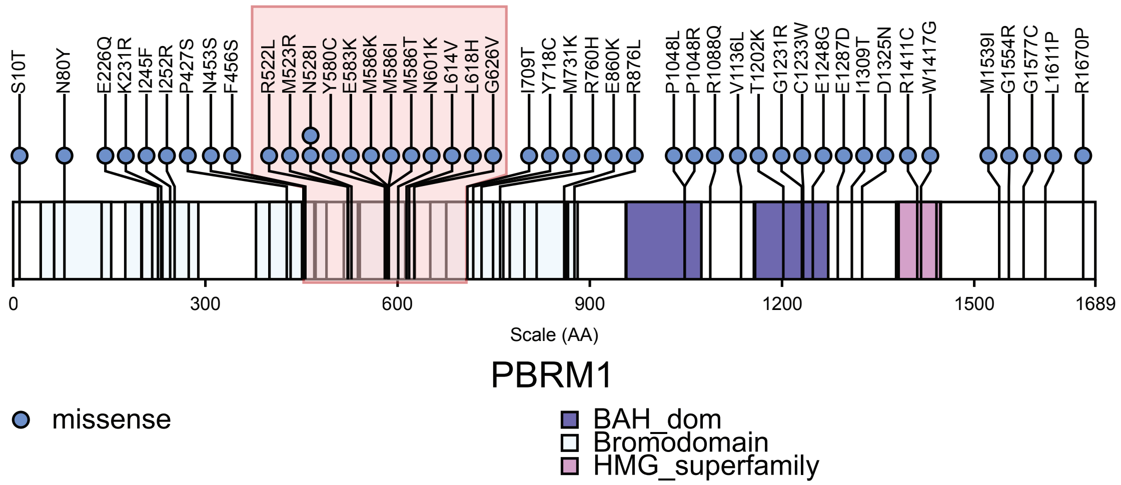
- Cohesin complex
- Condensin complex
- MHC Class I proteins

http://compbio.cs.brown.edu/pancancer/hotnet2

SWI/SNF complex





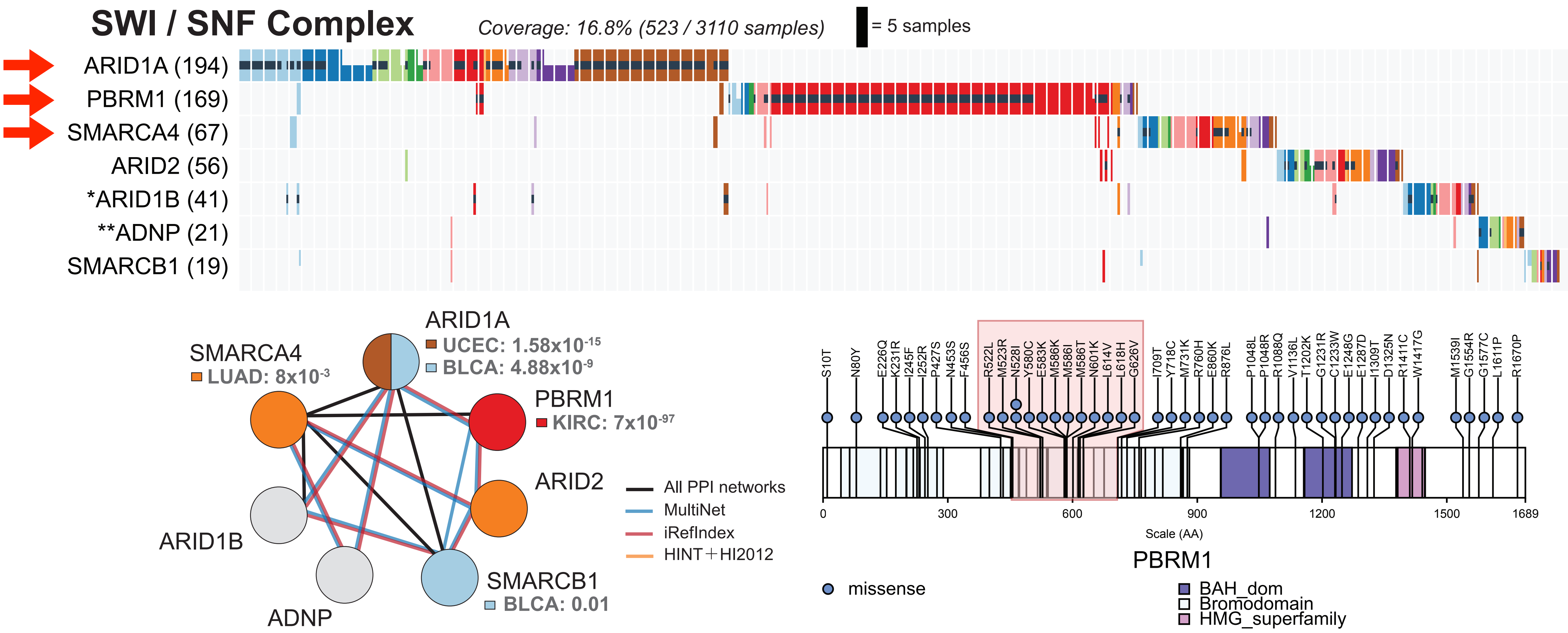


Legend

Cancer types: BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, LAML, LUAD, LUSC, OV, UCEC

Mutation types: Inactivating SNV, SNV, Amplification, Deletion

SWI/SNF complex

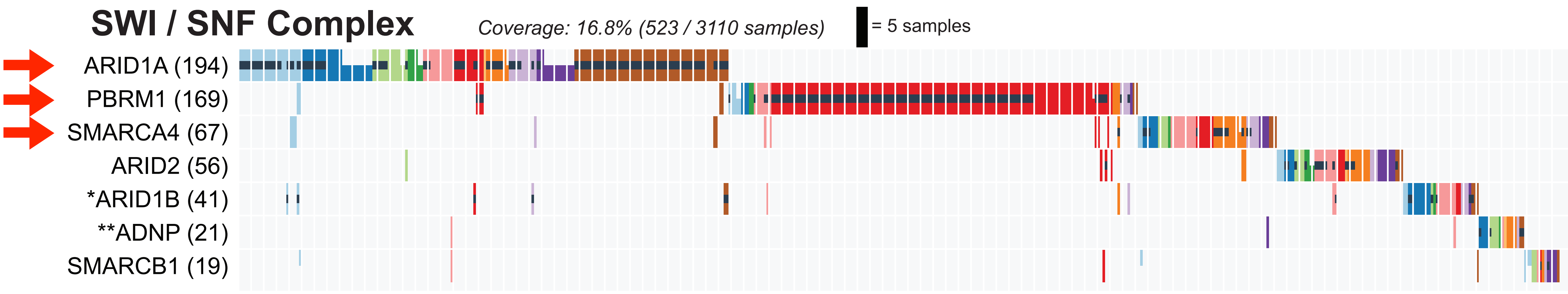


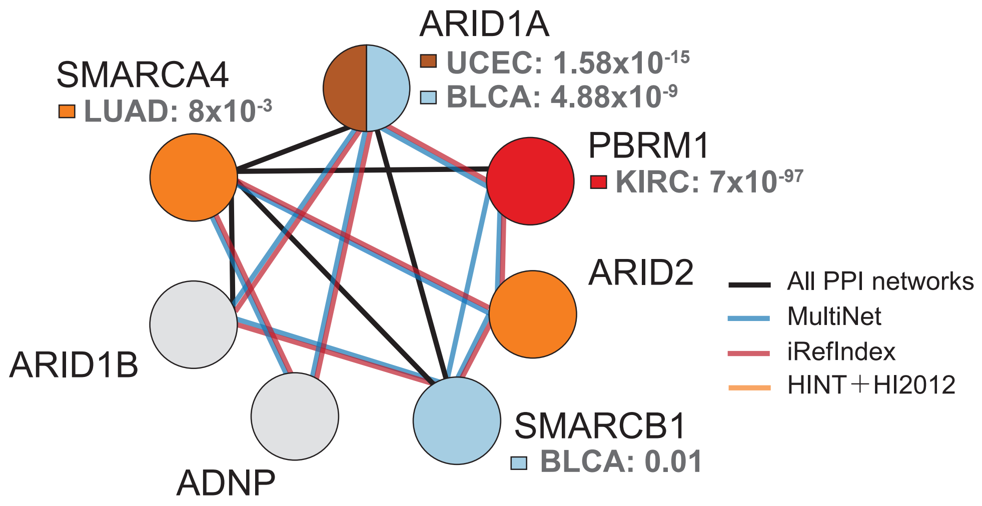
Legend

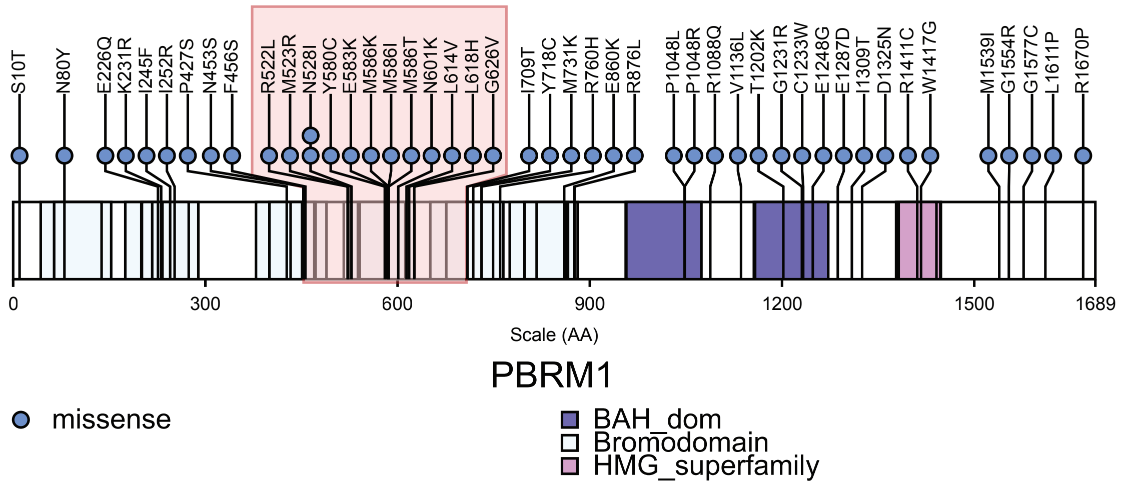
Cancer types: BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, LAML, LUAD, LUSC, OV, UCEC

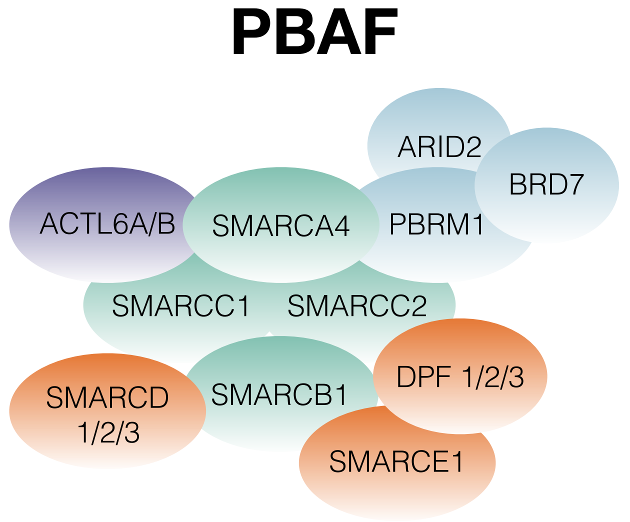
Mutation types: Inactivating SNV, SNV, Amplification, Deletion

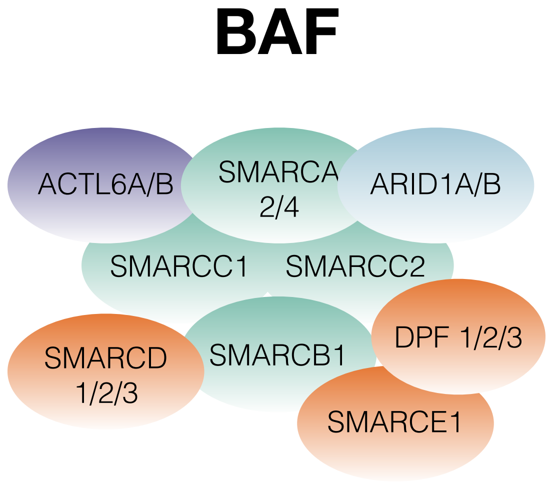
SWI/SNF complex











Wilson and Roberts.
Nature Reviews Cancer (2011)

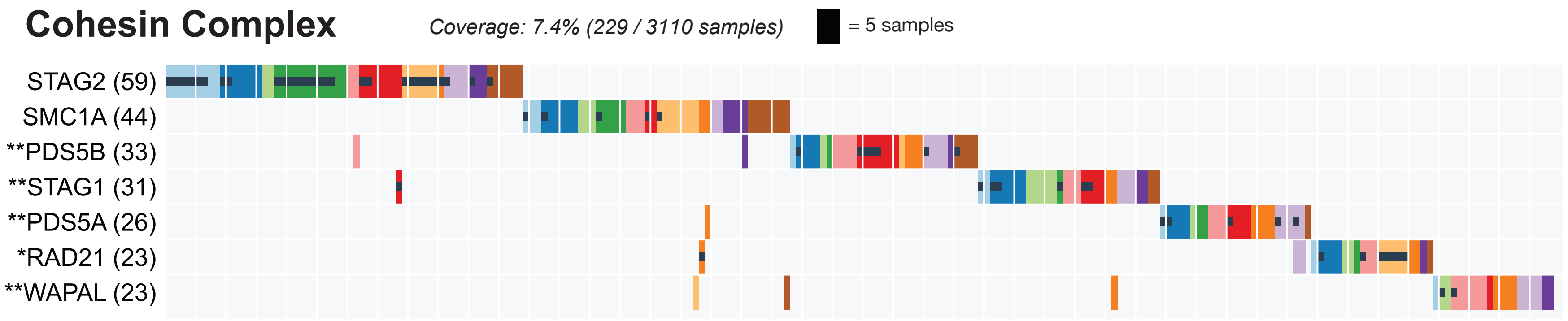
Involved in
nucleosome remodeling

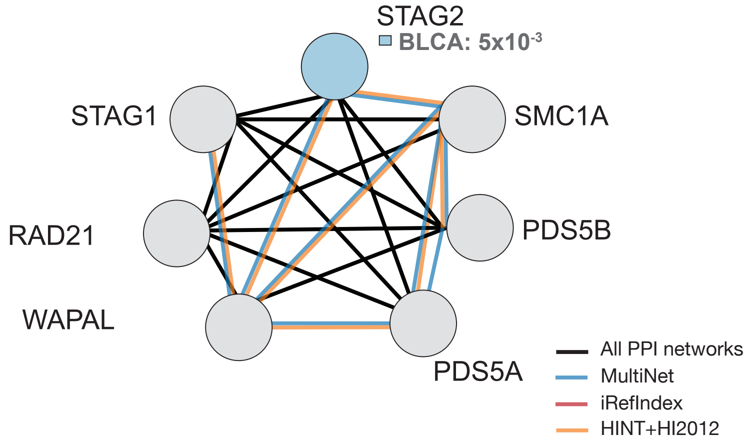
Legend

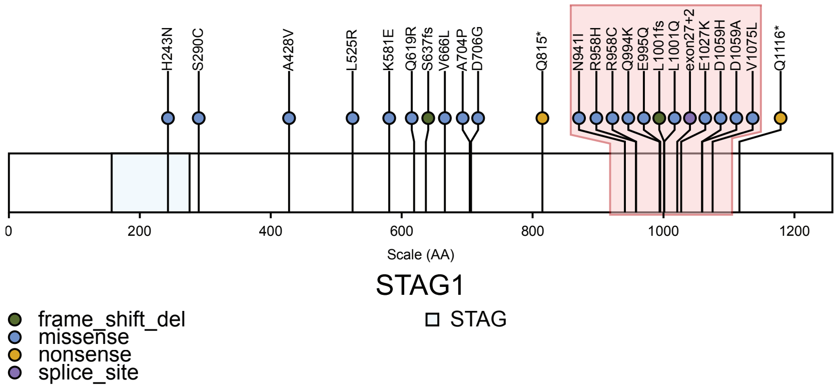
Cancer types: ■ BLCA ■ BRCA ■ COADREAD ■ GBM ■ HNSC ■ KIRC ■ LAML ■ LUAD ■ LUSC ■ OV ■ UCEC

Mutation types: ▌Inactivating SNV ▌SNV ▌Amplification ▌Deletion

Cohesin and condensin complexes

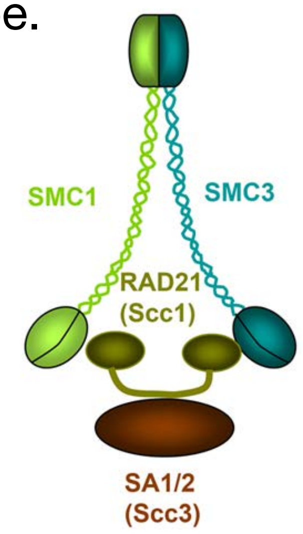






Cohesin complex

- 4/5 members of complex
- Involved in sister chromatid cohesion and gene regulation.
- Mutated in >4% of samples in each cancer type.

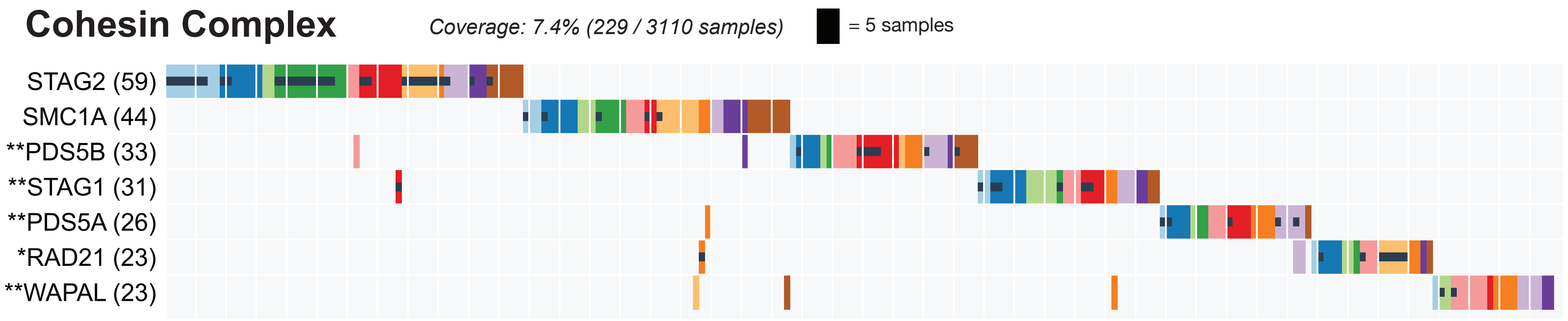


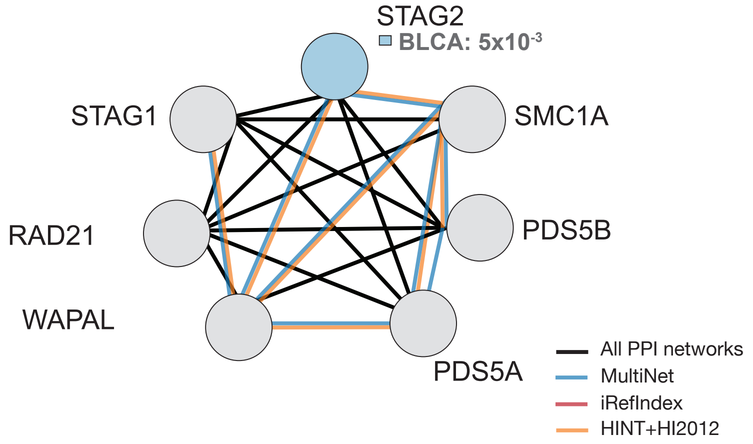
Legend

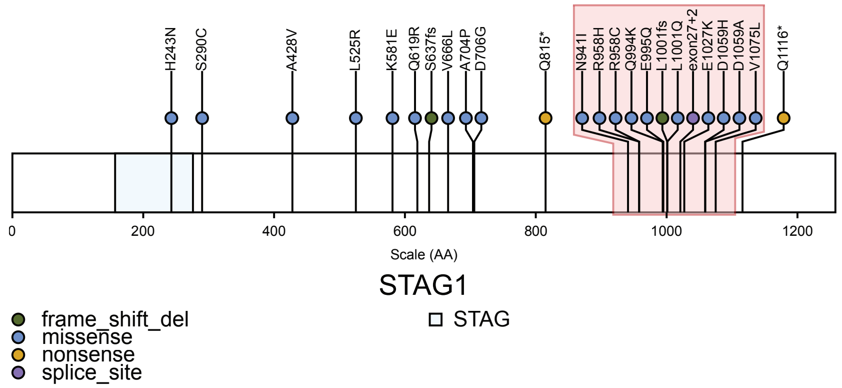
Cancer types: BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, LAML, LUAD, LUSC, OV, UCEC

Mutation types: Inactivating SNV, SNV, Amplification, Deletion

Cohesin and condensin complexes

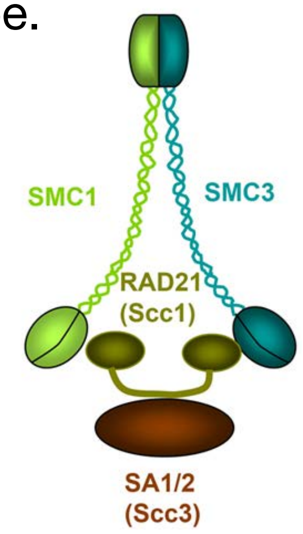


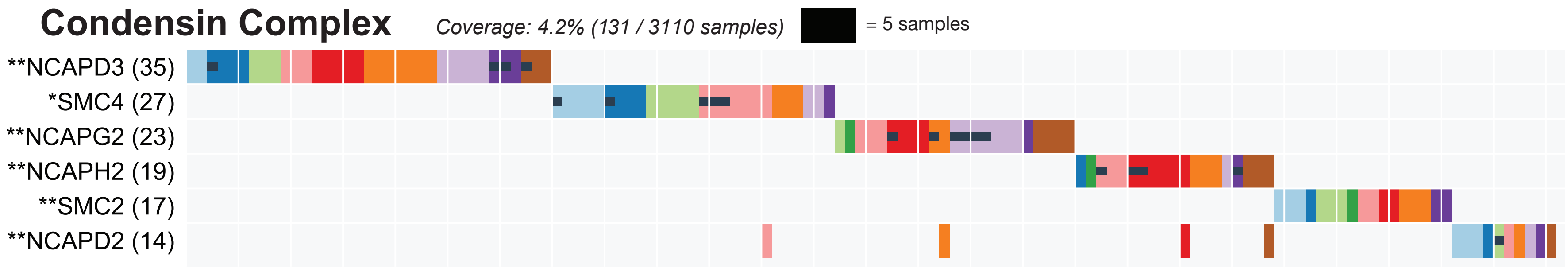


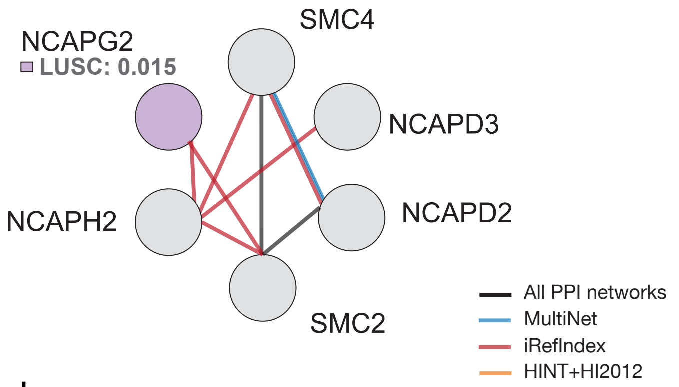


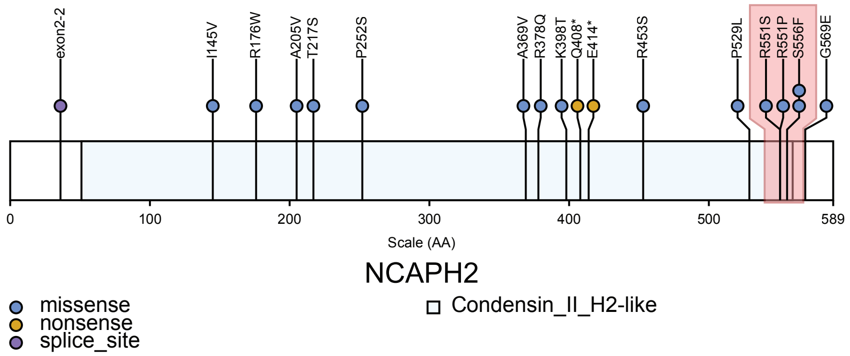
Cohesin complex

- 4/5 members of complex
- Involved in sister chromatid cohesion and gene regulation.
- Mutated in >4% of samples in each cancer type.









Condensin complex

- 6/8 members of complex
- Involved in sister chromatid condensation and gene regulation.
- Somatic mutations and expression validated using whole-genome sequencing and RNA-Seq

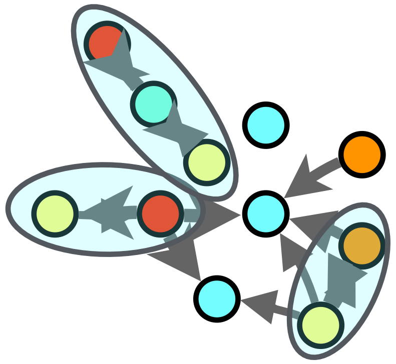
Legend

Cancer types: ▮ BLCA ▮ BRCA ▮ COADREAD ▮ GBM ▮ HNSC ▮ KIRC ▮ LAML ▮ LUAD ▮ LUSC ▮ OV ▮ UCEC

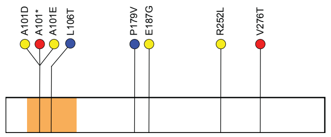
Mutation types: ▮ Inactivating SNV ▮ SNV ▮ Amplification ▮ Deletion

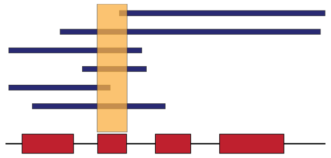
Outline

1. A new algorithm, HotNet2.



2. Application to TCGA Pan-Cancer data.





3. **Comparison of HotNet2 to similar methods.**

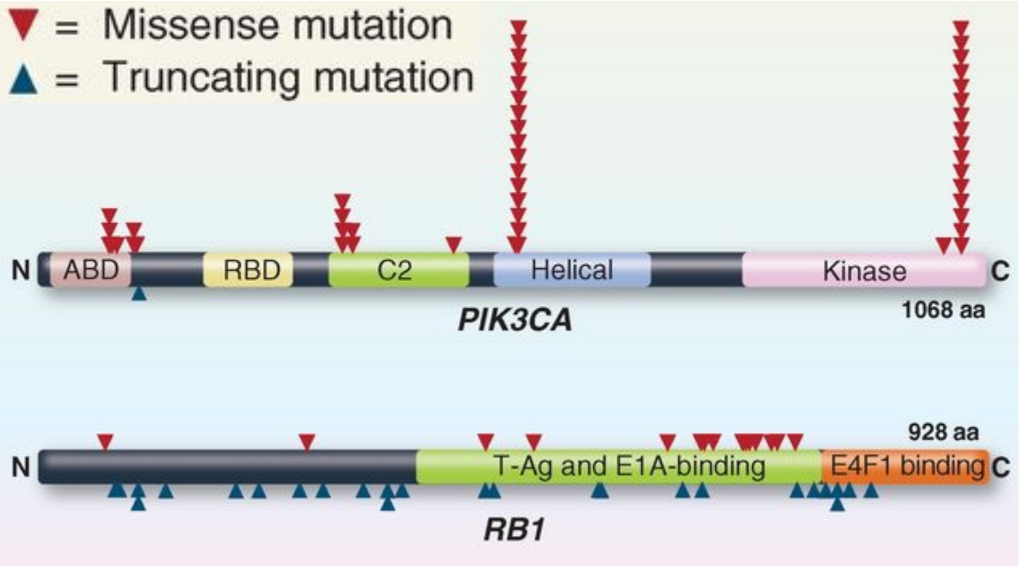
HotNet2 outperforms other methods on real data

No gold standard dataset → compare methods at identifying putative cancer genes

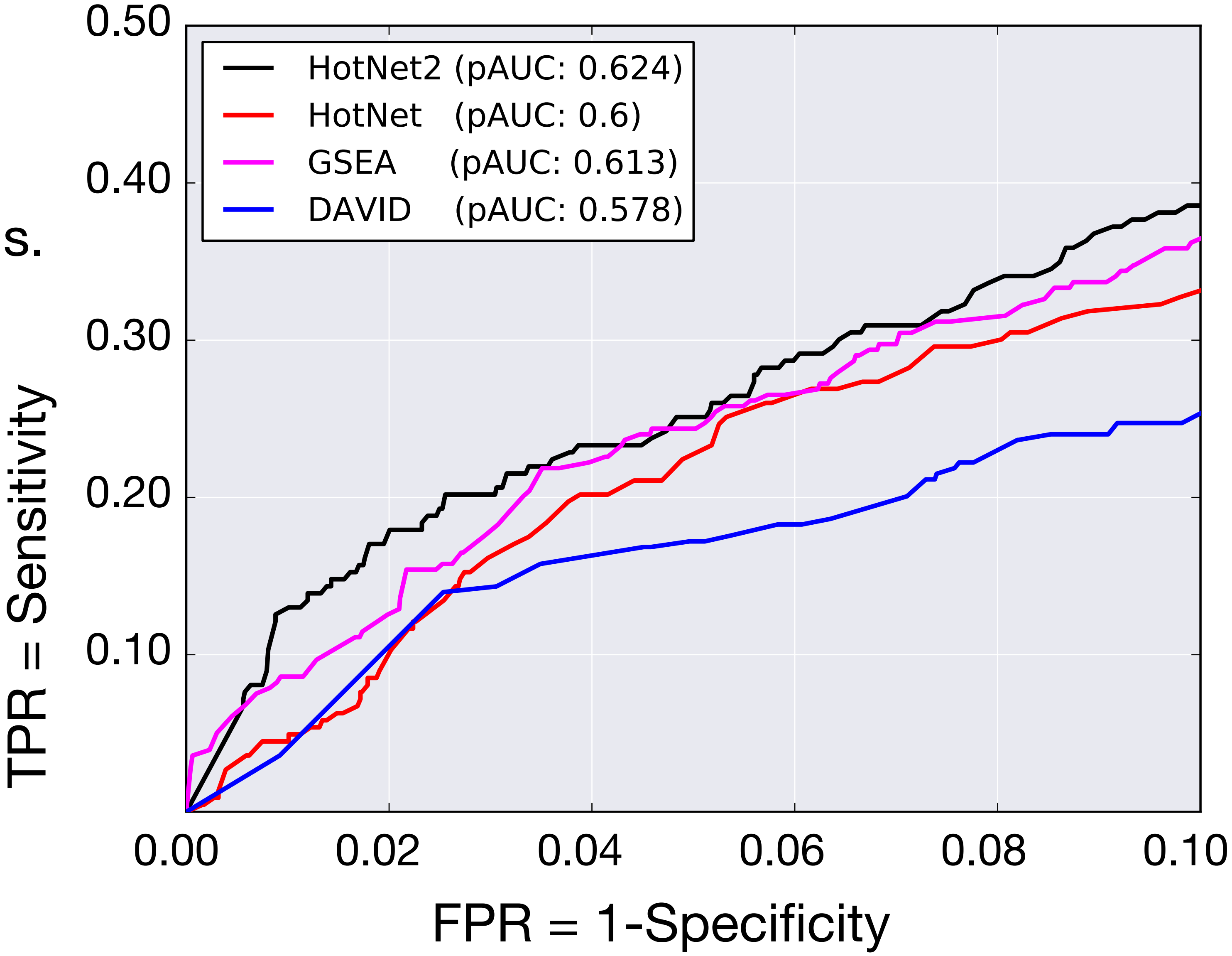
Dataset of *putative* cancer genes

Cancer genes have:

1. ≥ 20% truncating mutations; or,
2. ≥ 20% mutations clustered at a locus.



Vogelstein *et al*. (*Science*, 2013)

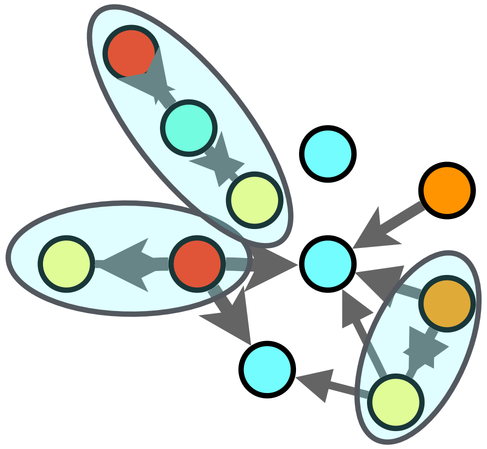


Summary

- HotNet2: Novel algorithm that analyzes topology and mutations simultaneously with asymmetric heat diffusion.

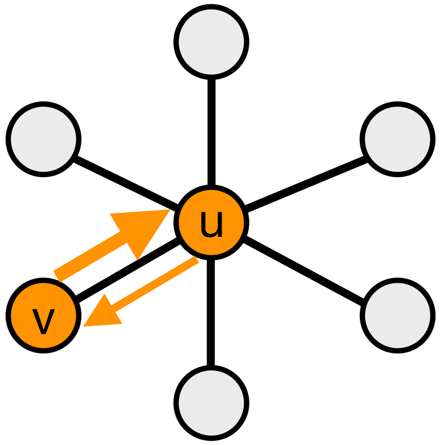
- Identifies known and novel pathways and complexes with frequently and rarely mutated genes on TCGA Pan-Cancer data.

- Future work:
  - Alternate graph partitioning algorithms?
  - Other applications: gene expression, GWAS, social networks, etc.



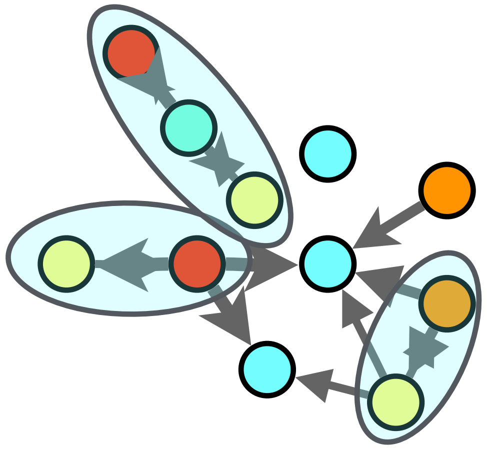
Summary

- HotNet2: Novel algorithm that analyzes topology and mutations simultaneously with asymmetric heat diffusion.



- Identifies known and novel pathways and complexes with frequently and rarely mutated genes on TCGA Pan-Cancer data.

- Future work:
  - Alternate graph partitioning algorithms?
  - Other applications: gene expression, GWAS, social networks, etc.



Acknowledgements

Research Group

Ben Raphael
Fabio Vandin
Hsin-Ta Wu
Jason R. Dobson
Matt Reyna
Jonathan Eldridge
Alexandra Papoutsaki
Jacob Thomas
Younhun Kim

BROWN

CCMB

Collaborators

Beifang Niu
Michael McLellan
Li Ding

THE GENOME INSTITUTE at Washington University

Michael Lawrence
Gad Getz

BROAD INSTITUTE

Nuria Lopez-Bigas
Abel Gonzalez-Perez
David Tamborero

upf. Universitat Pompeu Fabra Barcelona

Yuwei Chang
Greg Ryslik

Yale University

LUX ET VERITAS

Funding & Data

NSF Travel Fellowship to RECOMB 2015

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

NATIONAL INSTITUTES OF HEALTH

The Cancer Genome Atlas