Simultaneous Identification of Multiple Driver Pathways in Cancer

Mark D. M. Leiserson, *et.al*

Goal

- To distinguish the functional driver mutations responsible for cancer development from the random passenger mutations that have no consequences for cancer.

Distinguishing driver vs. passenger

- Strategies for identifying the driver mutations (Ding2014)
  - Recurrence and frequency assessment
  - Variant effect prediction
  - Pathway or network analysis

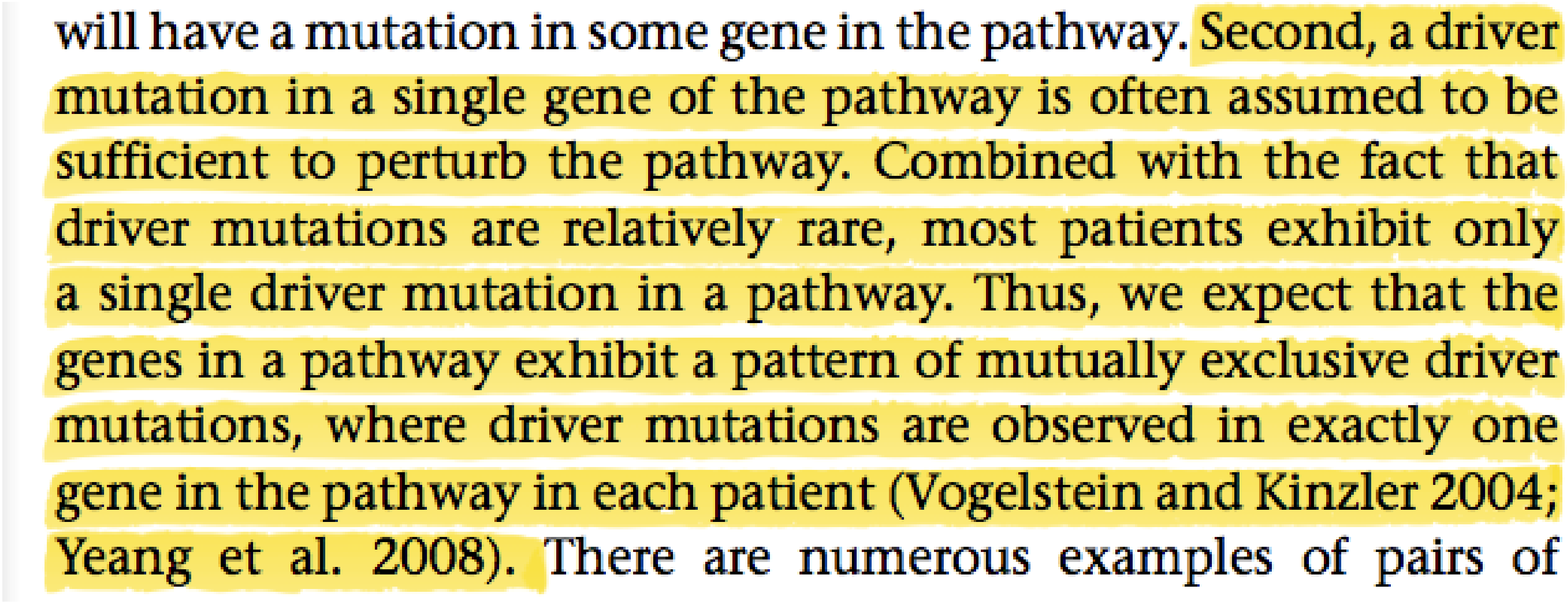
Multi-Dendrix

- Dendrix – De novo Driver Exclusivity
- **Important Assumption:**

1) High Coverage- most patients have at least one mutation in the set, i.e, set of potential mutated genes of a particular pathway

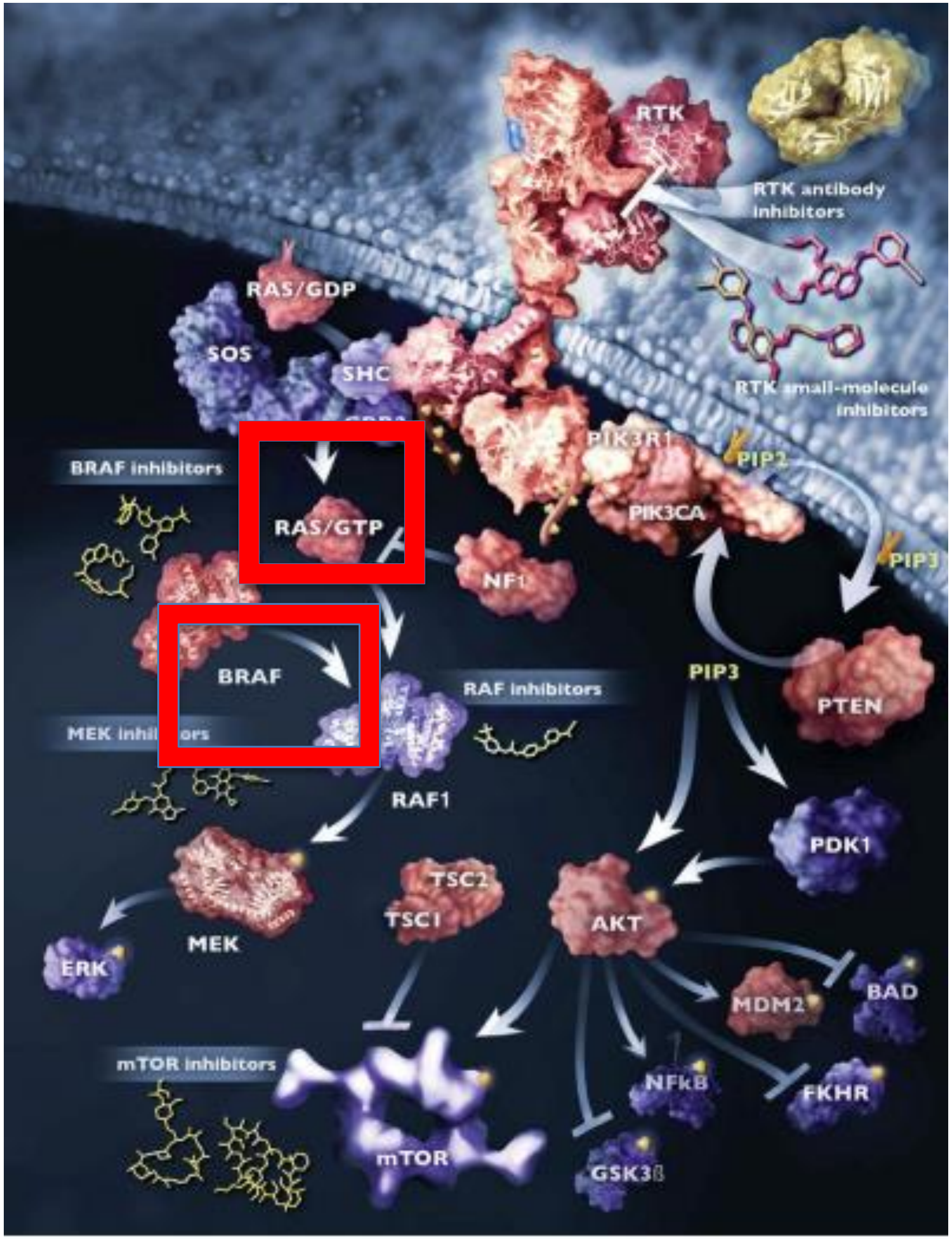
2) High Exclusivity- nearly all patients have no more than one mutation in the set

Justification by the author



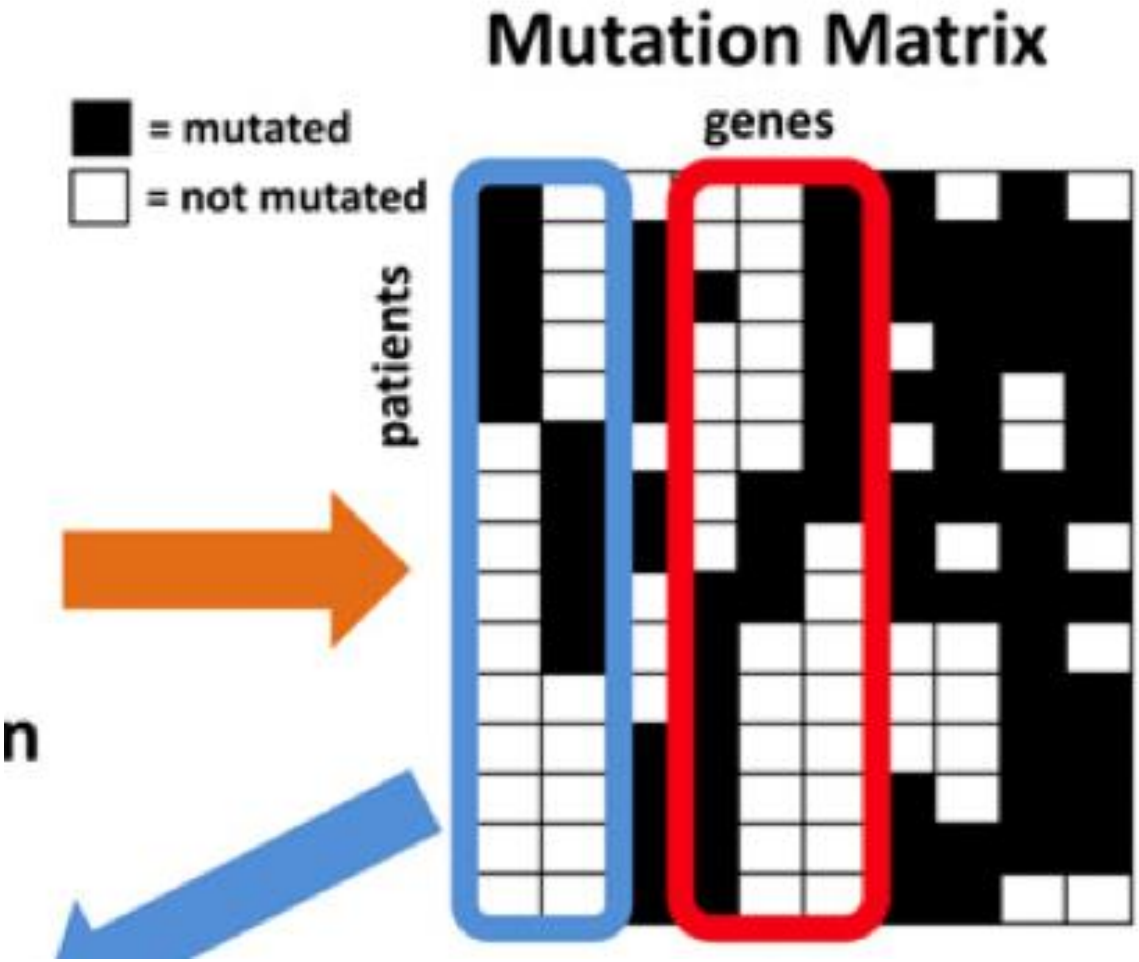
will have a mutation in some gene in the pathway. Second, a driver mutation in a single gene of the pathway is often assumed to be sufficient to perturb the pathway. Combined with the fact that driver mutations are relatively rare, most patients exhibit only a single driver mutation in a pathway. Thus, we expect that the genes in a pathway exhibit a pattern of mutually exclusive driver mutations, where driver mutations are observed in exactly one gene in the pathway in each patient (Vogelstein and Kinzler 2004; Yeang et al. 2008). There are numerous examples of pairs of

From Vandin, *et al,* 2012



Vogelstein2013

Dendrix - Method



From Vandin, *et al,* 2012

Dendrix Method

Maximum Coverage Exclusive Submatrix Problem: Given an m*n mutation matrix A and an integer k>0, find a mutually exclusive m*k submatrix of M of k columns (genes) of A with the largest number of nonzero rows (patients).

Coverage Overlap

Denote the set of patients in which g is mutated

$$\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|.$$

Denote the set of patients in which at least one of the genes in M is mutated

Weight

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|.$$

From Vandin, *et al,* 2012

Dendrix Method

Maximum Weight Submatrix Problem: Given an m * n mutation matrix A and an integer k >0, find the m * k column submatrix M of A that maximizes W (M).

Problem
Computationally Difficult to Solve

Size k = 6 of 20,000 genes

10^ 23 subsets

Solution
A greedy Algorithm for independent genes
Markov Chain Monte Carlo (MCMC)

From Vandin, *et al,* 2012

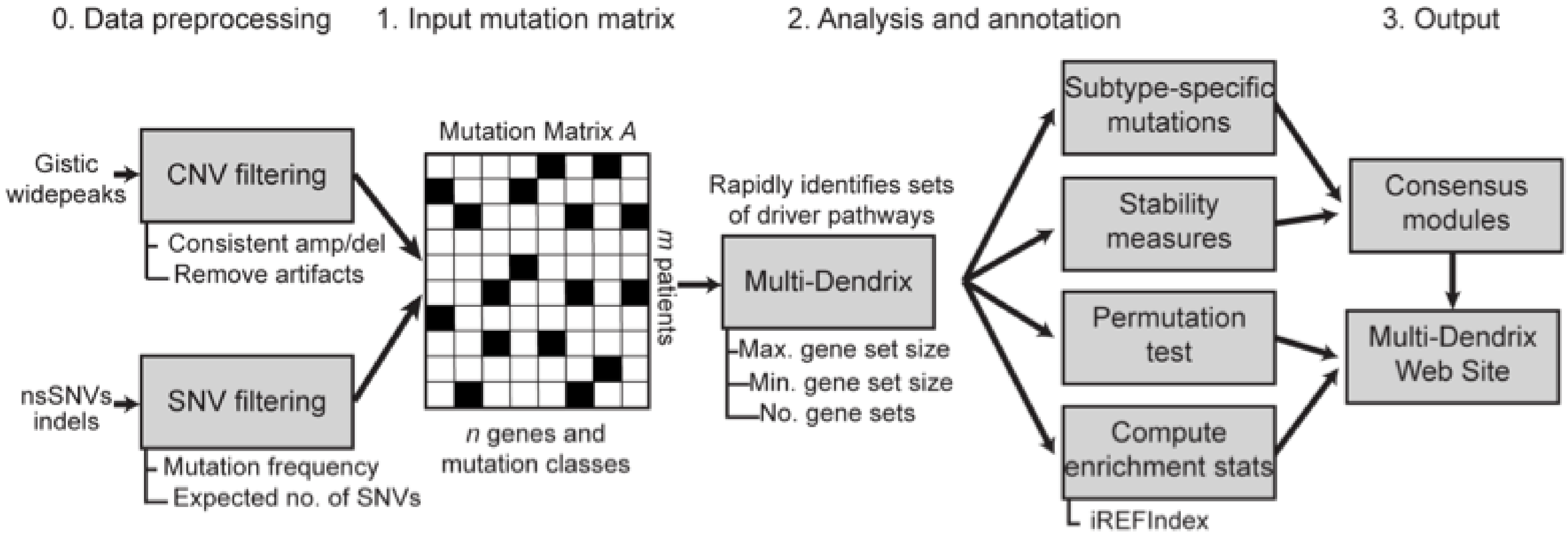
Limitation of Dendrix

- Mutations in different pathways may not be mutually exclusive.
- Mutations in different pathways may exhibit significant patterns of co-occurrence across patients.
- Solution -> Multi-Dendrix Algorithm

Multi-Dendrix Algorithm

- 1) Find sets of genes with high coverage as an integer linear program (ILP)
- 2) Generalize the ILP to simultaneously find multiple driver pathways
- 3) Additional Analysis: Subtype-specific mutations, stability measures, permutation test, compute enrichment states

The Multi-Dendrix Pipeline



Multi-Dendrix Method - the same as the first step of Dendrix

Maximum Coverage Exclusive Submatrix Problem: Given an m*n mutation matrix A and an integer k>0, find a mutually exclusive m*k submatrix of M of k columns (genes) of A with the largest number of nonzero rows (patients).

Coverage Overlap

Denote the set of patients in which g is mutated

$$\omega(M) = \sum_{g\in M}|\Gamma(g)| - |\Gamma(M)|.$$

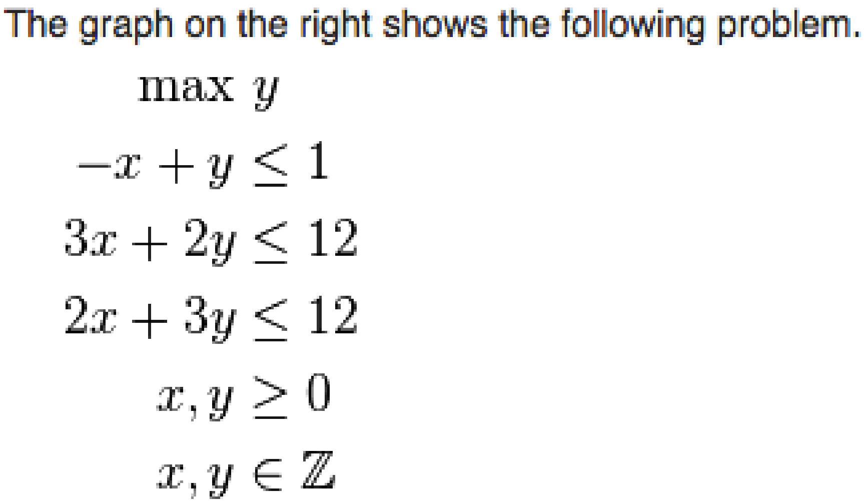
Denote the set of patients in which at least one of the genes in M is mutated

Weight

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g\in M}|\Gamma(g)|.$$

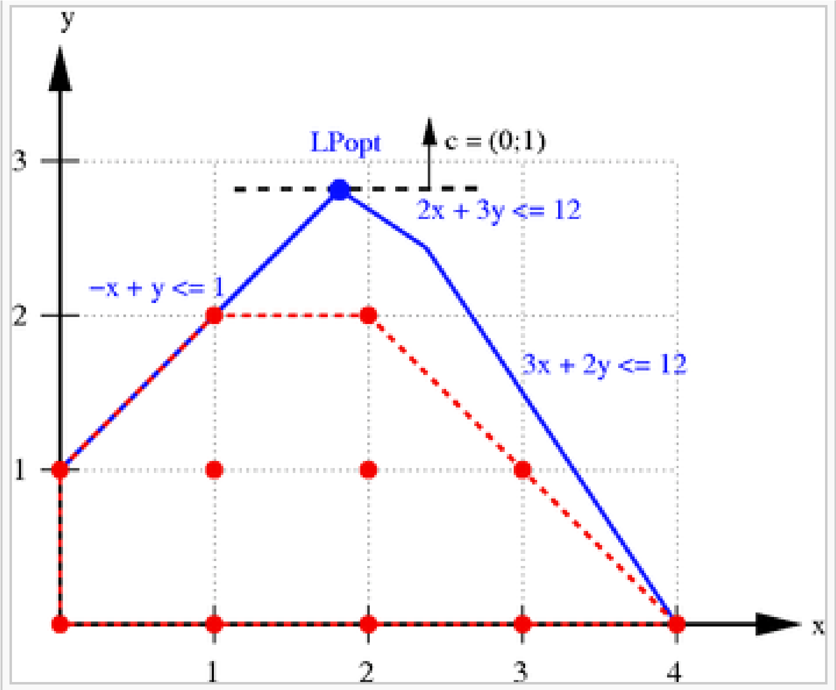
ILP- Integer Linear Programming

- Mathematical optimization or feasibility program where variables are restricted to be integers



The graph on the right shows the following problem.

$$\max\ y$$
$$-x + y \leq 1$$
$$3x + 2y \leq 12$$
$$2x + 3y \leq 12$$
$$x, y \geq 0$$
$$x, y \in \mathbb{Z}$$



IP polytope with LP relaxation

From Wikipedia

ILP for the Maximum Weight Submatrix Problem

Mutation matrix:

$$A_{ij} = \begin{cases} 1 & \text{if gene } j \text{ is mutated in patient } i \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

For each gene j, a gene set M is determined by

$$I_M(j) = \begin{cases} 1 & \text{if gene } j \text{ is a member of gene set } M, \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

For each patient I, the coverage is determined by

$$C_i(M) = \begin{cases} 1 & \text{if gene set } M \text{ is mutated in patient } i, \\ 0 & \text{otherwise.} \end{cases} \qquad (4)$$

Then, $Dendrix_{ILP}(k)$ is defined as follows:

$$\text{maximize} \sum_{i=1}^{m} \left(2 \cdot C_i(M) - \sum_{j=1}^{n} I_M(j) \cdot A_{ij} \right) \qquad \text{(5a)}$$

$$k_{\min} \leq \sum_{j=1}^{n} I_M(j) \leq k_{\max}.$$

←

$$\text{subject to} \sum_{j=1}^{n} I_M(j) = k \qquad \text{(5b)}$$

Denote the set of patients in which g is mutated

$$\omega(M) = \textstyle\sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|.$$

Denote the set of patients in which at least one of the genes in M is mutated

$$\left(\sum_{j=1}^{n} A_{ij} \cdot I_M(j) \right) \geq C_i(M), \qquad \text{(5c)}$$

$$\text{for } 1 \leq i \leq m.$$

Multiple Maximum Weight Submatrices Problem

Multiple Maximum Weight Submatrices Problem: Given an m*n mutation matrix A and an integer t>0, find a collection M = { M1, M2, …., Mt} of m*k column submatrices that maximizes $W'(\mathbf{M}) = \sum_{\rho=1}^{t} W(M_\rho)$.

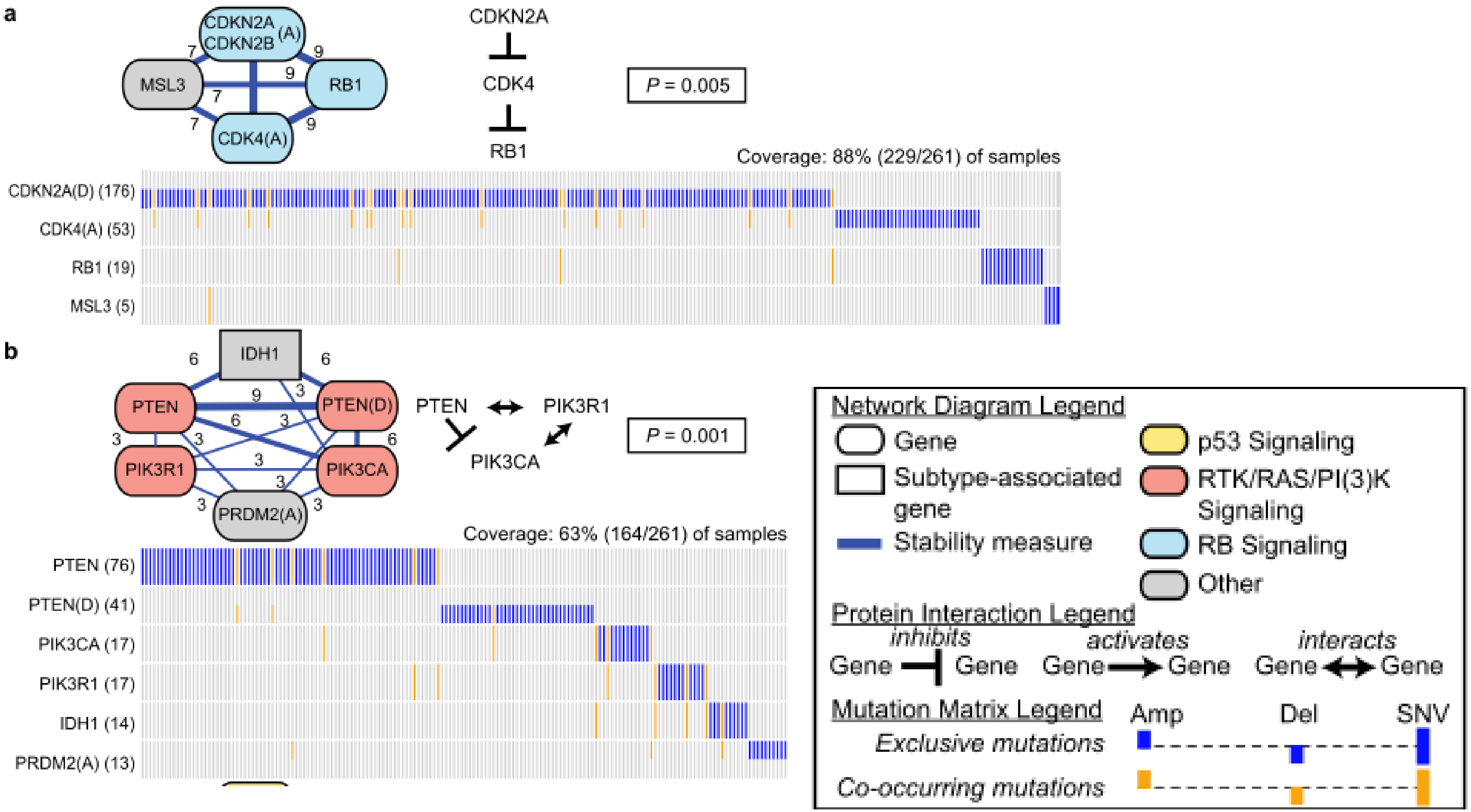
$$\text{maximize} \sum_{\rho=1}^{t} \sum_{i=1}^{m} \left(2 \cdot C_i(M_\rho) - \sum_{j=1}^{n} I_{M_\rho}(j) \cdot A_{ij} \right) \qquad (7a)$$

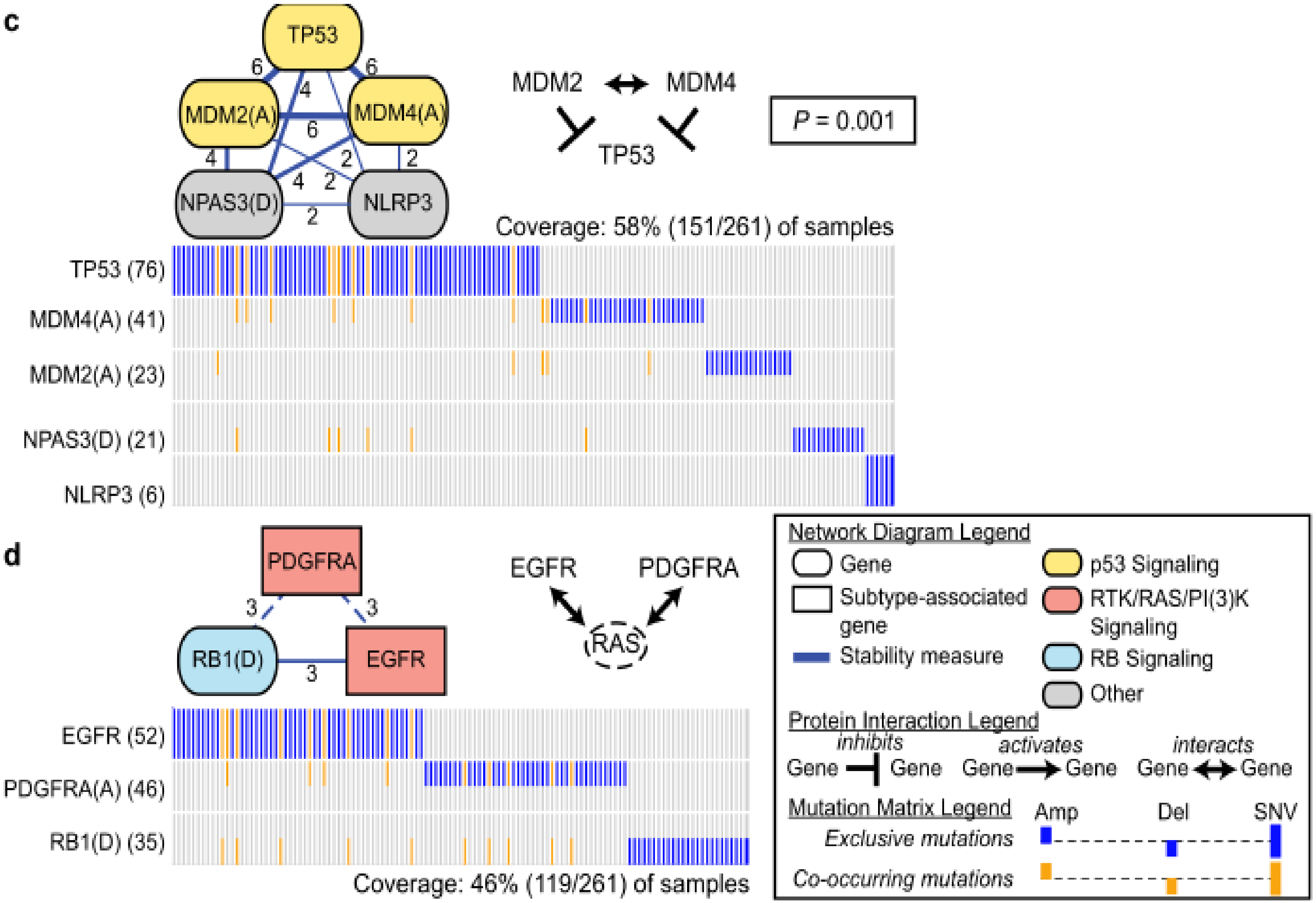
$$\text{subject to} \left(\sum_{j=1}^{n} I_{M_\rho}(j) \cdot A_{ij} \right) \geq C_i(M_\rho), \qquad (7b)$$

$$\text{for } 1 \leq i \leq m, 1 \leq \rho \leq t,$$

$$\sum_{\rho=1}^{t} I_{M_\rho}(j) \leq 1, 1 \leq j \leq m. \qquad (7c)$$

Multi-Dendrix Results on the GBM Dataset




c
TP53
MDM2(A)
MDM4(A)
NPAS3(D)
NLRP3
6
6
4
6
4
2
4
2
2
2
MDM2
MDM4
TP53
P = 0.001
Coverage: 58% (151/261) of samples
TP53 (76)
MDM4(A) (41)
MDM2(A) (23)
NPAS3(D) (21)
NLRP3 (6)
d
PDGFRA
RB1(D)
EGFR
3
3
3
EGFR
PDGFRA
RAS
EGFR (52)
PDGFRA(A) (46)
RB1(D) (35)
Coverage: 46% (119/261) of samples
Network Diagram Legend
Gene
Subtype-associated gene
Stability measure
p53 Signaling
RTK/RAS/PI(3)K Signaling
RB Signaling
Other
Protein Interaction Legend
inhibits
Gene ⊣ Gene
activates
Gene → Gene
interacts
Gene ↔ Gene
Mutation Matrix Legend
Amp
Del
SNV
Exclusive mutations
Co-occurring mutations

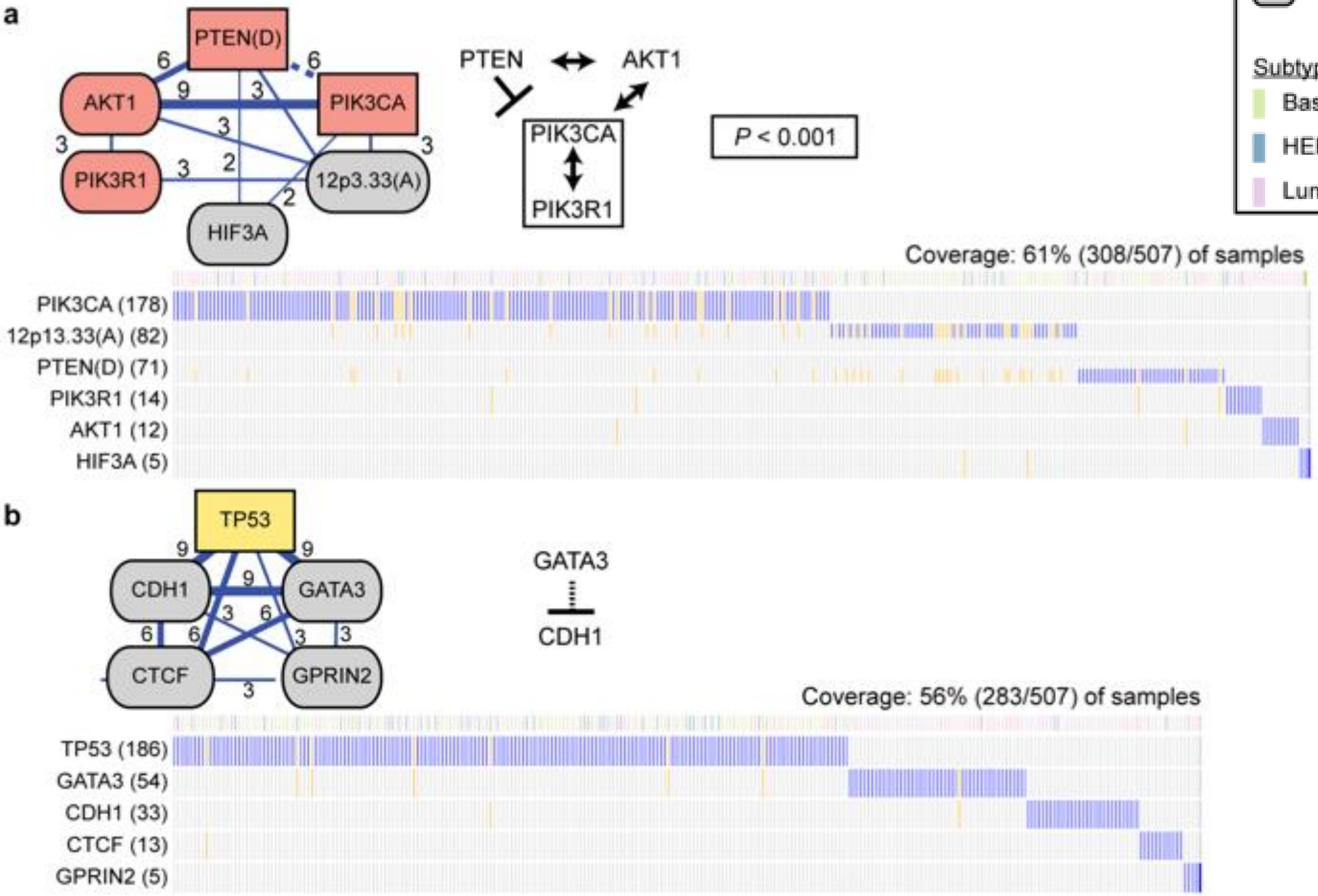
Multi-Dendrix Results on the BRCA Dataset

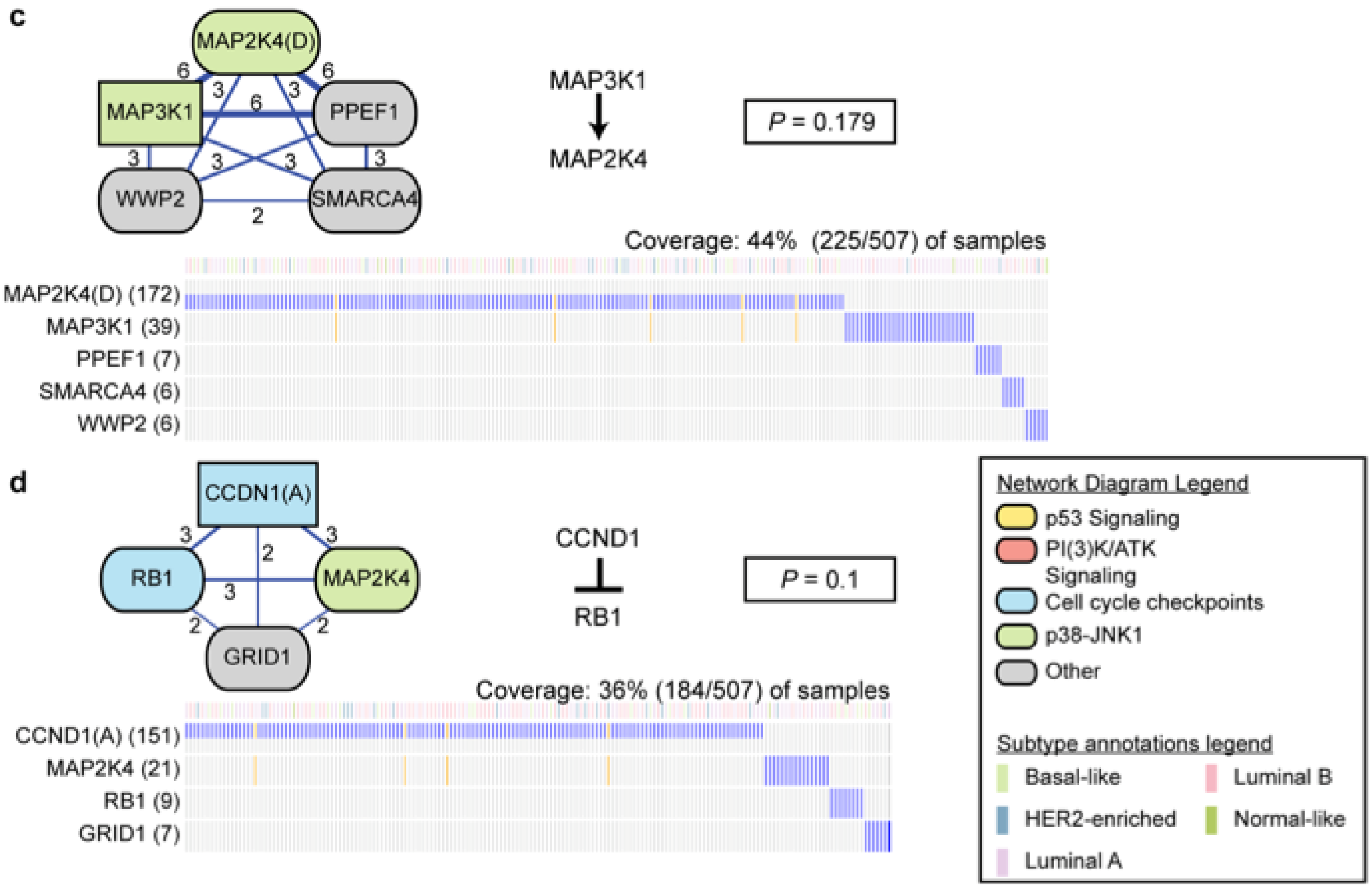
Network Diagram Legend

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

Subtype annotations legend

🟩 Basal-like	🟥 Luminal B
🟦 HER2-enriched	🟩 Normal-like
🟪 Luminal A	




c
MAP2K4(D)
MAP3K1
PPEF1
WWP2
SMARCA4
6
6
6
3
3
3
3
3
3
2
MAP3K1
MAP2K4
P = 0.179
Coverage: 44% (225/507) of samples
MAP2K4(D) (172)
MAP3K1 (39)
PPEF1 (7)
SMARCA4 (6)
WWP2 (6)
d
CCDN1(A)
RB1
MAP2K4
GRID1
3
3
2
3
2
2
CCND1
RB1
P = 0.1
Coverage: 36% (184/507) of samples
CCND1(A) (151)
MAP2K4 (21)
RB1 (9)
GRID1 (7)
Network Diagram Legend
p53 Signaling
PI(3)K/ATK Signaling
Cell cycle checkpoints
p38-JNK1
Other
Subtype annotations legend
Basal-like
Luminal B
HER2-enriched
Normal-like
Luminal A

- Thank you for your attention!