

# Identifying multi-hit cancer drivers without massive parallelization: a CP, MIP, and column generation framework

Rick Willemsen<sup>1</sup>   Tenindra Abeywickrama<sup>2</sup>   Ramu Anandakrishnan<sup>3</sup>

<sup>1</sup>Singapore University of Technology and Design, Singapore

<sup>2</sup>RIKEN Center for Computational Science, Japan

<sup>3</sup>Virginia Tech, Edward Via College of Osteopathic Medicine, USA

willemsen@ese.eur.nl

July 22, 2026

Cancer caused 9.7 million deaths worldwide in 2022<sup>1</sup>

---

<sup>1</sup>Bray et al. (2024)

# Multi-hit theory of carcinogenesis

## Multi-hit theory

Multiple mutated genes (called hits) are required to cause cancer<sup>2</sup>

## Estimated hit sizes

Ranges from 2 to 9, depending on the cancer type<sup>3</sup>

---

<sup>2</sup>Armitage and Doll (1954); Nordling (1953); Sutherland and Bailar III (1984)

<sup>3</sup>Little and Wright (2003); Zhang and Simon (2005); Tomasetti et al. (2015); Ashley (1969); Anandakrishnan et al. (2019)

# Multi-hit theory - example

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $t$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $n$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |

# Multi-hit theory - example

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $t$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $n$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |

# Multi-hit theory - example

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $t$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |
|     |  | 1     | 0     | 0     | 1     | 0     | 0     | 0     |

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $n$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |

# Multi-hit theory - example

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $t$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |
|     |  | 1     | 0     | 0     | 1     | 0     | 0     | 0     |
|     |  | 1     | 1     | 0     | 1     | 0     | 0     | 0     |

|     |  | genes |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|
|     |  | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
| $n$ |  | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |
|     |  | 0     | 0     | 0     | 1     | 0     | 0     | 0     |
|     |  | 0     | 0     | 1     | 1     | 0     | 0     | 0     |

# Multi-hit theory - example

tumor sample

genes

|                       | $g_1$ | $g_2$ | $g_3$              | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
|-----------------------|-------|-------|--------------------|-------|-------|-------|-------|
| $t$                   | 0     | 0     | 0                  | 0     | 0     | 0     | 0     |
|                       | 0     | 0     | 0                  | 1     | 0     | 0     | 0     |
|                       | 1     | 0     | 0                  | 1     | 0     | 0     | 0     |
|                       | 1     | 1     | 0                  | 1     | 0     | 0     | 0     |
| mutations of interest |       |       | harmless mutations |       |       |       |       |

normal sample

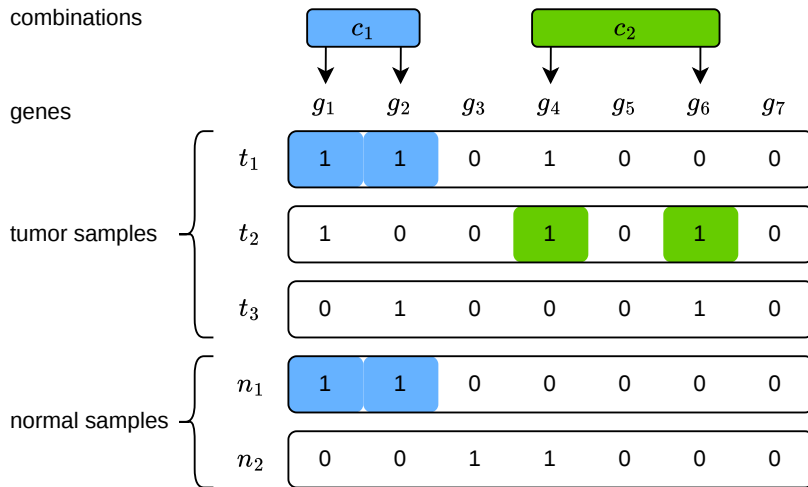
genes

|     | $g_1$ | $g_2$ | $g_3$              | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
|-----|-------|-------|--------------------|-------|-------|-------|-------|
| $n$ | 0     | 0     | 0                  | 0     | 0     | 0     | 0     |
|     | 0     | 0     | 0                  | 1     | 0     | 0     | 0     |
|     | 0     | 0     | 0                  | 1     | 0     | 0     | 0     |
|     | 0     | 0     | 1                  | 1     | 0     | 0     | 0     |
|     |       |       | harmless mutations |       |       |       |       |

# Multi-hit theory - example

| genes          |       | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ |
|----------------|-------|-------|-------|-------|-------|-------|-------|-------|
| tumor samples  | $t_1$ | 1     | 1     | 0     | 1     | 0     | 0     | 0     |
|                | $t_2$ | 1     | 0     | 0     | 1     | 0     | 1     | 0     |
|                | $t_3$ | 0     | 1     | 0     | 0     | 0     | 1     | 0     |
| normal samples | $n_1$ | 1     | 1     | 0     | 0     | 0     | 0     | 0     |
|                | $n_2$ | 0     | 0     | 1     | 1     | 0     | 0     | 0     |

# Multi-hit theory - example



# Table of contents

- 1 Existing methods
- 2 Classification problem
- 3 Column generation framework
- 4 Computational experiments
- 5 Conclusion & future work

# Table of contents

- 1 Existing methods
- 2 Classification problem
- 3 Column generation framework
- 4 Computational experiments
- 5 Conclusion & future work

# Existing methods

Exhaustive-based enumeration of gene combinations  $\implies$  greedy heuristic

- Dash et al. (2019)
- Al Hajri et al. (2020)
- Dash et al. (2021): 6,000 V100 GPUs ( $\sim 48$  million GPU cores)
- Oles et al. (2025): 280 nodes on the Summit supercomputer ( $\sim 11,000$  CPU cores)
- Prabhu et al. (2026): 3,072 nodes on the Fugaku supercomputer ( $\sim 147,000$  CPU cores)

We obtain near-optimal solutions on a regular laptop within a minute

# Contributions

Our contributions:

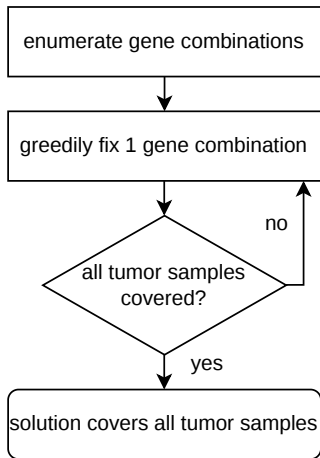
- Propose a column generation framework
- First to solve practical instances on regular laptop
- First to provide optimality bounds (price-and-branch heuristic)

# Table of contents

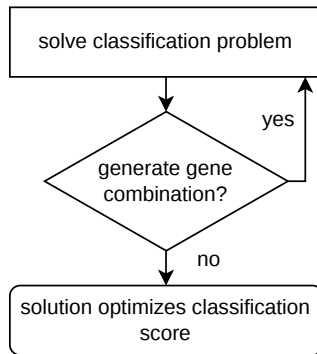
- 1 Existing methods
- 2 Classification problem
- 3 Column generation framework
- 4 Computational experiments
- 5 Conclusion & future work

# Classification problem

## existing methods



## intuition behind our approach



# Performance measures

- specificity
- sensitivity
- F1 score
- Matthews Correlation Coefficient (MCC)

## Linear objective function

Maximize the number of correctly classified tumor samples (true positives)

Minimize the number of times a normal sample is misclassified (approximating false positives)

# Table of contents

- 1 Existing methods
- 2 Classification problem
- 3 Column generation framework**
- 4 Computational experiments
- 5 Conclusion & future work

# Constraint programming formulation

$$\max \sum_{t \in T} x_t - \sum_{n \in N} y_n$$

$\Rightarrow$  max true positives, min false positives

$$\text{s.t. } x_t = \max_{c \in C_t} z_c$$

$$\forall t \in T$$

$\Rightarrow$  try to cover each tumor sample at least once

$$y_n = \sum_{c \in C_n} z_c$$

$$\forall n \in N$$

$\Rightarrow$  number of times a normal sample is covered

$$\sum_{c \in C} z_c \leq \beta$$

$\Rightarrow$  limit number of gene combinations to  $\beta$

$$x_t \in \{0, 1\}$$

$$\forall t \in T$$

$$y_n \in \mathbb{N}$$

$$\forall n \in N$$

$$z_c \in \{0, 1\}$$

$$\forall c \in C$$

$\Rightarrow$  select gene combinations

# Mixed integer programming formulation

$$\max \sum_{t \in T} x_t - \sum_{n \in N} y_n$$

$$\text{s.t. } x_t \leq \sum_{c \in C_t} z_c \quad \forall t \in T$$

$$\sum_{c \in C_n} z_c \leq y_n \quad \forall n \in N$$

$$\sum_{c \in C} z_c \leq \beta$$

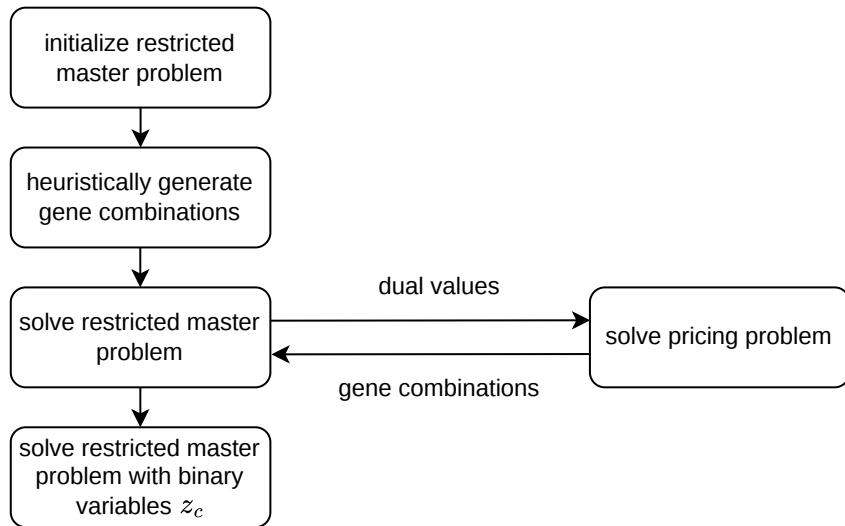
$$x_t \in \{0, 1\} \quad \forall t \in T$$

$$y_n \in \mathbb{N} \quad \forall n \in N$$

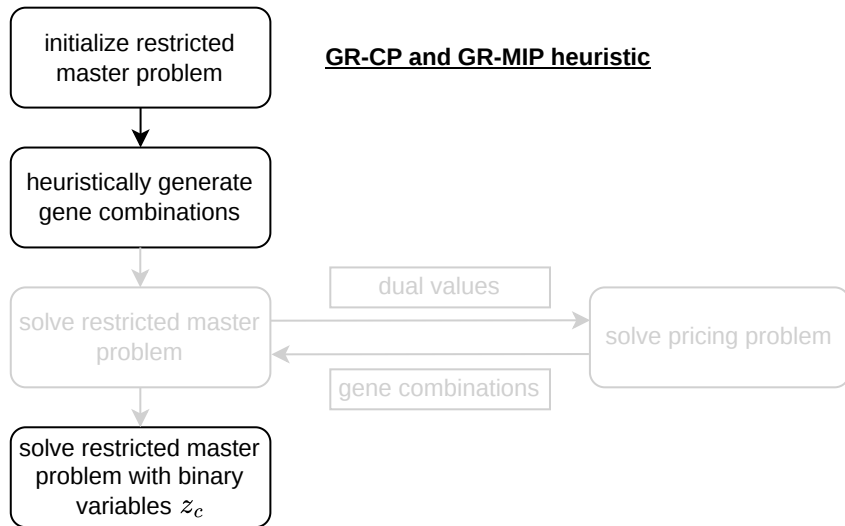
$$z_c \in \{0, 1\} \quad \forall c \in C$$

Many gene combinations ( $C$ )  $\implies$  column generation

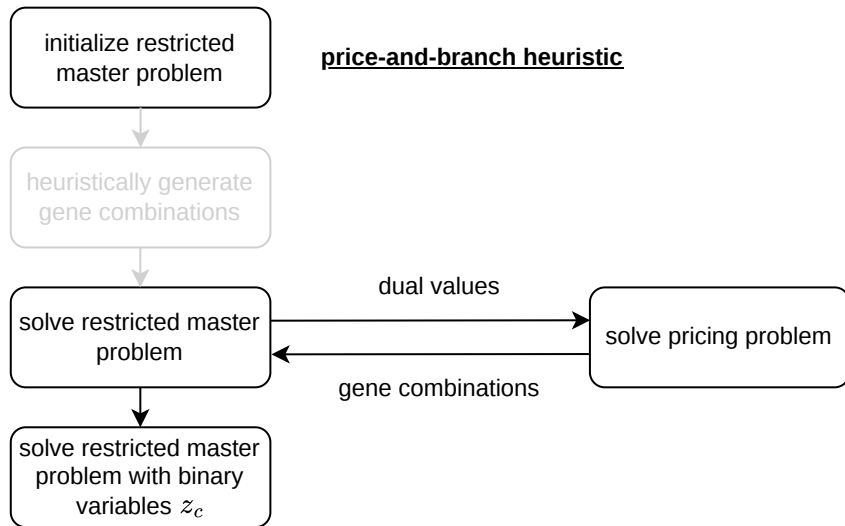
# Column generation framework



# Column generation framework



# Column generation framework



# Pricing problem formulation

$$\max \sum_{t \in T} \pi_t v_t - \sum_{n \in N} \mu_n v_n - \lambda$$

$\Rightarrow$  max tumor samples, min normal samples

$$\text{s.t. } \sum_{g \in G} u_g \geq \underline{k}$$

$\Rightarrow$  combination contains at least  $\underline{k}$  genes

$$\sum_{g \in G} u_g \leq \bar{k}$$

$\Rightarrow$  combination contains at most  $\bar{k}$  genes

$$v_t \leq 1 - \frac{1}{\bar{k}} \sum_{g \in G: \mathbf{M}_{tg}=0} u_g \quad \forall t \in T$$

$\Rightarrow$  tumor sample cannot be covered

$$v_n \geq 1 - \sum_{g \in G: \mathbf{M}_{ng}=0} u_g \quad \forall n \in N$$

$\Rightarrow$  normal sample is covered

$$v_t \in \{0, 1\} \quad \forall t \in T$$

$$v_n \in \{0, 1\} \quad \forall n \in N$$

$$u_g \in \{0, 1\} \quad \forall g \in G$$

$\Rightarrow$  select genes

# Table of contents

- 1 Existing methods
- 2 Classification problem
- 3 Column generation framework
- 4 Computational experiments**
- 5 Conclusion & future work

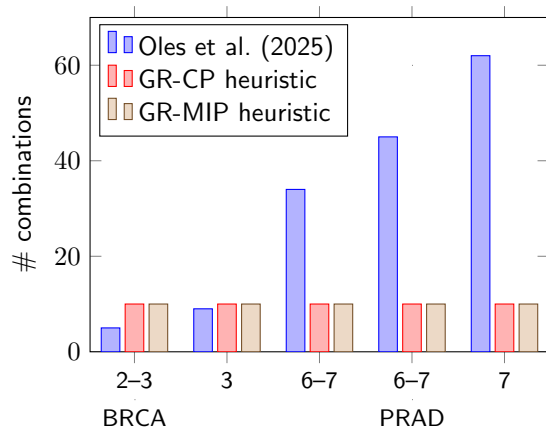
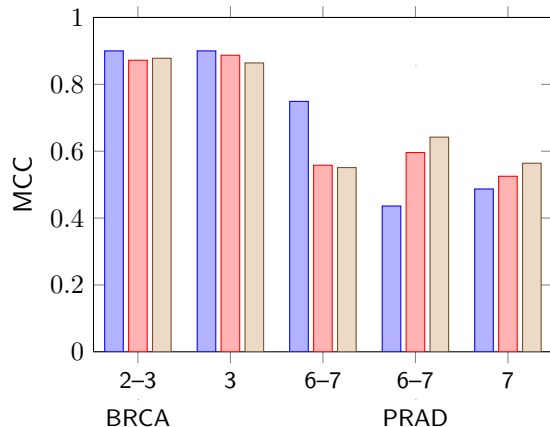
# Experimental setup

- The Cancer Genome Atlas<sup>5</sup>
  - Dataset A: 2 cancer types
  - Dataset B: 16 cancer types
- Hit size is fixed (Oles et al., 2025)
- Select at most  $\beta = 10$  gene combinations
- GR-MIP heuristic: greedily generate 100,000 gene combinations
- Train-test split (75%, 25%)

| dataset | cancer type | #hits       | #genes | #samples | #mutations |
|---------|-------------|-------------|--------|----------|------------|
| A       | BRCA        | 3, 2–3, 6–7 | 19,411 | 1242     | 837,207    |
|         | PRAD        | 6–7, 7      | 17,274 | 752      | 476,115    |
| B (avg) | -           | -           | 19,823 | 738      | 597,230    |

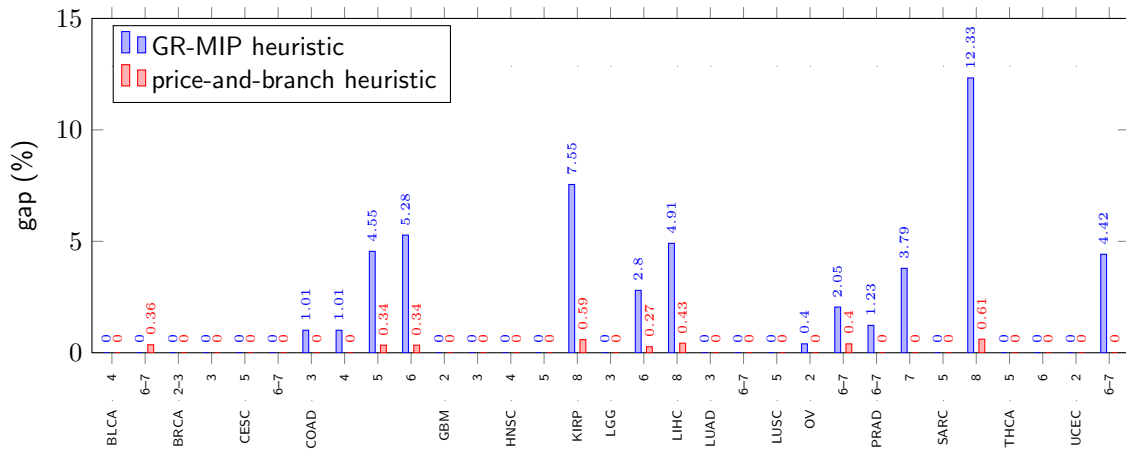
<sup>5</sup>Weinstein et al. (2013)

# Comparison to state-of-the-art – dataset A



Similar test set performance on a single CPU. CP within 20 minutes and MIP within 1 minute.

# Optimality gaps – dataset B



Price-and-branch found a valid upper bound for 31 out of 33 instances within 5 minutes

# Table of contents

- 1 Existing methods
- 2 Classification problem
- 3 Column generation framework
- 4 Computational experiments
- 5 Conclusion & future work

# Conclusion & future work

Identifying multi-hit combinations is less computationally demanding than previously believed

- More accessible to practitioners

Future work

- Explore complex biological constraints using constraint programming
- Exact method that guarantees optimality
- Prevent overfitting

# Thank You!

Willemsen, R. S. H., Abeywickrama, T., and Anandakrishnan, R. (2026). *Identifying multi-hit cancer drivers without massive parallelization: A CP, MIP, and column generation framework*. Accepted for publication at CP2026.



# References I

- Al Hajri, Q., Dash, S., Feng, W.-c., Garner, H. R., and Anandakrishnan, R. (2020). Identifying multi-hit carcinogenic gene combinations: Scaling up a weighted set cover algorithm using compressed binary matrix representation on a gpu. *Scientific Reports*, 10(1):2022.
- Anandakrishnan, R., Varghese, R. T., Kinney, N. A., and Garner, H. R. (2019). Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations. *PLoS Computational Biology*, 15(3):e1006881.
- Armitage, P. and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 8(1):1–12.
- Ashley, D. (1969). The two "hit" and multiple "hit" theories of carcinogenesis. *British Journal of Cancer*, 23(2):313.
- Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263.
- Dash, S., Al-Hajri, Q., Feng, W.-c., Garner, H. R., and Anandakrishnan, R. (2021). Scaling out a combinatorial algorithm for discovering carcinogenic gene combinations to thousands of gpus. In *Proceedings of the 35th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 837–846. IEEE.
- Dash, S., Kinney, N. A., Varghese, R. T., Garner, H. R., Feng, W.-c., and Anandakrishnan, R. (2019). Differentiating between cancer and normal tissue samples using multi-hit combinations of genetic mutations. *Scientific Reports*, 9(1):1005.

# References II

- Little, M. and Wright, E. (2003). A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Mathematical Biosciences*, 183(2):111–134.
- Nordling, C. (1953). A new theory on cancer-inducing mechanism. *British Journal of Cancer*, 7(1):68–72.
- Oles, V., Dash, S., and Anandakrishnan, R. (2025). Bigpicc: a graph-based approach to identifying carcinogenic gene combinations from mutation data. *BMC Bioinformatics*, 26(1):155.
- Prabhu, R., Vatai, E., Moussad, B., Jeannot, E., Anandakrishnan, R., Feng, W.-c., and Wahib, M. (2026). Looking for (genomic) needles in a haystack: Sparsity-driven search for identifying correlated genetic mutations in cancer. *arXiv preprint arXiv:2603.16721*.
- Sutherland, J. V. and Bailer III, J. C. (1984). The multihit model of carcinogenesis: etiologic implications for colon cancer. *Journal of Chronic Diseases*, 37(6):465–480.
- Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G., and Vogelstein, B. (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123.
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.
- Zhang, X. and Simon, R. (2005). Estimating the number of rate limiting genomic changes for human breast cancer. *Breast Cancer Research and Treatment*, 91(2):121–124.

# Performance measures

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{TN + FP}$$

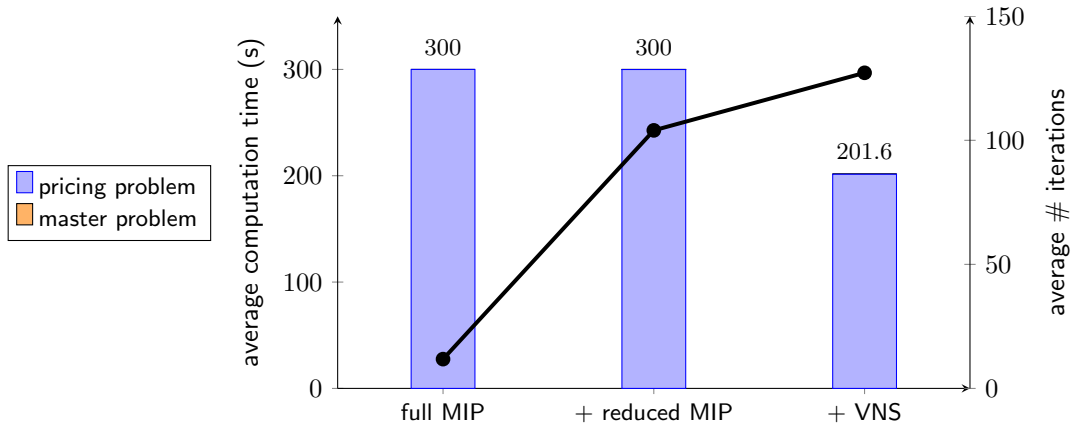
$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Note: precision is not used as a performance measure, but is used to define the F1 score

# Accelerating the pricing problem – dataset A



Pricing problem speedups are needed to obtain a valid upper bound within time limit