Optimization in the Race to a Liquid Biopsy

Kyra Gan
Cornell Tech

Su Jia
Cornell

Andrew Li
CMU

Sridhar Tayur
CMU

A Little Biology

- Cancer is caused by DNA mutations
  - Tumors contain mutated DNA

A Little Biology

- Cancer is caused by DNA mutations
  - Tumors contain mutated DNA
- **Cell-free DNA:** blood contains tiny amounts of mutated DNA
  - Concentration of one in ten-thousand
  - Should hypothetically be a cancer signal

The Race is On

- Recent academic successes based on this idea:

  [Cohen et al.] Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (2018)*.

  [Liu et al.] Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology (2020)*.

The Race is On

- Recent academic successes based on this idea:

  [Cohen et al.] Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (2018)*.

Forbes

Jan 18, 2018, 02:00pm EST

A New $500 Blood Test Could Detect Cancer Before Symptoms Develop

The Race is On

- Recent academic successes based on this idea:

  [Cohen et al.] Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (2018)*.


Forbes

Jan 18, 2018, 02:00pm EST

A New $500 Blood Test Could Detect Cancer Before Symptoms Develop

- …and biotech firms building on it:

GRAIL GUARDANT EXACT SCIENCES freenome

Nearing the Finish Line

- Grail's *Galleri* test

Nearing the Finish Line

- Grail's *Galleri* test
  - $949

The cost of the Galleri® test may vary depending on the healthcare practice or provider who orders the test. The list price for the Galleri test is $949.

Nearing the Finish Line

- Grail's *Galleri* test
  - $949
  - Detects 50+ cancer types
  - Specificity 99.5%

50+
cancer types

99.5%
specificity[1]

Nearing the Finish Line

- Grail's *Galleri* test
  - $949
  - Detects 50+ cancer types
  - Specificity 99.5%
  - Sensitivity 77%

Sensitivity

76.3% sensitivity in cancers that cause two-thirds of cancer deaths in the US [1,6,7]

Data Available Today

- DNA is 3 billion addresses long:

  ... AGCATGCAGTACGTACGTCACATTCGATCGATGG...

Data Available Today

- DNA is 3 billion addresses long:

  ... AGCATGCAGTACGTACGTCACATTCGATCGATGG...

- Mutations are with respect to a reference sequence

  ... AG**G**ATGCAGT**C**CGTACGTCACATTC**A**ATCGATGG...

Data Available Today

- DNA is 3 billion addresses long:

  ... AGCATGCAGTACGTACGTCACATTCGATCGATGG...

- Mutations are with respect to a reference sequence

  …00**1**0000000**1**0000000000000**1**00000000…

Data Available Today

- DNA is 3 billion addresses long:

  ... AGCATGCAGTACGTACGTCACATTCGATCGATGG...

- Mutations are with respect to a reference sequence

  …00**1**0000000**1**0000000000000**1**00000000…

- Many (hundreds of thousands) of tumors have been sequenced

Data Available Today

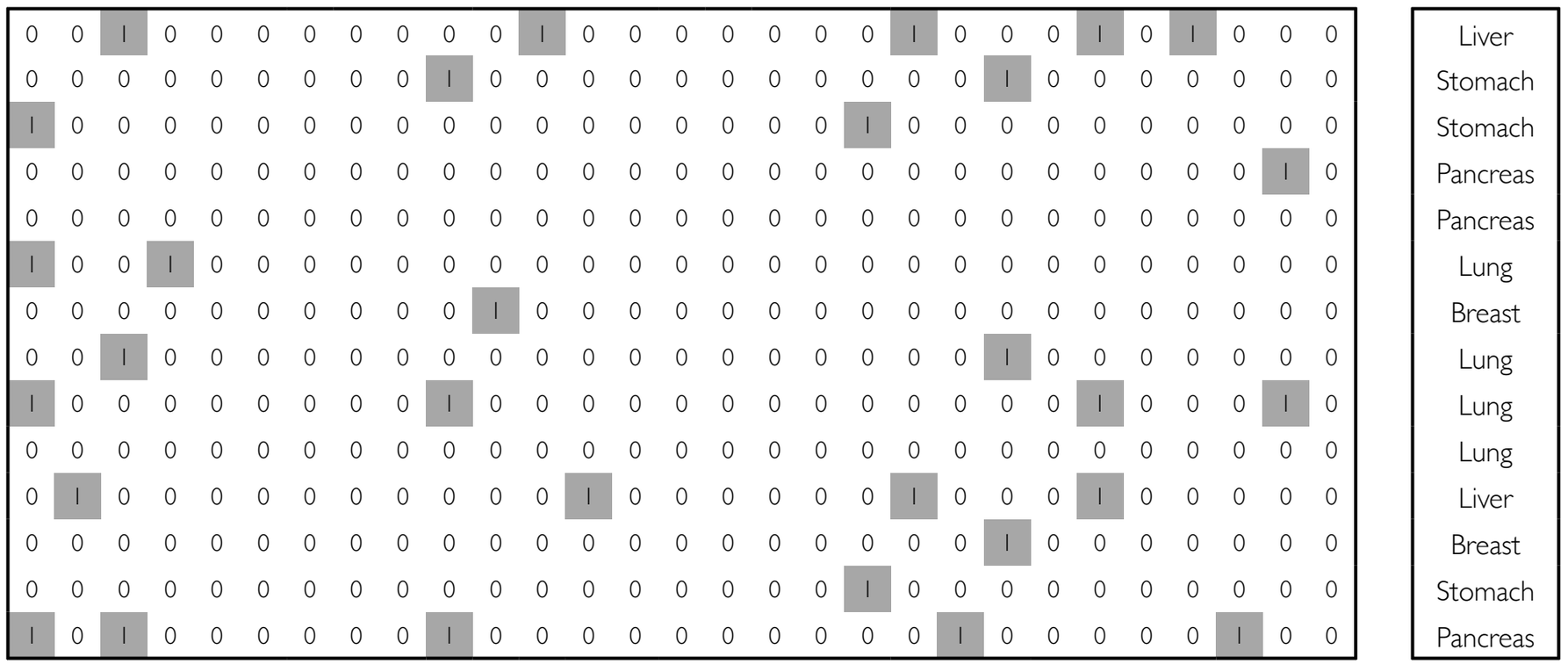
- DNA is 3 billion addresses long:

  ... AGCATGCAGTACGTACGTCACATTCGATCGATGG...

- Mutations are with respect to a reference sequence

  …00**1**0000000**1**000000000000000**1**00000000…

- Many (hundreds of thousands) of tumors have been sequenced



Challenge: Cost

- A back-of-the-envelope calculation:
  - Test must cost at most $\$10^2$

0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	Liver
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Stomach
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Stomach
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	Pancreas
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Pancreas
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Lung
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Breast
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Lung
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	Lung
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Lung
0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	Liver
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Breast
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Stomach
1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	Pancreas

Challenge: Cost

- A back-of-the-envelope calculation:
  - Test must cost at most $\$10^2$
  - Sequencing an address ten-thousand times costs $\$10^{-2}$

0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0

Liver
Stomach
Stomach
Pancreas
Pancreas
Lung
Breast
Lung
Lung
Lung
Liver
Breast
Stomach
Pancreas

Challenge: Cost

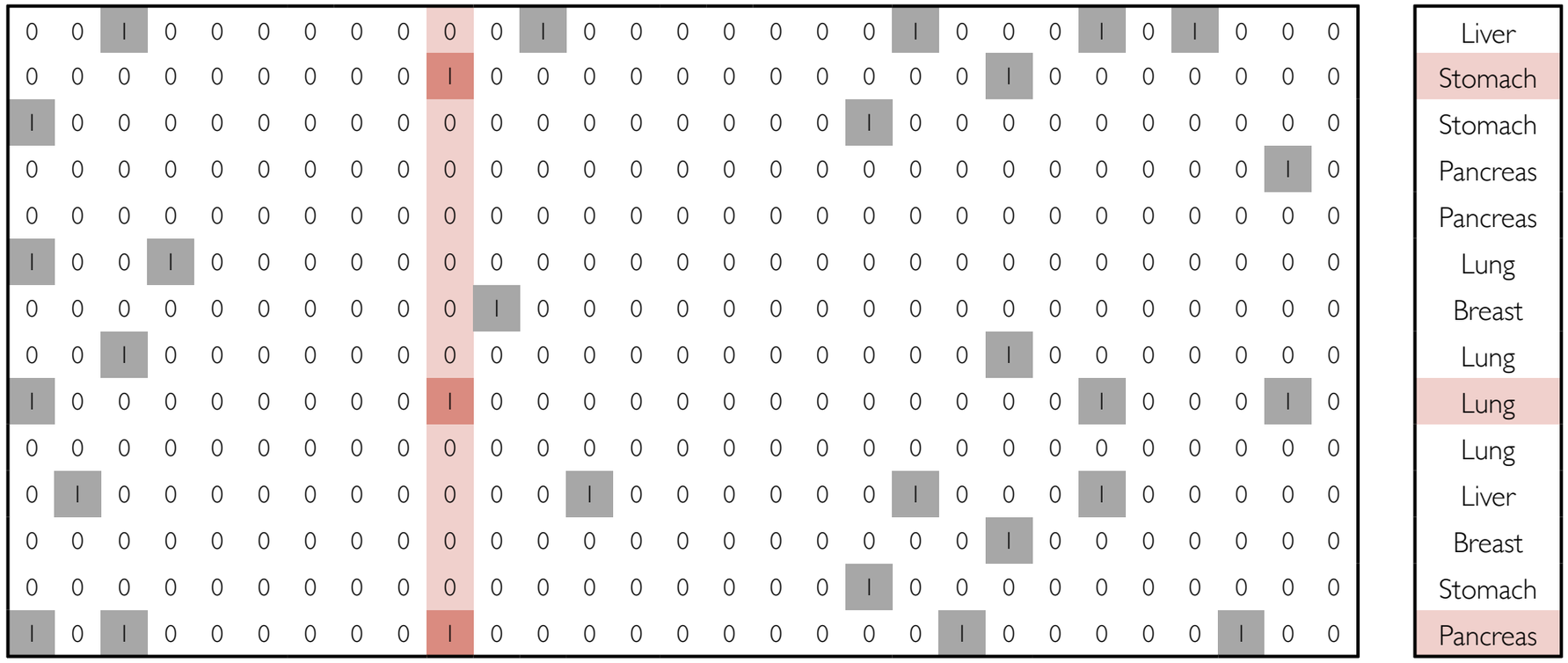
- A back-of-the-envelope calculation:
  - Test must cost at most $\$10^2$
  - Sequencing an address ten-thousand times costs $\$10^{-2}$
  - Thus, can only use a **panel** of ~10^4 addresses

0	0	I	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0	I	0	0	0	I	0	I	0	0	0
0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0	0	0	0	0	0	0	I	0	0	0	I	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	I	0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	I	0	0	0	I	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	0	0	0	0
I	0	I	0	0	0	0	0	0	I	0	0	0	0	0	0	0	0	0	0	I	0	0	0	0	0	I	0	0

Liver
Stomach
Stomach
Pancreas
Pancreas
Lung
Breast
Lung
Lung
Lung
Liver
Breast
Stomach
Pancreas

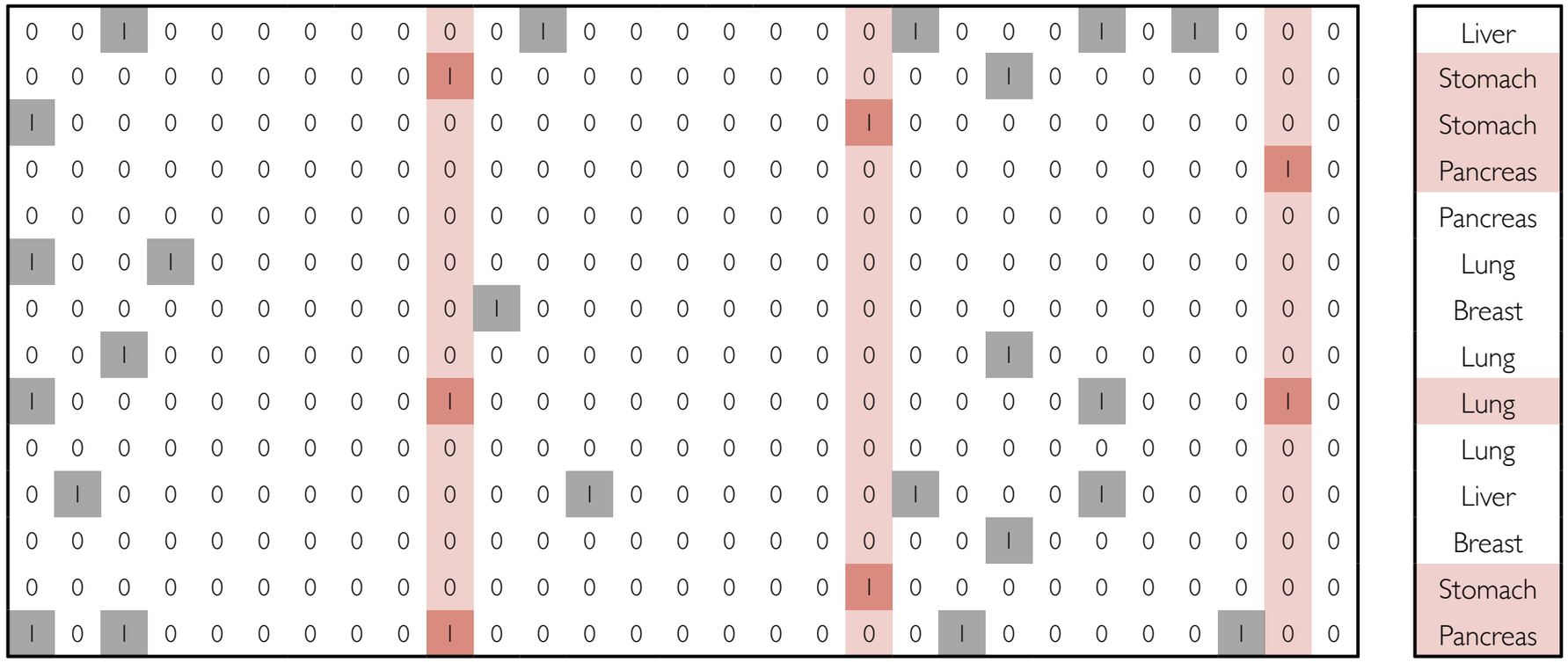
Challenge: Cost

- A back-of-the-envelope calculation:
  - Test must cost at most $\$10^2$
  - Sequencing an address ten-thousand times costs $\$10^{-2}$
  - Thus, can only use a **panel** of ~10^4 addresses

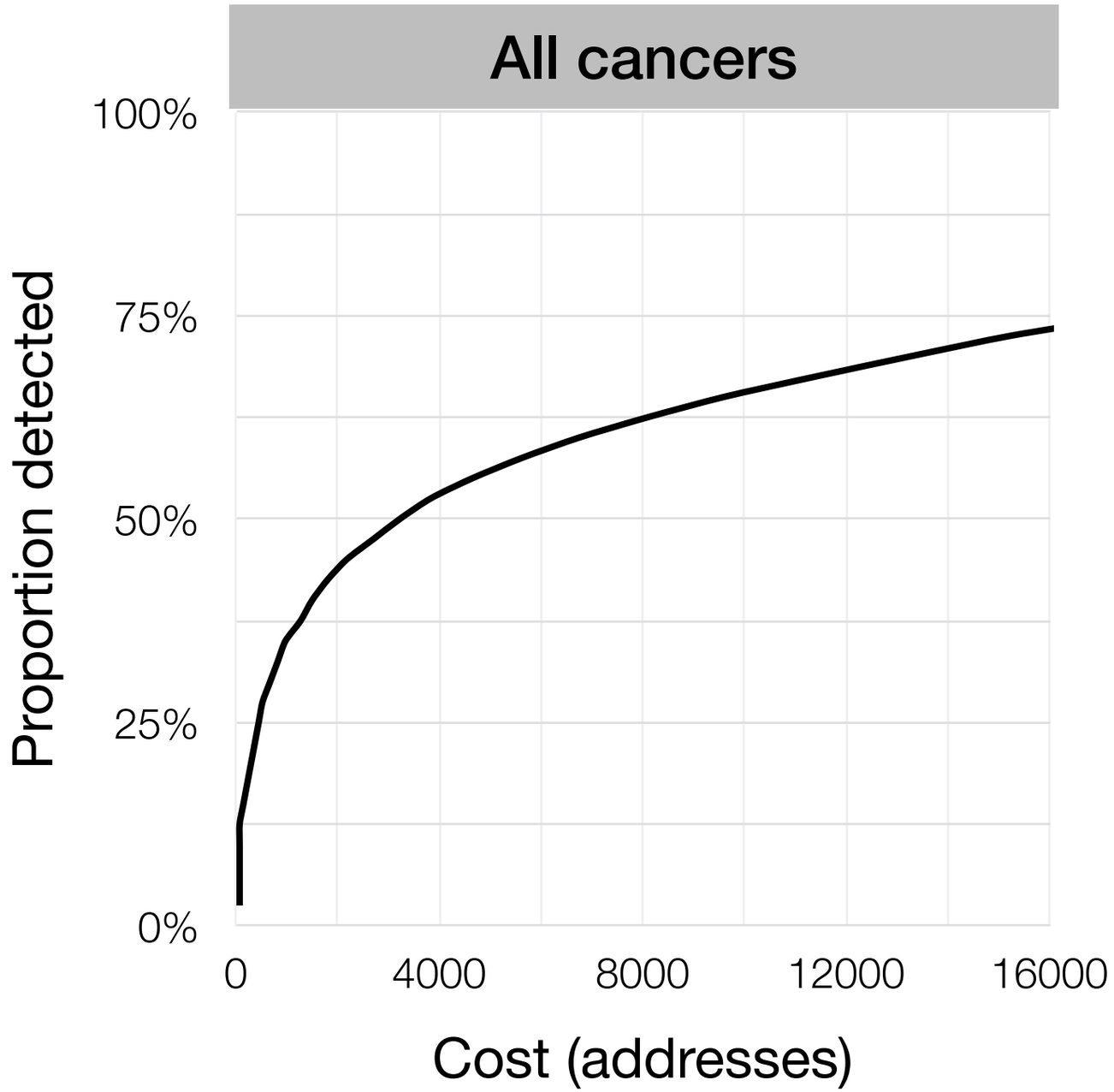


Challenge: Cost

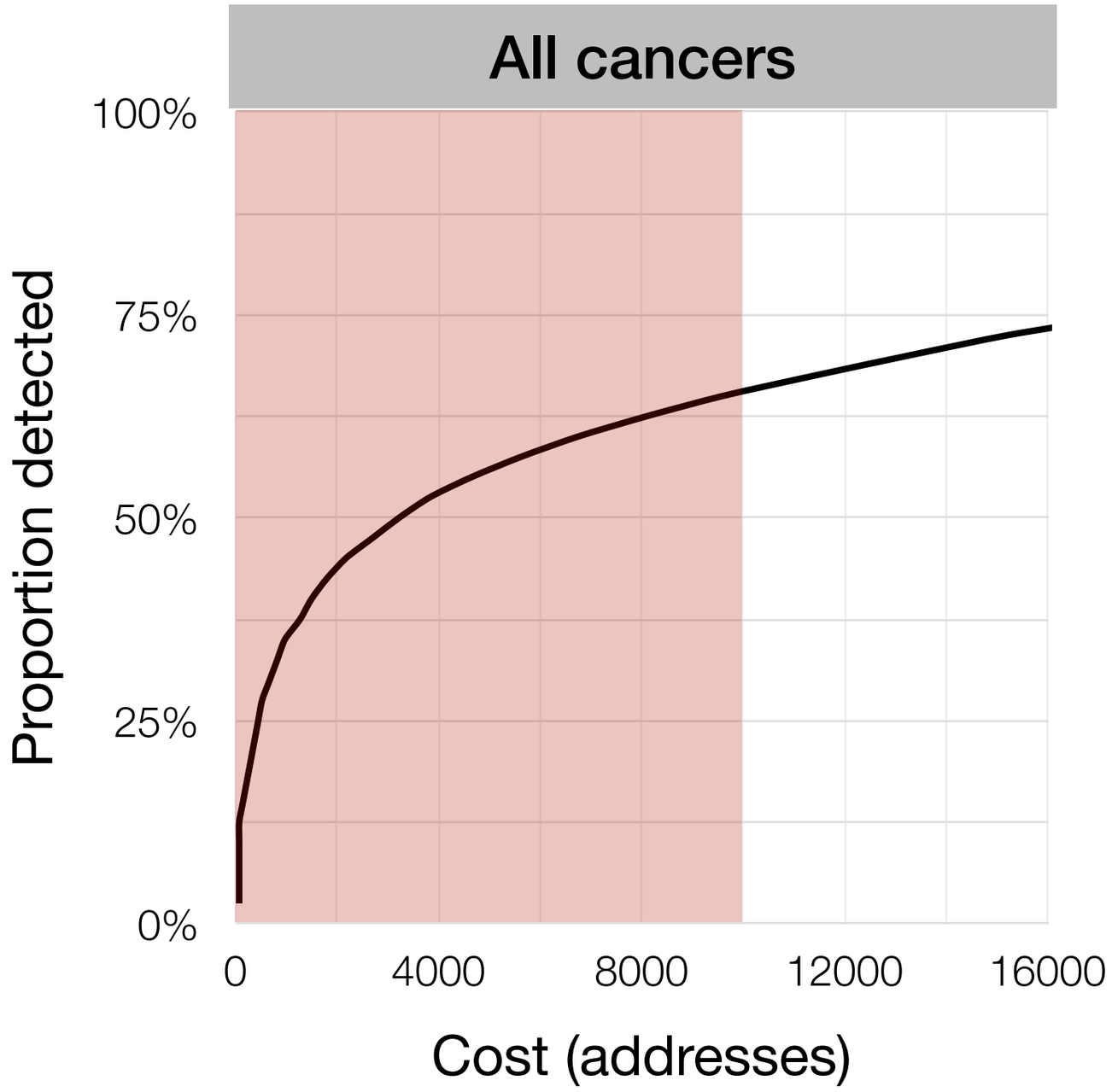
- A back-of-the-envelope calculation:
  - Test must cost at most $\$10^2$
  - Sequencing an address ten-thousand times costs $\$10^{-2}$
  - Thus, can only use a **panel** of ~10^4 addresses



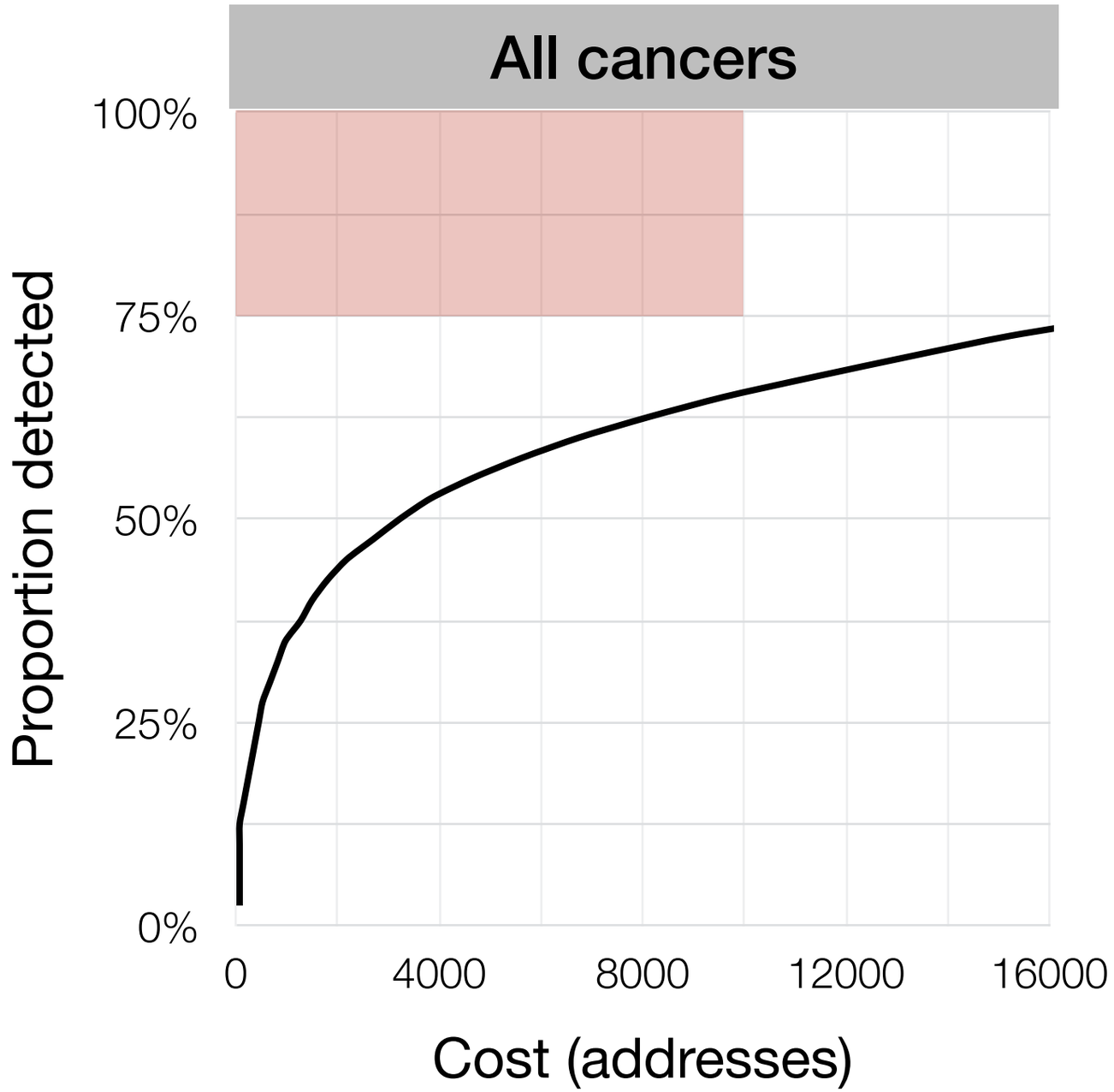
DNA Panels from [Cohen et al.]



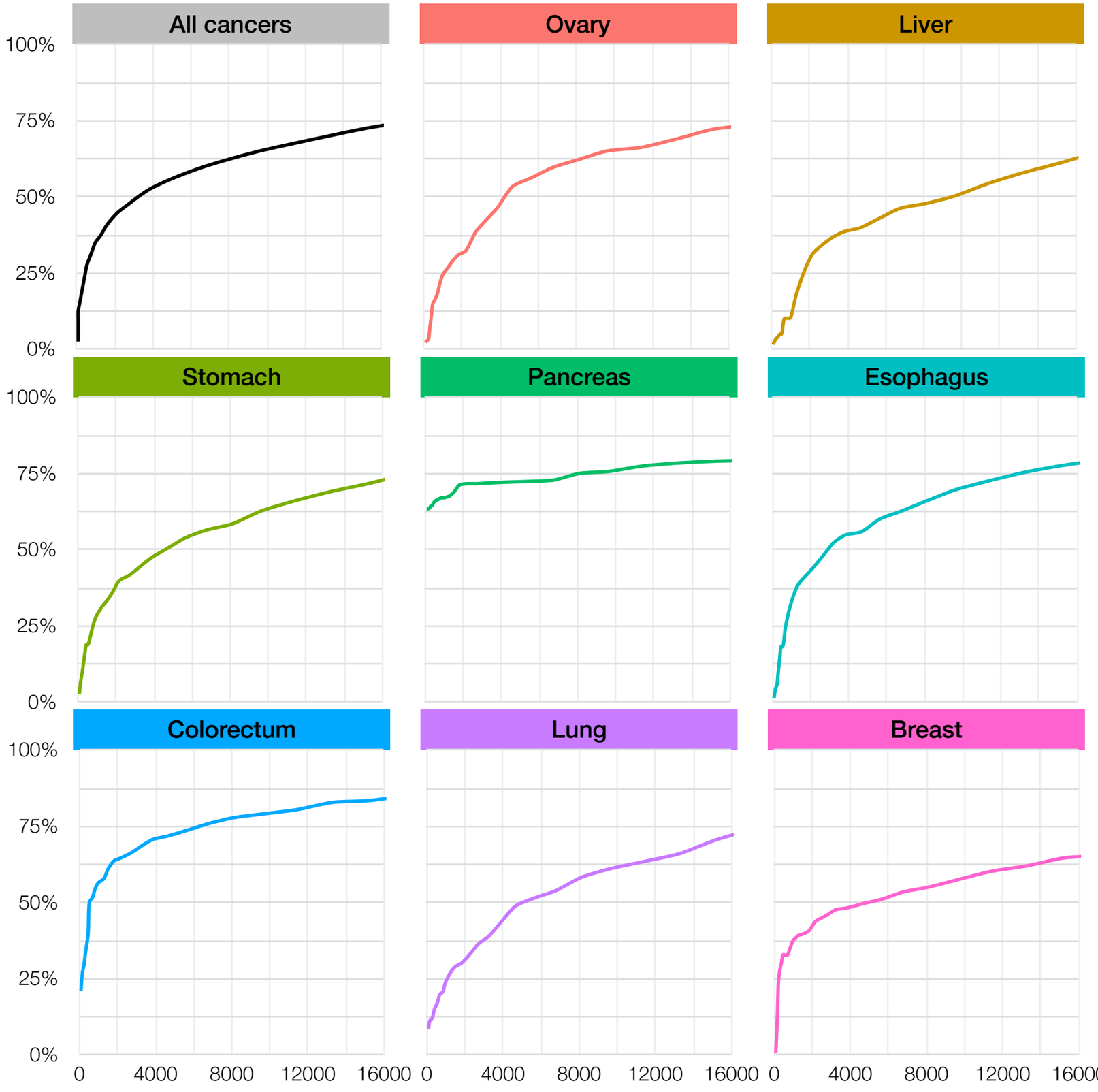
DNA Panels from [Cohen et al.]



DNA Panels from [Cohen et al.]

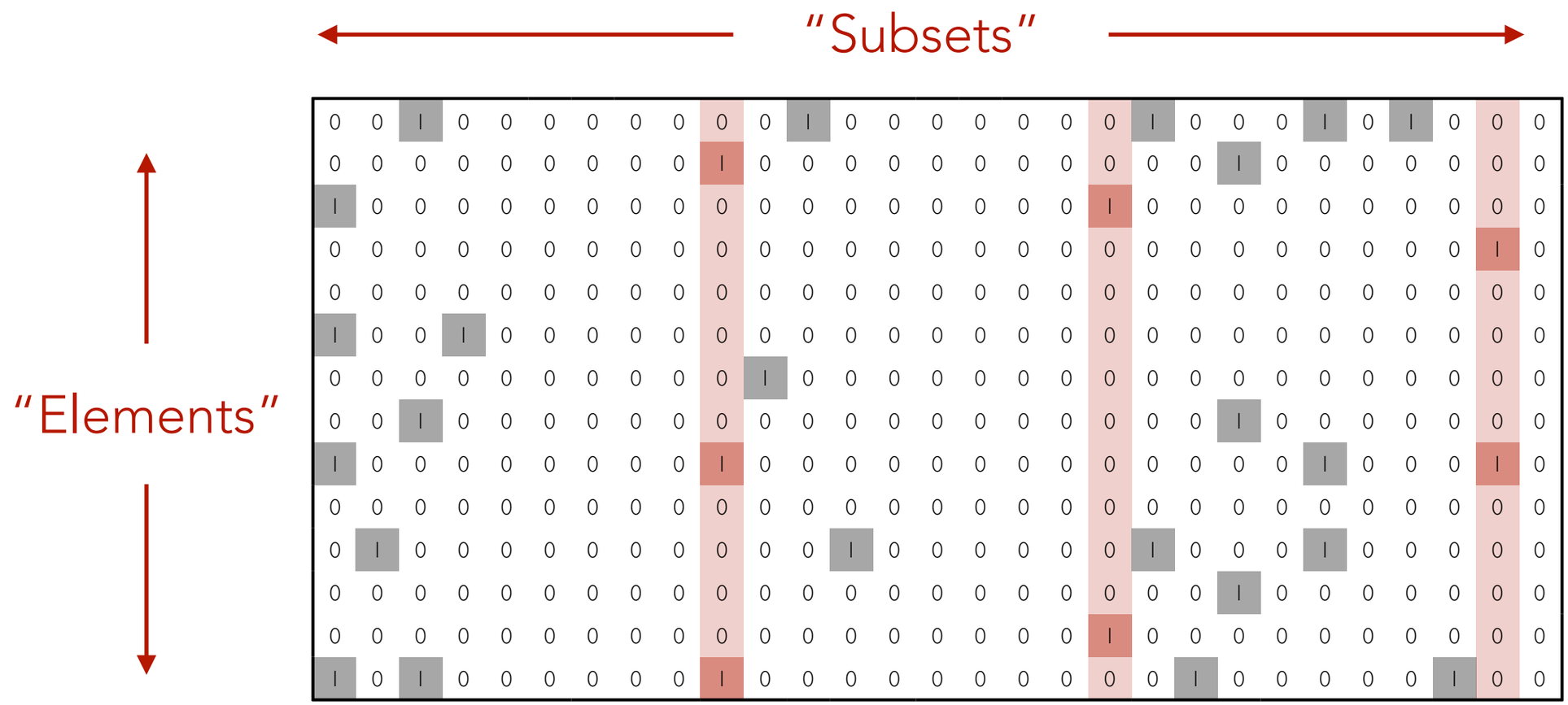


DNA Panels from [Cohen et al.]



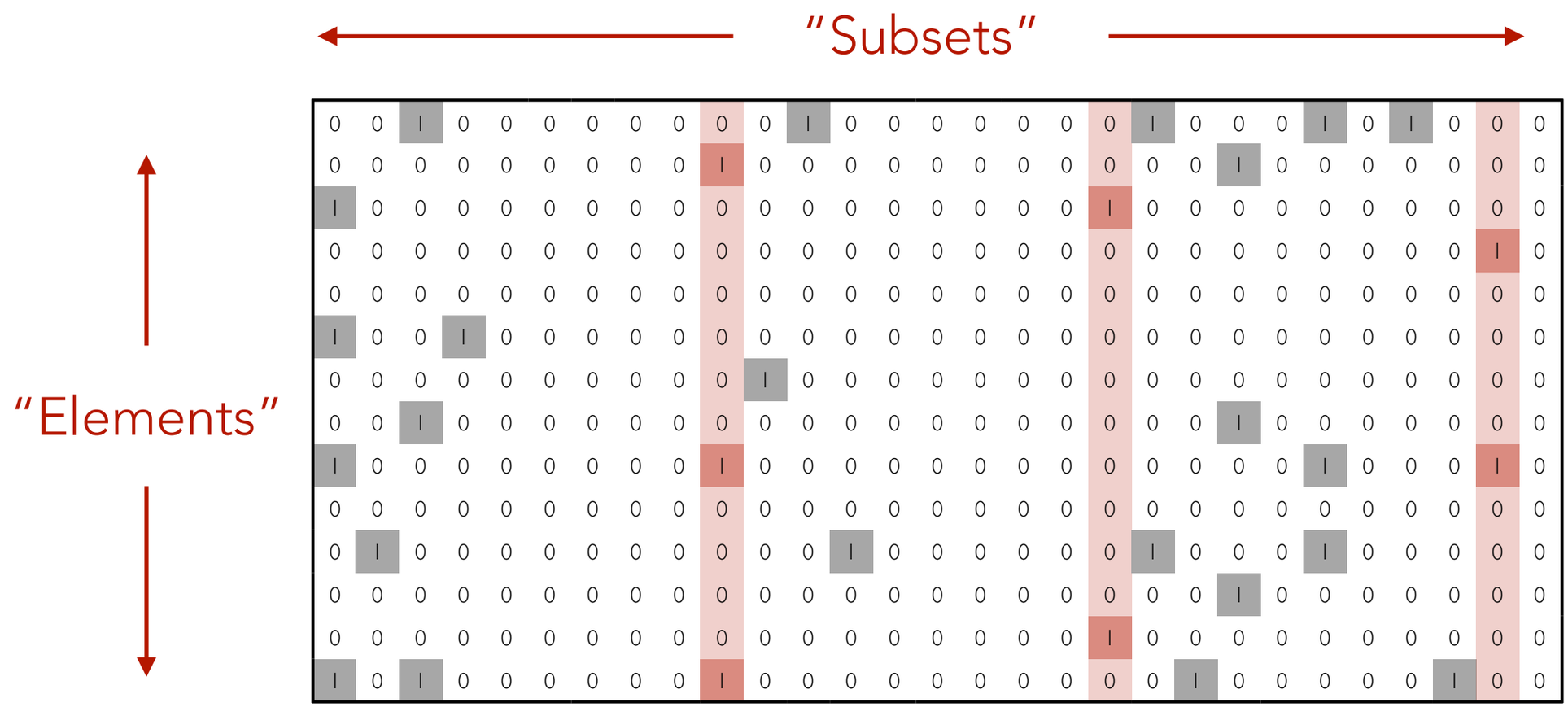
This is an optimization problem!

- *Max Cover*, to be specific



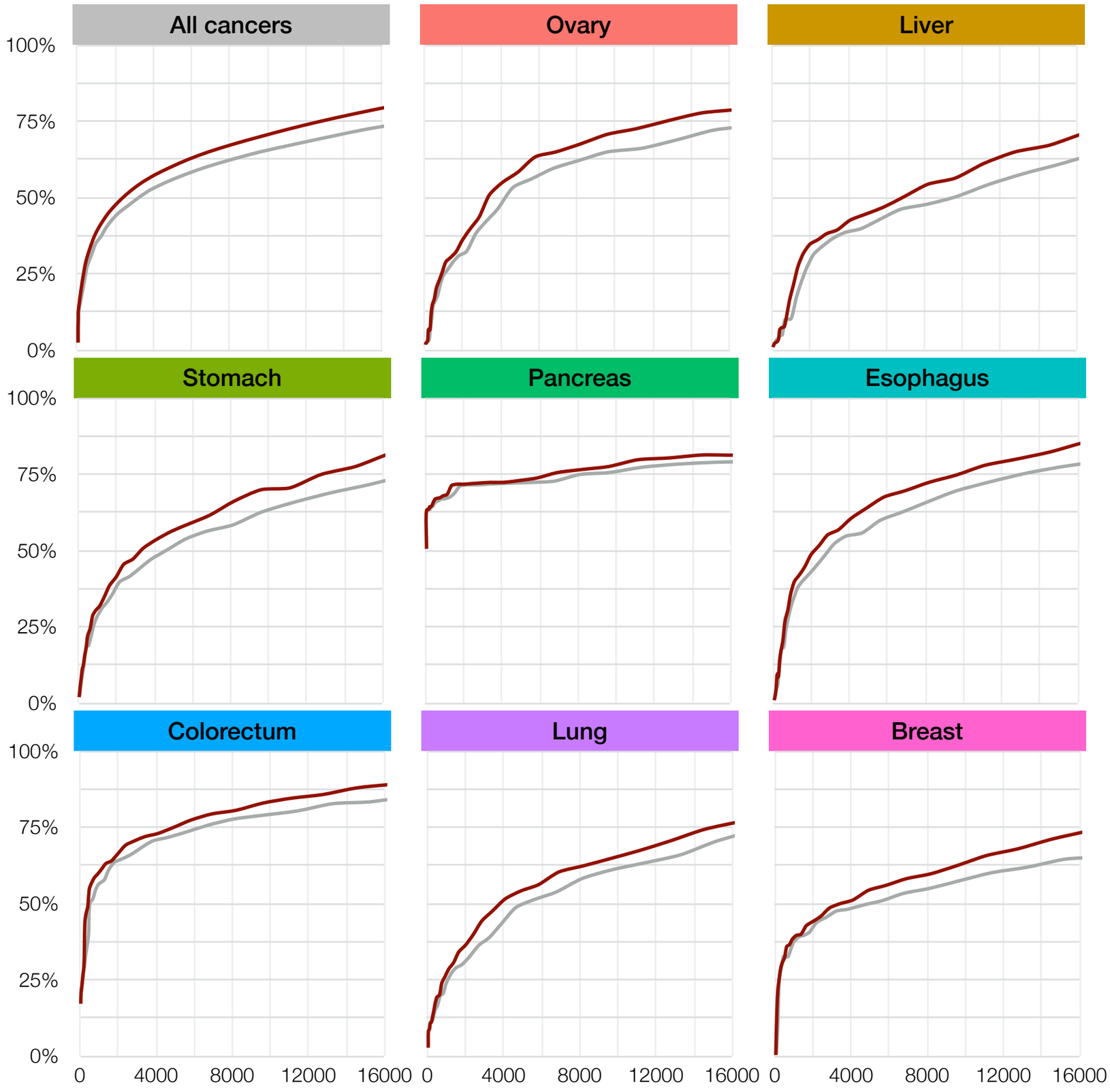
This is an optimization problem!

- *Max Cover*, to be specific

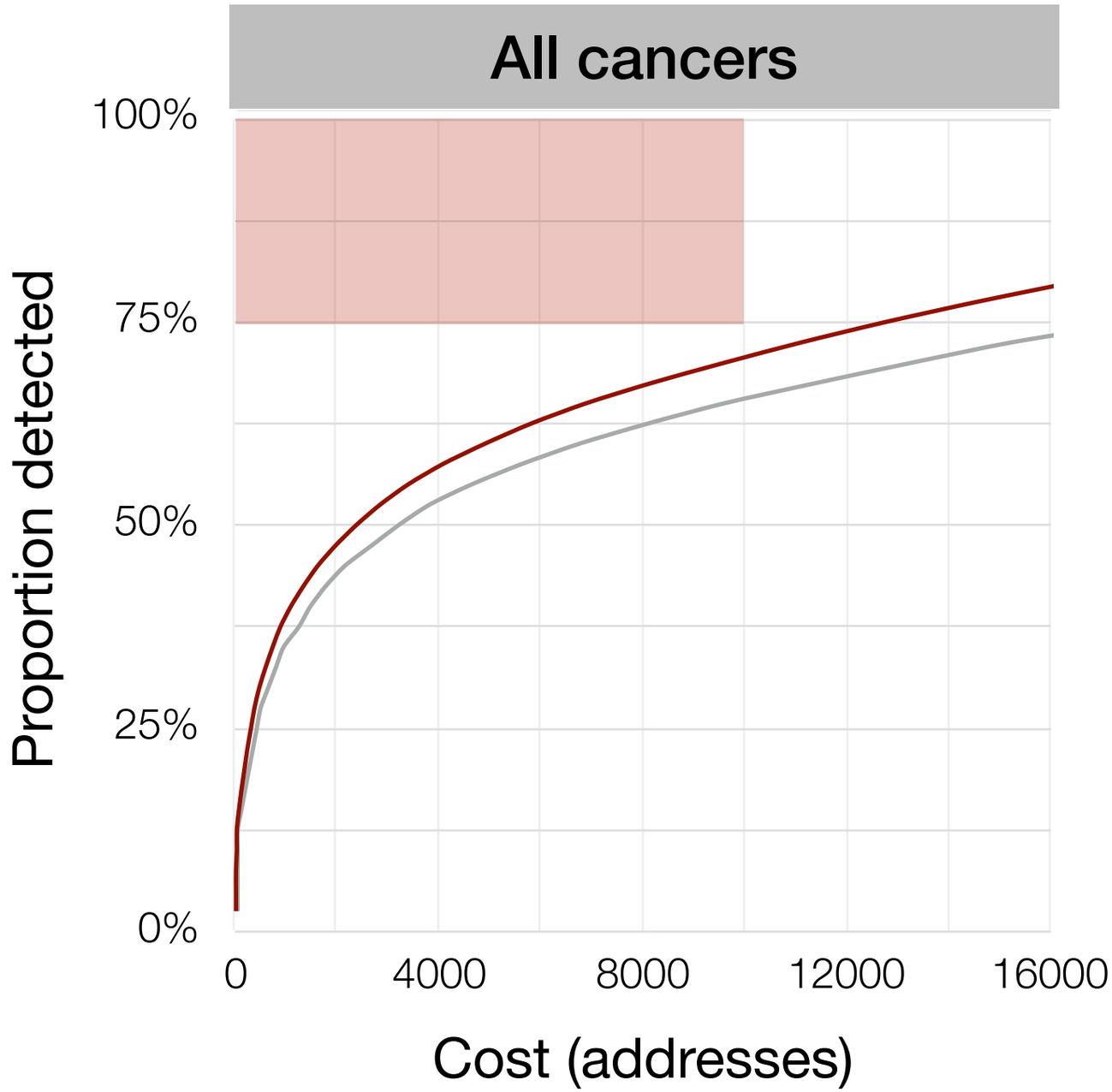


- Huge instance (10^4 elements, 10^9 subsets)
  - Still solves in Gurobi

Optimal Panels



Optimal Panels



Adaptive Panels

- **Adaptivity:** perform the test in stages
- Opportunity: the $100 constraint only needs to hold in expectation
  - Higher accuracy at same average cost

Adaptive Panels

- **Adaptivity:** perform the test in stages
- Opportunity: the $100 constraint only needs to hold in expectation
  - Higher accuracy at same average cost
- Challenge: this problem is hard (theoretically)

Adaptive Panels

- **Adaptivity:** perform the test in stages
- Opportunity: the $100 constraint only needs to hold in expectation
  - Higher accuracy at same average cost
- Challenge: this problem is hard (theoretically) and hard (practically)

A Simple Model for DNA Mutations

- Cancer types $t = 1, \ldots, T$

A Simple Model for DNA Mutations

- Cancer types $t = 1, \ldots, T$
- DNA addresses $a = 1, \ldots, A$

A Simple Model for DNA Mutations

- Cancer types $t = 1, \ldots, T$
- DNA addresses $a = 1, \ldots, A$

Addresses

Types

0.09	0.80	0.68	0.34	0.59	0.56	0.66	0.10	0.07	0.14	0.80	0.74	0.05	0.25	0.90	0.30	0.58	0.62	0.18	0.12	0.12	0.30	0.61	0.27	0.38	0.88	0.62
0.26	0.17	0.68	0.16	0.25	0.01	0.78	0.56	0.05	0.82	0.97	0.49	0.65	0.99	0.73	0.85	0.45	0.60	0.25	0.28	0.83	0.54	0.29	0.85	0.66	0.48	0.69
0.70	0.84	0.47	0.69	0.98	0.66	0.41	0.65	0.51	0.97	0.82	0.92	0.74	0.75	0.53	0.34	0.40	0.67	0.12	0.08	0.58	0.62	0.41	0.41	0.66	0.99	0.83
0.01	0.28	0.03	0.46	0.50	0.26	0.73	0.44	0.68	0.74	0.31	0.31	0.62	0.92	0.04	0.61	0.74	0.47	0.44	0.45	0.87	0.24	0.68	0.55	0.28	0.34	0.87
0.18	0.03	0.57	0.14	0.07	0.99	0.17	0.66	0.02	0.05	0.87	0.80	0.03	0.45	0.75	0.43	0.07	0.06	0.51	0.57	0.16	0.96	0.85	0.89	0.91	0.06	0.06
0.29	0.18	0.53	0.81	0.64	0.69	0.83	0.30	0.06	0.94	0.26	0.15	0.19	0.92	0.35	0.36	0.99	0.97	0.14	0.10	0.51	0.11	0.84	0.81	0.66	0.00	0.42
0.01	0.96	0.49	0.02	0.52	0.97	0.21	0.82	0.49	0.09	0.60	0.46	0.93	0.07	0.55	0.06	0.74	0.74	0.24	0.88	0.92	0.10	0.58	0.09	0.78	0.36	0.96
0.75	0.45	0.64	0.67	0.26	0.15	0.72	0.17	0.49	0.05	0.36	0.86	0.77	0.86	0.90	0.88	0.13	0.99	0.08	0.34	0.65	0.95	0.73	0.13	0.44	0.21	0.42
0.14	0.42	0.31	0.99	0.63	0.43	0.62	0.12	0.78	0.77	0.92	0.36	0.80	0.89	0.33	0.29	0.62	0.99	0.80	0.85	0.07	0.67	0.13	0.26	0.57	0.58	0.56
0.24	0.48	0.13	0.55	0.16	0.56	0.89	0.18	0.11	0.43	0.65	1.00	0.57	0.52	0.47	0.89	0.03	0.72	0.49	0.36	0.65	0.34	0.71	0.19	0.15	0.56	0.52

$$P \in \mathbb{R}^{T \times A}$$

A Simple Model for DNA Mutations

- Cancer types $t = 1, \ldots, T$
- DNA addresses $a = 1, \ldots, A$

Addresses

Types

0.09	0.80	0.68	0.34	0.59	0.56	0.66	0.10	0.07	0.14	0.80	0.74	0.05	0.25	0.90	0.30	0.58	0.62	0.18	0.12	0.12	0.30	0.61	0.27	0.38	0.88	0.62
0.26	0.17	0.68	0.16	0.25	0.01	0.78	0.56	0.05	0.82	0.97	0.49	0.65	0.99	0.73	0.85	0.45	0.60	0.25	0.28	0.83	0.54	0.29	0.85	0.66	0.48	0.69
0.70	0.84	0.47	0.69	0.98	0.66	0.41	0.65	0.51	0.97	0.82	0.92	0.74	0.75	0.53	0.34	0.40	0.67	0.12	0.08	0.58	0.62	0.41	0.41	0.66	0.99	0.83
0.01	0.28	0.03	0.46	0.50	0.26	0.73	0.44	0.68	0.74	0.31	0.31	0.62	0.92	0.04	0.61	0.74	0.47	0.44	0.45	0.87	0.24	0.68	0.55	0.28	0.34	0.87
0.18	0.03	0.57	0.14	0.07	0.99	0.17	0.66	0.02	0.05	0.87	0.80	0.03	0.45	0.75	0.43	0.07	0.06	0.51	0.57	0.16	0.96	0.85	0.89	0.91	0.06	0.06
0.29	0.18	0.53	0.81	0.64	0.69	0.83	0.30	0.06	0.94	0.26	0.15	0.19	0.92	0.35	0.36	0.99	0.97	0.14	0.10	0.51	0.11	0.84	0.81	0.66	0.00	0.42
0.01	0.96	0.49	0.02	0.52	0.97	0.21	0.82	0.49	0.09	0.60	0.46	0.93	0.07	0.55	0.06	0.74	0.74	0.24	0.88	0.92	0.10	0.58	0.09	0.78	0.36	0.96
0.75	0.45	0.64	0.67	0.26	0.15	0.72	0.17	0.49	0.05	0.36	0.86	0.77	0.86	0.90	0.88	0.13	0.99	0.08	0.34	0.65	0.95	0.73	0.13	0.44	0.21	0.42
0.14	0.42	0.31	0.99	0.63	0.43	0.62	0.12	0.78	0.77	0.92	0.36	0.80	0.89	0.33	0.29	0.62	0.99	0.80	0.85	0.07	0.67	0.13	0.26	0.57	0.58	0.56
0.24	0.48	0.13	0.55	0.16	0.56	0.89	0.18	0.11	0.43	0.65	1.00	0.57	0.52	0.47	0.89	0.03	0.72	0.49	0.36	0.65	0.34	0.71	0.19	0.15	0.56	0.52

$$P \in \mathbb{R}^{T \times A}$$

- Sequencing address a on individual of type t:
  - Yields observation $\sim \mathrm{Ber}(P_{ta})$

Problem: Active Sequential Hypothesis Testing

- Unknown cancer type drawn according to (known) prior

Problem: Active Sequential Hypothesis Testing

- Unknown cancer type drawn according to (known) prior
- **Partial Adaptivity:** select a sequence of addresses
  - Same sequence always used
  - Stop anytime, and "guess" the cancer type

Problem: Active Sequential Hypothesis Testing

- Unknown cancer type drawn according to (known) prior
- **Partial Adaptivity:** select a sequence of addresses
  - Same sequence always used
  - Stop anytime, and "guess" the cancer type
- Constraint: correctly identify type with probability at least $1 - \delta$
- Objective: minimize expected cost (number of addresses used)

Theoretical Guarantee

- Recall:
  - A = number of addresses (3 billion)
  - T = number of cancer types (10-100)

Theoretical Guarantee

- Recall:
  - A = number of addresses (3 billion)
  - T = number of cancer types (10-100)
- L = maximum length of test sequence (can think of this as ~10,000)

Theoretical Guarantee

- Recall:
  - A = number of addresses (3 billion)
  - T = number of cancer types (10-100)
- L = maximum length of test sequence (can think of this as ~10,000)

	Runtime	Cost Guarantee
Brute Force	A^L	OPT

Theoretical Guarantee

- Recall:
  - A = number of addresses (3 billion)
  - T = number of cancer types (10-100)
- L = maximum length of test sequence (can think of this as ~10,000)

	Runtime	Cost Guarantee
Brute Force	A^L	OPT
LP Heuristic [Naghshvar,Javidi'13]	$O(AT^2)$	None

Theoretical Guarantee

- Recall:
  - A = number of addresses (3 billion)
  - T = number of cancer types (10-100)
- L = maximum length of test sequence (can think of this as ~10,000)

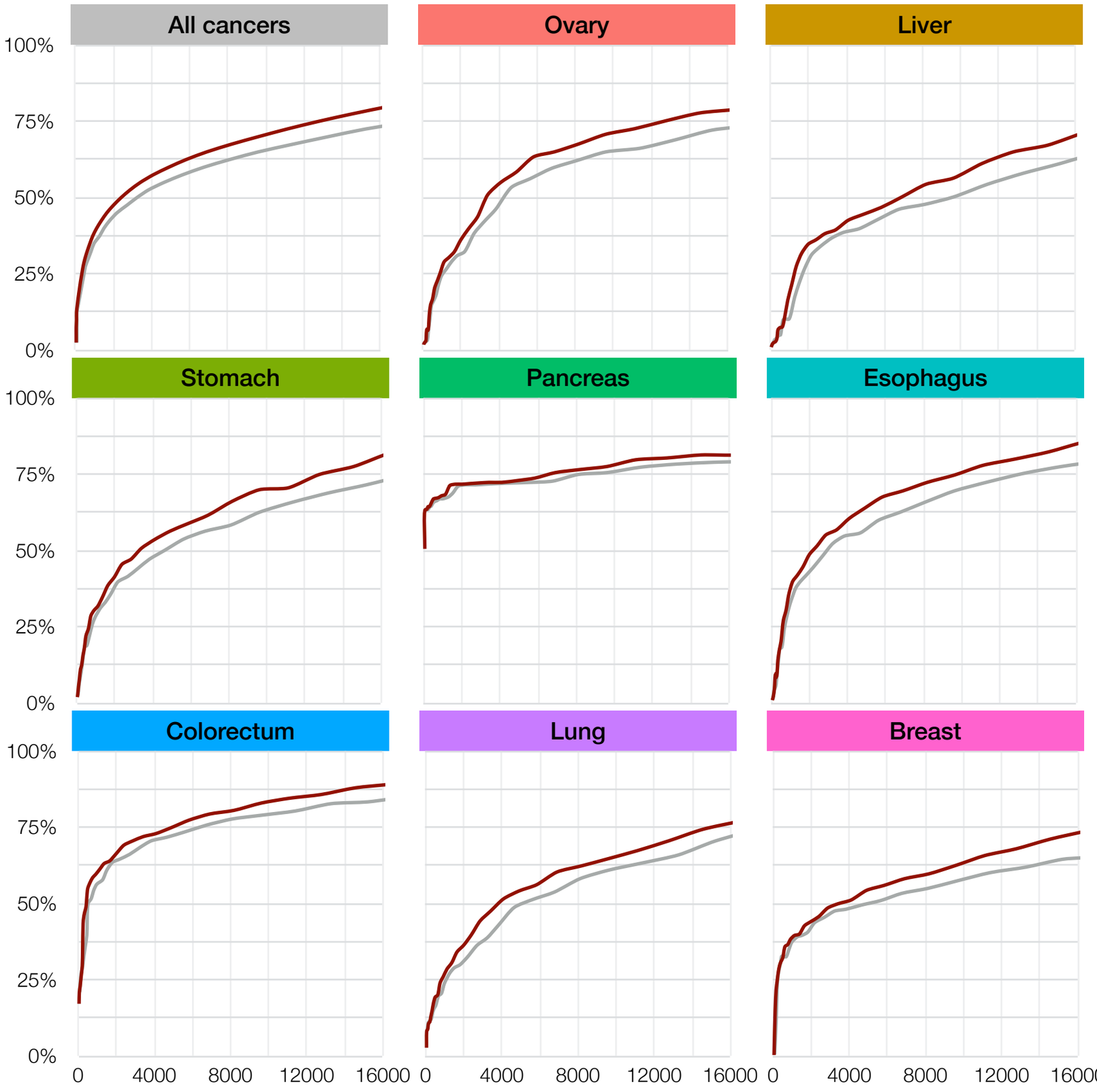
	Runtime	Cost Guarantee
Brute Force	A^L	OPT
LP Heuristic [Naghshvar,Javidi'13]	$O(AT^2)$	None
Our Algorithm	$O(ATL)$	$O(\log T)\mathrm{OPT}$

Theoretical Guarantee

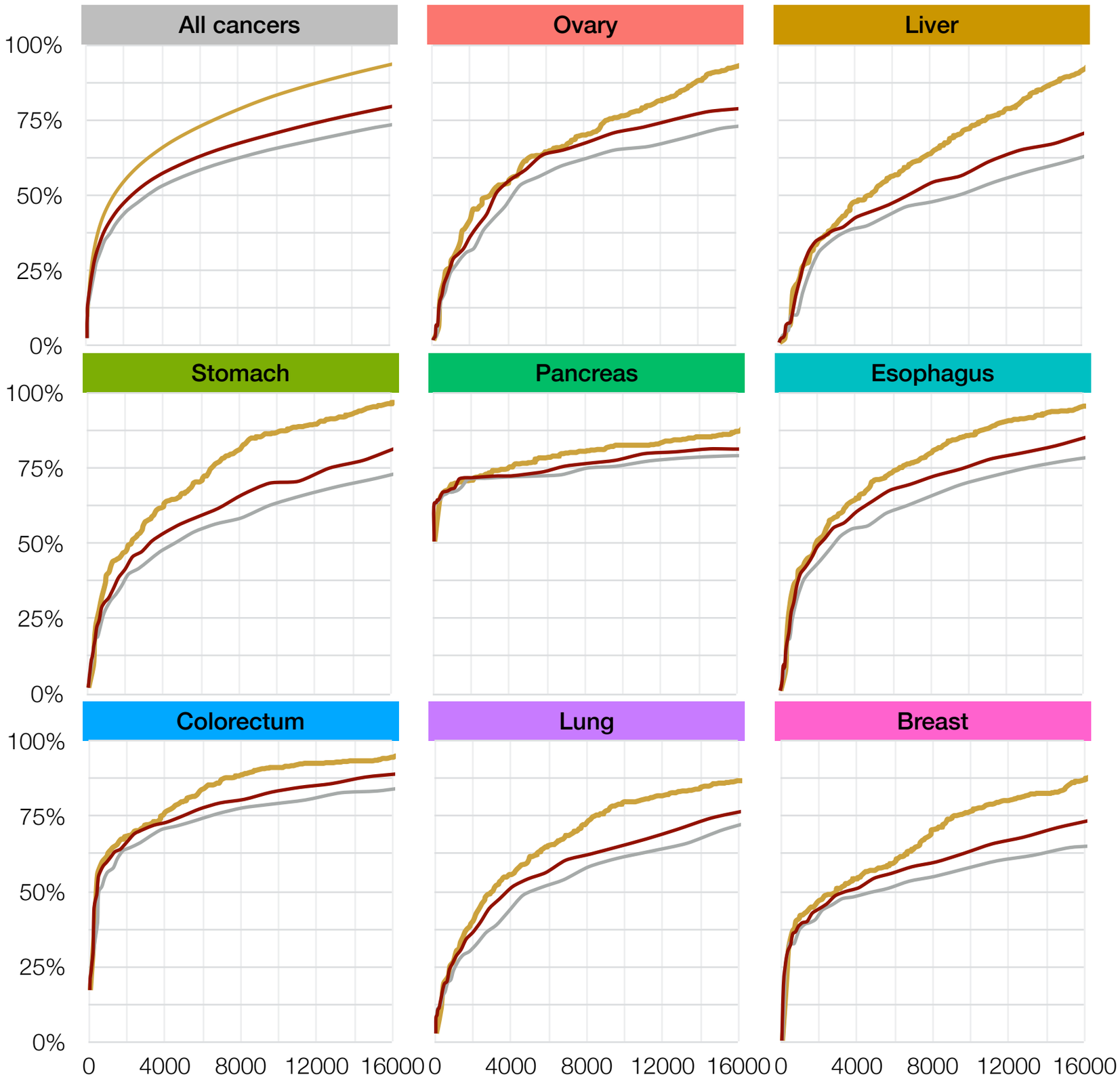
- Recall:
  - A = number of addresses (3 billion)
  - T = number of cancer types (10-100)
- L = maximum length of test sequence (can think of this as ~10,000)

	Runtime	Cost Guarantee
Brute Force	A^L	OPT
LP Heuristic [Naghshvar,Javidi'13]	$O(AT^2)$	None
Our Algorithm	$O(ATL)$	$O(\log T + \log\log \delta^{-1})\mathrm{OPT}$

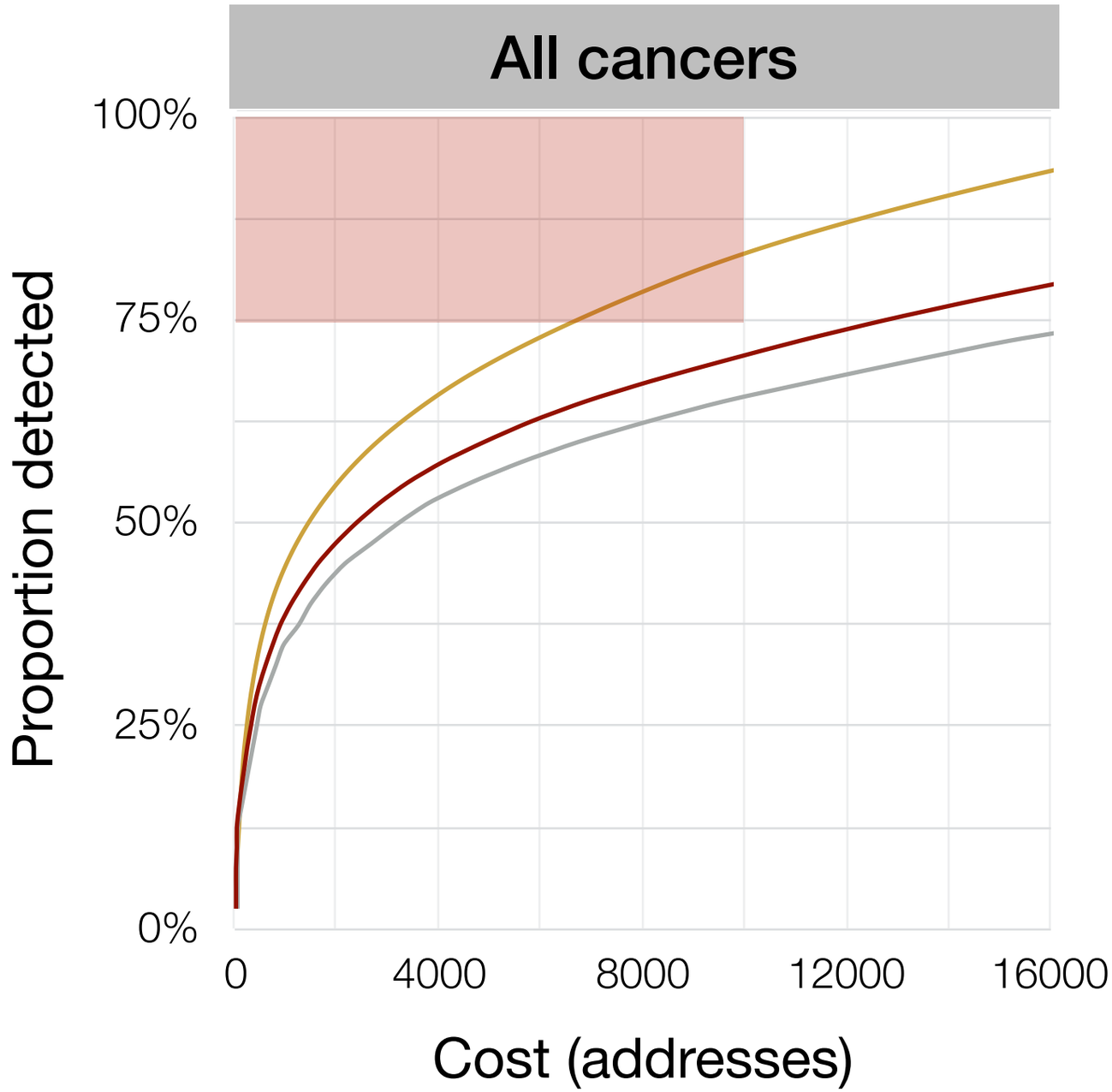
Optimal (Non-adaptive) Panels



(Sub-optimal) Adaptive Panels



(Sub-optimal) Adaptive Panels



Thanks!