ROBUSTNESS OF FITTED MUTATIONAL SIGNATURE EXPOSURES IN SINGLE-CELL DATA

Deciphering Cancer Heterogeneity with Machine Learning

AUTHOR Rebecca Nys
RESPONSIBLE PROFESSOR Dr. Joana Gonçalves
SUPERVISORS Sara Costa, Ivan Stresec

TU Delft

INTRODUCTION

- **Intratumor heterogeneity:** different cancer cells within one tumor carry distinct mutations [1]
- **Clinical impact:** therapies that target specific mutations may fail on certain cells [2]
- **Mutational signatures:** characteristic mutation patterns left by certain processes (e.g. UV light, defective DNA repair) [3]
- **Quantifying exposure:** de novo extraction (NMF) [4] or fitting (known signatures) → traditionally on bulk-sequencing data
- **Single-cell view:** scRNA-seq → mutations per cell, finer insight but low coverage & dropouts
- **Objective:** test how much missing data destabilises fitted mutational signature exposures at the single-cell level

METHODOLOGY

- **Dataset:** 688 scRNA-seq VCFs (one breast cancer tumor)
  - Variant calling → Liu et al. pipeline [5]; quality filtering → GATK best practices
  - Each VCF lists chromosome, position, ref / alt base
- **Signature fitting:** SigProfilerAssignment (COSMIC v3.4 SBS96, GRCh38)
  - Signatures (matrix P) fixed; exposures (matrix E) estimated per cell
  - Exposures normalised so each cell's values sum to 1
- **Simulating data loss**
  - First fit signatures to original data as baseline
  - Randomly delete 5% of mutations in every cell & refit signature exposures
  - Repeat 20 perturb-refit cycles with different random seeds
  - Repeat also for 10%, 20% and 40%

Deletion level:	5%	10%	20%	40%
SBS1	100%	100%	100%	100%
SBS5	100%	100%	100%	100%
SBS12	100%	100%	100%	100%
SBS26	100%	100%	100%	100%
SBS40c	100%	100%	100%	100%
SBS54	100%	100%	100%	100%
SBS87	35%	45%	100%	100%
SBS93	5%	10%	50%	95%
SBS37	15%	20%	50%	95%
SBS17a			30%	100%
SBS51			10%	40%
SBS21			5%	55%
SBS57			15%	90%
SBS19			5%	70%
SBS31				10%
SBS7d				50%
SBS23				15%
SBS33				15%
SBS32				15%
SBS88				20%
SBS7a				5%
SBS11				5%
SBS92				5%
SBS7b				5%

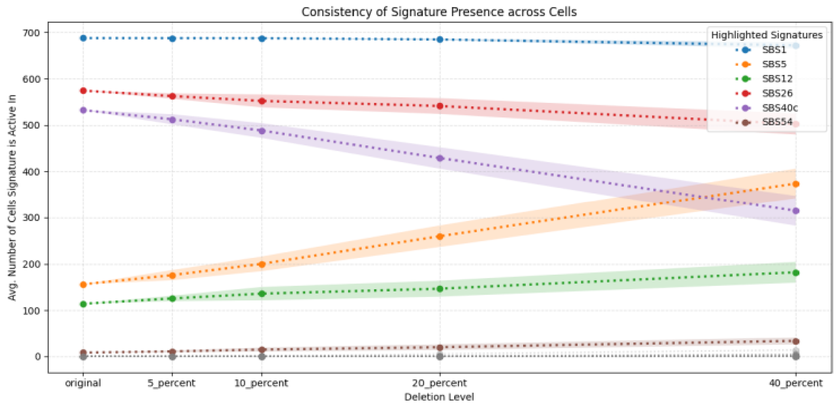
RESULTS

Signature presence in the dataset

- Fraction of perturbation runs in which each signature is detected (>0 exposure in ≥1 cell)
- Strong biological signal is recoverable
- Overfitting: extra signatures likely model noise
- Similar signatures might be confused

Consistency of signature presence across cells

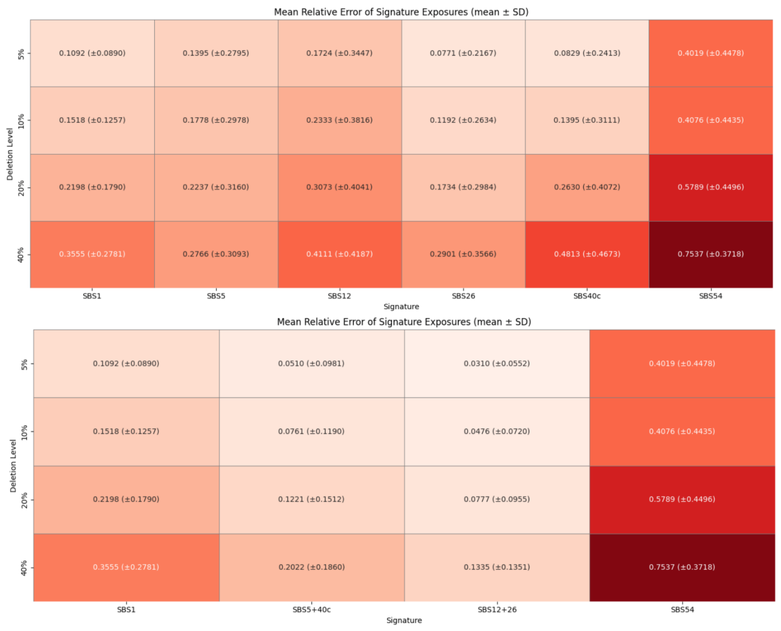
- For each signature, count active cells (exposure > 0) → average over 20 runs
- Distinct signatures seem more stable
- Possible swapping of similar signatures
- Cells with low mutation counts + flatter signatures could be more fragile



Per-signature MRE relative to original exposures

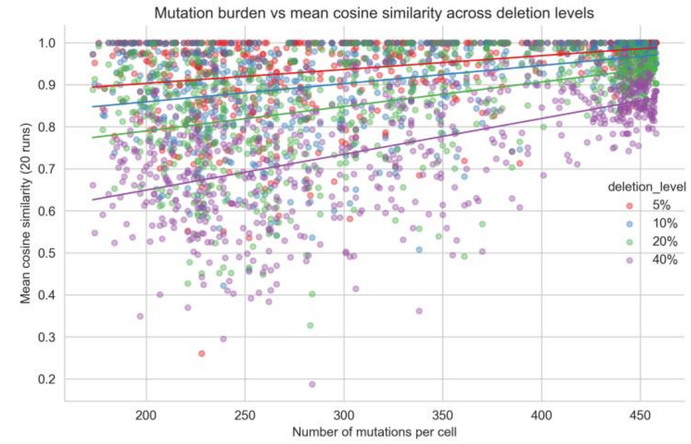
- For this signature, how much do the exposures deviate from the original on average across all cells?
- Lower data loss → varies more from cell to cell; higher data loss → more consistent shift across all cells
- Merging similar signatures cuts MRE by ≈58% → indicates signature swapping

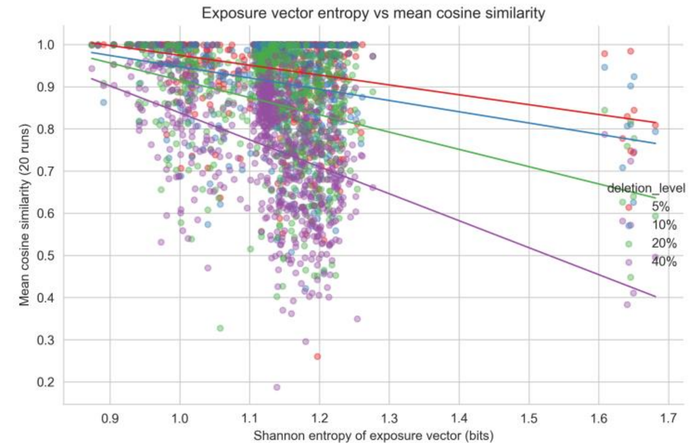
$$\mathrm{MRE}_{s,c} = \frac{1}{20}\sum_{i=1}^{20}\frac{|x_i - y|}{y}$$



Per-cell cosine similarity between exposure vectors

- Per-cell cosine similarity between original exposure vector and 20 perturbed vectors
- Similarities drop as mutation loss rises, but some cells drift at 5% while others stay stable even at 40%
- More mutations → higher similarity (ρ ≈ 0.38–0.59)
- Low-entropy exposure vector → higher similarity (ρ ≈ −0.53 at 40%)





Limitations

- Single tumor, one fitting tool, uniform random dropout
- COSMIC v3.4 SBS96 library derived from bulk genomes

Future directions

- Expand to other tumor types & mutational burdens
- Track reconstruction error & run-to-run exposure consistency
- Biased dropout (chromosome-specific) & simulate noise
- Biological/clinical validation

REFERENCES

[1] N. McGranahan and C. Swanton, "Clonal heterogeneity and tumor evolution: Past, present, and the future," Cell, vol. 168, pp. 613–628, Feb. 2017.
[2] M. Greaves, "Evolutionary determinants of cancer," Cancer discovery, vol. 5, no. 8, pp. 806–820, 2015.
[3] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering signatures of mutational processes operative in human cancer," Cell Reports, vol. 3, pp. 246–259, Jan. 2013.
[4] J. G. Tate et al., "Cosmic: the catalogue of somatic mutations in cancer," Nucleic Acids Research, vol. 47, pp. D941–D947, Jan. 2019.
[5] X. Liu, J. I. Griffiths, I. Bishara, J. Liu, A. H. Bild, and J. T. Chang, "Phylogenetic inference from single-cell RNA-seq data," Sci. Rep., vol. 13, no. 1, p. 12854, Aug. 2023.