Mutation signature analysis with genome location awareness

Feiyang Huang
Tri-Institutional Training Program in Computational Biology and Medicine
Weill Cornell Medicine
New York, NY
feh4005@med.cornell.edu

Abstract

Mutation signature analysis have emerged as standard practice for inferring latent biological processes from genomic alterations in cancer. Classically, mutation signatures are derived through non-negative matrix factorization of mutation count matrices generated from whole genomes. Recently, researchers have leveraged tensor factorization to extract mutation signatures that incorporate transcription state and epigenetic information, but mutation patterns remain unexplored in the context of chromosomal location. Here, I present a **location-aware mutation signature extraction method using tensor factorization**. I show that tensor mutation signatures **recapitulate known signatures while offering more nuanced explanations of the mutation profiles** of non-small cell lung cancer.

1 Introduction

Mutation in cancer arise due to a variety of biological and environmental processes, leading to distinct patterns that enables cancer classification, cell-type of origin identification, fitness prediction etc. Mutation signature analysis aims to 1) derive latent factors that contributes to a tumour's mutation profile through an admixture modeling approach, and 2) provide biological explanations for the latent factors.

Mutations are commonly classified into single nucleotide variations (SNVs), doublets (DNVs), insertions and deletions (indels), and larger-scale structural variants (SVs). SNVs are represented as tri-nucleotide contexts in the form of [5' BASE][REFERENCE][VARIANT][3' BASE]. C>A and G>T mutations are collapsed into a single class, resulting in 96 unique tri-nucleotide contexts. For *de novo* extraction of single nucleotide variation (SNV) mutation signatures (Section 4.2), mutation counts are aggregated per sample to form a $n \times 96$ count matrix and non-negative matrix factorization (NMF) is performed [1]. Growing bodies of high-quality whole genome sequences (WGS) of cancer have enabled iterative expansion of mutation signatures with biological associations [2, 3]. Mutation signatures have been successfully applied to the cancer diagnostics, inferring tumour-of-origin, and drug sensitivity prediction [4, 5].

While mutation signature analysis have demonstrated usefulness, it aggregates mutation at a coarse (per sample) level and under-utilizes sequencing information such as the position, strandedness, and chromatin state. Previous works have linked mutation signatures to biological processes in a location-specific manner in cancer. Aggregated regional mutation density is associated with the chromosome accessibility of the cell-type-of-origin of the cancer [6]. Additionally, the dominant mutation signature changes along the genome in a cancer-specific manner that may be indicative of evolutionary processes [7]. Therefore, representing mutations with location awareness provides an opportunity to model relevant biological processes in cancer formation and progression.

Coarse aggregation of mutations also lead to unrealistic assumptions when mutation signatures are used in practice. In clinical and biological discovery settings, signatures derived from whole genome sequencing are used to factor mutation profiles from whole exome sequencing (WES) and even targeted panel sequencing, which sub-samples the genome. This usage implicitly assumes that the regions covered by WES or panel sequencing preserves the genome-wide mutation profile. This assumption is incorrect because sequencing panels explicitly select for highly mutated driver genes, while the exome is subject to evolutionary pressures.

Inspired by the recent development of a tensor factorization framework for mutation signature extraction [8], I present a tensor factorization model for location-aware mutation signature (henceforth referred to as *tensor signature*). While Vohinger *et al* focus on the strandedness and chromatin state of mutations, I explore the localization of tensor signatures across chromosomal regions. To the best of my knowledge, this is the first attempt at location-aware tensor signature extraction.

2 Location-aware mutation signature modeling

The standard mutation signature inference is formulated as a non-negative matrix factorization (NMF) problem over a 2D matrix

$$\min_{\hat{M}} ||M - \hat{M}||, \quad \hat{M} = AB^T, \quad M \in \mathbb{R}^{n \times m}, A \in \mathbb{R}^{n \times r}, B^T \in \mathbb{R}^{r \times m} \tag{1}$$

where M is the mutation matrix of n rows for each cancer sample, $m = 96$ columns for each of the 96 tri-nucleotide sequences, and each M_{ij} contains the count of mutations for sample i and tri-nucleotide sequence j. NMF yields r mutation signatures, each of them a relative loading over the 96 tri-nucleotide sequences, represented by B. A is the relative attribution of each sample's observed mutations to the r mutation signatures.

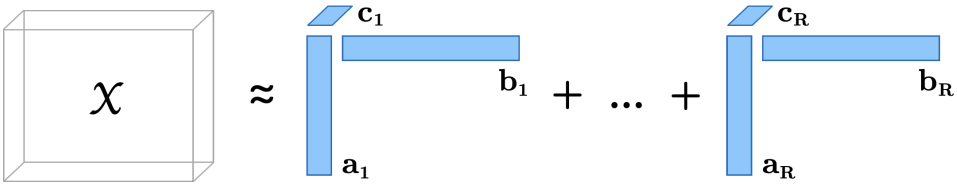

Figure 1: **CP Decomposition** [9].

To incorporate location awareness, we augment the 2D mutation matrix with a third dimension representing location, which is implemented as binned chromosomal positions. Mutation signature inference is now a tensor decomposition problem over a 3D tensor. We adopt the canonical polyadic (CP) decomposition formulation (Figure 1).

$$\min_{\hat{M}} ||M - \hat{M}||, \quad \hat{M} = \sum_{r=1}^{R} a_r \odot b_r \odot c_r, \quad A \in \mathbb{R}^{n \times r}, B \in \mathbb{R}^{m \times r}, C \in \mathbb{R}^{p \times r} \tag{2}$$

where A is the sample to signature mapping and B is the signature to tri-nucleotide sequence mapping, similar to the 2D case, and C is the signature to genome position mapping. In other words, the entry corresponding to sample i, tri-nucleotide sequence j and position k is

$$m_{ijk} \approx \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \tag{3}$$

The CP implementation in `Tensorly` was used [10, 11]. I enforced non-negativity to enable biological interpretation of the factors, and adopt the hierarchical alternating least-squared optimization to enable convergence within a reasonable time [12].

3 Datasets

3.1 Genomic data

In this section, I introduce the datasets used for model pre-training and benchmarking experiments. Cancers have been shown to carry unique somatic mutations and individual mutation signatures have been well studied. For modeling simplicity, I only use single nucleotide variations (SNVs) to encode each tumour, and discard indels and other forms of chromosomal aberrations. Additional sequencing information such as coverage, and clinical information such as tissue type and primary/metastasis sites are available in each dataset.

MSK-IMPACT MSK-IMPACT [13] is a FDA-approved targeted panel which sequences somatic and germ-line alterations in over 500 genes. Since normal peripheral blood samples are collected from the same patients, somatic mutations can be separated from germline mutations. `MuTect` was used to call somatic mutations from 67207 solid tumour samples with a total of 2,521,939 SNVs. The mean tumour mutation burden is 37.

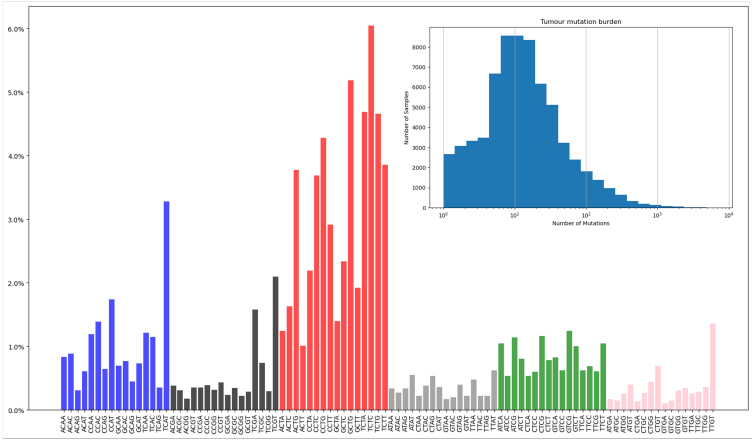

Figure 2: **SNVs in MSK-IMPACT cohort**

TRACERx 421 The TRACERx cohort consists primarily of patients diagnosed for non-small cell lung cancer (NSCLC) in the United Kingdom. Whole exome sequencing was performed on samples which passed quality checks [14] and 694 samples from 125 patients were open-sourced [15], from which 659,339 SNVs were identified (Figure 3). The mean tumour mutation burden is 950.

3.2 COSMIC Mutation Signature

COSMIC curates mutation signatures extracted through NMF, and their associated etiology. COSMIC signature v3.3 is used, and signatures corresponding to known sequencing artifacts are removed, namely: SBS43, SBS45, SBS46, SBS47, SBS48, SBS49, SBS50,SBS51, SBS52, SBS53, SBS54, SBS55, SBS56, SBS57, SBS58,SBS59, SBS60, SBS95.

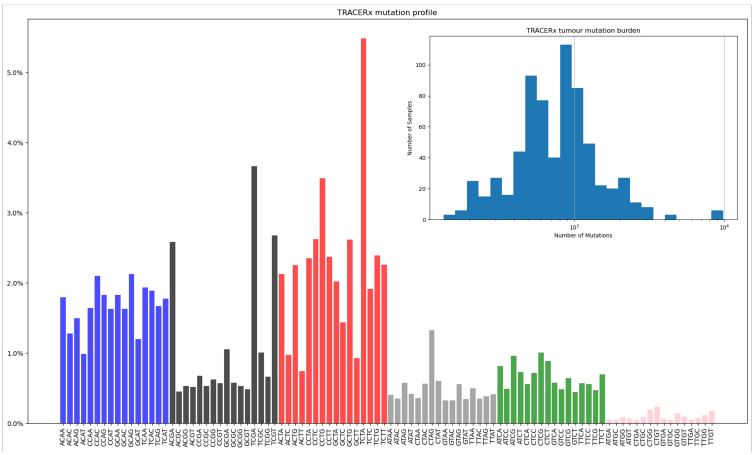

Figure 3: **SNVs in TRACERx 421 cohort**

4 Results

4.1 NSCLC mutation profile construction is sensitive to sequencing modality

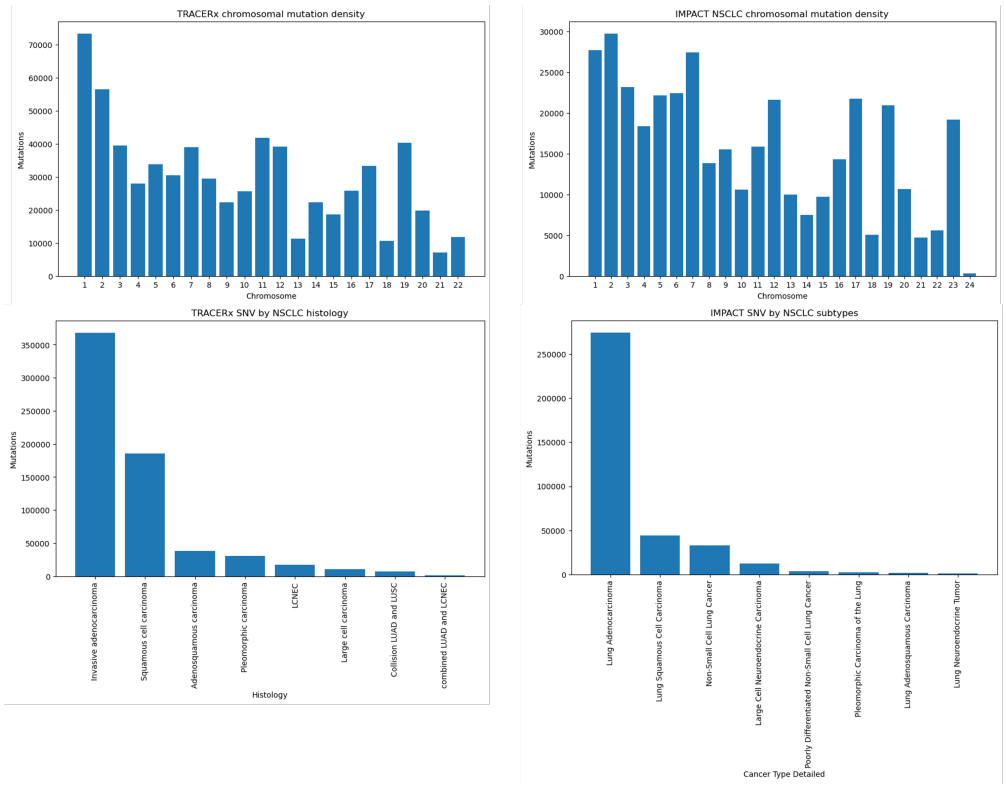

Figure 4: **Sequencing modality produces different chromosomal mutation profiles** Chromosomal mutation density and mutation profile of 96 tri-nucleotide context frequencies of NSCLC.

To understand the bias to mutation profiles introduced by sequencing technologies, I compare the mutation density of NSCLC produced by MSK-IMPACT, a targeted panel sequencing of 500 cancer driver genes, and TRACERx, a whole-exome sequencing dataset (Figure 4). Chromosome X and Y are excluded from the TRACERx panel but not MSK-IMPACT. The overall mutation density

per chromosome is lower in IMPACT, as expected from it's lower coverage. The relative mutation density of chromosome 1 is noticeable lower in IMPACT compared to TRACERx. A majority of mutations from both datasets come from the adenocarcinoma subtype. The number of mutations originating from squamous cell carcinoma is around 50% of adenocarcinoma in TRACERx, but significantly less in IMPACT.

Mutation signature inference was performed with `SigProfiler` using COSMIC 3.3 single base substitution SBS (Figure 5). The top-10 mutation signature by tumour mutation burden for TRACERx is SBS4, SBS25, SBS5, SBS39, SBS24, SBS89, SBS29, SBS95, SBS92, SBS30. The top-10 mutation signature by tumour mutation burden for MSK-IMPACT is SBS4, SBS3, SBS95, SBS94, SBS92, SBS39, SBS40, SBS25, SBS89, SBS37. Only 5 out of the top-10 mutation signatures are common for both IMPACT and TRACERx. The only common top-5 mutation signature for both MSK-IMPACT and TRACERx is SBS4, which corresponds to tobacco-smoking and aging.

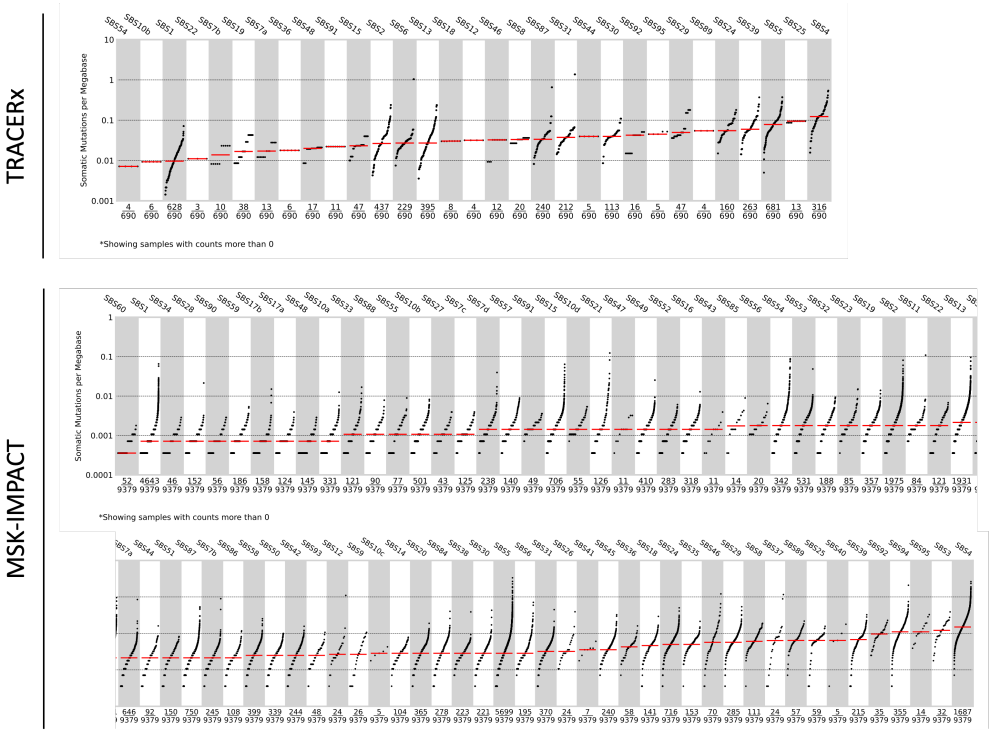

Figure 5: **SigProfiler tumour mutation burden analysis**. Mutation signatures in ascending order of tumour mutation burden. The fraction of samples containing each signature at the bottom.

4.2 Tensor signatures reveal positional variations across the genome

Tensor signatures are extracted from the TRACERx dataset. The TRACERx dataset contains 690 exomes. 2867 positions are formed by concatenating chromosomes and binning at 1Mb resolution. Together with 96 unique tri-nucleotide contexts, these form a count tensor of $(690 \times 96 \times 543)$. First, I empirically optimize the number of factors by performing CP decomposition with rank = $[2, 3, 5, 10]$ (Figure 6). Choosing 2 factors lead to optimal reconstruction loss for this dataset, while 10 factors are excessive since the TRACERx dataset contains only NSCLC samples and is relatively homogeneous. Furthermore, the first two factors persist as the number of factors are increased, while additional factors are strongly attributed to only a small number of samples (Figure 7).

Encouragingly, **the positional attributions of the 2 highlighted tensor signature show considerable variations across the genome**, which should be further investigated.

Figure 6: **Optimization of number of factors**: Reconstruction loss is calculated by flattening the reconstructed tensor along the position dimension into a 2D matrix, and calculating the amount of unexplained variance of the truth mutation matrix.

4.3 Tensor signatures recapitulate known mutation signatures

To sanity check that the tensor signatures could recover meaningful biology, we compare the tensor signatures to known COSMIC mutation signatures using the cosine similarity between the mutation dimension of extracted tensor signatures and known mutation signatures (Figure 8). Encouragingly, the **tensor signatures share a high degree of similarity with known lung cancer mutation signatures** such as SBS2, SBS4, SBS13, SBS29, and SBS24. These signatures have been associated with AID/APOBEC family cytidine deaminases activity, tobacco use, and aflatoxin exposure, which have known attributions to lung cancer.

Interestingly, as the number of factors are increase from 2 to 5, additional tensor factors are highly similar SBS2 and SBS13, which are linked to AID/APOBEC family cytidine deaminases activity (Figure 7, 8). These signatures are strongly attributed to a few sample and may indicate a sub-type of lung cancer within the TRACERx cohort.

5 Discussions

Here, I presented a tensor factorization framework for deriving position-aware mutation signatures. Using a whole-exome NSCLC dataset, I demonstrated that the derived tensor signatures have a high degree of similarity to plausible COSMIC signatures and show interesting variations in positional attributions. While previous works have demonstrated regional variations of mutation signature across the genome, this is the first attempt at *de novo* extraction of mutation signatures that take regional variations into account. It will be interesting to generate and test biological hypothesis underlying such regional variations in the future.

This project is limited by the homogeneity of the TRACERx dataset, which a single cancer type dataset primarily designed for lineage tracing of metastasis. While the MSK-IMPACT dataset contains diverse cancer types, the mutation count per sample (mean = 37) is too small for meaningful signature extraction. Therefore, the next logical step is to apply tensor factorization for tensor signature extraction from pan-cancer WGS/WES datasets, such as the PCAWG or TCGA dataset. Further work to improve the runtime of tensor factorization will enable factorization of larger datasets at higher positional resolution (smaller bin size).

Lastly, it is not clear how to best use components of the tensor signature for inferring sample attributions given an unseen dataset that has different sequencing coverage along the genome. Advances here will enable a more principled framework for using signatures derived from WGS for inference on WES and panel sequencing, which has high clinical relevance.

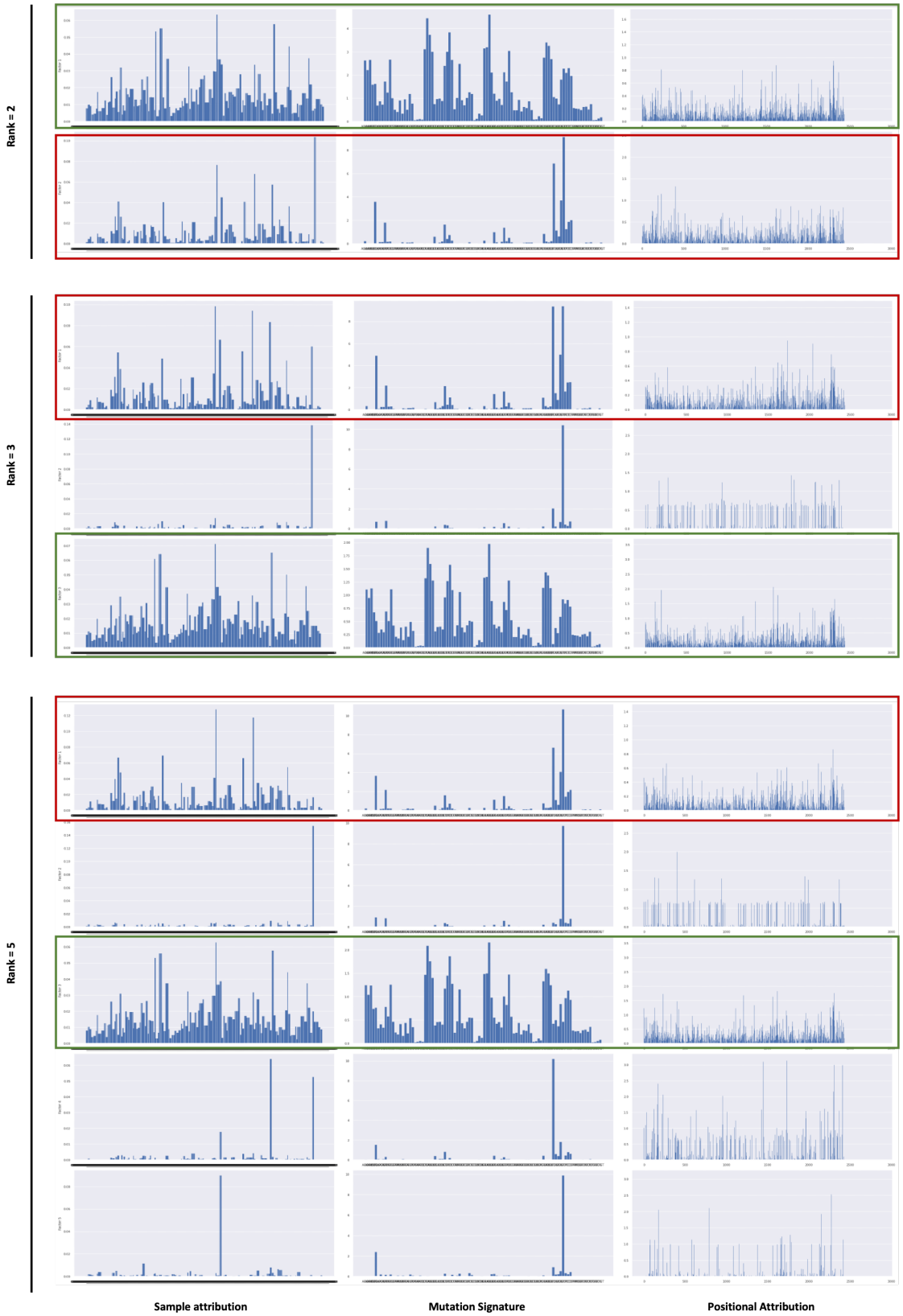

Figure 7: **Tensor mutation factors**: CP decomposition of TRACERx whole exome sequencing. Red and green factors persist while additional factors have sparse sample attributions.

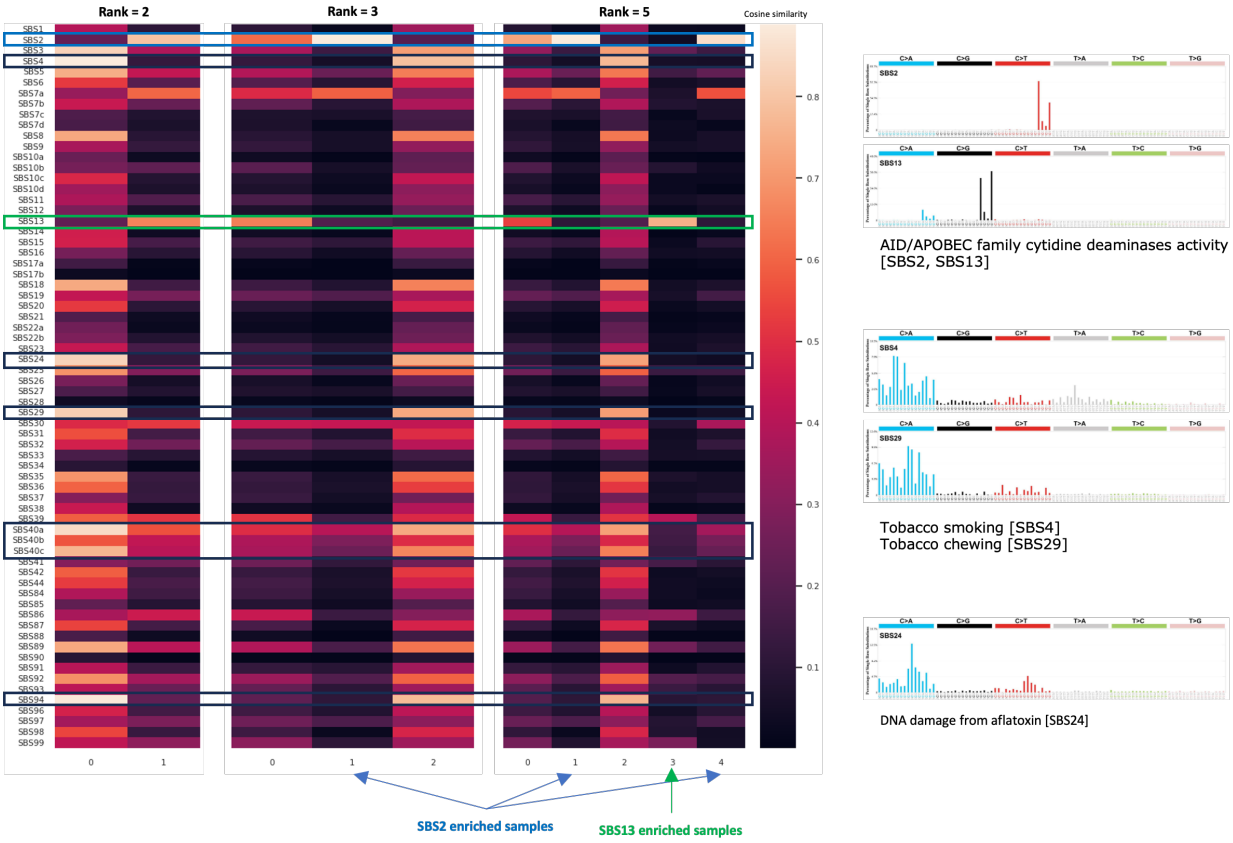

Figure 8: **Comparison with COSMIC signatures**: After CP decomposition, the mutation vector of the tensor signatures are compared to known COSMIC signatures using cosine similarity.

References

1. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. en. *Nature* **500.** Number: 7463 Publisher: Nature Publishing Group, 415–421. ISSN: 1476-4687. https://www.nature.com/articles/nature12477 (2023) (Aug. 2013).
2. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. en. *Nature* **578.** Number: 7793 Publisher: Nature Publishing Group, 94–101. ISSN: 1476-4687. https://www.nature.com/articles/s41586-020-1943-3 (2023) (Feb. 2020).
3. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47,** D941–D947. ISSN: 0305-1048. https://doi.org/10.1093/nar/gky1015 (2023) (Jan. 2019).
4. Levatić, J., Salvadores, M., Fuster-Tormo, F. & Supek, F. Mutational signatures are markers of drug sensitivity of cancer cells. en. *Nat Commun* **13.** Number: 1 Publisher: Nature Publishing Group, 2926. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-30582-3 (2023) (May 2022).
5. Van Hoeck, A., Tjoonk, N. H., van Boxtel, R. & Cuppen, E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* **19,** 457. ISSN: 1471-2407. https://doi.org/10.1186/s12885-019-5677-2 (2023) (May 2019).
6. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. eng. *Nature* **518,** 360–364. ISSN: 1476-4687 (Feb. 2015).
7. Timmons, C., Morris, Q. & Harrigan, C. F. Regional mutational signature activities in cancer genomes. en. *PLOS Computational Biology* **18.** Publisher: Public Library of Science, e1010733. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010733 (2023) (Dec. 2022).
8. Vöhringer, H., Hoeck, A. V., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. en. *Nat Commun* **12.** Number: 1 Publisher: Nature Publishing Group, 3628. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-23551-9 (2023) (June 2021).
9. Rabanser, S., Shchur, O. & Günnemann, S. *Introduction to Tensor Decompositions and their Applications in Machine Learning* arXiv:1711.10781 [cs, stat]. Nov. 2017. http://arxiv.org/abs/1711.10781 (2023).
10. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. *Journal of Machine Learning Research* **20,** 1–6. ISSN: 1533-7928. http://jmlr.org/papers/v20/18-277.html (2023) (2019).
11. Shashua, A. & Hazan, T. *Non-negative tensor factorization with applications to statistics and computer vision* en. in *Proceedings of the 22nd international conference on Machine learning - ICML '05* (ACM Press, Bonn, Germany, 2005), 792–799. ISBN: 978-1-59593-180-1. http://portal.acm.org/citation.cfm?doid=1102351.1102451 (2023).
12. Gillis, N. & Glineur, F. Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization. *Neural Computation* **24.** arXiv:1107.5194 [cs, math], 1085–1105. ISSN: 0899-7667, 1530-888X. http://arxiv.org/abs/1107.5194 (2023) (Apr. 2012).
13. Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of Molecular Diagnostics* **17,** 251–264. ISSN: 1525-1578. https://www.sciencedirect.com/science/article/pii/S1525157815000458 (2023) (May 2015).
14. Frankell, A. M. *et al.* The evolution of lung cancer and impact of subclonal selection in TRACERx. en. *Nature* **616.** Number: 7957 Publisher: Nature Publishing Group, 525–533. ISSN: 1476-4687. https://www.nature.com/articles/s41586-023-05783-5 (2023) (Apr. 2023).

15. Al Bakir, M. *et al.* The evolution of non-small cell lung cancer metastases in TRACERx. en. *Nature* **616.** Number: 7957 Publisher: Nature Publishing Group, 534–542. ISSN: 1476-4687. https://www.nature.com/articles/s41586-023-05729-x (2023) (Apr. 2023).

Acknowledgements I am grateful to Dr Wesley Tansey for his teaching and mentorship. I would like to thank Dr Quaid Morris for providing the computational resources used for this project. I also thank Leah Morales and Divya Koyyalagunta for data generation and curation help.