mutSignatures: An R Package For Extraction And Analysis Of Cancer Mutational Signatures

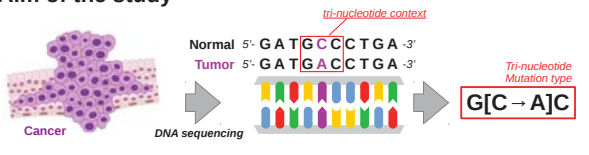
Damiano Fantini[1,2], Vania Vidimar[3], Yanni Yu[1], and Joshua J. Meeks[1,2]

[1.] Department of Urology, Northwestern University, Feinberg School of Medicine, [2.] Robert H. Lurie Comprehensive Cancer Center, and
[3.] Department of Microbiology-Immunology, Northwestern University, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

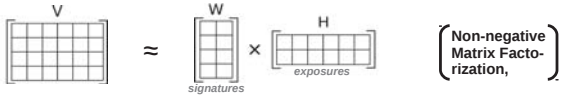
Background

- Genetic instability is one of the **hallmarks of cancer**. Neoplastic cells accumulate somatic mutations in their genomes, resulting in aberrant homeostasis, cancer cell survival, and proliferation
- Different genetic instability processes result in **distinct non-random patterns of DNA mutations**, also known as mutational signatures
- The interest in the identification of **mutational signatures** and the corresponding genetic instability processes is rapidly growing because these signatures are footprints of the molecular aberrations occurring in tumors [1, 2], and hence may be prognostic of clinical outcomes and support customized anti-cancer treatments in the future

Aim of the study

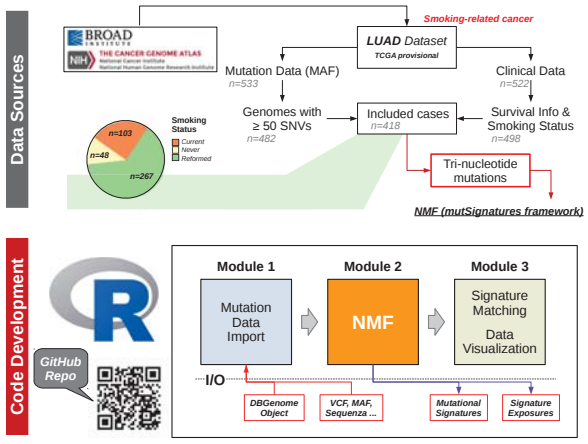


● What are the most common patterns of tri-nucleotide mutations occurring in human tumors? ● Are they prognostic of clinical outcomes?



We developed '*mutSignatures*', an integrated R-based computational framework aimed at deciphering DNA mutational signatures. Our software provides advanced functions for importing DNA variants, computing tri-nucleotide or non-standard mutation types, and extracting mutational signatures via non-negative matrix factorization (NMF) [3]. Additionally, our framework supports deconvolution of catalogs of DNA variants against known mutational signatures (*https://cran.r-project.org/web/packages/mutSignatures/index.html*).

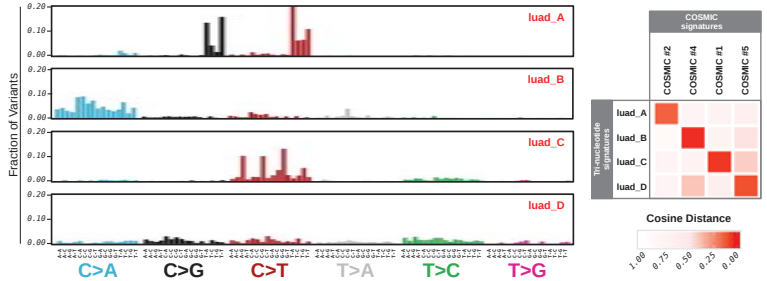
Data Sources and Implementation



Results

Identification of mutational signatures from lung adenocarcinoma genomes

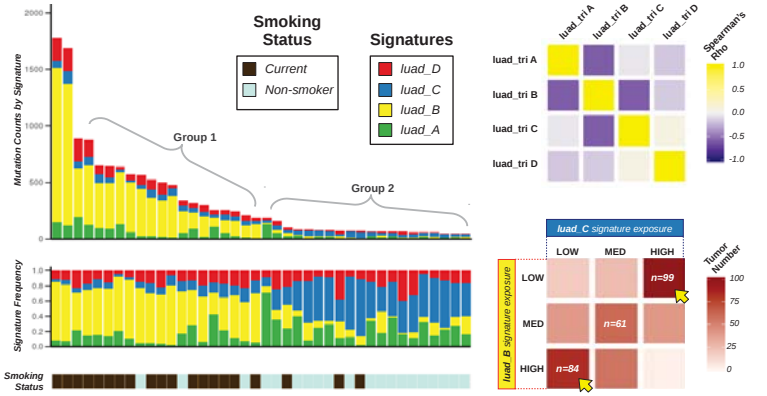
We used the *mutSignatures* framework to extract tri-nucleotide mutational signatures from the lung adenocarcinoma (*LUAD*) TCGA dataset. Mutational signatures are the basic mutational patterns that contribute to DNA mutagenesis in the lung adenocarcinoma genomes, and correspond to the *W* matrix in the NMF equation. In the LUAD TCGA dataset, we found four mutational signatures. The reliability of our method was assessed by comparing our results to the mutational signatures previously identified in lung cancer using the original MATLAB-based framework developed by the Sanger Institute [2]. Our signatures matched those reported before, namely signatures COSMIC (Catalogue Of Somatic Mutations In Cancer) 1, 2, 4, and 5.



Analysis of tri-nucleotide mutational signatures extracted from the LUAD TCGA dataset. *A) Barplots summarizing the mutational profiles of mutational signatures. Relative mutation frequency (y-axis) of every mutation type (x-axis) is visualized. B) Heatmap comparing tri-nucleotide mutational signatures extracted from the LUAD TCGA dataset with known mutational signatures from COSMIC (Sanger Institute). Color intensity tracks with the value of cosine distance.*

Signature Exposures in smokers and non-smokers affected by lung adenocarcinoma

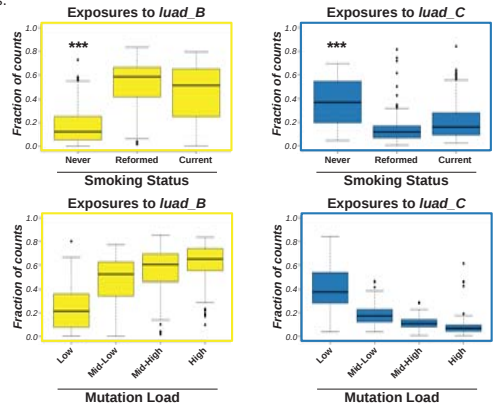
We analyzed the mutational signature exposures across lung cancer patients. Exposures estimate how many mutations were the consequence of each mutational signature in each sample, and correspond to the *H* matrix in the NMF equation. Analysis of signature exposures revealed two groups in the data: *i)* tumors enriched in *luad_B* signature (yellow bars), usually having high mutation burden (group 1); and *ii)* tumors depleted in *luad_B* signature, usually featuring low total number of DNA mutations (group 2). Further analyses showed that in lung adenocarcinoma, signatures *luad_B* (yellow bars) and *luad_C* (blue bars) were inversely correlated and displayed a trend toward mutual exclusion.



Signature Exposures in smokers and non-smokers affected by lung adenocarcinoma. *A) Exposures to mutational signatures that were de novo extracted from LUAD TCGA. A limited number (n=40, including 20 random genomes from smokers and 20 random genomes from life-long non-smokers) of lung cancer samples are displayed. Each bar represents a tumor and the vertical axis denotes the total (top barplot) or the relative (central barplot) number of mutations imputed to each signature (highlighted by colors). The patient smoking status key is shown below the barplots. B) Heatmap showing Spearman correlation coefficients (Rho) across signature exposures in the Lung Adenocarcinoma dataset. C) Heatmap highlighting the distribution of exposures to luad_B (y-axis) and luad_C (x-axis) signatures in LUAD TCGA genomes. Exposures to both signatures were tertile-discretized (low, medium, and high), and then orthogonally analyzed. Tumors belonging to each of the 9 possible groups were counted.*

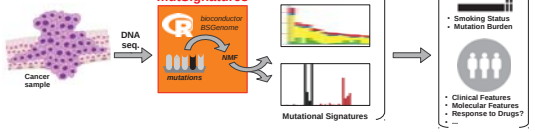
Mutational signatures are linked to clinical and molecular parameters

We examined the association of mutational signatures with clinical or molecular features, specifically patient smoking status, and tumor mutation burden. We confirmed the observation that mutations corresponding to the pattern of signature luad_b accumulated in high mutation burden genomes, as well as in smokers (either reformed, or current) as compared to non-smokers. On the contrary, exposures to luad_C were higher in low-mutation burden samples, as well as genomes from non-smoking patients. This confirmed that mutational signatures can be prognostic of selected clinical and molecular features.



Mutational correlate with clinical parameters in lung adenocarcinomas. *A, B) Boxplots showing relative exposure to signatures luad_B (A) and luad_C (B) according to discretized mutation burden. Mutation burden was quartile-discretized. Correlation between relative exposures and binned mutation burden was computed by Kendall's rank correlation test. Kendall's coefficients (tau) were tau=0.4563 (p-val < 2.2e-16) for luad_B signature (A), and tau=-0.6240 (p-val < 2.2e-16) for luad_C signature (B). C, D) Boxplots showing relative exposures to signatures luad_B (C) and luad_C (D) in LUAD genomes according to patient smoking status. Groups were compared by t-test.*

Conclusions



Our software can be used for the identification of mutational determinants of cancer, supports the analysis of signature-associated molecular and clinical features, and has the potential of revealing insights into tumor biology and treatment.

Acknowledgments

JJM is supported by grant BX003692 and the John P. Hanson Foundation for Cancer Research at the Robert H. Lurie Comprehensive Cancer Center of Northwestern University. DF, JJM designed the research project. DF, YY acquired and prepared the data. DF developed software, wrote R extension. DF and VV performed data analyses, and data visualization.

References

[1] **Fantini D**, Glaser AP, Rimar KJ, et al. ***Oncogene***. 2018 Jan 25
[2] Alexandrov LB, Nik-Zainal S, Wedge DC, et al. ***Nature***. 2013 Aug 22; 500:415-21
[3] Lee DD and Seung S. ***Nature***. 1999 Oct 21; 401:788–791