**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 deconvolving 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 mutation 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) Bistatics summarizing the mutational profiles of lung adenocarcinoma, relative mutation frequency (y-axis) of each mutation type (x-axis) is illustrated. (B) Pairplot comparing tri-nucleotide mutational signatures extracted from the LUAD TCGA dataset with initial 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 W matrix in the NMF equation. Analysis of signature exposures revealed two groups in the data: (i) tumors enriched in load_B signature (yellow bars), usually having high mutation burden (group 1), and (ii) tumors depleted in load_B signature, usually featuring low total number of DNA mutations (group 2). Further analyses showed that in lung adenocarcinoma, signatures load_B (yellow bars) and load_C (blue bars) were inversely correlated and displayed a trend toward mutual exclusion.

**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**
- JKM is supported by grant B003692 and the John P. Hanson Foundation for Cancer Research at the Robert H. Lurie Comprehensive Cancer Center of Northwestern University. DJM 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**