**MetaMutationalSigs: Comparison of mutational signature refitting results made easy**

**Pandey Palash 1\*Sanjeevani Arora2., Gail Rosen1.1 Drexel University, Philadelphia PA; 2 Fox chase cancer center, Philadelphia, PA**

**Abstract:**

**Summary:**

The analysis of mutational signatures is becoming increasingly common in cancer genetics, with emerging implications in cancer evolution, classification, treatment decision and prognosis. ~~Mutations that result in cancers are caused by several mutational processes; mutational signature analysis can identify the contribution of these processes to observed mutational patterns.~~ Recently, several packages have been developed for mutational signature analysis, with each using different methodology and yielding significantly different results. Because of the nontrivial differences in tools’ refitting results, researchers may desire to survey and compare the available tools, in order to objectively evaluate the results for their specific research question, such as which mutational signatures are prevalent in different cancer types. There is a need for a software that can aggregate results from different refitting packages and present them in a user-friendly way to facilitate effective comparison of mutational signatures.

**Availability and implementation:**

MetaMutationalSigs is implemented using R and python and is available for installation using Docker and available at: <https://github.com/PalashPandey/MetaMutationalSigs>

**Contact:**

Palash Pandey (pp535@drexel.edu).

**Supplementary information:**

More information about the package including test data and results are availab;e at <https://github.com/PalashPandey/MetaMutationalSigs>

**Introduction:**

~~Cancerous~~ Mutational signature analysis provides an operative framework to understand the somatic evolution of cancer from normal tissue [19, 21, 22, 23, 24]. From the earliest phases of neoplastic changes, cells may acquire several types of mutations in the form of single nucleotide variants, insertions and deletions, copy number changes and chromosomal aberrations. These mutations are ~~hypothesized to be~~ caused by multiple mutational processes operative in cancer leaving behind specific footprints in the DNA that can by captured by ~~tumor~~ mutational signature analysis[4,20]. It is becoming increasingly evident that these ~~tumor~~ mutational signatures are not only important for understanding cancer evolution but also may have therapeutic implications, thus this an a very active and important area of research [1,2,3,20].

The basic idea behind mutational signatures is that mutational processes create specific patterns of mutations. Thus, it follows that if one can identify these patterns in a given sample then they can essentially detect the corresponding mutational processes. The possible mutations are grouped into 6 mutation types based on the base where the mutation was observed. These 6 mutation types are C>A, C>G, C>T, T>A, T>C, and T>G. Now, these 6 types of mutations are further divided based on their location, i.e., other bases that are in their immediate proximity giving us the 96 mutation types that are termed the single based substitution context. Alexandrov et al. first developed and applied this idea to The Cancer Genome Atlas (TCGA) data and identified the first iteration of 30 single base substitution (SBS) signatures ~~termed as COSMIC signatures~~, which were compiled into the Catalogue Of Somatic Mutations In Cancer (COSMIC) and came to be used as the de facto reference for signature refitting [7,5]. The initial study was then expanded to the analysis of data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project [7], resulting in two additional ~~signatures using different contexts~~ signature classes with multiple signatures in each class. These new classes are double base substitutions (DBS) signatures and insertions/deletion (ID) signatures, which are in addition to the reference SBS signatures [8].

The mutational signature analysis workflows involves multiple steps that require different amounts of time and processing power. Briefly, the workflow starts from Binary Alignment Format (BAM) files that are aligned to a reference genome and then proceeds to the variant calling step which outputs the Variant Calling Format (VCF) files. These steps are usually very resource-intensive and thus do not allow for much experimentation on personal computers (the downstream steps of variant filtering and annotation are much faster). The final step, the mutational signature analysis, is the least resource-intensive and, therefore, is easier for users to compare multiple methods on their desktop. Therefore, to facilitate comprehensive mutational signature analyses, we developed the package, MetaMutationalSigs. We developed a wrapper for 4 typically-used refitting packages [10,11,12,13], that have various underlying methodologies, such as Bayesian inference, non-negative least squares, and quadratic programming. Here, we have developed a standard format for inputs and outputs for easy interoperability and effective comparison, respectively. With our previous experience in visualization of genomic data [18], we have also implemented standard visualizations for the results of all mutational signature packages to ensure easy analysis. MetaMutationalSigs software is easy to install and use through Docker.

**Approach:**

The two major methods typically used for mutational signature analysis are signature refitting and de-novo signature extraction. Signature refitting methods try to reconstruct the observed mutational pattern in the sample (the frequencies of 96 types of mutations) using linear combinations of known signatures (COSMIC Legacy SBS and COSMIC V3 SBS, ID, DBS, etc.), these methods work quite well on small sample sizes (such as single samples) and are widely used with small datasets [6]. Signature extraction methods infer signatures from a given dataset ~~of samples~~, and then compare the extracted signatures with known reference signatures. Each extracted signature is assigned to a known signature if their cosine similarity exceeds a set threshold, otherwise signatures with similarity less than the threshold are ignored [4]. There are a few important caveats to signature extraction as recently discussed in [9]: 1) a novel signature can be very similar to several reference signatures and the assignment is not always perfect and 2) the threshold for assignment plays a crucial role but is not widely agreed upon, using a different threshold can change the assignment [9].

We chose signature refitting as our primary task and implemented high performing packages as identified in [9] that were implemented in R. We implement DeconstructSigs [10], MutationalPatterns [11], Sigfit [12], Sigminer [13], these tools build up on other tools such as [14, 15]. Our package outputs several data files in comma, separated values (CSV) format ready for further analysis and visualization using external packages along with visualizations of the signature exposures as described in **Table 1**.

We use root mean squared error (RMSE) between the reconstructed and actual signals (a metric commonly used in signal processing [17]] )as our performance metric for comparison of these packages.

**Discussion:**

The massive increase in the number of software packages has made managing dependencies quite burdensome, coupled with incompatible data formats for signature matrices can make mutational signature refitting results difficult and hard to compare. Our package provides an easy way of performing these setup related tasks so that more focus can be placed on the analysis. Investigators should keep in mind that refitting approaches need a priori knowledge about the samples for effective interpretation [16], and the results should not be used as-is without a sanity check.

Future work for this project would focus on expanding the tool to work with more packages and keep the reference signatures updated as new versions are released. Due to the open-source nature of the project, we also welcome additional feature requests using the project link on GitHub <https://github.com/PalashPandey/MetaMutationalSigs>

# Conflict of Interest.

# **The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.** Ethics Statement.

Not applicable to this study.

# Data Availability Statement.

All test data used is open source and is available with the software at GitHub <https://github.com/PalashPandey/MetaMutationalSigs>

**Funding.**

P.P. and G.R. was supported by NSF awards # 1936791 and #1919691, Fox Chase Cancer Center Risk Assessment Program Funds. S.A. was supported by DOD W81XWH-18-1-0148 (to S.A.).

**References:**

[1] Iqbal, W., Demidova, E., Serrao, S., ValizadehAslani, T., Rosen, G., & Arora, S. (2020). RRM2B is frequently amplified across multiple tumor types: non-oncogenic addiction and therapeutic opportunities. doi: 10.1101/2020.09.10.291567

[2] Campbell, B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., & de Borja, R. et al. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. *Cell*, *171*(5), 1042-1056.e10. doi: 10.1016/j.cell.2017.09.048

[3] Chung, J., Maruvka, Y., Sudhaman, S., Kelly, J., Haradhvala, N., & Bianchi, V. et al. (2020). DNA polymerase and mismatch repair exert distinct microsatellite instability signatures in normal and malignant human cells. *Cancer Discovery*, CD-20-0790. doi: 10.1158/2159-8290.cd-20-0790

[4] Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Reports. 2013;3: 246–259. pmid:23318258

[5] Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research. 2017;45: D777–D783. pmid:27899578

[6] [dataset] Cancer Genome Atlas Research Network et al. “The Cancer Genome Atlas Pan-Cancer analysis project.” Nature genetics vol. 45,10 (2013): 1113-20. doi:10.1038/ng.2764

[7] [dataset] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium., Campbell, P.J., Getz, G. et al. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020)

[8] Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AW, Boot A, et al. The Repertoire of Mutational Signatures in Human Cancer. bioRxiv. 2018; https://www.nature.com/articles/s41586-020-1943-3[9] Omichessan H, Severi G, Perduca V (2019) Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. PLOS ONE 14(9): e0221235

[10] Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology. 2016;17. pmid:26899170

[11] Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Medicine. 2018;10. pmid:29695279

[12] Gori K, Baez-Ortega A. sigfit: flexible Bayesian inference of mutational signatures. bioRxiv. 2018

[13] Wang, Shixiang, et al. “Copy number signature analyses in prostate cancer reveal distinct etiologies and clinical outcomes” medRxiv (2020)

[14] Mayakonda, Anand, et al. “Maftools: efficient and comprehensive analysis of somatic variants in cancer.” Genome research 28.11 (2018): 1747-1756

[15] Huang X, Wojtowicz D, Przytycka TM. Detecting presence of mutational signatures in cancer with confidence. Bioinformatics. 2018;34: 330–337

[16] Maura, F., Degasperi, A., Nadeu, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications* 10, 2969 (2019)

[17] G. Rosen, "Comparison of Autoregressive Measures for DNA Sequence Similarity," 2007 IEEE International Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland, 2007, pp. 1-4, doi: 10.1109/GENSIPS.2007.4365814

[18] Yemin Lan, J. Calvin Morrison, Ruth Hershberg, Gail L. Rosen, POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes, Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D625–D632, <https://doi.org/10.1093/nar/gkt1094>

[19] Robinson, P., Coorens, T., Palles, C., Mitchell, E., Abascal, F., & Olafsson, S. et al. (2020). Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases. doi: 10.1101/2020.06.23.167668

[20] Alexandrov, L., Kim, J., Haradhvala, N., Huang, M., Tian Ng, A., & Wu, Y. et al. (2020). The repertoire of mutational signatures in human cancer. Nature, 578(7793), 94-101. doi: 10.1038/s41586-020-1943-3

[21] Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 574, 532–537, doi: 10.1038/s41586-019-1672-7 (2019)

[22] Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature 574, 538–542, doi: 10.1038/s41586-019-1670-9 (2019).

[23] Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature 578, 266–272, doi: 10.1038/s41586-020-1961-1 (2020).[24] Moore, L. et al. The mutational landscape of normal human endometrial epithelium. Nature 580, 640–646, doi: 10.1038/s41586-020-2214-z (2020).



**Figure 1. Workflow and results for metaMutationalSigs.** A) The Workflow for mutational signature analysis, starting with a BAM file of a sequenced genome or exome and followed by variant calling, filtering, and annotation. Our tool MetaMutationalSigs analyzes signatures found in a variant calling file. B) Heatmap of euclidean distance between the predicted exposures of COSMIC legacy vs. COSMIC v3 signatures by different tools for a sample. With the legacy signatures, tools are generally less in agreement in their resulting signature contributions, while with COSMIC v3 signatures, mutationalPatterns and Sigfit are more in agreement and different from Sigflow and deconstructSigs, which are in great agreement with each other. C) Reconstruction error (RMSE) using for each tool and reference signatures (lower values are better) – while RMSE does not change for mutationalPatterns or deconstructSigs, Sigflow and Sigfit RMSE significantly drops with COSMIC v3. D) Heatmaps of COSMIC legacy signature contributions vs. COSMIC v3 exposures, one row per sample. The first two samples, while different, yield identical signature contributions while sample 3 has some differences. The COSMIC V3 refitting reveals more reference signatures may be playing a role in the overall signature than is found with the legacy signatures.

**Table 1. Summary of result files.**

|  |  |  |
| --- | --- | --- |
| File Name | Format | Description |
| Heatmap\_exposures\_all\_sigs\_legacy.pdf | pdf | Contributions for all COSMIC Legacy 30 SBS signatures to the overall signature. |
| Heatmap\_exposures\_all\_sigs\_SBS.pdf | pdf | Exposures for all COSMIC V3 SBS signatures. |
| Heatmap\_COSMIC\_legacy.pdf | pdf | Heatmap for difference between the predicted contributions by different tools. One for each sample. |
| Heatmap\_COSMIC\_V3.pdf | pdf | Heatmap for difference between the predicted contributions by different tools. One for each sample. |
| legacy\_pie\_charts.html | html | Interactive pie charts of COSMIC legacy contribution exposures, per sample and for each tool. |
| sbs\_pie\_charts.html | html | Interactive pie charts of COSMIC V3 signature exposures, per sample and for each tool. |
| legacy\_rmse\_scatter\_plot.png | png | Reconstruction error using COSMIC Legacy signatures for each tool. |
| sbs\_rmse\_scatter\_plot.png | png | Reconstruction error using COSMIC V3 signatures for each tool. |
| toolname\_results\legacy\_sample\_error.csv | csv | Data used to create the bar plot. |
| toolname\_results\legacy\_sample\_contribution.csv | csv | Data used to create heatmap and pie chart. |
| toolname\_results\sbs\_sample\_error.csv | csv | Data used to create the bar plot. |
| toolname\_results\sbs\_sample\_contribution.csv | csv | Data used to create heatmap and pie chart. |