MetaMutationalSigs: Comparison of mutational signature refitting results made easy

Pandey Palash [1*]Sanjeevani Arora[1,] ., Gail Rosen[1].[1]

Any other affiliations to add below?? Sanjee is 1 and 3. Palash will be 1, 2 and ???, Gail is 2.

[1]Cancer Prevention and Control Program, Fox Chase Cancer Center, Philadelphia, PA, USA
[2]Department of Electrical and Computer Engineering, College of Engineering, Drexel University, Philadelphia, PA, USA
[3]Department of Radiation Oncology, Fox Chase Cancer Center, Philadelphia, PA, USA
[‡]Authors contributed equally

***Correspondence:**
Gail Rosen, PhD
ADD INFO

OR

Sanjeevani Arora, PhD
Assistant Professor
Cancer Prevention and Control Program
Fox Chase Cancer Center
333 Cottman Avenue
Philadelphia, PA 19111
Tel. 215-214-3956
Fax: 215-728-4333
Email: Sanjeevani.Arora@fccc.edu

Commented [SA1]: EDIT this to reflect everyone's affiliations with numbers AND co-corresponding with an *

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 available at https://github.com/PalashPandey/MetaMutationalSigs

Commented [R2]: Somatic mutations in cancerous tumors are thought to be caused by independent mutational processes?

Commented [SA3]: Does this belong here or at the end? Check journal format

Commented [R4]: Palash, since I'm a little more permanent at Drexel, I think that I should be the contact author. But I know this makes submitting the manuscript difficult. Let's talk about this.

Commented [SA5]: Shouldn't this all be in sections below on data availability? Check journal format?

Introduction.

~~Cancerous~~ Mutational signature analysis provides an operative framework to understand the somatic evolution of cancer from normal tissue (Robinson et al 2020; Alexandrov et al, 2020; Brunner et al., 2019; Yoshida et al., 2020; Moore et al.,2020). 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(Alexandrov et al.,2013; Alexandrov et al, 2020). 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 a very active and important area of research (Iqbal et al., 2020; Campbell et al., 2017; Chung et al., 2020; Alexandrov et al., 2020).

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 providing the 96 mutation types that are termed the single base substitution (SBS) context. Alexandrov et al. first developed and applied this idea to The Cancer Genome Atlas (TCGA) data and identified the first iteration of 30 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, we refer to these as COSMIC legacy SBS signatures (Campbell et al., 2020; Forbes et al., 2017). The initial study was then expanded to the analysis of data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (Campbell et al., 2020), resulting in two additional signature classes with multiple signatures in each class. These new classes are COSMIC V3 SBS, double base substitutions (DBS) signatures and insertions/deletion (ID) signatures, which are in (Alexandrov et al., 2018).

The mutational signature analysis workflow 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

Commented [SA6]: cite-
https://www.biorxiv.org/content/10.1101/2020.06.23.167668v1.full
Olafsson, S. et al. The mutational profile and clonal landscape of the inflammatory bowel disease affected colon. bioRxiv (2019)

- *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).*
CrossRefPubMedGoogle Scholar
- *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).*
CrossRefPubMedGoogle Scholar
- *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).*
CrossRefGoogle Scholar
- *Moore, L. et al. The mutational landscape of normal human endometrial epithelium. Nature 580, 640–646, doi: 10.1038/s41586-020-2214-z (2020).*

Commented [SA7]: Cite Alexandrov LB 2013 and this paper https://www.nature.com/articles/s41586-020-1943-3

Commented [SA8]: Also cite-
https://www.nature.com/articles/s41586-020-1943-3

Commented [SA9]: include SBS here..

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 refitting analyses, we developed the package, MetaMutationalSigs. We developed a wrapper for 4 typically used refitting packages (Rosenthal et al., 2016; Blokzijil et. al. 2018; Gori et. al., 2018; Wang et al., 2020), 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 (Yemin et al., 2014) 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.

Commented [R10]: Refitting?

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, 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 (Alexandrov et al.,2013). There are a few important caveats to signature extraction as recently discussed in (Omichessan et al., 2019): 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 (Omichessan et al., 2019).

Commented [SA11]: Maybe use a diff word here…

Commented [R12]: What happens to truly novel signatures?

We chose signature refitting as our primary task and implemented high performing packages as identified in (Omichessan et al., 2019) that were implemented in R using common input matrix generated using SigProfilerMatrixGenerator (Bergstrom et al., 2019). We implement

DeconstructSigs (Rosenthal et al., 2016), MutationalPatterns (Blokzijil et. al. 2018) , Sigfit (Gori et. al., 2018), Sigminer (Wang et al., 2020), these tools build up on other tools such as (Mayakonda et al., 2018; Huang et al., 2018). 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 contributions 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 (Rosen, 2007) as our performance metric for comparison of these packages.

Commented [R13]: Let's change this to signature contributions throughout the manuscript

Commented [R14]: This reference: https://ieeexplore.ieee.org/abstract/document/4365814

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 (Maura et al., 2019), 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:

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

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

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

Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, Alexandrov LB. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics. 2019 Aug 30;20(1):685. doi: 10.1186/s12864-019-6041-2. PMID: 31470794; PMCID: PMC6717374.

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

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).

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

[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

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

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

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

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

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

[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)

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

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)

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

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

Moore, L. et al. The mutational landscape of normal human endometrial epithelium. Nature 580, 640–646, doi: 10.1038/s41586-020-2214-z (2020).

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

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

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

Wang, Shixiang, et al. "Copy number signature analyses in prostate cancer reveal distinct etiologies and clinical outcomes" medRxiv (2020)

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

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).

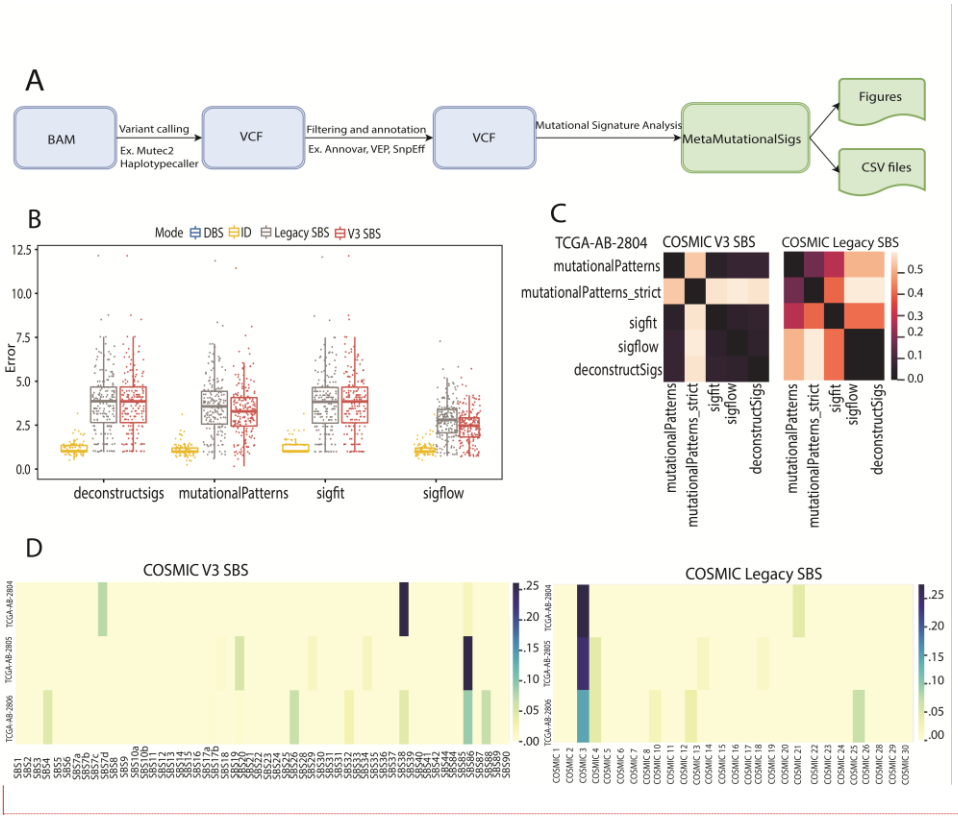

Figure 1. Workflow and results for metaMutationalSigs. A) The Workflow for mutational signature analysis, starting with a BAM file of a sequenced genome or exome and followed by variant calling, filtering, and annotation. Our tool MetaMutationalSigs analyzes signatures found in a variant calling file. B) Heatmap of Euclidean distance between the predicted contributions of COSMIC legacy SBS vs. COSMIC v3 SBS 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 SBS signature contributions vs. COSMIC v3 SBS, 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.

Commented [R15]: how's the combo (to enlarge) and the scatterplot going. If combo'd change to A->D … 4 panels. Make sure that you change legend in the RMSE panel to say COSMIC Legacy and COSMIC V3; get rid of F label. I would say to get rid of test1 or test 2.. they seem to get the same "exposure" signatures

Table 1. Summary of result files.

Commented [R16]: I guess I'd like to go over this with you.

File Name	Format	Description
Heatmap_contributions_all_sigs_legacy.pdf	pdf	Contributions for all COSMIC Legacy SBS signatures to the overall signature.
Heatmap_contributions_all_sigs_SBS.pdf	pdf	Contributions 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 SBS contribution, per sample and for each tool.
sbs_pie_charts.html	html	Interactive pie charts of COSMIC V3 SBS signature contributions, per sample and for each tool.
legacy_rmse_bar_plot.pdf	pdf	Reconstruction error using COSMIC Legacy SBS signatures for each tool.
sbs_rmse_bar_plot.png	pdf	Reconstruction error using COSMIC V3 SBS 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.