MetaMutationalSigs: Comparison of mutational signature refitting results made easy

Palash Pandey [1,2], Sanjeevani Arora[1,3*], Gail Rosen[2*]

[1]Cancer Prevention and Control Program, Fox Chase Cancer Center, Philadelphia, PA, USA
[2]Ecological and Evolutionary Signal-processing and Informatics Laboratory, 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

***Correspondence:**
Gail Rosen, PhD
Professor
Ecological and Evolutionary Signal-processing and Informatics Laboratory
Department of Electrical and Computer Engineering
College of Engineering
3141 Chestnut St.
Drexel University
Philadelphia, PA, 19104
USA
Telephone: 215-895-0400
Email: glr26@drexel.edu

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
Email: Sanjeevani.Arora@fccc.edu

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. 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/EESI/MetaMutationalSigs

Contact:

Gail Rosen (glr26@drexel.edu).

Supplementary information:

More information about the package including test data and results are available at https://github.com/EESI/MetaMutationalSigs

Introduction.

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 caused by multiple mutational processes operative in cancer leaving behind specific footprints in the DNA that can by captured by mutational signature analysis(Alexandrov et al.,2013; Alexandrov et al, 2020). It is becoming increasingly evident that these 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 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 v2 signatures 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

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 (Lan et al., 2014), we have 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 (Omichessan et al., 2019). 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 and using a different threshold can change the assignment (Omichessan et al., 2019).

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 compare packages using the root mean squared error (RMSE) between the reconstructed and actual signals. RMSE is a performance metric commonly used in signal processing (Rosen, 2007).

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, MetaMutationalSigs, provides a simplified approach for performing the 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 and each package for effective interpretation (Maura et al., 2019), and the results should not be used as-is without an assessment of the cell biology and genomics.

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/EESI/MetaMutationalSigs

Conflict of Interest.

S.A. performs collaborative research (with no funding) with Caris Life Sciences, Foundation Medicine, Inc., Ambry Genetics and Invitae Corporation. S.A. has several patents and/or pending patents related to colorectal cancer diagnostics/treatment. All other authors declare no competing interests.

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/EESI/MetaMutationalSigs

Funding.

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

References:

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., … Tian Ng, A. W. (2020). The repertoire of mutational signatures in human cancer. Nature, 578(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Reports, 3(1), 246–259. https://doi.org/10.1016/j.celrep.2012.12.008

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., … Tian Ng, A. W. (2020). The repertoire of mutational signatures in human cancer. Nature, 578(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Bergstrom, E. N., Huang, M. N., Mahto, U., Barnes, M., Stratton, M. R., Rozen, S. G., & Alexandrov, L. B. (2019). SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics, 20(1). https://doi.org/10.1186/s12864-019-6041-2

Blokzijl, F., Janssen, R., van Boxtel, R., & Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Medicine, 10(1). https://doi.org/10.1186/s13073-018-0539-0

Brunner, S. F., Roberts, N. D., Wylie, L. A., Moore, L., Aitken, S. J., Davies, S. E., … Campbell, P. J. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature, 574(7779), 538–542. https://doi.org/10.1038/s41586-019-1670-9

Campbell, B. B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., … Shlien, A. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. Cell, 171(5), 1042-1056.e10. https://doi.org/10.1016/j.cell.2017.09.048

[dataset] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., … Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764

Chung, J., Maruvka, Y. E., Sudhaman, S., Kelly, J., Haradhvala, N. J., Bianchi, V., … Tabori, U. (2020). DNA polymerase and mismatch repair exert distinct microsatellite instability signatures in normal and malignant human cells. Cancer Discovery, CD-20-0790. https://doi.org/10.1158/2159-8290.cd-20-0790

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Campbell, P. J. (2016). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research, 45(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121

Gori, K., & Baez-Ortega, A. (2018). sigfit: flexible Bayesian inference of mutational signatures. Cold Spring Harbor Laboratory. https://doi.org/10.1101/372896

Huang, X., Wojtowicz, D., & Przytycka, T. M. (2017). Detecting presence of mutational signatures in cancer with confidence. Bioinformatics, 34(2), 330–337. https://doi.org/10.1093/bioinformatics/btx604

[dataset] (2020). Pan-cancer analysis of whole genomes. Nature, 578(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Iqbal, W., Demidova, E. V., Serrao, S., ValizadehAslani, T., Rosen, G., & Arora, S. (2021). RRM2B Is Frequently Amplified Across Multiple Tumor Types: Implications for DNA Repair, Cellular Survival, and Cancer Therapy. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.628758

Lan, Y., Morrison, J. C., Hershberg, R., & Rosen, G. L. (2013). POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes. Nucleic Acids Research, 42(D1), D625–D632. https://doi.org/10.1093/nar/gkt1094

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R. J., Sanders, M. A., Moore, L., … Stratton, M. R. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. Nature, 574(7779), 532–537. https://doi.org/10.1038/s41586-019-1672-7

Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., … Bolli, N. (2019). A practical guide for mutational signature analysis in hematological malignancies. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-11037-8

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Research, 28(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

Moore, L., Leongamornlert, D., Coorens, T. H. H., Sanders, M. A., Ellis, P., Dentro, S. C., … Stratton, M. R. (2020). The mutational landscape of normal human endometrial epithelium. Nature, 580(7805), 640–646. https://doi.org/10.1038/s41586-020-2214-z

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. https://doi.org/10.1371/journal.pone.0221235

Robinson, P. S., Coorens, T. H. H., Palles, C., Mitchell, E., Abascal, F., Olafsson, S., … Stratton, M. R. (2020). Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2020.06.23.167668

Rosen, G. (2007). Comparison of Autoregressive Measures for DNA Sequence Similarity. 2007 IEEE International Workshop on Genomic Signal Processing and Statistics. Presented at the 2007 IEEE International Workshop on Genomic Signal Processing and Statistics. https://doi.org/10.1109/gensips.2007.4365814

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology, 17(1). https://doi.org/10.1186/s13059-016-0893-4

Wang, S., Li, H., Song, M., He, Z., Wu, T., Wang, X., … Liu, X.-S. (2020). Copy number signature analyses in prostate cancer reveal distinct etiologies and clinical outcomes. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2020.04.27.20082404

Yoshida, K., Gowers, K. H. C., Lee-Six, H., Chandrasekharan, D. P., Coorens, T., Maughan, E. F., … Campbell, P. J. (2020). Tobacco smoking and somatic mutations in human bronchial epithelium. Nature, 578(7794), 266–272.https://doi.org/10.1038/s41586-020-1961-1

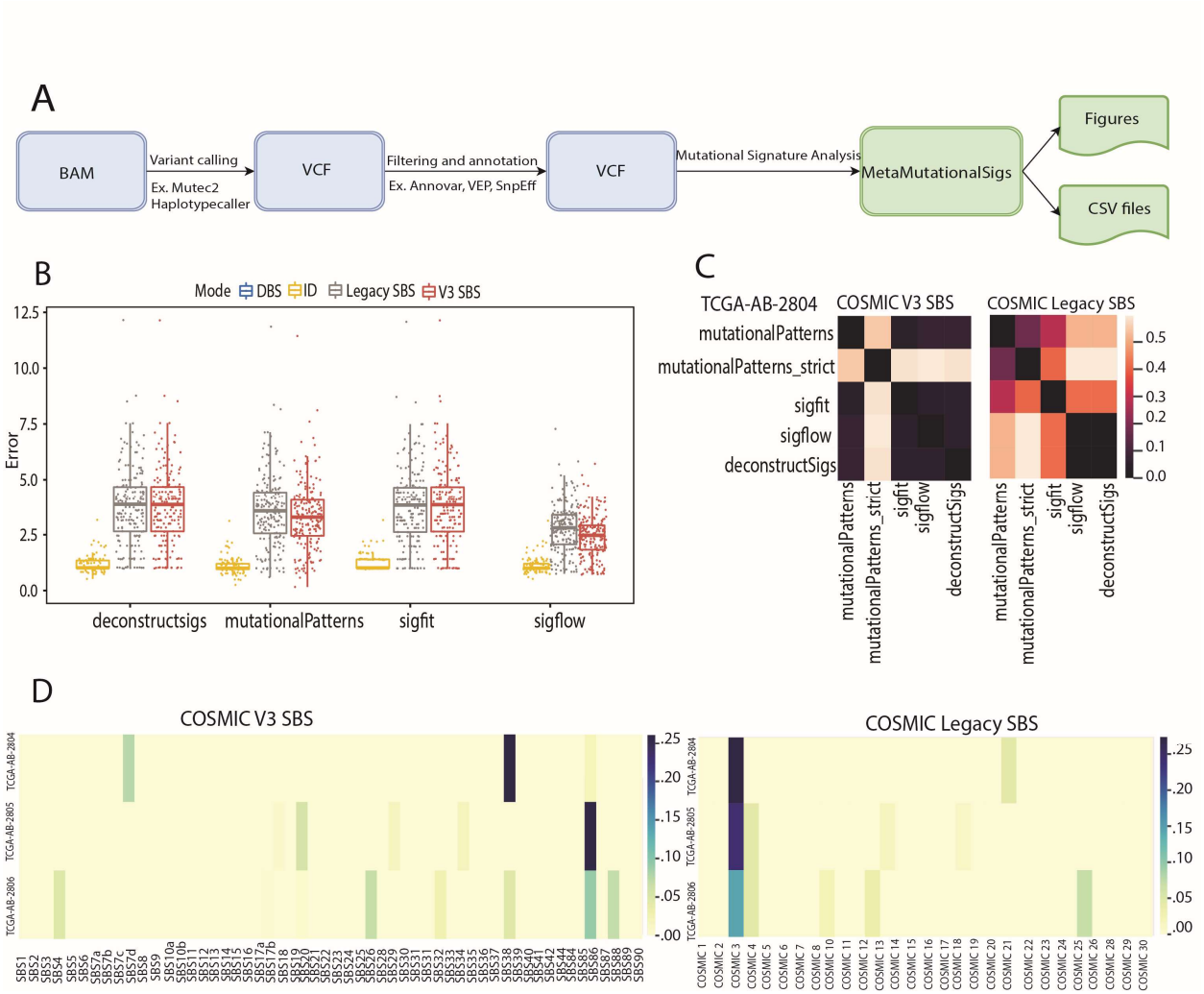
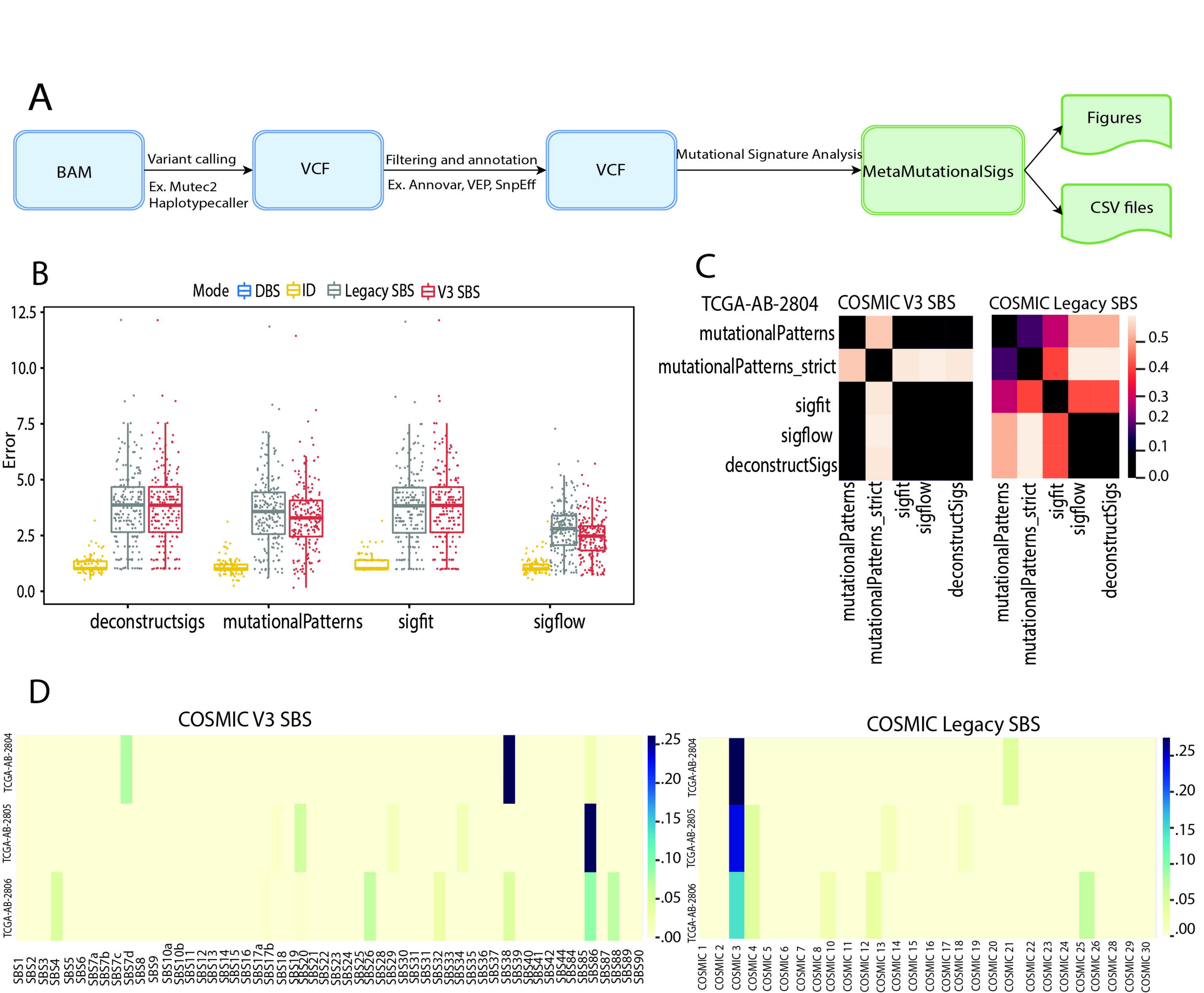


Figure 1. Workflow and results for metaMutationalSigs. A) The workflow for mutational signature analysis, starting with a BAM file of a sequenced genome or exome and followed by variant calling, filtering, and annotation. Our tool, MetaMutationalSigs, analyzes signatures found in a VCF. **B**) RMSE for each tool and reference signatures (lower values are better) for SBS and IDs (no tool predicted DBS for the samples used). While RMSE does not change for deconstructSigs and Sigfit, the RMSE significantly drops for mutationalPatterns and Sigflow with COSMIC v3. **C**) Heatmap of Euclidean distance between the predicted contributions of COSMIC v3 SBS vs. COSMIC legacy SBS signatures by different tools for the same TCGA patient sample. With the legacy signatures, tools are generally less in agreement in their resulting signature contributions, while with COSMIC v3 signatures, the standard use tools are all in agreement with each other. Sigflow had the lowest RMSE and was selected for analysis in **D.** **D**) Heatmaps of COSMIC v3 SBS vs. COSMIC legacy SBS mutational signature contributions using whole-exome sequence data from three TCGA patients with acute myeloid leukemia. Here, each row is a patient sample. Left. COSMIC v3 SBS refitting provides different dominant signature contributions, TCGA-AB-2804: unknown etiology, TCGA-AB-2805 and TCGA-AB-2806: unknown

chemotherapy and different DNA mismatch repair signatures, SBS20 and 26, respectively. COSMIC Legacy SBS refitting provides signature 3 (failure of double-strand break-repair by homologous recombination) as the dominant signature for all samples. The COSMIC v3 SBS refitting reveals multiple mutational processes may be playing a role in the overall signature contribution than is found with the COSMIC Legacy SBS refitting. =

Table 1. Summary of result files.

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.


A
BAM
Variant calling
Ex. Mutec2
Haplotypecaller
VCF
Filtering and annotation
Ex. Annovar, VEP, SnpEff
VCF
Mutational Signature Analysis
MetaMutationalSigs
Figures
CSV files
B
Mode DBS ID Legacy SBS V3 SBS
Error
deconstructsigs
mutationalPatterns
sigfit
sigflow
C
TCGA-AB-2804
COSMIC V3 SBS
COSMIC Legacy SBS
mutationalPatterns
mutationalPatterns_strict
sigfit
sigflow
deconstructSigs
D
COSMIC V3 SBS
COSMIC Legacy SBS
TCGA-AB-2804
TCGA-AB-2805
TCGA-AB-2806