Genome analysis

MetaMutationalSigs: comparison of mutational signature refitting results made easy

Palash Pandey\textsuperscript{1,2}, Sanjeevani Arora\textsuperscript{2,3,*} and Gail L. Rosen\textsuperscript{1}*

\textsuperscript{1}Ecological and Evolutionary Signal-Processing and Informatics Laboratory, Department of Electrical and Computer Engineering, College of Engineering, Drexel University, Philadelphia, PA 19104, USA, \textsuperscript{2}Cancer Prevention and Control Program, Fox Chase Cancer Center, Philadelphia, PA 19111, USA and \textsuperscript{3}Department of Radiation Oncology, Fox Chase Cancer Center, Philadelphia, PA 19111, USA

*To whom correspondence should be addressed.

Received on April 16, 2021; revised on December 7, 2021; editorial decision on February 1, 2022; accepted on February 9, 2022

Abstract

Motivation: 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 non-trivial 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.

Results: Due to the need for effective comparison of refitting mutational signatures, we introduce a user-friendly software that can aggregate and visually present results from different refitting packages.

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: Sanjeevani.Arora@fccc.edu or glr26@drexel.edu

1 Introduction

Mutational signature analysis provides an operative framework to understand the somatic evolution of cancer from normal tissue (Alexandrov et al., 2020; Brunner et al., 2019; Moore et al., 2020; Robinson et al., 2020; Yoshida 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 specific footprints in the DNA that can by captured by mutational signature analysis (Alexandrov et al., 2013, 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 (Alexandrov et al., 2020; Campbell et al., 2017; Chung et al., 2021; Iqbal et al., 2021).

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 six single mutation types based on the base where the mutation was observed. These six single mutation types are C $\rightarrow$ A, C $\rightarrow$ G, C $\rightarrow$ T, T $\rightarrow$ A, T $\rightarrow$ C and T $\rightarrow$ G. Now, these six types of single mutations are further divided based on their context, e.g. one base preceding and
methods on their desktop. Therefore, to facilitate comprehensive mutational signature refitting analyses, we developed the package, MetaMutationalSigs, to analyze the mutational signatures in the VCF files. We developed a wrapper for four typically used refitting packages (Bloknijl et al., 2018; Gori and Baez-Ortega, 2018; Rosenthal et al., 2016; Wang et al., 2020), that have diverse underlying methodologies, including multiple linear regression, non-negative least squares (NNLS), Bayesian inference and simulated annealing (SA), respectively. 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.

2 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 (Ormichessan 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 Ormichessan et al. (2019): (i) a novel signature can be very similar to several reference signatures and the assignment is not always perfect and (ii) the threshold for assignment plays a crucial role but is not widely agreed upon and using a different threshold can change the assignment (Ormichessan et al., 2019).

We chose signature refitting as our primary task because refitting techniques use COSMIC signatures that are well-established, are able to analyze signatures in smaller sets of samples than de-novo techniques and are computationally less intense than de-novo techniques. We implemented high performing packages as identified in Ormichessan et al. (2019) that were implemented in R using a common input matrix generated using SigProfilerMatrixGenerator (Bergstrom et al., 2019). While other techniques exist, including convenient web-based tools, such as Mutalisk (Lee et al., 2018), that refits using a maximum likelihood estimation of the signature contributions, and Signal (Degasperi et al., 2020), which uses quadratic programming or SA, there may be a desire to run the mutational signature analysis on local machines, that do not require uploading data to a third party. To address this problem, we implement a software package to compare four mutational signature analyzers—DeconstructSigs (Rosenthal et al., 2016), MutationalPatterns (Bloknijl et al., 2018), Sigfit (Gori and Baez-Ortega, 2018), Siger (Wang et al., 2020), which build up on other tools such as Mayakonda et al. (2018) and Huang et al. (2018). DeconstructSigs is the most cited method and uses a multiple linear regression model, with coefficients constrained to positive values, to find contributions of mutational signatures to an overall signature. MutationalPatterns is also popular and uses NNLS optimization to estimate the mutational signature weighting. Sigfit uses Bayesian inference to perform fitting. Siger uses an SA method to mutational signature fitting. Between multiple linear regression, optimization via NNLS and SA, and Bayesian estimation, users can survey how a variety of techniques can estimate COSMIC signatures in sample(s).

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. In Figure 1A, we illustrate the workflow of the analysis (including preprocessing steps in blue and our package’s steps in green). In Figure 1B, we compare packages using the root mean squared error (RMSE) between the reconstructed and actual signals for 188 myeloid leukemia (LAML) patients obtained from The Cancer Genome Atlas portal (Weinstein et al., 2013). RMSE is a performance metric commonly used in signal processing (Rosen, 2007). In Figure 1C, we plot heatmaps of distances between methods for predicting signature contributions for SBS V2 (Legacy) and V3. In Figure 1D, we plot the heatmap of SBS contributions to overall signature reconstruction for three of the LAML patient samples.

3 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/ESI-MetaMutationalSigs.
Funding

P.P. and G.R. were supported by the NSF [awards #1936791 and #1919691], and P.P. was also supported by the Fox Chase Cancer Center Risk Assessment Program Funds. S.A. was supported by the DOD W81XWH-18-1-0148 award and a CEP award from the Yale Head and Neck Cancer NIH SPORE.

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 cancer diagnostics/ treatment. All other authors declare no competing interests.

References

Alexandrov,L.B. et al. (2013) Deciphering signatures of mutational processes operative in human cancer. Cell Rep., 3, 246–259.

Alexandrov,L.B. et al.; PCAWG Consortium. (2020) The repertoire of mutational signatures in human cancer. Nature, 578, 94–101.

Bergstrom,E.N. et al. (2019) SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics, 20, 685.

Blokzijl,F. et al. (2018) MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med., 10, 33.

Brunner,S.F. et al. (2019) Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature, 574, 530–542.

Camplin,B.B. et al. (2017) Comprehensive analysis of hypermutation in human cancer. Cell, 171, 1042–1056.e10.

Campbell,R.P. et al. (2020) Pan-Cancer analysis of whole genomes. Nature, 578, 82–93.

Chung,J. et al. (2021) DNA polymerase and mismatch repair exert distinct microsatellite instability signatures in normal and malignant human cells. Cancer Disc., 11, 1176–1191.

Degasperi,A. et al. (2020) A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. Nat. Cancer, 1, 249–263.

Forbes,S.A. et al. (2017) COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res., 45, D777–D783.

Gori,K. and Baez-Ortega,A. (2018) Sigfit: Flexible Bayesian Inference of Mutational Signatures. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. doi:10.1101/372896

Huang,X. et al. (2018) Detecting presence of mutational signatures in cancer with confidence. Bioinformatics, 34, 330–337.

Iqbal,W. et al. (2021) RRM2B is frequently amplified across multiple tumor types: implications for DNA repair, cellular survival, and cancer therapy. Front. Genet., 12, 628758.

Lan,Y. et al. (2014) POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes. Nucleic Acids Res., 42, D625–D632.

Lee,J. et al. (2018) Mutalisk: a web-based somatic MUTation AnalyLyS toolkit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Res., 46, W102–W108.
Maura,F. et al. (2019) A practical guide for mutational signature analysis in hematological malignancies. Nat. Commun., 10, 2969.
Mayakonda,A. et al. (2018) Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res., 28, 1747–1756.
Moore,L. et al. (2020) The mutational landscape of normal human endometrial epithelium. Nature, 580, 640–646.
Omichessan,H. et al. (2019) Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. PLoS One, 14, e0221235.
Robinson,P.S. et al. (2020) Elevated Somatic Mutation Burdens in Normal Human Cells due to Defective DNA Polymerases. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. https://doi.org/10.1101/2020.06.23.167668

Rosen,G. (2007) Comparison of autoregressive measures for DNA sequence similarity. In: IEEE International Workshop on Genomic Signal Processing and Statistics, pp. 1–4, https://doi.org/10.1109/GENSIPS.2007.4365814.
Rosenthal,R. et al. (2016) DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol., 17, 31.
Wang,S. et al. (2020) Copy Number Signature Analyses in Prostate Cancer Reveal Distinct Etiologies and Clinical Outcomes. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. https://doi.org/10.1101/2020.04.27.20082404
Weinstein,J.N.; Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer Analysis project. Nat. Genet., 45, 1113–1120.
Yoshida,K. et al. (2020) Tobacco smoking and somatic mutations in human bronchial epithelium. Nature, 578, 266–272.