MetaMutationalSigs

Abstract:

Motivation:

Mutations that result in cancers are caused by several mutational processes; mutational signature analysis can identify the effect of mutational processes in cancers. There are several packages available now for mutational signature analysis and they all use different approaches and give nontrivially different results. Because of the differences in their results, it is important for researchers to survey the available tools and select a few that best suit their research application. There is a need for a software that can aggregate results from different packages and present them in a user-friendly way to facilitate effective comparison.

Results:

We developed this package MetaMutationalSigs to facilitate comprehensive mutational signature analysis by implementing a wrapper for different packages and providing a standard format for their outputs so that they can be effectively compared. We have also standardized the input formats accepted by various packages to ease interoperability. We also provide standard visualizations for the results of all packages to ensure easy analysis. MetaMutationalSigs is easy to install and use through Docker, a package manager that automates the dependencies.

Introduction:

Cancers acquire several 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 tumor mutation signature analysis. It is becoming increasingly evident that these tumor mutation 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].

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 toTCGA [7] data to identify the first iteration of 30 mutational signatures termed COSMIC signatures [5], which came be used as the de facto reference for signature refitting. They then expanded their analysis to more data from PCAWG [7] resulting in multiple signatures using different contexts, such as double base and indels, this set of reference signatures is called SBS signatures for this package [8].

The mutational signature analysis workflow involves multiple steps that require different amounts of time and processing power. Briefly, the workflow starts from BAM files that are aligned to a reference genome and then proceeds to the variant calling step which outputs the VCF files. These steps are usually very resource-intensive and thus do not allow for much experimentation, the downstream steps of variant filtering and annotation are much faster The final step is the mutational signature analysis, which is the least resource-intensive and thus allows users to try multiple methods.

Here, to facilitate comprehensive mutational signature analysis we developed the package, MetaMutationalSigs. We create a wrapper for multiple packages [10,11,12,13] typically used for mutation signature analysis, and provide a standard format for their outputs so that they can be effectively compared. We have also standardized the input formats accepted by various packages to ease interoperability. We also implement standard visualizations for the results of all 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 recreate the observed mutational pattern in the sample (the frequencies of 96 types of mutations) using the linear combinations of known signatures (COSMIC 30, SBS, ID, etc.), these methods work quite well on small sample sizes and single samples and are widely used as such [6]. Signature extraction methods try to find new signatures from a given dataset using a set of samples, the newly extracted signatures are then compared with known reference signatures and the novel signature is assigned to a known signature if their cosine similarity exceeds a set threshold [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 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 CSV format ready for further analysis and visualization using external packages along with visualizations of the signature exposures as described in table 1.

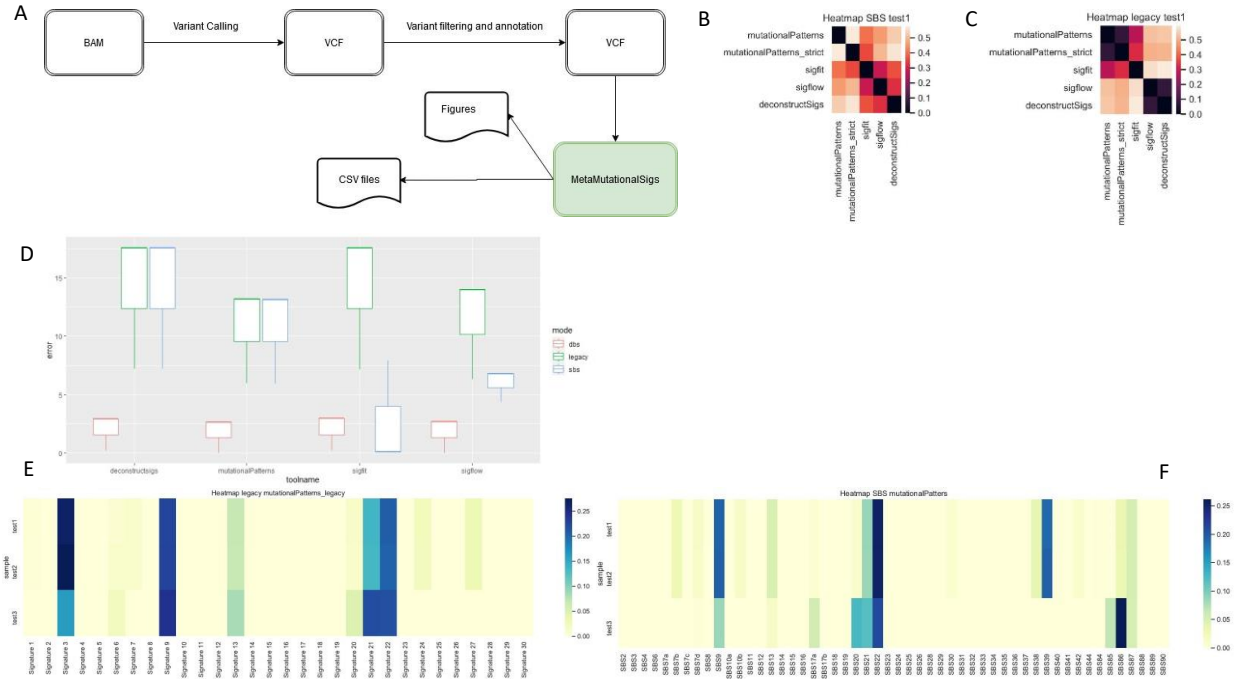
We use the RMSE of the reconstruction error as our performance metric for comparison of these packages.

Table 1.

File Name	Format	Description
Heatmap_exposures_all_sigs_legacy.pdf	pdf	Exposures for all COSMIC 30 signatures.
Heatmap_exposures_all_sigs_SBS.pdf	pdf	Exposures for all COSMIC SBS signatures.
Heatmap_legacy.pdf	pdf	Heatmap for difference between the predicted exposures by different tools. One for each sample.
Heatmap_SBS.pdf	pdf	Heatmap for difference between the predicted exposures by different tools. One for each sample.
legacy_pie_charts.html	html	Interactive pie charts of 30 legacy signature exposures, per sample and for each tool.
sbs_pie_charts.html	html	Interactive pie charts of SBS signature exposures, per sample and for each tool.
legacy_rmse_bar_plot.png	png	Reconstruction error using 30 legacy COSMIC signatures for each tool.
sbs_rmse_bar_plot.png	png	Reconstruction error using COSMIC SBS signatures for each tool.
toolname_results\legacy_sample_error.csv	csv	Data used to create the bar plot.
toolname_results\legacy_sample_exposure.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_exposure.csv	csv	data used to create heatmap and pie chart.

Figure 1



A) Workflow for mutational signature analysis. Our tool MetaMutationalSigs is at the final level of analysis. B) Heatmap of Euclidean distance between the predicted exposures of COSMIC legacy signatures by different tools for a sample. C) Heatmap of Euclidean distance between the predicted exposures of COSMIC SBS signatures by different tools for a sample. D) RMSE of the reconstruction error using for each tool and reference signatures, lower is better. Prevalence of SBS:ID:DBS::100:10:1. E) Heatmap of COSMIC legacy signature exposures, one row per sample. F) Heatmap of COSMIC SBS signature exposures, one row per sample.

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 analysis difficult and hard to reproduce. Our package provides an easy way of performing these setup related tasks so one can focus more 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

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] 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] 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; 322859

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. *Nat Commun* 10, 2969 (2019)