The Mutational Signature Comprehensive Analysis Toolkit (musicatk) for the discovery, prediction, and exploration of mutational signatures

Aaron Chevalier[1,2], Shiyi Yang[1], Masanao Yajima[3], Josh Campbell[1]

1. Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA.
2. Bioinformatics Program, Boston University, Boston, MA.
3. Department of Mathematics & Statistics, Boston University, Boston, MA.

Abstract
Mutational signatures are patterns of somatic alterations in the genome caused by carcinogenic exposures or aberrant cellular processes. To provide a comprehensive workflow for preprocessing, analysis, and visualization of mutational signatures we created the Mutational Signature Comprehensive Analysis Toolkit (*musicatk*) package. *musicatk* enables users to select different schemas for counting mutation types and easily combine count tables from different schemas. Multiple distinct methods are available to deconvolute signatures and exposures or to predict exposure in individual samples for a pre-existing set of signatures. An extensive set of visualizations are provided to facilitate exploratory analysis of signatures and exposures. These include a comparison of discovered signatures to those in the COSMIC database, using UMAP to embed tumors in two dimensions, and plotting of exposure distributions across user-defined annotations. Overall, *musicatk* will enable users to gain novel insights into the patterns of mutational signature observed in cancer.

Introduction
Somatic mutations to the genome can be caused by exposure to environmental carcinogens or aberrant cellular processes[1,2]. A “mutational signature” is a specific pattern of mutation types caused by a particular mutational process. The set of mutations observed in a tumor genome can be the result of multiple mutational processes active during the course of tumor development. Therefore, deconvolution is needed to determine which processes have occurred across a group of tumor genomes. Recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium characterized a large cohort of whole-exome and whole-genome samples with single-base-substitution (SBS), doublet-base-substitution (DBS), and small insertion-and-deletion (INDEL) mutational schemas using NMF-based methods[3]. While some software packages have been previously developed to perform mutational signature inference, they cannot quantify the latest set of mutation schema from COSMIC, lack functionality for comprehensive exploratory analysis, or have limited functionality for predicting exposures to pre-defined signatures in new samples[3–9] **(Supplemental File 1)**. The *musicatk* package provides functionality to streamline the steps of mutational inference and has several extra features to enhance exploratory analysis. The major steps of mutational signature inference are 1) importing the variants, 2) building and combining count tables based on different types of mutation schema, 3) performing discovery and/or prediction of signatures and exposures, and 4) using visualization for exploratory analysis of the results. We provide an overview of this functionality and present an exploratory analysis of tumors from The Cancer Genome Atlas (TCGA).

Methods
Importing and processing of mutations

The *musicatk* package has functions to read mutations from various input formats including several file types and R objects. Mutation Annotation Formats (MAFs) are read from files or from the R class “MAF” created by the maftools package. Variant Call Formats (VCFs) can be read from files or from the R class “CollapsedVCF” created by the VariantAnnotation package. Additionally, variant information stored in a data.frame or data.table can be used as input **(Figure 1A)**. To streamline the processing of variants, mutation profiles from multiple tumors in different format can be automatically read and combined into a musica object. Variants are stored in an efficient table within the musica object to allow for efficient storage and retrieval of a large number of mutations **(Supplementary Figure 1)**.

We also include functions that automatically parse different types of mutation motifs within each tumor and create count tables that are used in downstream analysis. Tables that can be calculated by *musicatk* include SBS into 96 motifs, SBS with transcription strand orientation into 192 motifs, SBS with replication strand orientation into 192 motifs, DBS into 78 motifs, and INDELs into 83 motifs **(Figure 1B)**. User-defined functions that create custom count tables can also be added to the object. Importantly, multiple tables can be concatenated to create new combined tables. For example, users have the flexibility to utilize the most common SBS-96 motif table or they can combine the SBS-96 motif, DBS-78 motif, and INDEL-83 tables and perform downstream analyses as a more generalized form of the PCAWG Composite signatures[3].

Discovery of mutational signatures

Deconvolution is the process of decomposing a matrix of mutation counts per tumor into a matrix of mutational signatures and mutational exposures. We support both Latent Dirichlet Allocation (LDA)[10] and Non-Negative Matrix Factorization (NMF)[11] algorithms. Both algorithms can be applied to any count table. One challenging aspect of mutational signature discover is the process of determining the appropriate number of signature (*K*).

We might want to explore K in several circumstances. For instance, in mixed datasets which combine multiple tumor types we may wish to discover different signatures and different K values in different tumor types. Even in single sets we may wish to find the optimal number of signatures. To facilitate these types of discovery, *musicatk* provides a wrapper that allows users to explore with multiple choices of *K* and then compare the results with metrics such as reconstruction error (NMF and LDA), log-likelihood (LDA), or perplexity (LDA).

Prediction of mutational signatures

Prediction of exposures for existing signatures can be performed on any count table given that the mutation motif schema is the same. We include wrappers for tools such as deconstructSigs and decompTumor2Sig[4,8]. We also implement a Bayesian algorithm based on LDA where exposures are estimated based on a fixed set of signatures **(Supplemental File 2)**. To allow for prediction using previously defined signatures from the COSMIC database[3], we include objects for COSMIC V2 signatures (SBS-96 motif schema) and COSMIC V3 signatures (SBS-96, DBS-78, and INDEL-83 motif schemas). One challenge in the prediction of existing signatures in new tumors is that not all of the signatures will be present in the new dataset. including non-active signatures may cause additional noise in the estimates for signatures that are present within the dataset. We implemented a two-step procedure that chooses active signatures for a dataset based on several filters. In the first step, exposures are estimated using all signatures. In the second step, active signatures are chosen if they are dominant in some samples, or active above a customizable threshold in most samples. Signatures that are present in moderate

levels across many tumors or highly present in a small number of tumors are more likely to be active in the dataset[3]. As with signature discovery, users can perform the 2-step prediction within subgroups of tumors supplying a categorical annotation such as tumor type.

Visualization

Visualization is an important part of the workflow of generating and inspecting mutational signatures and tumor exposures. *musicatk* has standard plots that show the probability of each motifs in each signature or the composition of exposures in each sample **(Figure 2A)**. We can sort exposures in multiple ways, either by counts, by sample name, or by one or more signatures, so that we can see clustering of exposures. We can also zoom in on the top samples in our sorting to see the breakdown of a subset of samples. Exposures can be grouped by a sample annotation such as tumor type **(Figure 2B, 2C)**. This allows grouping and comparison either by signature or by annotation type. As with the exposure plots, visualization can show raw or normalized counts, and samples can be sorted by count, sample name, or signatures name(s). Functionality is provided for correlating sets of signatures between different signature result objects. This can be used to easily perform comparisons to COSMIC V2 and V3 signatures, which are including as objects within the package **(Figure 2D)**. To view relationships between tumors in two dimensions, the Uniform Manifold Approximation and Projections (UMAPs)[12] algorithm can be generated from normalized signature exposures **(Figure 2B, 2C)**. The UMAP can be colored by annotations (e.g. Tumor Type) or colored by signature exposure.

Results

We applied the LDA-based method to predict COSMIC V3 SBS signatures in a Pan-Cancer dataset from TCGA. 37 signatures were found to be active in at least one tumor type. We generated a UMAP plot in order to explore the patterns of signatures across tumors **(Figure 3)**. Some signatures were present in nearly half of samples (SBS1, SBS5, SBS40), some in a few tumor types (SBS25, SBS3, SBS32, SBS4, SBS46), some in single tumor types (SBS7a, SBS7b, SBS9), some in subsets of multiple tumor types (SBS13, SBS15, SBS21, SBS2, SBS19, SBS18, SBS16, SBS19, SBS30, SBS37, SBS39, SBS44, SBS46, SBS8, SBS6), and a few signatures are present in only a few samples (SBS9, SBS10a, SBS10b, SBS12, SBS17, SBS22, SBS26, SBS38, SBS36, SBS35, SBS33, SBS42, SBS59).

APOBEC signatures (SBS2 and SBS13) were present in a subset of tumors BRCA, CESC, BLCA, HNSC and UCEC, and distantly in a subset of LUAD and LUSC tumors. In general, both APOBEC signatures were present in the same samples. The only exception was that SBS2 was present in a subset of SKCM tumors. Interestingly, in addition to the UCEC samples clustered around the previously mentioned APOBEC-high samples, UCEC samples are split into three additional clusters. The top samples are clustered because of their exposure to the unknown origin SBS39 signature. The middle cluster includes samples from a few other tumor types including COAD because of their POLE (SBS10 a/b) exposure. The bottom cluster has high levels of a defective DNA mismatch repair signature (SBS44). Two other MMR signatures, SBS15 and SBS21, were present in COAD and STAD clusters, respectively, separated from their main clusters but near each other, despite disjoint exposure status. Their proximity in the UMAP is due to their co-exposure to SBS6, another MMR signature.

In conclusion, the *musicatk* package provides a comprehensive set of preprocessing utilities, access to several discover and prediction tools, and functions for downstream analysis of the patterns of mutational signatures in a cohort of tumors.

Figure Legends

Figure 1. Overview of workflow for mutational signature analysis. A) Workflow allows for loading and combining data from multiple sources, discovery and prediction of signatures and exposures followed by additional downstream analysis **B)** Table creation allows for flexibility and customization in prediction and deconvolution of signatures and exposures.

Figure 2. Comprehensive visualization of mutational signature results. A) This is the standard exposure and signature plots widely used. But we can see they don't give us a lot of information B) Plotting by annotation we can see Signature 3 uniquely has no tumor type representing high exposure. UMAP visualization shows a melanoma cluster and a strange additional cluster C) This strange UMAP cluster represents high signature 3 exposure across tumor type. Plotting the top proportion of samples in each annotation sorted by signature 3 exposure, we can see COAD and UCEC both contain small numbers of samples with very high exposure D) To recap, we've discovered a signature that makes up a large proportion of counts for a small subset of samples, especially colon and uterine samples. We could guess COSMIC Signature 10 (POLE) based on these facts, and we are able to confirm it by using the *musicatk* built-in COSMIC signature comparison tool.

Figure 3. UMAPs of TCGA sample exposure to COSMIC v3 signatures. A) UMAP colored by tumor type with cluster labels. **B)** Set of four UMAPs colored by exposure to SBS2, SBS6, SBS10a, SBS44 respectively with blue being low exposure and red being high exposure.

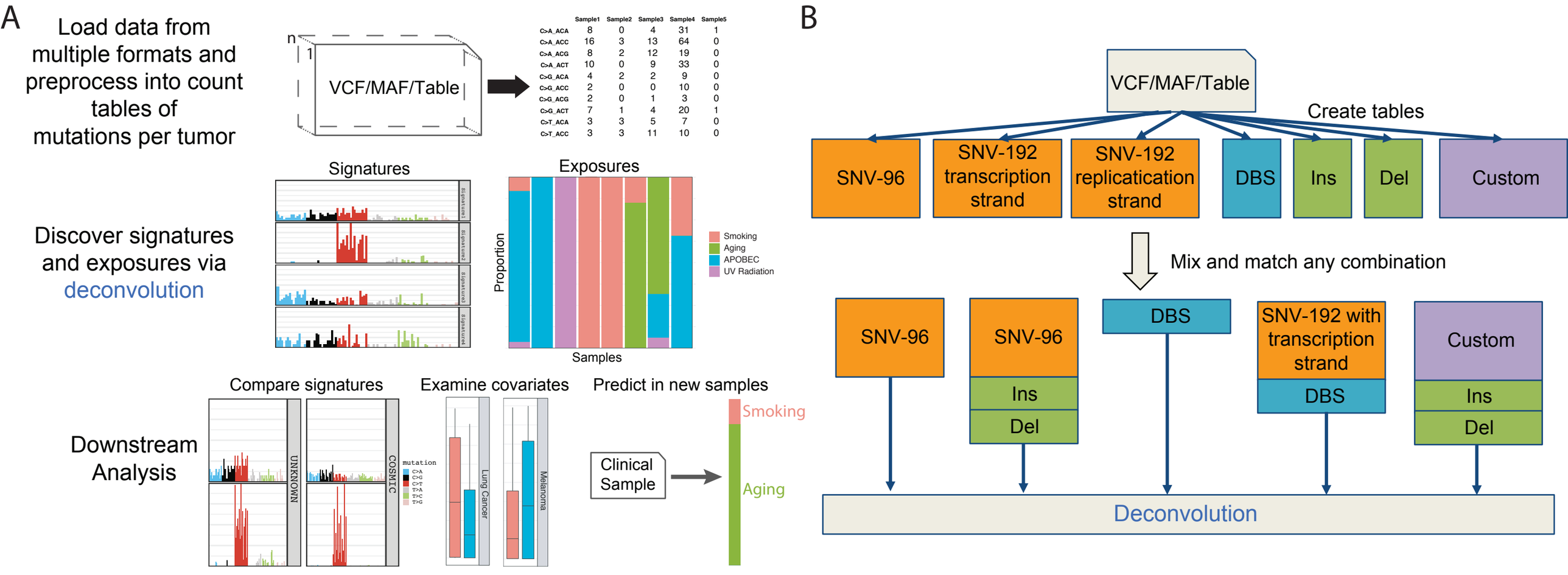
Acknowledgements

This work was funded by the Informatics Technology for Cancer Research (ITCR) R21 CA226188-01A1 (J.D.C. and M.Y.).

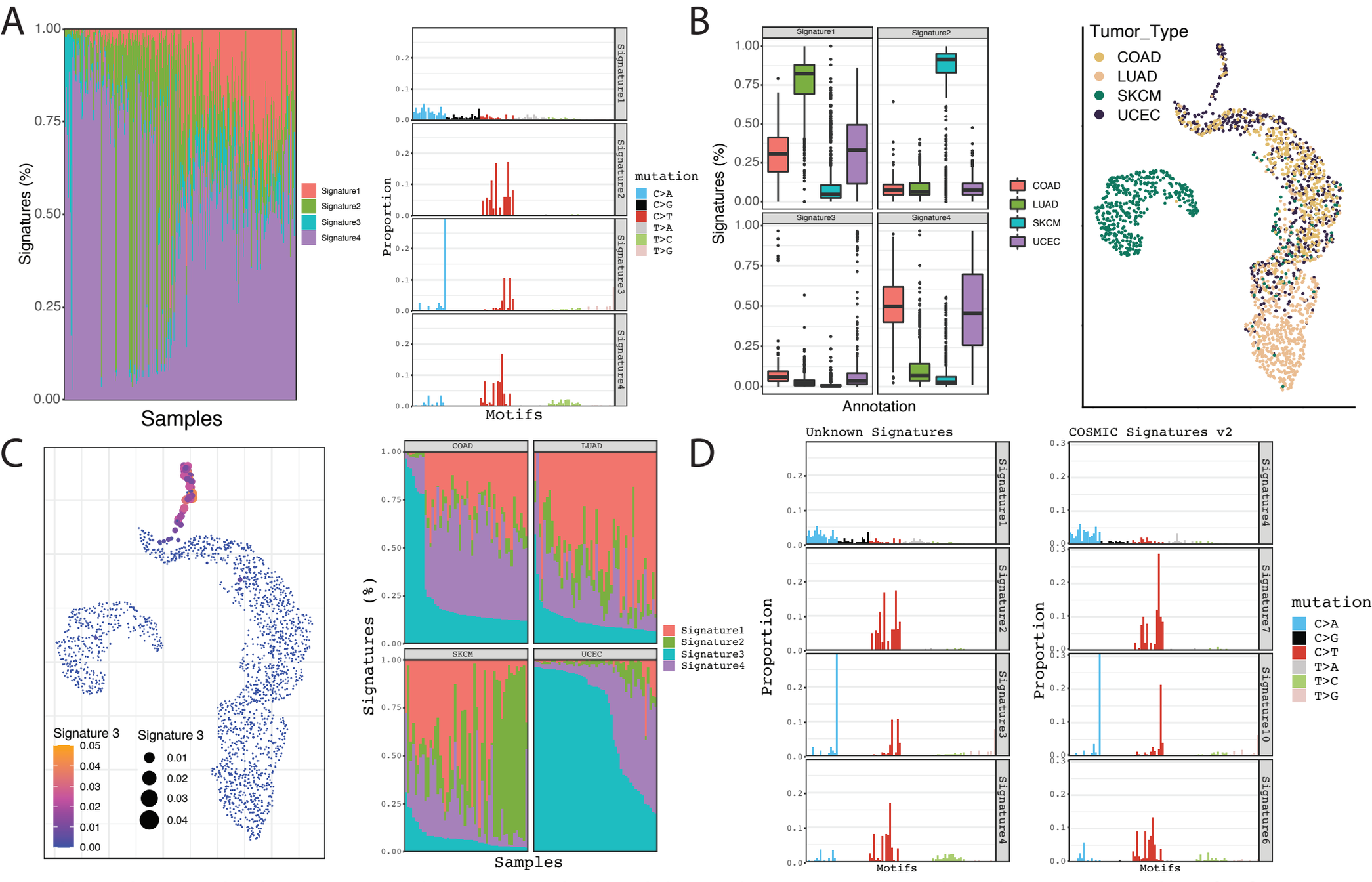
References

1. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415 (2013).
3. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
4. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
5. Teresa Przytycka Research Page. Available at: https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#signatureestimation. (Accessed: 16th November 2020)
6. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
7. YAPSA: Yet Another Package for Signature Analysis version 1.16.0 from Bioconductor. Available at: https://rdrr.io/bioc/YAPSA/. (Accessed: 16th November 2020)
8. Krüger, S. & Piro, R. M. DecompTumor2Sig: Identification of mutational signatures active in individual tumors. *BMC Bioinformatics* **20**, 152 (2019).
9. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures.

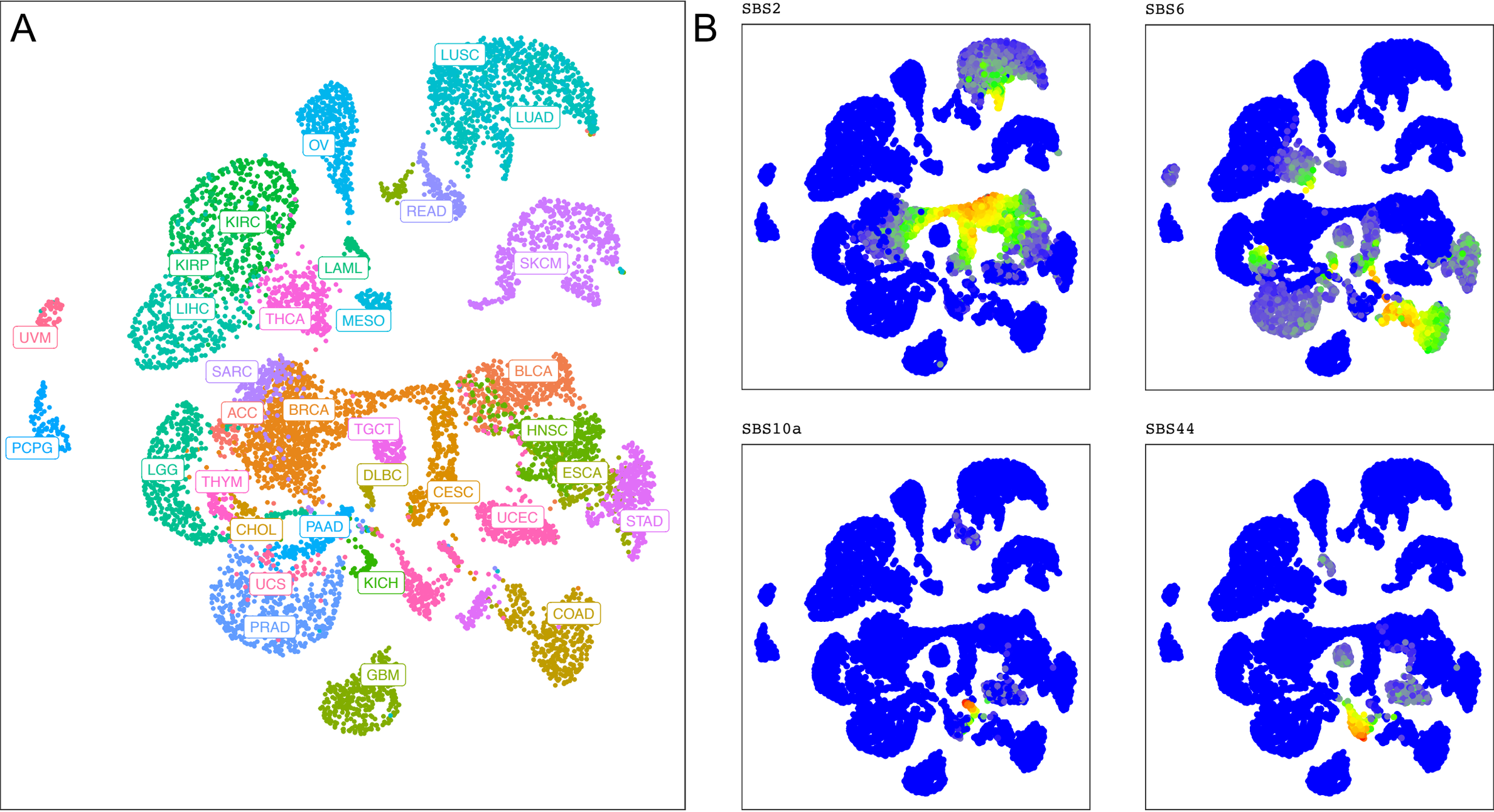
bioRxiv 372896 (2018). doi:10.1101/372896
10. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
11. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
12. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (2018).



A) Workflow allows for loading and combing data from multiple sources, discovery and prediction of signatures and exposures followed by additional downstream analysis B) Table creation allows for flexibility and customization in deconvolution of signatures/exposures



A) This is the standard exposure and signature plots widely used. But we can see they don't give us a lot of information. B) Plotting by annotation we can see Signature 3 uniquely has no tumor type representing high exposure. UMAP visualization shows a melanoma cluster and a strange additional cluster. C) This strange UMAP cluster represents high signature 3 exposure across tumor types. Plotting the top proportion samples in each annotation sorted by signature 3 exposure we can see COAD and UCEC both contain small number of samples with very high exposure. D) To recap, we've discovered a signature that makes up a large proportion of counts a small subset of samples, especially colon and uterine samples. We could guess COSMIC Signature 10 (POLE) based on these facts, but using our built-in COSMIC signature comparison we a able to comfirm it.



A) UMAP with samples colored by tumor type. B) Set of four UMAPs with samples colored by exposure to SBS2, SBS6, SBS10a, SBS44 respectively with blue being low exposure and red being high exposure.