A new approach to the challenging problem of mutational signature attribution

Nanhai Jiang[1,2], Yang Wu[1,2], Steven G. Rozen[1,2,3*]

[1] Centre for Computational Biology, Duke-NUS Medical School, Singapore 169857, Singapore

[2] Programme in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore 169857, Singapore

[3] Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, 27710, United States

*To whom correspondence may be addressed:

Steven G. Rozen
Department of Biostatistics and Bioinformatics, Duke University School of Medicine
2424 Erwin Road, Suite 11092
Hock Plaza Box 2721
Durham, North Carolina, 27710, United States
steverozen@pm.me

ABSTRACT

Mutational signatures are characteristic patterns of mutations caused by endogenous mutational processes or by exogenous mutational exposures. Much research has focused on the problem of inferring mutational signatures as latent variables in somatic mutation data from multiple tumors. However, the problem of determining which signatures are present in a given sample and how many mutations each signature is responsible for has received negligible attention. In particular, there has been little systematic benchmarking of various approaches to this problem. This problem is referred to as “signature attribution” in a single sample. We show that this is a challenging problem, because there are often many combinations of signatures that can reconstruct the mutational spectrum of a given sample reasonably well. We benchmarked the accuracy of five approaches to signature attribution, including a new approach we call Presence Attribute Signature Activity (PASA), on large synthetic data sets. These data sets recapitulated the single-base, insertion-deletion, and doublet-base mutational signature repertoires of 9 cancer types. For single-base substitution mutations, PASA outperformed other approaches on all the cancer types combined. Interestingly, however, the ranking of approaches varied by cancer type. For doublet-base substitutions and small insertions and deletions, the ranking of approaches was more stable, with PASA outperforming other approaches in most, but not all of the nine cancer types. For all mutation types, the ranking of approaches varied by cancer type, and no approach achieved both high precision and recall. We believe these observations reflect the inherent challenges in signature attribution.

INTRODUCTION

Different mutational processes can generate characteristic patterns of mutations; these are termed as mutational signatures [1]. The causes of mutations can be endogenous, e.g. deamination of genomic 5-methyl cytosines [2] or defective polymerase epsilon proofreading [3], or exogenous, e.g. exposure to aristolochic acid [4,5] or tobacco smoke [6]. Mutational signatures can serve as biomarkers for mutagenic exposures that increase cancer risk [5,7–9]. More generally, mutational signature analysis can shed light on cancer causes, prognosis, and prevention and can provide insights into the mechanisms of DNA damage and repair [10–13].

Much prior research has concentrated on discovering mutational signatures, which is not the main topic of this manuscript. One common approach to discovering signatures is to expose cultured cells to a suspected mutagen, and then sequence the exposed culture [14–16]. In addition, machine learning methods can discover mutational signatures in large databases of somatic mutations from tumors [1,17]. This process is often referred to as *signature extraction*. The approach depends on the model that a mutational spectrum can be explained as a linear combination of mutations generated by mutational signatures (Figure 1). The number of mutations due to a particular signature is referred to as the signature's *activity*. Signature extraction discovers mutational signatures as latent variables that can parsimoniously explain sets of mutational spectra [18–20]. Several benchmarking studies have systematically examined the accuracy of different approaches to signature extraction [20–23]. To-date, experimental methods and in silico signature extraction together have identified > 100 reference mutational signatures [24].

In addition to the discovery of mutational signatures, another important task is to estimate the presence of existing mutational signatures and their activities in a

mutational spectrum, a task that commonly called *signature attribution*. This is often a challenging task. Absent critical review of output, signature attribution software can generate results that are useless for understanding the underlying biology of mutagenesis and its consequences. For example, one study reported that nearly 50% of lung tumors in never smokers (mostly adenocarcinomas) have the SBS3 signature (Fig. 4 in reference [25]). SBS3 is the result of deficient homologous-recombination-based DNA damage repair, and the same study (contradictorily) reported that HRDetect [12] detected homologous recombination deficiency in only 16% of the tumors (Extended Data Fig. 8a in reference [25]). Furthermore, Alexandrov et al. [1] detected SBS3 in only 8% of lung adenocarcinomas. SBS3 is especially prone to this kind of error, which is also shown in Extended Data Fig. 3a in reference [26], in which high proportions of tumors of almost all cancer types have SBS3, an implausible result in light of the actual prevalences of homologous recombination deficiency across cancer types. A related issue is signature attribution software often includes small activities of signatures due to implausible exposures. For example, one study reported the signature of UV exposure not only in cells from skin melanomas and in skin fibroblasts, but also in cells from every tissue, including kidney, liver and skeletal muscle (Fig. 3b in reference [27]). Furthermore, beyond these sorts of implausible results, challenges remain. We show below that it is often possible to reconstruct the mutational spectrum of a sample using dozens or more different combinations of signatures, all of which yield reasonably good reconstructions.

Despite the importance of mutational signature attribution, there has been little benchmarking of software for this task [21,28,29]. In addition, previous studies of which we are aware studied only single-base substitution (SBS) mutational

signatures and neglected doublet-base substitution (DBS) signatures and insertion-deletion (ID) signatures.

Here, we present PASA (Presence-based Attribution of Signature Activity), a new, statistically-grounded algorithm for mutational signature attribution for three types of mutational signatures – SBS, DBS, and ID – in a sample. We assessed the accuracy of PASA and four other prominent approaches: SigProfilerAssignment (SigPro, "signature assignment" is another term for signature attribution) [28], Signature Fit Multi-Step (FitMS) in SignatureToolsLib [17], MutationalPatterns (MP) [30] and Mutational Signature Attribution (MSA) [31] using synthetic SBS, DBS and ID data from 2,700 tumors representing 9 cancer types.

MATERIALS AND METHODS

Preliminary definitions

The *mutational spectrum* of one sample (tumor) is a 1-column matrix, $D \in N_{\geq 0}^{k \times 1} = [d_1\ d_2 \ldots d_k]^T$, where k is the number of mutation types and each d_k is the number of mutations of that type. For example, for the common case of single-base substitutions in the context of preceding and following bases, the mutation types are ACA→AAA, ACC→AAC, …, CCA→CAA, …, TTT→TGT. By convention, a mutated base is represented by the pyrimidine of the Watson-Crick base pair, and therefore there are six substitution subtypes: C→A, C→G, C→T, T→A, T→C, and T→G. There are altogether 96 types of single-base substitutions in the context of preceding and following bases (6 types of substitutions x 4 types of preceding base x 4 types of following base). The term "SBS signature" is usually understood to mean the signature of single-base substitutions in the context of preceding and following

bases. The classification of doublet-base substitutions (DBS) is detailed at https://www.synapse.org/#!Synapse:syn11801895. For small insertions and deletions (ID), the classification is described at https://www.synapse.org/#!Synapse:syn11801742.

A *mutational signature* is a multinomial probability vector $\vec{h} \in R_{\geq 0}^{k}$, with $\sum_k h = 1$, which represents the characteristic proportions of the mutation types that are generated by one mutational process. Each element inside $\vec{h}$ is the probability of observing one mutation of that particular mutation type. In this model, each mutational process generates mutations of different mutation types by sampling from the multinomial distribution that is the process's signature. Since in general multiple mutational processes generate mutations in a tumor, in this model, the spectrum, D, is the sum of mutations in each mutation type generated by different processes.

Given a matrix $H \in R_{\geq 0}^{k \times g} = [\overrightarrow{h_1}\ \overrightarrow{h_2} \ldots \overrightarrow{h_g}]$ of g mutational signatures, the task of *signature attribution* is to find a non-negative activity matrix $A \in R_{\geq 0}^{g \times 1} = [a_1\ a_2\ \ldots\ a_g]^T$ that approximately reconstructs the original tumor spectrum using input signatures H, i.e. such that $D \approx H \times A$. Most approaches seek to minimize the L2 norm of $D - (H \times A)$.

Exhaustive enumeration of stomach cancer attributions and their accuracy

We selected 14 mutational signatures that were common in stomach adenocarcinoma (Stomach-AdenoCA in reference [1]). For each stomach adenocarcinoma in the Pan Cancer Analysis of Whole Genomes mutation data analyzed in [1], for the non-empty power set of these signatures, we used the function mSigTools::optimize_exposure_QP to select an attribution that minimizes

the Euclidean distance to the given spectrum. We excluded attributions with a contribution of < 0.5 mutation from any signature as redundant with attributions lacking that signature altogether. Code and attributions with cosine similarity > 0.98 are available at https://github.com/Rozen-Lab/sig_attribution_paper_code/common_code/figure_2/.

PASA algorithm

The PASA algorithm for signature attribution takes as input a mutational spectrum, D, to be reconstructed from a set of g possible signatures represented by a matrix H, as above. Also as above, the algorithm returns a column matrix, A, of signature activities that can reconstruct D. For signature attribution in a given cancer, H normally consists of the set of signatures previously observed in that cancer type.

The PASA algorithm proceeds in two steps.

Step 1: Signature presence tests

In Step 1, PASA conducts a signature presence test for each signature $\overrightarrow{h_i} \in H$ to exclude signatures that are not statistically likely to be present in the tumor sample. This is a likelihood ratio test, in which the null hypothesis is that reconstruction of the spectrum without $\overrightarrow{h_i}$ (a reduced model) is as good as the reconstruction with all input signatures (full model). The test first separately searches for activities that optimize the likelihood P(D | A), where A is either $\left[a_1 \ \dots \ a_{i-1} \ a_{i+1} \dots \ a_g\right]^T$ (reduced model) or $\left[a_1 \ \dots \ a_g\right]^T$ (full model). The search is implemented by the R nloptr package [32] (https://cran.r-project.org/package=nloptr, see function OptimizeExposure in the mSigAct package). The likelihood ratio test then depends on the test statistic $\lambda = -2\left(\log l_{reduced} - \log l_{full}\right)$, where $\log l_{reduced}$ and $\log l_{full}$ are the log likelihoods of

the reduced and full model respectively. λ follows a chi-squared distribution with 1 degree of freedom from which a p-value can be determined. If p exceeds the significance level (default 0.05), the null hypothesis is not rejected and $\overrightarrow{h_\iota}$ is removed from the signature set that will be used in Step 2. The output of Step 1 is the set of signatures, $V = [\,\overrightarrow{v_1}\ \overrightarrow{v_2} \ldots \overrightarrow{v_t}\,],\ t \leq g$, that survived the signature presence test.

Step 2: Forward search

See Algorithm STEP2 of PASA.

Algorithm: STEP 2 of PASA Forward search for a minimal set of mutational signatures to reconstruct a mutational spectrum

INPUT $D \in N_{\geq 0}^{k \times 1}$ // The mutational spectrum of one sample as a 1-column matrix
$V \in R_{\geq 0}^{k \times t}$ // A matrix of signatures from which to reconstruct D, from Step 1
α // Significance level for comparing reconstruction of a spectrum
// using a set of signatures versus using a proper subset of that set

OUTPUT $A \in R_{>0}^{u \times 1}$ // Activities of $u \leq t$ signatures from V that can plausibly reconstruct D,
// in the sense that the likelihood ratio test for reconstructing D from A
// versus reconstructing D from V is not significant at significance level α

```
loglh_max , A_all = OptimizeActivity (D, V)                    // See definition below
X = V
W = [ ]                          // We will be moving signatures from X to W in a greedy search
                                 // to accumulate signatures in W for reconstructing D
FOR step in 1 to t − 1
    pValues = [ ]
    FOR j in 1 to numCols(X)
        Y = [W; X[ ,j]]                          // We will test a matrix of signatures, Y,
                                                 // consisting of W and the jth signature of X
        loglh , A = OptimizeActivity (D, Y )
        pValues[j] = LRT(loglh_max, loglh, df = t − step)   // Perform a likelihood ratio test
                                                 // comparing reconstruction of D with V
                                                 // versus reconstruction with Y = [W; X[ ,j]]
    ENDFOR
    index = whichmax(pValues)                    // Find the index of the signature that
                                                 // improves the reconstruction the most
    W = [W; X[ ,index]]                  // Add the signature with the largest p value to W
    X = X[ ,−index]                      // Remove the signature with the largest p value from X
    IF pValues[index] > α                // Reconstruction with V is not significantly
                                         // better than reconstruction with W
        loglh_opt , A_opt = OptimizeActivity (D, W)   // Recalculate the activity matrix for
                                                      // reconstruction with the signatures in W
        RETURN A_opt
    ENDIF
ENDFOR
RETURN A_all
```

```
FUNCTION OptimizeActivity (D, V)
    loglh , A = NLopt(D, V)      // Use non-linear numerical optimization to find activities, A,
                                 // that maximize loglh = log(P(D|A,V)); A is a 1-column matrix
    RETURN loglh , A
ENDFUNCTION OptimizeActivity
```

Running other approaches to signature attribution

The code for running other approaches and raw output are available at https://github.com/Rozen-Lab/sig_attribution_paper_code/analysis/. Importantly, for every approach and every cancer type, we allowed attribution with only the set of signatures previously observed in that cancer type as reported in reference [1]. Table S1 lists software versions and arguments.

Generation of synthetic mutation data

We used COSMIC [24] v3.2 reference mutational signatures and the signature activities estimated by Alexandrov et al. [1] to generate synthetic SBS, DBS and ID mutational spectra. Detailed methods are described in our previous publication [20]. Briefly, for each of the SBS, DBS, and ID mutation types, we generated 100 synthetic spectra for each of nine cancer types. To generate one synthetic spectrum of a particular cancer type, the code first selects the signatures present in the spectrum and the ground-truth activities of each signature as random draws from the distributions of these estimated from [1]. Then, the numbers of each mutation type are selected from a negative binomial distribution that is centered on the overall number of mutations due to the signature times the proportion of mutations of that mutation type in the signature. Code to generate the synthetic mutational spectra and the spectra themselves are at https://github.com/Rozen-Lab/sig_attribution_paper_code/synthetic_data/.

Definition of evaluation measures

For a given synthetic spectrum with given ground truth activities, let P be the number of signatures which have activity > 0. Let TP (true positive) be the number of signatures with activity > 0 that also have estimated activities > 0. Let FP (false positive) be the number of signatures with 0 activity, but that have estimated activity > 0

The evaluation measures for attribution of signatures for a synthetic spectrum of a given cancer type are:

Precision = $TP/(TP + FP)$

Recall = $TP\ /\ P$

Scaled Manhattan distance = $\sum_i |X_i - Y_i|/M$, where the X_i are the ground truth activities of all signatures known to occur in the given cancer type, the Y_i are the estimated activities, and M is the number of mutations in the sample

Combined = (1 – Scaled Manhattan distance) + Precision + Recall

RESULTS

Hundreds of reasonable signature attribution solutions present a challenge

As noted in the Introduction, signature attribution appears to be inherently challenging, and we show here that one reason for this is that often there are many different reasonable attributions of a given original spectrum. We take as examples 49 stomach cancer spectra (from [1]) that, using 14 SBS signatures common in this type of tumor, have attributions providing cosine similarities to the original spectrum

> 0.98. These spectra have a median of 257 distinct attributions with similarity above this similarity threshold (Figure 2A). For one stomach cancer, Figure 2B-F and Table S2 show 4 example attributions out of 189 attributions that exceed this similarity threshold. The relevant code and, for all 49 stomach cancers, the attributions that have cosine similarity > 0.98, are available at https://github.com/Rozen-Lab/sig_attribution_paper_code/common_code/figure_2/.

Perhaps one could simply take the attribution with the highest similarity to a given spectrum as the most likely true attribution for that spectrum. We explored this question by examining at all possible attributions for each of 95 synthetic stomach-adenocarcinoma spectra. For 93 out of these 95 spectra, the ground-truth attribution was inferior to the best alternative attribution – inferior in the sense of having lower cosine similarity to the given spectrum (Table S3). Indeed, there was a median of 15 alternative attributions that were superior to the ground truth attribution. We emphasize that the signatures available for attribution were only those actually found in this type of cancer.

Alternative attributions superior to the ground-truth attribution can exist because, although the number of mutations due to a given signature was known for each synthetic spectrum, the distribution of counts due to that signature across different mutation types (e.g. ACA → AAA, …, CTG → CAG, …, TTT → TGT), was sampled randomly. This was designed to simulate a model in which a mutational process generates a certain number of mutations according to a fixed multinomial distribution across mutation types, but the count of mutations of each mutation type varies due to random sampling.

Accuracy on SBS (single-base substitution) mutational spectra

We assessed the accuracy of SBS mutational signature attributions produced by PASA and four other approaches, SigProfilerAssignment (SigPro) [28], Signature Fit Multi-Step (FitMS) in SignatureToolsLib [17], MutationalPatterns (MP) [30] and Mutational Signature Attribution (MSA) [31] (Figure 3, Table S4). The Combined Score of PASA across all 900 synthetic SBS spectra was significantly higher than the scores of SigPro or FitMS (median: 2.69 versus 2.63 and 2.59, $p < 2.7 \times 10^{-8}$ and $p < 1.2 \times 10^{-12}$, respectively, 2-sided Wilcoxon rank-sum tests, significant at a Bonferroni-corrected alpha of 0.025). The Combined Scores of the remaining approaches were lower still.

Of note, the rankings for the signature attribution approaches varied across cancer type (Figure 3E). Thus, while PASA's median Combined Score ranked 1st in 6 of the 9 cancer types, it ranked 5th for Liver-HCC and 2nd for Lung-AdenoCA and Ovary-AdenoCA. As another example, MSA ranked last in 5 cancer types but ranked 2nd in Eso-AdenoCA.

All approaches other than MSA had precision of 1 on the majority of spectra but had median recall in the range of 0.75 to 0.80. By contrast, MSA's median precision was < 0.74 while median recall was 1. We speculate that this reflects a tacit consensus that for most application of signature attribution precision is more useful than recall. Therefore, we also investigated whether MSA's precision could be improved by adjusting its parameters. Indeed, we were able to increase its precision and Combined Score, at the cost of decreased recall (Figure 3). Running MSA with these parameters is labelled MSA_opt in the figures.

We also observed that, across tumor types, most approaches had the lowest median Combined Scores for Skin-Melanoma, mainly due to very low recall (Figure 3, S1, Table S4). We originally hypothesized that presence of SBS7a might interfere with detection of SBS7b, since both are dominated by C → T mutations thought to be caused by exposure to ultraviolet radiation. In fact, however, SBS5 and SBS1 were the most common false negatives in Skin-Melanoma (Table S5). For all approaches other than PASA and MSA, SBS5 was a false negative in over half of the spectra in which it was present. Similarly, for all approaches other than PASA and SigPro, SBS1 was a false negative in over half of the spectra in which it was present.

Accuracy on DBS (doublet-base substitution) mutational spectra

On synthetic DBS spectra, PASA had the highest Combined Score, which was significantly higher than that of FitMS, which had next highest (median: 2.92 versus 2.88, $p < 5.2 \times 10^{-6}$, 2-sided Wilcoxon rank-sum test, Figure 4, Table S6). The Combined Scores of other approaches were much lower. Recall for DBS data tended to be better than for SBSs, with medians of 1 for all approaches other than SigPro, although the mean recall for the 3 approaches with the highest Combined Scores ranged from 0.87 to 0.92. As was the case for SBSs, MSA with default parameters had low precision, but unlike the case for SBS, attempts to adjust MSA's parameters were not able to improve its precision. While SigPro performed well on synthetic SBS data (Figure 3), on synthetic DBS data it had lower recall than the other approaches (Figures 4 and S2, Table S6). SigPro's interface offers no parameters that might be tuned to improve its recall on DBS spectra.

For DBS spectra, as was the case for SBS spectra, there was variation in the ranking of the approaches across cancer types. Based on median Combined Scores,

PASA ranked 1st in 8 of the 9 cancer types (tying for 1st in Skin-Melanoma). As another example, FitMS ranked 2nd in 7 of the 9 cancer types, but ranked 1st in Eso-AdenoCA and tied for 1st in Skin-Melanoma.

Accuracy on ID (insertion and deletion) mutational spectra

On synthetic ID spectra, PASA had the largest Combined Score, which was significantly higher than that of FitMS, which had next highest (Figure 5, Table S7, median: 2.91 versus 2.84, $p < 1.1 \times 10^{-10}$, 2-sided Wilcoxon rank-sum test). The Combined Scores of the remaining tools were much lower. As was the case for SBS, MSA with default parameters had low precision. Attempts to adjust MSA's parameters resulted in a slight improvement in precision.

For ID spectra, as was the case for SBS and DBS spectra, there was variation in the ranking of the approaches across cancer types. PASA's median Combined Score ranked 1st in 8 of the 9 cancer types and ranked 2nd in ColoRect-AdenoCA.

CPU usage

We calculated the total CPU time used by the process and its children when running each approach to mutational signature attribution. On all three types of synthetic spectra (SBS, DBS, and ID), MSA required 4 orders of magnitude more CPU time than the least resource-intensive approaches, FitMS and MP. (Figure 6 and Table S8). This is mainly because the MSA algorithm creates simulations of the input data and then tests using each of 4 pre-specified thresholds (program parameter "weak_thresholds") to select the threshold for final signature attribution. For each proposed threshold, MSA evaluates results on at least 1,000 simulated spectra. After

a threshold is selected, MSA calculates confidence intervals for signature attribution by bootstrapping for each input spectrum. All these factors contributed to the substantial CPU resources required by MSA.

DISCUSSION

We have presented the first (to our knowledge) systematic benchmarking of signature attribution on all three of SBS, DBS, and ID mutational signatures, and we have presented a new method for attribution that is based on finding an attribution that maximizes the likelihood of a target spectrum. We assessed the accuracy of this new method, PASA (Presence Attribute Signature Activity) and of other four approaches [17,28,30,31,33], on a total of 2,700 synthetic spectra encompassing SBS, DBS, and ID mutation types.

While some previous studies [21,28,29] benchmarked accuracy on the SBS mutational signatures, we are not aware of any that have benchmarked attribution accuracy for DBS or ID signatures. We also point out that two of these studies did not examine accuracy from the point of view of precision or recall, and instead used mean squared error [21] or a variation on scaled Manhattan distance [29], neither of which can distinguish false negatives from false positives. We propose that understanding the propensity of approaches to include false positive signatures or exclude false negatives is important for applications of signature attribution. These would include molecular cancer epidemiology, for which one might want to determine with certainty whether the signature of a particular mutagen is enriched in a particular group of cancers [5,7–9]. They would also include efforts to understand the mutational exposures or processes responsible for oncogenic mutations [8,11]. In addition, the accuracy measures in [21] and [29] may have little power to distinguish

the accuracy of different signature attribution methods. For example [21] stated that "[a]ll methods give almost identical results"; see also Figure 9 in [21].

We also demonstrated that attribution is a challenging task. First, we showed that, for a given spectrum, there are often over one hundred possible alternative attributions that yield reasonably good similarity to the spectrum (Figure 2, Table S2). Second, we showed that, for a given synthetic spectrum, there can be many incorrect attributions that provide more similar reconstructions than the correct, ground-truth attribution (Table S3).

For SBS, DBS, and ID mutational signatures, across all cancer types together, the new algorithm presented here, PASA (Presence Attribute Signature Activity) was more accurate than the other four approaches [17,28,30,31,33] (Figures 3 to 5). However, the ranking for different approaches varied by cancer type, and, for example, PASA was the most accurate approach in 6 of the 9 cancer types (Figure 3, Table S4). We speculate that this is a consequence of the inherent challenges of mutational signature attribution. Signature attribution remains an open area, and advances might depend partly on integrating data from all thee mutation types (SBS, DBS, ID) or on incorporating prior evidence on signature prevalence and activity in different cancer types.

DATA AND CODE AVAILABILITY

The R code for the PASA algorithm is freely available at https://github.com/steverozen/mSigAct. The version reported here is V3.0.1-branch, which can be installed with the R call

```
remotes::install_github(repo = "steverozen/mSigAct", ref = "v3.0.1-branch")
```

The PASA algorithm is implemented in the function mSigAct::PresenceAttributeSigActivity.

All other code and data for this paper are freely available at https://github.com/Rozen-Lab/sig_attribution_paper_code.

FUNDING

This work was supported by National Medical Research Council, Singapore, grants NMRC/CIRG/1422/2015 and MOH-000032/MOH-CIRG18may-0004 to S.G.R, and the Singapore Ministries of Health and Education via the Duke-NUS Signature Research Programmes.

AUTHOR CONTRIBUTIONS

N.J. : Design of PASA algorithm, synthetic data generation, running approaches to signature attribution, formal analysis, visualization, writing, reviewing, editing.

Y.W. : Running approaches to signature attribution.

S.G.R. : Funding acquisition, conceptualization, supervision, resources, design of PASA algorithm, visualization, writing, reviewing, editing.

REFERENCES

1. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. Nature 2020; 578:94–101
2. Cooper DN, Mort M, Stenson PD, et al. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Human Genomics 2010; 4:406
3. Rayner E, van Gool IC, Palles C, et al. A panoply of errors: polymerase proofreading domain mutations in cancer. Nat Rev Cancer 2016; 16:71–81
4. Hoang ML, Chen C-H, Chen P-C, et al. Aristolochic Acid in the Etiology of Renal Cell Carcinoma. Cancer Epidemiology, Biomarkers & Prevention 2016; 25:1600–1608
5. Poon SL, Pang S-T, McPherson JR, et al. Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool. Science Translational Medicine 2013; 5:197ra101-197ra101
6. Alexandrov LB, Ju YS, Haase K, et al. Mutational signatures associated with tobacco smoking in human cancer. Science 2016; 354:618–622
7. Huang MN, Yu W, Teoh WW, et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. Genome Res. 2017;
8. Ng AWT, Poon SL, Huang MN, et al. Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. Science Translational Medicine 2017; 9:eaan6446
9. Dziubańska-Kusibab PJ, Berger H, Battistini F, et al. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. Nat Med 2020; 26:1063–1069

10. Volkova NV, Meier B, González-Huici V, et al. Mutational signatures are jointly shaped by DNA damage and repair. Nat Commun 2020; 11:2169
11. Boot A, Liu M, Stantial N, et al. Recurrent mutations in topoisomerase IIα cause a previously undescribed mutator phenotype in human cancers. Proceedings of the National Academy of Sciences 2022; 119:e2114024119
12. Davies H, Glodzik D, Morganella S, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med 2017; 23:517–525
13. Grolleman JE, De Voer RM, Elsayed FA, et al. Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. Cancer Cell 2019; 35:256-266.e5
14. Boot A, Huang MN, Ng AWT, et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. Genome Res. 2018; 28:654–665
15. Kucab JE, Zou X, Morganella S, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell 2019; 177:821-836.e16
16. Koh G, Zou X, Nik-Zainal S. Mutational signatures: experimental design and analytical framework. Genome Biology 2020; 21:37
17. Degasperi A, Zou X, Dias Amarante T, et al. Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. Science 2022; 376:abl9283
18. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013; 500:415–421
19. Roberts, Nicola D. Patterns of somatic genome rearrangement in human cancer. PhD thesis, https://doi.org/10.17863/CAM.22674. 2018;
20. Liu M, Wu Y, Jiang N, et al. mSigHdp: hierarchical Dirichlet process mixture modeling for mutational signature discovery. NAR Genom Bioinform 2023; 5:lqad005
21. Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. PLOS ONE 2019; 14:e0221235
22. Wu Y, Chua EHZ, Ng AWT, et al. Accuracy of mutational signature software on correlated signatures. Sci Rep 2022; 12:390
23. Islam SMA, Díaz-Gay M, Wu Y, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genomics 2022; 2:100179
24. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 2019; 47:D941–D947
25. Zhang T, Joubert P, Ansari-Pour N, et al. Genomic and evolutionary classification of lung cancer in never smokers. Nat Genet 2021; 53:1348–1359
26. Priestley P, Baber J, Lolkema MP, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 2019; 575:210–216
27. Franco I, Helgadottir HT, Moggio A, et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. Genome Biology 2019; 20:285
28. Díaz-Gay M, Vangara R, Barnes M, et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. Bioinformatics 2023; 39:btad756
29. Medo M, Medová M. A comprehensive comparison of tools for fitting mutational signatures [preprint: http://arxiv.org/abs/2310.01562]. 2023;
30. Manders F, Brandsma AM, de Kanter J, et al. MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genomics 2022; 23:134

31. Senkin S. MSA: reproducible mutational signature attribution with confidence based on simulations. BMC Bioinformatics 2021; 22:540
32. Johnson, Steven G. The NLopt nonlinear-optimization package http://github.com/stevengj/nlopt.
33. Moody S, Senkin S, Islam SMA, et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. Nat Genet 2021; 53:1553–1563

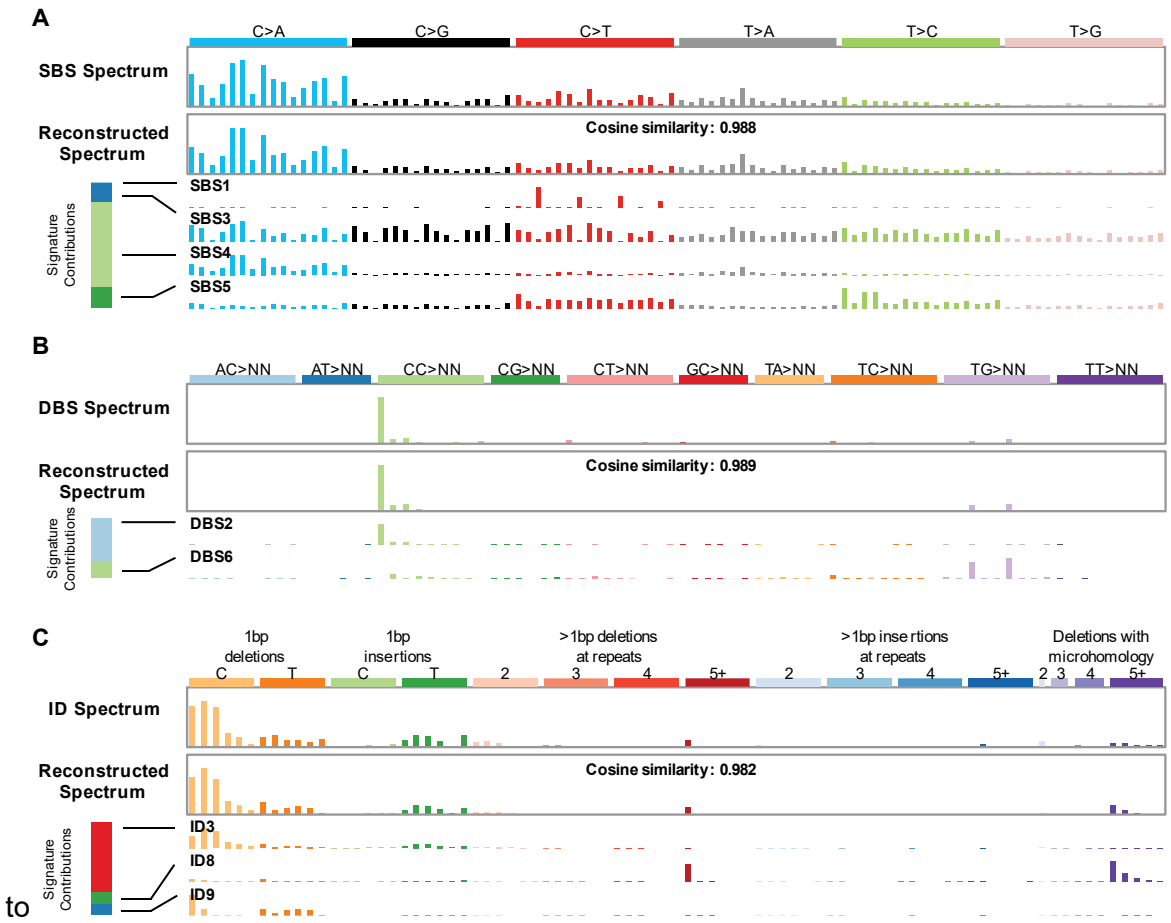

Figure 1. The task of mutational signature attribution is to find signatures that can reconstruct the mutational spectrum well**. (A)** Example of an SBS spectrum that can be reconstructed with a cosine similarity of 0.988 from four signatures. The bar at the left shows that mutational signature SBS4 contributed the most mutations to this spectrum. This signature is associated with tobacco smoking. **(B)** The DBS spectrum from the same tumor can be reconstructed with a cosine similarity of 0.989 from two signatures. The bar at the left shows that mutational signature DBS2, which is also associated with tobacco smoking, contributed the most mutations to this spectrum. **(C)** The ID spectrum from the same tumor can be reconstructed with a cosine similarity of 0.982 from three signatures. The bar at the left shows that mutational signature ID3 contributed the most mutations to this spectrum. Like the SBS signature SBS4, this ID signature is also associated with tobacco smoking. Spectra from Lung-AdenoCA::SP52667 [1]. The *x* axes of all panels follow the conventions described at https://cancer.sanger.ac.uk/signatures/.

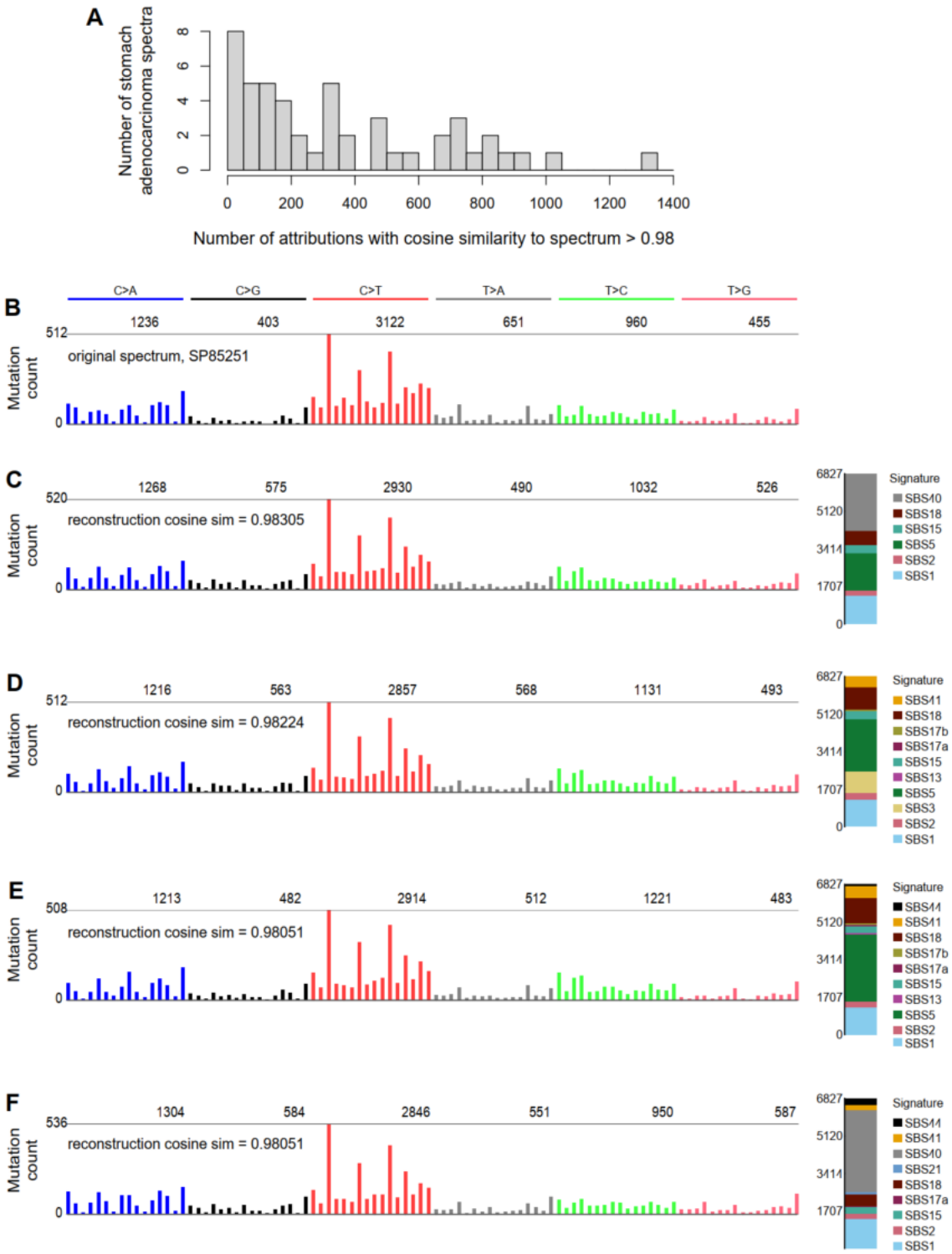

Figure 2. Many mutational spectra have numerous distinct attributions with cosine similarity > 0.98. Shown here is the example of stomach adenocarcinomas (Stomach-AdenoCA) from the Pan Cancer Analysis of Whole Genomes mutation data analyzed in [1]. **(A)** Numbers of stomach adenocarcinoma spectra by number of distinct attributions with cosine similarity to the spectrum > 0.98. **(B)** One example spectrum from those analyzed in panel A. The *x* axis of this and the following panels follow the conventions described at https://cancer.sanger.ac.uk/signatures/. **(C-F)** Several example attributions that have cosine similarity to the spectrum in panel B > 0.98. Code for this figure and the attributions are available at https://github.com/Rozen-Lab/sig_attribution_paper_code/common_code/figure_2/

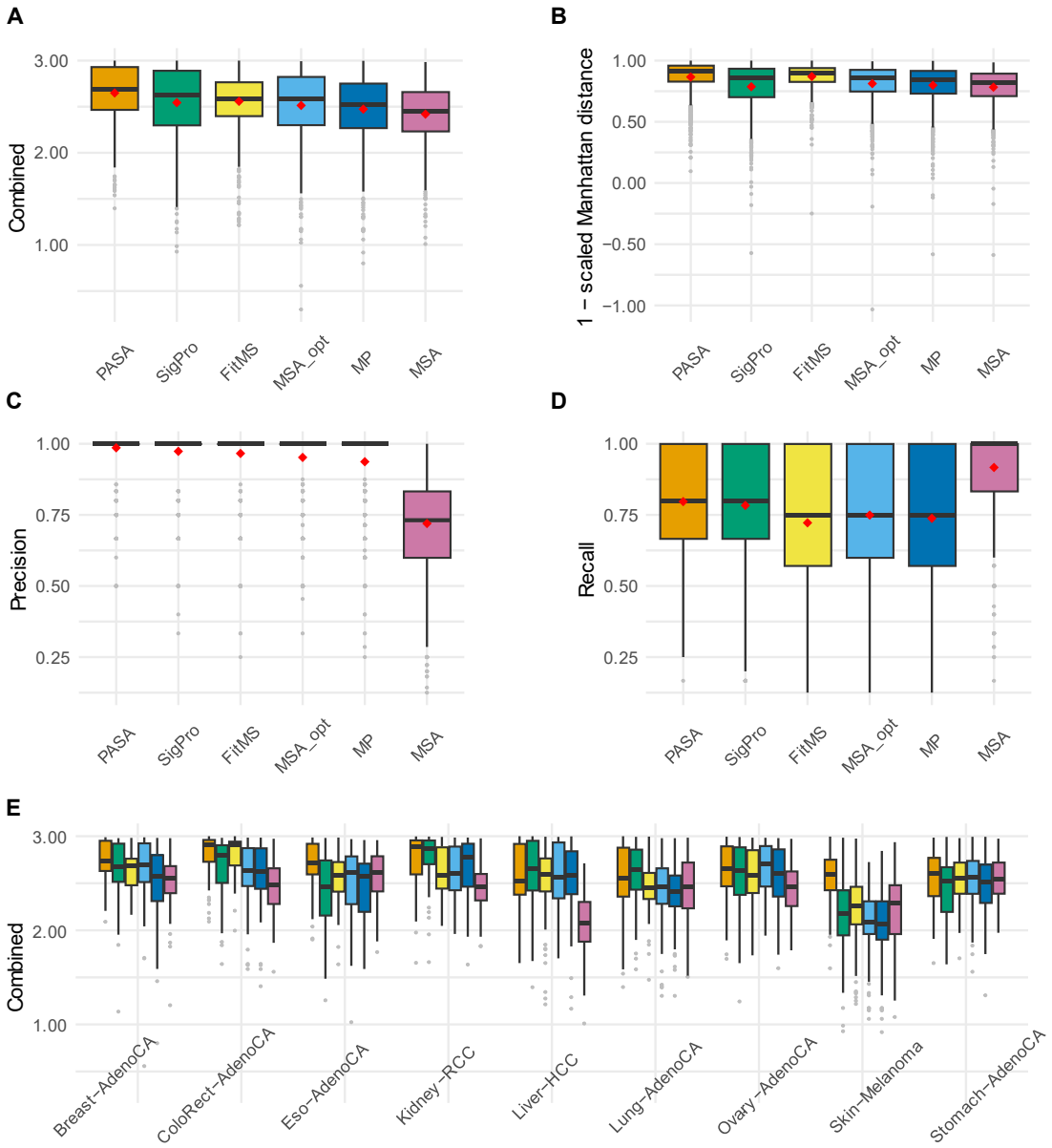

Figure 3. Accuracy of mutational signature attribution approaches on synthetic SBS spectra. **(A-D)** Accuracy measures over all synthetic SBS spectra. Combined Score is the sum of 1 – scaled Manhattan distance, precision and recall. Scaled Manhattan distance is the Manhattan distance between the spectrum and the reconstructed spectrum divided by the total mutation count. Dark horizontal lines indicate medians, red diamonds indicate means. **(E)** Combined Scores by each cancer type. PASA, Presence Attribute Signature Activity; SigPro, SigProfilerAssignment; FitMS, Signature Fit Multi-Step; MP, MutationalPatterns; MSA, Mutational Signature Attribution with default parameters; MSA_opt, Mutational Signature Attribution with the weak_thresholds parameter tuned for higher precision (Table S1). Abbreviations for cancer types are as in reference [1].

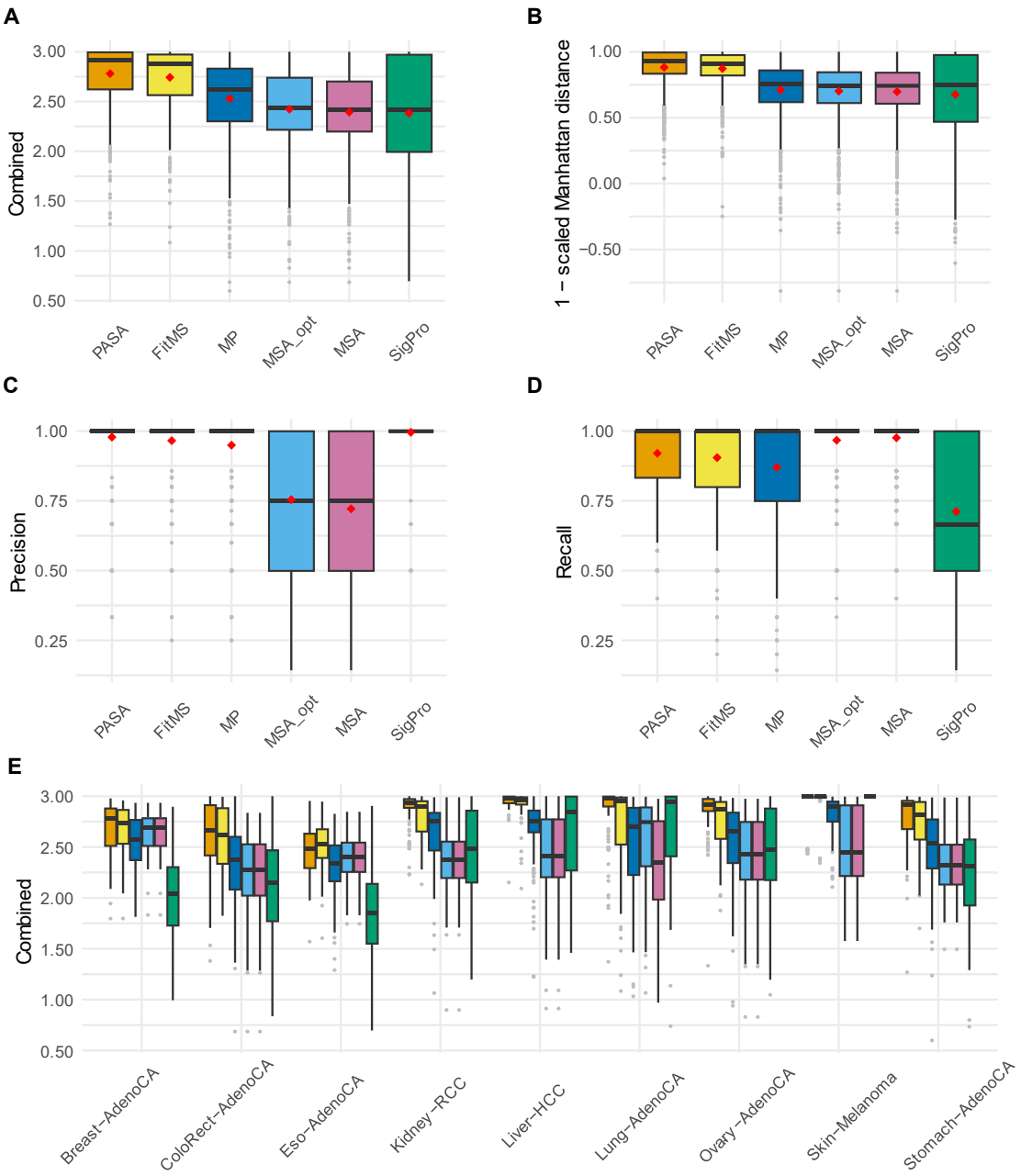

Figure 4. Accuracy of signature attribution approaches on synthetic DBS spectra**.** Abbreviations are as in Figure 3.

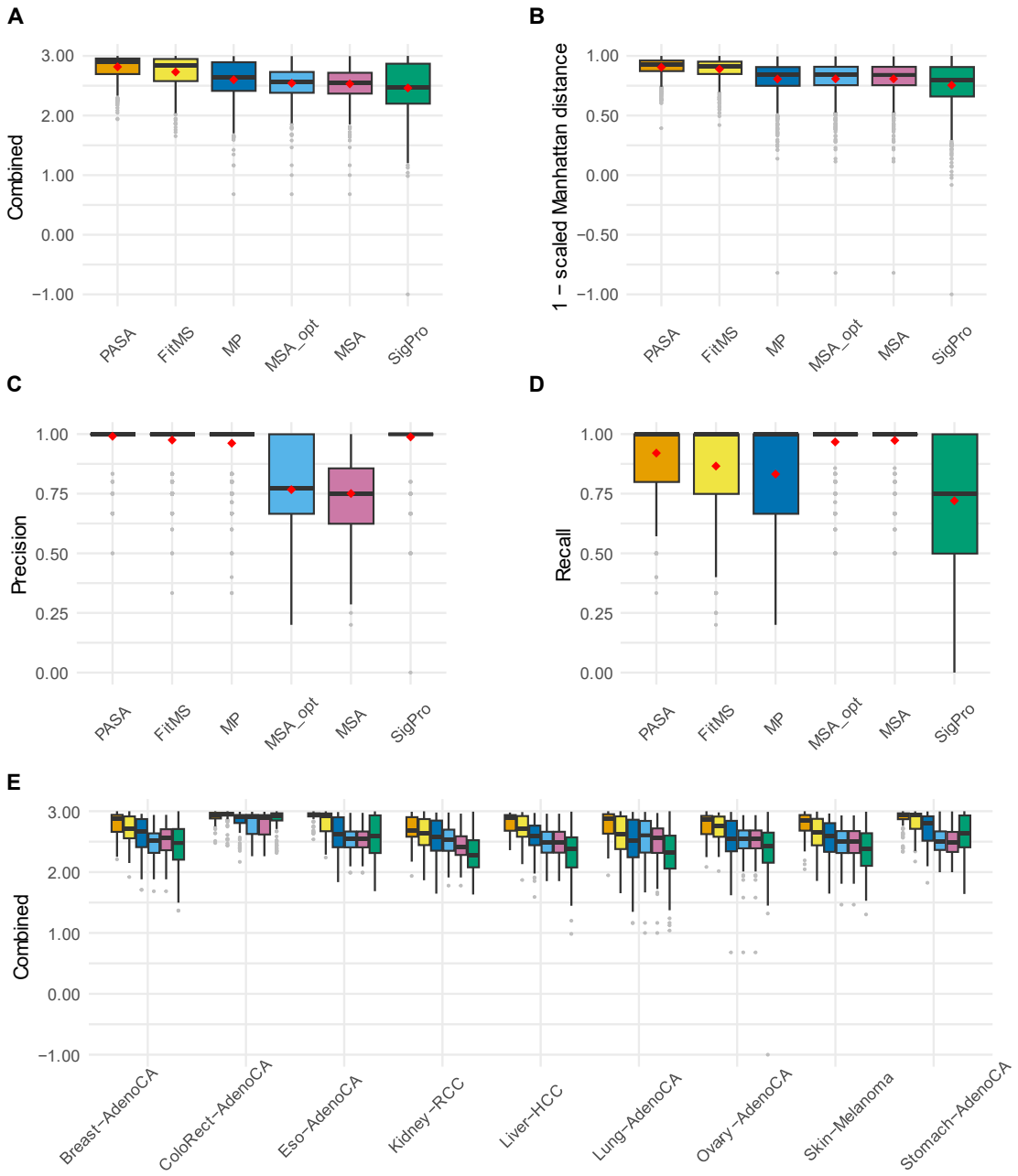

Figure 5. Accuracy of signature attribution approaches on synthetic ID (insertion and deletion) spectra**.** Abbreviations are as in Figure 3.

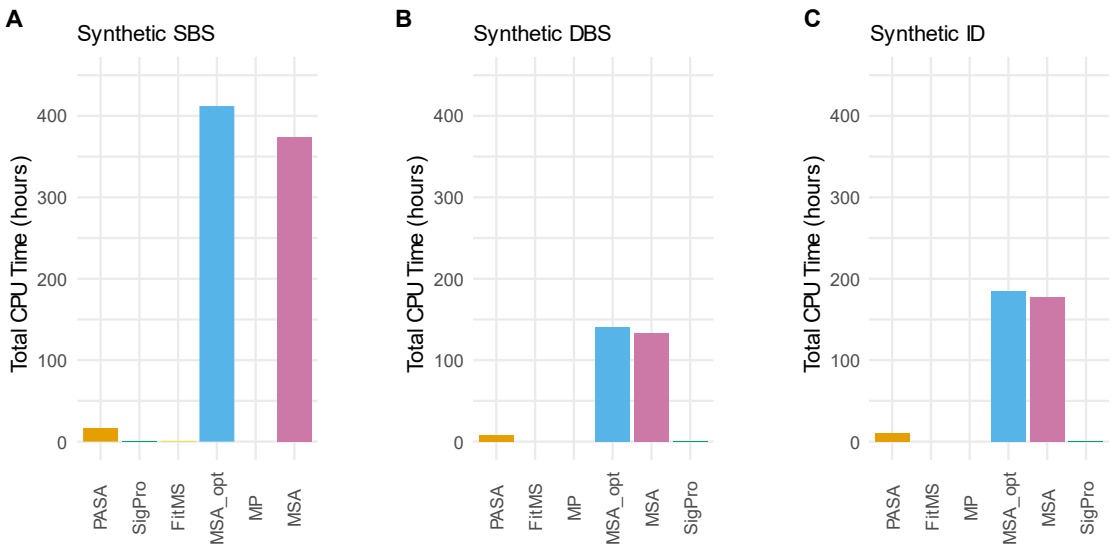

Figure 6. Total CPU time for running approaches to signature attribution on synthetic **(A)** SBS, **(B)** DBS and **(C)** ID mutational spectra**.** The approaches are ordered by decreasing median Combined Score (Figures 3 to 5). Abbreviations for attribution approaches are as in Figure 3.

Supplementary Figures for

A new approach to the challenging problem of mutational signature attribution

Nanhai Jiang[1,2], Yang Wu[1,2], Steven G. Rozen[1,2,3*]

[1] Centre for Computational Biology, Duke-NUS Medical School, Singapore 169857, Singapore

[2] Programme in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore 169857, Singapore

[3] Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, 27710, United States

*To whom correspondence may be addressed:

Steven G. Rozen
Department of Biostatistics and Bioinformatics, Duke University School of Medicine
2424 Erwin Road, Suite 11092
Hock Plaza Box 2721
Durham, North Carolina, 27710, United States
steverozen@pm.me

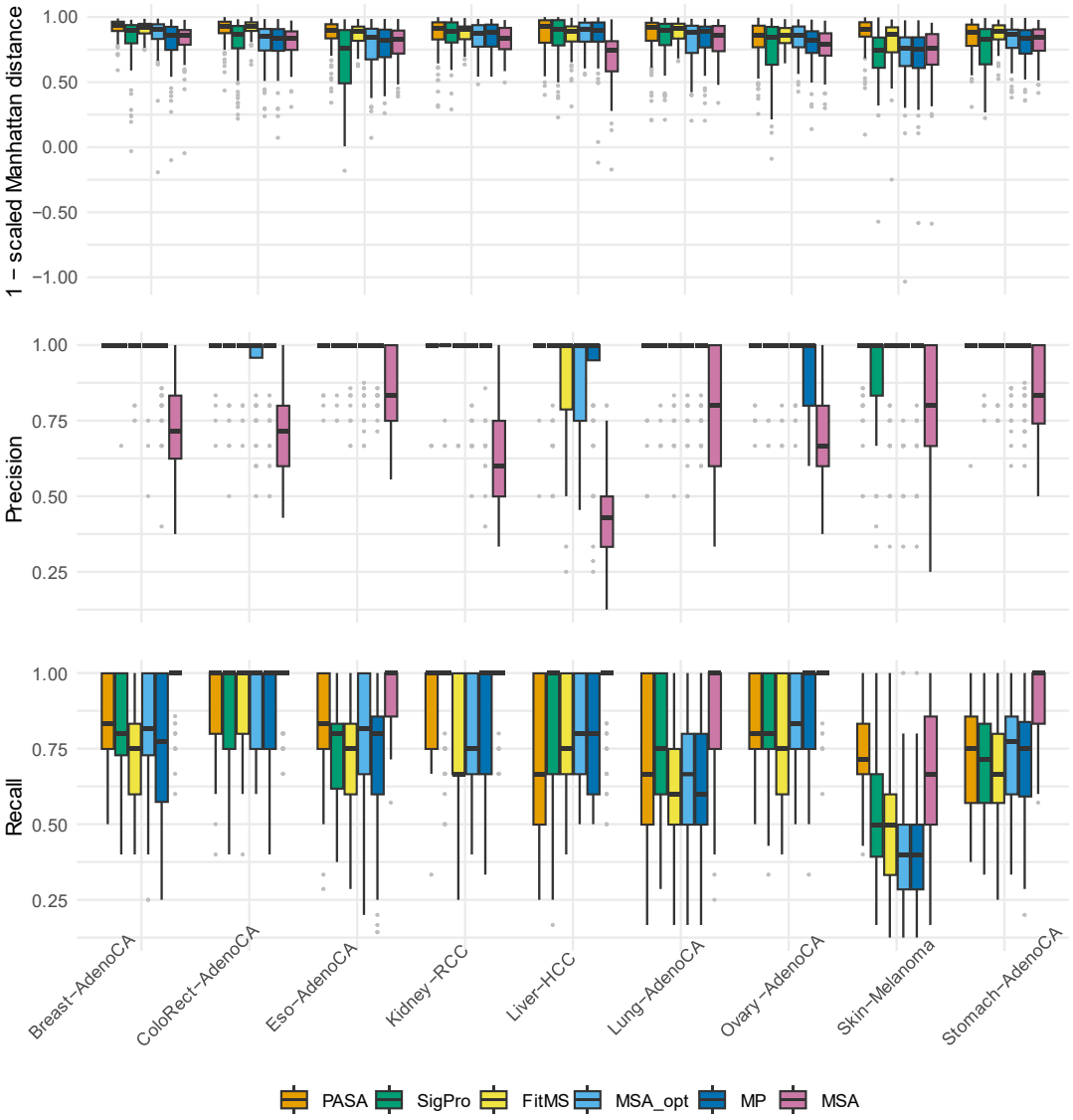

Figure S1. Accuracy of mutational signature attribution approaches on synthetic SBS data for each cancer type**.** Scaled Manhattan distance is the Manhattan distance between the spectrum and the reconstructed spectrum, divided by total mutation count. The attribution approaches are ordered by descending median of the Combined score for all cancer types from highest to lowest. See main Figure 3 for more details. PASA, Presence Attribute Signature Activity; SigPro, SigProfilerAssignment; FitMS, Signature Fit Multi-Step; MP, MutationalPatterns; MSA, Mutational Signature Attribution with default parameters; MSA_opt, Mutational Signature Attribution with the weak_thresholds parameter tuned for higher precision (Table S1). Abbreviations of cancer types are as in Alexandrov et al., 2020, doi: 10.1038/s41586-020-1943-3

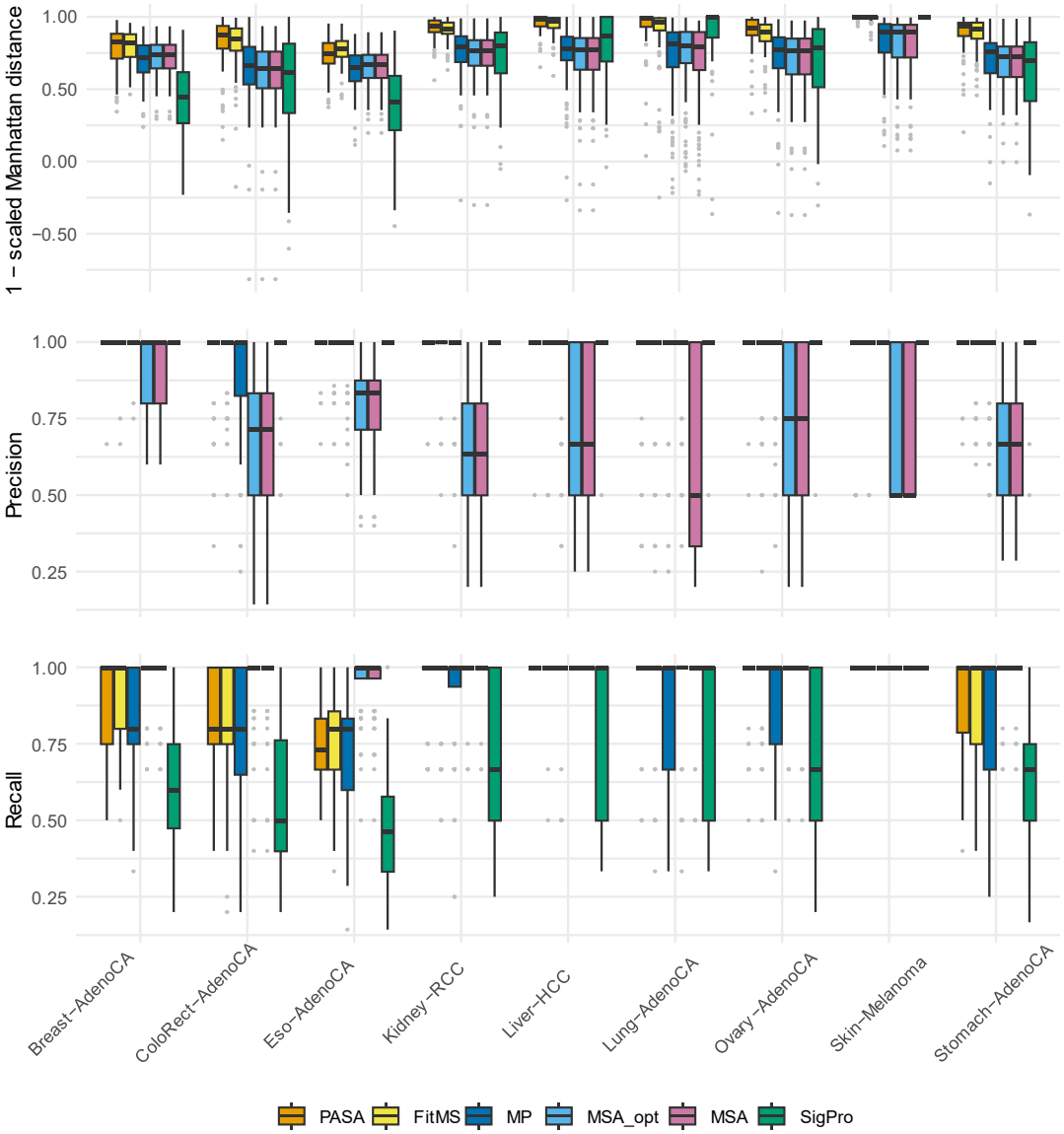

Figure S2. Accuracy of mutational signature attribution approaches on synthetic DBS data for each cancer type**.** Scaled Manhattan distance is the Manhattan distance between the spectrum and the reconstructed spectrum, divided by total mutation count. The attribution approaches are ordered by descending median of the Combined Score for all cancer types from highest to lowest. See Figure 4 for more details. Abbreviations are as in Figure S1.

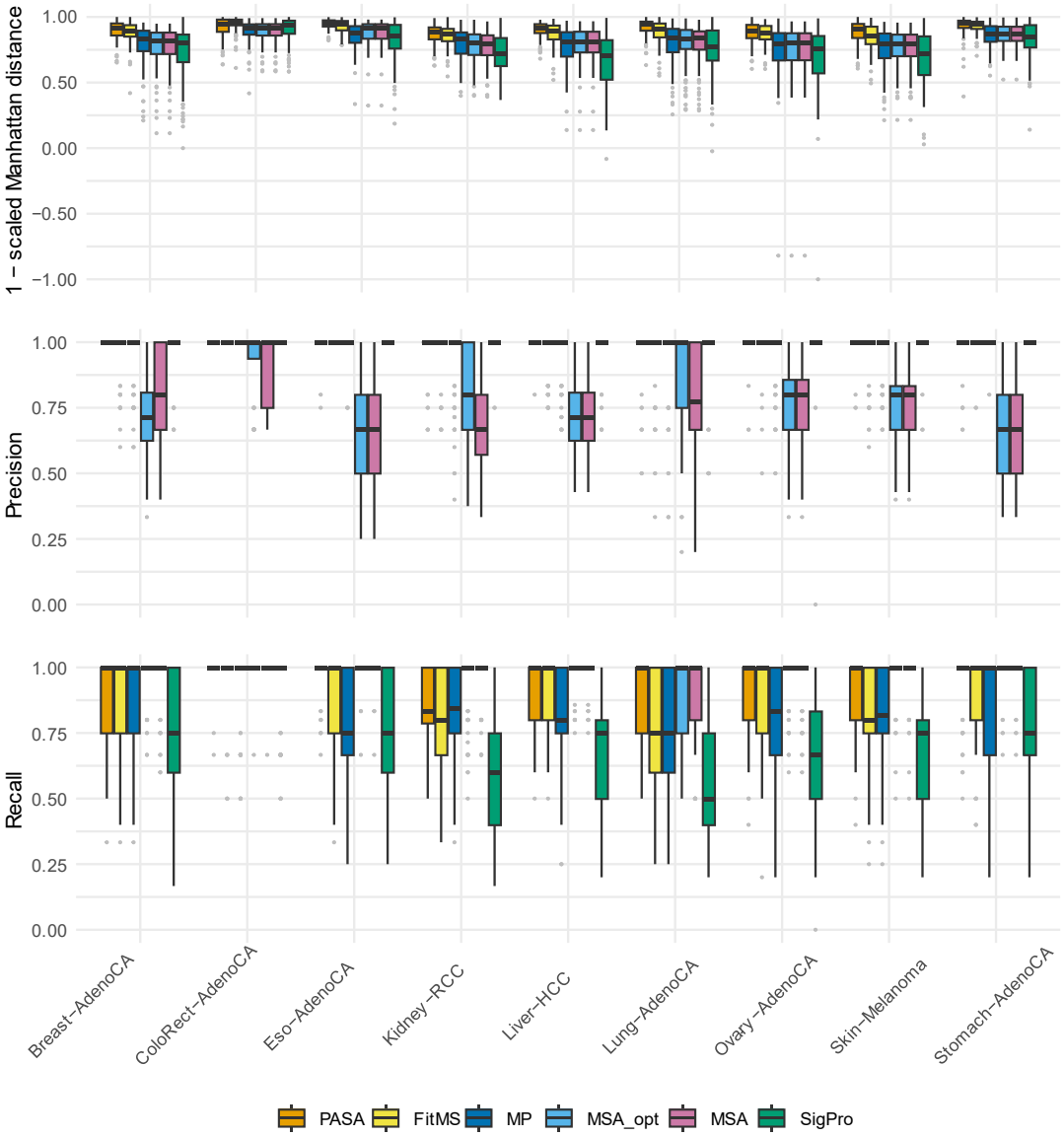

Figure S3. Accuracy of mutational signature attribution approaches on synthetic ID data for each cancer type**.** Scaled Manhattan distance is the Manhattan distance between the spectrum and the reconstructed spectrum, divided by total mutation count. The attribution approaches are ordered by descending median of the Combined Score for all cancer types from highest to lowest. See Figure 5 for more details. Abbreviations are as in Figure S1.