**Description of tests of SignatureAnalyzer and SigProfiler on synthetic mutational spectra**

**Draft paragraph for revised main text**

We tested SignatureAnalyzer and SigProfiler on 8 collections of synthetic data sets in which known signature profiles were used to generate catalogs of synthetic mutational spectra. Both approaches performed well in re-extracting the known signatures in realistically complex data. However, the tests highlighted the importance of and challenges in selecting the number of signatures, because extracted signatures discordant from the known input usually arose from difficulty in selecting the correct number of signatures. Moreover, the tests confirmed that use of NMF-based approaches to extract signatures is not a purely automatic process. Instead, signature extraction requires human judgement that considers all available data and evidence, and reasonable priors, as well as form of human-guided sensitivity analysis and that confirms that highly similar signatures are extracted from different groupings of tumors and that examines any discordant results. (All of these approaches were used in the determination of the signature profiles we report in this paper). These findings from tests on synthetic data are consistent with abundant literature on NMF-based approaches in other domains, which widely acknowledges that the choice of the number of factors in the W matrix (for mutational signatures, this is the number of extracted signatures) is rarely amenable to complete automation [1-4].

More details are provided in methods, and all results are at https://doi.org/10.7303/syn18497223 (summarized at <https://doi.org/10.7303/syn18511087.1>).

References:

1. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. Computational statistics & data analysis. 2007 Sep 15;52(1):155-73.

2. Cichocki A, Zdunek R, Phan AH, Amari SI. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons; 2009 Jul 10.

3. Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS computational biology. 2008 Jul 25;4(7):e1000029.

4. Steyvers M, Griffiths T. Probabilistic topic models. Handbook of latent semantic analysis. 2007 Jan;427(7):424-40.

**Draft Methods subsection**

**Introduction**

Our goal was to test SignatureAnalyzer (SA) and SigProfiler (SP) on realistic synthetic data. We operationally defined "realistic" as corresponding to either SA's or SP's analysis of the PCAWG genome data. SA’s reference signature profiles were based on “COMPOSITE” signatures, consisting of 1536 strand-agnostic single base substitutions (SBSs) , 78 doublet base substitutions and 83 types of small insertions and deletions, for a total of 1697 mutation types. SP’s reference analysis was based on strand-agnostic single base substitution in the context of one 5’ and one 3’ base; we term this “SBS96” data. For each test, we generated two sets of "realistic" data: *SP-realistic*, based on SP's reference signatures and attributions, and *SA-realistic*, based on SA's reference signatures and attributions, as well as two other types of data that involved using SA profiles with SP attributions and vice versa.

**Generating synthetic data - overview**

Synthetic data for sets of synthetic tumors of a given cancer type, *t*, were generated based on three parameters that were in turn based on the observed statistics for each signature, *s*, in a cancer type, *t*:

π, the proportion of tumours of cancer type *t* with *s*

μ, the mean of log10 of the number of *s* mutations in those *t* tumors that have s

σ, the standard deviation of log10 of the numbers of *s* mutations in those *t* tumours that have *s*

To generate synthetic data,

(i) the proportion of tumours affected by *s* was drawn from the binomial distribution based on π,

(ii) the number of mutations due to *s* in an affected tumor was drawn from a normal distribution based on μ and σ.

The code used to generate the synthetic data and summarize SignatureAnalyzer and SigProfiler results is available as the SynSig package: <https://github.com/steverozen/SynSig/tree/v0.2.0>

**Description of each synthetic data set**

**i. 1,000 mutational spectra modelled on signatures found in one human cancer type (pancreatic adenocarcinoma).**

<https://doi.org/10.7303/syn18500212.1>

**ii. 2,700 mutational spectra consisting of 300 synthetic spectra from each of 9 different cancer types.** These consist of 300 synthetic spectra from each of bladder transitional cell carcinoma, esophageal adenocarcinoma, breast adenocarcinoma, lung squamous cell carcinoma, renal cell carcinoma, ovarian adenocarcinoma, osteosarcoma, cervical adenocarcinoma, and stomach adenocarcinoma.

<https://doi.org/10.7303/syn18500213.1>

**iii. Mutational spectra generated from combinations of flat, relatively featureless mutational signatures -- version 1,** 1000 synthetic tumours comprised of 500 synthetic Kidney-RCCs (high prevalence and mutation load from SBS5 and SBS40 signatures) and 500 synthetic ovarian adenocarcinomas (high prevalence of and mutation load from SBS3). This data set embodies tumours with high prevalences of the main flat signatures, SBS3, SBS5, and SBS40, in a somewhat realistic context, as SBS3 is mostly not found in tumors with SBS5 and SBS40.

<https://doi.org/10.7303/syn18500214.1>

**iv. Mutational spectra generated from combinations of flat, relatively featureless mutational signatures -- version 2,** 1000 synthetic spectra all constructed entirely from SBS3, SBS5, and SBS40, using mutational loads modeled on kidney-RCC (SBS5 and SBS40) and ovarian adenocarcinoma (SBS3). Most synthetic spectra have contributions from all three signatures.

<https://doi.org/10.7303/syn18500215.1>

**v. Mutational spectra generated from signatures with overlapping and potentially interfering profiles - version 1.** 500 synthetic bladder transitional cell carcinomas (high in SBS2 and SBS13) and 500 synthetic skin melanomas (high in SBS7a,b,c,d). The potential interference is between

SBS2 (mainly C > T) and SBS7a,b (mainly C > T).

<https://doi.org/10.7303/syn18500217.1>

**vi. Mutational spectra generated from signatures with overlapping and potentially interfering profiles - version 2.** 1000 synthetic tumours composed from SBS2 and SBS7a,b Mutation load distributions drawn from bladder transitional cell carcinoma (SBS2) and from skin melanoma (SBS7a,b). Most spectra contain both signatures. The potential interference is between SBS2 (mainly C > T) and SBS7a,b (mainly C > T).

<https://doi.org/10.7303/syn18500216.1>

**vii. Mutational spectra generated from combinations of signatures conferring high and low mutation burdens.** Based on 500 synthetic non-hypermutated tumours (parameters for SBS1 and SBS5 estimated from colorectal and uterine adenocarcinomas) and 500 hyper-mutated tumours (parameters for SBS26 and SBS44 estimated from non hypermutated colorectal and uterine adenocarcinomas). High and low mutation burden tumours are segregated for SignatureAnalyzer (which does low mutation burden tumours first, then high-burden tumours). SigProfiler analyses all tumours together.

<https://doi.org/10.7303/syn18500218.1>

<https://doi.org/10.7303/syn18500219.1>

<https://doi.org/10.7303/syn18500216.1>

**viii. A set of 30 random 96-feature mutational signature profiles and a set of 30 random 1697-feature signature profiles (mimicking COMPOSITE signatures, which have 1697 mutation types).** Each of these are used in two types of exposures, one with more (mean ~15.6) signatures per tumor and one with fewer (mean ~4) signatures per tumor.

<https://doi.org/10.7303/syn18500221.1>