mSigHdp: hierarchical Dirichlet process mixture modeling for mutational signature discovery

Mo Liu[1,2,†], Yang Wu[1,2,†], Nanhai Jiang[1,2], Arnoud Boot[1,2], Steven G. Rozen[1,2,*]

[1] Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore

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

[†] Denotes equal contribution.

[*] To whom correspondence may be addressed: SGR: steve.rozen@duke-nus.edu.sg

Abstract

Mutational signatures are characteristic patterns of mutations caused by endogenous or exogenous mutational processes. These signatures can be discovered by analyzing mutations in a large set of samples – usually somatic mutations in tumor samples. Most approaches to mutational-signature discovery are based on non-negative matrix factorization. Alternatively, signatures can be inferred using hierarchical Dirichlet process (HDP) mixture models, an approach that has been relatively little explored. These models assign mutations to clusters and view each cluster of mutations as being generated from a particular mutational process. Here we describe mSigHdp, an improved approach to using HDP mixture models to discover mutational signatures. We benchmarked mSigHdp and several other programs on realistic synthetic data. For single-base mutations, mSigHdp and the widely used SigProfilerExtractor discovered signatures better than the other 2 programs and had different strengths: mSigHdp discovered more of the signatures present in the synthetic data, while SigProExtractor discovered fewer false positives. mSigHdp was better able to discover rare signatures, which may be an advantage, since most common signatures have probably already been discovered. For small insertion-and-deletion mutations, mSigHdp discovered signatures better than the other 3 programs. Thus, mSigHdp is an advance for discovering rare SBSs mutational signatures and for discovering small insertion-and-deletion mutational signatures.

Availability: mSigHdp is available at GitHub public repositories https://github.com/steverozen/mSigHdp and https://github.com/steverozen/hdpx. Synthetic data, code for generating the synthetic data, code for running the mutational-signature discovery programs and analyzing their results, the main outputs of the programs,

and code for generating the data figures in this paper are available at https://github.com/Rozen-Lab/Liu_et_al_Sup_Files.

Supplementary information: Supplementary information is provided on the bioRxiv web site for this manuscript.

Introduction

Mutational signatures are characteristic patterns of mutations caused by endogenous factors or by exposure to exogenous mutagens (Alexandrov et al., 2020). Examples of the former are signatures due to deamination of 5-methylcytosine or to activated APOBECs. Examples of the latter are the signatures caused by tobacco smoking or by ultraviolet radiation. The concept of mutational signatures can be applied to various types and classifications of mutations. Single base mutations (single base substitutions, SBSs) are most often classified in the context of the preceding and following bases while considering the mutation to have occurred from a pyrimidine (C or T) (Figure 1A). The mutation classes are, for example, ACA → AAA, ACA → AGA, … CCA → CAA, …, TTT → TGT. Insertion and deletion mutations (indels) are classified first by their length. Indels of 1 base pair are classified according to the inserted or deleted base and the length of the polynucleotide repeat (if any) in which the indel occurred. Indels of length > 1 base pair are classified according to (i) their length, (ii) whether the indel is in a repeat and the length of the repeat and, (iii) for deletions, if there is microhomology between the ends of the deleted sequence and the immediately surrounding sequence (Figure 1B). The classification of indels involves several subtleties, and we refer interested readers to the full description at https://www.synapse.org/#!Synapse:syn11801742. For a given biological sample, the counts of mutations in each of the mutational classes constitute the *mutational spectrum* of that sample. Since the samples are usually tumors, for simplicity of exposition, we will refer to biological samples as tumors, understanding that we can discover mutational signatures in other types of biological samples as well.

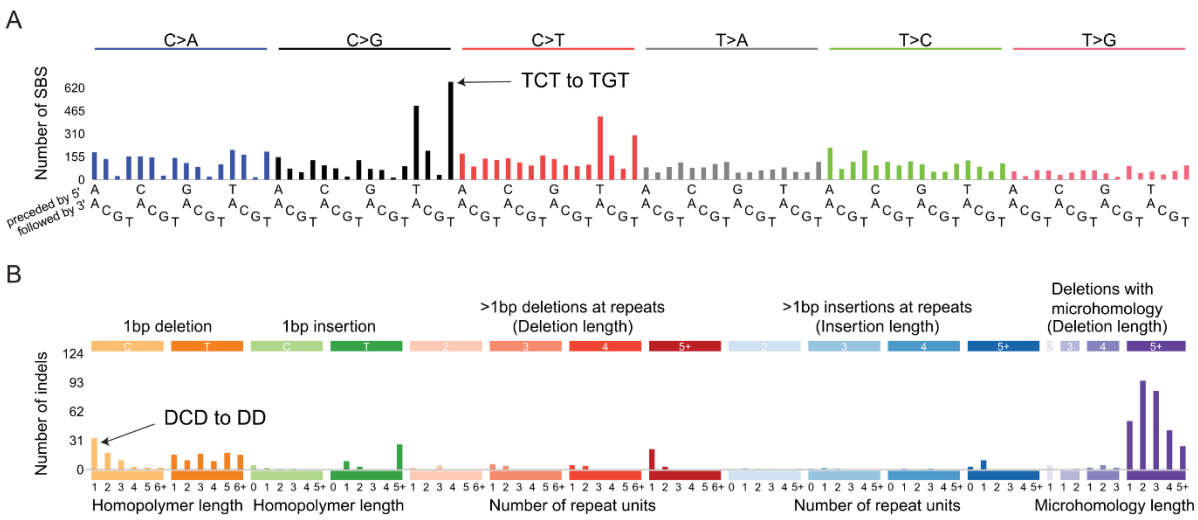

Figure 1 Examples of (A) SBS and (B) indel mutational spectra in a breast cancer. The height of each vertical bar indicates the number of mutations of a particular mutation class, for example, in panel A, 617 mutations from TCT to TGT and in panel B, 34 mutations from DCD to DD (deletion of one C in the middle of two non-C nucleotides, denoted D). This is Breast−AdenoCA::SP116341 from Alexandrov et al. (2020) and ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020).

For our purposes, a *mutational signature* consists of the proportions of mutations in each of the mutation classes, as generated by a particular mutational process. More abstractly, a mutational signature is a multinomial probability vector that maps mutation classes to their probabilities. We view the mutational spectrum of 1 tumor as the mutation-class-wise sum of mutations caused by several mutational processes (Figure 2). In this model, each mutational process generates individual mutations by sampling from the multinomial distribution parameterized by its signature, and the spectrum consists of the sum of mutations in each mutational class generated by all the mutational processes operating in the tumor. Abstractly, the task of mutational signature discovery is to infer a set of signatures that can reasonably reconstruct the spectra of a set of tumors. To be

useful, this set of signatures should approximately correspond to underlying biological processes.

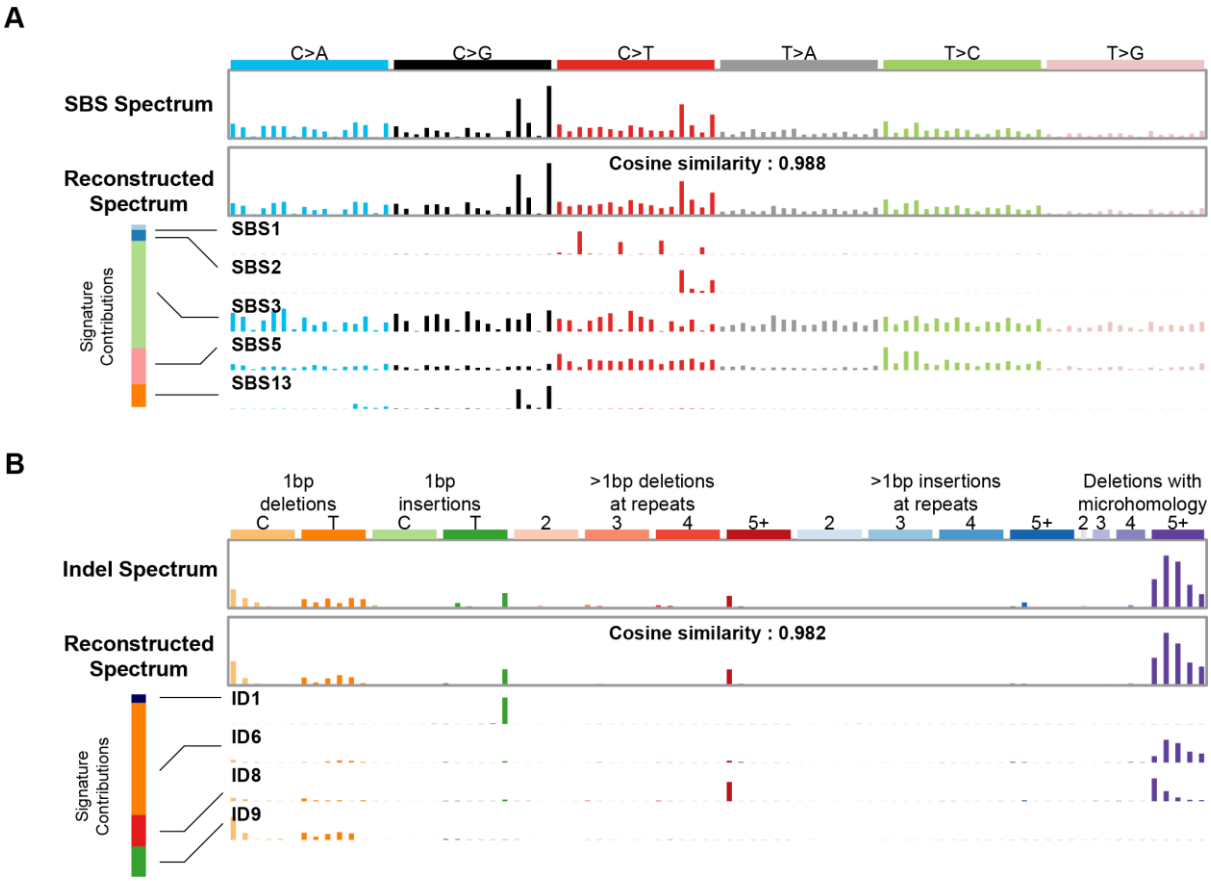

Figure 2 Mutational spectra can be explained as a linear combination of mutations generated by mutational signatures.

(A) Example of an SBS spectrum that can be reconstructed with a cosine similarity of 0.988 from 5 signatures. The bar at the left shows that mutational signature SBS3 contributed the most mutations to this spectrum. This signature is due to defective homologous-recombination-based DNA repair. The contributions of signatures SBS2 and SBS13 (both due to the activity of endogenous APOBEC cytidine deaminases) are visible in the original and reconstructed spectra. **(B)** The indel spectrum from the same tumor can be reconstructed with a cosine similarity of 0.982 from 4 signatures. The bar at the left shows that mutational signature ID6 contributes the most mutations to this spectrum. Like SBS3, this signature is also due to defective homologous-recombination-based DNA repair. Spectra from Breast−AdenoCA::SP116341 (Alexandrov et al., 2020; ICGC/TCGA Pan-Cancer Analysis of

Whole Genomes Consortium, 2020), plotting code in https://github.com/Rozen-Lab/Liu_et_al_Sup_Files.

Early work on discovering mutational signatures relied on approaches based on non-negative matrix factorization (NMF) (Alexandrov, Nik-Zainal, Wedge, Campbell, & Stratton, 2013; Nik-Zainal et al., 2012) and many subsequent efforts also used NMF based approaches. To discover mutational signatures, NMF views the tumor spectra as constituting a matrix, with rows being the mutation classes (e.g., for SBSs, CCA → CAA), with the columns being tumors, and with each cell containing the number of mutations of one mutation class for one tumor. NMF then derives an approximate factorization in which the matrix of observed mutation counts is the product of two matrices. The first is a matrix of K signatures in which each signature is a column. The second is a matrix in which each element is the number of mutations due to a particular signature in a particular tumor, which we call the *exposure* of that signature in that tumor. Alexandrov and colleagues developed a series of implementations of this concept (Alexandrov et al., 2020; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013; Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013; Islam et al., 2020; Nik-Zainal et al., 2012; Nik-Zainal et al., 2016) and indeed signatures discovered by the version in (Alexandrov et al., 2020) are the basis for a widely used reference set of mutational signatures (https://cancer.sanger.ac.uk/signatures/). The most recent incarnation of this concept is SigProfilerExtractor, which was assessed along with 13 other approaches in a benchmarking study on synthetic SBS data (Islam et al., 2020).

As is common in unsupervised learning, a key challenge in mutational signature discovery is determining the number of items, in this case signatures, that might be present in a set of tumors. NMF-based approaches to mutational signature discovery have deployed several automated or semi-automated approaches to this challenge. SigProfilerExtractor carries out NMF on multiple bootstrapped replicates of the input matrix over a pre-defined

range of K (the number of signatures to discover). By default, it selects K based on a combination of (i) a silhouette analysis of the clustering of signatures discovered across the bootstrap replicates and (ii) the reconstruction errors – how well discovered signatures can reconstruct the input spectra.

In addition to NMF-based approaches, a few approaches to mutational signature discovery have been inspired by probabilistic topic modeling in document classification (Funnell et al., 2019; Roberts, 2018; Robinson, Sharan, & Leiserson, 2019). Probabilistic topic modeling is isomorphic to mutational signature discovery: documents correspond to tumors, topics correspond to mutational signatures, and occurrences of individual words in a text correspond to individual mutations in a tumor. In these abstract generative models, topics generate words in characteristic proportions just as mutational signatures generate mutations of different classes in characteristic proportions. Several approaches to probabilistic topic modeling require specification of the number of topics to infer, but hierarchical Dirichlet process (HDP) mixture models estimate the number of topics (or mutational signatures) automatically. Teh, Jordan, Beal, and Blei (2006) showed how to use HDP models for Bayesian estimation of topics and their associated words, and they implemented this approach in C and MATLAB. Roberts (2018) showed how to apply this approach to mutational signature discovery, re-implemented the MATLAB portion in R, and then tested this implementation on several mutational-signature-discovery tasks.

Unlike mutational discovery by NMF, signature discovery by HDP mixture models focuses on individual mutations. Mutations are assigned to clusters, a cluster contains a set of mutations that were generated by one mutational process, and the classes of the generated mutations in the cluster depend on multinomial draws from the mutational

signature of the process. In Teh and colleagues' implementation, the HDPs are represented by "Chinese restaurant franchises", and Gibbs sampling estimates the posterior distributions of (i) the number of mutation clusters and (ii) the assignment of observed mutations to the clusters. Because of the representation of HDPs as Chinese restaurant franchises, multiple clusters of mutations have similar proportions of counts across the mutation classes, i.e. similar "signature profiles". We combine mutation clusters with similar signature profiles both within and across Gibbs samples, and we interpret the signature profile of the combined cluster as the mutational signature that generated the mutations in the constituent clusters.

We used the C code from Teh et al. (2006) and built upon the R code from Roberts (2018) (https://github.com/nicolaroberts/hdp) by (i) reducing extremely high R memory usage in the representation of one of the prior distributions, (ii) correcting a problem with the timing of R garbage collection in the R / C interface, and (iii) providing built-in parallelization. Beyond these technical improvements, our conceptual contributions consist of reducing the number of false positives by adjusting some of the prior distributions, as described below, and a new approach to grouping mutation clusters with similar signature profiles that avoids grouping mutation clusters with dissimilar signature profiles. This approach stemmed from a more explicit conceptualization of the Gibbs samples as capturing the posterior distribution of the number of mutational processes (i.e., the number of mutational signatures) in addition to capturing the distribution of the proportions of mutations in each mutation class.

Here we describe this implementation, termed mSigHdp, and describe the results of benchmarking mSigHdp and 3 additional programs, SigProfilerExtractor, SignatureAnalyzer,

and signeR, on synthetic SBS and indel mutation data (Haradhvala et al., 2018; Islam et al., 2020; Kasar et al., 2015; Kim et al., 2016; Rosales, Drummond, Valieris, Dias-Neto, & da Silva, 2017). The 3 additional programs were those with the best F1 scores in Islam et al. (2020).

Method

Synthetic data generation

We generated synthetic SBS and indel mutation data based on the signature profiles from COSMIC (v3.2, https://cancer.sanger.ac.uk/signatures/) and on the numbers of mutations due to each signature in each tumor (the exposure matrix) in the "platinum" genomes in The ICGC TCGA Pan-Cancer Analysis Genomes Consortium (2020), as estimated by Alexandrov et al. (2020). The rows of an exposure matrix are signatures, and the columns are tumors. Each cell contains the number of mutations caused by a particular signature in a particular tumor.

For each cancer type, we took parameters for the synthetic data from the corresponding exposure matrix in Alexandrov et al. (2020) as follows:

Let p_s be the probability of occurrence of a signature, s, in a tumor of that cancer type. We estimate this from the proportion of tumors of that type in the exposure matrix that have the signature (i.e. have mutation counts > 0 for s in the tumor).

Let μ_s be the expected mean of the number of mutations due to s in the exposure matrix for those tumors of the given cancer type that have exposure to s.

Let $size_s$ be the negative-binomial dispersion parameter of the number of mutations due to s, estimated from the dispersion of the number of mutations due to s in the exposure matrix for those tumors of the given cancer type that have exposure to s.

We used the R package *fitdistrplus* (Delignette-Muller & Dutang, 2015) to get maximum likelihood estimates of parameters μ_s and $size_s$.

To gain insight into the sensitivity of methods to different levels of resampling noise in the synthetic input data, we generated 3 versions of each synthetic tumor, one without statistical noise, one with moderate resampling noise, and one with higher resampling noise, which we think better reflects variance in real data (Supplementary Figures S1, S2). To generate a single noiseless synthetic spectrum for one tumor type, we decide whether s is present by a random draw from $Bernoulli(p_s)$. If s is present, we draw the number of mutations m_s due to s from $NB(\mu_s, size_s)$ where NB denotes a negative-binomial distribution. For each signature, s, present in the tumor, we multiplied m_s by the signature profile vector to get the numbers of mutations in each mutation class, $m_{s,t}$, where t is a mutation class (e.g., CCA → CAA). Then, to generate the number of mutations in mutation class t in the full synthetic spectrum, we used $\sum_s m_{s,t}$ and rounded it to an integer.

To create a version of a synthetic tumor spectrum with moderate resampling noise, for each mutation class, t, and signature, s, we added negative-binomial noise with dispersion parameter 100 by drawing $m_{s,t,100}$, the number of mutations of class t due to s, from $NB\left(m_{s,t}, 100\right)$. To construct the synthetic spectrum, we set the number of mutations for each t as $\sum_s m_{s,t,100}$ and rounded it to an integer. The generation of a synthetic tumor spectrum with higher resampling noise was done analogously, but with negative-binomial dispersion parameter 30 for SBS data and of 10 for indel data.

We chose nine cancer types: Breast-AdenoCA, ColoRect-AdenoCA, Eso-AdenoCA, Kidney-RCC, Liver-HCC, Lung-AdenoCA, Ovary-AdenoCA, Skin-Melanoma and Stomach-AdenoCA to provide a broad representation of signatures in real data. The synthetic SBS data contain 9 cancer types and 23 COSMIC SBS signatures, and the synthetic indel data contain the same 9 cancer types and 11 COSMIC indel signatures (Supplementary Table S1). The synthetic data and the code to generate the data are at https://github.com/Rozen-Lab/Liu_et_al_Sup_Files.

Running programs for signature discovery

Code to run the signature discovery programs and assess their output is at https://github.com/Rozen-Lab/Liu_et_al_Sup_Files. See Supplementary Table S2 for program versions and arguments. Because all programs tested depend on random sampling, results often vary depending on the initial random seed. Therefore, we ran each program on 5 different seeds.

Assessing discovered signatures

We assessed how well the discovered signatures reflected the ground truth signatures in the synthetic mutation as follows. We first matched all the discovered signatures to the ground truth signatures using function TP_FP_FN_avg_sim (https://github.com/steverozen/ICAMSxtra). This function returns the number of true-positive discovered signatures (TP), the number of false-positive discovered signatures (FP), and the number of false negatives (FN, the ground truth signatures that were not discovered), and the average cosine similarity over the true positives. The function first computes cosine distances (1 – cosine similarity) and sets cosine distances > 0.1 to a large value, L. It then uses the "Hungarian algorithm" (Kuhn, 1955) to find a matching between

discovered and ground truth signatures that minimizes the total cosine distance. Matches with cosine distance L are then discarded, after which discovered signatures that match a ground truth signature are considered true positive matches, any remaining discovered signatures are considered false positives, and any ground truth signatures with no matching discovered signature are considered false negatives.

We then assess a set of discovered signatures by three measures:

1. Positive predictive value, $PPV \stackrel{\text{def}}{=} \frac{TP}{TP+FP}$
2. True positive rate, $TPR \stackrel{\text{def}}{=} \frac{TP}{TP+FN}$
3. The mean of the cosine similarities of true positive signatures to their matching ground truth signatures.

As a summary assessment, we define the *Composite Measure* as the sum of *PPV*, *TPR* and the mean cosine similarity.

mSigHdp implementation notes

The package hdpx (https://github.com/steverozen/hdpx/) contains the original C code from Teh et al. (2006) and much of the R code from Roberts (2018). However, in hdpx, the algorithm for finding sets of mutation clusters with similar signature profiles is completely new. We also added three diagnostic plots to help assess the reliability of each discovered signature. The first is a scatterplot of the proportion of mutations due to a particular signature versus the count of mutations due to the signature for each input spectrum (Supplementary Figure S3A). The second is a plot of the signature plus 5 input spectra that have the highest proportions of mutations due to that signature (Supplementary Figure S3B). We describe the third plot below. The package mSigHdp

(https://github.com/steverozen/mSigHdp/) provides functions for parallelizing and checkpointing the burnin and Gibbs sampling. Checkpointing allows one to extend the burnin if chains have not reached equilibrium distributions. mSigHdp uses the core R function parallel::mclapply for parallelizing over multiple threads, and because mclapply can only use 1 thread on Microsoft Windows, it is not practical to run mSigHdp on Windows.

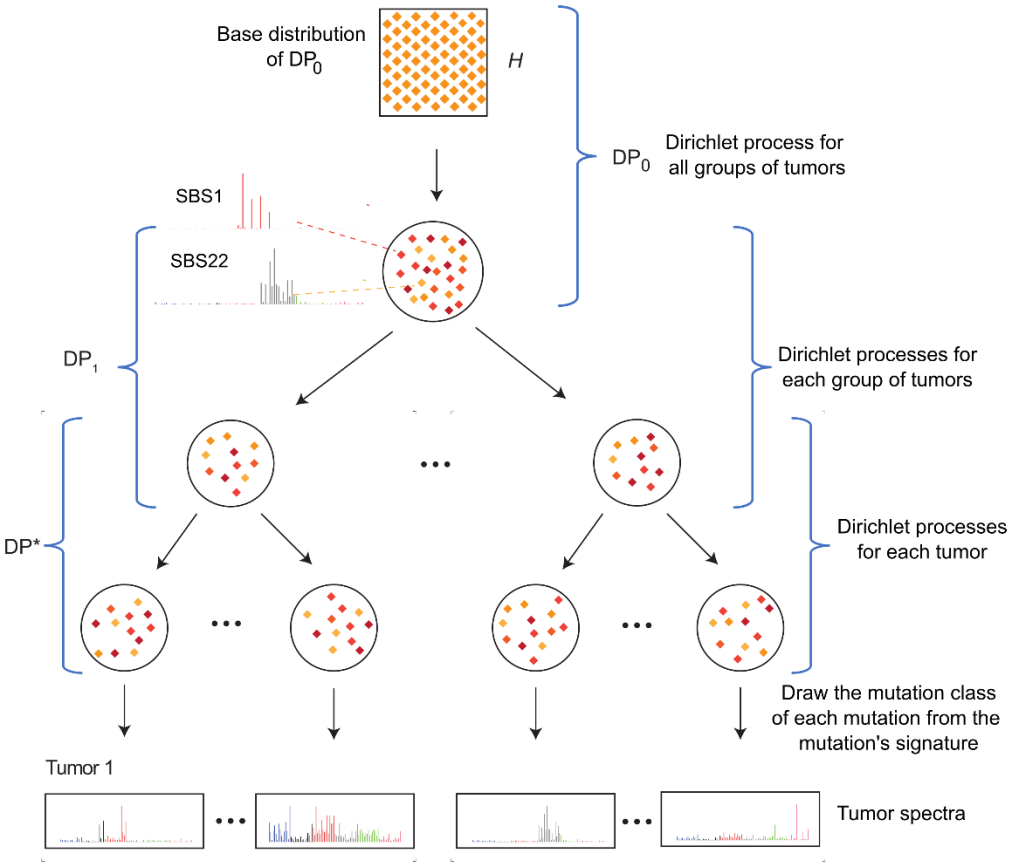

Figure 3 Schematic of a 3-layer HDP mixture model for mutational signature discovery. Please see the text for a full description. Each large circle represents a distribution over signatures in a Dirichlet process. H is the base distribution of Dirichlet process DP_0, DP_0 is the base distribution of DP_1, and DP_1 is the base distribution of DP^*. The small orange dots inside each circle represent individual signatures, with different shades of color representing the probability of that signature in the distribution. Two SBS signatures, SBS1 and SBS22, are shown at DP_0 as examples. Example observed spectra are shown at the bottom. mSigHdp also supports 2-layer HDP mixture models, in which all tumors have the same parent Dirichlet process, which has H as its base distribution.

Results

Discovering mutational signatures with hierarchical Dirichlet mixture models

Hierarchical Dirichlet process (HDP) mixture modeling and its application to mutational signature discovery are described in (Roberts, 2018; Teh et al., 2006). Figure 3 shows an HDP mixture model of mutations generated from mutational signatures. In this conceptualization, signatures model real-world mutational processes. To generate a mutation in, for example, tumor 1 (lower left), one first selects a mutational signature from tumor 1's Dirichlet process, DP*. This mutational signature is probably one already known to DP*, but it might be a new signature drawn from DP*'s base distribution, DP_1. Dirichlet process concentration parameters (not shown in Figure 3) partly govern whether an existing signature is selected or whether a new signature is drawn from the base distribution. If the signature is drawn from DP_1, the signature is also probably already known to DP_1, but if not, it is drawn from DP_1's base distribution, DP_0. If the signature is drawn from DP_0, it is once again probably already known to DP_0, but if not, it is drawn from DP_0's base distribution, H. H is a distribution of an infinite number of possible signatures that have a tendency to a moderate amount of spikiness, formally a symmetric Dirichlet distribution (not a Dirichlet process) with all concentration parameters = 0.5 (Teh et al., 2006). Once the signature of the mutation is known, the mutation's class is selected by a multinomial draw from the signature. (Recall that a signature is a multinomial probability vector.) The number of possible signatures in H is unbounded, but depending on the Dirichlet process concentration parameters, brand new signatures are rarely drawn from H.

Teh and colleagues represented a hierarchical Dirichlet process as a "Chinese restaurant franchise". Based on this representation, they devised and implemented a

Bayesian scheme whereby the posterior distribution of the assignment of mutations to clusters generated by a single signature could be calculated by Gibbs sampling. This scheme also allowed estimation of the concentration parameters of the Dirichlet processes in the model.

The Chinese restaurant franchise is a generative model. Each restaurant (tumor) contains a potentially unbounded number of tables and shares the menu of dishes with other restaurants. Dishes represent mutational signatures (which model mutational processes and the characteristic frequencies at which the processes generate mutations of different classes). Each table is served only 1 dish (signature) but each dish can be served at multiple tables. To generate a new mutation in a tumor – represented by a customer (mutation) arriving at a restaurant – the customer is assigned to either (1) a table with other customers already sharing a single dish or (2) an empty table for which a dish will be selected. The choice of an existing table or a new table is governed by a Dirichlet process, including its concentration parameter. If the customer was assigned to an empty table, the dish for that table will be either an existing dish (a mutational signature that has already been generated) or a new dish (one that is not served at any other table), corresponding to a new mutational signature, again according to a Dirichlet process.

The Gibbs sampling scheme estimates the posterior distribution of the assignment of customers to tables (which also defines the assignment of mutations to signatures, since each table has a single signature). From the signature, one can also estimate the distribution of the classes of the individual mutations (e.g., CCA → CAA). Each Gibbs sample contains several items, among which we focus on the number of mutations of each mutation class associated with each signature across all tables in all restaurants; here we will call these

“mutation clusters”. Teh et al. (2006) termed this a “cluster of category counts” in, where their “category” corresponds to “mutation class” here. In the software, these are represented as matrices in which rows are mutation classes and columns represent different mutation clusters due to different signatures.

A minor issue with this representation is that because the same “dish” (mutational signature) can be served at multiple tables even in one restaurant, we need to combine mutation clusters with nearly identical “profiles” – the proportions of mutations in each mutation class. We accomplish this by class-wise summation of mutation counts. A more challenging issue is that the Gibbs samples represent signatures by their realizations as mutation clusters in many Gibbs samples of the posterior distribution (Figure 4A). These realizations must be combined to derive point estimates of the signature profiles while noting which signature profiles are common across the Gibbs samples and therefore well supported.

The approach in Roberts (2018), Section 4.2.2, “Extracting Consensus Signatures”, was fairly involved, and in our tests was prone to grouping of mutation clusters (analogous to the groups in Figure 4B) that had dissimilar signature profiles (i.e., dissimilar proportions of mutations in each mutation class). After experimenting with several clustering techniques, we settled on divisive hierarchical clustering (Kaufman & Rousseeuw, 2005)(function diana in R package cluster, https://cran.r-project.org/package=cluster) to find groups of mutation clusters with similar signature profiles (Figure 4B).

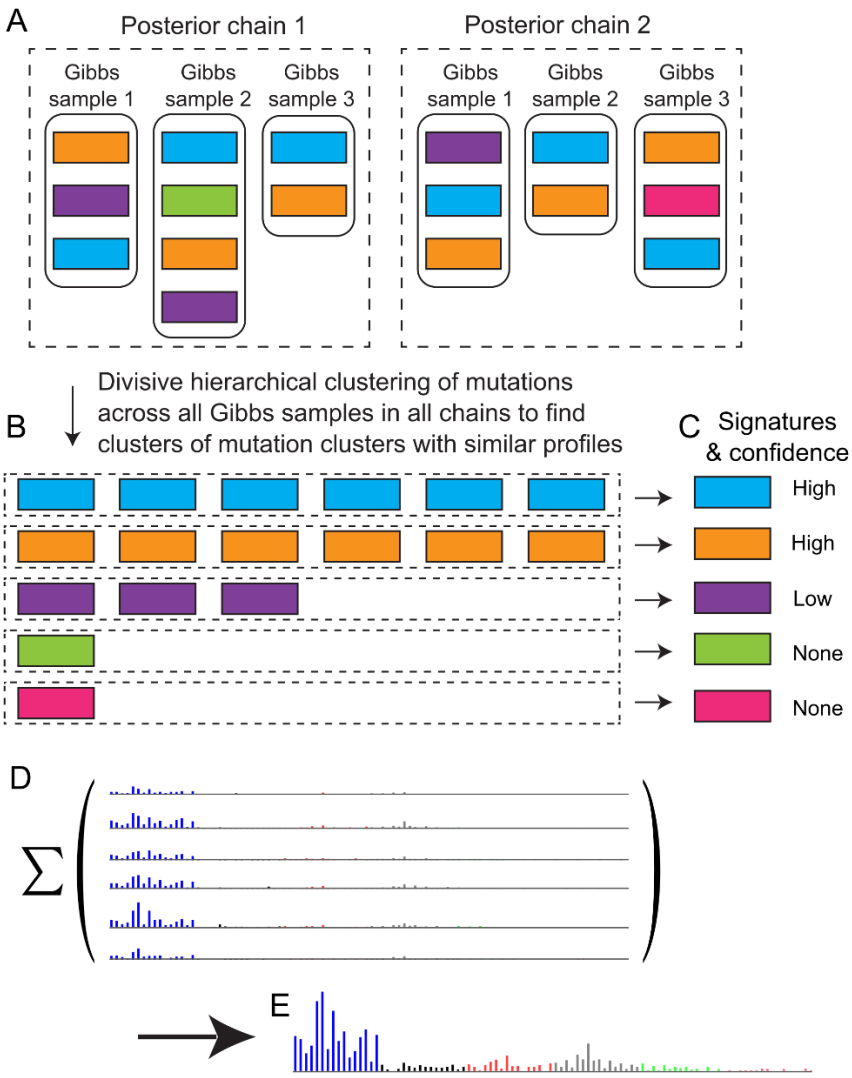

Figure 4 Combining and interpreting mutation clusters collected by Gibbs sampling. **(A)** Gibbs samples from 2 posterior chains in highly simplified form. Each rectangle represents a mutation cluster, and mutation clusters with similar profiles (similar proportions of counts of each mutation class) have the same color. A typical configuration would be 20 posterior chains in parallel, with 200 Gibbs samples per chain and with 100 iterations between every 2 Gibbs samples. (**B)** Each row (dashed rectangle) is a group of mutation clusters with similar signature profiles (proportions of mutations in each mutation class) as identified by divisive hierarchical clustering across all Gibbs samples from all sampling chains. We interpret these groups as containing mutation clusters from one signature. **(C)** To infer signatures, in each group of mutation clusters (dashed rectangles in panel B), the counts of mutations of each class are summed, and then these counts are divided by the total number of mutations to yield mutation signatures. **(D, E)** One row of B

and C (e.g. the row of blue rectangles) in detail. Each blue rectangle in B is a mutation cluster that is represented by its mutational spectrum in D. The height of each bar in the spectrum indicates the number of mutations of a particular class. Panel E shows the results of summing the mutations of each mutation class over the mutational spectra in D. Dividing by the total number of mutations (not shown) yields the signature profile, corresponding to the bule rectangle in C. We have high confidence in signatures that are represented by mutation clusters in all or most of the Gibbs samples, i.e., that are common in the posterior distribution. In this study and generally, we use a cutoff of 90% of Gibbs samples as indicating high confidence.

In Figure 4C, the “blue” and “orange” signatures derived from mutation clusters found in all 6 posterior samples in Figure 4A. Because these are samples from the posterior distribution of mutation clusters generated by a mutational processes, this indicates strong evidence for the existence of the process generating the “blue” and “orange” mutational signatures. By contrast, the “green” and “pink” mutational signatures in panel 4C derive from mutation clusters present in only 1 posterior sample. Thus, there is little evidence for these mutational signatures. mSigHdp provides a diagnostic plot that shows which Gibbs samples contributed to each signature (Supplementary Figure S4). Finally, we take the signature profiles of the mutation-class-wise sums of mutations in all mutation clusters in each group in Figure 4B as the estimate of the underlying mutational signature (Figure 4C,D,E, Supplementary Figure S5).

The HDP mixture model’s estimate of the number of signatures is influenced by the prior gamma distributions of the Dirichlet-process concentration parameters. The implementation in (Roberts, 2018) appeared to have a high false discovery rate on real data, with a PPV of roughly 0.5 as estimated from Robert’s Figures 4.13 and D.17. Roberts (2018) tested different gamma shape parameters, but all with a gamma beta (rate) of 1. Based on subsequent analysis, we believe these parameters would favor a posterior distribution with too many signatures. (The high false discovery rates might also have stemmed partly from the algorithm for finding groups of mutation clusters due to one signature, which sometimes combined mutation clusters with dissimilar signature profiles.) In any case, for mSigHdp, the prior distributions of the Dirichlet-process concentration parameters strongly influences the number of signatures discovered and therefore the false discovery rate (Liu, 2021). Based on previous analysis on a different set of synthetic data (Liu, 2021), for the SBS signature discovery results here, we used a much more stringent beta parameter of 20. (The

location of the gamma distribution decreases as beta increases. The mean of a gamma distribution with shape 1 and beta 1 is nearly 1, and for shape 1 and beta 20 the mean is ~0.05. So compared to a beta of 1, the beta of 20 encourages smaller concentration parameters, which shifts the posterior distribution of the number of signatures to lower values i.e., fewer signatures.) For signature discovery in the indel data, we tested several values of beta on the synthetic indel data used here, and we a beta of 50 (Supplementary Figure S6).

Discovering SBS signatures

We assessed how well mSigHdp and the 3 other programs discovered SBS mutational signatures according to several measures (Figure 5, Table 1). To assess the stability of the programs' results, we ran each program with five different random seeds. For the realistic data, the Composite Measures of mSigHdp and SigProfilerExtractor were very similar, while the Composite Measures of the other 2 programs were substantially lower. Cosine similarities varied little across all 4 programs, and therefore did not strongly influence the Composite Measures. For synthetic data sets with resampling noise, all programs generated different results with different starting random seeds.

Despite similar overall Composite Measures, mSigHdp and SigProfilerExtractor had notably different strengths. Broadly, SigProfilerExtractor had more false negatives (lower TPR) and mSigHdp had more false positives (lower PPV). Among the false negatives in the realistic data, SigProfilerExtractor failed to discover SBS5 and SBS16 at all 5 random seeds and at 4 seeds failed to discover SBS12 (Supplementary Table S3). The reason seems to be that SigProfilerExtractor did not separate these signatures but instead combined them in a false positive signature that is mixture of these signature (Supplementary Figure S7).

SigProfilerExtractor also missed SBS29, SBS35, SBS38, SBS41, and sometimes SBS22 in runs with realistic noise (Supplementary Table S3). In contrast, mSigHdp discovered all these signatures except SBS29 in all runs. SigProfilerExtractor's false-negative signatures tended to be those present in smaller proportions of the synthetic tumors, with median proportions of 0.0278 for false negatives versus 0.194 for true positives ($p < 0.016$ by two-sided Wilcoxon rank-sum test, Supplementary Table S4).

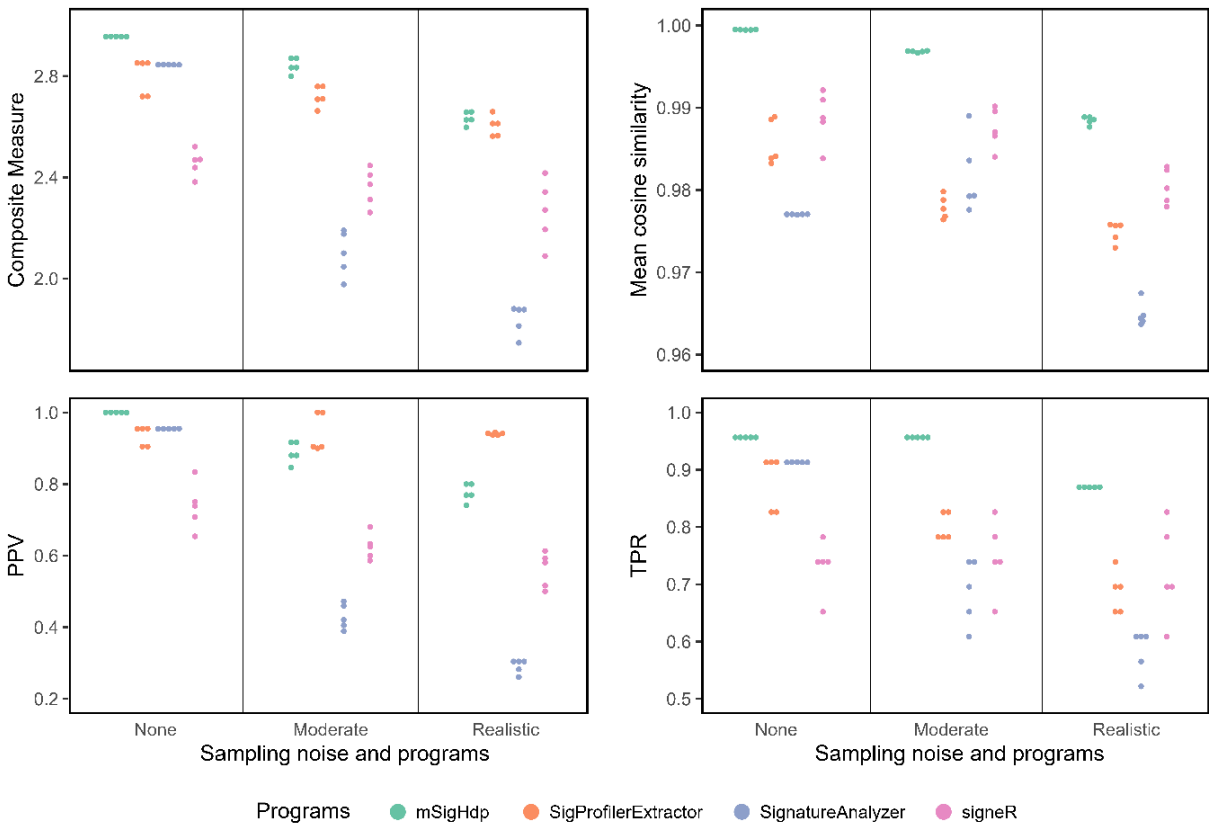

Figure 5 Results on synthetic SBS data. Composite Measure is the sum of PPV (positive predictive value), TPR (true positive rate) and mean cosine similarity. Each dot represents one measure for one program for one random seed. Sampling noise refers to sampling variance introduced into the synthetic data. On the *x* axis, "Moderate" and "Realistic" indicate sampling variance from a negative-binomial distributions with size parameters of 100 and 30, respectively. The rankings of the 4 programs were similar at all 3 levels of

sampling noise, with the notable exception that mSigHdp and SigProfilerExtractor were essentially tied on the Composite Measure at realistic sampling noise.

Noise	Approach	Mean PPV	Mean TPR	Mean cosine similarity	Composite Measure	
					Mean	SD
None	mSigHdp	**1.00**	**0.96**	**0.999**	**2.96**	**0.00**
	SP	0.94	0.88	0.986	2.80	0.07
	SA	0.96	0.91	0.977	2.84	**0.00**
	signeR	0.74	0.73	0.989	2.46	0.05
Moderate	mSigHdp	0.89	**0.96**	**0.997**	**2.84**	**0.03**
	SP	**0.94**	0.80	0.978	2.72	0.04
	SA	0.43	0.69	0.982	2.10	0.09
	signeR	0.63	0.75	0.987	2.36	0.07
Realistic	mSigHdp	0.78	**0.87**	**0.988**	**2.63**	**0.03**
	SP	**0.94**	0.69	0.975	2.60	0.04
	SA	0.29	0.58	0.965	1.84	0.06
	signeR	0.56	0.72	0.980	2.26	0.13

Table 1 Results on synthetic SBS data. Means are over the 5 different random seeds. Bold text indicates the best results in each column. PPV: positive predictive value, SA: SignatureAnalyzer, SD: standard deviation over the 5 random seeds, SP: SigProfilerExtractor, TPR: true positive rate.

SigProfilerExtractor may have traded relatively high false negative rates in exchange for very low false positive rates. Supporting this hypothesis, the signatures that SigProfilerExtractor discovered at the K that it had selected $+2$ ($K + 2$ signatures total) included an additional false negative and an additional false positive (Supplementary Table S5).

Unlike SigProfilerExtractor, mSigHdp often had false-positive signatures that consisted primarily of a single mutation class; some of these may have compensated for variability in the relative mutation counts in particular mutation classes (Supplementary

Figure S8). In addition, mSigHdp discovered neither SBS7a nor SBS7b in the data with realistic noise (Supplementary Table S3). We believe that part of the issue with these two signatures is that while the negative-binomial size parameter of 30 seems to generate realistic reconstruction errors for most tumor types, for the skin melanoma tumors (which are dominated by SBS7a and SBS7b), it generates reconstruction errors that are too large (Supplementary Figure S9). In synthetic data with a negative-binomial size parameter of 100 (less dispersed data), mSigHdp discovered SBS7a with high cosine similarity, but it still failed to discover SBS7b and generated some false positives probably related to these signatures (Supplementary Table S7). SigProfilerExtractor also had difficulty with SBS7a: the extracted signature was quite divergent from SBS7a in data with noise (average cosine similarity 0.909 [moderate] and 0.928 [realistic], Supplementary Figure S10).

Discovering indel signatures

We assessed how well mSigHdp and the 3 other programs discovered indel mutational signatures according to the same measures used for assessing discovery of SBS signatures (Figure 6, Table 2). We focus on the results for the realistic level of sampling noise. mSigHdp had the highest Composite Measure (mean 2.91 versus 2.75 for SigProfilerExtractor, $p = 0.012$, two-sided Wilcoxon rank-sum test over the results of the 5 different random seeds), while the other two programs had substantially lower Composite Measures. As with the SBS data, the PPV and TPR were the main contributors to differences in the Composite Measure across the 4 programs, while average cosine similarities were similar for the 4 programs. SigProfilerExtractor and mSigHdp had similar PPVs. SigProfilerExtractor had 0 or 1 false discovery, which was a signature similar to ID5 but with one mutation class missing and excessively high proportions of 3 other mutation classes (Supplementary Table S6,

Supplementary Figure S11). mSigHdp had 0, 1, or 2 false discoveries. One of these occurred in 4 runs and might reflect high variation in one mutation class stemming from sampling noise in other signatures. The other false positive occurred in 1 run for reasons that are not clear.

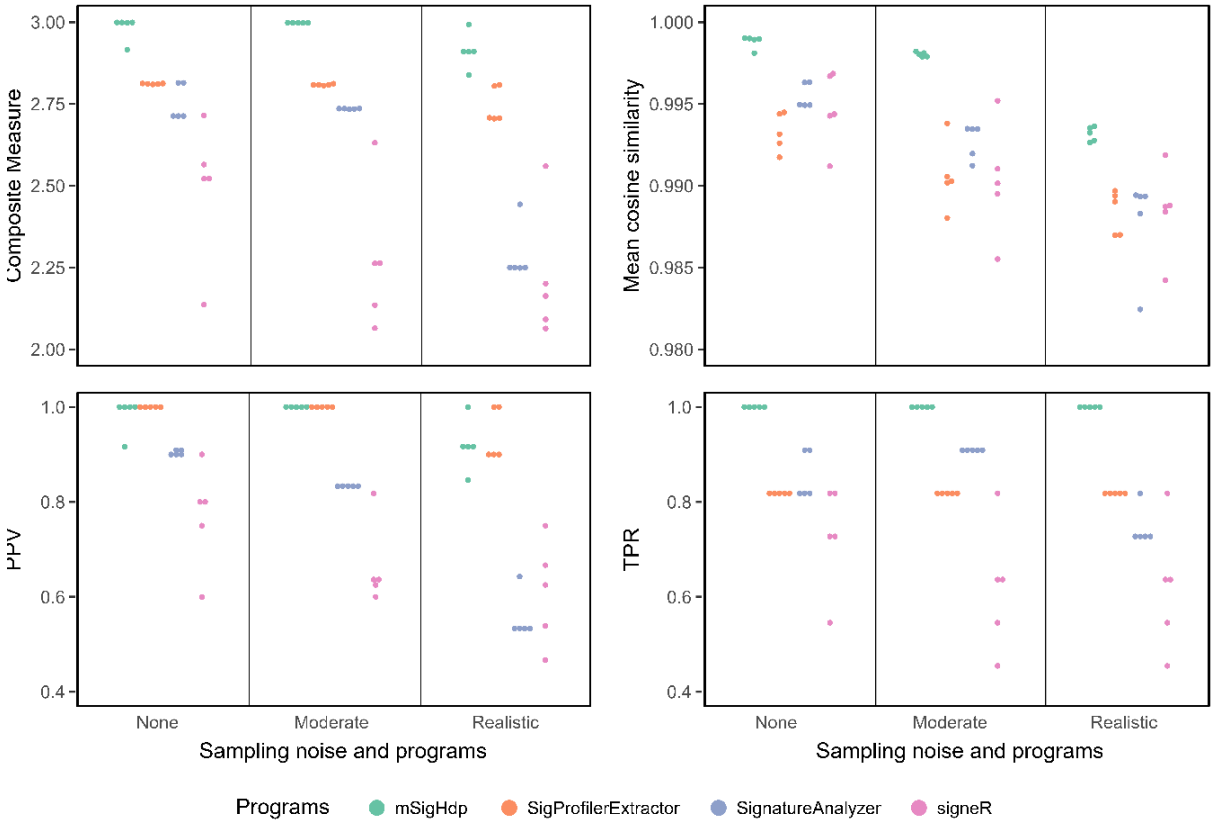

Figure 6 Results on synthetic indel data. Composite Measure is the sum of PPV (positive predictive value), TPR (true positive rate) and mean cosine similarity. Each dot represents one measure for one program for one random seed. Sampling noise level refers to sampling variance introduced into the synthetic data. On the x axis, “Moderate” indicates sampling variance from a negative-binomial distribution with size parameter of 100. “Realistic” indicates sampling variance from a size parameter of 10. The rankings of the 4 programs were similar at all 3 levels of sampling noise.

Noise	Approach	Mean PPV	Mean TPR	Mean cosine similarity	Composite Measure	
					Mean	SD
None	mSigHdp	0.98	**1.00**	**0.999**	**2.98**	0.04
	SP	**1.00**	0.82	0.993	2.81	**0.00**
	SA	0.90	0.86	0.995	2.75	0.06
	signeR	0.77	0.73	0.995	2.49	0.21
Moderate	mSigHdp	**1.00**	**1.00**	**0.998**	**3.00**	**0.00**
	SP	**1.00**	0.82	0.991	2.81	**0.00**
	SA	0.83	0.91	0.993	2.74	**0.00**
	signeR	0.66	0.62	0.99	2.27	0.22
Realistic	mSigHdp	0.92	**1.00**	**0.993**	**2.91**	**0.05**
	SP	**0.94**	0.82	0.988	2.75	**0.05**
	SA	0.56	0.75	0.988	2.29	0.09
	signeR	0.61	0.62	0.988	2.22	0.20

Table 2 Results on synthetic indel data. Means are over the 5 different random seeds. Bold text indicates the best results in each column. PPV: positive predictive value, SA: SignatureAnalyzer, SD: standard deviation over the 5 random seeds, SP: SigProfilerExtractor, TPR: true positive rate.

In terms of false negatives, SigProfilerExtractor consistently failed to discover indel signature ID11, and sometimes failed to discover indel signatures ID5 or ID13 (Supplementary Table S3). ID11 and ID13 are rare in the synthetic mutation, data, while ID5 is common (Supplementary Table S4). Unlike the SBS signatures, the indel signatures discovered by SigProfilerExtractor at $2 +$ the selected K included 2 additional false positives and the same number of false negatives (Supplementary Table S5). Thus, it does not seem that the indel false negatives mainly stemmed from incorrect estimates of K, the number of signatures. For mSigHdp, there were no false-negative indel signatures.

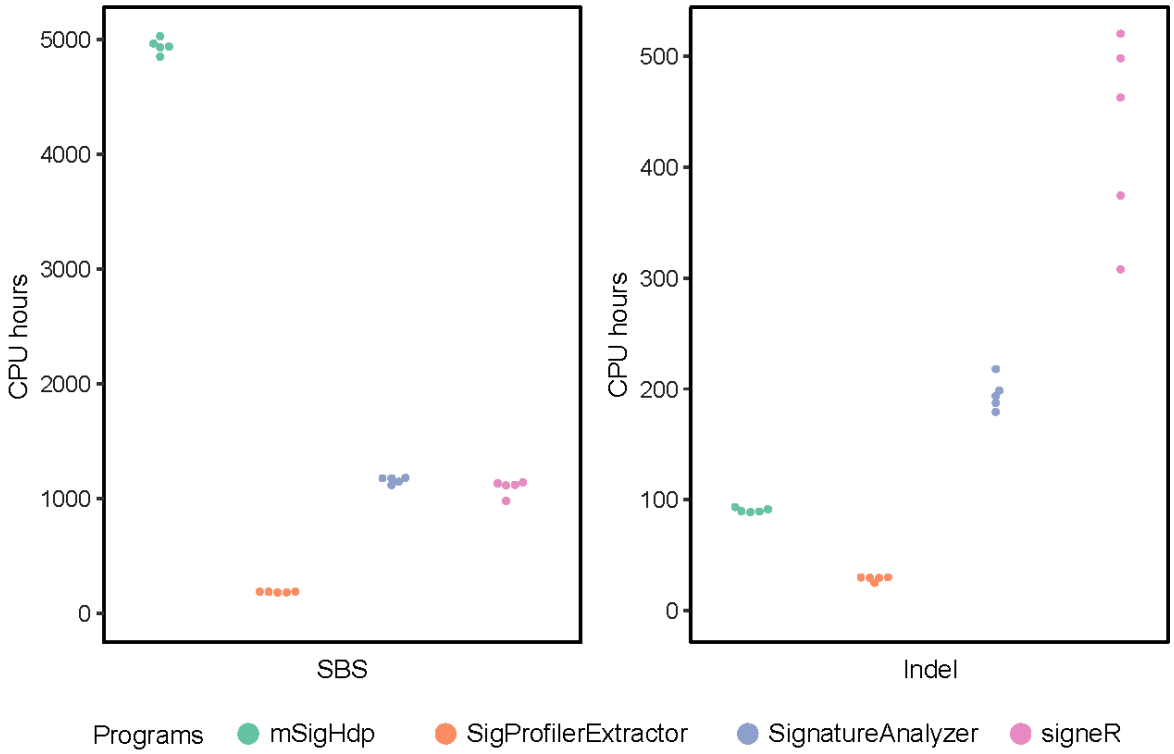

Figure 7 CPU usage on the realistic synthetic data sets. CPU usage on AMD EPYC 7552 processors running at 3170 MHz, measured by the times (Python) or system.time (R) functions. Times include user and system times for parent and child processes.

CPU usage

For the realistic synthetic SBS data, mSigHdp required much more CPU time than the other programs. For the realistic synthetic indel data, mSigHdp used more CPU than SigProfilerExtractor, but substantially less than the other 2 programs (Figure 7). The difference in mSigHdp's CPU usage between SBS and indel data seems partly due to the difference in the number of mutations in the two data sets (1.5×10^7 versus 2.3×10^6), since the time needed for each iteration of the Gibbs sampler depends partly on the number of mutations in the input data set. In addition, we used a longer burnin for SBSs

than for the indels data (100,000 versus 30,000 iterations). For both data sets the subsequent sampling used 20,000 iterations (200 samples, one sample every 100 iterations).

Discussion

Here, we have described mSigHdp, software that uses hierarchical Dirichlet process mixture models to discover mutational signatures. We evaluated the ability of mSigHdp and 3 other programs to discover mutational signatures in synthetic SBS and indel mutation data. In the SBS, mSigHdp and SigProfilerExtractor were substantially better able to discover signatures than the other 2 programs. mSigHdp and SigProfilerExtractor had different strengths, with mSigHdp less susceptible to false negatives and SigProfilerExtractor less susceptible to false positives (Figure 5, Table 1). mSigHdp was better able to discover rare signatures, which may be an advantage, since most common signatures have probably already been discovered (Supplementary Tables S3,S4). In indel mutation data, mSigHdp was better able to discover signatures than the other 3 programs (Figure 6, Table 2).

While there has been some systematic benchmarking of mutational signature discovery methods on SBS data (Islam et al., 2020; Omichessan, Severi, & Perduca, 2019; Wu, Chua, Ng, Boot, & Rozen, 2022), we are unaware of previous benchmarking efforts on indel data. Based on analysis of reconstruction errors, we believe that the amount of resampling noise in the synthetic SBS and indel mutation data sets used here more closely resembles variability in real data than some previously used benchmarking data (Supplementary Figures S1,S2). Thus, the synthetic mutation data sets here, and especially the indel data set, are resources for developers of mutational-signature discovery methods.

The results presented here also lead to some practical suggestions for mutational-signature discovery, especially now that many mutational signatures are well established

(Alexandrov et al., 2020)(https://cancer.sanger.ac.uk/signatures/). A common application would be *de novo* analysis of ≥ 100 mutational spectra to understand what signatures are present and discover possible new signatures. Signature discovery in this context is not a lightweight or purely algorithmic endeavor. Indeed, in practice, signature discovery has often depended on analyses by multiple programs run on various different subsets of the input data and on human judgement drawing on many sources (Alexandrov et al., 2020). It would be wise to do several independent analyses, since all methods studied here discovered slightly different sets of signatures when run with different starting random seeds (Figures 5, 6). Certainly, one must consider the totality of evidence, including known signatures, to decide which potentially novel discovered signatures are true positives. Alexandrov et al. (2020) present a more extensive discussion of the criteria that can be applied to this decision.

In conclusion, the results of benchmarking on realistic data here indicate that mSigHdp is an advance in the state of the art for discovering rare SBSs mutational signatures and for discovering indel signatures.

Competing interests Y.W. and S.G.R. are authors on Islam et al. (2020). The other authors declare no competing interests.

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

Acknowledgements We thank the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes

Network for making the mutational spectra, exposures, and signatures available.

References

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., . . . PCAWG Consortium. (2020). The repertoire of mutational signatures in human cancer. *Nature, 578*(7793), 94-101. doi:10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., . . . Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature, 500*(7463), 415-421. doi:10.1038/nature12477

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep, 3*(1), 246-259. doi:10.1016/j.celrep.2012.12.008

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software, 64*(4), 1 - 34. doi:10.18637/jss.v064.i04

Funnell, T., Zhang, A. W., Grewal, D., McKinney, S., Bashashati, A., Wang, Y. K., & Shah, S. P. (2019). Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput Biol, 15*(2), e1006799. doi:10.1371/journal.pcbi.1006799

Haradhvala, N. J., Kim, J., Maruvka, Y. E., Polak, P., Rosebrock, D., Livitz, D., . . . Getz, G. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun, 9*(1), 1746. doi:10.1038/s41467-018-04002-4

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature, 578*(7793), 82-93. doi:10.1038/s41586-020-1969-6

Islam, S. M. A., Wu, Y., Díaz-Gay, M., Bergstrom, E. N., He, Y., Barnes, M., . . . Alexandrov, L. B. (2020). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor, doi:10.1101/2020.12.13.422570. *bioRxiv*. doi:10.1101/2020.12.13.422570

Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., . . . Brown, J. R. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun, 6*, 8866. doi:10.1038/ncomms9866

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data An Introduction to Cluster Analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Kim, J., Mouw, K. W., Polak, P., Braunstein, L. Z., Kamburov, A., Kwiatkowski, D. J., . . . Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet, 48*(6), 600-606. doi:10.1038/ng.3557

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly, 2*(1-2), 83-97.

Liu, M. (2021). *Computational Approaches to Studying the Causes and Consequences of Mutational Signatures.* (PhD). Duke-NUS Medical School, National University of Singapore, Retrieved from https://scholarbank.nus.edu.sg/handle/10635/213416

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., . . . Breast Cancer Working Group of the International Cancer Genome, C. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell, 149*(5), 979-993. doi:10.1016/j.cell.2012.04.024

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., . . . Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature, 534*(7605), 47-54. doi:10.1038/nature17676

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. doi:10.1371/journal.pone.0221235

Roberts, N. D. (2018). *Patterns of somatic genome rearrangement in human cancer.* (PhD Thesis). Cambridge University, Cambridge, England, United Kingdom. Retrieved from https://www.repository.cam.ac.uk/bitstream/handle/1810/275454/Roberts-2018-PhD.pdf

Robinson, W., Sharan, R., & Leiserson, M. D. M. (2019). Modeling clinical and molecular covariates of mutational process activity in cancer. *Bioinformatics, 35*(14), i492-i500. doi:10.1093/bioinformatics/btz340

Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E., & da Silva, I. T. (2017). signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics, 33*(1), 8-16. doi:10.1093/bioinformatics/btw572

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association, 101*(476), 1566-1581. doi:10.1198/016214506000000302

The ICGC TCGA Pan-Cancer Analysis Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature, 578*(7793), 82.

Wu, Y., Chua, E. H. Z., Ng, A. W. T. N., Boot, A., & Rozen, S. G. (2022). Accuracy of mutational signature software on correlated signatures. *Scientific Reports, 12*(1), 390. doi: 10.1038/s41598-021-04207-6