Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance

Omichessan Hanane (1,2), Severi Gianluca (1,2,3), Perduca Vittorio (4,1)

1. CESP (UMR INSERM 1018), Université Paris-Saclay, UPS, UVSQ, Villejuif Cedex, F-94805, France
2. Gustave Roussy, Villejuif, F-94805, France
3. Cancer Epidemiology Centre, Cancer Council Victoria, and Centre for Epidemiology and Biostatistics, Melbourne School for Population and Global Health, The University of Melbourne, Australia
4. Laboratoire de Mathématiques Appliquées – MAP5 (UMR CNRS 8145), Université Paris Descartes, 75006, Paris, France

Abstract

Mutational signatures refer to patterns in the occurrence of somatic mutations that reflect underlying mutational processes. To date, after the analysis of tens of thousands of genomes and exomes from about 40 different cancers types, 30 mutational signatures characterized by a unique probability profile across the 96 mutation types have been identified, validated and listed on the COSMIC (Catalogue of Somatic Mutations in Cancer) website. Simultaneously with this development, the last few years saw the publication of several concurrent methods (mathematical algorithms implemented in publicly available software packages) for either the quantification of the contribution of prespecified signatures (e.g. COSMIC signatures) in a given cancer genome or the identification of new signatures from a sample of cancer genomes. A review about existing

computational tools has been recently published to guide researchers and practitioners in conducting their mutational signatures analysis, however, other tools have been introduced since its publication and, to date, there has not been a systematic evaluation and comparison of the performance of such tools. In order to fill this gap, we carry on an empirical evaluation study of all available packages to date, using both real and simulated data.

Keywords: *DNA; sequencing; cancer; mutational signatures; decomposition algorithms; probabilistic models; simulation study*

Introduction

It is well-known and established that cancers are the results of changes that have occurred in the DNA sequence of the genomes of cancer cells [1]. Such changes that include Single Nucleotide Variants (SNVs), insertions or deletions, Copy Number Variation (CNVs) and chromosomal aberrations, have been termed as "somatic mutations" to distinguish them with those inherited and transmitted from parents (germline mutations). With the development and the improvement of sequencing technologies collectively referred to as *High-Throughput Sequencing (HTS)* and the availability of cancer genomes and exomes data from most human cancers, much has been learnt about these mutations and the set of genes (oncogenes and tumour suppressor genes) that operate in human cancers. Beyond the relatively small number of “driver” mutations, those that confer a selective advantage for tumour development and progression, the vast majority of mutations, so-called “passenger” mutations, are “neutral” that means they do not affect cancer cells’ survival. Together, passenger and driver mutations constitute a record of all cumulative DNA damage and repair activities occurred during the cellular lineage of the cancer cell [2].

Among all somatic mutations, a particular focus has been placed on substitutions which have been classified in six types according to the mutated base, which by convention and to avoid redundancy due to the complementary strand, is always a pyrimidine (C or T). Such 6 substitutions (C>A, C>G, C>T, T>A, T>C and T>G) may be further classified in different types when considering the sequence pattern in which they are located (sequence context). For practical reasons, the sequence context is defined using the 5' and 3' bases proximal to the mutated base, that results in substitutions being classified in 96 types ($6 * 4 * 4$). In this paper we adopt this convention and use "mutations" to refer to these 96 types of contextualized substitutions. Accordingly, the mutational catalogue or spectrum in a cancer genome can be defined as a probability distribution over these 96 types of mutations, giving the relative frequency of each type with respect to the total mutational burden. [2,3] .

It has been hypothesised that specific patterns of mutations are the results of the action of particular underlying mutational process. To identify such patterns, computational models, such as matrix decomposition algorithms or probabilistic models, have been developed. The first one was published in 2013 by Alexandrov and colleagues [2,3], and, as for most of all the other models that followed, is based on the idea that a mutational signature can be seen as a probability distribution of the 96 types of mutations, each contributing to the final catalogue.

Such model, applied to tens of thousands of genomes and exomes from 40 different cancers types from large data repositories such as TCGA (The Cancer Genome Atlas), has led to the identification of 30 mutational signatures characterized by a unique probability profile across the 96 mutation types. These validated mutational signatures are listed on the COSMIC (Catalogue Of Somatic Mutations In Cancer) website [4].

In the following years several other mathematical methods and computational tools have been proposed to detect mutational signatures and estimate their contribution to a given catalogue. These methods can be grouped in two categories with different goals. The first class of methods aims to discover novel signatures while the second class aims to detect the known and validated mutational signatures in the mutational catalogue of a given sample. The approaches used in the first class are referred to as "*de novo*" while those in the second class as "refitting". All methods have been implemented in open source tools, mainly R packages, but some of them are available through command line, the Galaxy project or a web interface.

Signatures identified with de novo methods can be compared to reference signatures (i.e. those listed in COSMIC) through measures such as cosine [5] or bootstrapped cosine similarity [6], which is a distance metric between two non-zeros vectors. The issue around the results from these comparisons is the choice of a cut-off, which implies the level at which two signatures will be considered as similar or not [3,7]

Recently, [8] published a paper that has been limited to the summary of current methodologies, in particular the mathematical models and computational techniques, which form the basis of mutational signature analysis. As rightly pointed out in [9,10], the problem of identifying mutational signatures from cancer genomes is a classification task, for which there is no guarantee of finding the underlying truth because of its unsupervised learning nature. Actually, there has not been a systematic evaluation of the performance of the existing methods and the problem of choosing an appropriate cosine similarity cut-off has not been addressed so far.

This work was designed to evaluate and compare the different methods as date, using simulated as well as real-world cancers data, and assessed the ability of each tools to properly detect the known mutational signatures underlying a mutational catalogue.

Mathematical definition of mutational catalogues and signatures

In this section we briefly review the mathematical framework and notations based on the original model developed by Alexandrov and colleagues[2]. The data structure that is used as input by most current methods is the *mutational catalogue,* or *spectrum,* of a genome (or exome), with the counts for each of the 96 different mutation *features,* or *types.* In most approaches, the notion of mutation type refers to base substitutions within the trinucleotide context described above.

The mutational catalogue of a genome (or exome) g is defined as the vector $(m_g^1, \dots, m_g^K)^T$, where each m_g^k is the number of mutations of type k found in the genome and K, the number of possible mutation types, is equal to 96. Note that in this setting, information about mutation locations in the sequence is lost. The catalogue is built by comparing the sequence to a reference sequence in order to detect mutations and then by simply counting the occurrences of each type. The reference sequence can either be a standard reference (e.g. the assembly GRCh38) or a sequence from a "normal" tissue from the same individual (e.g. DNA from blood or from normal tissue surrounding tumours when available). For the purposes of the present work we will use the generic term "samples" for both genomes and exomes as the concepts and models used may be applied to both.

The basic idea underlying all computational models proposed is that the mutational catalogue of a sample results from the combination of all the mutational processes operative during lifetime, and therefore it can be seen as the weighted superposition of simpler mutational *signatures,* each corresponding to a specific process. The weight is larger if the process has a larger role in the final catalogue of mutations: for example, for exposures to mutagens that last longer and/or are more intense.

Formally, the signature of a mutational process n is a vector $p_n = (p_n^1, \dots, p_n^K)^T$, where each p_n^k represents the probability that the mutational process will induce a mutation of type k. In other words, p_n^k is the expected relative frequency of type k mutations in genomes exposed to n. Note that $\sum_{k=1}^{K} p_n^k = 1$ and $0 \leq p_n^k \leq 1$ for all k.

The intensity of the exposure to a mutational process n in a sample g is measured by the number of mutations e_g^n in g that are due to n. For this reason, e_g^n is referred to as the "exposure" of g to n. The expected number of mutations of type k due to the process n in sample g is therefore $p_n^k e_g^n$. If sample g has been exposed to N mutational processes, then the total number of mutations of type k is

$$m_g^k = \sum_{n=1}^{N} p_n^k \, e_g^n + \epsilon_g^k, \qquad (1)$$

where ϵ_g^k is an error term reflecting sampling variability and non-systematic errors in sequencing and/or subsequent analysis.

Matrix notation is effectively used when dealing with several genomes and signatures. In this situation, the collection of samples G is represented by the $K \times G$ matrix, with catalogues in columns:

$$M = \begin{pmatrix} m_1^1 & m_2^1 & \dots & m_G^1 \\ \vdots & \vdots & & \vdots \\ m_1^K & m_2^K & \dots & m_G^K \end{pmatrix}, \qquad (2)$$

the N signatures are represented by the $K \times N$ matrix

$$P = \begin{pmatrix} p_1^1 & p_2^1 & \dots & p_N^1 \\ \vdots & \vdots & & \vdots \\ p_1^K & p_2^K & \dots & p_N^K \end{pmatrix}, \qquad (3)$$

and the exposures by the $N \times G$ matrix

$$E = \begin{pmatrix} e_1^1 & e_2^1 & \dots & e_G^1 \\ \vdots & \vdots & & \vdots \\ e_1^N & e_2^N & \dots & e_G^N \end{pmatrix}. \qquad (4)$$

Equation (1) then becomes

$$M \approx P \times E$$

where we omitted the error term.

***de novo* extraction vs. refitting**

de novo signature extraction methods aim at estimating P and E given M. Non-negative matrix factorization (NMF) is an appealing solution to this non-supervised learning problem, because, by definition, all involved matrices are non-negative [11]. NMF identifies two matrices P and E that minimize the distance between M and $P \times E$. In particular, NMF finds an approximated solution to the non-convex optimization problem

$$\underset{P \geq 0,\, E \geq 0}{\operatorname{argmin}} ||M - P \times E||_F^2, \qquad (5)$$

where the Frobenius matrix norm of the error term is considered. NMF requires the number of signatures N, an unknown parameter, to be predefined. An approach for selecting this parameter consists in obtaining a factorization of M for several of its values and then choosing the best N with respect to some performance measure such as the reconstruction error or the overall reproducibility. NMF is at the core of the Welcome Trust Sanger Institute (WTSI) Mutational Signature Framework, the first published method for signature extraction [2]. An alternative to numerical approaches based on NMF is given by statistical modelling and algorithms. With these

latter approaches, the number of mutations of a given type can be modelled by a Poisson distribution

$$m_g^k \sim \mathcal{P}\left(\sum_{n=1}^{N} p_n^k \, e_g^n\right)$$

where mutational processes are assumed to be mutually independent. In order to estimate E and P, it has been proposed to consider E as latent data and P as a matrix of unknown parameter and to apply an expectation-maximization algorithm [12] or use Bayesian approaches [13,14]. One important advantage of statistical approaches is the availability of model selection techniques for the choice of N. We refer to the recent review [8] for a comprehensive overview of the mathematical and computational aspects of existing methods, including alternative techniques not mentioned above.

With respect to de novo extraction of signatures, the refitting approaches consider that the signatures P are known and the goal is to estimate E given M and P. Refitting can be done for individual mutational catalogues (i.e. individual samples) and, from a linear algebra perspective, can be seen as the problem of projecting a catalogue living in the K-dimensional vector space (the space spanned by all mutation types) onto its subset of all linear combinations of the given mutational signatures with non-negative coefficients (the cone spanned by the given signatures).

A current practice consists in first performing a de novo extraction of signatures followed by a comparison of the newly identified signatures with the reference signatures (e.g. COSMIC signatures) by means of a similarity score, typically cosine similarity ranging from 0 (completely different) to 1 (identical) [3,6]. Finally, a "novel" signature is considered to reflect a specific reference signature if the similarity is larger than a fixed cut-off. If similarity is observed with more than one reference signature, the one with the largest value of similarity is chosen, supplementary

Figure 1. This assignment step crucially depends on the choice of the cut-off h that has been so far inconsistent in the literature with some studies using a value of 0.75 [7] whereas others 0.80 [15,16]. Another difficulty is that different signatures might have very close cosine similarity, as it happens also between COSMIC signatures, supplementary Figure 2, so that a unique assignment is not always possible.

Overview of the available tools

An equivalent number of *de novo* and refitting methods exist and all of them are available as open source tools, mainly as R packages, or web interfaces (Table 1). The typical input of the tools is a file including the mutations counts but some tools derive the mutation counts from ad-hoc input files that may include for each individual a list of mutated bases, their position within the genome and the corresponding bases from a reference genome. The typical format of such input files is MAF, VCF or less common formats such as MPF and MFVF. For biologists or those that are not familiar with programming, a set of tools were also developed and provided with user-friendly interfaces (see also Table 1). Some tools include additional features such as the possibility to search for specific patterns of mutations (e.g. APOBEC-related mutations [17]) and differentially expressed genes [5].

de novo approaches

Most tools that have been developed to identify mutational signatures were based on decomposition algorithms including NMF or Bayesian NMF [13,14]. The original model developed by Alexandrov was based on NMF and was implemented in MATLAB [2] and is available also as R package developed independently [18]. Another tool that utilises NMF is Maftools that is one of the few *de novo* tools that allows systematic comparison with the 30 validated signatures in

COSMIC by computing cosine similarity and assigning the identified signatures to the COSMIC one with the highest cosine similarity [5].

Another tool named SomaticSignatures allows the identification of mutational signatures through Principal Component Analysis (PCA) in addition to NMF [17]. For the sake of our formal comparison of the performance of the tools, we have tested SomaticSignatures only with the NMF implementation as with PCA the factors are orthogonal and the values inside the matrix can potentially be null or negatives, which biologically is unrealistic.

In addition to decomposition methods, a probabilistic model based on Expectation Maximization (EM) has been proposed to identify mutational signatures and it is implemented in the EMu tool [12]. The EM algorithm infers the number of mutational processes operative in a mutational catalogue and their individual signatures by considering that the input samples are independent.

Another tool that uses a probabilistic model named mixed-membership model is pmsignature [19]. This tool utilises a flexible approach that at the same time reduces the number of estimated parameters and allows to modify key contextual parameters and in particular increases the number of flanking bases. The latter feature makes it possible to consider more-fine grained signatures and could potentially lead to new biological insights[20].

For the purpose of the comparison with the other tools, we set to one the number of flanking bases, and therefore to 96 mutation types.

EMu, SigneR and pmsignature (and the refitting tool deconstructSigs) have the capacity to take into account “opportunity” matrices. For each analysed sequence, the corresponding opportunities are the frequencies of the types of mutations that are found in the same region of the reference genome or in a sequence from a normal tissue of the same individual. By “normalizing” the input

mutational catalogues with respect to the expected mutation frequencies in the analysed regions, accounting for opportunity matrices reduce the bias of the inferred signatures.

Refitting with known mutational signatures

In addition to the identification of novel mutational signatures, we are often interested in evaluating whether a signature observed in an individual tumour is one of an established set of signatures (e.g. the COSMIC signatures). This task is performed by "refitting tools" that are based on the estimation of the "best" combination of established signatures that explains the observed mutational catalogue that is identified by projecting the latter (i.e. mapping) into the multidimensional space of all non-negative linear combinations of the N established signatures.

The deconstructSigs [21] tool searches for the best linear combination of the established signatures through an iterative process based on multiple linear regression aimed at minimizing the distance between the linear combination of the signatures and the mutational catalogue. All the other tools minimise the distance through equivalent approaches based on quadratic programming [22–24], non-negative least square [25] linear combination decomposition [26] and simulated annealing [22].

A faster implementation

We propose an alternative implementation of the decomposition performed in [25,26] based on a simple geometric framework. Finding the linear decomposition of the input catalogue M on a set of given signatures minimizing the distance can be seen as the problem of projecting M on the geometric cone whose edges are the reference signatures. We propose to solve this problem by applying the very efficient R package called coneproj [27]. More details about our algorithm, which

we called MutationalCone, together with the R code implementing it, can be found in the Supplementary Material.

Data and experimental settings

Real cancers data from TCGA

In order to evaluate the performance of the available algorithms on real data, we used the exome sequences in The Cancer Genome Atlas (TCGA) repository (https://cancergenome.nih.gov/) for four cancer types: breast cancer, lung adenocarcinoma, B-cell lymphoma, and melanoma.

Mutation Annotation Format (MAF) files with the whole-exome somatic mutation datasets from these cohorts were downloaded from the portal gdc.cancer.gov on 6 March 2018. Data were annotated with MuSE [28] and the latest human reference genome (GRCh38). Mutational catalogues from these cohorts were obtained by counting the number of different mutation types using maftools [5]. The distribution of the number of mutations for each sample and separately for each cancer type is depicted in Figure 1.

According to the COSMIC website, 13 and 7 signatures have been found for breast cancer and lung adenocarcinoma respectively, 6 for B-cell lymphomas and 5 for melanoma.

Simulated data

The first key assumption of our model of simulated mutational catalogues is that the number of mutations in a sample g that are induced by process n follows a zero-inflated Poisson (ZIP) distribution. According to this two-component mixture model, e_g^n is either 0 with probability π or is sampled according to a Poisson distribution $\mathcal{P}(\lambda)$ with probability $1-\pi$. Such a model depends

on two parameters: the expectation λ of the Poisson component, and the probability π of extra zeros. The ZIP model allows for frequent zeroes and is therefore more suitable for modelling a heterogeneous situation where some samples are not exposed to a given mutational process ($e_g^n = 0$) while some others are ($e_g^n > 0$). Realistically, the mutation counts due to process n in each of the G samples, $e_1^n, \dots, e_G^n$, are assumed to be independent and identically distributed according to a ZIP model where the expectation of the Poisson component is specific to n:

$$e_g^n \sim \mathrm{ZIP}(\lambda_n, \pi), \text{for all } n = 1, \dots, N.$$

Note that the expected number of mutations in sample g due to process n is $(1-\pi)\lambda_n$. This flexibility given by process-specific average counts is the second important characteristics of our model and reflects the possibility that the mutagenic actions of different processes are intrinsically different with respect to their intensity. Obviously, it would have been possible to do one step further and allow for parameters $\lambda_{n,g}$ specific to both processes and samples, thus representing the realistic situation in which the exposures of different samples to the same process have different duration (e.g. smokers/non-smokers). However, this would have resulted in too many parameters to tune, thus making it difficult to interpret the results of our simulation study. For the same reason we considered one fixed value of π.

To explicit the dependence of λ_n on both the average total number r of mutations in a sample and the relative contribution of n, we impose the parametrization $\lambda_n(1-\pi) = q_n r$, where q_n is the average percentage share of mutations due to the process n. Note that in practice, several simulated samples will have a total number of mutations significantly lower than r, due to the effect of the zero-components of the ZIP models used to build the total catalogues. This formulation allows us

to study empirically the performance of the signature detection method as a function of the average number of mutations r.

Our simulation protocol is the following:

1. We choose N signatures from the COSMIC database, thus obtaining the matrix P.
2. For each sample g and process n we sample e_g^n from a ZIP distribution with parameters $\lambda_n = q_n r/(1-\pi)$ and π and obtain E. Here q_n, r and π are fixed parameters to set.
3. We compute the product $P \times E$. In order to obtain the final simulated catalogue M, we add some noise to the latter matrix by taking $m_g^k \sim \mathcal{P}\big((P \times E)_g^k\big)$.

We generate four alternative sets of simulated catalogues, referred to as profiles 1, 2, 3 and 4, each set mimicking a particular cancer: breast cancer, lung adenocarcinoma, B-cell lymphoma and melanoma. In order to do so, for each tumour type, we applied MutationalCone to the corresponding TCGA datasets and we calculated the mean contribution across all samples of each signature known to contribute to the specific cancer type q_n. Signatures with $q_n = 0$ do not contibute to the final catalogue and are not in the matrix *P* above. Supplementary Figure 3 depicts the resulting four sets of configurations $(q_1, \dots, q_n)$ used for the simulations. We fix the frequency of structural null contributions to the catalogues to $\pi = 0.6$ in all simulations. This value was chosen because it leads to a limited number of hypermutated catalogues, as it is often encountered in practice. At last, we vary the number of r from as little as 10 to as much as 100,000 mutations. This allows us to study the performance of methods on a large spectrum of catalogues: from a limited number of mutations as in exomes, to a very large number, as in cancer genome full sequences.

For each of the four tumour types and for each value of r, we simulate a catalogue matrix with G samples. Figure 2 shows examples with $G = 30$ for three values of r.

Methods for measuring the performance of algorithms

All methods for identifying signatures find solutions to the minimization problem (5). A straightforward way to measure the accuracy of the reconstructed catalogue is, therefore, to calculate the Frobenius norm of the reconstruction error

$$||M - \hat{M}||_F^2 = \sum_{g=1}^{G} \sum_{n=1}^{N} (m_g^k - \hat{m}_g^k)^2,$$

where $\hat{M} = \hat{P} \times \hat{E}$ is the matrix of catalogues reconstructed from the estimated signature and exposure matrices. Some of the algorithms involve stochastic steps such as resampling and/or random draws of initial parameters. For these algorithms, one simple way to assess the robustness of the estimates is to look at the variability of the reconstruction error when the same catalogues are analyzed several times with the same algorithm.

In order to make decisions about whether an extracted signature is the same as validated signatures (e.g. COSMIC signatures) a cut-off for cosine similarity needs to be defined. We decided to apply five different cut-offs from .75 to .95 and termed as "new" all identified mutational signatures that have a cosine similarity lower than our cut-off value.

Specificity and sensitivity for de novo extraction and assignment. In most applications, signature extraction is done in two steps: first, signatures are found using a de novo extraction tool and then for the extracted signatures a cosine similarity is calculated with each of the COSMIC

signatures. In order to measure the performance of both these steps combined, we used simulated catalogues and computed false and true positive rates and false and true negative rates. In a simulated catalogue, the set of *true* signatures $p_1, \dots, p_N$ that do contribute to the catalogue are known, thus allowing the comparison of the latter to the estimated signatures $\hat{p}_{i_1}, \dots, \hat{p}_{i_N}$. Note that, for the sake of simplicity, we explicitly look for the same number of signatures that we used to simulate the catalogues, not addressing questions about model selection performance.

Estimated signatures that belong to the set of "true signatures" are considered as true positives, while all "true signatures" that are not extracted count as false negatives. False positives are all estimated signatures that do not have a match in the set of "true signatures". This can happen for two reasons: the estimated signature is assigned to a COSMIC signature not used to build the catalogue, or it is not sufficiently similar to any COSMIC signature. This last situation usually takes place when setting a very high cosine similarity threshold h. In this case, signatures that have maximal cosine similarity lower that the cutoff, will be termed as "new". At last, true negatives are all COSMIC signatures not used for the simulation, nor estimated. From all this counts we compute specificity (number of true negatives divided by the total number of negatives) and sensitivity (number of true positives divided by the total number of positives).

In our empirical study, for each simulation setting described in the Simulated data section (that is for each profile given by a choice of proportions $(q_1, \dots, q_N)$ and for a choice of total number of mutations r) we built 50 replicates, each made of a matrix of G samples. Then we extracted signatures from all replicates with a given tool. Importantly, we set the expected number of signatures to be the same as the number of signatures used in the simulation, thus not addressing the problem of model selection. Then we compared the extracted signatures to the COSMIC

signatures using a cosine similarity threshold h. Finally, we computed specificity and sensitivity and obtained Monte-Carlo estimates based on the means over all replicates.

Bias of refitting procedures. Because refitting algorithms assume that the matrix of signatures is known, a simple way to assess their performance is to look at the bias of the exposure estimates e_g^n. In order to do so, we simulated 50 replicates each consisting of one lymphoma-mimicking (Profile 2) catalogue with an average number of mutations set to $r = 10^4$. Then, for each process n, we obtained Monte-Carlo estimates of the bias $E[\hat{e}_g^n] - e_g^n$ by averaging the differences $\hat{e}_g^n - e_g^n$ over all replicates. A global measure of performance that considers all exposure estimates is given by the mean squared error (MSE), that is the expected value of the loss function $\sum_{n=1}^{30}(\hat{e}_g^n - e_g^n)^2$. We obtained Monte-Carlo estimates of the mean squared error by averaging the loss function values across all 50 replicates and calculate asymptotic confidence intervals.

Analyses were performed using R 3.5.1 on a Linux Ubuntu (14.04) machine with 32 GB RAM and Intel Xeon E5-2630 v4 CPUs @ 2.2 GHz x 16.

Comparisons of algorithm performance

Performance of de novo tools

Figure 3 shows the distribution of the reconstruction error when a given computational tool is applied several times to the same real trinucleotide matrix. Reconstruction errors show limited variability due to stochastic steps in the algorithms and no variability whatsoever for EMu and

Maftools. All methods under evaluation are roughly equivalent in terms of their ability to properly reconstruct the initial matrix of mutational catalogues. This is not surprising, given that all methods are meant to solve the optimization problem in (5). Error value also appears to be cancer-dependent. This is expected because cancers differ with regards to the total number of mutations and the number of operating mutational signatures, making the decomposition more or less difficult.

We used realistic simulations to evaluate the performance of each method for de novo extraction followed by a classification step in which the extracted signatures are assigned to the most similar COSMIC signature.

Figures 4 and 5 show the specificity and sensitivity of such two-stage procedure as functions of the number of samples G in each catalogue and the cosine similarity cut-off h, while Figures 6 and 7 show the specificity and sensitivity as functions of the number of mutations in each catalogue and h.

We do not see very large differences in the tools' performance (specificity and sensitivity) between profiles with respect to the number of samples even though those measures are significantly lower for the Profile 1 samples which are more characterized by sparsely mutated sequences (Supplementary figure 3). For Profiles 2, 3 and 4 there is almost no gain in specificity and sensitivity when the sample size increases whereas for the "harder" Profile 1, performance measures increase for larger samples sizes (Supplementary figure 3).

Sensitivity improves with the average number of mutations. There is a large variability in sensitivity according to the cancer profiles and methods but in most scenarios, sensitivity is below 0.75. In general, sensitivity is higher for profiles characterized by a signature with a stronger effect than other contributing signatures (e.g. Profile 2 vs. Profile 4). Specificity is less variable and generally high. The specificity of all methods is generally slightly lower for profile 1 samples that

were simulated using many mutational signatures with heterogeneous relative contributions (Supplementary figure 3).

Specificity and sensitivity slightly deteriorate for higher cut-off values. This is expected because by setting a higher cut-off, the number of found signatures that are not similar enough to COSMIC signatures increases. These signatures are considered as false positives (that is true COSMIC signatures that are considered novel by the method), leading to a greater number of false positives and therefore to a lower specificity. Moreover, if the cut-off is too stringent, the number of false negatives will be high because some signatures used for the simulations are correctly found but do not score a high enough cosine similarity and therefore count as false negatives. This will make the resulting sensitivity low.

Figure 8 shows the running time when tools are applied to real lung datasets with a varying number of samples. While all methods show a fast-growing running time with increasing number of samples, SomaticSignatures and maftools are much faster than the others for more than 100 samples, making it possible to analyse large number of samples in few seconds. For example, for two hundred samples, the slowest method (SigneR), the running time is 913.72s while for the fastest (maftools), the value is 5.97s.

.

Performance of refitting tools

Figure 9 shows the distribution of the differences between the estimated and true contribution of all n signatures $e_g^1, \ldots, e_g^{30}$ for the different refitting methods under evaluation. Sample catalogues were simulated mimicking Lung cancer profiles (Profile 3), with signatures 1,2,4,5,6, 13 and 17 actually contributing as shown in supplementary figure 3. All methods give almost identical results.

By comparison with the true exposure profile given in supplementary figure 3, it is clear that all refitting methods provide good estimates of the contributions of all but signatures 1,5 and, 6 and to a lesser extent signature 4. Moreover, all methods correctly estimate a zero contribution for signatures 3 and 16 even though these are very similar to signature 5, supplementary Figure 2.

Interestingly, we note that signatures 2 and 13 (both attributed to APOBEC activity) are well identified by all methods. This finding is in line with previous claims about the stability of these two signatures.

In terms of running time, deconstructSigs and SignatureEstimation based on simulated annealing are more than two orders of magnitude slower than the other methods, (figure 10). All other methods run in a fraction of second. As expected, the running time increases linearly with the number of samples. MutationalCone, our custom implementation of the solution to the optimization problem solved by YAPSA and MutationalPatterns outperforms all other methods. The second fastest method is SignatureEstimation based on Quadratic Programing. As example, for two hundred samples, the execution time of deconstrucSigs is 86.148s and for MutationalCone is 0.028s.

Discussion

In this work we complement and expand a recent review of the available methods to identify mutational signatures [29] by empirically comparing their performance, using both real world TCGA data and simulated data. To our knowledge, this is the first formal comparison of both de novo discovery and refitting of mutational signatures. The results of this work lead to a better understanding of the strengths and limitations of each method as well as to the identification of the key parameters influencing their performance, namely the number of mutations and the "complexity" of the contributing signatures.

Since the publication of the first work about mutational signatures in 2013 [2], multiple algorithms have been developed, leading to similar but not identical results, a source of concern for researchers interested in this type of analysis. Conceptually, this is not surprising: mutational signatures are naturally defined in terms of non-negative matrix factorization, a well-known ill-posed problem (a unique solution does not exist). Although this limitation has cast doubts on the biological validity of mutational signatures, this has been somehow validated using experimental and computational approaches by Zho et al. [27].

We note that the paper by X. Li [28] is accompanied by an online technical report with a comparison of the performance of NMF algorithms in terms of stability and reconstruction error (http://lixiangchun.github.io/msig/), however this work was based solely on real data.

An original aspect of this paper is the empirical performance study and in particular the probabilistic model of mutational catalogues used to generate simulations. This model is based on

the zero-inflated Poisson distribution that allows for sparse contribution of signatures and thus makes it possible to build mutation count data that are more realistic than the pure Poisson model previously considered [12,13]. We suggest that developers should assess their new methods on simulations based on realistic models such as ours. Another innovative aspect of our work is the assessment of the performance of de novo approaches. This is in general a slippery problem given the non-supervised learning nature of NMF and concurrent methods. We addressed it by building confusion matrices after comparing the "true" signatures used for simulations with the newly identified signatures in terms of their cosine similarity. At last, we introduced MutationalCone, a new implementation of existing refitting tools that is based on a very simple geometric model and proved to be faster than all other methods. The R code implementing MutationalCone can be found in the Supplementary Material.

The results of our study can be helpful to researchers planning mutational signature analysis of real data. In particular we show that the performance of de novo methods depends on the complexity of the analysed sequences, the number of mutations and to a lesser degree the number of samples analysed. We showed that cancer profiles for which it is "easier" to detect signatures are characterised by one predominant signature only (as in our profiles 2 and 3), while cancers with more complex profiles, characterized by non-negligible contributions by several signatures (profiles 1 and 4), are more difficult to identify. Indeed, recent evidence shows that the majority of cancers harbour a large number of mutational signatures [20] and therefore belong to the latter scenario.

With regards to the mutation number, we observe that with the number of mutations that could be found in some cancer exomes the performance is generally poor (i.e. low specificity and sensitivity). A clear example is illustrated in a recent comparison between a smaller set of genomes (G=2,780) and a much larger sample of exomes (G=19,184), where it is shown that in the former

set a much larger number is signatures is detected than in the latter [20]. This problem is likely to be mitigated if counts were normalized by the expected number of each type's trinucleotides in the analysed region under healthy condition, that is if an opportunity matrix was provided. We do not address this important aspect in our comparison study as only a few methods can account for opportunity matrices.

We showed that in comparing found signatures with COSMIC signatures, the choice of a cosine similarity cut-off has a relatively low impact on the overall performance. If the object is to identify novel signatures it would preferable to choose a lower value (0.75 or less). On the contrary, if the aim is to validate existing signatures, we recommend to turn to refitting methods.

Not all the de novo methods we evaluated offer the possibility to automatically choose the number of signatures to be found. For instance, the popular SomaticSignatures only provides a graphical visualisation of the residual sum of squares for several choices of the number of signatures; the user can choose the optimal number by identifying the inflexion point. For this reason, this crucial aspect was not addressed in our empirical assessment. Similarly, we only considered mutation types defined by the trinucleotide motifs, as currently only pmsignature[19] can consider more than one flanking basis on each side of the substitution.

For well-studied cancers, refitting approaches are a faster and more powerful alternative to de novo methods, even with just one input sample. As the COSMIC database has been built and validated by analysing tens of thousands of sequences of most cancer types, we recommend borrowing strength from previous studies and using refitting tools when carrying on standard analysis not aimed at the discovery of new signatures.

As a conclusive note, we recommend systematically performing a sensitivity analysis based on the application of one or more alternative methods based on different algorithms, in order to assess the robustness of claims.

Table

Software	Available platform/model	Input files	Additional features
De novo approaches			
EMu	Command line/EM algorithm	Mutation counts file Relative to the other tools, the counts file is transposed (the rows correspond to the samples)	- Opportunity matrix - Selection of the optimal number of signatures - Enrichment and depletion analysis
Maftools	R-Bioconductor /NMF	Mutation Annotation Format	- Genomic visualization - Cosine similarity - Selection of the optimal number of signatures - Group-wise comparisons (differential analysis) - APOBEC enrichment analysis
bayesNMF	R/bayesian NMF	Mutation counts file	- Selection of the optimal number of signatures
mutSignatures	R/NMF	Mutation counts file	
SomaticSignatures	R/NMF and PCA	Variant Call Format	- Group-wise comparisons - Genomic visualization - Hierarchical clustering
signeR	R/bayesian NMF	Variant Call Format	- Opportunity matrix - Selection of the optimal number of signatures
pmsignature	R/mixed-membership model	- Mutation Position Format -Mutation Feature Vector Format	- Reduction of complexity - Mutation types defined by one or two flanking bases - Selection of the optimal number of signatures - Transcriptional strand bias - Background signature

Refitting approaches			
decompTumor2Sig	R/quadratic programming	- Variant Call Format - Mutation Position Format - Mutation Feature Vector Format	- Converts a set of Alexandrov signatures[3] to Shiraishi signatures[19] - Decomposes a mutational catalogue in Shiraishi signatures
deconstructSigs	R/linear regression	Mutation counts file	- Group-wise comparison (differential analysis) - Opportunity matrix
MutationalCone	R/cone projection	Mutation counts file	
MutationalPatterns	R/Non-Negative Least Squares	Mutation counts file	- Also de novo identification - Cosine similarity comparison - Strand bias analyses - Enrichment and depletion
SignatureEstimation	R/quadratic programming and simulated alienation	Mutation counts file	
QPsig	R/quadratic programming	Mutation counts file	
YAPSA	R/Linear Combination Decomposition	Mutation counts file	- Cut-off for normalized exposure - Enrichment and depletion
Pipelines and web-interfaces			
mSignatureDB[6]	Web-interface	- Variant Call Format - Mutation Annotation Format - TSV	- Refitting and de novo identification - Bootstrapped cosine similarity - Comparison with either hg19 or hg38
MuSiCa[30]	Web-interface	- Variant Call Format - Mutation Annotation Format - TSV - Excel	- Refitting and de novo identification - Cosine similarity - Samples classification
MutaGene[31]	Web-interface	TCGA and ICGC data	- Refitting and de novo identification - Clustering of samples according to mutational profiles - Identification of potential driver's mutations
Mutalisk[32]	Web-interface	Variant Call Format	- Refitting and de novo identification - Transcriptional strand bias - Localization of kaetegis - Histones modifications - Cosine similarity comparison
Mutspec[33]	Galaxy pipeline/NMF	Variant Call Format	- de novo identification - Includes MS analysis in mouse cancer

Figures

Figure 1: Number of mutations of samples in the four TCGA cohorts. The number of analysed exomes for each cancer type is 963 (Breast), 563 (Lung adenocarcinoma), 37 (Lymphoma B-cell) and 465 (Melanoma).

Figure 2: Sample of simulated catalogues. Catalogues are simulated according to the model described in the text. Tuning parameters are the average number of mutations, here $r = 10^2, 10^4$ and 10^5 and the cancer type specific average contribution of each signature (see Supplementary figure 3).

Figure 3: Reconstruction errors and their variability due to stochastic steps in the algorithms. Each program under evaluation is applied 50 times on the same matrix of real catalogues; boxplots represent the distribution of the squared Frobenius distance between the original catalogue and its reconstruction. Boxplots look like flat segments because of the scale of the y-axis.

Figure 4 and 5: Performance of extraction methods and mapping on COSMIC signatures as the number of analysed catalogues and the cosine cut-off *h* vary. Specificity and sensitivity are estimated from 50 replicates each made of G genomes. The average number of mutations in each catalogue is $r = 10{,}000$. The model used to simulate realistic replicates according to the four cancer types and the estimation methods are described in the section Data and experimental settings.

Figures 6 and 7: Performance of extraction methods and mapping on COSMIC signatures as the average number of mutations and the cosine cut-off *h* vary. Specificity and sensitivity

are estimated from 50 replicates each made of $G = 30$ catalogues. The model used to simulate realistic replicates according to the four cancer types and the estimation methods are described in the section Data and experimental settings.

Figure 8 : Running times of refitting tools. Methods were applied to subsets of the TCGA Lung cohort of different sizes.

Figure 9: Bias of the estimates of each signature contribution for several refitting methods. For each signature, the bias estimates are obtained by averaging the exposure estimates across 50 samples. Mean square errors, together with 95% confidence intervals, are reported on the top of each plot. Simulations were done according to the model described in the Data and experimental settings section.

Figure 10: Running times of refitting tools. Methods were applied to subsets of the TCGA Lung cohort of different sizes.

Acknowledgements

We acknowledge the support of a grant on epigenetic signatures and cancer from La Ligue contre le Cancer (Appel à projets 2016 'Recherche en Épidémiologie'). HO is supported by a PhD fellowship from the French Institut National du Cancer (INCa_11330). VP would like to thank Grégory Nuel for the useful discussions around zero-inflated models.

Key points

- Mutational signatures are an interesting approach to study the process of carcinogenis and identify factors linked to it.
- The sets of tools that have been developed to identify mutational signatures provide similar results and point out the validity of the concept.
- The performance of such tools depends on parameters such as the number of mutations and should be carefully evaluated. For this we recommend the application of more than one method.
- When carrying on standard analysis not aimed at the discovery of new signatures, we recommend borrowing strength from previous studies and using refitting rather than de novo tools.

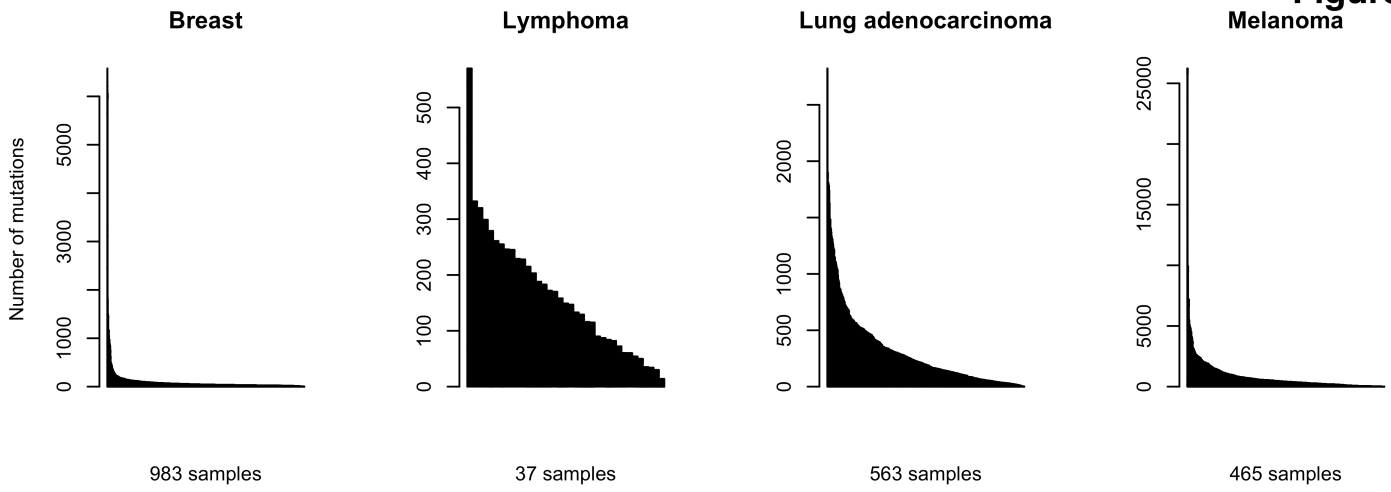
References

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458:719–724
2. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Reports 2013; 3:246–259
3. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013; 500:415–421
4. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research 2017; 45:D777–D783
5. Mayakonda A, Koeffler HP. Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. bioRxiv 2016; 052662
6. Huang P-J, Chiu L-Y, Lee C-C, et al. mSignatureDB: a database for deciphering mutational signatures in human cancers. Nucleic Acids Research 2018; 46:D964–D970
7. Letouzé E, Shinde J, Renault V, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nature Communications 2017; 8:
8. Baez-Ortega A, Gori K. Computational approaches for discovery of mutational signatures in cancer. Briefings in Bioinformatics 2017;
9. Rogozin IB, Pavlov YI, Goncearenco A, et al. Mutational signatures and mutable motifs in cancer genomes. Briefings in Bioinformatics 2017;
10. Nik-Zainal S, Morganella S. Mutational Signatures in Breast Cancer: The Problem at the

DNA Level. Clinical Cancer Research 2017; 23:2617–2629
11. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999; 401:788
12. Fischer A, Illingworth CJ, Campbell PJ, et al. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biology 2013; 14:R39
13. Rosales RA, Drummond RD, Valieris R, et al. signeR: an empirical Bayesian approach to mutational signature discovery. Bioinformatics 2017; 33:8–16
14. Kim J, Mouw KW, Polak P, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nature Genetics 2016; 48:600–606
15. Kim SY, Jung S-H, Kim MS, et al. Genomic profiles of a hepatoblastoma from a patient with Beckwith-Wiedemann syndrome with uniparental disomy on chromosome 11p15 and germline mutation of APC and PALB2. Oncotarget 2017; 8:
16. Han M-R, Shin S, Park H-C, et al. Mutational signatures and chromosome alteration profiles of squamous cell carcinomas of the vulva. Experimental & Molecular Medicine 2018; 50:e442
17. Gehring JS, Fischer B, Lawrence M, et al. SomaticSignatures: inferring mutational signatures from single-nucleotide variants: Fig. 1. Bioinformatics 2015; btv408
18. Fantini D, Huang S, Asara JM, et al. Chromatin association of XRCC5/6 in the absence of DNA damage depends on the XPE gene product DDB2. Molecular Biology of the Cell 2017; 28:192–200
19. Shiraishi Y, Tremmel G, Miyano S, et al. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. PLOS Genetics 2015; 11:e1005657
20. Alexandrov L, Kim J, Haradhvala NJ, et al. The Repertoire of Mutational Signatures in Human Cancer. 2018;
21. Rosenthal R, McGranahan N, Herrero J, et al. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology 2016; 17:
22. Huang X, Wojtowicz D, Przytycka TM. Detecting presence of mutational signatures in cancer with confidence. Bioinformatics 2018; 34:330–337
23. Krüger S, Piro RM. Identification of mutational signatures active in individual tumors.
24. Lynch AG. Decomposition of mutational context signatures using quadratic programming methods. F1000Research 2016; 5:1253
25. Blokzijl F, Janssen R, van Boxtel R, et al. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Medicine 2018; 10:
26. Huebschmann D, Gu Z, Schelsner M. YAPSA: Yet Another Package for Signature Analysis R package version 1.6.0. 2015;
27. Liao X, Meyer MC. **coneproj** : An *R* Package for the Primal or Dual Cone Projections with Routines for Constrained Regression. Journal of Statistical Software 2014; 61:
28. Fan Y, Xi L, Hughes DST, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biology 2016; 17:
29. Perduca V, Omichessan H, Baglietto L, et al. Mutational and epigenetic signatures in cancer tissue linked to environmental exposures and lifestyle: Current Opinion in Oncology 2018; 30:61–67
30. Díaz-Gay M, Vila-Casadesús M, Franch-Expósito S, et al. Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. BMC Bioinformatics 2018; 19:
31. Goncearenco A, Rager SL, Li M, et al. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic Acids Research 2017; 45:W514–W522

32. Lee J, Lee AJ, Lee J-K, et al. Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Research 2018;
33. Ardin M, Cahais V, Castells X, et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinformatics 2016; 17:

Figure 1


Breast
Lymphoma
Lung adenocarcinoma
Melanoma
Number of mutations
0
1000
3000
5000
0
100
200
300
400
500
0
500
1000
2000
0
5000
15000
25000
983 samples
37 samples
563 samples
465 samples

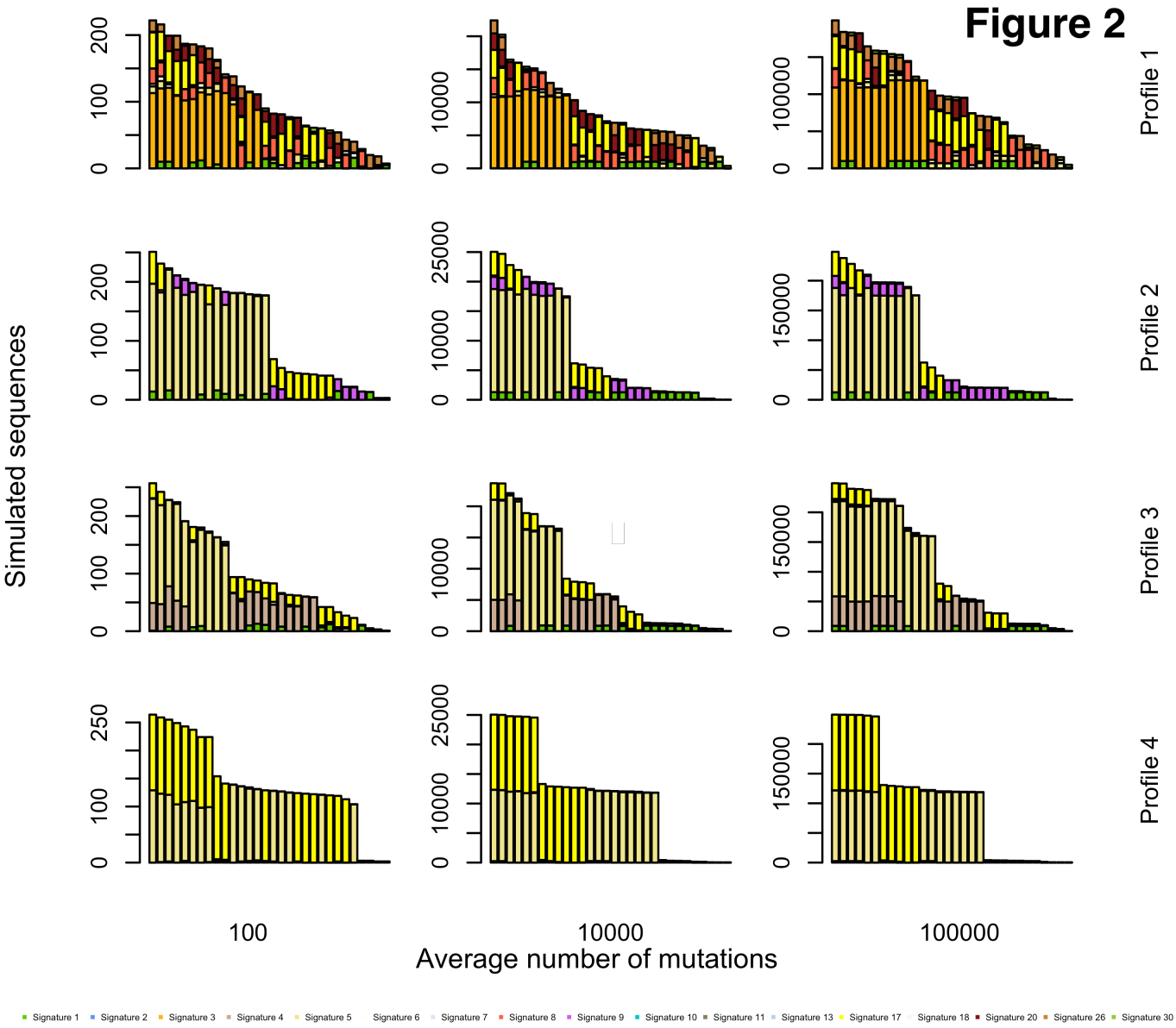


Figure 2
Simulated sequences
Profile 1
Profile 2
Profile 3
Profile 4
0
100
200
250
10000
25000
100000
150000
100
10000
100000
Average number of mutations
Signature 1
Signature 2
Signature 3
Signature 4
Signature 5
Signature 6
Signature 7
Signature 8
Signature 9
Signature 10
Signature 11
Signature 13
Signature 17
Signature 18
Signature 20
Signature 26
Signature 30

Figure 3

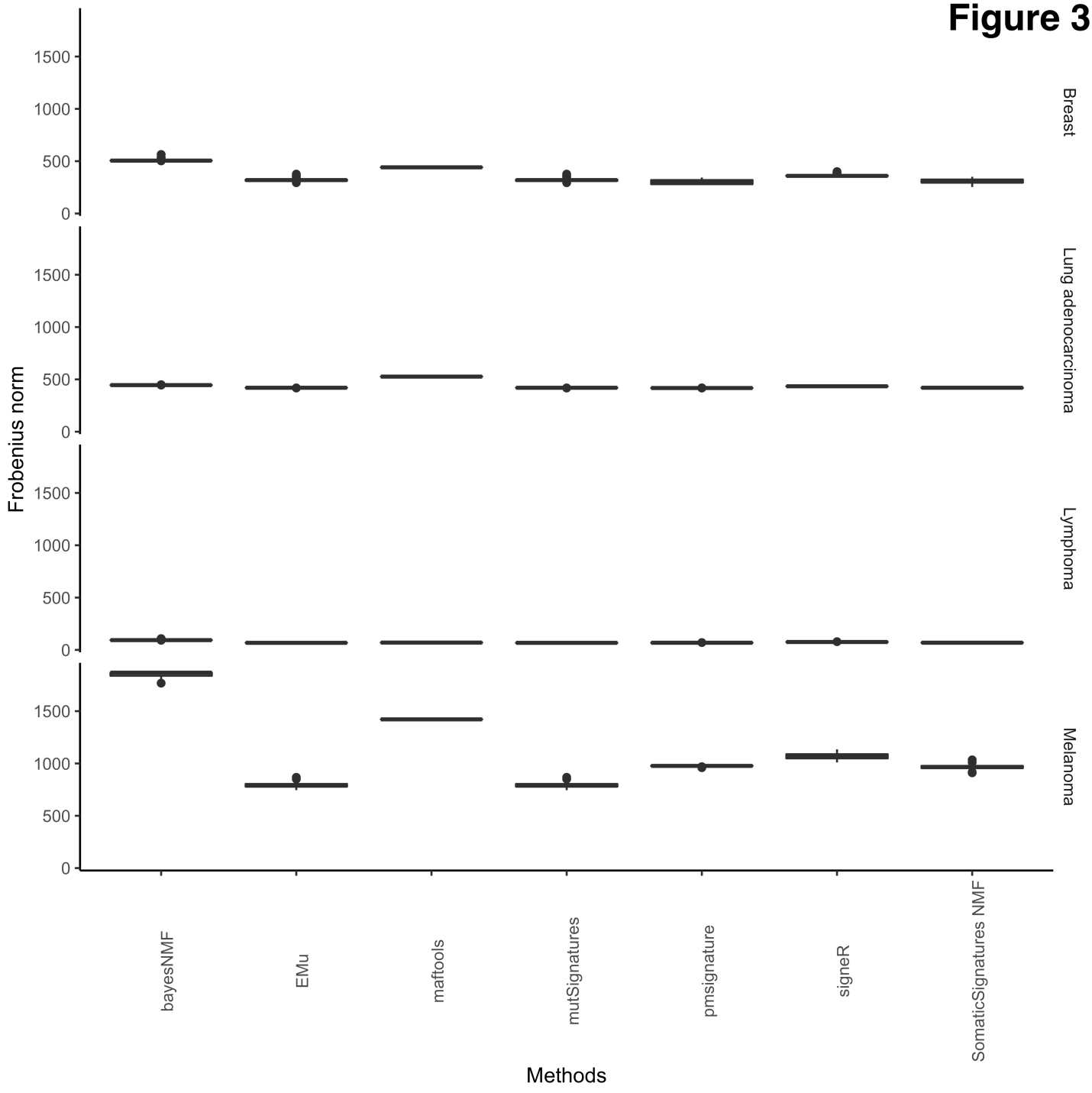

Frobenius norm
1500
1000
500
0
Breast
Lung adenocarcinoma
Lymphoma
Melanoma
bayesNMF
EMu
maftools
mutSignatures
pmsignature
signeR
SomaticSignatures NMF
Methods

Figure 4

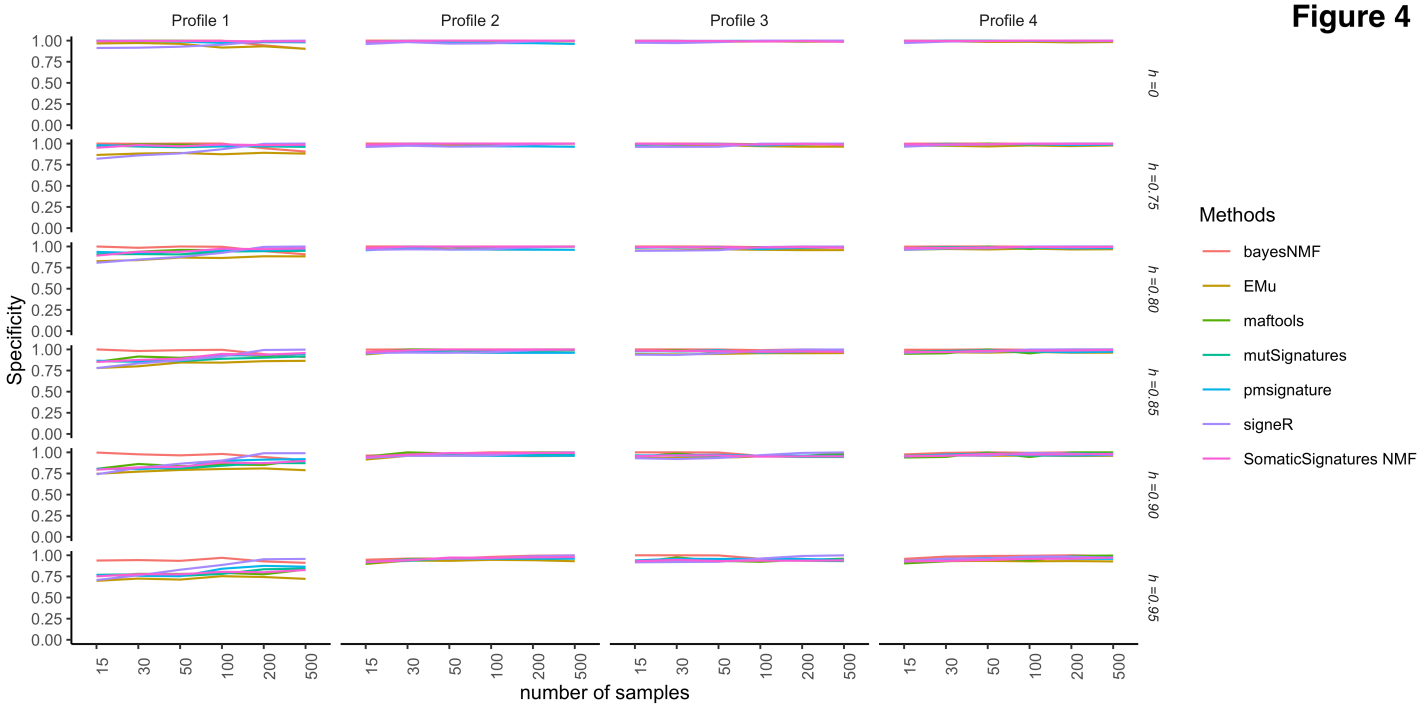

Profile 1
Profile 2
Profile 3
Profile 4
Specificity
1.00
0.75
0.50
0.25
0.00
h =0
h =0.75
h =0.80
h =0.85
h =0.90
h =0.95
15
30
50
100
200
500
number of samples
Methods
bayesNMF
EMu
maftools
mutSignatures
pmsignature
signeR
SomaticSignatures NMF

Figure 5

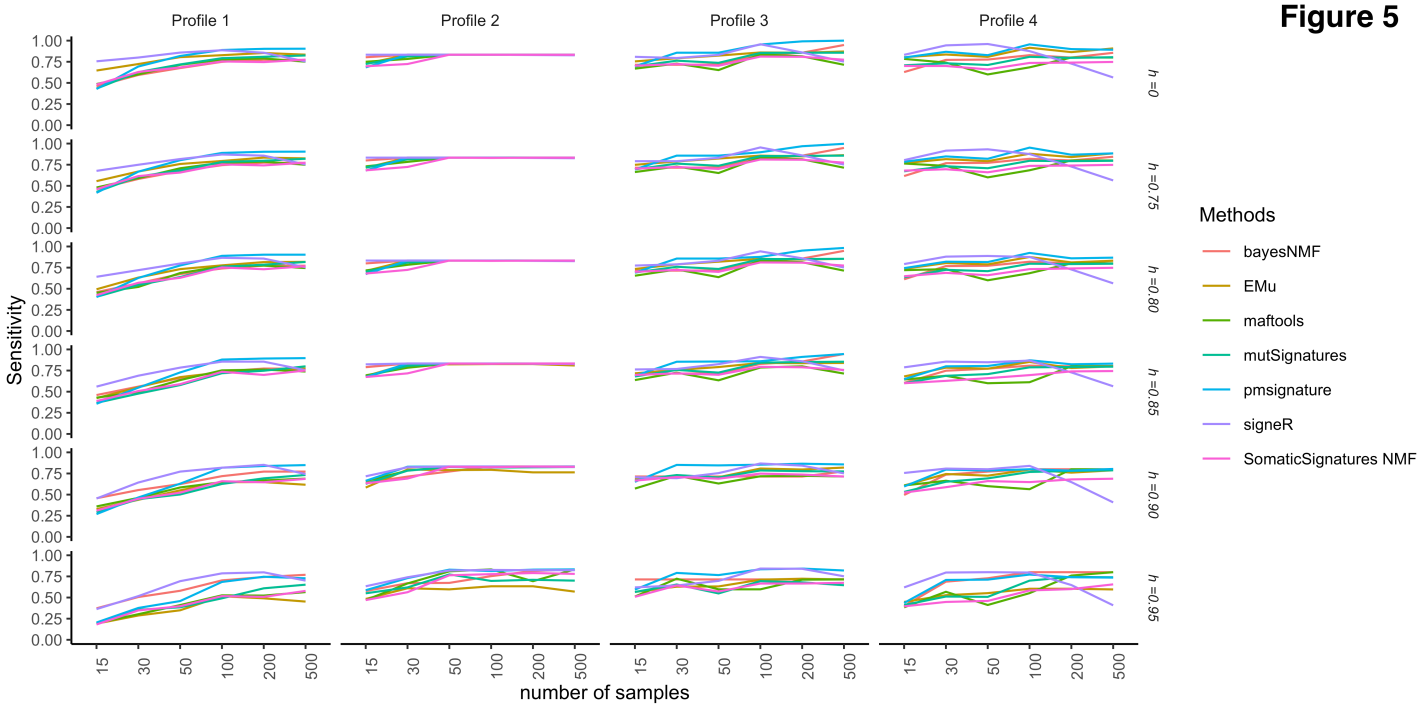

Profile 1
Profile 2
Profile 3
Profile 4
Sensitivity
1.00
0.75
0.50
0.25
0.00
h =0
h =0.75
h =0.80
h =0.85
h =0.90
h =0.95
15
30
50
100
200
500
number of samples
Methods
bayesNMF
EMu
maftools
mutSignatures
pmsignature
signeR
SomaticSignatures NMF

Figure 6

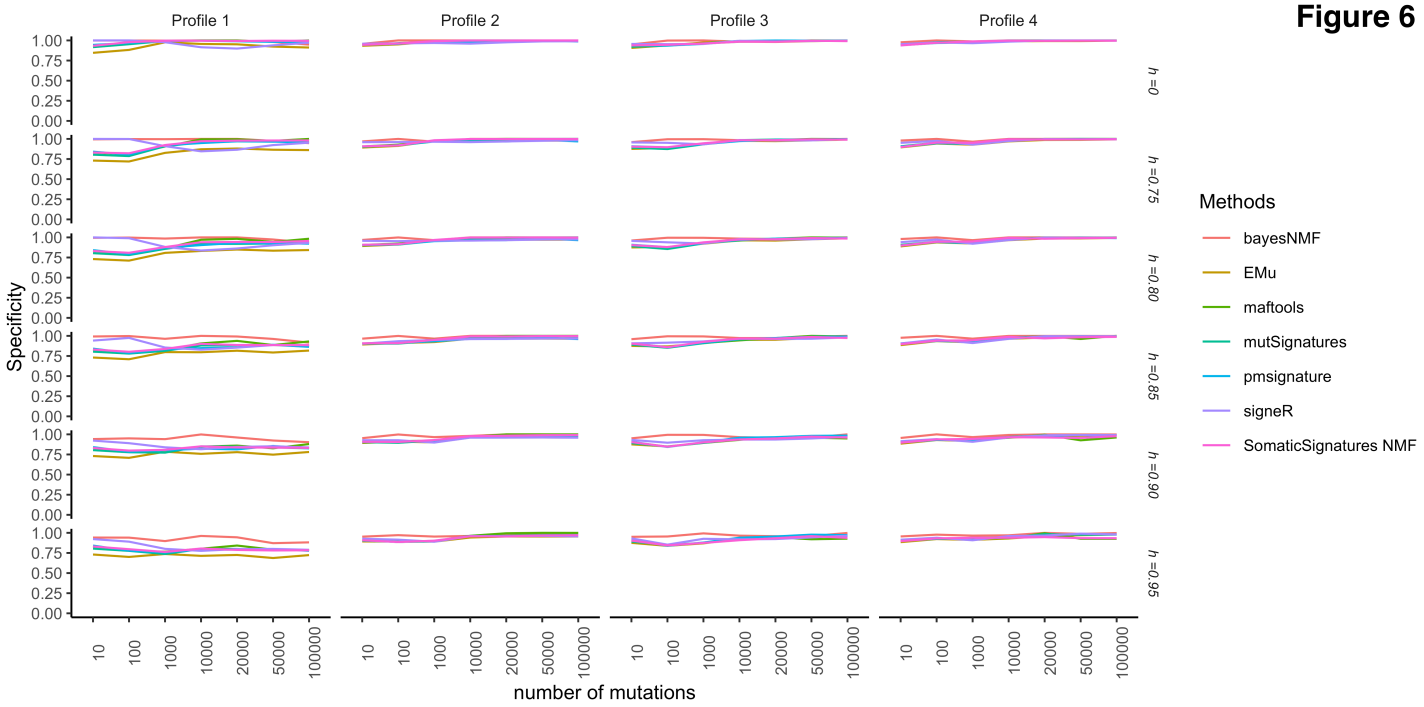
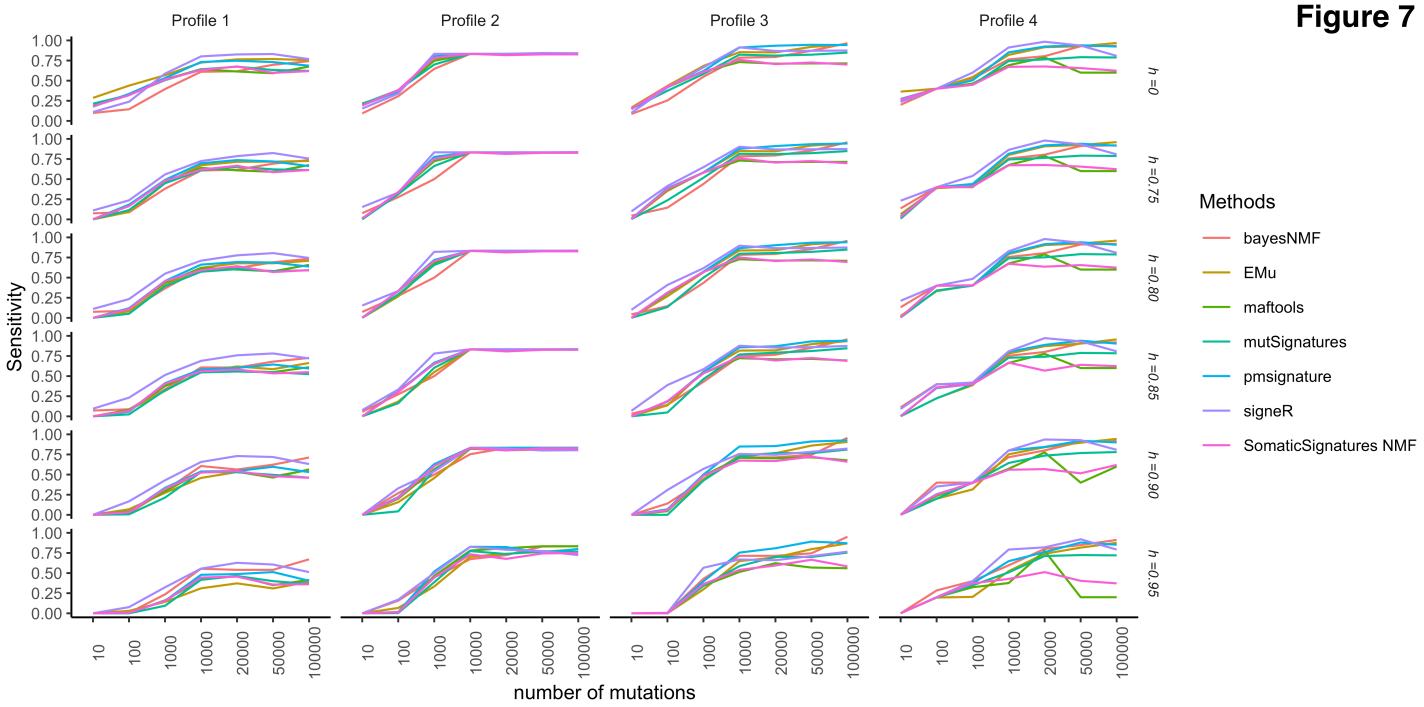

Profile 1
Profile 2
Profile 3
Profile 4
Specificity
1.00
0.75
0.50
0.25
0.00
h =0
h =0.75
h =0.80
h =0.85
h =0.90
h =0.95
10
100
1000
10000
20000
50000
100000
number of mutations
Methods
bayesNMF
EMu
maftools
mutSignatures
pmsignature
signeR
SomaticSignatures NMF

Figure 7


Profile 1
Profile 2
Profile 3
Profile 4
Sensitivity
1.00
0.75
0.50
0.25
0.00
h =0
h =0.75
h =0.80
h =0.85
h =0.90
h =0.95
10
100
1000
10000
20000
50000
100000
number of mutations
Methods
bayesNMF
EMu
maftools
mutSignatures
pmsignature
signeR
SomaticSignatures NMF

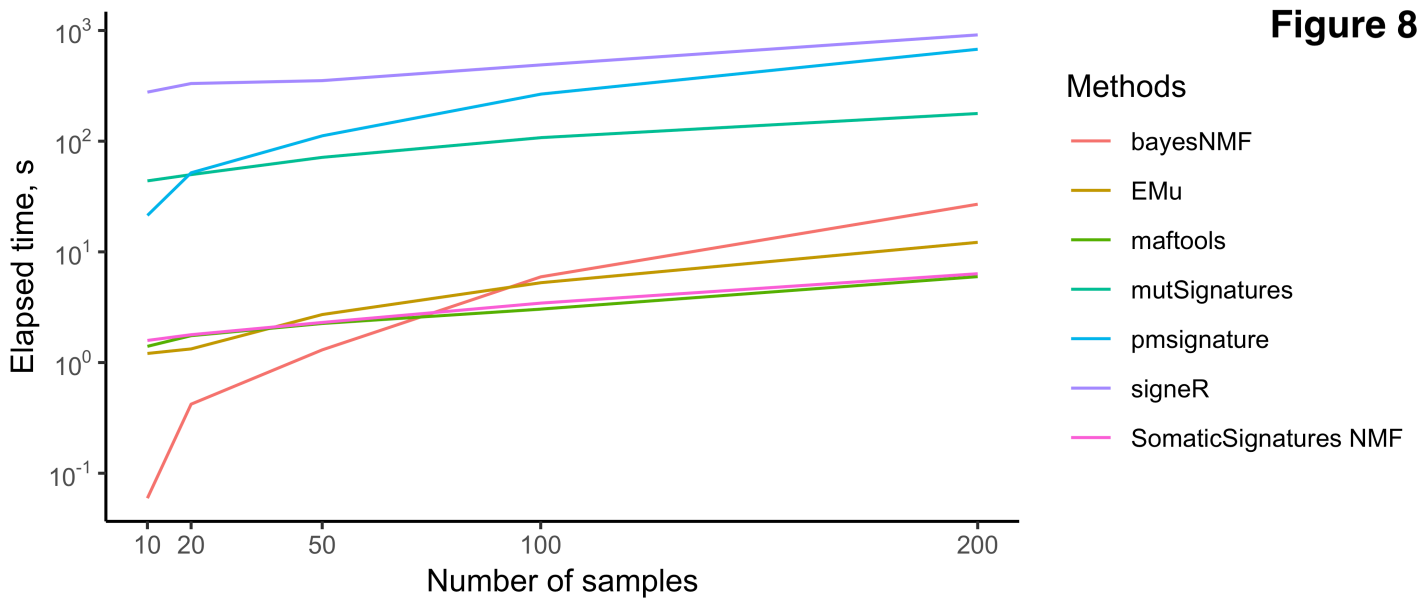
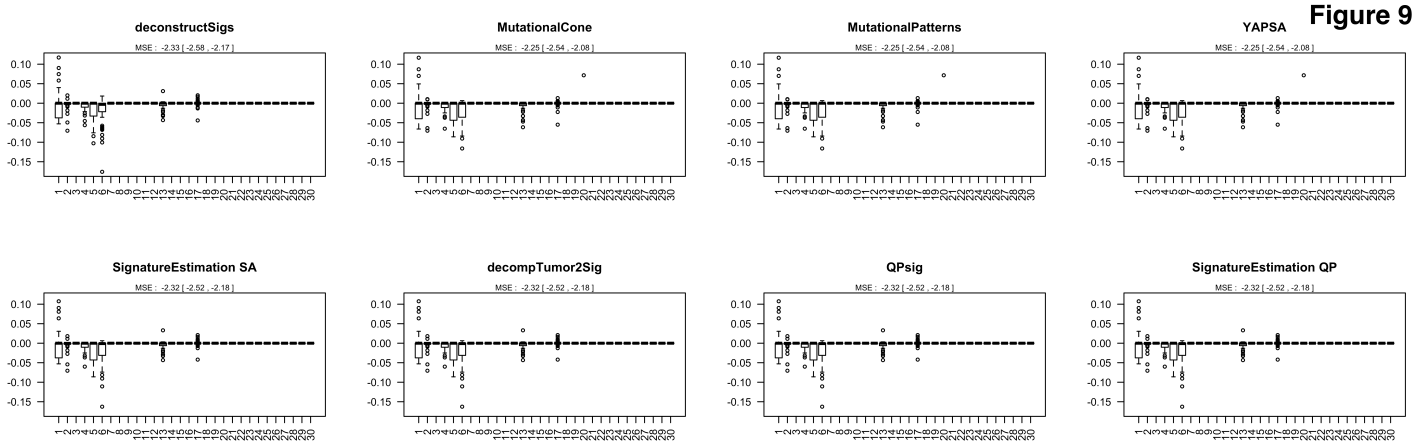


Figure 8
Elapsed time, s
10^3
10^2
10^1
10^0
10^{-1}
10
20
50
100
200
Number of samples
Methods
bayesNMF
EMu
maftools
mutSignatures
pmsignature
signeR
SomaticSignatures NMF

Figure 9


deconstructSigs
MSE : -2.33 [-2.58 , -2.17]
MutationalCone
MSE : -2.25 [-2.54 , -2.08]
MutationalPatterns
MSE : -2.25 [-2.54 , -2.08]
YAPSA
MSE : -2.25 [-2.54 , -2.08]
SignatureEstimation SA
MSE : -2.32 [-2.52 , -2.18]
decompTumor2Sig
MSE : -2.32 [-2.52 , -2.18]
QPsig
MSE : -2.32 [-2.52 , -2.18]
SignatureEstimation QP
MSE : -2.32 [-2.52 , -2.18]

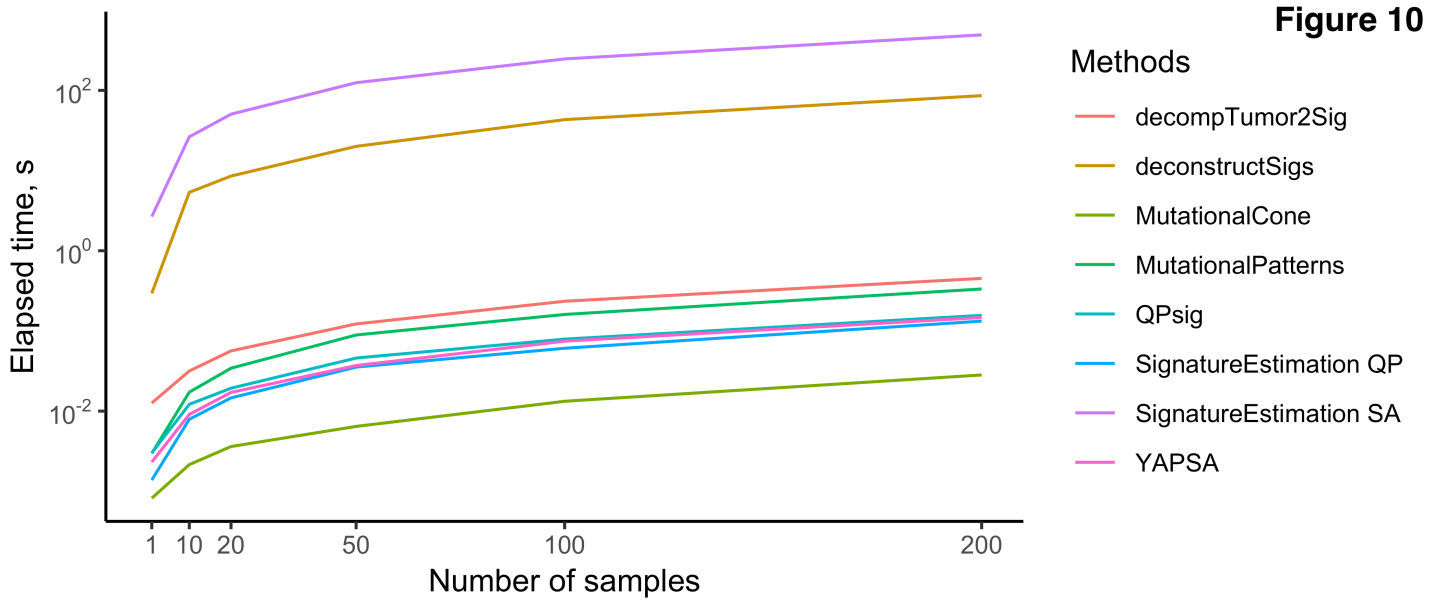


Figure 10
Elapsed time, s
10^2
10^0
10^{-2}
1
10
20
50
100
200
Number of samples
Methods
decompTumor2Sig
deconstructSigs
MutationalCone
MutationalPatterns
QPsig
SignatureEstimation QP
SignatureEstimation SA
YAPSA