Uncovering novel mutational signatures by *de novo* extraction with SigProfilerExtractor

S M Ashiqul Islam[1,2,3], Yang Wu[4], Marcos Díaz-Gay[1,2,3], Erik N Bergstrom[1,2,3], Yudou He[1,2,3], Mark Barnes[1,2,3], Mike Vella[5], Jingwei Wang[6], Jon W Teague[6], Peter Clapham[6], Sarah Moody[6], Sergey Senkin[7], Yun Rose Li[8], Laura Riva[6], Tongwu Zhang[9], Andreas J Gruber[10,11], Raviteja Vangara[12], Christopher D Steele[13], Burçak Otlu[1,2,3], Azhar Khandekar[1,2,3], Ammal Abbasi[1,2,3], Laura Humphreys[6], Natalia Syulyukina[2], Samuel W Brady[14], Boian S Alexandrov[11], Nischalan Pillay[13,15], Jinghui Zhang[14], David J Adams[6], Iñigo Marticorena[6], David C Wedge[10,11], Maria Teresa Landi[9], Paul Brennan[7], Michael R Stratton[6], Steven G Rozen[4], and Ludmil B Alexandrov[1,2,3*]

Affiliations

[1]Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, 92093, USA
[2]Department of Bioengineering, UC San Diego, La Jolla, CA, 92093, USA
[3]Moores Cancer Center, UC San Diego, La Jolla, CA, 92037, USA
[4]Centre for Computational Biology and Programme in Cancer & Stem Cell Biology, Duke NUS Medical School, 169857, Singapore
[5]NVIDIA Corporation, 2788 San Tomas Expressway, Santa Clara, CA, 95051, USA
[6]Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK
[7]Genetic Epidemiology Group, International Agency for Research on Cancer, 69372 Lyon CEDEX 08, France
[8]Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, 94158, USA
[9]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, 20892, USA
[10]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7LF, UK
[11]Manchester Cancer Research Centre, The University of Manchester, Manchester, M20 4GJ, UK
[12]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA
[13]Research Department of Pathology, Cancer Institute, University College London, London, WC1E 6BT, UK
[14]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA
[15]Department of Cellular and Molecular Pathology, Royal National Orthopaedic Hospital NHS Trust, Stanmore, Middlesex, HA7 4LP, UK

[*]Correspondence should be addressed to L2alexandrov@health.ucsd.edu.

ABSTRACT

Mutational signature analysis is commonly performed in genomic studies surveying cancer and normal somatic tissues. Here we present SigProfilerExtractor, an automated tool for accurate *de novo* extraction of mutational signatures for all types of somatic mutations. Benchmarking with a total of 33 distinct scenarios encompassing 1,106 simulated signatures operative in more than 200,000 synthetic genomes demonstrates that SigProfilerExtractor outperforms ten other tools across all datasets with and without noise. For simulations with 5% noise, reflecting high-quality genomic datasets, SigProfilerExtractor outperforms other approaches by elucidating between 20% and 50% more true positive signatures while yielding more than 5-fold less false positive signatures. Applying SigProfilerExtractor to 2,778 whole-genome sequenced cancers reveals three previously missed mutational signatures. Two of the signatures are confirmed in independent cohorts with one of these signatures associating with tobacco smoking. In summary, this report provides a reference tool for analysis of mutational signatures, a comprehensive benchmarking of bioinformatics tools for extracting mutational signatures, and several novel mutational signatures including a signature putatively attributed to direct tobacco smoking mutagenesis in bladder cancer and in normal bladder epithelium.

INTRODUCTION

De novo extraction of mutational signatures[1] is an unsupervised machine learning approach where a matrix, ***M***, which corresponds to the somatic mutations in a set of cancer genomes under a mutational classification[2], is approximated by the product of two low-rank matrices, ***S*** and ***A***. The matrix ***S*** reflects the set of mutational signatures while the matrix ***A*** encompasses the activities of the signatures; an activity corresponds to the number of mutations contributed by a signature in a sample. Algorithmically, *de novo* extraction of mutational signatures has relied on nonnegative matrix factorization (NMF)[3] or on approaches mathematically analogous to NMF[4-6]. The main advantage of NMF over other factorization approaches is its ability to yield nonnegative factors that are part of the original data, thus, allowing interpretation of the identified nonnegative factors[3]. Biologically, mutational signatures extracted from cancer genomes have been attributed to exposures to environmental carcinogens, failure of DNA repair pathways, infidelity/deficiency of replicating polymerases, iatrogenic events, and others[7-14].

Since we introduced the mathematical concept of mutational signatures[1], a number of computational frameworks have been developed for performing *de novo* extraction of mutational signatures (**Table 1**)[15-25]. Notably, the majority of existing *de novo* extraction tools *(i)* predominately support the simplest mutational classification, *viz.*, SBS-96 which encompasses single base substitutions with their immediate 5' and 3' sequence context[2]; *(ii)* lack automatic selection for the number of mutational signatures; *(iii)* do not identify a robust solution leading to different results following re-analysis of the same dataset; *(iv)* require pre-selection of a large number of priors and/or hyperparameters; *(v)* do not decompose *de novo* signatures to the set of

79 reference COSMIC signatures[11]. Importantly, there has been no extensive benchmark of the
80 existing tools for *de novo* extraction leading to uncertainty in regard to their performance.
81
82 To address these limitations, here, we present SigProfilerExtractor – a reference tool for *de novo*
83 extraction of mutational signatures. SigProfilerExtractor allows analysis of all types of
84 mutational classifications, performs automatic selection of the number of signatures, yields
85 robust solutions, requires only minimum setup, and decomposes *de novo* extracted signatures to
86 known COSMIC signatures. A comprehensive benchmark including 2,879 applications of
87 SigProfilerExtractor and ten other tools across a total of 33 distinct scenarios reveals that
88 SigProfilerExtractor is robust to noise and it outperforms all other computational tools for *de*
89 *novo* extraction of mutational signatures (**Supplementary Tables 1–3**). Applying
90 SigProfilerExtractor to the recently published set of 2,778 whole-genome sequenced cancers
91 from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project[26] elucidates three novel
92 signatures that were not found in the original PCAWG analysis of mutational signatures[11]. Two
93 of the signatures are confirmed in independent cohorts and a putative etiology of tobacco-
94 associated mutagenesis is attributed to one of these signatures.
95

RESULTS

Overview of SigProfilerExtractor and its implementation

SigProfilerExtractor is implemented as a Python package, with an R wrapper, allowing users to run it in both Python and R environments: https://github.com/AlexandrovLab/SigProfilerExtractor . The tool is also extensively documented including a detailed Wiki page: https://osf.io/t6j7u/wiki/home/. By default, the tool requires only a single parameter – the input dataset containing the mutational catalogues of interest. SigProfilerExtractor supports most commonly used formats outputted by variant calling algorithms (*e.g.*, VCF and MAF), which are internally converted to a matrix, ***M,*** by SigProfilerMatrixGenerator[2]. SigProfilerExtractor can also be applied to a text file containing a matrix, ***M***, thus supporting nonnegative matrix factorization for any custom matrix dataset. By default, the tool decomposes the matrix ***M*** searching for an optimal solution between 1 and 25 mutational signatures (**Figure 1*a***). For each decomposition, SigProfilerExtractor performs 500 independent factorizations and, for each repetition, the matrix ***M*** is first Poisson resampled and normalized and, subsequently, factorized with the multiplicative update NMF algorithm[3] by minimizing an objective function based on the Kullback–Leibler divergence measure[27] (**Figure 1*b***). Custom partition clustering, that utilizes the Hungarian algorithm[28] for comparing different repetitions, is applied to the 500 factorizations to identify stable solutions[29] (**Figure 1*b***). Specifically, SigProfilerExtractor selects the centroids of stable clusters as optimal solutions, thus, making these solutions resistant to fluctuations in the input data and to the lack of uniqueness of NMF due to the potential existence of multiple convergent stationary points in the solution[30]. Lastly, when applicable, the optimal set of *de novo* signatures are matched to the set

118 of reference COSMIC mutational signatures (**Figure 1*c***) with any *de novo* signature reported as
119 novel when it cannot be decomposed by a combination of known COSMIC signatures.
120
121 ***Framework for benchmarking tools for de novo extraction of mutational signatures***
122 To allow comprehensive benchmarking of tools for *de novo* extraction of mutational signatures,
123 more than 200,000 synthetic cancer genomes were generated with known ground-truth
124 mutational signatures (**Supplementary Note 1**). These synthetic data included 32 distinct
125 noiseless scenarios and one scenario with five different levels of noise. Each scenario contained
126 between 3 and 40 known signatures operative in 200 to 3,000 simulated cancer genomes
127 (**Supplementary Tables 1–3**). Some of the scenarios were generated up to 20 times to account
128 for variability in the simulated data. The majority of noiseless scenarios (20/32) were based on
129 SBS-96 mutational classification; 12 scenarios based on extended mutational classifications, *i.e.,*
130 matrices with more than 96 mutational channels, were also included (**Supplementary Table 3**).
131 To avoid bias in evaluating each tool's performance, three sets of SBS-96 mutational signatures
132 were used for generating the synthetic data: *(i)* COSMICv3 reference signatures[11]; *(ii)* SA
133 signatures previously extracted by SignatureAnalyzer[11]; and *(iii)* randomly generated signatures.
134 For presentation simplicity, scenarios were labeled based on their complexity as easy, medium,
135 or hard. Easy scenarios were generated using ≤5 signatures and provide a good indication of each
136 tool's performance on approximately 7.4% of human cancer types (*e.g.*, pediatric brain tumors).
137 Medium scenarios contained 11 to 21 signatures and biologically reflect 15.9% of cancer types
138 (*e.g.*, cervical cancer). Hard scenarios have more than 25 signatures and reflect 59.5% of human
139 cancer types (*e.g.*, breast, lung, liver, *etc.*) as well as pan-cancer datasets. In addition to the 32

noiseless scenarios, one SBS-96 scenario with five different levels of noise (average noise per sample ranging between 0% and 10%) was included in the benchmark (**Supplementary Note 1**).

To compare the performance between different tools for *de novo* extraction of mutational signatures, we developed a standard set of evaluation metrics (**Supplementary Figure 1**). Specifically, each *de novo* extracted signature is classified as either a *true positive* (TP), *false positive* (FP), or *false negative* (FN) signature. An extracted signature is considered TP if it matches one of the ground-truth signatures above a cosine similarity threshold of 0.90. In contrast, a signature is classified as FP when it has a maximum cosine similarity below 0.90 with all ground-truth signatures. Lastly, FN signatures are ground-truth signatures that were not detected in the data. These standard metrics allow calculating each tool's precision, sensitivity, and F_1 score. Precision is defined as $\frac{TP}{TP+FP}$, sensitivity as $\frac{TP}{TP+FN}$, and F_1 score is the harmonic mean of the precision and sensitivity: $2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$.

Benchmarking SigProfilerExtractor and ten other tools using SBS-96 noiseless data

SigProfilerExtractor and ten other tools (**Table 1**) were first applied to all noiseless scenarios based on the SBS-96 mutational classification. Each tool was applied to each scenario by using its suggested method for selecting the number of operative signatures and, with the exception of SignatureAnalyzer which lacks this capability, by forcing each tool to extract the known number of ground-truth signatures. Results from the suggested approach reflect the expected outcome from running a tool on an unknown dataset, while results from the forced approach allow understanding limitations in each tool's implementation. Our evaluation reveals that most tools are able to successfully extract mutational signatures from easy scenarios with the majority of F_1

163 scores between 0.90 and 1.00 (**Figure 2*a***). This is perhaps unsurprising as many of these tools
164 used synthetic data with ≤5 signatures to evaluate their performance in the respective original
165 publications[15-25]. In contrast, medium scenarios have proven to be a challenge for the majority of
166 tools with only SigProfilerExtractor and SignatureAnalyzer exhibiting an F_1 score above 0.90.
167 All tools had worst performance for the hard set of scenarios with F_1 scores below 0.70; only
168 SigProfilerExtractor had an F_1 score above 0.85 (**Figure 2*a***).
169
170 To evaluate whether the type of ground-truth signatures affects the *de novo* extraction, we
171 compared the ratio of F_1 scores (rF_1) from scenarios generated using COSMIC, SA, or random
172 signatures (**Figure 2*b***). Most tools had similar performance ($rF_1 \approx 1$) between COSMIC and
173 random signatures and worst performance with SA signatures ($rF_1 < 1$). SomaticSignatures was an
174 exception as it performed well on random signatures but had similarly suboptimal performance
175 on COSMIC and SA signatures. SigProfilerExtractor outperformed all other tools regardless of
176 whether the synthetic data were generated using COSMIC, SA, or random signatures
177 (**Supplementary Table 1**).
178
179 To examine the performance of *de novo* extraction between the suggested and forced selection of
180 the total number of signatures, we evaluated rF_1 across all medium and hard scenarios (**Figure**
181 **2*c***). SigProfilerExtractor exhibited almost identical F_1 scores between the suggested and forced
182 selection indicating a good performance of the automatic selection algorithm. Most other tools
183 had similar F_1 scores between the suggested and forced selection albeit with more variability
184 across the different scenarios (**Figure 2*c***). For example, MutSpec had $rF_1 \approx 1$ in both medium and
185 hard scenarios indicating that MutSpec is performing worse than SigProfilerExtractor (**Figure**

186 **2*a***) not because of its algorithm for selecting the total number of signatures but likely due to its
187 implementation of the utilized numerical factorization. SigneR (hard scenarios), SigMiner
188 (medium), and SigFit (all) had lower F_1 scores for automatic solutions compared to forced
189 solutions ($rF_1<1$), thus, indicating that their automatic approaches for selecting the total number
190 of signatures are not optimally performing (**Figure 2*c***). Surprisingly, EMu had higher F_1 scores
191 for automatic solutions in some hard scenarios. Considering the overall performance of EMu
192 (**Figure 2*a***), this outcome likely reflects the lack of convergence during the minimization of the
193 EMu objective function for certain number of signatures in the hard scenarios.
194
195 Overall, across all suggested extractions from noiseless medium and hard scenarios,
196 SigProfilerExtractor outperformed all other tools. SigProfilerExtractor was able to identify
197 between 7% and 25% more true positive signatures while yielding between 3.7-and 16-fold less
198 false positive signatures compared to the next six best performing tools: SignatureAnalyzer,
199 SigneR, MutationalPatterns, MutSpec, SomaticSignatures, and SignatureTools (**Figure 2*d*** and
200 **Supplementary Table 1**).
201
202 ***Extended benchmarking of SigProfilerExtractor and the other six top performing tools***
203 The reported comparisons for SBS-96 scenarios rely on a cosine similarity ≥0.90 for determining
204 TP signatures and <0.90 for determining FP signatures. Note that a cosine similarity ≥0.90 is
205 highly unlikely to happen purely by chance (p-value = 5.90×10^{-9}) as two random nonnegative
206 vectors are expected to have an average cosine similarity of 0.75 purely by chance[31].
207 SigProfilerExtractor's performance does not depend on the specific value of the cosine similarity
208 threshold (**Figure 3*a***) as the tool consistently outperforms other bioinformatics approaches for

209 almost any value of the threshold above 0.80 (p-value: 0.057). Cosine similarity thresholds
210 below 0.80 were not explored as extracted signature may be similar to ground-truth signatures
211 purely by chance.
212
213 Additional benchmarking was performed by generating 12 scenarios simulated using between 3
214 and 30 signatures with an extended number of mutational channels (**Supplementary Note 1**).
215 SigProfilerExtractor and SignatureAnalyzer are the only two tools that support analysis of
216 custom size matrices and also provide GPU support (**Table 1**), thus, allowing analysis of data
217 with extended number of mutational channels within a reasonable timeframe. In contrast, all
218 other tools rely solely on CPU implementations with full runs expected to take many months for
219 each tool applied to these scenarios (**Table 1**). SigProfilerExtractor and SignatureAnalyzer
220 exhibited similar performance on the extended noiseless scenarios to that observed on SBS-96
221 noiseless scenarios. Overall, SigProfilerExtractor outperformed SignatureAnalyzer with average
222 F1 scores of 0.92 and 0.85, respectively (**Supplementary Table 2**).
223
224 To further compare SigProfilerExtractor with the other six top performing tools, we applied each
225 tool to a dataset with 30 ground-truth SBS-96 signatures operative in 1,000 genomes and random
226 noise between 0% and 10%. Analysis for each noise level was repeated 20 times to account for
227 any variability in the noise. SigProfilerExtractor, SomaticSignatures, MutSpec,
228 SignatureToolsLib were robust to noise with mostly unaffected performance (**Figure 3*b*** and
229 **Supplementary Table 3**). In contrast, SignatureAnalyzer, SigneR, and MutationalPatterns were
230 susceptible to noise (**Figure 3*b***). For example, 2.5% noise reduced SignatureAnalyzer's F_1 from
231 0.76 to 0.66 while 10% noise reduced its F_1 to 0.07. Similarly, 10% noise reduced the F_1 of

232 SigneR from 0.61 to 0.43 and the F_1 of MutationalPatterns from 0.60 to 0.37.
233 SignatureAnalyzer's reduced performance on data with noise is due to its automated approach
234 for selecting total number of signatures. SignatureAnalyzer uses automatic relevance
235 determination[32] for selecting the number of signatures with this number increasing from 26 (no
236 noise; 30 ground-truth signatures) to 96 signatures (10% noise; **Supplementary Table 3**). In
237 contrast, SigneR and MutationalPatterns exhibit similar performance between forced and
238 suggested solutions on data with noise (**Supplementary Table 3**) indicating that their reduced
239 performance is likely due to the numerical implementation of their respective factorization
240 approaches.
241
242 SigProfilerExtractor outperformed all other tools regardless of the levels of noise. Simulations
243 with 5% noise reflect genomics datasets with ~0.95 average sensitivity and precision of single
244 base substitutions, similar to the recently published PCAWG cohort which has 95% sensitivity
245 (90% confidence interval, 88–98%) and 95% precision (90% confidence interval, 71–99%)[26].
246 For simulations with 5% noise, SigProfilerExtractor was able to identify between 20% and 50%
247 more true positive signatures while yielding more than 5-fold less false positive signatures
248 compared to the next six best performing tools: SignatureAnalyzer, SigneR, MutationalPatterns,
249 MutSpec, SomaticSignatures, and SignatureTools (**Figure 3*c*** and **Supplementary Table 3**).
250
251 ***Analysis of 2,778 whole-genome sequenced human cancers with SigProfilerExtractor***
252 To demonstrate its ability to yield novel biological results, SigProfilerExtractor was applied to
253 the recently published set of 2,778 whole-genome sequenced cancers[26]. As previously done in
254 our original PCAWG analysis of mutational signatures[11], extraction of mutational signatures was

255 performed within each cancer type as well as across all samples (**Supplementary Data**). In
256 addition to all previously detected signatures[11], three novel mutational signatures were identified
257 in the PCAWG dataset: SBS92, SBS93, and SBS94 (**Figure 4** and **Supplementary Table 4**).
258
259 Signature SBS92 was found predominately in PCAWG bladder cancers; the signature was
260 characterized by T>C mutations with strong transcriptional strand bias consistent with damage
261 on purines for all types of single base substitutions (**Figure 4*a***). Signature SBS92 was 9-fold
262 elevated (**Figure 4*d***; p-value: 7.6×10^{-3} using Wilcoxon rank sum test) in bladder cancers of ever
263 smokers compared to never smokers in the PCAWG cohort. An almost identical signature was
264 identified by re-analyzing a recently published cohort of 88 whole-genome sequenced
265 microbiopsies of histologically normal urothelium[33] with the similarity extending to both
266 trinucleotide context and transcriptional strand bias (**Figure 4*a***; cosine similarity: 0.98; p-value <
267 10^{-256}). Indeed, this signature was reported in the original publication and found to be enriched in
268 smokers[33]. In our analysis, SBS92 was found to be 3-fold elevated in the normal urothelium of
269 tobacco ever smokers compared to never smokers (**Figure 4*d***; p-value: 8.3×10^{-3} using
270 Wilcoxon rank sum test).
271
272 Signature SBS93 was identified almost exclusively in PCAWG stomach cancers. SBS93 was
273 characterized by T>C and T>G mutations with a strand bias consistent with damage on
274 pyrimidines for Tp<u>T</u>pA contexts (mutated base underlined; **Figure 4*b***). *De novo* extraction from
275 the Mutographs cohort of 552 whole-genome sequenced esophageal squamous cell carcinomas[34],
276 a cancer type not included in the PCAWG dataset[26], identified an analogous mutational signature
277 with the similarity extending to both trinucleotide context and transcriptional strand bias (**Figure**

278 **4*b***; cosine similarity: 0.88; p-value: 1.1 x 10^{-6}). Signature SBS94 was found at high levels in a
279 single colorectal PCAWG cancer with smaller contributions to another 8 colorectal cancers. The
280 pattern of SBS94 was characterized by C>A mutations with a strand bias indicative of damage
281 on guanine (**Figure 4*c***). Validation of somatic mutations by visual inspection confirmed that
282 98% of mutations contributed by SBS94 are likely real. Signatures SBS93 and SBS94 did not
283 associate with any of the available PCAWG metadata[26] and their etiologies remain unknown.
284

DISCUSSION

The performed large-scale benchmarking demonstrates that SigProfilerExtractor outperforms ten other tools for *de novo* extraction of mutational signatures for noiseless datasets as well as for datasets containing matrices with different levels of random noise. Importantly, SigProfilerExtractor generates almost no false positive signatures while still identifying a higher number of true positive signatures when compared to any of the other tools (**Figure 2*d*** and **Figure 3*c***). *De novo* extraction of mutational signatures relies both on a factorization approach and on a model selection algorithm for determining the total number of operative signatures (**Figure 1**). Benchmarking with forced model selection, where tools were required to extract the known number of ground-truth mutational signatures, reveals that SigProfilerExtractor's factorization performs better when compared to the factorizations of other tools (**Figure 2*a*** and **Supplementary Tables 1-3**). Similarly, benchmarking with suggested model selection, which most closely matches analysis of a real dataset with unknown number of signatures, further demonstrates SigProfilerExtractor's ability to reveal novel biological results (**Figure 2*a*** and **Supplementary Tables 1-3**).

While our benchmarking evaluated ten additional tools, six of the ten tools internally rely on the same computational engine. Maftools, MutationalPatterns, MutSpec, SignatureToolsLib, SigMiner, and SomaticSignatures use the NMF R package[35] to perform their factorization (**Table 1**), albeit with slightly different hyperparameters and, in some cases, distinct pre-processing of the input matrix. Predictably, these six tools have similar performance across many of the scenarios (**Supplementary Tables 1-3**). SigFit also uses a previously developed nonnegative factorization method, *viz.*, Stan R package[36]. In contrast, EMu, SignatureAnalyzer, SigneR, and

308 SigProfilerExtractor provide original implementations of their factorization algorithms (**Table**
309 **1**). EMu was originally developed and tested on small datasets (*e.g.,* 21 breast genomes)[15] and its
310 benchmarking performance is perhaps unsurprising considering the large number of synthetic
311 samples used in all scenarios. Surprisingly, the original implementations of SignatureAnalyzer
312 and SigneR were susceptible to noise, yielding high numbers of false-positive signatures (**Figure**
313 **3*b***).
314
315 Six of the other ten tools did not provide an automatic approach for selecting the total number of
316 operative signatures in a dataset (**Table 1**). Instead, these tools offered methodologies for
317 manually selecting the optimal number of signatures bringing user-dependence and arbitrariness
318 in selecting solutions. EMu, SigFit, SignatureAnalyzer, SigMiner, SigneR, and
319 SigProfilerExtractor provided capabilities for automatically selecting the total number of
320 signatures. EMu and SigneR select the total number of signatures using Bayesian information
321 criterion (BIC)[37], while SignatureAnalyzer and SigMiner utilize automatic relevance
322 determination (ARD)[32]. SigFit's selection approach is based on the Elbow method[38].
323 SigProfilerExtractor leverages a modified version of the NMFk selection approach which was
324 previously tested on more than 55,000 synthetic random matrices with pre-determined latent
325 factors and shown to outperform other model selection approaches[39]. Importantly, our
326 simulations demonstrate that SigProfilerExtractor's model selection is robust to noise while the
327 implemented BIC and ARD approaches are affected even by low levels of noise (**Figure 3*b***).
328
329 In addition to outperforming ten other tools on simulated datasets, SigProfilerExtractor is able to
330 reveal additional biological results as demonstrated by identifying three novel signatures from

331 reanalysis of the PCAWG dataset. Importantly, SigProfilerExtractor identifies signature SBS92
332 (**Figure 4**) which is associated with tobacco smoking in whole-genome sequenced bladder
333 cancers and in whole-genome sequenced microbiopsies from normal bladder urothelium. The
334 strong transcriptional strand bias observed in SBS92 is indicative of an environmental mutagen
335 exposure that damages purines. Tobacco smoke is a complex mixture of at least 60 chemicals[14],
336 many capable of causing damage on purines. Interestingly, our and other prior analyses of exome
337 sequenced bladder cancers from The Cancer Genome Atlas (TCGA) project[14,40] did not reveal
338 SBS92. Reanalysis of the set of TCGA bladder cancer exomes[41] with SigProfilerExtractor was
339 also unable to detect SBS92 (**Supplementary Data**). We suspect that the lack of SBS92 in the
340 TCGA bladder cancers was due to the use of exome sequencing; note that SBS92 is
341 predominately found in intergenic regions (**Figure 4*a***) with most samples expected to have less
342 than 15 mutations from SBS92 in their exomes. To confirm this hypothesis, we downsampled the
343 whole-genome sequenced bladder cancers and the whole-genome sequenced microbiopsies from
344 normal bladder urothelium to exomes. SigProfilerExtractor's analysis of these downsampled
345 genomes was unable to detect SBS92 confirming that exome sequencing is insufficient to
346 identify signature SBS92 (**Supplementary Data**).
347
348 In summary, here we report SigProfilerExtractor – a computational tool for *de novo* extraction of
349 mutational signatures. We demonstrate that SigProfilerExtractor outperforms ten other tools by
350 conducting the largest benchmarking of bioinformatics approaches for extracting mutational
351 signatures. Further, we apply SigProfilerExtractor to 2,778 whole-genome sequenced cancers
352 and reveal several novel mutational signatures including a signature putatively attributed to
353 tobacco smoking mutagenesis in bladder cancer and in normal bladder epithelium.
354

355 **ONLINE METHODS**

356 **Computational implementation of SigProfilerExtractor and its seven modules**

357 The implementation of SigProfilerExtractor can be separated into seven distinct modules which
358 are packaged together into a single bioinformatics tool. *Module 1* processes the initial input data,
359 which can be provided as either a mutational catalogue containing a set of somatic mutations or a
360 mutational matrix. *Module 2* is responsible for resampling and normalization of the mutational
361 matrix prior to performing nonnegative matrix factorization. *Module 3* performs matrix
362 factorization using nonnegative matrix factorization with multiple replicates. *Module 4* utilizes
363 custom clustering to derive consensus solutions and to perform model selection. *Module 5*
364 decomposes the derived set of *de novo* signatures to a set of previously derived COSMIC
365 signatures. *Module 6* is responsible for calculating the activities of different signatures in
366 individual samples. *Module 7* handles the extensive outputting and plotting of the different
367 analysis performed by SigProfilerExtractor. In principle, each of these modules allows extensive
368 customization. SigProfilerExtractor provides a seamless integration of these seven modules that
369 allows using them in an orchestrated and preconfigured manner with little input from a user.

370

371 ***Module 1: Processing of input mutational catalogues or input mutational matrices***

372 SigProfilerExtractor deciphers mutational signatures from a mutational matrix $\boldsymbol{M}$ with t rows
373 and n columns; rows represent mutational channels while columns reflect individual cancer
374 samples (**Figure 1*a***). The value of each cell in the matrix, $\boldsymbol{M}$, corresponds to the number of
375 somatic mutations from a particular mutational channel in a given sample. The mutational matrix
376 can be provided as a text file with the first column containing the names of the mutational
377 channels and the first row containing the names of the examined samples. Alternatively, users

378 can provide a mutational catalogue of somatic mutations in a commonly used format (*e.g.,* VCF,
379 MAF, *etc.*) and this mutational catalogue will be internally converted into the appropriate
380 mutational matrix by SigProfilerMatrixGenerator[2].
381
382 ***Module 2: Resampling of the input mutational matrix and normalizing the resampled matrix***
383 SigProfilerExtractor does not factorize the original input matrix. Rather, prior to performing
384 matrix factorization, SigProfilerExtractor performs independent Poisson resampling of the
385 original matrix for each replicate[1]. As such, the matrix factorized in each replicate is never the
386 same for a given value of *k* (**Figure 1*b***). The resampling is performed to ensure that Poisson
387 fluctuations of the matrix do not impact the stability of the factorization results. Additional
388 normalization is performed after resampling to overcome potential skewing of the factorization
389 from any hypermutators. SigProfilerExtractor supports four standard normalization methods[42]:
390 *(i)* Gaussian mixture model (GMM) normalization (default); *(ii)* 100X normalization; *(iii)* log2
391 normalization; *(iv)* no normalization. *No normalization* does not perform any additional
392 transformation on the Poisson resampled matrix. In *log2 normalization*, the sum of each column
393 in the matrix is derived and logarithm with base 2 is calculated for each of these sums. Each cell
394 in a column of the matrix is multiplied by the log2 of the column-sum and subsequently divided
395 by the original column sum. In *100X normalization*, the sum of each column in the matrix is
396 derived. For each column where the sum exceeds 100 times the number of mutational channels
397 (*i.e.*, 100 times the number of rows in the matrix), each cell in the column is multiplied by the
398 100 times the number of mutational channels and subsequently divided by the original column
399 sum. This normalization ensures that no sample has a total number of mutations above 100 times
400 the number of mutational channels. *GMM normalization* encompasses a two-step process. The

401 first step derives the normalization cutoff value in a data-driven manner using a Gaussian
402 mixture model (GMM). The second step normalizes the appropriate columns using the derived
403 cutoff value. The first step uses a GMM to separate the samples into two groups based on their
404 total number of mutations; the total number of mutations in a sample reflects the sum of a
405 column in the matrix. The group with larger number of samples is subsequently selected, and the
406 same process is applied iteratively until it converges. Convergence is achieved when the mean of
407 the two groups is separated by no more than four standard deviations of the larger group. A
408 cutoff value is derived as the average value plus two standard deviations from the total number
409 of somatic mutations in the last large group. If the derived cutoff value is below 100 times the
410 number of mutational channels, the cutoff value is adjusted to 100 times the number of
411 mutational channels. For each column where the sum exceeds the derived cutoff value, each cell
412 in the column is multiplied by the cutoff value and subsequently divided by the original column
413 sum. Note that no normalization is performed if the means of the first two groups are not
414 separate by at least four standard deviations. In all cases, columns with a sum of zero, reflecting,
415 genomes without any somatic mutations, are ignored to avoid division by zero.
416
417
418 ***Module 3: Matrix Factorization Using Nonnegative Matrix Factorization with Replicates***
419 By default, SigProfilerExtractor factorizes the matrix $\boldsymbol{M}$ with different ranks searching for an
420 optimal solution between k=1 and k=25 mutational signatures. For each value of k, by default,
421 the tool performs 500 independent nonnegative matrix factorizations of the normalized Poisson
422 resampled input matrix. Thus, for each value of k, SigProfilerExtractor generates 500 distinct
423 factorizations of normalized Poisson resampled matrices resulting into 500 different matrices $\boldsymbol{S}$,

424 each matrix reflecting the patterns of the *de novo* mutational signatures, and 500 different
425 matrices ***A***, each matrix reflecting the activities of the *de novo* mutational signatures (**Figure 1*b***).
426 To perform each of these factorizations, SigProfilerExtractor utilizes a custom implementation of
427 the multiplicative update algorithm[3]. Specifically, SigProfilerExtractor initializes the ***S*** and ***A***
428 matrices in the first step of the factorization using either random initial conditions (default) or
429 one of the derivatives of nonnegative double singular vector decomposition[43].
430 SigProfilerExtractor provides internal support for minimizing three different objective functions
431 based on: *(i)* generalized Kullback-Leibler updates (default); *(ii)* Euclidean updates; *(iii)* Itakura-
432 Saito updates. By default, the tool performs all factorization using multithreading of central
433 processing units (CPUs) and also provides support for factorization using graphics processing
434 units (GPUs) by leveraging PyTorch[44]. In all cases, by default, the implemented minimization
435 performs at least 10,000 iterations (also known as NMF updates or NMF multiplicative update
436 steps) with a maximum of 1,000,000 iterations. By default, the convergence tolerance of the
437 algorithm is set to 10^{-15}. Note that SigProfilerExtractor allows configuring all factorization
438 parameters.
439
440 ***Module 4: Custom partition clustering and performing model selection***
441 The previously described *Module 3* generates a number of sets with each set containing, by
442 default, 500 different matrices ***S***, where each matrix reflects the patterns of *de novo* mutational
443 signatures for a particular factorization of a normalized Poisson resampled matrix. One set,
444 containing 500 different matrices ***S***, is generated for each of the interrogated total number of
445 operative signatures, *k*, with a default range for *k* between 1 and 25 signatures. For each value of
446 *k*, *Module 4* first performs custom clustering of the ***S*** matrices and, subsequently, applies a

447 modified version of the NMFk model selection approach to select the optimal value of k[39]

448 (**Figure 1*b***). Specifically, for each value of k, the clustering is initialized with k random

449 centroids. One of the $\boldsymbol{S}$ matrices is randomly chosen, and its columns matched to the most similar

450 centroids with no two columns assigned to the same cluster. The process is repeated until the

451 columns of all $\boldsymbol{S}$ matrices in the set are assigned to their respective clusters. SigProfilerExtractor

452 implements the Hungarian algorithm[28] to pair consensus vectors from two matrices (*i.e.,* cluster

453 centroids and mutational signature from a matrix $\boldsymbol{S}$); the Hungarian algorithm maximizes the

454 total cosine similarities of all paired vectors between two matrices[28]. After assigning all columns

455 to a cluster, the centroids of each cluster are recalculated by evaluating the average of all

456 columns/vectors in a cluster. This process continues iteratively until the average silhouette

457 coefficient converges (*i.e.,* its value does not change by more than 10^{-12}). After convergence for a

458 given value of k, the centroids of the clusters are reported as consensus mutational signatures, an

459 overall reconstruction error is calculated for describing the original input matrix, $\boldsymbol{M}$, and stability

460 is calculated for each signature by computing the silhouette value[45] of the cluster corresponding

461 to that signature (**Figure 1*b***). The silhouette value of a cluster measures the similarities of the

462 objects assigned to that cluster compared to any other cluster. Silhouette values range from -1.0

463 to +1.0 with values above zero indicating that, on average, objects have a higher similarity with

464 their own cluster compared to their nearest clusters. Note that signatures with low stability

465 correspond to a lack of uniqueness of the NMF due to Poisson resampling and/or to the potential

466 existence of multiple convergent stationary points in the NMF solution[30].

467

468 Our custom clustering is performed for each of the interrogated total number of operative

469 signatures, k, with a default range for k between 1 and 25 signatures. After performing clustering,

470 for each value of *k*, one has derived: *(i)* the consensus set of mutational signatures; *(ii)* an overall
471 reconstruction error for describing the original input matrix; and *(iii)* stability value for each of
472 the identified consensus mutational signatures.
473
474 SigProfilerExtractor performs a solution selection based on the stability of signatures in a
475 solution and the ability of these signatures to reconstruct the original input matrix. By default,
476 SigProfilerExtractor will consider solutions stable if the signatures derived in the solution have
477 an average stability above 0.80 with no individual signature having stability below 0.20. To
478 reduce overfitting, the tool also measures the information gained from the extracted set of
479 signatures in each solution. SigProfilerExtractor compares, using Wilcoxon rank-sum tests, the
480 reconstruction errors across all samples from the stable solution with the greatest number of
481 signatures to the reconstruction errors across all samples from stable solutions with lower
482 number of signatures. Stable solutions with lower number of signatures are compared in a
483 decreasing order to their total number of signatures with comparison stopping if the Wilcoxon
484 rank-sum test yields a *p-value* below 0.05 (*i.e.*, reflecting that a solution does not describe the
485 original data as good as the stable solution with the greatest number of signatures). The stable
486 solution with lowest number of signatures and a Wilcoxon rank-sum test *p-value* above 0.05 is
487 selected as the optimal solution. If no solution has a Wilcoxon rank-sum test *p-value* above 0.05,
488 the stable solution with the greatest number of signatures is selected as the optimal solution. Note
489 that while SigProfilerExtractor selects an optimal solution, it outputs all the information
490 necessary to evaluate mutational signatures and their activities for all other stable and unstable
491 solutions.
492

493 ***Module 5: Decomposing de novo extracted signatures to known COSMIC signatures***

494 SigProfilerExtractor provides a module for decomposing each of the *de novo* extracted
495 mutational signatures to a set of previously derived signatures. By default, the tool decomposes
496 each of the signatures in the optimal solution to a set of COSMICv3 reference signatures[11] with
497 support for signatures of single base substitutions (SBS), doublet base substitutions (DBS), and
498 small insertions and deletions (ID). Since the SBS COSMICv3 reference signatures were derived
499 under the SBS-96 classification[2], any extended classification of single base substitutions (*e.g.*,
500 SBS-288 and SBS-1536)[2] is first collapsed to the SBS-96 classification and, subsequently,
501 decomposed to the COSMICv3 reference signatures[11]. The decomposition functionality
502 leverages nonnegative least square (NNLS) algorithm[46] as its main computational engine. A
503 mixture of addition and removal steps (add-remove functionality) were developed to estimate the
504 list of COSMIC signatures for a *de novo* signature. Specifically, for each *de novo* signature, a
505 COSMIC signature is iteratively added to a list of signatures used to explain the *de novo*
506 signature, NNLS is applied, and the signature which addition causes the greatest decrease of the
507 L2 error is selected. If this greatest decrease is more than a specific threshold (default value of
508 0.05) then the signature is included in the list of signatures used to explain the *de novo* signature.
509 The addition is immediately followed by a removal step. Each COSMIC signature in the list of
510 signatures used to explain the *de novo* signature are iteratively removed, NNLS is applied, and
511 the signature that causes the least decrease of the L2 error is selected. If this least decrease is less
512 than a specific threshold (default value of 1%) then the signature is removed from the list of
513 signatures used to explain the *de novo* signature. The addition and removal steps are iterated until
514 no signatures are added or removed from the list of signatures used to explain the *de novo*
515 signature. Several previously implemented rules for mutational signatures are incorporated by

516 default in the decomposition module[11]. Specifically, for signatures of single base substitutions:
517 *(i)* the list of signatures used to explain the *de novo* signature is initialized with clock-like
518 signatures SBS1 and SBS5;[12] *(ii)* biologically connected signatures are included as previously
519 done in Ref [11] (*e.g.,* if SBS17a is included in the list then SBS17b is also included the list). The
520 decomposition module is highly customizable as it allows changing all default parameters as
521 well as adding additional new rules or removing existing rules for inclusion and exclusion of
522 particular signatures.
523
524 ***Module 6: Evaluating activities of mutational signatures in individual samples***
525 *De novo* extracted and COSMIC derived signatures are refitted to individual samples using
526 nonnegative least squares (NNLS)[46]. *Module 6* internally utilizes the add-remove functionality of
527 *Module 5* with each sample in the original matrix, ***M***, being individually examined. For *de*
528 *novo* mutational signatures, all *de novo* signatures are initially added to the list of signatures used
529 to explain the sample and a removal step with a cutoff of 2% is applied. To assign COSMIC
530 signatures in a sample, the module first derives the set of *de novo* signatures in that sample.
531 Decomposition to the COSMICv3 signatures using *Module 5* is performed for each of the *de*
532 *novo* signatures and the identified COSMICv3 signatures are refitted using the add-remove
533 functionality with a removal and addition cutoffs set at 5%. Finally, the activity matrix is
534 constructed by combining the activity vectors generated for all samples in the dataset.
535
536 ***Module 7: Outputting and plotting of analysis results***
537 All previous modules make use of *Module 7* for outputting and plotting of the generated results.
538 It should be noted that SigProfilerExtractor provides extensive output for the interrogated total

539 number of operative signatures, *k*, with a default range of *k* between 1 and 25 signatures. For
540 each value of *k*, SigProfilerExtractor outputs the set of operative *de novo* mutational signatures,
541 the activities of the operative signatures, and an extensive set of information related to individual
542 samples, individual *de novo* signatures, and the overall convergence of the factorization and
543 clustering. *Module 7* also provides additional information when ran in debug mode. In addition
544 to outputting information, SigProfilerExtractor also generates a bouquet of plots both for each
545 value of *k* as well as for the suggested optimal solution. SigProfilerExtractor utilizes all
546 previously implemented plots in SigProfilerPlotting[2] as well as includes a number of newly
547 developed plots.
548
549
550 **Analysis of the genomics data from cancer and normal somatic tissues**
551 For all examined cancer and normal somatic tissues, *de novo* extraction of mutational signatures
552 was performed with SigProfilerExtractor with default parameters using two distinct mutational
553 classifications: SBS-96 and SBS-288. The SBS-96 mutation classification incorporates the six
554 types of single base substitutions: C>A, C>G, C>T, T>A, T>C, and T>G. Each type of single
555 base substitution is further separated into 16 subtypes determined by the four possible bases 5'
556 and 3' adjacent to each mutated base. The SBS-288 mutation classification extends the SBS-96
557 mutation classification by adding additional information for each of the 96 subtypes.
558 Specifically, SBS-288 incorporates whether a single base substitution is in non-
559 transcribed/intergenic DNA, on the transcribed strand of a gene, or on the untranscribed strand of
560 the gene. *De novo* extraction was performed separately for all examined datasets. Specifically,
561 SigProfilerExtractor was applied: *(i)* to all 2,778 whole-genome sequenced cancers from the Pan-
562 Cancer Analysis of Whole Genomes project[26]; *(ii)* to all samples in each of the 37 cancer types

563 of Pan-Cancer Analysis of Whole Genomes project[26] with each cancer type examined separately;
564 *(iii)* to all 88 whole-genome sequenced microbiopsies of histologically normal urothelium[33]; *(iv)*
565 to the complete set of bladder cancers from TCGA[41]; *(v)* to exome downsampling of all bladder
566 whole-genome sequenced cancers from the Pan-Cancer Analysis of Whole Genomes project[26];
567 *(vi)* to exome downsampling of all 88 whole-genome sequenced microbiopsies of histologically
568 normal urothelium[33]. In all cases, the mutational catalogues of each samples were taken from the
569 respective original publications. The results from all performed *de novo* extractions can be found
570 in **Supplementary Data**. Downsampling of a whole-genome sequenced sample to a whole-
571 exome was performed using SigProfilerMatrixGenerator[2].
572
573 **Additional approaches for miscellaneous analysis**
574 Synthetic scenarios were labeled as easy, medium, and hard based on the number of operative
575 signatures in each scenario. Based on our most recent analysis of mutational signatures in 82
576 cancer types[11], approximately 7.4% of human cancer types have 5 or less signatures (reflected in
577 simulations of easy scenarios), 15.9% have 11 to 21 signatures (medium scenarios), and 59.5%
578 have 25 or more signatures (hard scenarios). Note that 17.2% of cancer types have either
579 between 5 and 10 signatures or between 22 and 24 signatures.
580
581 Cosine similarity was used to compare the profiles of different mutational signatures. P-values
582 can be attributed to cosine similarities based on a null hypothesis of uniform random distribution
583 of nonnegative vectors[31].
584

585 Briefly, the prevalence of somatic mutations in a whole-exome sample was calculated based on
586 the identified mutations in protein coding genes and assuming that an average whole-exome has
587 sufficient coverage of 30.0 megabase-pairs in protein coding genes. The prevalence of somatic
588 mutations in a whole-genome sample was calculated based on all identified mutations and
589 assuming that an average whole-genome has sufficient coverage of 3.00 gigabase-pairs.
590
591 All methods related to the generation of the benchmarking scenarios and the application of the
592 different tools to these scenarios can be found in **Supplementary Note 1**.
593

TABLES

Tool Name	Input	Platform	Factorization Approach			Selection Approach			Supported Contexts		COSMIC Comparison
			Method	Computational Engine	GPU	Manual	Automatic	Automatic Algorithm	Mutational Catalogue Support	Plotting	
EMu[15]	Matrix	C++	EM	Original implementation[15]	No	Yes	Yes^	BIC[37]	SBS-96	No	No
Maftools[16]	Matrix MAF	R-Bioconductor	NMF	NMF R package[35]	No	Yes	No	-	SBS-96	SBS-96	1-to-1
MutationalPatterns[17]	Matrix VCF	R-Bioconductor	NMF	NMF R package[35]	No	Yes	No	-	SBS-96 SBS-192	SBS-96 SBS-192	1-to-1
MutSpec[18]	Matrix VCF Custom	Galaxy Perl R	NMF	NMF R package[35]	No	Yes	No	-	SBS-96 SBS-192	SBS-96 SBS-192	1-to-1
SigFit[19]	Matrix	R	Bayesian Inference	Stan R package[36]	No	Yes	Yes^	Elbow method[38]	SBS-96	SBS-96 SBS-192	1-to-1
SigMiner[20]	Matrix MAF	R	[automatic] Bayesian NMF [manual] NMF	[automatic] SignatureAnalyzer implementation[21] [manual] NMF R package[35]	No	Yes^	Yes	ARD[32]	SBS-96 DBS-78 ID-83	Generic	1-to-1
SignatureAnalyzer[21,22]	Matrix MAF	R [CPU][18] Python [GPU][19]	Bayesian NMF	Original implementation[21,22]	Yes	No	Yes	ARD[32]	SBS-96 DBS-78 ID-83	SBS-96 DBS-78 ID-83	1-to-1
SignatureToolsLib[23]	Matrix VCF Custom	R	NMF	NMF R package[35]	No	Yes	No	-	SBS-96 DBS-78 ID-83 SV-32	SBS-96 SV-32 Generic	No
SigneR[24]	Matrix VCF	R-Bioconductor C++	Bayesian NMF	Original implementation[24]	No	Yes	Yes^	BIC[37]	SBS-96	SBS-96	No
SigProfilerExtractor	Matrix VCF MAF Custom	Python R wrapper	NMF	[current report] Original implementation	Yes	Yes	Yes^	NMFk[39]	SBS-96 DBS-78 ID-83 Others[2] Any	SBS-96 DBS-78 ID-83 SV-32 Others[2] Generic	1-to-many
SomaticSignatures[25]	Matrix VCF	R-Bioconductor	NMF PCA	NMF R package[35] pcaMethods R package[47]	No	Yes	No	-	SBS-96	SBS-96	No

Table 1: Overview of bioinformatics tools for *de novo* extraction of mutational signatures. Tools are ordered alphabetically. Notations: ^ denotes the default approach for selecting the total number of signatures when a tool supports both manual and automatic selection; 1-to-1 refers to one *de novo* signature being matched with exactly one COSMIC signature; 1-to-many refers to

601 one *de novo* signature being matched with a combination of one or more COSMIC signatures.
602 Abbreviations: MAF: mutation annotation format; VCF: variant call format; EM: expectation-
603 maximization algorithm; NMF: nonnegative matrix factorization; ARD: automatic relevance
604 determination; BIC: Bayesian information criterion; COSMIC: catalogue of somatic mutations in
605 cancer; SBS: single base substitutions; DBS: doublet base substitutions; ID: small insertions and
606 deletions; SV: structural variants.
607

608 **FIGURE LEGENDS**

609 **Figure 1. Overview of SigProfilerExtractor.** ***(a)*** SigProfilerExtractor's general workflow is
610 outlined starting from an input of somatic mutations and resulting in an output of *de novo*
611 mutational signatures. An example is shown for a solution with three *de novo* signatures.
612 Somatic mutations are first converted into a mutational matrix. Subsequently, the matrix is
613 factorized with different ranks using nonnegative matrix factorization. Model selection is applied
614 to identify the optimal factorization rank based on each solution's stability and its reconstruction
615 of the original data. ***(b)*** Schematic representation for an example decomposition with a
616 factorization rank of $k=3$. By default, SigProfilerExtractor performs 500 independent
617 nonnegative matrix factorizations with the matrix ***M*** being Poisson resampled and normalized
618 (denoted by ^) prior to each factorization. Partition clustering of the 500 factorizations is used to
619 evaluate the factorization stability rank, measured in silhouette values; clustering can also be
620 presented as two-dimensional projections revealing more similar mutational signatures as shown
621 for the three example signatures. The centroid of the clustered solutions (denoted by –) is
622 compared to the original matrix ***M***. ***(c)*** All identified *de novo* signatures are matched to a
623 combination of known COSMIC mutational signatures. An example is given for *de novo*
624 extracted signature SBS96B which matches a combination of COSMIC signatures SBS1, SBS2,
625 and SBS13.

626

627 **Figure 2. Benchmarking of bioinformatics tools for *de novo* extraction of mutational**
628 **signatures using SBS-96 noiseless scenarios.** ***(a)*** Average precision (x-axes), sensitivities (y-
629 axes), and F_1 scores (harmonic mean of precision and sensitivity; red curves) are shown across
630 the three types of scenarios. Different tools are displayed using circles and triangles with

631 different colors. Circles are used to display results for suggested model selection, which most
632 closely matches analysis of a real dataset. Triangles are used to display results for forced model
633 selection, where tools were required to extract the known total number of ground-truth
634 mutational signatures. All triangles are located on the diagonal as the forced model selection
635 results in equal numbers of false positive and false negative signatures. ***(b)*** Evaluating the effect
636 of ground-truth signatures on the *de novo* extraction by different tools (x-axes). Ratio of F_1
637 scores (y-axes) with confidence intervals were calculated for medium complexity scenarios
638 simulated using COSMIC, SA, or random signatures. Ratio of approximately 1.00 indicates a
639 similar performance between different types of signatures. ***(c)*** Evaluating the performance of *de*
640 *novo* extraction between suggested and forced selection for different tools (x-axes). Ratio of F_1
641 scores (y-axes) with confidence intervals were calculated for all medium and hard scenarios.
642 Ratio of approximately 1.00 indicates a similar performance between suggested and forced
643 model selection. ***(d)*** Summary of the performance for the top seven tools on medium and hard
644 SBS-96 noiseless scenarios with suggested model selection. Y-axes reflect F_1 score (left plot),
645 sensitivity (middle plot), and false discovery rate (right plot), respectively. Results from
646 SignatureAnalyzer are not displayed in panels ***(a)***, ***(b)***, and ***(c)*** for forced model selection as
647 SignatureAnalyzer does not support predefined/forced solution with a specific total number of
648 signatures.
649
650 **Figure 3. Additional evaluations of the top seven bioinformatics tools for *de novo* extraction**
651 **of mutational signatures.** ***(a)*** Average F_1 scores for the top seven tools based on different
652 thresholds for cosine similarity in suggested medium and hard scenarios; thresholds for cosine
653 similarity are used for determining true positive signatures (**Supplementary Figure 1**). X-axes

654 reflect the cosine similarity thresholds, while the Y-axes correspond to the average F_1 scores
655 corresponding to cosine similarity thresholds. ***(b)*** Precision and sensitivity of the top seven tools
656 for SBS-96 scenarios with different levels of noise. Noise levels reflect the average number of
657 somatic mutations in a cancer genome affected by additive white Gaussian noise; for example,
658 1% noise corresponds to approximately 1% of mutations in a sample being due to noise. ***(c)***
659 Summary of the performance of the top seven tools on SBS-96 scenarios with 5% noise. Y-axes
660 reflect F_1 score (left plot), sensitivity (middle plot), and false discovery rate (right plot),
661 respectively.
662
663 **Figure 4. Novel signatures identified in the PCAWG cohort of 2,778 whole-genome**
664 **sequenced cancers.** Mutational signatures are displayed using 96-plots. Single base substitutions
665 are shown using the six subtypes of substitutions: C>A, C>G, C>T, T>A, T>C, and T>G.
666 Underneath each subtype are 16 bars reflecting the sequence contexts determined by the four
667 possible bases 5' and 3' to each mutated base. Additional information whether mutations from a
668 signature are in non-transcribed/intergenic DNA, on the transcribed strand of a gene, or on the
669 untranscribed strand of the gene is provided adjacent to the 96 plots. ***(a)*** Mutational profile of
670 signature SBS92 derived from the PCAWG cohort (top). Confirmation of the profile of signature
671 SBS92 (bottom) by analysis of an independent whole-genome sequenced set of microbiopsies of
672 histologically normal urothelium[33]. ***(b)*** Mutational profile of signature SBS93 derived from the
673 PCAWG cohort (top). Confirmation of the profile of signature SBS93 (bottom) by analysis of an
674 independent whole-genome sequenced set of esophageal squamous cell carcinomas[34]. ***(c)***
675 Mutational profile of signature SBS94 derived from the PCAWG cohort. Signature SBS94 was
676 not identified in any additional independent cohort. ***(d)*** Bars are used to display average values

677 for numbers of somatic substitutions per megabase (Mb) attributed to signature SBS92 in bladder
678 cancer and normal bladder urothelium. Green bars represent never smokers, whereas blue bars
679 correspond to ever smokers. Error bars correspond to 95% confidence intervals. Each p-value is
680 based on a Wilcoxon rank sum test.
681

682 **SUPPLEMENTARY INFORMATION**

683 **Supplementary Figure 1. Standard set of performance metrics used for benchmarking all**
684 **bioinformatics tools.** An example demonstrating the derivation of *true positive* (TP), *false*
685 *positive* (FP), or *false negative* (FN) signatures for a tool applied to a synthetic dataset generated
686 using 6 ground truth signatures (termed, Ground Truth Signatures 1 through 6). The tool extracts
687 4 signatures (termed, Extracted Signatures A through D). In this example, an extracted signature
688 is considered a true positive if it matches one of the ground-truth signatures with a cosine
689 similarity threshold of at least 0.90.

690

691 **Supplementary Table 1. Detailed performance metrics after applying each tool across all**
692 **SBS-96 noiseless synthetic scenarios.** Performance metrics are calculated as per *Supplementary*
693 *Figure 1*. An extracted signature is considered a true positive if it matches one of the ground-
694 truth signatures with a cosine similarity threshold of at least 0.90.

695

696 **Supplementary Table 2. Detailed performance metrics after applying the seven best**
697 **performing tools across SBS-96 synthetic scenarios with different levels of noise.**
698 Performance metrics are calculated as per *Supplementary Figure 1*. An extracted signature is
699 considered a true positive if it matches one of the ground-truth signatures with a cosine similarity
700 threshold of at least 0.90.

701

702 **Supplementary Table 3. Detailed performance metrics of applying SigProfilerExtractor**
703 **and SignatureAnalyzer to extended synthetic scenarios.** Performance metrics are calculated

704 as per *Supplementary Figure 1*. An extracted signature is considered a true positive if it matches
705 one of the ground-truth signatures with a cosine similarity threshold of at least 0.90.
706
707 **Supplementary Table 4. Profiles of three novel mutational signatures identified in the**
708 **PCAWG cohort of 2,778 whole-genome sequenced cancers.** The profiles of the novel
709 mutational signatures are reported using the SBS-288 classification which incorporates the
710 trinucleotide context and strand information (intergenic region, untranscribed strand, or
711 transcribed strand) for each type of single base substitution. The SBS-288 classification can be
712 easily collapsed to the commonly used SBS-96 classification.
713
714 **Supplementary Note 1. Detailed description of the performed benchmarking.** The
715 supplementary note provides extensive details about each of the generated synthetic scenarios as
716 well as about applying each of the tools to these scenarios. The results from applying all tools to
717 all scenarios, including appropriate input and out files, can be found in **Supplementary Data**.
718
719 **Supplementary Data**
720 All results from the benchmarking with synthetic datasets, including the appropriate input used
721 to run each of the tools as well as the output generated by each of the tools, can be found at:
722 ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/Benchmark/.
723
724 All results from the *de novo* extraction of mutational signatures from the PCAWG dataset can be
725 found at: ftp://alexandrovlab-
726 ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/PCAWG_Reanalysis/.

All results from the *de novo* extraction of mutational signatures for confirming the patterns of the novel signatures for additional datasets can be found at: ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/Confirmation_of_Novel_Signatures/.

All results from the *de novo* extraction of mutational signatures from downsampling of whole-genome sequenced samples to whole-exomes can be found at: ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/Downsampling_of_whole_genomes/

ACKNOWLEDGEMENTS

The authors would like to thank Allan Balmain (UC San Francisco) for the many useful discussions as well as Ville Mustonen (University of Helsinki) and Israel Tojal Da Silva (A.C. Camargo Cancer Center) for help in configuring EMu and SigneR, respectively. This work was supported by Cancer Research UK Grand Challenge Award C98/A24032 (LBA, PB, and MRS), Wellcome grant reference 206194 (MRS), and US National Institute of Health grants R01MH116281-01A1 (BSA) and R01ES030993-01A1 (LBA). This work was also supported by Singapore National Medical Research Council grants NMRC/CIRG/1422/2015 and MOH-000032/MOHCIRG18may-0004 and the Singapore Ministry of Health via the Duke-NUS Signature Research Programmes. LBA is an Abeloff V Scholar and he is supported by an Alfred P. Sloan Research Fellowship. Research at UC San Diego was also supported by a Packard Fellowship for Science and Engineering to LBA. AJG was funded by a postdoctoral fellowship (grant nr. P2BSP3_178591). NP receives funding through the Cancer Research UK Clinician Scientist Fellowship scheme and is supported by University College London Cancer Institute. Research at Los Alamos National Laboratory was conducted under Contract No. 89233218CNA000001 by the U.S. Department of Energy's National Nuclear Security Administration and supported by Laboratory Directed Research and Development (LDRD) grant 20190020DR (BSA). CDS is supported by the GEM consortium and acknowledges funding for this work through a Cancer Research UK travel grant. The funders had no roles in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

LBA and SMAI designed both SigProfilerExtractor's methodology and the performed analyses with help from NP, JZ, DJA, IM, BSA, LH, DCW, MTL, PB, MRS, and SGR. SMAI developed SigProfilerExtractor with help from MV, ENB, YH, CDS, RV, and JW. All synthetic benchmarking datasets were generated by YW and SGR. SMAI documented SigProfilerExtractor and performed the benchmarking of all tools on synthetic data with help from MDG, MB, BO, AK, and AA. Additional validations, confirmations, and applications of SigProfilerExtractor to real and synthetic datasets were performed by SMAI, SM, SS, YRL, NS, LR, TZ, AJG, YH, CDS, and SWB. LBA directed the overall research and wrote the manuscript with help from SMAI and input from all other authors. All authors read and approved the final manuscript.

COMPETING INTERESTS

MV is an employee of NVIDIA corporation. All other authors declare no competing interests.

TOOL AVAILABILITY

SigProfilerExtractor and all its modules are open source and freely available for use under the permissive 2-clause BSD license. SigProfilerExtractor and its modules are implemented in Python with an R wrapper package allowing users to run the tool from an R environment. SigProfilerExtractor can be installed using the PyPI package manager from https://pypi.org/project/SigProfilerExtractor/ or downloaded from GitHub from https://github.com/AlexandrovLab/SigProfilerExtractor. The R version of the tool can be downloaded from https://github.com/AlexandrovLab/SigProfilerExtractorR. A detailed wiki

page including installation, usage, and explanation of result is provided at https://osf.io/t6j7u/wiki/home/. SigProfilerExtractor is compatible with Windows, Linux, Unix, and macOS operating systems.

REFERENCE

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

2 Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685, doi:10.1186/s12864-019-6041-2 (2019).

3 Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791, doi:10.1038/44565 (1999).

4 Févotte, C. & Cemgil, A. T. in *2009 17th European Signal Processing Conference.* 1913-1917.

5 Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1-22, doi:10.1111/j.2517-6161.1977.tb01600.x (1977).

6 Suri, P. & Roy, N. R. in *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT).* 1-5.

7 Alexandrov, L. B. Understanding the origins of human cancer. *Science* **350**, 1175, doi:10.1126/science.aad7363 (2015).

8 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

9 Petljak, M. & Alexandrov, L. B. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* **37**, 531-540, doi:10.1093/carcin/bgw055 (2016).

10 Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat Genet* **51**, 1732-1740, doi:10.1038/s41588-019-0525-5 (2019).

11 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).

12 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).

13 Alexandrov, L. B., Nik-Zainal, S., Siu, H. C., Leung, S. Y. & Stratton, M. R. A mutational signature in gastric cancer suggests therapeutic strategies. *Nat Commun* **6**, 8683, doi:10.1038/ncomms9683 (2015).

14 Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618-622, doi:10.1126/science.aag0299 (2016).

15 Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol* **14**, R39, doi:10.1186/gb-2013-14-4-r39 (2013).

16 Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747-1756, doi:10.1101/gr.239244.118 (2018).

17 Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**, 33, doi:10.1186/s13073-018-0539-0 (2018).

18 Ardin, M. *et al.* MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17**, 170, doi:10.1186/s12859-016-1011-z (2016).
19 Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv*, 372896, doi:10.1101/372896 (2020).
20 Wang, S. *et al.* Copy number signature analyses in prostate cancer reveal distinct etiologies and clinical outcomes. *medRxiv*, 2020.2004.2027.20082404, doi:10.1101/2020.04.27.20082404 (2020).
21 Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun* **6**, 8866, doi:10.1038/ncomms9866 (2015).
22 Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* **20**, 228, doi:10.1186/s13059-019-1836-7 (2019).
23 Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat Cancer* **1**, 249-263, doi:10.1038/s43018-020-0027-5 (2020).
24 Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8-16, doi:10.1093/bioinformatics/btw572 (2017).
25 Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673-3675, doi:10.1093/bioinformatics/btv408 (2015).
26 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).
27 Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Statist.* **22**, 79-86, doi:10.1214/aoms/1177729694 (1951).
28 Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**, 83-97, doi:10.1002/nav.3800020109 (1955).
29 Huang, K., Sidiropoulos, N. D. & Swami, A. Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition. *IEEE Transactions on Signal Processing* **62**, 211-224, doi:10.1109/TSP.2013.2285514 (2014).
30 Lin, C. On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks* **18**, 1589-1596, doi:10.1109/TNN.2007.895831 (2007).
31 Bergstrom, E. N., Barnes, M., Martincorena, I. & Alexandrov, L. B. Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics* **21**, 438, doi:10.1186/s12859-020-03772-3 (2020).
32 Tan, V. Y. F. & Févotte, C. Automatic Relevance Determination in Nonnegative Matrix Factorization with the /spl beta/-Divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1592-1605, doi:10.1109/TPAMI.2012.240 (2013).
33 Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75-82, doi:10.1126/science.aba8347 (2020).
34 Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence. *bioRxiv, 372XXX* (2020).
35 Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367, doi:10.1186/1471-2105-11-367 (2010).

36 Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76**, doi:10.18637/jss.v076.i01 (2017).

37 Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461-464, doi:10.1214/aos/1176344136 (1978).

38 Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267-276, doi:10.1007/bf02289263 (1953).

39 Benjamin, N., Raviteja, V., Miguel, A. H.-H., Svetlana, K. & Boian, A. A neural network for determination of latent dimensionality in Nonnegative Matrix Factorization. *Machine Learning: Science and Technology* (2020).

40 Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606, doi:10.1038/ng.3557 (2016).

41 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322, doi:10.1038/nature12965 (2014).

42 Shalabi, L. A. & Shaaban, Z. in *2006 International Conference on Dependability of Computer Systems.* 207-214.

43 Žitnik, M. & Zupan, B. Nimfa: A python library for nonnegative matrix factorization. *The Journal of Machine Learning Research* **13**, 849-853 (2012).

44 Lew, J. *et al.* in *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS).* 151-152.

45 Aranganayagi, S. & Thangavel, K. in *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007).* 13-17 (IEEE).

46 Franc, V., Hlaváč, V. & Navara, M. in *Computer Analysis of Images and Patterns.* (eds André Gagalowicz & Wilfried Philips) 407-414 (Springer Berlin Heidelberg).

47 Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. pcaMethods--a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164-1167, doi:10.1093/bioinformatics/btm069 (2007).

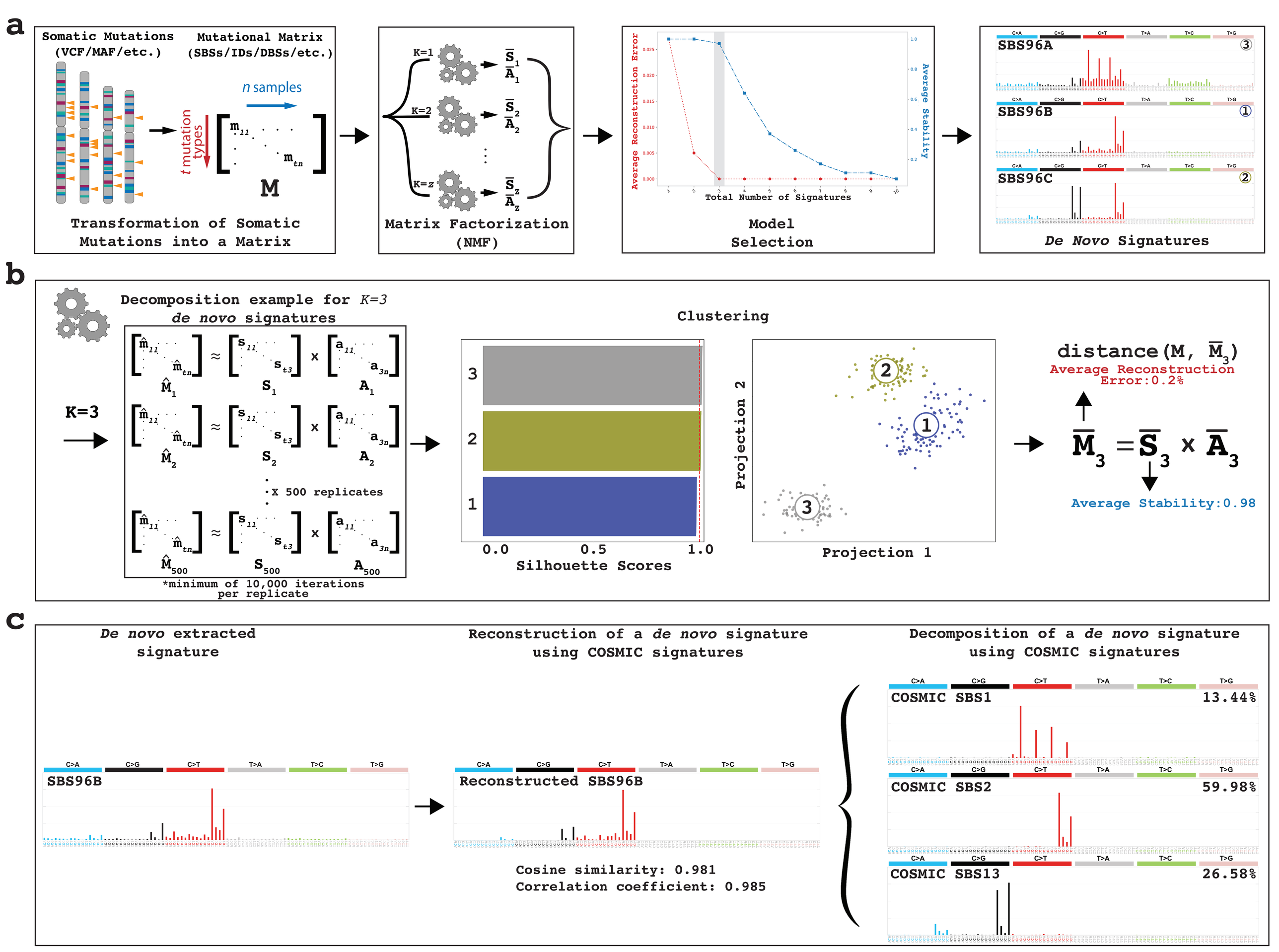
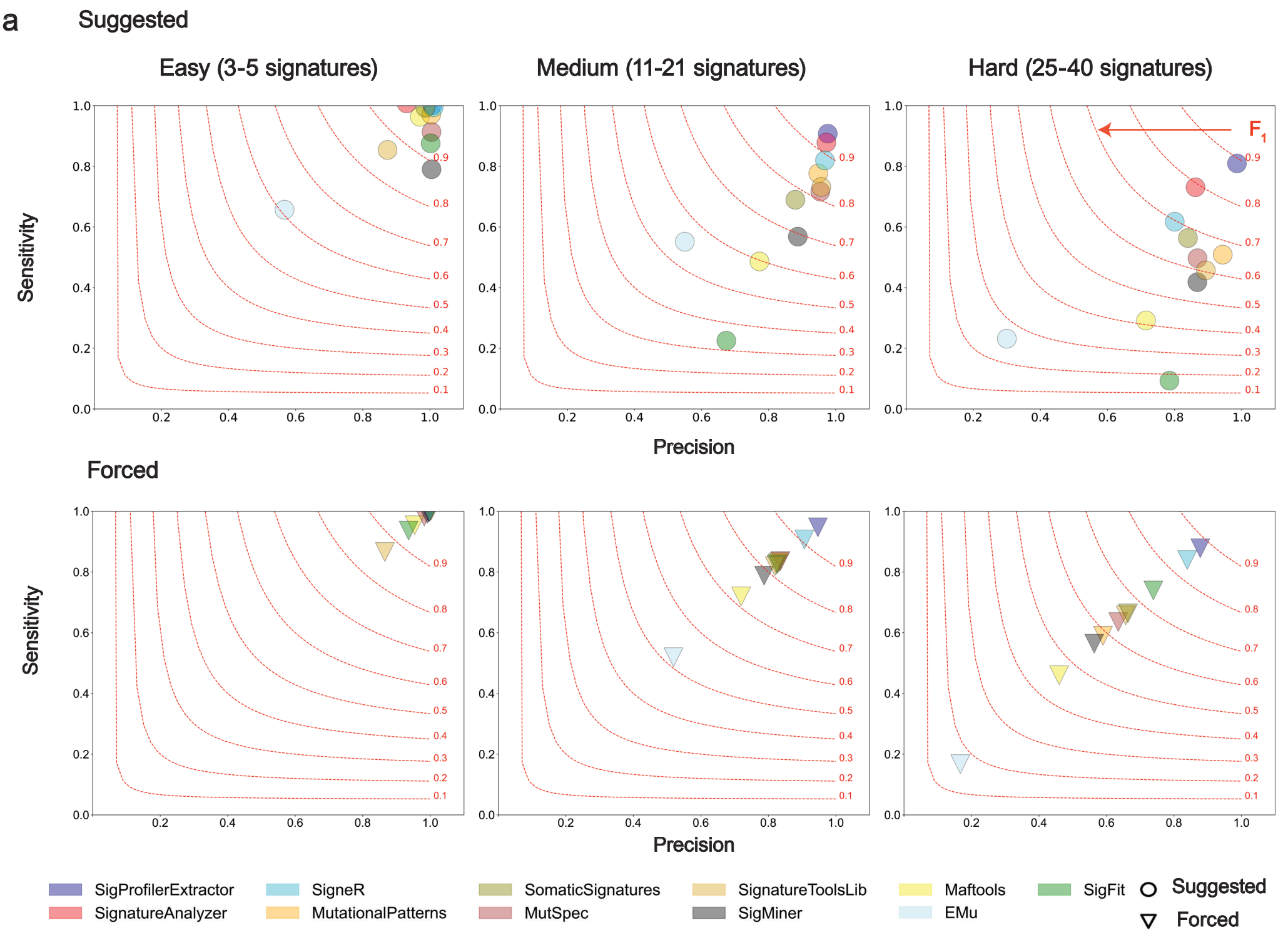
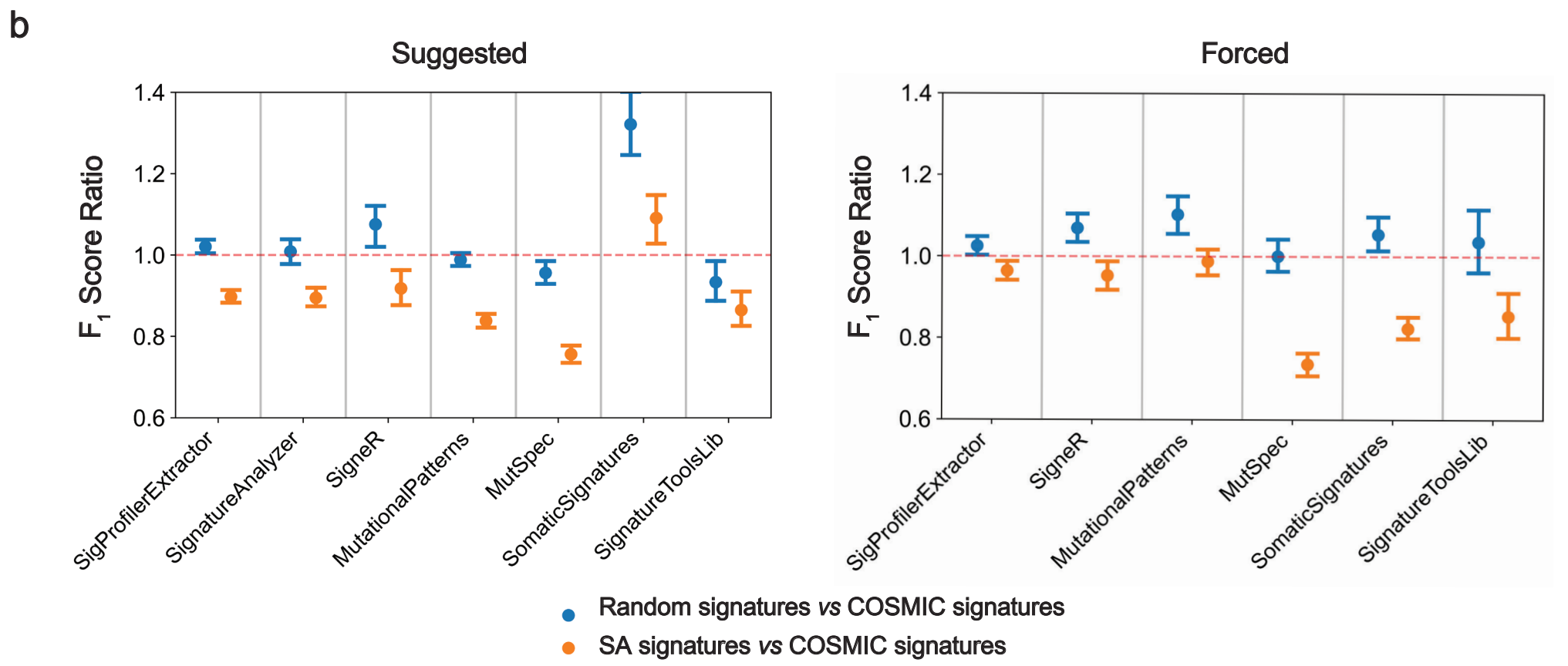
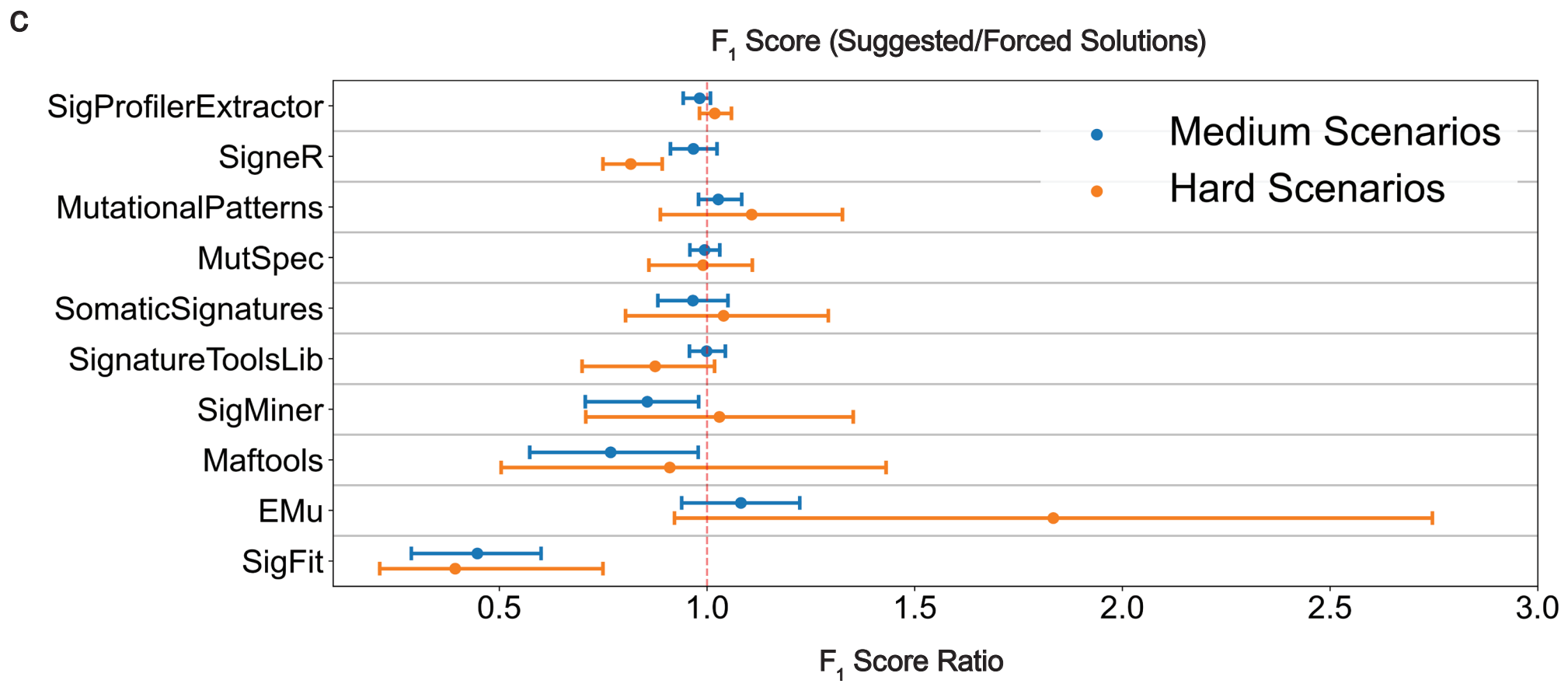
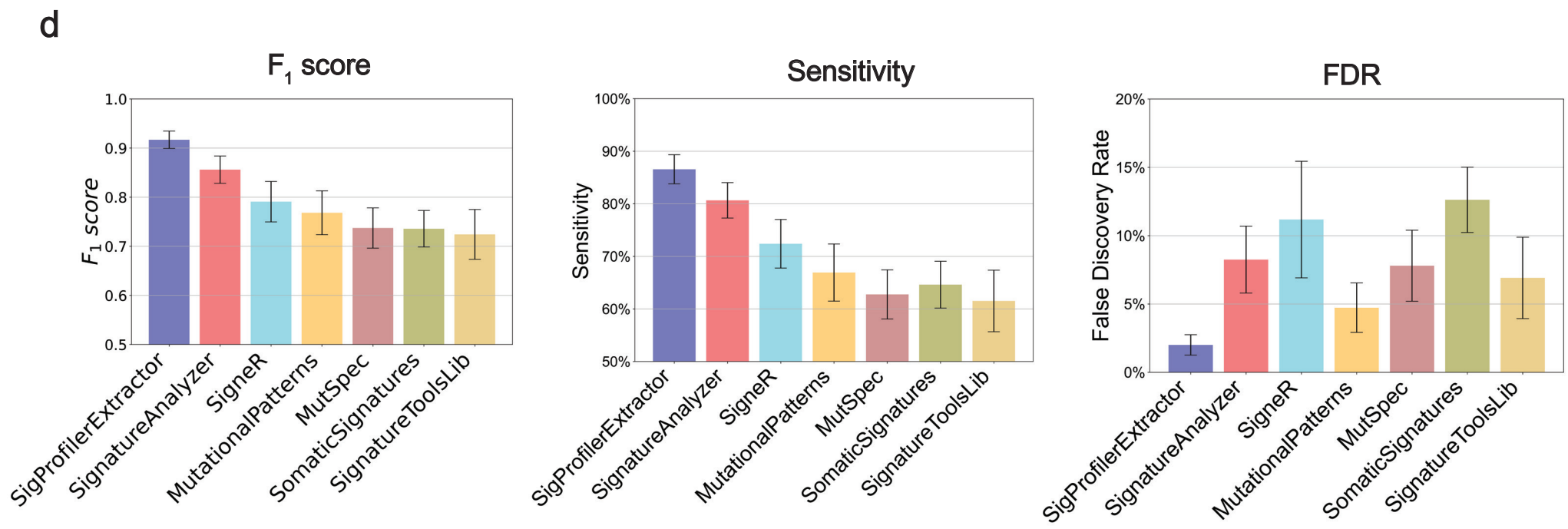

a
Somatic Mutations (VCF/MAF/etc.)
Mutational Matrix (SBSs/IDs/DBSs/etc.)
n samples
t mutation types
M
Transformation of Somatic Mutations into a Matrix
K=1
K=2
K=z
Matrix Factorization (NMF)
Average Reconstruction Error
Average Stability
Total Number of Signatures
Model Selection
SBS96A
SBS96B
SBS96C
De Novo Signatures
b
Decomposition example for K=3 de novo signatures
K=3
X 500 replicates
*minimum of 10,000 iterations per replicate
Clustering
Silhouette Scores
Projection 1
Projection 2
distance(M, M3)
Average Reconstruction Error:0.2%
Average Stability:0.98
c
De novo extracted signature
SBS96B
Reconstruction of a de novo signature using COSMIC signatures
Reconstructed SBS96B
Cosine similarity: 0.981
Correlation coefficient: 0.985
Decomposition of a de novo signature using COSMIC signatures
COSMIC SBS1 13.44%
COSMIC SBS2 59.98%
COSMIC SBS13 26.58%

Fig.2


a
Suggested
Easy (3-5 signatures)
Medium (11-21 signatures)
Hard (25-40 signatures)
Sensitivity
Precision
F1
Forced
SigProfilerExtractor
SignatureAnalyzer
SigneR
MutationalPatterns
MutSpec
SomaticSignatures
SignatureToolsLib
SigMiner
Maftools
EMu
SigFit
Suggested
Forced


b
Suggested
Forced
F1 Score Ratio
SigProfilerExtractor
SignatureAnalyzer
SigneR
MutationalPatterns
MutSpec
SomaticSignatures
SignatureToolsLib
Random signatures vs COSMIC signatures
SA signatures vs COSMIC signatures


c
F1 Score (Suggested/Forced Solutions)
SigProfilerExtractor
SigneR
MutationalPatterns
MutSpec
SomaticSignatures
SignatureToolsLib
SigMiner
Maftools
EMu
SigFit
Medium Scenarios
Hard Scenarios
F1 Score Ratio


d
F1 score
Sensitivity
FDR
False Discovery Rate
SigProfilerExtractor
SignatureAnalyzer
SigneR
MutationalPatterns
MutSpec
SomaticSignatures
SignatureToolsLib

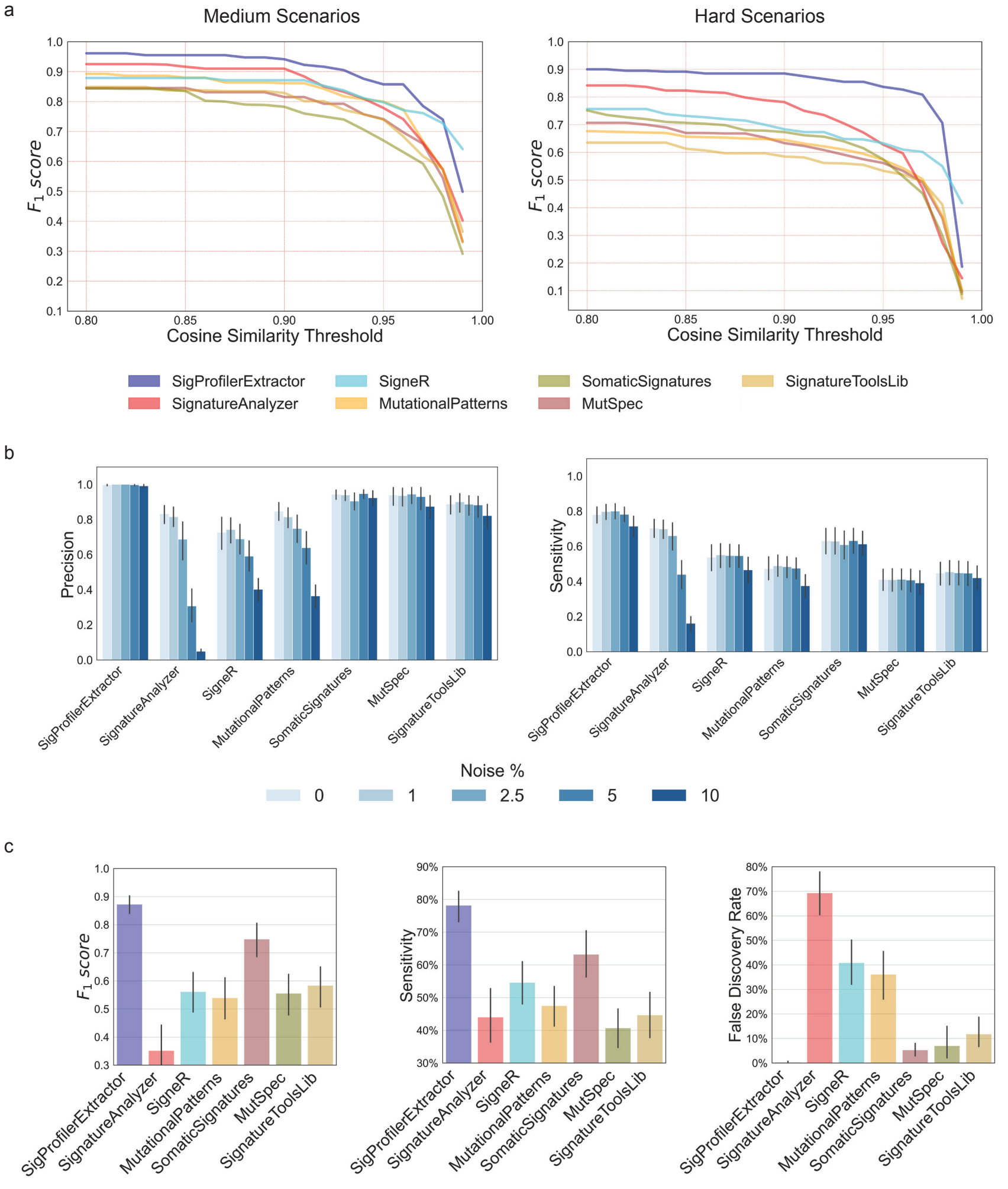
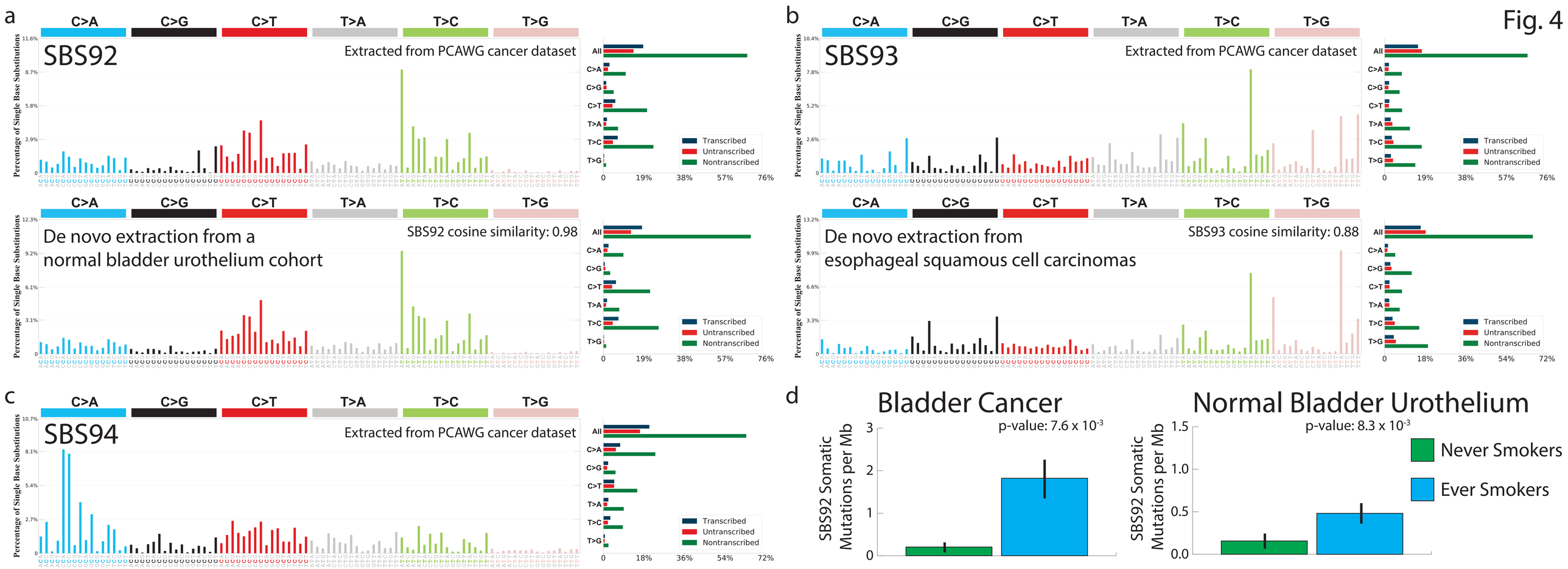

a
Medium Scenarios
Hard Scenarios
F_1 score
Cosine Similarity Threshold
SigProfilerExtractor
SignatureAnalyzer
SigneR
MutationalPatterns
SomaticSignatures
MutSpec
SignatureToolsLib
b
Precision
Sensitivity
Noise %
0
1
2.5
5
10
c
F_1 score
Sensitivity
False Discovery Rate

Fig. 4


a
SBS92
Extracted from PCAWG cancer dataset
De novo extraction from a normal bladder urothelium cohort
SBS92 cosine similarity: 0.98
b
SBS93
Extracted from PCAWG cancer dataset
De novo extraction from esophageal squamous cell carcinomas
SBS93 cosine similarity: 0.88
c
SBS94
Extracted from PCAWG cancer dataset
C>A
C>G
C>T
T>A
T>C
T>G
Percentage of Single Base Substitutions
All
Transcribed
Untranscribed
Nontranscribed
d
Bladder Cancer
p-value: 7.6 x 10-3
Normal Bladder Urothelium
p-value: 8.3 x 10-3
SBS92 Somatic Mutations per Mb
Never Smokers
Ever Smokers