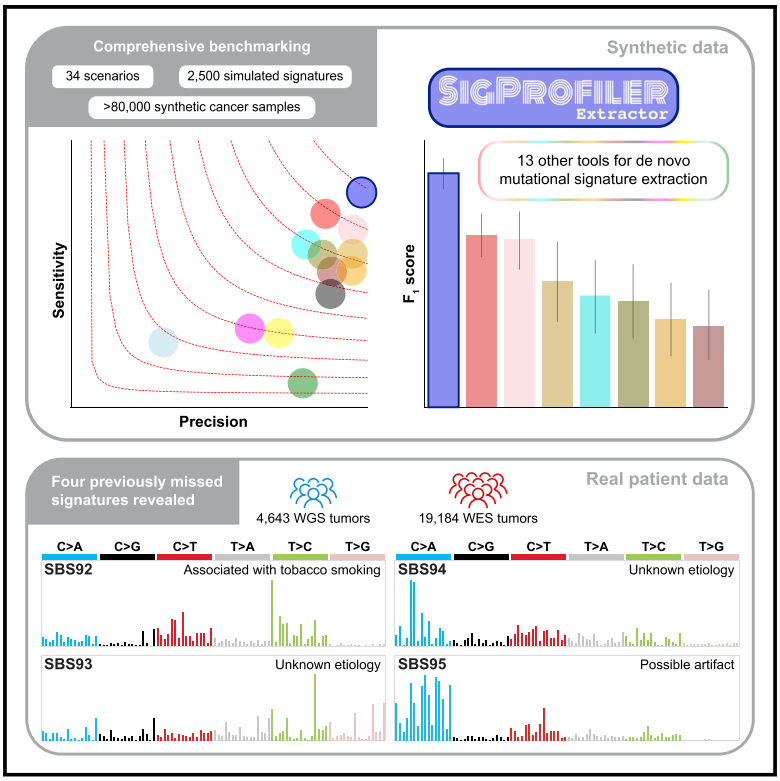
Cell Genomics

Technology

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

Graphical abstract



Highlights

- Most advanced bioinformatics tool for *de novo* extraction of mutational signatures
- Comprehensive benchmarking of 14 *de novo* extraction tools with and without noise
- Analysis of 23,827 sequenced cancers revealing four novel mutational signatures
- Novel signature attributed to direct tobacco smoking mutagenesis in bladder tissues

Authors

S.M. Ashiqul Islam, Marcos Díaz-Gay, Yang Wu, ..., Michael R. Stratton, Steven G. Rozen, Ludmil B. Alexandrov

Correspondence

l2alexandrov@health.ucsd.edu

In brief

Islam et al. present SigProfilerExtractor, a novel computational tool for *de novo* extraction of mutational signatures. Using more than 80,000 synthetic cancer samples, the authors demonstrate that SigProfilerExtractor outperforms 13 existing approaches. Applying SigProfilerExtractor to more than 23,000 sequenced human cancers reveals novel insights about the processes leading to somatic mutation accumulation and tumorigenesis, including a novel mutational signature related to tobacco smoking in bladder cancer.

Islam et al., 2022, Cell Genomics *2*, 100179
November 9, 2022
https://doi.org/10.1016/j.xgen.2022.100179

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

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

S.M. Ashiqul Islam,[1,2,3,17] Marcos Díaz-Gay,[1,2,3,17] Yang Wu,[4] Mark Barnes,[1,2,3] Raviteja Vangara,[1,2,3] Erik N. Bergstrom,[1,2,3] Yudou He,[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,[10,11,16] Christopher D. Steele,[12] Burçak Otlu,[1,2,3] Azhar Khandekar,[1,2,3] Ammal Abbasi,[1,2,3] Laura Humphreys,[6] Natalia Syulyukina,[2] Samuel W. Brady,[13] Boian S. Alexandrov,[14] Nischalan Pillay,[12,15] Jinghui Zhang,[13] David J. Adams,[6] Iñigo Martincorena,[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[1,2,3,18,*]

[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, Singapore 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, Cedex 08, 69372 Lyon, France
[8]Departments of Radiation Oncology and Cancer Genetics, City of Hope Comprehensive Cancer Center, Duarte, CA, 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]Research Department of Pathology, Cancer Institute, University College London, London WC1E 6BT, UK
[13]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[14]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[15]Department of Cellular and Molecular Pathology, Royal National Orthopaedic Hospital NHS Trust, Stanmore, Middlesex HA7 4LP, UK
[16]Department of Biology, University of Konstanz, Universitaetsstrasse 10, D-78464 Konstanz, Germany
[17]These authors contributed equally
[18]Lead contact
*Correspondence: l2alexandrov@health.ucsd.edu
https://doi.org/10.1016/j.xgen.2022.100179

SUMMARY

Mutational signature analysis is commonly performed in cancer genomic studies. Here, we present SigProfilerExtractor, an automated tool for *de novo* extraction of mutational signatures, and benchmark it against another 13 bioinformatics tools by using 34 scenarios encompassing 2,500 simulated signatures found in 60,000 synthetic genomes and 20,000 synthetic exomes. For simulations with 5% noise, reflecting high-quality datasets, SigProfilerExtractor outperforms other approaches by elucidating between 20% and 50% more true-positive signatures while yielding 5-fold less false-positive signatures. Applying SigProfilerExtractor to 4,643 whole-genome- and 19,184 whole-exome-sequenced cancers reveals four novel signatures. Two of the signatures are confirmed in independent cohorts, and one of these signatures is associated with tobacco smoking. In summary, this report provides a reference tool for analysis of mutational signatures, a comprehensive benchmarking of bioinformatics tools for extracting signatures, and several novel mutational signatures, including one putatively attributed to direct tobacco smoking mutagenesis in bladder tissues.

INTRODUCTION

The somatic mutations found in a cancer genome are the cumulative result of all endogenous and exogenous mutational processes that have been operative through the lineage of a cancer cell.[1] By examining the types of mutations in *TP53* across cancers, early studies demonstrated that specific environmental carcinogens exhibit characteristic patterns of somatic mutations.[2] The explosion of next-generation sequencing data from cancer genomes[3] and the development of novel computational approaches[4] have allowed separating the signatures of individual mutagenic processes operative in cancer. Large-scale analyses of cancer genomes have revealed more than 100 distinct signatures, with some attributed to exposures to environmental carcinogens, failure of DNA-repair pathways, infidelity/deficiency of replicating polymerases, iatrogenic events, and others.[5–12] Moreover, mutational signatures have been utilized for both cancer prevention and cancer treatment.[13,14]

De novo extraction of mutational signatures[4] is an unsupervised machine-learning approach where a matrix, *M*, which

corresponds to the somatic mutations in a set of cancer samples under a mutational classification,[15] 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 cancer sample. Algorithmically, *de novo* extraction of mutational signatures has relied on nonnegative matrix factorization (NMF)[16] or on approaches mathematically analogous to NMF.[17–19] 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 biological interpretation of the identified nonnegative factors.[16]

Since we introduced the mathematical concept of mutational signatures,[4] multiple computational frameworks were developed for *de novo* extraction of mutational signatures (Table 1).[12,20,22,24,25,27,28,31,32,34–36,38,40] Notably, the majority of existing tools (1) predominately support the simplest mutational classification, viz., SBS-96, which encompasses single base substitutions with their immediate 5′ and 3′ sequence context;[15] (2) lack automatic selection for the number of signatures; (3) do not identify a robust solution; (4) require pre-selection of a large number of hyperparameters; and (5) do not decompose *de novo* signatures to the set of more than 100 reference signatures available at the Catalog of Somatic Mutations in Cancer (COSMIC) database.[12,42] Importantly, there has been no extensive benchmark of the existing tools for *de novo* extraction leading to uncertainty regarding their performance.

To address these limitations, here we present SigProfiler Extractor—a reference tool for *de novo* extraction of mutational signatures. SigProfilerExtractor allows analysis of all types of mutational classifications, performs automatic selection of the number of signatures, yields robust solutions, requires only minimum setup, and decomposes *de novo* extracted signatures to known COSMIC signatures. A comprehensive benchmark including 3,608 unique matrix decompositions with SigProfilerExtractor and 13 other tools across a total of 34 distinct scenarios reveals that SigProfilerExtractor is robust to noise and that it outperforms all other computational tools for *de novo* extraction of mutational signatures (Tables S1, S2, S3, S4, and S5). Applying SigProfilerExtractor to the recently published set of 2,778 whole-genome-sequenced (WGS) cancers from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project[43] and an additional curated collection of 1,865 WGS and 19,184 whole-exome-sequenced (WES) cancers (Table S8) elucidates four novel mutational signatures. Two of the signatures are confirmed in independent cohorts, and a putative etiology of tobacco-associated mutagenesis is attributed to one of these signatures (SBS92).

RESULTS

Overview of SigProfilerExtractor

SigProfilerExtractor is implemented as a Python package, with an R wrapper, allowing users to run it in both Python and R environments (STAR Methods). By default, the tool requires only a single parameter—the input dataset containing the mutational catalogs of interest. SigProfilerExtractor supports most used formats outputted by variant-calling algorithms, which are internally converted[15] into a matrix, *M*. By default, the tool decomposes the matrix *M* searching for an optimal solution for the number of operative signatures, *k*, between 1 and 25 mutational signatures (Figure 1A). For each decomposition, SigProfilerExtractor performs 100 independent factorizations and, for each repetition, the matrix *M* is first Poisson resampled and normalized and, subsequently, factorized with the multiplicative update NMF algorithm[16] by minimizing an objective function based on the Kullback-Leibler divergence measure[44] (Figure 1B). Custom partition clustering, which utilizes the Hungarian algorithm[45] for comparing different repetitions, is applied to the 100 factorizations to identify stable solutions.[46] Specifically, the centroids of stable clusters are selected as optimal solutions, thus making these solutions resistant to fluctuations in the input data and the lack of uniqueness of NMF.[47] Lastly, when applicable, the optimal set of *de novo* signatures are matched to the set of reference COSMIC signatures (Figure 1C), with any *de novo* signature reported as novel when it cannot be decomposed by a combination of known COSMIC signatures.

Framework for benchmarking tools for *de novo* extraction

To benchmark tools for *de novo* extraction of mutational signatures, more than 60,000 unique synthetic cancer genomes and 20,000 cancer exomes were generated with known ground-truth mutational signatures (STAR Methods). These synthetic data included 32 noiseless scenarios and two scenarios with different levels of noise. Each scenario contained between 3 and 39 known signatures operative in 200 to 2,700 simulated cancer genomes (Tables S1, S2, S3, S4, and S5). Some scenarios were generated up to 20 times to account for variability in the simulations. While most noiseless scenarios (20/32) were based on SBS-96 mutational classification, we also generated 12 scenarios using extended classifications, i.e., matrices with more than 96 mutational channels (Table S2). To avoid bias in evaluating each tool's performance, three sets of SBS-96 signatures were used in generating the synthetic data: (1) COSMICv3 reference signatures,[12] (2) signatures previously extracted by SignatureAnalyzer (SA),[12] and (3) randomly generated signatures. Most of the noiseless scenarios were designed to mimic the activities of mutational signatures in specific cancer types, with four scenarios emulating a single cancer type, 16 scenarios a combination of two cancer types, and two scenarios mimicking the analysis of a pan-cancer dataset. In addition, randomly generated signatures displaying different distributions and exposures were used in 10 noiseless scenarios and in the noise scenarios, which were generated up to 20 times. Some of the scenarios included combinations of signatures that represent a challenge for *de novo* extraction, including mutational signatures with overlapping profiles in specific contexts or exhibiting flat featureless profiles. For presentation simplicity, scenarios were labeled based on their complexity as easy, medium, or hard. Easy scenarios were

Table 1. Overview of bioinformatics tools for *de novo* extraction of mutational signatures

Tool name	Input	Platform	Factorization method	Factorization engine	GPU	Manual selection	Automatic selection	Automatic algorithm	Mutational catalog support	Plotting support	COSMIC comparison
EMu[20]	matrix	C++	EM	original implementation[20]	no	yes	yes[a]	BIC[21]	SBS-96	no	no
Maftools[22]	matrix, MAF	R-Bioconductor	NMF	NMF R package[23]	no	yes	no	–	SBS-96	SBS-96	1 to 1
Mutational Patterns[24]	matrix, VCF	R-Bioconductor	NMF	NMF R package[23]	no	yes	no	–	SBS-96, SBS-192	SBS-96, SBS-192	1 to 1
MutSignatures[25]	matrix, VCF, MAF	R	NMF	Brunet et al.[26]	no	no	no	–	SBS-96	SBS-96	1 to 1
MutSpec[27]	matrix, VCF, custom	Galaxy, Perl, R	NMF	NMF R package[23]	no	yes	no	–	SBS-96, SBS-192	SBS-96, SBS-192	1 to 1
SigFit[28]	matrix	R	Bayesian inference	Stan R package[29]	no	yes	yes[a]	Elbow method[30]	SBS-96	SBS-96, SBS-192	1 to 1
SigMiner[31]	matrix, MAF	R	(automatic) Bayesian NMF, (manual) NMF	(automatic) Signature Analyzer implementation,[32] (manual) NMF R package[23]	no	yes[a]	yes	ARD[33]	SBS-96, DBS-78, ID-83	generic	1 to 1
Signature Analyzer[32,34]	matrix, MAF	R (CPU),[18] Python (GPU)[19]	Bayesian NMF	original implementation[32,34]	yes	no	yes	ARD[33]	SBS-96, DBS-78, ID-83	SBS-96, DBS-78, ID-83	1 to 1
Signature ToolsLib[35]	matrix, VCF, custom	R	NMF	NMF R package[23]	no	yes	no	–	SBS-96, DBS-78, ID-83, SV-32	SBS-96, SV-32, generic	1 to 2
SigneR[36]	matrix, VCF	R-Bioconductor, C++	Bayesian NMF	original implementation[36]	no	yes	yes[a]	BIC[21]	SBS-96	SBS-96	no
SigProfiler Extractor	matrix, VCF, MAF, custom	Python, R wrapper	NMF	(current work) original implementation	yes	yes	yes[a]	NMFk[37]	SBS-96, DBS-78, ID-83, CN-48, others,[15] any	SBS-96, DBS-78, ID-83, CN-48, SV-32, others,[15] generic	1 to many

(*Continued on next page*)

Table 1. ***Continued***

Tool name	Input	Platform	Factorization method	Factorization engine	GPU	Manual selection	Automatic selection	Automatic algorithm	Mutational catalog support	Plotting support	COSMIC comparison
SigProfiler_PCAWG[12]	matrix, VCF, MAF, custom	Python, MATLAB	NMF	Brunet et al.[26]	no	yes	no	–	SBS-96, DBS-78, ID-83, others,[15] any	SBS-96, DBS-78, ID-83	no
Somatic Signatures[38]	matrix, VCF	R-Bioconductor	NMF, PCA	NMF R package[23] pcaMethods R package[39]	no	yes	no	–	SBS-96	SBS-96	no
Tensor Signatures[40]	VCF	Python	NTF	TensorFlow[41]	yes	yes	yes[a]	BIC[21]	tensor	SBS-96 with strand bias	no

Tools are ordered alphabetically. 1 to 1 refers to one *de novo* signature being matched with exactly one COSMIC signature; 1 to 2 refers to one *de novo* signature being matched with a combination of up to two COSMIC signatures; 1 to many refers to one *de novo* signature being matched with a combination of one or more COSMIC signatures. MAF, mutation annotation format; VCF, variant call format; EM, expectation maximization algorithm; NMF, nonnegative matrix factorization; PCA, principal component analysis; NTF, nonnegative tensor factorization; ARD, automatic relevance determination; BIC, Bayesian information criterion; COSMIC, catalog of somatic mutations in cancer; SBS, single base substitutions; DBS, doublet base substitutions; ID, small insertions and deletions; CN, copy number; SV, structural variants.

[a]The default approach for selecting the total number of signatures when a tool supports both manual and automatic selection.

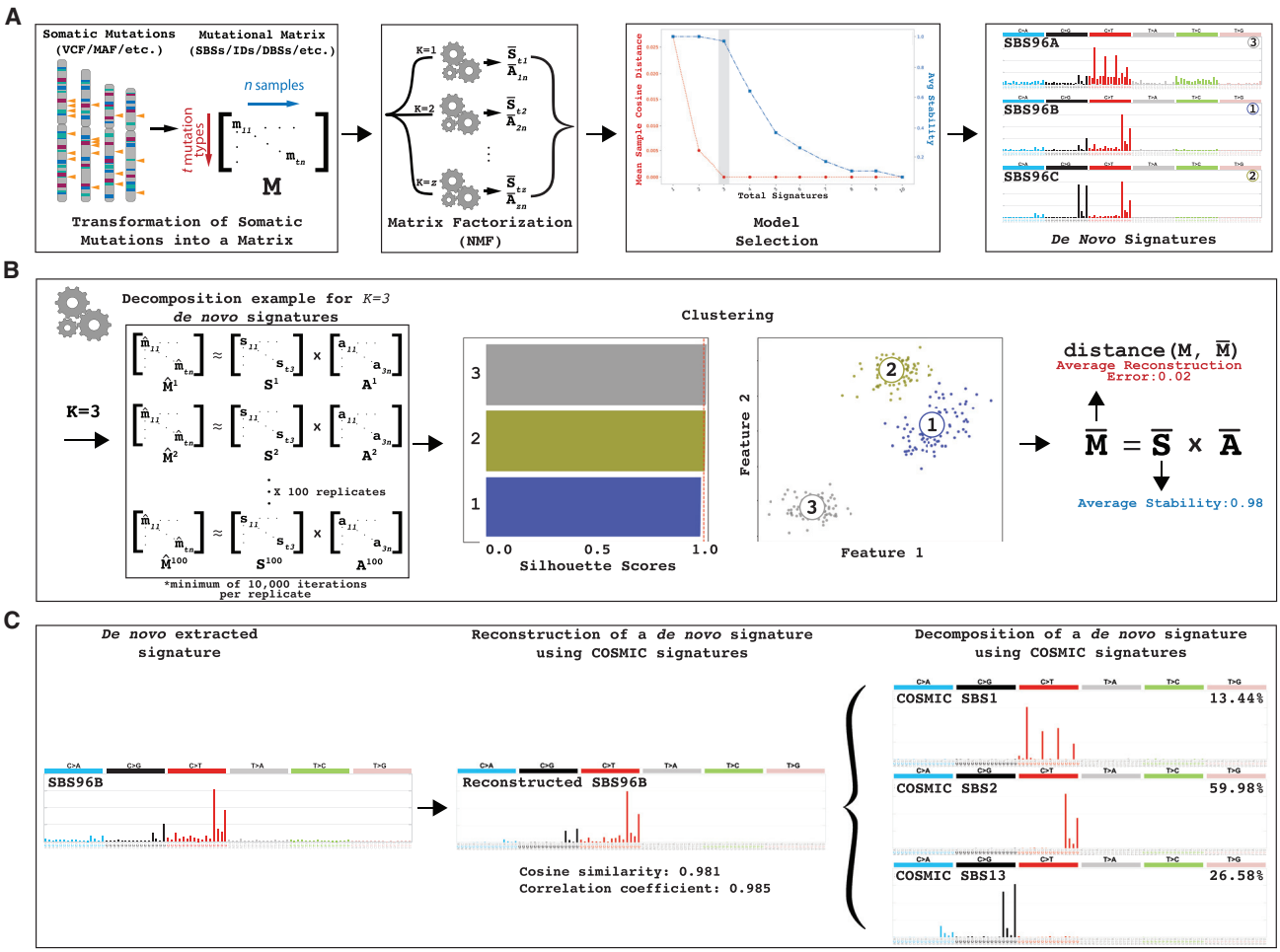


Figure 1. Overview of SigProfilerExtractor

(A) SigProfilerExtractor's general workflow is outlined starting from an input of somatic mutations and resulting in an output of *de novo* mutational signatures. An example is shown for a solution with three *de novo* signatures. Somatic mutations are first converted into a mutational matrix *M*. Subsequently, the matrix is factorized with different ranks using nonnegative matrix factorization. Model selection is applied to identify the optimal factorization rank based on each solution's stability and its reconstruction of the original data.

(B) Schematic representation for an example decomposition with a factorization rank of $k = 3$ reflecting three operative mutational signatures. By default, SigProfilerExtractor performs 100 independent nonnegative matrix factorizations with the matrix *M* being Poisson resampled and normalized (denoted by "^") prior to each factorization. Partition clustering of the 100 factorizations is used to evaluate the factorization stability rank, measured in silhouette values; clustering can also be presented as two-dimensional projections revealing more similar mutational signatures as shown for the three example signatures. The centroid of the clustered solutions (denoted by "–") is compared with the original matrix *M*.

(C) All identified *de novo* signatures are matched to a combination of known COSMIC mutational signatures. An example is given for *de novo* extracted signature SBS96B, which matches a combination of COSMIC signatures SBS1, SBS2, and SBS13.

generated using $\leq$5 signatures and provide a good indication of each tool's performance on approximately 7.4% of human cancer types (e.g., brain tumors). Medium scenarios contained 11 to 21 signatures and biologically reflect 15.9% of cancer types (e.g., cervical cancer). Hard scenarios have $\geq$25 signatures and reflect 59.5% of human cancer types (e.g., breast cancer) as well as pan-cancer datasets. In addition to the 32 noiseless scenarios, one whole-genome SBS-96 scenario with five different levels of noise, ranging between 0% and 10%, was included in the benchmark (STAR Methods). Further, an SBS-96-based whole-exome scenario with 5% noise was also included.

To compare the performance between different tools, we developed a standard set of evaluation metrics (Figure S1). 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 corresponds to a combined metric, defined as the harmonic mean of the precision and sensitivity: $2 * \frac{Precision*Sensitivity}{Precision+Sensitivity}$

Benchmarking using SBS-96 noiseless WGS data

SigProfilerExtractor and 13 other tools (Table 1) were first applied to all noiseless WGS scenarios based on the SBS-96 mutational classification. The 13 tools include SignatureAnalyzer (SA) and SigProfiler_PCAWG, a legacy MATLAB/Python version of SigProfilerExtractor, which were jointly used in the PCAWG analysis of mutational signatures and the derivation of the COSMICv3 set of reference signatures.[12] Except for MutSignatures, which can only decompose a matrix for a fixed number of signatures, all other tools were applied to each scenario by using their suggested methods for selecting the number of operative signatures. Apart from SA, which lacks this capability, all tools were also forced 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 can successfully extract mutational signatures from easy scenarios with the majority of F_1 scores >0.90 (Figure 2A). This is perhaps unsurprising, as many of these tools used synthetic data with $\leq$5 signatures to evaluate their performance.[20,22,24,27,28,31,32,34–36,38] In contrast, medium scenarios have proven to be a challenge for most tools with only SigProfilerExtractor, SigProfiler_PCAWG, and SA exhibiting F_1 scores >0.90. All tools had worst performance for the hard scenarios with F_1 scores below 0.80; only SigProfilerExtractor had an F_1 score of ~0.90 (Figure 2A).

To evaluate whether the type of ground-truth signatures affects the *de novo* extraction, we compared the ratio of F_1 scores (rF_1) from scenarios generated using COSMIC, SA, or random signatures (Figure 2B). Most tools had similar performance ($rF_1 \approx 1$) between COSMIC and random signatures and worst performance with SA signatures ($rF_1 < 1$). SomaticSignatures was an exception, as it performed well on random signatures but had similarly suboptimal performance on COSMIC and SA signatures. SigProfilerExtractor outperformed all other tools regardless of whether the synthetic data were generated using COSMIC, SA, or random signatures (Table S1).

To examine the performance of *de novo* extraction between the suggested and forced selection of the total number of signatures, we evaluated rF_1 across all medium and hard scenarios (Figure 2C). SigProfilerExtractor exhibited almost identical F_1 scores in suggested and forced selection, indicating a good performance of the automatic selection algorithm. Most other tools had similar F_1 scores between the suggested and forced selection, albeit with more variability across the different scenarios (Figure 2C). For example, MutSpec, one of the multiple tools based on NMF factorization, had $rF_1 \approx 1$ in both medium and hard scenarios, indicating that MutSpec is performing worse than SigProfilerExtractor (Figure 2A) not because of its algorithm for selecting the total number of signatures but likely due to its factorization approach. Other tools obtained lower F_1 scores for suggested solutions compared with forced solutions ($rF_1 < 1$), including SigneR and SigProfiler_PCAWG in the case of hard scenarios, SigMiner and Maftools for medium scenarios, and TensorSignatures and SigFit for both medium and hard scenarios. Lower F_1 scores for suggested solutions indicate that the different approaches used by these tools for selecting the number of signatures are not optimally performing (Figure 2C). Surprisingly, EMu, the only tool based on the expectation maximization algorithm,[20] had higher F_1 scores for automatic solutions in some hard scenarios. Considering the overall performance of EMu (Figure 2A), this outcome likely reflects the lack of convergence during the minimization of the EMu objective function for some hard scenarios.

Overall, across all suggested extractions from noiseless WGS hard scenarios reflecting ~60% of human cancer types, SigProfilerExtractor outperformed all other tools. SigProfilerExtractor was able to identify between 10% and 37% more TP signatures while yielding between 2.7- and 16-fold less FP signatures compared with the next seven best-performing tools (Figure 2D; Table S1).

Extended benchmarking of the top-performing tools

The reported comparisons for SBS-96 scenarios rely on a cosine similarity $\geq$0.90 for determining TP signatures and <0.90 for determining FP signatures. Note that a cosine similarity $\geq$0.90 is highly unlikely to happen purely by chance ($p = 5.90 \times 10^{-9}$), as two random nonnegative vectors are expected to have an average cosine similarity of 0.75 purely by chance.[48] Importantly, SigProfilerExtractor's performance does not depend on the specific value of the cosine similarity threshold (Figure 3A), as the tool consistently outperforms other approaches for TP thresholds above 0.80 ($p = 0.057$). Cosine similarity thresholds below 0.80 were not explored, as extracted signatures may be similar purely by chance.

Additional benchmarking was performed by generating 12 scenarios simulated using between 3 and 30 signatures with an extended number of mutational channels (STAR Methods). SigProfilerExtractor and SA are the only two tools that support analysis of custom-size matrices and provide GPU support (Table 1), thus allowing analysis of data with an extended number of mutational channels within a reasonable time frame. In contrast, all other matrix factorization tools rely solely on CPU implementations, with full runs expected to take many months for each tool applied to these scenarios. Overall, SigProfilerExtractor outperformed SA with average F_1 scores of 0.92 and 0.85, respectively (Table S2).

To further compare SigProfilerExtractor with the other seven top-performing tools, we applied each tool to a dataset with 30 ground-truth SBS-96 signatures operative in 1,000 genomes and random noise between 0% and 10%. Analysis for each noise level was repeated 20 times to account for variability in the noise generation. SigProfilerExtractor, SomaticSignatures, MutSpec, and SignatureToolsLib were robust to noise, with mostly unaffected performance (Figure 3B; Table S3). In contrast, SigProfiler_PCAWG, SA, SigneR, and MutationalPatterns were susceptible to noise (Figure 3B). For example, 2.5% noise reduced SA's F_1 from 0.76 to 0.66, while 10% noise reduced its F_1 to 0.07. Similarly, 10% noise reduced the F_1 of SigProfiler_PCAWG from 0.71 to 0.58, the F_1 of SigneR from 0.61 to 0.43, and the F_1 of MutationalPatterns from 0.60 to

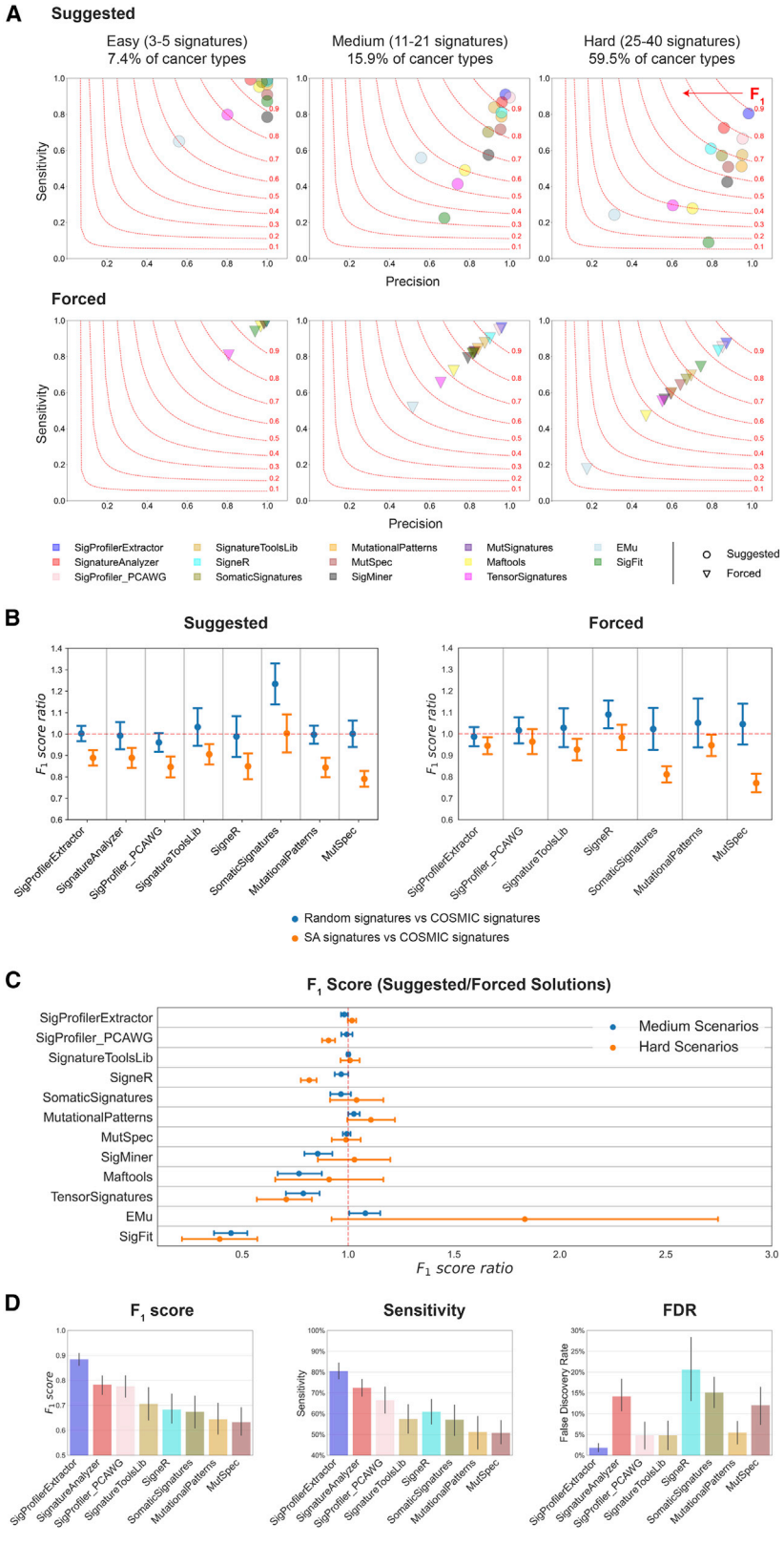


Figure 2. Benchmarking of bioinformatics tools for *de novo* extraction of mutational signatures using SBS-96 noiseless scenarios

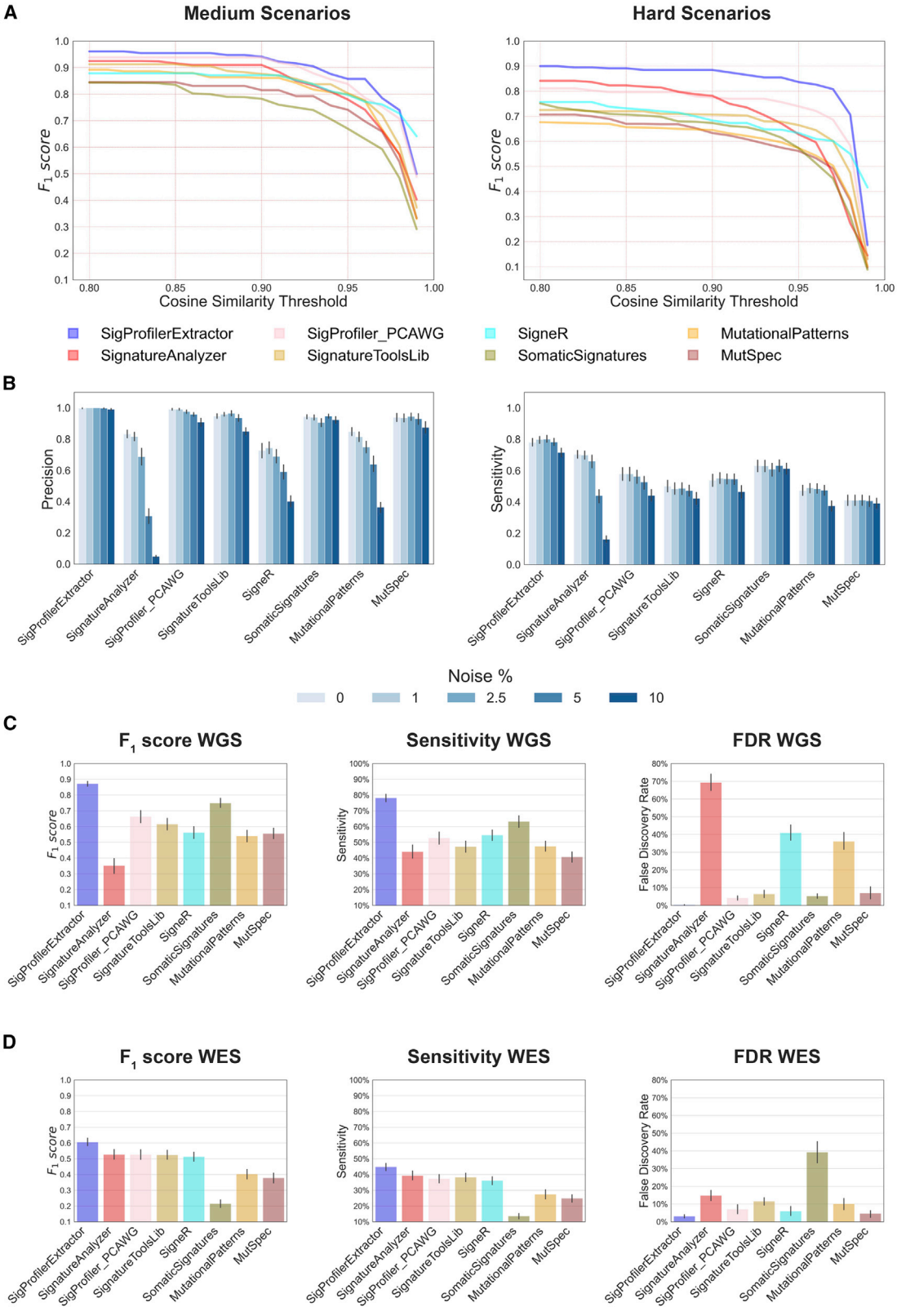
(A) Average precision (x axes), sensitivities (y axes), and F_1 scores (harmonic mean of precision and sensitivity; red curves) are shown across the three types of scenarios. Different tools are displayed using circles and triangles with different colors. Circles are used to display results for suggested model selection, which most closely matches analysis of a real dataset. Triangles are used to display results for forced model selection, where tools were required to extract the known total number of ground-truth mutational signatures. All triangles are located on the diagonal, as the forced model selection results in equal numbers of false-positive and false-negative signatures.

(B) Evaluating the effect of ground-truth signatures on the *de novo* extraction by different tools (x axes). Ratio of F_1 scores (y axes) with standard errors of the mean were calculated for medium complexity scenarios simulated using COSMIC, SA, or random signatures. Ratio of approximately 1.00 indicates a similar performance between different types of signatures.

(C) Evaluating the performance of *de novo* extraction between suggested and forced selection for different tools (x axes). Ratio of F_1 scores (y axes) with standard errors of the mean was calculated for all medium and hard scenarios. Ratio of approximately 1.00 indicates a similar performance between suggested and forced model selection.

(D) Summary of the performance for the top eight tools on hard SBS-96 noiseless scenarios with suggested model selection. Vertical axes reflect F_1 score (left plot), sensitivity (middle plot), and false discovery rate (right plot), respectively. Error bars correspond to standard errors of the mean.

Results from SignatureAnalyzer and MutSignatures are not displayed in (A)–(C) for forced and suggested model selections, respectively, as the tools do not support these types of analyses.



(legend on next page)

0.37. SA's reduced performance on data with noise is due to its automated approach for selecting total number of signatures. SA uses automatic relevance determination (ARD)[33] for selecting the number of signatures, with this number increasing from 26 (no noise; 30 ground-truth signatures) to 96 signatures (10% noise; Table S3). In contrast, SigProfiler_PCAWG, SigneR, and MutationalPatterns exhibit similar performance between forced and suggested solutions on data with noise (Table S3), indicating that their reduced performance is likely due to their factorization approaches.

SigProfilerExtractor outperformed all other tools regardless of noise levels. Simulations with 5% noise reflect genomics datasets with ~0.95 average sensitivity and precision of single base substitutions, similar to the recently published PCAWG cohort, which has 95% sensitivity (90% confidence interval, 88%–98%) and 95% precision (71%–99%).[43] For WGS simulations with 5% noise, SigProfilerExtractor was able to identify between 20% and 50% more TP signatures while yielding more than 5-fold less FP signatures compared with the next seven best-performing tools (Figure 3C; Table S3).

To assess the ability of the top-performing tools to extract *de novo* mutational signatures from exome sequencing data, a WES benchmarking dataset, encompassing 20,000 unique synthetic cancer exomes, was generated by downsampling the WGS noise scenario with 5% noise. Exome data were challenging for all the *de novo* mutational signature extraction tools, resulting in a significant decrease in performance (Figure 3D). The average F_1 score for all tools dropped from 0.61 for WGS simulations with 5% noise to 0.46 for WES simulations with 5% noise. Specifically, only SigProfilerExtractor showed an average F_1 score above 0.60, with no other tool showing an F_1 score above 0.53 (Figure 3D). SA was the only tool exhibiting an increased performance in WES compared with WGS in the 5% noise scenario, suggesting that the ARD approach was optimized for exome data (Table S5).

Lastly, simulations with 5% noise were additionally considered for benchmarking the different options provided by SigProfilerExtractor for performing *de novo* extraction. Specifically, we evaluated the effect of normalizing the input data (Gaussian mixture model [GMM], 100X, log2, and no normalization), the three different types of multiplicative updates for the NMF algorithm (Kullback-Leibler, Euclidean, or Itakura-Saito), and the two options for initializing the *S* and *A* matrices in the first step of the factorization: random initialization or nonnegative double singular vector decomposition (NNDSVD) initialization (STAR Methods). Overall, the objective function based on Kullback-Leibler updates outperformed the other two, independently of the normalization or initialization methods (Figure S2; Table S6). Regarding the four normalization methods, GMM, 100X, and log2 yielded comparable results, whereas running SigProfilerExtractor without previous transformation of the Poisson resampled matrix led to a significant drop in overall performance. The results obtained for the two different initialization methods, random and NNDSVD, differed depending on the other parameters. Nevertheless, they did not exhibit significant variations in the case of the top-performing NMF approach based on Kullback-Leibler updates and normalization using either GMM, 100X, or log2 transformation (Figure S2; Table S6).

Reanalysis of 4,643 WGS and 19,184 WES human cancers

To demonstrate its ability to yield novel biological results, SigProfilerExtractor was applied to the recently published set of 2,778 WGS cancers from the PCAWG project.[43] Additionally, we applied SigProfilerExtractor to an extended cohort of another 1,865 WGS and 19,184 WES cancers, encompassing data from The Cancer Genome Atlas (TCGA)[49] as well as 261 other published studies and 35 different ICGC projects (Table S8). As previously done in our original PCAWG analysis of mutational signatures,[12] extraction of mutational signatures was performed within each cancer type and across all samples (STAR Methods). In addition to all previously detected signatures,[12] our direct application of SigProfilerExtractor revealed three novel mutational signatures in the PCAWG dataset: SBS92, SBS93, and SBS94. Further, a novel signature was also identified exclusively in the extended cohort: SBS95 (Figure 4; Table S7).

Signature SBS92 was found predominately in PCAWG bladder cancers; the signature was characterized by T>C mutations with strong transcriptional strand asymmetry consistent with damage on purines for all types of substitutions (Figure 4A). Signature SBS92 was 9-fold elevated (Figure 4B; $p = 7.6 \times 10^{-3}$; Wilcoxon rank-sum test) in bladder cancers of ever smokers compared with never smokers. An almost identical signature was identified by reanalyzing a recently published cohort of 88 WGS microbiopsies of histologically normal urothelium,[50] with the similarity extending to both trinucleotide context and transcriptional strand asymmetry (Figure 4A; cosine similarity: 0.98; $p < 10^{-32}$). Consistently, SBS92 was found to be 3-fold elevated in the normal urothelium of tobacco ever smokers compared with never smokers (Figure 4B; $p = 8.3 \times 10^{-3}$; Wilcoxon rank-sum test).

Signature SBS93 was identified almost exclusively in WGS stomach cancers, both from PCAWG[12] and the extended cohort.[51] SBS93 was characterized by T>C and T>G mutations with a strand asymmetry consistent with damage on pyrimidines for TpTpA contexts (mutated base underlined; Figure 4C).

Figure 3. Additional evaluations of the top eight bioinformatics tools for *de novo* extraction of mutational signatures

(A) Average F_1 scores for the top eight tools based on different thresholds for cosine similarity in suggested medium and hard scenarios; thresholds for cosine similarity are used for determining true-positive signatures (Figure S1). Horizontal axes reflect the cosine similarity thresholds, while vertical axes correspond to the average F_1 scores corresponding to cosine similarity thresholds.

(B) Precision and sensitivity of the top eight tools for SBS-96 WGS scenarios with different levels of noise. Noise levels reflect the average number of somatic mutations in a cancer genome affected by additive white Gaussian noise; for example, 1% noise corresponds to approximately 1% of mutations in a sample being due to noise. Error bars correspond to standard errors of the mean.

(C and D) Summary of the performance of the top eight tools on SBS-96 (C) WGS and (D) WES scenarios with 5% noise. Vertical axes reflect F_1 score (left plot), sensitivity (middle plot), and false discovery rate (right plot), respectively. Error bars correspond to standard errors of the mean.

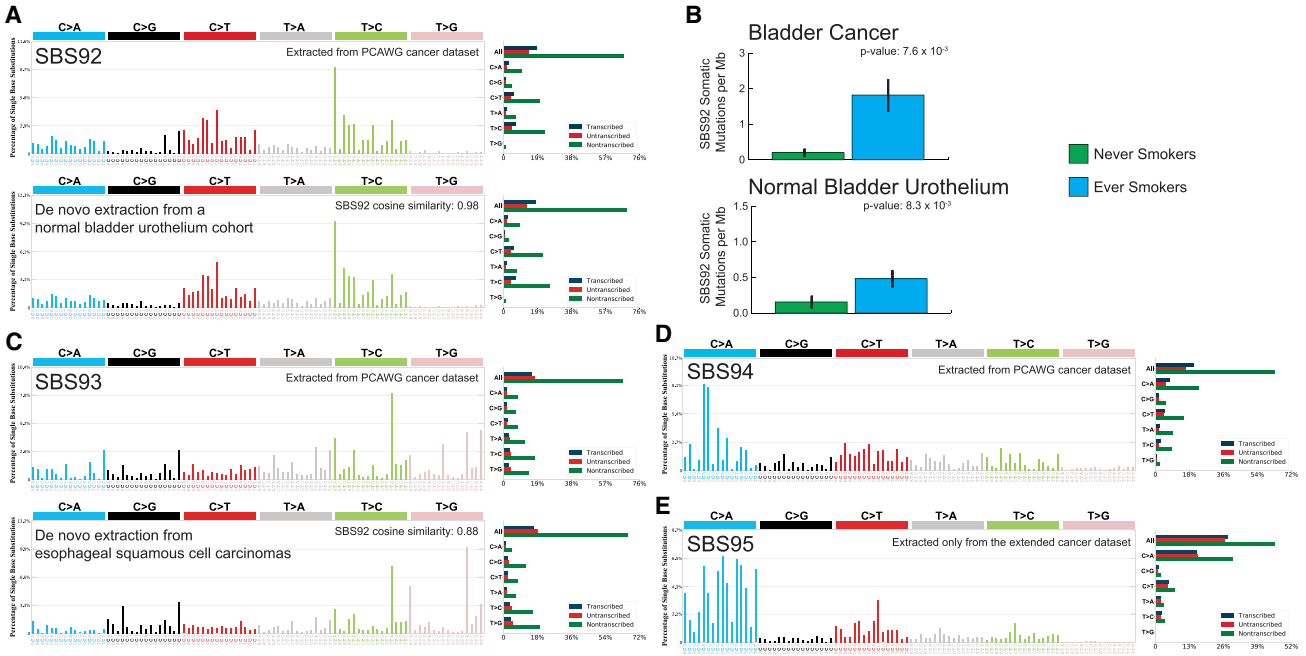


Figure 4. Novel signatures identified in a cohort of 4,643 WGS and 19,184 WES cancers

Mutational signatures are displayed using 96 plots. Single base substitutions are shown using the six subtypes of substitutions: C>A, C>G, C>T, T>A, T>C, and T>G. Underneath each subtype are 16 bars reflecting the sequence contexts determined by the four possible bases 5′ and 3′ to each mutated base. Additional information whether mutations from a signature are in nontranscribed/intergenic DNA, on the transcribed strand of a gene, or on the untranscribed strand of the gene is provided adjacent to the 96 plots.

(A) Mutational profile of signature SBS92 derived from the PCAWG cohort (top). Confirmation of the profile of signature SBS92 (bottom) by analysis of an independent WGS set of microbiopsies of histologically normal urothelium.[50]

(B) Bars are used to display average values for numbers of somatic substitutions per Mb attributed to signature SBS92 in bladder cancer and normal bladder urothelium. Green bars represent never smokers, whereas blue bars correspond to ever smokers. Error bars correspond to 95% confidence intervals. Each p value is based on a Wilcoxon rank-sum test.

(C) Mutational profile of signature SBS93 derived from the PCAWG cohort (top). Confirmation of the profile of signature SBS93 (bottom) by analysis of an independent WGS set of esophageal squamous cell carcinomas.[43]

(D) Mutational profile of signature SBS94 derived from the PCAWG cohort.

(E) Mutational profile of signature SBS95 derived only from liver hepatocellular carcinomas of the extended cohort. Signatures SBS94 and SBS95 were not identified in any additional independent cohort.

De novo extraction from the Mutographs cohort of 552 WGS esophageal squamous cell carcinomas,[52] a cancer type not included in the PCAWG dataset,[43] identified an analogous mutational signature, with the similarity extending to both trinucleotide context and transcriptional strand asymmetry (Figure 4C; cosine similarity: 0.88; $p = 1.1 \times 10^{-6}$). Signature SBS94 was found at high levels in a single colorectal PCAWG cancer, with smaller contributions to another eight colorectal cancers. The pattern of SBS94 was characterized by C>A mutations with a strand asymmetry indicative of damage on guanine (Figure 4D). Validation of somatic mutations by visual inspection confirmed that 98% of mutations contributed by SBS94 are likely real. Signatures SBS93 and SBS94 did not associate with any of the available PCAWG metadata,[43] and their etiologies remain unknown. Signature SBS95 was only identified in a set of 109 WGS liver hepatocellular carcinomas from the extended cohort, with a profile characterized by C>A mutations and a bias toward the genic regions in comparison to the intergenic (Figure 4E). SBS95 was found as the predominant signature in five samples from the ICGC LINC-JP project, with modest contributions to another 24 samples. The lack of any associations or validations in external cohorts does not allow us to independently confirm signature SBS95, and, following our standard protocol, we have classified SBS95 as a possible artifactual signature.

DISCUSSION

The performed large-scale benchmarking demonstrates that SigProfilerExtractor outperforms 13 other tools for *de novo* extraction of mutational signatures for noiseless datasets as well as for datasets containing different levels of random noise, including synthetic data emulating WGS and WES cancers. Importantly, SigProfilerExtractor generates almost no FP signatures while still identifying a higher number of TP signatures when compared with any of the other tools (Figures 2D, 3C, and 3D). *De novo* extraction 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 signatures,

reveals that SigProfilerExtractor's factorization performs better when compared with the factorizations of other tools (Figure 2B; Tables S1, S2, and S3). Similarly, benchmarking with suggested model selection, which most closely matches analysis of a real dataset, further demonstrates SigProfilerExtractor's ability to reveal novel biological results (Figure 2A; Tables S1, S2, S3, S4, and S5). Interestingly, SigProfilerExtractor outperforms other tools when extracting correlated mutational signatures[53] and signatures with overlapping profiles for specific contexts. In scenarios 5, 6, 9, and 10 (based on COSMIC signatures SBS2, SBS7a, and SBS7b, which share specific subtypes of C>T mutations), SigProfilerExtractor exhibited an average F_1 score of 0.96, while the next best tools had F_1 scores <0.90 (Table S1).

While our benchmarking evaluated 13 additional tools, 6 of the 13 tools internally rely on the same computational engine. Maftools, MutationalPatterns, MutSpec, SignatureToolsLib, SigMiner, and SomaticSignatures use the NMF R package[23] 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 (Tables S1, S2, S3, S4, and S5). SigProfiler_PCAWG and MutSignatures utilize similar implementations of NMF.[26] TensorSignatures makes use of the standard factorization algorithms included in TensorFlow.[41] SigFit uses a previously developed nonnegative factorization method.[29] In contrast, EMu, SA, SigneR, and SigProfilerExtractor provide original implementations of their factorization algorithms (Table 1). EMu was originally developed and tested on small datasets,[20] and its benchmarking performance is perhaps unsurprising considering the large number of synthetic samples used in all scenarios. Surprisingly, the original implementations of SA and SigneR were susceptible to noise, yielding high numbers of FP signatures (Figure 3B).

While SigProfilerExtractor and SigProfiler_PCAWG, the latter used in the PCAWG analysis,[12] share names, their computational engines are completely different. SigProfilerExtractor provides a fast-converging custom implementation of the multiplicative update algorithm,[16] supporting three different objective functions and a GPU-based factorization implemented using PyTorch.[54] In contrast, SigProfiler_PCAWG relies on a previously developed method by Brunet et al.[26] for analysis of gene-expression data. SigProfilerExtractor supports automate noise-resistant selection of the matrix decomposition rank based on the Hungarian algorithm[45] and the NMFk model selection approach,[37] while SigProfiler_PCAWG does not provide an automate selection (Table 1). Importantly, SigProfilerExtractor also implements different normalization options preventing hypermutated tumors from skewing the factorization.

Seven of the tools did not provide an automatic approach for selecting the total number of signatures (Table 1). Instead, most of these tools offered methodologies for manual selection, thus, bringing user dependence and arbitrariness in selecting solutions. EMu, TensorSignatures, and SigneR automatically select the total number of signatures using Bayesian information criterion (BIC),[21] while SA and SigMiner utilize ARD.[33] SigFit's selection approach is based on the Elbow method.[30] SigProfilerExtractor leverages a modified version of the NMFk selection approach.[37] Importantly, our simulations demonstrate that SigProfilerExtractor's model selection is robust to noise, while the implemented BIC and ARD approaches are affected even by low levels of noise (Figure 3B).

High noise levels had limited effect on SigProfilerExtractor, causing the tool to miss some of the ground-truth signatures used to generate the synthetic datasets. Indeed, the average number of detected signatures dropped from 23.45 in the replicates without noise to 21.60 in those with 10% noise while maintaining a similar high precision (0.998 and 0.992, respectively; Figure 3B; Table S3). However, for other tools, the number of signatures either rose significantly with noise, leading to a notable increase in the FP signatures identified (MutationalPatterns, SA, and SigneR), or were kept stable but with a decrease in precision (MutSpec, SomaticSignatures, and SignatureToolsLib; Table S3). To deeply characterize the shape of the FP signatures identified by the different tools, we applied the Shannon equitability index to the results of the noise benchmark with suggested model selection (STAR Methods). Interestingly, the three tools showing a significant increase of FP signatures with noise (MutationalPatterns, SA, and SigneR) also showed a decrease in the Shannon equitability index (Table S4). In the case of SA, 5% noise reduced the average Shannon equitability for the FP signatures from 0.826 to 0.572, while 10% noise reduced it to 0.337 (in the range of the sparsest COSMIC signatures). This behavior was also found at lower levels for MutationalPatterns and SigneR. A similar trend was found for the average number of FP signatures detected, increasing with 10% noise from 4.40 to 90.95 for SA and from 2.50 to 19.70 and 6.15 to 20.85 for MutationalPatterns and SigneR, respectively (Table S3). These findings indicate that the higher the number of signatures detected in these tools, the higher the possibility to obtain more sparse FP signatures. On the other hand, the tools that maintain a similar number of detected signatures independently of the noise level (MutSpec, SomaticSignatures, and SignatureToolsLib) showed similar values for the Shannon equitability of their FP calls. In all cases, average values exceeded 0.88, indicating that mostly flat signatures are erroneously called by these tools (Table S4). In the case of SigProfilerExtractor, the average Shannon equitability of FP without noise and for all noise levels was, in all cases, over 0.92, following a similar trend as the previously mentioned tools. However, it is worth noting that only three FP signatures were detected by SigProfilerExtractor in all 20 replicates with 10% noise (600 total ground-truth signatures), whereas, for example, 417 and 1,819 were found for SigneR and SA, respectively (Table S3).

In addition to outperforming 13 other tools on simulated datasets, SigProfilerExtractor can reveal additional biological results, as demonstrated by identifying four novel signatures from the reanalysis of 23,827 sequenced cancers from the PCAWG and the extended datasets. Importantly, SigProfilerExtractor identified signature SBS92 (Figure 4), which is associated with tobacco smoking in WGS bladder cancers and in WGS microbiopsies from normal bladder urothelium. The strong transcriptional strand bias observed in SBS92 is indicative of an environmental mutagen exposure that damages purines. Tobacco smoke is a complex mixture of at least 60 chemicals,[11] many capable of causing damage on purines. Interestingly, our and other prior

analyses of exome-sequenced bladder cancers from TCGA[11,55] did not reveal SBS92. Reanalysis of the set of TCGA bladder WES cancers[56] with SigProfilerExtractor was also unable to detect SBS92 (STAR Methods). We suspect that the lack of SBS92 in the TCGA bladder cancers was due to the use of exome sequencing; note that SBS92 is predominately found in intergenic regions (Figure 4A), with most samples expected to have less than 15 mutations from SBS92 in their exomes. To confirm this hypothesis, we downsampled the WGS bladder cancers and the WGS microbiopsies from normal bladder urothelium to exomes. SigProfilerExtractor's analysis of these downsampled genomes was unable to detect SBS92, confirming that exome sequencing is insufficient to identify signature SBS92 (STAR Methods).

In summary, here we report SigProfilerExtractor—a computational tool for *de novo* extraction of mutational signatures. We demonstrate that SigProfilerExtractor outperforms 13 other tools by conducting the largest benchmarking of bioinformatics approaches for extracting mutational signatures. Further, we apply SigProfilerExtractor to 4,643 WGS and 19,184 WES cancers and reveal four novel mutational signatures, including a signature putatively attributed to tobacco smoking mutagenesis in bladder cancer and in normal bladder epithelium.

Limitations of the study

In this study, we assumed that mutational signatures are linearly and independently accumulating across the genomic landscape. While this assumption is likely correct for most signatures of small mutational events,[4] such as substitutions and small insertions and deletions, it will be likely violated for signatures of larger mutational events including most copy-number signatures.[57] In addition, a prior study has shown that the pattern of at least one substitution signature is not a superposition of individual alterations.[58] Our current benchmarking ignores such scenarios, as they tend to be found in a small number of cancers with concurrent loss of both polymerase proofreading and mismatch repair.[58] Lastly, this study focused on benchmarking the *de novo* extraction of mutational signatures from large sets of tumor samples, and it did not consider the assignment of signatures to a single cancer genome. Future benchmarking efforts will be required to evaluate the ability of different tools to accurately assign known mutational signatures to individual cancers.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100179.

ACKNOWLEDGMENTS

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 (L.B.A., P.B., and M.R.S.) and Wellcome grant reference 206194 (M.R.S.), as well as US National Institutes of Health grants R01MH116281-01A1 (B.S.A.), R01ES030993-01A1 (L.B.A.), R01ES032547 (L.B.A.), and R01CA269919 (L.B.A.). 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. L.B.A. 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 L.B.A. A.J.G. was funded by a postdoctoral fellowship (grant no. P2BSP3_178591). I.M. is funded by Cancer Research UK (C57387/A21777) and the Wellcome Trust. Y.R.L. was supported by the NCI F32 training grant and NCI K12 career development award. N.P. 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 US Department of Energy's National Nuclear Security Administration and was supported by Laboratory Directed Research and Development (LDRD) grant 20190020DR (B.S.A.). C.D.S. 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

L.B.A., S.M.A.I., and M.D.-G. designed both SigProfilerExtractor's methodology and performed analyses with help from N.P., J.Z., D.J.A., I.M., B.S.A., L.H., D.C.W., M.T.L., P.B., M.R.S., and S.G.R. S.M.A.I. and M.D.-G. developed SigProfilerExtractor with help from M.B., R.V., M.V., E.N.B., Y.H., C.D.S., and J.W. All synthetic benchmarking datasets were generated by Y.W. and S.G.R. S.M.A.I. and M.D.-G. documented SigProfilerExtractor and performed the benchmarking of all tools on synthetic data with help from M.B., R.V., B.O., A.K., and A.A. Additional validations, confirmations, and applications of SigProfilerExtractor to real and synthetic datasets were performed by M.D.-G., S.M.A.I., M.B., R.V., S.M., S.S., Y.R.L., N.S., L.R., T.Z., A.J.G., Y.H., C.D.S., and S.W.B. L.B.A. directed the overall research and wrote the manuscript with help from S.M.A.I. and M.D.-G. All authors read, provided input, and approved the final manuscript.

DECLARATION OF INTERESTS

M.V. is an employee of NVIDIA corporation. L.B.A. is a compensated consultant and has equity interest in io9, LLC. His spouse is an employee of

Biotheranostics, Inc. L.B.A. and B.S.A. are inventors of a US patent 10,776,718. E.N.B. and L.B.A. declare US provisional applications with serial numbers 63/289,601 and 63/269,033. L.B.A. and A.A. declare US provisional patent applications with serial numbers 63/366,392 and 63/367,846. All other authors declare no competing interests.

Received: June 6, 2021
Revised: April 10, 2022
Accepted: August 31, 2022
Published: September 23, 2022

REFERENCES

1. Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724. https://doi.org/10.1038/nature07943.
2. Hollstein, M., Hergenhahn, M., Yang, Q., Bartsch, H., Wang, Z.-Q., and Hainaut, P. (1999). New approaches to understanding p53 gene tumor mutation spectra. Mutat. Res. *431*, 199–209. https://doi.org/10.1016/s0027-5107(99)00162-1.
3. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558. https://doi.org/10.1126/science.1235122.
4. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013). Deciphering signatures of mutational processes operative in human cancer. Cell Rep. *3*, 246–259. https://doi.org/10.1016/j.celrep.2012.12.008.
5. Alexandrov, L.B. (2015). Understanding the origins of human cancer. Science *350*, 1175. https://doi.org/10.1126/science.aad7363.
6. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421. https://doi.org/10.1038/nature12477.
7. Petljak, M., and Alexandrov, L.B. (2016). Understanding mutagenesis through delineation of mutational signatures in human cancer. Carcinogenesis *37*, 531–540. https://doi.org/10.1093/carcin/bgw055.
8. Pich, O., Muiños, F., Lolkema, M.P., Steeghs, N., Gonzalez-Perez, A., and Lopez-Bigas, N. (2019). The mutational footprints of cancer therapies. Nat. Genet. *51*, 1732–1740. https://doi.org/10.1038/s41588-019-0525-5.
9. Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407. https://doi.org/10.1038/ng.3441.
10. Alexandrov, L.B., Nik-Zainal, S., Siu, H.C., Leung, S.Y., and Stratton, M.R. (2015). A mutational signature in gastric cancer suggests therapeutic strategies. Nat. Commun. *6*, 8683. https://doi.org/10.1038/ncomms9683.
11. Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science *354*, 618–622. https://doi.org/10.1126/science.aag0299.
12. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101. https://doi.org/10.1038/s41586-020-1943-3.
13. Brady, S.W., Gout, A.M., and Zhang, J. (2022). Therapeutic and prognostic insights from the analysis of cancer mutational signatures. Trends Genet. *38*, 194–208. https://doi.org/10.1016/j.tig.2021.08.007.
14. Georgeson, P., Pope, B.J., Rosty, C., Clendenning, M., Mahmood, K., Joo, J.E., Walker, R., Hutchinson, R.A., Preston, S., Como, J., et al. (2021). Evaluating the utility of tumour mutational signatures for identifying hereditary colorectal cancer and polyposis syndrome carriers. Gut *70*, 2138–2149. https://doi.org/10.1136/gutjnl-2019-320462.
15. Bergstrom, E.N., Huang, M.N., Mahto, U., Barnes, M., Stratton, M.R., Rozen, S.G., and Alexandrov, L.B. (2019). SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genom. *20*, 685. https://doi.org/10.1186/s12864-019-6041-2.
16. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791. https://doi.org/10.1038/44565.
17. Févotte, C., and Cemgil, A.T. (2009). Nonnegative matrix factorizations as probabilistic inference in composite models. IEEE *24-28*, 1913–1917.
18. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B *39*, 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
19. Suri, P., and Roy, N.R. (2017). Comparison between LDA & NMF for event-detection from large text stream data. IEEE *9-10*, 1–5.
20. Fischer, A., Illingworth, C.J.R., Campbell, P.J., and Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. *14*, R39. https://doi.org/10.1186/gb-2013-14-4-r39.
21. Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. *6*, 461–464. https://doi.org/10.1214/aos/1176344136.
22. Mayakonda, A., Lin, D.C., Assenov, Y., Plass, C., and Koeffler, H.P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. *28*, 1747–1756. https://doi.org/10.1101/gr.239244.118.
23. Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinf. *11*, 367. https://doi.org/10.1186/1471-2105-11-367.
24. Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. *10*, 33. https://doi.org/10.1186/s13073-018-0539-0.
25. Fantini, D., Vidimar, V., Yu, Y., Condello, S., and Meeks, J.J. (2020). MutSignatures: an R package for extraction and analysis of cancer mutational signatures. Sci. Rep. *10*, 18217.
26. Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA *101*, 4164–4169. https://doi.org/10.1073/pnas.0308531101.
27. Ardin, M., Cahais, V., Castells, X., Bouaoun, L., Byrnes, G., Herceg, Z., Zavadil, J., and Olivier, M. (2016). MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinf. *17*, 170. https://doi.org/10.1186/s12859-016-1011-z.
28. Gori, K., and Baez-Ortega, A. (2020). sigfit: flexible Bayesian inference of mutational signatures. Preprint at bioRxiv. https://doi.org/10.1101/372896.
29. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. J. Stat. Softw. *76*. https://doi.org/10.18637/jss.v076.i01.
30. Thorndike, R.L. (1953). Who belongs in the family? Psychometrika *18*, 267–276. https://doi.org/10.1007/bf02289263.
31. Wang, S., Li, H., Song, M., Tao, Z., Wu, T., He, Z., Zhao, X., Wu, K., and Liu, X.S. (2021). Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. PLoS Genet. *17*, e1009557. https://doi.org/10.1371/journal.pgen.1009557.
32. Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., Lawrence, M.S., Kiezun, A., Fernandes, S.M., Bahl, S., et al. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat. Commun. *6*, 8866. https://doi.org/10.1038/ncomms9866.

33. Tan, V.Y.F., and Févotte, C. (2013). Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. IEEE Trans. Pattern Anal. Mach. Intell. *35*, 1592–1605. https://doi.org/10.1109/TPAMI.2012.240.

34. Taylor-Weiner, A., Aguet, F., Haradhvala, N.J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E.M., and Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. Genome Biol. *20*, 228. https://doi.org/10.1186/s13059-019-1836-7.

35. Degasperi, A., Amarante, T.D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., Morganella, S., Nanda, A.S., Badja, C., Koh, G., et al. (2020). A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat. Cancer *1*, 249–263. https://doi.org/10.1038/s43018-020-0027-5.

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

37. Benjamin, N., Raviteja, V., Miguel, A.H.-H., Svetlana, K., and Boian, A. (2020). A neural network for determination of latent dimensionality in Nonnegative Matrix Factorization. Mach. Learn.: Sci. Technol. https://doi.org/10.48550/arXiv.2006.12402.

38. Gehring, J.S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics *31*, 3673–3675. https://doi.org/10.1093/bioinformatics/btv408.

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

40. Vöhringer, H., Hoeck, A.V., Cuppen, E., and Gerstung, M. (2021). Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. Nat. Commun. *12*, 3628. https://doi.org/10.1038/s41467-021-23551-9.

41. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2016). TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at arXiv. https://doi.org/10.48550/arXiv.1603.04467.

42. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.

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

44. Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. Ann. Math. Statist. *22*, 79–86. https://doi.org/10.1214/aoms/1177729694.

45. Kuhn, H.W. (1955). The Hungarian method for the assignment problem. Nav. Res. Logist. *2*, 83–97.

46. Huang, K., Sidiropoulos, N.D., and Swami, A. (2014). Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition. IEEE Trans. Signal Process. *62*, 211–224. https://doi.org/10.1109/TSP.2013.2285514.

47. Lin, C. (2007). On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Trans. Neural Netw. *18*, 1589–1596. https://doi.org/10.1109/TNN.2007.895831.

48. Bergstrom, E.N., Barnes, M., Martincorena, I., and Alexandrov, L.B. (2020). Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. BMC Bioinf. *21*, 438. https://doi.org/10.1186/s12859-020-03772-3.

49. Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. *6*, 271–281.e7. https://doi.org/10.1016/j.cels.2018.03.002.

50. Lawson, A.R.J., Abascal, F., Coorens, T.H.H., Hooks, Y., O'Neill, L., Latimer, C., Raine, K., Sanders, M.A., Warren, A.Y., Mahbubani, K.T.A., et al. (2020). Extensive heterogeneity in somatic mutation and selection in the human bladder. Science *370*, 75–82. https://doi.org/10.1126/science.aba8347.

51. Wang, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H.N., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat. Genet. *46*, 573–582. https://doi.org/10.1038/ng.2983.

52. Moody, S., Senkin, S., Islam, S.M.A., Wang, J., Nasrollahzadeh, D., Cortez Cardoso Penha, R., Fitzgerald, S., Bergstrom, E.N., Atkins, J., He, Y., et al. (2021). Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. Nat. Genet. *53*, 1553–1563. https://doi.org/10.1038/s41588-021-00928-6.

53. Wu, Y., Chua, E.H.Z., Ng, A.W.T., Boot, A., and Rozen, S.G. (2022). Accuracy of mutational signature software on correlated signatures. Sci. Rep. *12*, 390. https://doi.org/10.1038/s41598-021-04207-6.

54. Lew, J., Shah, D.A., Pati, S., Cattell, S., Zhang, M., Sandhupatla, A., Ng, C., Goli, N., Sinclair, M.D., Rogers, T.G., and Aamodt, T.M. (2019). Analyzing machine learning workloads using a detailed GPU simulator. IEEE *24-26*, 151–152.

55. Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D'Andrea, A., and Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. *48*, 600–606. https://doi.org/10.1038/ng.3557.

56. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315–322. https://doi.org/10.1038/nature12965.

57. Steele, C.D., Abbasi, A., Islam, S.M.A., Bowes, A.L., Khandekar, A., Haase, K., Hames-Fathi, S., Ajayi, D., Verfaillie, A., Dhami, P., et al. (2022). Signatures of copy number alterations in human cancer. Nature *606*, 984–991. https://doi.org/10.1038/s41586-022-04738-6.

58. Haradhvala, N.J., Kim, J., Maruvka, Y.E., Polak, P., Rosebrock, D., Livitz, D., Hess, J.M., Leshchiner, I., Kamburov, A., Mouw, K.W., et al. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat. Commun. *9*, 1746. https://doi.org/10.1038/s41467-018-04002-4.

59. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979–993. https://doi.org/10.1016/j.cell.2012.04.024.

60. Royer-Bertrand, B., Torsello, M., Rimoldi, D., El Zaoui, I., Cisarova, K., Pescini-Gobert, R., Raynaud, F., Zografos, L., Schalenbourg, A., Speiser, D., et al. (2016). Comprehensive genetic landscape of uveal melanoma by whole-genome sequencing. Am. J. Hum. Genet. *99*, 1190–1198. https://doi.org/10.1016/j.ajhg.2016.09.008.

61. Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278. https://doi.org/10.1016/j.cell.2012.06.023.

62. Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell *150*, 1107–1120. https://doi.org/10.1016/j.cell.2012.08.029.

63. Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214–220. https://doi.org/10.1038/nature09744.

64. Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510. https://doi.org/10.1038/nature10738.

65. Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature *488*, 43–48. https://doi.org/10.1038/nature11213.

66. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54. https://doi.org/10.1038/nature17676.

67. Love, C., Sun, Z., Jima, D., Li, G., Zhang, J., Miles, R., Richards, K.L., Dunphy, C.H., Choi, W.W.L., Srivastava, G., et al. (2012). The genetic landscape of mutations in Burkitt lymphoma. Nat. Genet. *44*, 1321–1325. https://doi.org/10.1038/ng.2468.

68. Zhang, J., Wu, G., Miller, C.P., Tatevossian, R.G., Dalton, J.D., Tang, B., Orisme, W., Punchihewa, C., Parker, M., Qaddoumi, I., et al. (2013). Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. Nat. Genet. *45*, 602–612. https://doi.org/10.1038/ng.2611.

69. Sato, Y., Yoshizato, T., Shiraishi, Y., Maekawa, S., Okuno, Y., Kamura, T., Shimamura, T., Sato-Otsubo, A., Nagae, G., Suzuki, H., et al. (2013). Integrated molecular analysis of clear-cell renal cell carcinoma. Nat. Genet. *45*, 860–867. https://doi.org/10.1038/ng.2699.

70. Behjati, S., Tarpey, P.S., Presneau, N., Scheipl, S., Pillay, N., Van Loo, P., Wedge, D.C., Cooke, S.L., Gundem, G., Davies, H., et al. (2013). Distinct H3F3A and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone. Nat. Genet. *45*, 1479–1482. https://doi.org/10.1038/ng.2814.

71. Behjati, S., Tarpey, P.S., Sheldon, H., Martincorena, I., Van Loo, P., Gundem, G., Wedge, D.C., Ramakrishna, M., Cooke, S.L., Pillay, N., et al. (2014). Recurrent PTPRB and PLCG1 mutations in angiosarcoma. Nat. Genet. *46*, 376–379. https://doi.org/10.1038/ng.2921.

72. Wu, G., Diaz, A.K., Paugh, B.S., Rankin, S.L., Ju, B., Li, Y., Zhu, X., Qu, C., Chen, X., Zhang, J., et al. (2014). The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. Nat. Genet. *46*, 444–450. https://doi.org/10.1038/ng.2938.

73. Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., et al. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat. Genet. *48*, 500–509. https://doi.org/10.1038/ng.3547.

74. Walter, M.J., Shen, D., Ding, L., Shao, J., Koboldt, D.C., Chen, K., Larson, D.E., McLellan, M.D., Dooling, D., Abbott, R., et al. (2012). Clonal architecture of secondary acute myeloid leukemia. N. Engl. J. Med. *366*, 1090–1098. https://doi.org/10.1056/NEJMoa1106968.

75. Roberts, K.G., Li, Y., Payne-Turner, D., Harvey, R.C., Yang, Y.L., Pei, D., McCastlain, K., Ding, L., Lu, C., Song, G., et al. (2014). Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. N. Engl. J. Med. *371*, 1005–1015. https://doi.org/10.1056/NEJMoa1403088.

76. Beà, S., Valdés-Mas, R., Navarro, A., Salaverria, I., Martín-Garcia, D., Jares, P., Giné, E., Pinyol, M., Royo, C., Nadeu, F., et al. (2013). Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. Proc. Natl. Acad. Sci. USA *110*, 18250–18255. https://doi.org/10.1073/pnas.1314608110.

77. Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T.D., Gong, Z., Gao, H., Hao, K., Willard, M.D., Xu, J., et al. (2013). Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res. *23*, 1422–1433. https://doi.org/10.1101/gr.154492.113.

78. Tirode, F., Surdez, D., Ma, X., Parker, M., Le Deley, M.C., Bahrami, A., Zhang, Z., Lapouble, E., Grossetête-Lalami, S., Rusch, M., et al. (2014). Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. Cancer Discov. *4*, 1342–1353. https://doi.org/10.1158/2159-8290.CD-14-0622.

79. Jusakul, A., Cutcutache, I., Yong, C.H., Lim, J.Q., Huang, M.N., Padmanabhan, N., Nellore, V., Kongpetch, S., Ng, A.W.T., Ng, L.M., et al. (2017). Whole-genome and epigenomic landscapes of etiologically distinct subtypes of cholangiocarcinoma. Cancer Discov. *7*, 1116–1135. https://doi.org/10.1158/2159-8290.CD-17-0368.

80. Cho, J., Bass, A.J., Lawrence, M.S., Cibulskis, K., Cho, A., Lee, S.N., Yamauchi, M., Wagle, N., Pochanard, P., Kim, N., et al. (2014). Colon cancer-derived oncogenic EGFR G724S mutant identified by whole genome sequence analysis is dependent on asymmetric dimerization and sensitive to cetuximab. Mol. Cancer *13*, 141. https://doi.org/10.1186/1476-4598-13-141.

81. Johansson, P., Aoude, L.G., Wadt, K., Glasson, W.J., Warrier, S.K., Hewitt, A.W., Kiilgaard, J.F., Heegaard, S., Isaacs, T., Franchina, M., et al. (2016). Deep sequencing of uveal melanoma identifies a recurrent mutation in PLCB4. Oncotarget *7*, 4624–4631. https://doi.org/10.18632/oncotarget.6614.

82. Ramsay, A.J., Martínez-Trillos, A., Jares, P., Rodríguez, D., Kwarciak, A., and Quesada, V. (2013). Next-generation sequencing reveals the secrets of the chronic lymphocytic leukemia genome. Clin. Transl. Oncol. *15*, 3–8. https://doi.org/10.1007/s12094-012-0922-z.

83. Richter, J., Schlesner, M., Hoffmann, S., Kreuz, M., Leich, E., Burkhardt, B., Rosolowski, M., Ammerpohl, O., Wagener, R., Bernhart, S.H., et al. (2012). Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat. Genet. *44*, 1316–1320. https://doi.org/10.1038/ng.2469.

84. Scarlett, C.J., Salisbury, E.L., Biankin, A.V., and Kench, J. (2011). Precursor lesions in pancreatic cancer: morphological and molecular pathology. Pathology *43*, 183–200. https://doi.org/10.1097/PAT.0b013e3283445e3a.

85. Rausch, T., Jones, D.T.W., Zapatka, M., Stütz, A.M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P.A., et al. (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell *148*, 59–71. https://doi.org/10.1016/j.cell.2011.12.013.

86. Zhao, Y., Yang, J., Chen, Z., Gao, Z., Zhou, F., Li, X., Li, J., Shi, S., Feng, X., Sun, N., et al. (2014). Identification of somatic alterations in stage I lung adenocarcinomas by next-generation sequencing. Genes Chromosomes Cancer *53*, 289–298. https://doi.org/10.1002/gcc.22138.

87. Duns, G., Hofstra, R.M.W., Sietzema, J.G., Hollema, H., van Duivenbode, I., Kuik, A., Giezen, C., Jan, O., Bergsma, J.J., Bijnen, H., et al. (2012). Targeted exome sequencing in clear cell renal cell carcinoma tumors suggests aberrant chromatin regulation as a crucial step in ccRCC development. Hum. Mutat. *33*, 1059–1062. https://doi.org/10.1002/humu.22090.

88. Arai, E., Sakamoto, H., Ichikawa, H., Totsuka, H., Chiku, S., Gotoh, M., Mori, T., Nakatani, T., Ohnami, S., Nakagawa, T., et al. (2014). Multilayer-omics analysis of renal cell carcinoma, including the whole exome, methylome and transcriptome. Int. J. Cancer *135*, 1330–1342. https://doi.org/10.1002/ijc.28768.

89. Samman, M., Wood, H.M., Conway, C., Stead, L., Daly, C., Chalkley, R., Berri, S., Senguven, B., Ross, L., Egan, P., et al. (2015). A novel genomic signature reclassifies an oral cancer subtype. Int. J. Cancer *137*, 2364–2373. https://doi.org/10.1002/ijc.29615.

90. Al-Hebshi, N.N., Li, S., Nasher, A.T., El-Setouhy, M., Alsanosi, R., Blancato, J., and Loffredo, C. (2016). Exome sequencing of oral squamous cell carcinoma in users of Arabian snuff reveals novel candidates for driver genes. Int. J. Cancer *139*, 363–372. https://doi.org/10.1002/ijc.30068.

91. Yip, S., Butterfield, Y.S., Morozova, O., Chittaranjan, S., Blough, M.D., An, J., Birol, I., Chesnelong, C., Chiu, R., Chuah, E., et al. (2012). Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish

oligodendrogliomas from other cancers. J. Pathol. *226*, 7–16. https://doi.org/10.1002/path.2995.

92. Castellarin, M., Milne, K., Zeng, T., Tse, K., Mayo, M., Zhao, Y., Webb, J.R., Watson, P.H., Nelson, B.H., and Holt, R.A. (2013). Clonal evolution of high-grade serous ovarian carcinoma from primary to recurrent disease. J. Pathol. *229*, 515–524. https://doi.org/10.1002/path.4105.
93. Jones, S., Wang, T.L., Kurman, R.J., Nakayama, K., Velculescu, V.E., Vogelstein, B., Kinzler, K.W., Papadopoulos, N., and Shih, I.M. (2012). Low-grade serous carcinomas of the ovary contain very few point mutations. J. Pathol. *226*, 413–420. https://doi.org/10.1002/path.3967.
94. Bashashati, A., Ha, G., Tone, A., Ding, J., Prentice, L.M., Roth, A., Rosner, J., Shumansky, K., Kalloger, S., Senz, J., et al. (2013). Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. J. Pathol. *231*, 21–34. https://doi.org/10.1002/path.4230.
95. Chong, I.Y., Cunningham, D., Barber, L.J., Campbell, J., Chen, L., Kozarewa, I., Fenwick, K., Assiotis, I., Guettler, S., Garcia-Murillas, I., et al. (2013). The genomic landscape of oesophagogastric junctional adenocarcinoma. J. Pathol. *231*, 301–310. https://doi.org/10.1002/path.4247.
96. Szuhai, K., de Jong, D., Leung, W.Y., Fletcher, C.D.M., and Hogendoorn, P.C.W. (2014). Transactivating mutation of the MYOD1 gene is a frequent event in adult spindle cell rhabdomyosarcoma. J. Pathol. *232*, 300–307. https://doi.org/10.1002/path.4307.
97. Jiao, Y., Yonescu, R., Offerhaus, G.J.A., Klimstra, D.S., Maitra, A., Eshleman, J.R., Herman, J.G., Poh, W., Pelosof, L., Wolfgang, C.L., et al. (2014). Whole-exome sequencing of pancreatic neoplasms with acinar differentiation. J. Pathol. *232*, 428–435. https://doi.org/10.1002/path.4310.
98. Kim, T.M., Jung, S.H., Baek, I.P., Lee, S.H., Choi, Y.J., Lee, J.Y., Chung, Y.J., and Lee, S.H. (2014). Regional biases in mutation screening due to intratumoural heterogeneity of prostate cancer. J. Pathol. *233*, 425–435. https://doi.org/10.1002/path.4380.
99. Flynn, A., Benn, D., Clifton-Bligh, R., Robinson, B., Trainer, A.H., James, P., Hogg, A., Waldeck, K., George, J., Li, J., et al. (2015). The genomic landscape of phaeochromocytoma. J. Pathol. *236*, 78–89. https://doi.org/10.1002/path.4503.
100. Demeure, M.J., Aziz, M., Rosenberg, R., Gurley, S.D., Bussey, K.J., and Carpten, J.D. (2014). Whole-genome sequencing of an aggressive BRAF wild-type papillary thyroid cancer identified EML4-ALK translocation as a therapeutic target. World J. Surg. *38*, 1296–1305. https://doi.org/10.1007/s00268-014-2485-3.
101. Reuss, D.E., Piro, R.M., Jones, D.T.W., Simon, M., Ketter, R., Kool, M., Becker, A., Sahm, F., Pusch, S., Meyer, J., et al. (2013). Secretory meningiomas are defined by combined KLF4 K409Q and TRAF7 mutations. Acta Neuropathol. *125*, 351–358. https://doi.org/10.1007/s00401-013-1093-x.
102. Schweizer, L., Koelsche, C., Sahm, F., Piro, R.M., Capper, D., Reuss, D.E., Pusch, S., Habel, A., Meyer, J., Göck, T., et al. (2013). Meningeal hemangiopericytoma and solitary fibrous tumors carry the NAB2-STAT6 fusion and can be diagnosed by nuclear expression of STAT6 protein. Acta Neuropathol. *125*, 651–658. https://doi.org/10.1007/s00401-013-1117-6.
103. Zhang, L., Zhou, Y., Cheng, C., Cui, H., Cheng, L., Kong, P., Wang, J., Li, Y., Chen, W., Song, B., et al. (2015). Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. Am. J. Hum. Genet. *96*, 597–611. https://doi.org/10.1016/j.ajhg.2015.02.017.
104. Babushok, D.V., Perdigones, N., Perin, J.C., Olson, T.S., Ye, W., Roth, J.J., Lind, C., Cattier, C., Li, Y., Hartung, H., et al. (2015). Emergence of clonal hematopoiesis in the majority of patients with acquired aplastic anemia. Cancer Genet. *208*, 115–128. https://doi.org/10.1016/j.cancergen.2015.01.007.
105. Roberts, K.G., Morin, R.D., Zhang, J., Hirst, M., Zhao, Y., Su, X., Chen, S.C., Payne-Turner, D., Churchman, M.L., Harvey, R.C., et al. (2012). Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. Cancer Cell *22*, 153–166. https://doi.org/10.1016/j.ccr.2012.06.005.
106. Chen, X., Stewart, E., Shelat, A.A., Qu, C., Bahrami, A., Hatley, M., Wu, G., Bradley, C., McEvoy, J., Pappo, A., et al. (2013). Targeting oxidative stress in embryonal rhabdomyosarcoma. Cancer Cell *24*, 710–724. https://doi.org/10.1016/j.ccr.2013.11.002.
107. Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, 886–895. https://doi.org/10.1016/j.cell.2012.02.025.
108. Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., Auclair, D., Li, L., Place, C., et al. (2012). A landscape of driver mutations in melanoma. Cell *150*, 251–263. https://doi.org/10.1016/j.cell.2012.06.024.
109. Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714–726. https://doi.org/10.1016/j.cell.2013.01.019.
110. Pasqualucci, L., Khiabanian, H., Fangazio, M., Vasishtha, M., Messina, M., Holmes, A.B., Ouillette, P., Trifonov, V., Rossi, D., Tabbò, F., et al. (2014). Genetics of follicular lymphoma transformation. Cell Rep. *6*, 130–140. https://doi.org/10.1016/j.celrep.2013.12.027.
111. Chen, X., Bahrami, A., Pappo, A., Easton, J., Dalton, J., Hedlund, E., Ellison, D., Shurtleff, S., Wu, G., Wei, L., et al. (2014). Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. Cell Rep. *7*, 104–112. https://doi.org/10.1016/j.celrep.2014.03.003.
112. Nordentoft, I., Lamy, P., Birkenkamp-Demtröder, K., Shumansky, K., Vang, S., Hornshøj, H., Juul, M., Villesen, P., Hedegaard, J., Roth, A., et al. (2014). Mutational context and diverse clonal development in early and late bladder cancer. Cell Rep. *7*, 1649–1663. https://doi.org/10.1016/j.celrep.2014.04.038.
113. Lindberg, J., Mills, I.G., Klevebring, D., Liu, W., Neiman, M., Xu, J., Wikström, P., Wiklund, P., Wiklund, F., Egevad, L., and Grönberg, H. (2013). The mitochondrial and autosomal mutation landscapes of prostate cancer. Eur. Urol. *63*, 702–708. https://doi.org/10.1016/j.eururo.2012.11.053.
114. Cutcutache, I., Suzuki, Y., Tan, I.B., Ramgopal, S., Zhang, S., Ramnarayanan, K., Gan, A., Lee, H.H., Tay, S.T., Ooi, A., et al. (2015). Exome-wide sequencing shows low mutation rates and identifies novel mutated genes in seminomas. Eur. Urol. *68*, 77–83. https://doi.org/10.1016/j.eururo.2014.12.040.
115. Brabrand, S., Johannessen, B., Axcrona, U., Kraggerud, S.M., Berg, K.G., Bakken, A.C., Bruun, J., Fosså, S.D., Lothe, R.A., Lehne, G., and Skotheim, R.I. (2015). Exome sequencing of bilateral testicular germ cell tumors suggests independent development lineages. Neoplasia *17*, 167–174. https://doi.org/10.1016/j.neo.2014.12.005.
116. Leich, E., Weißbach, S., Klein, H.U., Grieb, T., Pischimarov, J., Stühmer, T., Chatterjee, M., Steinbrunn, T., Langer, C., Eilers, M., et al. (2013). Multiple myeloma is affected by multiple and heterogeneous somatic mutations in adhesion- and receptor tyrosine kinase signaling molecules. Blood Cancer J. *3*, e102. https://doi.org/10.1038/bcj.2012.47.
117. Menezes, J., Salgado, R.N., Acquadro, F., Gómez-López, G., Carralero, M.C., Barroso, A., Mercadillo, F., Espinosa-Hevia, L., Talavera-Casañas, J.G., Pisano, D.G., et al. (2013). ASXL1, TP53 and IKZF3 mutations are present in the chronic phase and blast crisis of chronic myeloid leukemia. Blood Cancer J. *3*, e157. https://doi.org/10.1038/bcj.2013.54.
118. Menezes, J., Makishima, H., Gomez, I., Acquadro, F., Gómez-López, G., Graña, O., Dopazo, A., Alvarez, S., Trujillo, M., Pisano, D.G., et al. (2013). CSF3R T618I co-occurs with mutations of splicing and epigenetic genes and with a new PIM3 truncated fusion gene in chronic neutrophilic leukemia. Blood Cancer J. *3*, e158. https://doi.org/10.1038/bcj.2013.55.

119. Lasho, T.L., Finke, C.M., Zblewski, D., Patnaik, M., Ketterling, R.P., Chen, D., Hanson, C.A., Tefferi, A., and Pardanani, A. (2015). Novel recurrent mutations in ethanolamine kinase 1 (ETNK1) gene in systemic mastocytosis with eosinophilia and chronic myelomonocytic leukemia. Blood Cancer J. *5*, e275. https://doi.org/10.1038/bcj.2014.94.

120. Yu, C., Yu, J., Yao, X., Wu, W.K.K., Lu, Y., Tang, S., Li, X., Bao, L., Li, X., Hou, Y., et al. (2014). Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. Cell Res. *24*, 701–712. https://doi.org/10.1038/cr.2014.43.

121. Emmerich, D., Zemojtel, T., Hecht, J., Krawitz, P., Spielmann, M., Kühnisch, J., Kobus, K., Osswald, M., Heinrich, V., Berlien, P., et al. (2015). Somatic neurofibromatosis type 1 (NF1) inactivation events in cutaneous neurofibromas of a single NF1 patient. Eur. J. Hum. Genet. *23*, 870–873. https://doi.org/10.1038/ejhg.2014.210.

122. Greif, P.A., Yaghmaie, M., Konstandin, N.P., Ksienzyk, B., Alimoghaddam, K., Ghavamzadeh, A., Hauser, A., Graf, A., Krebs, S., Blum, H., and Bohlander, S.K. (2011). Somatic mutations in acute promyelocytic leukemia (APL) identified by exome sequencing. Leukemia *25*, 1519–1522. https://doi.org/10.1038/leu.2011.114.

123. Lilljebjörn, H., Rissler, M., Lassen, C., Heldrup, J., Behrendtz, M., Mitelman, F., Johansson, B., and Fioretos, T. (2012). Whole-exome sequencing of pediatric acute lymphoblastic leukemia. Leukemia *26*, 1602–1607. https://doi.org/10.1038/leu.2011.333.

124. Menezes, J., Acquadro, F., Wiseman, M., Gómez-López, G., Salgado, R.N., Talavera-Casañas, J.G., Buño, I., Cervera, J.V., Montes-Moreno, S., Hernández-Rivas, J.M., et al. (2014). Exome sequencing reveals novel and recurrent mutations with clinical impact in blastic plasmacytoid dendritic cell neoplasm. Leukemia *28*, 823–829. https://doi.org/10.1038/leu.2013.283.

125. Bateman, C.M., Alpar, D., Ford, A.M., Colman, S.M., Wren, D., Morgan, M., Kearney, L., and Greaves, M. (2015). Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. Leukemia *29*, 58–65. https://doi.org/10.1038/leu.2014.177.

126. Kontro, M., Kuusanmäki, H., Eldfors, S., Burmeister, T., Andersson, E.I., Bruserud, O., Brümmendorf, T.H., Edgren, H., Gjertsen, B.T., Itälä-Remes, M., et al. (2014). Novel activating STAT5B mutations as putative drivers of T-cell acute lymphoblastic leukemia. Leukemia *28*, 1738–1742. https://doi.org/10.1038/leu.2014.89.

127. Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature *455*, 1069–1075. https://doi.org/10.1038/nature07423.

128. Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809–813. https://doi.org/10.1038/nature08489.

129. Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999–1005. https://doi.org/10.1038/nature08989.

130. Lee, W., Jiang, Z., Liu, J., Haverty, P.M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K.P., Bhatt, D., et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature *465*, 473–477. https://doi.org/10.1038/nature09004.

131. Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C.K., Stephens, P., Davies, H., Jones, D., Lin, M.L., Teague, J., et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature *469*, 539–542. https://doi.org/10.1038/nature09639.

132. Puente, X.S., Pinyol, M., Quesada, V., Conde, L., Ordóñez, G.R., Villamor, N., Escaramis, G., Jares, P., Beà, S., González-Díaz, M., et al. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature *475*, 101–105. https://doi.org/10.1038/nature10113.

133. Morin, R.D., Mendez-Lago, M., Mungall, A.J., Goya, R., Mungall, K.L., Corbett, R.D., Johnson, N.A., Severson, T.M., Chiu, R., Field, M., et al. (2011). Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature *476*, 298–303. https://doi.org/10.1038/nature10351.

134. Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69. https://doi.org/10.1038/nature10496.

135. Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163. https://doi.org/10.1038/nature10725.

136. Molenaar, J.J., Koster, J., Zwijnenburg, D.A., van Sluis, P., Valentijn, L.J., van der Ploeg, I., Hamdi, M., van Nes, J., Westerman, B.A., van Arkel, J., et al. (2012). Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature *483*, 589–593. https://doi.org/10.1038/nature10910.

137. Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395–399. https://doi.org/10.1038/nature10933.

138. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400–404. https://doi.org/10.1038/nature11017.

139. Berger, M.F., Hodis, E., Heffernan, T.P., Deribe, Y.L., Lawrence, M.S., Protopopov, A., Ivanova, E., Watson, I.R., Nickerson, E., Ghosh, P., et al. (2012). Melanoma genome sequencing reveals frequent PREX2 mutations. Nature *485*, 502–506. https://doi.org/10.1038/nature11071.

140. Grasso, C.S., Wu, Y.M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature *487*, 239–243. https://doi.org/10.1038/nature11125.

141. Seshagiri, S., Stawiski, E.W., Durinck, S., Modrusan, Z., Storm, E.E., Conboy, C.B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B.S., et al. (2012). Recurrent R-spondin fusions in colon cancer. Nature *488*, 660–664. https://doi.org/10.1038/nature11282.

142. Jones, D.T.W., Jäger, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.J., Pugh, T.J., Hovestadt, V., Stütz, A.M., et al. (2012). Dissecting the genomic complexity underlying medulloblastoma. Nature *488*, 100–105. https://doi.org/10.1038/nature11284.

143. Pugh, T.J., Weeraratne, S.D., Archer, T.C., Pomeranz Krummel, D.A., Auclair, D., Bochicchio, J., Carneiro, M.O., Carter, S.L., Cibulskis, K., Erlich, R.L., et al. (2012). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature *488*, 106–110. https://doi.org/10.1038/nature11329.

144. Biankin, A.V., Waddell, N., Kassahn, K.S., Gingras, M.C., Muthuswamy, L.B., Johns, A.L., Miller, D.K., Wilson, P.J., Patch, A.M., Wu, J., et al. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature *491*, 399–405. https://doi.org/10.1038/nature11547.

145. Murtaza, M., Dawson, S.J., Tsui, D.W.Y., Gale, D., Forshew, T., Piskorz, A.M., Parkinson, C., Chin, S.F., Kingsbury, Z., Wong, A.S.C., et al. (2013). Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature *497*, 108–112. https://doi.org/10.1038/nature12065.

146. Wang, L., Yamaguchi, S., Burstein, M.D., Terashima, K., Chang, K., Ng, H.K., Nakamura, H., He, Z., Doddapaneni, H., Lewis, L., et al. (2014). Novel somatic and germline mutations in intracranial germ cell tumours. Nature *511*, 241–245. https://doi.org/10.1038/nature13296.

147. Shi, H., Moriceau, G., Kong, X., Lee, M.K., Lee, H., Koya, R.C., Ng, C., Chodon, T., Scolyer, R.A., Dahlman, K.B., et al. (2012). Melanoma whole-exome sequencing identifies (V600E)B-RAF amplification-mediated

acquired B-RAF inhibitor resistance. Nat. Commun. *3*, 724. https://doi.org/10.1038/ncomms1727.

148. India Project Team of the International Cancer Genome Consortium (2013). Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat. Commun. *4*, 2873. https://doi.org/10.1038/ncomms3873.

149. Fernandez-Cuesta, L., Peifer, M., Lu, X., Sun, R., Ozretić, L., Seidal, D., Zander, T., Leenders, F., George, J., Müller, C., et al. (2014). Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. Nat. Commun. *5*, 3518. https://doi.org/10.1038/ncomms4518.

150. Torrezan, G.T., Ferreira, E.N., Nakahata, A.M., Barros, B.D.F., Castro, M.T.M., Correa, B.R., Krepischi, A.C.V., Olivieri, E.H.R., Cunha, I.W., Tabori, U., et al. (2014). Recurrent somatic mutation in DROSHA induces microRNA profile changes in Wilms tumour. Nat. Commun. *5*, 4039. https://doi.org/10.1038/ncomms5039.

151. Nikolaev, S.I., Garieri, M., Santoni, F., Falconnet, E., Ribaux, P., Guipponi, M., Murray, A., Groet, J., Giarin, E., Basso, G., et al. (2014). Frequent cases of RAS-mutated Down syndrome acute lymphoblastic leukaemia lack JAK2 mutations. Nat. Commun. *5*, 4654. https://doi.org/10.1038/ncomms5654.

152. Rakheja, D., Chen, K.S., Liu, Y., Shukla, A.A., Schmid, V., Chang, T.C., Khokhar, S., Wickiser, J.E., Karandikar, N.J., Malter, J.S., et al. (2014). Somatic mutations in DROSHA and DICER1 impair microRNA biogenesis through distinct mechanisms in Wilms tumours. Nat. Commun. *2*, 4802. https://doi.org/10.1038/ncomms5802.

153. Nikolaev, S., Santoni, F., Garieri, M., Makrythanasis, P., Falconnet, E., Guipponi, M., Vannier, A., Radovanovic, I., Bena, F., Forestier, F., et al. (2014). Extrachromosomal driver mutations in glioblastoma and low-grade glioma. Nat. Commun. *5*, 5690. https://doi.org/10.1038/ncomms6690.

154. Litchfield, K., Summersgill, B., Yost, S., Sultana, R., Labreche, K., Dudakia, D., Renwick, A., Seal, S., Al-Saadi, R., Broderick, P., et al. (2015). Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. Nat. Commun. *6*, 5973. https://doi.org/10.1038/ncomms6973.

155. Pinto, E.M., Chen, X., Easton, J., Finkelstein, D., Liu, Z., Pounds, S., Rodriguez-Galindo, C., Lund, T.C., Mardis, E.R., Wilson, R.K., et al. (2015). Genomic landscape of paediatric adrenocortical tumours. Nat. Commun. *6*, 6302. https://doi.org/10.1038/ncomms7302.

156. Ma, X., Edmonson, M., Yergeau, D., Muzny, D.M., Hampton, O.A., Rusch, M., Song, G., Easton, J., Harvey, R.C., Wheeler, D.A., et al. (2015). Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat. Commun. *6*, 6604. https://doi.org/10.1038/ncomms7604.

157. Witkiewicz, A.K., McMillan, E.A., Balaji, U., Baek, G., Lin, W.C., Mansour, J., Mollaee, M., Wagner, K.U., Koduru, P., Yopp, A., et al. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat. Commun. *6*, 6744. https://doi.org/10.1038/ncomms7744.

158. Guo, G., Gui, Y., Gao, S., Tang, A., Hu, X., Huang, Y., Jia, W., Li, Z., He, M., Sun, L., et al. (2011). Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nat. Genet. *44*, 17–19. https://doi.org/10.1038/ng.1014.

159. Nikolaev, S.I., Rimoldi, D., Iseli, C., Valsesia, A., Robyr, D., Gehrig, C., Harshman, K., Guipponi, M., Bukach, O., Zoete, V., et al. (2011). Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat. Genet. *44*, 133–139. https://doi.org/10.1038/ng.1026.

160. Quesada, V., Conde, L., Villamor, N., Ordóñez, G.R., Jares, P., Bassaganyas, L., Ramsay, A.J., Beà, S., Pinyol, M., Martínez-Trillos, A., et al. (2011). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat. Genet. *44*, 47–52. https://doi.org/10.1038/ng.1032.

161. Stark, M.S., Woods, S.L., Gartside, M.G., Bonazzi, V.F., Dutton-Regester, K., Aoude, L.G., Chow, D., Sereduk, C., Niemi, N.M., Tang, N., et al. (2011). Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. Nat. Genet. *44*, 165–169. https://doi.org/10.1038/ng.1041.

162. Zang, Z.J., Cutcutache, I., Poon, S.L., Zhang, S.L., McPherson, J.R., Tao, J., Rajasegaran, V., Heng, H.L., Deng, N., Gan, A., et al. (2012). Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat. Genet. *44*, 570–574. https://doi.org/10.1038/ng.2246.

163. Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I.B., Calderaro, J., Bioulac-Sage, P., Letexier, M., Degos, F., et al. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat. Genet. *44*, 694–698. https://doi.org/10.1038/ng.2256.

164. Ong, C.K., Subimerb, C., Pairojkul, C., Wongkham, S., Cutcutache, I., Yu, W., McPherson, J.R., Allen, G.E., Ng, C.C.Y., Wong, B.H., et al. (2012). Exome sequencing of liver fluke-associated cholangiocarcinoma. Nat. Genet. *44*, 690–693. https://doi.org/10.1038/ng.2273.

165. Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. *44*, 685–689. https://doi.org/10.1038/ng.2279.

166. Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K.A., Hosoda, F., Nguyen, H.H., Aoki, M., Hosono, N., Kubo, M., Miya, F., et al. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat. Genet. *44*, 760–764. https://doi.org/10.1038/ng.2291.

167. Peña-Llopis, S., Vega-Rubín-de-Celis, S., Liao, A., Leng, N., Pavía-Jiménez, A., Wang, S., Yamasaki, T., Zhrebker, L., Sivanand, S., Spence, P., et al. (2012). BAP1 loss defines a new class of renal cell carcinoma. Nat. Genet. *44*, 751–759. https://doi.org/10.1038/ng.2323.

168. Krauthammer, M., Kong, Y., Ha, B.H., Evans, P., Bacchiocchi, A., McCusker, J.P., Cheng, E., Davis, M.J., Goh, G., Choi, M., et al. (2012). Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat. Genet. *44*, 1006–1014. https://doi.org/10.1038/ng.2359.

169. Huang, J., Deng, Q., Wang, Q., Li, K.Y., Dai, J.H., Li, N., Zhu, Z.D., Zhou, B., Liu, X.Y., Liu, R.F., et al. (2012). Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat. Genet. *44*, 1117–1121. https://doi.org/10.1038/ng.2391.

170. Peifer, M., Fernández-Cuesta, L., Sos, M.L., George, J., Seidel, D., Kasper, L.H., Plenker, D., Leenders, F., Sun, R., Zander, T., et al. (2012). Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat. Genet. *44*, 1104–1110. https://doi.org/10.1038/ng.2396.

171. Rudin, C.M., Durinck, S., Stawiski, E.W., Poirier, J.T., Modrusan, Z., Shames, D.S., Bergbower, E.A., Guan, Y., Shin, J., Guillory, J., et al. (2012). Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat. Genet. *44*, 1111–1116. https://doi.org/10.1038/ng.2405.

172. Le Gallo, M., O'Hara, A.J., Rudd, M.L., Urick, M.E., Hansen, N.F., O'Neil, N.J., Price, J.C., Zhang, S., England, B.M., Godwin, A.K., et al. (2012). Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nat. Genet. *44*, 1310–1315. https://doi.org/10.1038/ng.2455.

173. Sausen, M., Leary, R.J., Jones, S., Wu, J., Reynolds, C.P., Liu, X., Blackford, A., Parmigiani, G., Diaz, L.A., Jr., Papadopoulos, N., et al. (2013). Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. Nat. Genet. *45*, 12–17. https://doi.org/10.1038/ng.2493.

174. Piazza, R., Valletta, S., Winkelmann, N., Redaelli, S., Spinelli, R., Pirola, A., Antolini, L., Mologni, L., Donadoni, C., Papaemmanuil, E., et al.

(2013). Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. Nat. Genet. *45*, 18–24. https://doi.org/10.1038/ng.2495.

175. De Keersmaecker, K., Atak, Z.K., Li, N., Vicente, C., Patchett, S., Girardi, T., Gianfelici, V., Geerdens, E., Clappier, E., Porcu, M., et al. (2013). Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat. Genet. *45*, 186–190. https://doi.org/10.1038/ng.2508.

176. Robinson, D.R., Wu, Y.M., Kalyana-Sundaram, S., Cao, X., Lonigro, R.J., Sung, Y.S., Chen, C.L., Zhang, L., Wang, R., Su, F., et al. (2013). Identification of recurrent NAB2-STAT6 gene fusions in solitary fibrous tumor by integrative sequencing. Nat. Genet. *45*, 180–185. https://doi.org/10.1038/ng.2509.

177. Pugh, T.J., Morozova, O., Attiyeh, E.F., Asgharzadeh, S., Wei, J.S., Auclair, D., Carter, S.L., Cibulskis, K., Hanna, M., Kiezun, A., et al. (2013). The genetic landscape of high-risk neuroblastoma. Nat. Genet. *45*, 279–284. https://doi.org/10.1038/ng.2529.

178. Holmfeldt, L., Wei, L., Diaz-Flores, E., Walsh, M., Zhang, J., Ding, L., Payne-Turner, D., Churchman, M., Andersson, A., Chen, S.C., et al. (2013). The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat. Genet. *45*, 242–252. https://doi.org/10.1038/ng.2532.

179. Meyer, J.A., Wang, J., Hogan, L.E., Yang, J.J., Dandekar, S., Patel, J.P., Tang, Z., Zumbo, P., Li, S., Zavadil, J., et al. (2013). Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. Nat. Genet. *45*, 290–294. https://doi.org/10.1038/ng.2558.

180. Ramsay, A.J., Quesada, V., Foronda, M., Conde, L., Martínez-Trillos, A., Villamor, N., Rodríguez, D., Kwarciak, A., Garabaya, C., Gallardo, M., et al. (2013). POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. Nat. Genet. *45*, 526–530. https://doi.org/10.1038/ng.2584.

181. Dulak, A.M., Stojanov, P., Peng, S., Lawrence, M.S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S.E., Shefler, E., et al. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat. Genet. *45*, 478–486. https://doi.org/10.1038/ng.2591.

182. Ho, A.S., Kannan, K., Roy, D.M., Morris, L.G.T., Ganly, I., Katabi, N., Ramaswami, D., Walsh, L.A., Eng, S., Huse, J.T., et al. (2013). The mutational landscape of adenoid cystic carcinoma. Nat. Genet. *45*, 791–798. https://doi.org/10.1038/ng.2643.

183. Martin, M., Maßhöfer, L., Temming, P., Rahmann, S., Metz, C., Bornfeld, N., van de Nes, J., Klein-Hitpass, L., Hinnebusch, A.G., Horsthemke, B., et al. (2013). Exome sequencing identifies recurrent somatic mutations in EIF1AX and SF3B1 in uveal melanoma with disomy 3. Nat. Genet. *45*, 933–936. https://doi.org/10.1038/ng.2674.

184. Scholl, U.I., Goh, G., Stölting, G., de Oliveira, R.C., Choi, M., Overton, J.D., Fonseca, A.L., Korah, R., Starker, L.F., Kunstman, J.W., et al. (2013). Somatic and germline CACNA1D calcium channel mutations in aldosterone-producing adenomas and primary aldosteronism. Nat. Genet. *45*, 1050–1054. https://doi.org/10.1038/ng.2695.

185. Makishima, H., Yoshida, K., Nguyen, N., Przychodzen, B., Sanada, M., Okuno, Y., Ng, K.P., Gudmundsson, K.O., Vishwakarma, B.A., Jerez, A., et al. (2013). Somatic SETBP1 mutations in myeloid malignancies. Nat. Genet. *45*, 942–946. https://doi.org/10.1038/ng.2696.

186. Azizan, E.A.B., Poulsen, H., Tuluc, P., Zhou, J., Clausen, M.V., Lieb, A., Maniero, C., Garg, S., Bochukova, E.G., Zhao, W., et al. (2013). Somatic mutations in ATP1A1 and CACNA1D underlie a common subtype of adrenal hypertension. Nat. Genet. *45*, 1055–1060. https://doi.org/10.1038/ng.2716.

187. Guo, G., Sun, X., Chen, C., Wu, S., Huang, P., Li, Z., Dean, M., Huang, Y., Jia, W., Zhou, Q., et al. (2013). Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. Nat. Genet. *45*, 1459–1463. https://doi.org/10.1038/ng.2798.

188. Balbás-Martínez, C., Sagrera, A., Carrillo-de-Santa-Pau, E., Earl, J., Márquez, M., Vazquez, M., Lapi, E., Castro-Giner, F., Beltran, S., Bayés, M., et al. (2013). Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. Nat. Genet. *45*, 1464–1469. https://doi.org/10.1038/ng.2799.

189. Jiao, Y., Pawlik, T.M., Anders, R.A., Selaru, F.M., Streppel, M.M., Lucas, D.J., Niknafs, N., Guthrie, V.B., Maitra, A., Argani, P., et al. (2013). Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. Nat. Genet. *45*, 1470–1473. https://doi.org/10.1038/ng.2813.

190. Robinson, D.R., Wu, Y.M., Vats, P., Su, F., Lonigro, R.J., Cao, X., Kalyana-Sundaram, S., Wang, R., Ning, Y., Hodges, L., et al. (2013). Activating ESR1 mutations in hormone-resistant metastatic breast cancer. Nat. Genet. *45*, 1446–1451. https://doi.org/10.1038/ng.2823.

191. Waterfall, J.J., Arons, E., Walker, R.L., Pineda, M., Roth, L., Killian, J.K., Abaan, O.D., Davis, S.R., Kreitman, R.J., and Meltzer, P.S. (2014). High prevalence of MAP2K1 mutations in variant and IGHV4-34-expressing hairy-cell leukemias. Nat. Genet. *46*, 8–10. https://doi.org/10.1038/ng.2828.

192. Brastianos, P.K., Taylor-Weiner, A., Manley, P.E., Jones, R.T., Dias-Santagata, D., Thorner, A.R., Lawrence, M.S., Rodriguez, F.J., Bernardo, L.A., Schubert, L., et al. (2014). Exome sequencing identifies BRAF mutations in papillary craniopharyngiomas. Nat. Genet. *46*, 161–165. https://doi.org/10.1038/ng.2868.

193. Palomero, T., Couronné, L., Khiabanian, H., Kim, M.Y., Ambesi-Impiombato, A., Perez-Garcia, A., Carpenter, Z., Abate, F., Allegretta, M., Haydu, J.E., et al. (2014). Recurrent mutations in epigenetic regulators, RHOA and FYN kinase in peripheral T cell lymphomas. Nat. Genet. *46*, 166–170. https://doi.org/10.1038/ng.2873.

194. Gerlinger, M., Horswell, S., Larkin, J., Rowan, A.J., Salm, M.P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C.R., et al. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat. Genet. *46*, 225–233. https://doi.org/10.1038/ng.2891.

195. Yoo, H.Y., Sung, M.K., Lee, S.H., Kim, S., Lee, H., Park, S., Kim, S.C., Lee, B., Rho, K., Lee, J.E., et al. (2014). A recurrent inactivating mutation in RHOA GTPase in angioimmunoblastic T cell lymphoma. Nat. Genet. *46*, 371–375. https://doi.org/10.1038/ng.2916.

196. Taylor, K.R., Mackay, A., Truffaux, N., Butterfield, Y., Morozova, O., Philippe, C., Castel, D., Grasso, C.S., Vinci, M., Carvalho, D., et al. (2014). Recurrent activating ACVR1 mutations in diffuse intrinsic pontine glioma. Nat. Genet. *46*, 457–461. https://doi.org/10.1038/ng.2925.

197. Assié, G., Letouzé, E., Fassnacht, M., Jouinot, A., Luscap, W., Barreau, O., Omeiri, H., Rodriguez, S., Perlemoine, K., René-Corail, F., et al. (2014). Integrated genomic characterization of adrenocortical carcinoma. Nat. Genet. *46*, 607–612. https://doi.org/10.1038/ng.2953.

198. Kohsaka, S., Shukla, N., Ameur, N., Ito, T., Ng, C.K.Y., Wang, L., Lim, D., Marchetti, A., Viale, A., Pirun, M., et al. (2014). A recurrent neomorphic mutation in MYOD1 defines a clinically aggressive subset of embryonal rhabdomyosarcoma associated with PI3K-AKT pathway mutations. Nat. Genet. *46*, 595–600. https://doi.org/10.1038/ng.2969.

199. Kakiuchi, M., Nishizawa, T., Ueda, H., Gotoh, K., Tanaka, A., Hayashi, A., Yamamoto, S., Tatsuno, K., Katoh, H., Watanabe, Y., et al. (2014). Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. Nat. Genet. *46*, 583–587. https://doi.org/10.1038/ng.2984.

200. Zhang, L., Chen, L.H., Wan, H., Yang, R., Wang, Z., Feng, J., Yang, S., Jones, S., Wang, S., Zhou, W., et al. (2014). Exome sequencing identifies somatic gain-of-function PPM1D mutations in brainstem gliomas. Nat. Genet. *46*, 726–730. https://doi.org/10.1038/ng.2995.

201. Lin, D.C., Meng, X., Hazawa, M., Nagata, Y., Varela, A.M., Xu, L., Sato, Y., Liu, L.Z., Ding, L.W., Sharma, A., et al. (2014). The genomic landscape of nasopharyngeal carcinoma. Nat. Genet. *46*, 866–871. https://doi.org/10.1038/ng.3006.

202. Li, M., Zhang, Z., Li, X., Ye, J., Wu, X., Tan, Z., Liu, C., Shen, B., Wang, X.A., Wu, W., et al. (2014). Whole-exome and targeted gene sequencing

of gallbladder carcinoma identifies recurrent mutations in the ErbB pathway. Nat. Genet. *46*, 872–876. https://doi.org/10.1038/ng.3030.

203. Lim, W.K., Ong, C.K., Tan, J., Thike, A.A., Ng, C.C.Y., Rajasegaran, V., Myint, S.S., Nagarajan, S., Nasir, N.D.M., McPherson, J.R., et al. (2014). Exome sequencing identifies highly recurrent MED12 somatic mutations in breast fibroadenoma. Nat. Genet. *46*, 877–880. https://doi.org/10.1038/ng.3037.

204. Gao, Y.B., Chen, Z.L., Li, J.G., Hu, X.D., Shi, X.J., Sun, Z.M., Zhang, F., Zhao, Z.R., Li, Z.T., Liu, Z.Y., et al. (2014). Genetic landscape of esophageal squamous cell carcinoma. Nat. Genet. *46*, 1097–1102. https://doi.org/10.1038/ng.3076.

205. Giannakis, M., Hodis, E., Jasmine Mu, X., Yamauchi, M., Rosenbluh, J., Cibulskis, K., Saksena, G., Lawrence, M.S., Qian, Z.R., Nishihara, R., et al. (2014). RNF43 is frequently mutated in colorectal and endometrial cancers. Nat. Genet. *46*, 1264–1266. https://doi.org/10.1038/ng.3127.

206. Durinck, S., Stawiski, E.W., Pavía-Jiménez, A., Modrusan, Z., Kapur, P., Jaiswal, B.S., Zhang, N., Toffessi-Tcheuyap, V., Nguyen, T.T., Pahuja, K.B., et al. (2015). Spectrum of diverse genomic alterations define non-clear cell renal carcinoma subtypes. Nat. Genet. *47*, 13–21. https://doi.org/10.1038/ng.3146.

207. Bai, H., Harmancı, A.S., Erson-Omay, E.Z., Li, J., Coşkun, S., Simon, M., Krischek, B., Özduman, K., Omay, S.B., Sorensen, E.A., et al. (2016). Integrated genomic characterization of IDH1-mutant glioma malignant progression. Nat. Genet. *48*, 59–66. https://doi.org/10.1038/ng.3457.

208. Bonilla, X., Parmentier, L., King, B., Bezrukov, F., Kaya, G., Zoete, V., Seplyarskiy, V.B., Sharpe, H.J., McKee, T., Letourneau, A., et al. (2016). Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. Nat. Genet. *48*, 398–406. https://doi.org/10.1038/ng.3525.

209. Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D.I.S., Zairis, S., Abate, F., Liu, Z., Elliott, O., Shin, Y.J., et al. (2016). Clonal evolution of glioblastoma under therapy. Nat. Genet. *48*, 768–776. https://doi.org/10.1038/ng.3590.

210. Yan, X.J., Xu, J., Gu, Z.H., Pan, C.M., Lu, G., Shen, Y., Shi, J.Y., Zhu, Y.M., Tang, L., Zhang, X.W., et al. (2011). Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat. Genet. *43*, 309–315. https://doi.org/10.1038/ng.788.

211. Wei, X., Walia, V., Lin, J.C., Teer, J.K., Prickett, T.D., Gartner, J., Davis, S., NISC Comparative Sequencing Program; Stemke-Hale, K., Davies, M.A., et al. (2011). Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat. Genet. *43*, 442–446. https://doi.org/10.1038/ng.810.

212. Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V.A., Grunn, A., Messina, M., Elliot, O., et al. (2011). Analysis of the coding genome of diffuse large B-cell lymphoma. Nat. Genet. *43*, 830–837. https://doi.org/10.1038/ng.892.

213. Li, M., Zhao, H., Zhang, X., Wood, L.D., Anders, R.A., Choti, M.A., Pawlik, T.M., Daniel, H.D., Kannangai, R., Offerhaus, G.J.A., et al. (2011). Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat. Genet. *43*, 828–829. https://doi.org/10.1038/ng.903.

214. Gui, Y., Guo, G., Huang, Y., Hu, X., Tang, A., Gao, S., Wu, R., Chen, C., Li, X., Zhou, L., et al. (2011). Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nat. Genet. *43*, 875–878. https://doi.org/10.1038/ng.907.

215. Bass, A.J., Lawrence, M.S., Brace, L.E., Ramos, A.H., Drier, Y., Cibulskis, K., Sougnez, C., Voet, D., Saksena, G., Sivachenko, A., et al. (2011). Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat. Genet. *43*, 964–968. https://doi.org/10.1038/ng.936.

216. Wang, K., Kan, J., Yuen, S.T., Shi, S.T., Chu, K.M., Law, S., Chan, T.L., Kan, Z., Chan, A.S.Y., Tsui, W.Y., et al. (2011). Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat. Genet. *43*, 1219–1223. https://doi.org/10.1038/ng.982.

217. Tzoneva, G., Perez-Garcia, A., Carpenter, Z., Khiabanian, H., Tosello, V., Allegretta, M., Paietta, E., Racevskis, J., Rowe, J.M., Tallman, M.S., et al. (2013). Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nat. Med. *19*, 368–371. https://doi.org/10.1038/nm.3078.

218. Pugh, T.J., Yu, W., Yang, J., Field, A.L., Ambrogio, L., Carter, S.L., Cibulskis, K., Giannikopoulos, P., Kiezun, A., Kim, J., et al. (2014). Exome sequencing of pleuropulmonary blastoma reveals frequent biallelic loss of TP53 and two hits in DICER1 resulting in retention of 5p-derived miRNA hairpin loop sequences. Oncogene *33*, 5295–5302. https://doi.org/10.1038/onc.2014.150.

219. Furukawa, T., Sakamoto, H., Takeuchi, S., Ameri, M., Kuboki, Y., Yamamoto, T., Hatori, T., Yamamoto, M., Sugiyama, M., Ohike, N., et al. (2015). Whole exome sequencing reveals recurrent mutations in BRCA2 and FAT genes in acinar cell carcinomas of the pancreas. Sci. Rep. *5*, 8829. https://doi.org/10.1038/srep08829.

220. Tahara, T., Yamamoto, E., Madireddi, P., Suzuki, H., Maruyama, R., Chung, W., Garriga, J., Jelinek, J., Yamano, H.O., Sugai, T., et al. (2014). Colorectal carcinomas with CpG island methylator phenotype 1 frequently contain mutations in chromatin regulators. Gastroenterology *146*, 530–538.e5. https://doi.org/10.1053/j.gastro.2013.10.060.

221. Murphy, S.J., Hart, S.N., Lima, J.F., Kipp, B.R., Klebig, M., Winters, J.L., Szabo, C., Zhang, L., Eckloff, B.W., Petersen, G.M., et al. (2013). Genetic alterations associated with progression from pancreatic intraepithelial neoplasia to invasive pancreatic tumor. Gastroenterology *145*, 1098–1109.e1. https://doi.org/10.1053/j.gastro.2013.07.049.

222. Xue, R., Li, R., Guo, H., Guo, L., Su, Z., Ni, X., Qi, L., Zhang, T., Li, Q., Zhang, Z., et al. (2016). Variable intra-tumor genomic heterogeneity of multiple lesions in patients with hepatocellular carcinoma. Gastroenterology *150*, 998–1008. https://doi.org/10.1053/j.gastro.2015.12.033.

223. Sawada, G., Niida, A., Uchi, R., Hirata, H., Shimamura, T., Suzuki, Y., Shiraishi, Y., Chiba, K., Imoto, S., Takahashi, Y., et al. (2016). Genomic landscape of esophageal squamous cell carcinoma in a Japanese population. Gastroenterology *150*, 1171–1182. https://doi.org/10.1053/j.gastro.2016.01.035.

224. Gerlinger, M., Rowan, A.J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892. https://doi.org/10.1056/NEJMoa1113205.

225. Beuschlein, F., Fassnacht, M., Assié, G., Calebiro, D., Stratakis, C.A., Osswald, A., Ronchi, C.L., Wieland, T., Sbiera, S., Faucz, F.R., et al. (2014). Constitutive activation of PKA catalytic subunit in adrenal cushing's syndrome. N. Engl. J. Med. *370*, 1019–1028. https://doi.org/10.1056/NEJMoa1310359.

226. Klampfl, T., Gisslinger, H., Harutyunyan, A.S., Nivarthi, H., Rumi, E., Milosevic, J.D., Them, N.C.C., Berg, T., Gisslinger, B., Pietra, D., et al. (2013). Somatic mutations of calreticulin in myeloproliferative neoplasms. N. Engl. J. Med. *369*, 2379–2390. https://doi.org/10.1056/NEJMoa1311347.

227. Nangalia, J., Massie, C.E., Baxter, E.J., Nice, F.L., Gundem, G., Wedge, D.C., Avezov, E., Li, J., Kollmann, K., Kent, D.G., et al. (2013). Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. N. Engl. J. Med. *369*, 2391–2405. https://doi.org/10.1056/NEJMoa1312542.

228. Wagle, N., Grabiner, B.C., Van Allen, E.M., Amin-Mansour, A., Taylor-Weiner, A., Rosenberg, M., Gray, N., Barletta, J.A., Guo, Y., Swanson, S.J., et al. (2014). Response and acquired resistance to everolimus in anaplastic thyroid cancer. N. Engl. J. Med. *371*, 1426–1433. https://doi.org/10.1056/NEJMoa1403352.

229. Wu, J., Jiao, Y., Dal Molin, M., Maitra, A., de Wilde, R.F., Wood, L.D., Eshleman, J.R., Goggins, M.G., Wolfgang, C.L., Canto, M.I., et al. (2011).

Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways. Proc. Natl. Acad. Sci. USA *108*, 21188–21193. https://doi.org/10.1073/pnas.1118046108.

230. Zhang, J., Grubor, V., Love, C.L., Banerjee, A., Richards, K.L., Mieczkowski, P.A., Dunphy, C., Choi, W., Au, W.Y., Srivastava, G., et al. (2013). Genetic heterogeneity of diffuse large B-cell lymphoma. Proc. Natl. Acad. Sci. USA *110*, 1398–1403. https://doi.org/10.1073/pnas.1205299110.

231. Xu, L., Gu, Z.H., Li, Y., Zhang, J.L., Chang, C.K., Pan, C.M., Shi, J.Y., Shen, Y., Chen, B., Wang, Y.Y., et al. (2014). Genomic landscape of CD34+ hematopoietic cells in myelodysplastic syndrome and gene mutation profiles as prognostic markers. Proc. Natl. Acad. Sci. USA *111*, 8589–8594. https://doi.org/10.1073/pnas.1407688111.

232. Rossi, D., Trifonov, V., Fangazio, M., Bruscaggin, A., Rasi, S., Spina, V., Monti, S., Vaisitti, T., Arruga, F., Famà, R., et al. (2012). The coding genome of splenic marginal zone lymphoma: activation of NOTCH2 and other pathways regulating marginal zone development. J. Exp. Med. *209*, 1537–1551. https://doi.org/10.1084/jem.20120904.

233. Fabbri, G., Khiabanian, H., Holmes, A.B., Wang, J., Messina, M., Mullighan, C.G., Pasqualucci, L., Rabadan, R., and Dalla-Favera, R. (2013). Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. J. Exp. Med. *210*, 2273–2288. https://doi.org/10.1084/jem.20131448.

234. Xiong, D., Li, G., Li, K., Xu, Q., Pan, Z., Ding, F., Vedell, P., Liu, P., Cui, P., Hua, X., et al. (2012). Exome sequencing identifies MXRA5 as a novel cancer gene frequently mutated in non-small cell lung carcinoma from Chinese patients. Carcinogenesis *33*, 1797–1805. https://doi.org/10.1093/carcin/bgs210.

235. Mimaki, S., Totsuka, Y., Suzuki, Y., Nakai, C., Goto, M., Kojima, M., Arakawa, H., Takemura, S., Tanaka, S., Marubashi, S., et al. (2016). Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. Carcinogenesis *37*, 817–826. https://doi.org/10.1093/carcin/bgw066.

236. Kuhn, E., Wu, R.C., Guan, B., Wu, G., Zhang, J., Wang, Y., Song, L., Yuan, X., Wei, L., Roden, R.B.S., et al. (2012). Identification of molecular pathway aberrations in uterine serous carcinoma by genome-wide analyses. J. Natl. Cancer Inst. *104*, 1503–1513. https://doi.org/10.1093/jnci/djs345.

237. Aydin, I.T., Melamed, R.D., Adams, S.J., Castillo-Martin, M., Demir, A., Bryk, D., Brunner, G., Cordon-Cardo, C., Osman, I., Rabadan, R., and Celebi, J.T. (2014). FBXW7 mutations in melanoma and a new therapeutic paradigm. J. Natl. Cancer Inst. *106*, dju107. https://doi.org/10.1093/jnci/dju107.

238. Aihara, K., Mukasa, A., Gotoh, K., Saito, K., Nagae, G., Tsuji, S., Tatsuno, K., Yamamoto, S., Takayanagi, S., Narita, Y., et al. (2014). H3F3A K27M mutations in thalamic gliomas from young adult patients. Neuro. Oncol. *16*, 140–146. https://doi.org/10.1093/neuonc/not144.

239. Das, D., Kaur, I., Ali, M.J., Biswas, N.K., Das, S., Kumar, S., Honavar, S.G., Maitra, A., Chakrabarti, S., and Majumder, P.P. (2014). Exome sequencing reveals the likely involvement of SOX10 in uveal melanoma. Optom. Vis. Sci. *91*, e185–192. https://doi.org/10.1097/OPX.0000000000000309.

240. Wang, L., Tsutsumi, S., Kawaguchi, T., Nagasaki, K., Tatsuno, K., Yamamoto, S., Sang, F., Sonoda, K., Sugawara, M., Saiura, A., et al. (2012). Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Res. *22*, 208–219. https://doi.org/10.1101/gr.123109.111.

241. Liu, J., Lee, W., Jiang, Z., Chen, Z., Jhunjhunwala, S., Haverty, P.M., Gnad, F., Guan, Y., Gilbert, H.N., Stinson, J., et al. (2012). Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. Genome Res. *22*, 2315–2327. https://doi.org/10.1101/gr.140988.112.

242. Seo, J.S., Ju, Y.S., Lee, W.C., Shin, J.Y., Lee, J.K., Bleazard, T., Lee, J., Jung, Y.J., Kim, J.O., Shin, J.Y., et al. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. *22*, 2109–2119. https://doi.org/10.1101/gr.145144.112.

243. Jia, P., Jin, H., Meador, C.B., Xia, J., Ohashi, K., Liu, L., Pirazzoli, V., Dahlman, K.B., Politi, K., Michor, F., et al. (2013). Next-generation sequencing of paired tyrosine kinase inhibitor-sensitive and -resistant EGFR mutant lung cancer cell lines identifies spectrum of DNA changes associated with drug resistance. Genome Res. *23*, 1434–1445. https://doi.org/10.1101/gr.152322.112.

244. Totoki, Y., Yoshida, A., Hosoda, F., Nakamura, H., Hama, N., Ogura, K., Yoshida, A., Fujiwara, T., Arai, Y., Toguchida, J., et al. (2014). Unique mutation portraits and frequent COL2A1 gene alteration in chondrosarcoma. Genome Res. *24*, 1411–1420. https://doi.org/10.1101/gr.160598.113.

245. Andersson, E., Eldfors, S., Edgren, H., Ellonen, P., Väkevä, L., Ranki, A., and Mustjoki, S. (2014). Novel TBL1XR1, EPHA7 and SLFN12 mutations in a Sezary syndrome patient discovered by whole exome sequencing. Exp. Dermatol. *23*, 366–368. https://doi.org/10.1111/exd.12405.

246. Furney, S.J., Turajlic, S., Fenwick, K., Lambros, M.B., MacKay, A., Ricken, G., Mitsopoulos, C., Kozarewa, I., Hakas, J., Zvelebil, M., et al. (2012). Genomic characterisation of acral melanoma cell lines. Pigment Cell Melanoma Res. *25*, 488–492. https://doi.org/10.1111/j.1755-148X.2012.01016.x.

247. Jones, S., Wang, T.L., Shih, I.M., Mao, T.L., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L.A., Jr., Vogelstein, B., et al. (2010). Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science *330*, 228–231. https://doi.org/10.1126/science.1196333.

248. Choi, M., Scholl, U.I., Yue, P., Björklund, P., Zhao, B., Nelson-Williams, C., Ji, W., Cho, Y., Patel, A., Men, C.J., et al. (2011). K+ channel mutations in adrenal aldosterone-producing adenomas and hereditary hypertension. Science *331*, 768–772. https://doi.org/10.1126/science.1198785.

249. Jiao, Y., Shi, C., Edil, B.H., de Wilde, R.F., Klimstra, D.S., Maitra, A., Schulick, R.D., Tang, L.H., Wolfgang, C.L., Choti, M.A., et al. (2011). DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. Science *331*, 1199–1203. https://doi.org/10.1126/science.1200609.

250. Agrawal, N., Frederick, M.J., Pickering, C.R., Bettegowda, C., Chang, K., Li, R.J., Fakhry, C., Xie, T.-X., Zhang, J., Wang, J., et al. (2011). Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science *333*, 1154–1157. https://doi.org/10.1126/science.1206923.

251. Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A., et al. (2011). The mutational landscape of head and neck squamous cell carcinoma. Science *333*, 1157–1160. https://doi.org/10.1126/science.1208130.

252. Bettegowda, C., Agrawal, N., Jiao, Y., Sausen, M., Wood, L.D., Hruban, R.H., Rodriguez, F.J., Cahill, D.P., McLendon, R., Riggins, G., et al. (2011). Mutations in CIC and FUBP1 contribute to human oligodendroglioma. Science *333*, 1453–1455. https://doi.org/10.1126/science.1210557.

253. Iyer, G., Hanrahan, A.J., Milowsky, M.I., Al-Ahmadie, H., Scott, S.N., Janakiraman, M., Pirun, M., Sander, C., Socci, N.D., Ostrovnaya, I., et al. (2012). Genome sequencing identifies a basis for everolimus sensitivity. Science *338*, 221. https://doi.org/10.1126/science.1226344.

254. Clark, V.E., Erson-Omay, E.Z., Serin, A., Yin, J., Cotney, J., Özduman, K., Avşar, T., Li, J., Murray, P.B., Henegariu, O., et al. (2013). Genomic analysis of non-NF2 meningiomas reveals mutations in TRAF7, KLF4, AKT1, and SMO. Science *339*, 1077–1080. https://doi.org/10.1126/science.1233009.

255. Sato, Y., Maekawa, S., Ishii, R., Sanada, M., Morikawa, T., Shiraishi, Y., Yoshida, K., Nagata, Y., Sato-Otsubo, A., Yoshizato, T., et al. (2014). Recurrent somatic mutations underlie corticotropin-independent

Cushing's syndrome. Science *344*, 917–920. https://doi.org/10.1126/science.1252328.

256. Yu, J., Wu, W.K.K., Li, X., He, J., Li, X.X., Ng, S.S.M., Yu, C., Gao, Z., Yang, J., Li, M., et al. (2015). Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. Gut *64*, 636–645. https://doi.org/10.1136/gutjnl-2013-306620.

257. Nichols, A.C., Chan-Seng-Yue, M., Yoo, J., Agrawal, S.K., Starmans, M.H.W., Waggott, D., Harding, N.J., Dowthwaite, S.A., Palma, D.A., Fung, K., et al. (2013). A case report and genetic characterization of a massive acinic cell carcinoma of the parotid with delayed distant metastases. Case Rep. Oncol. Med., 270362. https://doi.org/10.1155/2013/270362.

258. Abaan, O.D., Polley, E.C., Davis, S.R., Zhu, Y.J., Bilke, S., Walker, R.L., Pineda, M., Gindin, Y., Jiang, Y., Reinhold, W.C., et al. (2013). The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. Cancer Res. *73*, 4372–4382. https://doi.org/10.1158/0008-5472.CAN-12-3342.

259. Nikolaev, S.I., Sotiriou, S.K., Pateras, I.S., Santoni, F., Sougioultzis, S., Edgren, H., Almusa, H., Robyr, D., Guipponi, M., Saarela, J., et al. (2012). A single-nucleotide substitution mutator phenotype revealed by exome sequencing of human colon adenomas. Cancer Res. *72*, 6279–6289. https://doi.org/10.1158/0008-5472.CAN-12-3869.

260. Seki, M., Yoshida, K., Shiraishi, Y., Shimamura, T., Sato, Y., Nishimura, R., Okuno, Y., Chiba, K., Tanaka, H., Kato, K., et al. (2014). Biallelic DICER1 mutations in sporadic pleuropulmonary blastoma. Cancer Res. *74*, 2742–2749. https://doi.org/10.1158/0008-5472.CAN-13-2470.

261. Guo, G., Chmielecki, J., Goparaju, C., Heguy, A., Dolgalev, I., Carbone, M., Seepo, S., Meyerson, M., and Pass, H.I. (2015). Whole-exome sequencing reveals frequent genetic alterations in BAP1, NF2, CDKN2A, and CUL1 in malignant pleural mesothelioma. Cancer Res. *75*, 264–269. https://doi.org/10.1158/0008-5472.CAN-14-1008.

262. Pickering, C.R., Zhou, J.H., Lee, J.J., Drummond, J.A., Peng, S.A., Saade, R.E., Tsai, K.Y., Curry, J.L., Tetzlaff, M.T., Lai, S.Y., et al. (2014). Mutational landscape of aggressive cutaneous squamous cell carcinoma. Clin. Cancer Res. *20*, 6582–6592. https://doi.org/10.1158/1078-0432.CCR-14-1768.

263. Durinck, S., Ho, C., Wang, N.J., Liao, W., Jakkula, L.R., Collisson, E.A., Pons, J., Chan, S.W., Lam, E.T., Chu, C., et al. (2011). Temporal dissection of tumorigenesis in primary cancers. Cancer Discov. *1*, 137–143. https://doi.org/10.1158/2159-8290.CD-11-0028.

264. Koo, G.C., Tan, S.Y., Tang, T., Poon, S.L., Allen, G.E., Tan, L., Chong, S.C., Ong, W.S., Tay, K., Tao, M., et al. (2012). Janus kinase 3-activating mutations identified in natural killer/T-cell lymphoma. Cancer Discov. *2*, 591–597. https://doi.org/10.1158/2159-8290.CD-12-0028.

265. Dahlman, K.B., Xia, J., Hutchinson, K., Ng, C., Hucks, D., Jia, P., Atefi, M., Su, Z., Branch, S., Lyle, P.L., et al. (2012). BRAF(L597) mutations in melanoma are associated with sensitivity to MEK inhibitors. Cancer Discov. *2*, 791–797. https://doi.org/10.1158/2159-8290.CD-12-0097.

266. Agrawal, N., Jiao, Y., Bettegowda, C., Hutfless, S.M., Wang, Y., David, S., Cheng, Y., Twaddell, W.S., Latt, N.L., Shin, E.J., et al. (2012). Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. Cancer Discov. *2*, 899–905. https://doi.org/10.1158/2159-8290.CD-12-0189.

267. Lui, V.W.Y., Hedberg, M.L., Li, H., Vangara, B.S., Pendleton, K., Zeng, Y., Lu, Y., Zhang, Q., Du, Y., Gilbert, B.R., et al. (2013). Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. Cancer Discov. *3*, 761–769. https://doi.org/10.1158/2159-8290.CD-13-0103.

268. Furney, S.J., Pedersen, M., Gentien, D., Dumont, A.G., Rapinat, A., Desjardins, L., Turajlic, S., Piperno-Neumann, S., de la Grange, P., Roman-Roman, S., et al. (2013). SF3B1 mutations are associated with alternative splicing in uveal melanoma. Cancer Discov. *3*, 1122–1129. https://doi.org/10.1158/2159-8290.CD-13-0330.

269. Wagle, N., Van Allen, E.M., Treacy, D.J., Frederick, D.T., Cooper, Z.A., Taylor-Weiner, A., Rosenberg, M., Goetz, E.M., Sullivan, R.J., Farlow, D.N., et al. (2014). MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. Cancer Discov. *4*, 61–68. https://doi.org/10.1158/2159-8290.CD-13-0631.

270. Lee, R.S., Stewart, C., Carter, S.L., Ambrogio, L., Cibulskis, K., Sougnez, C., Lawrence, M.S., Auclair, D., Mora, J., Golub, T.R., et al. (2012). A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. J. Clin. Invest. *122*, 2983–2988. https://doi.org/10.1172/JCI64400.

271. Banck, M.S., Kanwar, R., Kulkarni, A.A., Boora, G.K., Metge, F., Kipp, B.R., Zhang, L., Thorland, E.C., Minn, K.T., Tentu, R., et al. (2013). The genomic landscape of small intestine neuroendocrine tumors. J. Clin. Invest. *123*, 2502–2508. https://doi.org/10.1172/JCI67963.

272. Greif, P.A., Dufour, A., Konstandin, N.P., Ksienzyk, B., Zellmeier, E., Tizazu, B., Sturm, J., Benthaus, T., Herold, T., Yaghmaie, M., et al. (2012). GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. Blood *120*, 395–403. https://doi.org/10.1182/blood-2012-01-403220.

273. Walker, B.A., Wardell, C.P., Melchor, L., Hulkki, S., Potter, N.E., Johnson, D.C., Fenwick, K., Kozarewa, I., Gonzalez, D., Lord, C.J., et al. (2012). Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4;14) and t(11;14) myeloma. Blood *120*, 1077–1086. https://doi.org/10.1182/blood-2012-03-412981.

274. Green, M.R., Gentles, A.J., Nair, R.V., Irish, J.M., Kihira, S., Liu, C.L., Kela, I., Hopmans, E.S., Myklebust, J.H., Ji, H., et al. (2013). Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. Blood *121*, 1604–1611. https://doi.org/10.1182/blood-2012-09-457283.

275. Neumann, M., Heesch, S., Schlee, C., Schwartz, S., Gökbuget, N., Hoelzer, D., Konstandin, N.P., Ksienzyk, B., Vosberg, S., Graf, A., et al. (2013). Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. Blood *121*, 4749–4752. https://doi.org/10.1182/blood-2012-11-465138.

276. Messina, M., Del Giudice, I., Khiabanian, H., Rossi, D., Chiaretti, S., Rasi, S., Spina, V., Holmes, A.B., Marinelli, M., Fabbri, G., et al. (2014). Genetic lesions associated with chronic lymphocytic leukemia chemo-refractoriness. Blood *123*, 2378–2388. https://doi.org/10.1182/blood-2013-10-534271.

277. Chakraborty, R., Hampton, O.A., Shen, X., Simko, S.J., Shih, A., Abhyankar, H., Lim, K.P.H., Covington, K.R., Trevino, L., Dewal, N., et al. (2014). Mutually exclusive recurrent somatic mutations in MAP2K1 and BRAF support a central role for ERK activation in LCH pathogenesis. Blood *124*, 3007–3015. https://doi.org/10.1182/blood-2014-05-577825.

278. Gröschel, S., Sanders, M.A., Hoogenboezem, R., Zeilemaker, A., Havermans, M., Erpelinck, C., Bindels, E.M.J., Beverloo, H.B., Döhner, H., Löwenberg, B., et al. (2015). Mutational spectrum of myeloid malignancies with inv(3)/t(3;3) reveals a predominant involvement of RAS/RTK signaling pathways. Blood *125*, 133–139. https://doi.org/10.1182/blood-2014-07-591461.

279. Reichel, J., Chadburn, A., Rubinstein, P.G., Giulino-Roth, L., Tam, W., Liu, Y., Gaiolla, R., Eng, K., Brody, J., Inghirami, G., et al. (2015). Flow sorting and exome sequencing reveal the oncogenome of primary Hodgkin and Reed-Sternberg cells. Blood *125*, 1061–1072. https://doi.org/10.1182/blood-2014-11-610436.

280. Nannini, M., Astolfi, A., Urbini, M., Indio, V., Santini, D., Heinrich, M.C., Corless, C.L., Ceccarelli, C., Saponara, M., Mandrioli, A., et al. (2014). Integrated genomic study of quadruple-WT GIST (KIT/PDGFRA/SDH/RAS pathway wild-type GIST). BMC Cancer *14*, 685. https://doi.org/10.1186/1471-2407-14-685.

281. Shanmugam, V., Ramanathan, R.K., Lavender, N.A., Sinari, S., Chadha, M., Liang, W.S., Kurdoglu, A., Izatt, T., Christoforides, A., Benson, H., et al. (2014). Whole genome sequencing reveals potential targets for

therapy in patients with refractory KRASmutated metastatic colorectal cancer. BMC Med. Genomics *7*, 36. https://doi.org/10.1186/1755-8794-7-36.

282. Lee, Y.-S., Cho, Y.S., Lee, G.K., Lee, S., Kim, Y.-W., Jho, S., Kim, H.-M., Hong, S.-H., Hwang, J.-A., Kim, S.-y., et al. (2014). Genomic profile analysis of diffuse-type gastric cancers. Genome Biol. *15*, R55. https://doi.org/10.1186/gb-2014-15-4-r55.

283. Demeure, M.J., Craig, D.W., Sinari, S., Moses, T.M., Christoforides, A., Dinh, J., Izatt, T., Aldrich, J., Decker, A., Baker, A., et al. (2012). Cancer of the ampulla of Vater: analysis of the whole genome sequence exposes a potential therapeutic vulnerability. Genome Med. *4*, 56. https://doi.org/10.1186/gm357.

284. Alakus, H., Babicky, M.L., Ghosh, P., Yost, S., Jepsen, K., Dai, Y., Arias, A., Samuels, M.L., Mose, E.S., Schwab, R.B., et al. (2014). Genome-wide mutational landscape of mucinous carcinomatosis peritonei of appendiceal origin. Genome Med. *6*, 43. https://doi.org/10.1186/gm559.

285. Begg, C.B., Ostrovnaya, I., Carniello, J.V.S., Sakr, R.A., Giri, D., Towers, R., Schizas, M., De Brot, M., Andrade, V.P., Mauguen, A., et al. (2016). Clonal relationships between lobular carcinoma in situ and other breast malignancies. Breast Cancer Res. *18*, 66. https://doi.org/10.1186/s13058-016-0727-z.

286. Fisher, R., Horswell, S., Rowan, A., Salm, M.P., de Bruin, E.C., Gulati, S., McGranahan, N., Stares, M., Gerlinger, M., Varela, I., et al. (2014). Development of synchronous VHL syndrome tumors reveals contingencies and constraints to tumor evolution. Genome Biol. *15*, 433. https://doi.org/10.1186/s13059-014-0433-z.

287. Jhunjhunwala, S., Jiang, Z., Stawiski, E.W., Gnad, F., Liu, J., Mayba, O., Du, P., Diao, J., Johnson, S., Wong, K.-F., et al. (2014). Diverse modes of genomic alteration in hepatocellular carcinoma. Genome Biol. *15*, 436. https://doi.org/10.1186/s13059-014-0436-9.

288. Reimann, E., Kõks, S., Ho, X.D., Maasalu, K., and Märtson, A. (2014). Whole exome sequencing of a single osteosarcoma case—integrative analysis with whole transcriptome RNA-seq data. Hum. Genomics *8*, 20. https://doi.org/10.1186/s40246-014-0020-0.

289. Shankar, G.M., Taylor-Weiner, A., Lelic, N., Jones, R.T., Kim, J.C., Francis, J.M., Abedalthagafi, M., Borges, L.F., Coumans, J.-V., Curry, W.T., et al. (2014). Sporadic hemangioblastomas are characterized by cryptic VHL inactivation. Acta Neuropathol. Commun. *2*, 167. https://doi.org/10.1186/s40478-014-0167-x.

290. Newey, P.J., Nesbit, M.A., Rimmer, A.J., Attar, M., Head, R.T., Christie, P.T., Gorvin, C.M., Stechman, M., Gregory, L., Mihai, R., et al. (2012). Whole-exome sequencing studies of nonhereditary (sporadic) parathyroid adenomas. J. Clin. Endocrinol. Metab. *97*, E1995–E2005. https://doi.org/10.1210/jc.2012-2303.

291. Agrawal, N., Jiao, Y., Sausen, M., Leary, R., Bettegowda, C., Roberts, N.J., Bhan, S., Ho, A.S., Khan, Z., Bishop, J., et al. (2013). Exomic sequencing of medullary thyroid cancer reveals dominant and mutually exclusive oncogenic mutations in RET and RAS. J. Clin. Endocrinol. Metab. *98*, E364–E369. https://doi.org/10.1210/jc.2012-2703.

292. Crona, J., Delgado Verdugo, A., Maharjan, R., Stålberg, P., Granberg, D., Hellman, P., and Björklund, P. (2013). Somatic mutations in H-RAS in sporadic pheochromocytoma and paraganglioma identified by exome sequencing. J. Clin. Endocrinol. Metab. *98*, E1266–E1271. https://doi.org/10.1210/jc.2012-4257.

293. Kasaian, K., Chindris, A.M., Wiseman, S.M., Mungall, K.L., Zeng, T., Tse, K., Schein, J.E., Rivera, M., Necela, B.M., Kachergus, J.M., et al. (2015). MEN1 mutations in Hurthle cell (oncocytic) thyroid carcinoma. J. Clin. Endocrinol. Metab. *100*, E611–E615. https://doi.org/10.1210/jc.2014-3622.

294. Borad, M.J., Champion, M.D., Egan, J.B., Liang, W.S., Fonseca, R., Bryce, A.H., McCullough, A.E., Barrett, M.T., Hunt, K., Patel, M.D., et al. (2014). Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma. PLoS Genet. *10*, e1004135. https://doi.org/10.1371/journal.pgen.1004135.

295. Brohl, A.S., Solomon, D.A., Chang, W., Wang, J., Song, Y., Sindiri, S., Patidar, R., Hurd, L., Chen, L., Shern, J.F., et al. (2014). The genomic landscape of the Ewing Sarcoma family of tumors reveals recurrent STAG2 mutation. PLoS Genet. *10*, e1004475. https://doi.org/10.1371/journal.pgen.1004475.

296. Barber, L.J., Rosa Rosa, J.M., Kozarewa, I., Fenwick, K., Assiotis, I., Mitsopoulos, C., Sims, D., Hakas, J., Zvelebil, M., Lord, C.J., and Ashworth, A. (2011). Comprehensive genomic analysis of a BRCA2 deficient human pancreatic cancer. PLoS One *6*, e21639. https://doi.org/10.1371/journal.pone.0021639.

297. Zhou, D., Yang, L., Zheng, L., Ge, W., Li, D., Zhang, Y., Hu, X., Gao, Z., Xu, J., Huang, Y., et al. (2013). Exome capture sequencing of adenoma reveals genetic alterations in multiple cellular pathways at the early stage of colorectal tumorigenesis. PLoS One *8*, e53310. https://doi.org/10.1371/journal.pone.0053310.

298. Kim, S.C., Jung, Y., Park, J., Cho, S., Seo, C., Kim, J., Kim, P., Park, J., Seo, J., Kim, J., et al. (2013). A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers. PLoS One *8*, e55596. https://doi.org/10.1371/journal.pone.0055596.

299. Yost, S.E., Pastorino, S., Rozenzhak, S., Smith, E.N., Chao, Y.S., Jiang, P., Kesari, S., Frazer, K.A., and Harismendy, O. (2013). High-resolution mutational profiling suggests the genetic validity of glioblastoma patient-derived pre-clinical models. PLoS One *8*, e56185. https://doi.org/10.1371/journal.pone.0056185.

300. Hong, J.Y., Liu, X., Mao, M., Li, M., Choi, D.I., Kang, S.W., Lee, J., and La Choi, Y. (2013). Genetic aberrations in imatinib-resistant dermatofibrosarcoma protuberans revealed by whole genome sequencing. PLoS One *8*, e69752. https://doi.org/10.1371/journal.pone.0069752.

301. Parry, M., Rose-Zerilli, M.J.J., Gibson, J., Ennis, S., Walewska, R., Forster, J., Parker, H., Davis, Z., Gardiner, A., Collins, A., et al. (2013). Whole exome sequencing identifies novel recurrently mutated genes in patients with splenic marginal zone lymphoma. PLoS One *8*, e83244. https://doi.org/10.1371/journal.pone.0083244.

302. Lee, S.Y., Haq, F., Kim, D., Jun, C., Jo, H.J., Ahn, S.M., and Lee, W.S. (2014). Comparative genomic analysis of primary and synchronous metastatic colorectal cancers. PLoS One *9*, e90459. https://doi.org/10.1371/journal.pone.0090459.

303. Chen, J., Raju, G.S., Jogunoori, W., Menon, V., Majumdar, A., Chen, J.S., Gi, Y.J., Jeong, Y.S., Phan, L., Belkin, M., et al. (2016). Mutational profiles reveal an aberrant TGF-beta-CEA regulated pathway in colon adenomas. PLoS One *11*, e0153933. https://doi.org/10.1371/journal.pone.0153933.

304. Killela, P.J., Pirozzi, C.J., Reitman, Z.J., Jones, S., Rasheed, B.A., Lipp, E., Friedman, H., Friedman, A.H., He, Y., McLendon, R.E., et al. (2014). The genetic landscape of anaplastic astrocytoma. Oncotarget *5*, 1452–1457. https://doi.org/10.18632/oncotarget.1505.

305. Bruno, A., Boisselier, B., Labreche, K., Marie, Y., Polivka, M., Jouvet, A., Adam, C., Figarella-Branger, D., Miquel, C., Eimer, S., et al. (2014). Mutational analysis of primary central nervous system lymphoma. Oncotarget *5*, 5065–5075. https://doi.org/10.18632/oncotarget.2080.

306. Martin, D., Abba, M.C., Molinolo, A.A., Vitale-Cross, L., Wang, Z., Zaida, M., Delic, N.C., Samuels, Y., Lyons, J.G., and Gutkind, J.S. (2014). The head and neck cancer cell oncogenome: a platform for the development of precision molecular therapies. Oncotarget *5*, 8906–8923. https://doi.org/10.18632/oncotarget.2417.

307. Kannan, K., Inagaki, A., Silber, J., Gorovets, D., Zhang, J., Kastenhuber, E.R., Heguy, A., Petrini, J.H., Chan, T.A., and Huse, J.T. (2012). Whole-exome sequencing identifies ATRX mutation as a key molecular determinant in lower-grade glioma. Oncotarget *3*, 1194–1203. https://doi.org/10.18632/oncotarget.689.

308. Bettegowda, C., Agrawal, N., Jiao, Y., Wang, Y., Wood, L.D., Rodriguez, F.J., Hruban, R.H., Gallia, G.L., Binder, Z.A., Riggins, C.J., et al. (2013). Exomic sequencing of four rare central nervous system tumor types. Oncotarget *4*, 572–583. https://doi.org/10.18632/oncotarget.964.

309. Zheng, C.X., Gu, Z.H., Han, B., Zhang, R.X., Pan, C.M., Xiang, Y., Rong, X.J., Chen, X., Li, Q.Y., and Wan, H.Y. (2013). Whole-exome sequencing to identify novel somatic mutations in squamous cell lung cancers. Int. J. Oncol. *43*, 755–764. https://doi.org/10.3892/ijo.2013.1991.

310. Huang, Y., Zheng, J., Hu, J.D., Wu, Y.A., Zheng, X.Y., Liu, T.B., and Chen, F.L. (2014). Discovery of somatic mutations in the progression of chronic myeloid leukemia by whole-exome sequencing. Genet. Mol. Res. *13*, 945–953. https://doi.org/10.4238/2014.February.19.5.

311. Nichols, A.C., Chan-Seng-Yue, M., Yoo, J., Xu, W., Dhaliwal, S., Basmaji, J., Szeto, C.C.T., Dowthwaite, S., Todorovic, B., Starmans, M.H.W., et al. (2012). A pilot study comparing HPV-positive and HPV-negative head and neck squamous cell carcinomas by whole exome sequencing. ISRN Oncol., 809370. https://doi.org/10.5402/2012/809370.

312. Serra, V., Vivancos, A., Puente, X.S., Felip, E., Silberschmidt, D., Caratù, G., Parra, J.L., De Mattos-Arruda, L., Grueso, J., Hernández-Losa, J., et al. (2013). Clinical response to a lapatinib-based therapy for a Li-Fraumeni syndrome patient with a novel HER2V659E mutation. Cancer Discov. *3*, 1238–1244. https://doi.org/10.1158/2159-8290.CD-13-0132.

313. Xie, T., Cho, Y.B., Wang, K., Huang, D., Hong, H.K., Choi, Y.L., Ko, Y.H., Nam, D.H., Jin, J., Yang, H., et al. (2014). Patterns of somatic alterations between matched primary and metastatic colorectal tumors characterized by whole-genome sequencing. Genomics *104*, 234–241. https://doi.org/10.1016/j.ygeno.2014.07.012.

314. Tran, E., Turcotte, S., Gros, A., Robbins, P.F., Lu, Y.C., Dudley, M.E., Wunderlich, J.R., Somerville, R.P., Hogan, K., Hinrichs, C.S., et al. (2014). Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science *344*, 641–645. https://doi.org/10.1126/science.1251102.

315. Mouradov, D., Sloggett, C., Jorissen, R.N., Love, C.G., Li, S., Burgess, A.W., Arango, D., Strausberg, R.L., Buchanan, D., Wormald, S., et al. (2014). Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. Cancer Res. *74*, 3238–3247. https://doi.org/10.1158/0008-5472.CAN-14-0013.

316. Crompton, B.D., Stewart, C., Taylor-Weiner, A., Alexe, G., Kurek, K.C., Calicchio, M.L., Kiezun, A., Carter, S.L., Shukla, S.A., Mehta, S.S., et al. (2014). The genomic landscape of pediatric Ewing sarcoma. Cancer Discov. *4*, 1326–1341. https://doi.org/10.1158/2159-8290.CD-13-1037.

317. Nelson, D.S., Quispel, W., Badalian-Very, G., van Halteren, A.G.S., van den Bos, C., Bovée, J.V.M.G., Tian, S.Y., Van Hummelen, P., Ducar, M., MacConaill, L.E., et al. (2014). Somatic activating ARAF mutations in Langerhans cell histiocytosis. Blood *123*, 3152–3155. https://doi.org/10.1182/blood-2013-06-511139.

318. Herold, T., Metzeler, K.H., Vosberg, S., Hartmann, L., Röllig, C., Stölzel, F., Schneider, S., Hubmann, M., Zellmeier, E., Ksienzyk, B., et al. (2014). Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. Blood *124*, 1304–1311. https://doi.org/10.1182/blood-2013-12-540716.

319. Gambacorti-Passerini, C.B., Donadoni, C., Parmiani, A., Pirola, A., Redaelli, S., Signore, G., Piazza, V., Malcovati, L., Fontana, D., Spinelli, R., et al. (2015). Recurrent ETNK1 mutations in atypical chronic myeloid leukemia. Blood *125*, 499–503. https://doi.org/10.1182/blood-2014-06-579466.

320. Yu, W., McPherson, J.R., Stevenson, M., van Eijk, R., Heng, H.L., Newey, P., Gan, A., Ruano, D., Huang, D., Poon, S.L., et al. (2015). Whole-exome sequencing studies of parathyroid carcinomas reveal novel PRUNE2 mutations, distinctive mutational spectra related to APOBEC-catalyzed DNA mutagenesis and mutational enrichment in kinases associated with cell migration and invasion. J. Clin. Endocrinol. Metab. *100*, E360–E364. https://doi.org/10.1210/jc.2014-3238.

321. Mullighan, C.G., Su, X., Zhang, J., Radtke, I., Phillips, L.A.A., Miller, C.B., Ma, J., Liu, W., Cheng, C., Schulman, B.A., et al. (2009). Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N. Engl. J. Med. *360*, 470–480. https://doi.org/10.1056/NEJMoa0808253.

322. Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J.S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., et al. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N. Engl. J. Med. *365*, 1384–1395. https://doi.org/10.1056/NEJMoa1103283.

323. Shalabi, L.A., and Shaaban, Z. (2006). Normalization as a preprocessing engine for data mining and the approach of preference matrix. IEEE *25-27*, 207–214.

324. Žitnik, M., and Zupan, B. (2012). Nimfa: a python library for nonnegative matrix factorization. The Journal of Machine Learning Research *13*, 849–853.

325. Aranganayagi, S., and Thangavel, K. (2007). Clustering categorical data using silhouette coefficient as a relocating measure. IEEE, 13–17. https://doi.org/10.1109/ICCIMA.2007.328.

326. Franc, V., Hlaváč, V., and Navara, M. (2005). In Sequential Coordinate-Wise Algorithm for the Non-negative Least Squares Problem, A. Gagalowicz and W. Philips, eds. (Springer Berlin Heidelberg), pp. 407–414.

327. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
PCAWG and extended dataset	263 published studies and 35 ICGC projects[6,43,49,51,59–322]	Table S8
Normal urothelium dataset	Lawson et al., 2020[50]	EGA: EGAD00001006113 and EGAD00001006116
Bladder urothelial carcinoma dataset	The Cancer Genome Atlas Research Network, 2014[56]	https://gdc.cancer.gov/
Synthetically generated dataset	This paper	Figshare: https://doi.org/10.6084/m9.figshare.20409430
Software and algorithms		
EMu 1.5.2	Fischer et al., 2013[20]	https://github.com/andrej-fischer/EMu
Maftools 2.2.0	Mayakonda et al., 2018[22]	https://bioconductor.org/packages/release/bioc/html/maftools.html
MutationalPatterns 3.0.1	Blokzijl et al., 2018[24]	https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html
MutSignatures 2.1.1	Fantini et al., 2020[25]	https://CRAN.R-project.org/package=mutSignatures
MutSpec 2.0	Ardin et al., 2016[27]	https://github.com/IARCbioinfo/mutspec
SigFit 2.0.0	Gori et al., 2020[28]	https://github.com/kgori/sigfit
SigMiner 1.0.0	Wang et al., 2020[31]	https://github.com/ShixiangWang/sigminer
SignatureAnalyzer	Kasar et al., 2015[32]; Taylor-Weiner et al., 2019[34]	https://github.com/broadinstitute/SignatureAnalyzer-GPU
SignatureToolsLib 0.0.0.9000	Degasperi et al., 2020[35]	https://github.com/Nik-Zainal-Group/signature.tools.lib
SigneR 1.16.0	Rosales et al., 2016[36]	http://bioconductor.org/packages/release/bioc/html/signeR.html
SigProfiler_PCAWG (SigProExtractor) 0.0.5.48	Alexandrov et al., 2020[12]	https://pypi.org/project/sigproextractor/0.0.5.48/
SigProfilerExtractor 1.1.4	This paper	https://doi.org/10.5281/zenodo.6746540
SigProfilerExtractorR 1.1.0	This paper	https://doi.org/10.5281/zenodo.6941779
SigProfilerMatrixGenerator 1.2.4	Bergstrom et al., 2019[15]	https://github.com/AlexandrovLab/SigProfilerMatrixGenerator
SigProfilerSimulator 1.1.3	Bergstrom et al., 2020[48]	https://github.com/AlexandrovLab/SigProfilerSimulator
SomaticSignatures 2.26.0	Gehring et al., 2015[38]	https://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html
SynSigGen	Alexandrov et al., 2020[12]	https://github.com/steverozen/SynSigGen
TensorSignatures 0.5.0	Vöhringer et al., 2021[40]	https://github.com/sagar87/tensorsignatures
Other		
Results from the benchmarking with synthetic datasets, including the appropriate input used to run each of the tools as well as the generated output	This paper	https://doi.org/10.6084/m9.figshare.20409430
Results from the benchmarking of the different options available in SigProfilerExtractor for matrix normalization, NMF initialization, and NMF objective function	This paper	https://doi.org/10.6084/m9.figshare.20411483

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Results from the *de novo* extraction of mutational signatures from the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset	This paper	https://doi.org/10.6084/m9.figshare.20406279
Results from the *de novo* extraction of mutational signatures from the extended dataset	This paper	https://doi.org/10.6084/m9.figshare.20406326
Summarized collection of all input mutational matrices, as well as *de novo* extracted mutational signatures and activities for both PCAWG and extended datasets	This paper	https://doi.org/10.6084/m9.figshare.20293890
Results from the *de novo* extraction of mutational signatures for confirming the patterns of the novel signatures for additional datasets	This paper	https://doi.org/10.6084/m9.figshare.20406156
Results from the *de novo* extraction of mutational signatures from downsampling of whole-genome sequenced samples to whole-exomes	This paper	https://doi.org/10.6084/m9.figshare.20406276
Resource website for the COSMIC reference set of mutational signatures	Tate et al., 2019[42]	https://cancer.sanger.ac.uk/signatures/

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Ludmil B. Alexandrov (l2alexandrov@health.ucsd.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Our study analyzes synthetically generated data, as well as publicly available data from human subjects. The accessions numbers for the human datasets are listed in the key resources table and Table S8, and correspond to a total of 263 published studies as well as 35 ICGC projects.[6,43,49,51,59–322]

All results from the benchmarking with synthetic datasets, including the appropriate input used to run each of the tools as well as the output generated by each of the tools, can be found at: ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/Benchmark/ and Figshare: https://doi.org/10.6084/m9.figshare.20409430.

All results from the benchmarking of the different options available in SigProfilerExtractor for matrix normalization, NMF initialization, and NMF objective function can be found at: ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/Benchmark_Initialization_Normalization_Objective-Function/ and Figshare: https://doi.org/10.6084/m9.figshare.20411483.

All results from the *de novo* extraction of mutational signatures from the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset can be found at: ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/PCAWG_Reanalysis/ and Figshare: https://doi.org/10.6084/m9.figshare.20406279.

All results from the *de novo* extraction of mutational signatures from the extended dataset can be found at: ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/Extended_Cohort_Reanalysis/ and Figshare: https://doi.org/10.6084/m9.figshare.20406326.

A summarized collection of all input mutational matrices, as well as *de novo* extracted mutational signatures and activities for both PCAWG and extended datasets has also been deposited at Figshare: https://doi.org/10.6084/m9.figshare.20293890.

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/ and Figshare: https://doi.org/10.6084/m9.figshare.20406156.

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/ and Figshare: https://doi.org/10.6084/m9.figshare.20406276.

All original code has been deposited at GitHub (https://github.com/AlexandrovLab/SigProfilerExtractor and https://github.com/AlexandrovLab/SigProfilerExtractorR), PyPI (https://pypi.org/project/SigProfilerExtractor/) and Zenodo (https://doi.org/10.5281/zenodo.6746540 and https://doi.org/10.5281/zenodo.6941779). 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. 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 results is provided at https://osf.io/t6j7u/wiki/home/. SigProfilerExtractor is compatible with Windows, Linux, Unix, and macOS operating systems.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

No experimental models were utilized as part of this publication. No novel subjects were collected as part of this publication.

METHOD DETAILS

Computational implementation of SigProfilerExtractor and its seven modules

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

Module 1: Processing of input mutational catalogs or input mutational matrices

SigProfilerExtractor deciphers mutational signatures from a mutational matrix $\boldsymbol{M}$ with t rows and n columns; rows represent mutational channels while columns reflect individual cancer samples (Figure 1A). The value of each cell in the matrix, $\boldsymbol{M}$, corresponds to the number of somatic mutations from a particular mutational channel in each sample. The mutational matrix can be provided as a text file with the first column containing the names of the mutational channels and the first row containing the names of the examined samples, thus supporting nonnegative matrix factorization for any custom matrix dataset. Alternatively, users can provide a mutational catalog of somatic mutations in a commonly used format (*e.g.,* VCF, MAF, *etc.*) and this mutational catalog will be internally converted into the appropriate mutational matrix by SigProfilerMatrixGenerator.[15]

Module 2: Resampling of the input mutational matrix and normalizing the resampled matrix

SigProfilerExtractor does not factorize the original input matrix. Rather, prior to performing matrix factorization, SigProfilerExtractor performs independent Poisson resampling of the original matrix for each replicate.[4] As such, the matrix factorized in each replicate is never the same for a given value of k (Figure 1B). The resampling is performed to ensure that Poisson fluctuations of the matrix do not impact the stability of the factorization results. Additional normalization is performed after resampling to overcome potential skewing of the factorization from any hypermutators. SigProfilerExtractor supports four standard normalization methods[323]: *(i)* Gaussian mixture model (GMM) normalization (default); *(ii)* 100X normalization; *(iii)* log2 normalization; *(iv)* no normalization. *No normalization* does not perform any additional transformation on the Poisson resampled matrix. In *log2 normalization*, the sum of each column in the matrix is derived and logarithm with base 2 is calculated for each of these sums. Each cell in a column of the matrix is multiplied by the log2 of the column-sum and subsequently divided by the original column sum. In *100X normalization*, the sum of each column in the matrix is derived. For each column where the sum exceeds 100 times the number of mutational channels (*i.e.*, 100 times the number of rows in the matrix), each cell in the column is multiplied by the 100 times the number of mutational channels and subsequently divided by the original column sum. This normalization ensures that no sample has a total number of mutations above 100 times the number of mutational channels. *GMM normalization* encompasses a two-step process. The first step derives the normalization cutoff value in a data-driven manner using a Gaussian mixture model (GMM). The second step normalizes the appropriate columns using the derived cutoff value. The first step uses a GMM to separate the samples into two groups based on their total number of mutations; the total number of mutations in a sample reflects the sum of a column in the matrix. The group with larger number of samples is subsequently selected, and the same process is applied iteratively until it converges. Convergence is achieved when the mean of the two groups is separated by no more than four standard deviations of the larger group. A cutoff value is derived as the average value plus two standard deviations from the total number of somatic mutations in the last large group. If the derived cutoff

value is below 100 times the number of mutational channels, the cutoff value is adjusted to 100 times the number of mutational channels. For each column where the sum exceeds the derived cutoff value, each cell in the column is multiplied by the cutoff value and subsequently divided by the original column sum. Note that 100X normalization is performed if the means of the first two groups are not separated by at least four standard deviations. In all cases, fractional values after normalization are used as input for the factorization, and columns with a sum of zero, reflecting genomes without any somatic mutations, are ignored to avoid division by zero.

Module 3: Matrix factorization using nonnegative matrix factorization with replicates

By default, SigProfilerExtractor factorizes the matrix ***M*** with different ranks searching for an optimal solution between $k = 1$ and $k = 25$ mutational signatures. For each value of k, by default, the tool performs 100 independent nonnegative matrix factorizations of the normalized Poisson resampled input matrices. Thus, for each value of k, SigProfilerExtractor generates 100 distinct factorizations of normalized Poisson resampled matrices resulting into 100 different matrices ***S***, each matrix reflecting the patterns of the *de novo* mutational signatures, and 100 different matrices ***A***, each matrix reflecting the activities of the *de novo* mutational signatures (Figure 1B). To perform each of these factorizations, SigProfilerExtractor utilizes a custom implementation of the multiplicative update algorithm.[16] Specifically, SigProfilerExtractor initializes the ***S*** and ***A*** matrices in the first step of the factorization using either random initial conditions (default) or one of the derivatives of nonnegative double singular vector decomposition.[324] SigProfilerExtractor provides internal support for minimizing three different objective functions based on: *(i)* generalized Kullback-Leibler updates (default); *(ii)* Euclidean updates; *(iii)* Itakura-Saito updates. By default, the tool performs all factorizations using multithreading of central processing units (CPUs) and provides support for factorization using graphics processing units (GPUs) by leveraging PyTorch.[54] In all cases, by default, the implemented minimization performs at least 10,000 iterations (also known as NMF updates or NMF multiplicative update steps) with a maximum of 1,000,000 iterations. By default, the convergence tolerance of the algorithm is set to 10^{-15}. Note that SigProfilerExtractor allows reconfiguring all factorization parameters.

Module 4: Custom partition clustering and performing model selection

The previously described *Module 3* generates a number of sets with each set containing, by default, 100 different matrices ***S***, where each matrix reflects the patterns of *de novo* mutational signatures for a particular factorization of a normalized Poisson resampled matrix. One set, containing 100 different matrices ***S***, is generated for each of the interrogated total number of operative signatures, k, with a default range for k between 1 and 25 signatures. For each value of k, *Module 4* first performs custom clustering of the ***S*** matrices and, subsequently, applies a modified version of the NMFk model selection approach to select the optimal value of k[51] (Figure 1B). Specifically, for each value of k, the clustering is initialized with k random centroids. One of the ***S*** matrices is randomly chosen, and its columns matched to the most similar centroids with no two columns assigned to the same cluster. The process is repeated until the columns of all ***S*** matrices in the set are assigned to their respective clusters. SigProfilerExtractor implements the Hungarian algorithm[45] to pair consensus vectors from two matrices (*i.e.,* cluster centroids and mutational signature from a matrix ***S***); the Hungarian algorithm maximizes the total cosine similarities of all paired vectors between two matrices.[45] After assigning all columns to a cluster, the centroid of each cluster is recalculated by evaluating the average of all columns/vectors in a cluster. This process continues iteratively until the average silhouette coefficient converges (*i.e.,* its value does not change by more than 10^{-12}). After convergence for a given value of k, the centroids of the clusters are reported as consensus mutational signatures, an overall reconstruction error is calculated for describing the original input matrix, ***M***, and stability is calculated for each signature by computing the silhouette value[325] of the cluster corresponding to that signature (Figure 1B). The silhouette value of a cluster measures the similarities of the objects assigned to that cluster compared to any other cluster. Silhouette values range from −1.0 to +1.0 with values above zero indicating that, on average, objects have a higher similarity with their own cluster compared to their nearest clusters. Note that signatures with low stability correspond to a lack of uniqueness of the NMF due to Poisson resampling and/or to the potential existence of multiple convergent stationary points in the NMF solution.[47]

Our custom clustering is performed for each of the interrogated total number of operative signatures, k, with a default range for k between 1 and 25 signatures. After performing clustering, for each value of k, one has derived: *(i)* the consensus set of mutational signatures; *(ii)* an overall reconstruction error for describing the original input matrix; and *(iii)* stability value for each of the identified consensus mutational signatures.

SigProfilerExtractor performs a solution selection based on the stability of signatures in a solution and the ability of these signatures to reconstruct the original input matrix. By default, for whole-genome sequenced samples, SigProfilerExtractor will consider solutions stable if the signatures derived in the solution have an average stability above 0.80 with no individual signature having stability below 0.20 (0.70 and 0.10, respectively, are the recommended thresholds for extractions based on whole-exome sequenced samples). To reduce overfitting, the tool also measures the information gained from the extracted set of signatures in each solution. SigProfilerExtractor compares, using Wilcoxon rank-sum tests, the reconstruction errors across all samples from the stable solution with the greatest number of signatures to the reconstruction errors across all samples from stable solutions with lower number of signatures. Stable solutions with lower number of signatures are compared in a decreasing order to their total number of signatures with comparison stopping if the Wilcoxon rank-sum test yields a p *value* below 0.05 (*i.e.*, reflecting that a solution does not describe the original data as good as the stable solution with the greatest number of signatures). The stable solution with the lowest number of signatures and a Wilcoxon rank-sum test p *value* above 0.05 is selected as the optimal solution. If no solution has a Wilcoxon rank-sum test p *value* above 0.05, the stable solution with the greatest number of signatures is selected as the optimal solution. This test is not considered when extracting signatures from whole-exome sequenced samples, to favor sensitivity in low-mutation-count data.

Note that while SigProfilerExtractor selects an optimal solution, it outputs all the information necessary to evaluate mutational signatures and their activities for all other stable and unstable solutions.

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

SigProfilerExtractor provides a module for decomposing each of the *de novo* extracted mutational signatures to a set of previously derived signatures. By default, the tool decomposes each of the signatures in the optimal solution to a set of COSMICv3.1 reference signatures[12] with support for signatures of single base substitutions (SBS), doublet base substitutions (DBS), and small insertions and deletions (ID). Since the SBS COSMICv3.1 reference signatures were derived under the SBS-96 classification,[15] any extended classification of single base substitutions (*e.g.*, SBS-288 and SBS-1536)[15] is first collapsed to the SBS-96 classification and, subsequently, decomposed to the COSMICv3.1 reference signatures.[12] The decomposition functionality leverages the nonnegative least squares (NNLS) algorithm[326] as its main computational engine. A mixture of addition and removal steps (add-remove functionality) were developed to estimate the list of COSMIC signatures for a *de novo* signature. Specifically, for each *de novo* signature, a COSMIC signature is iteratively added to a list of signatures used to explain the *de novo* signature, NNLS is applied, and the signature which addition causes the greatest decrease of the L2 error is selected. If this greatest decrease is more than a specific threshold (default value of 5%) then the signature is included in the list of signatures used to explain the *de novo* signature. The addition is immediately followed by a removal step. Each COSMIC signature in the list of signatures used to explain the *de novo* signature are iteratively removed, NNLS is applied, and the signature that causes the least decrease of the L2 error is selected. If this least decrease is less than a specific threshold (default value of 1%) then the signature is removed from the list of signatures used to explain the *de novo* signature. The addition and removal steps are iterated until no signature is added or removed from the list of signatures used to explain the *de novo* signature. Several previously implemented rules for mutational signatures are incorporated by default in the decomposition module.[12] Specifically, for signatures of single base substitutions: *(i)* the list of signatures used to explain the *de novo* signature is initialized with the clock-like signatures SBS1 and SBS5;[9] *(ii)* biologically connected signatures are included as previously done in Ref [12] (*e.g.,* if SBS17a is included in the list then SBS17b is also included the list). The decomposition module is highly customizable as it allows changing all default parameters as well as adding additional new rules or removing existing rules for inclusion and exclusion of particular signatures.

Module 6: Evaluating activities of mutational signatures in individual samples

De novo extracted and COSMIC derived signatures are refitted to individual samples using nonnegative least squares (NNLS).[326] *Module 6* internally utilizes the add-remove functionality of *Module 5* with each sample in the original matrix, $\boldsymbol{M}$, being individually examined. For *de novo* mutational signatures, all *de novo* signatures are initially added to the list of signatures used to explain the sample and a removal step with a cutoff of 2% is applied. To assign COSMIC signatures in a sample, the module first derives the set of *de novo* signatures in that sample. Decomposition to the COSMICv3.1 signatures using *Module 5* is performed for each of the *de novo* signatures and the identified COSMICv3.1 signatures are refitted using the add-remove functionality with a removal and addition cutoffs set at 5%. Finally, the activity matrix is constructed by combining the activity vectors generated for all samples in the dataset.

Module 7: Outputting and plotting of analysis results

All previous modules make use of *Module 7* for outputting and plotting of the generated results. It should be noted that SigProfilerExtractor provides extensive output for the interrogated total number of operative signatures, *k*, with a default range of *k* between 1 and 25 signatures. For each value of *k*, SigProfilerExtractor outputs the set of operative *de novo* mutational signatures, the activities of the operative signatures, and an extensive set of information related to individual samples, individual *de novo* signatures, and the overall convergence of the factorization and clustering. *Module 7* also provides additional information when run in debug mode. In addition to outputting information, SigProfilerExtractor also generates a bouquet of plots both for each value of *k* as well as for the suggested optimal solution. SigProfilerExtractor utilizes all previously implemented plots in SigProfilerPlotting[15] as well as includes several newly developed visualizations.

Analysis of the genomics data from cancer and normal somatic tissues

For all examined whole-genome sequenced cancer and normal somatic tissues, *de novo* extraction of mutational signatures was performed with SigProfilerExtractor with default parameters using two distinct mutational classifications: SBS-96 and SBS-288. Only the SBS-96 classification was used for whole-exome sequenced data. The SBS-96 mutation classification incorporates the six types of single base substitutions: C>A, C>G, C>T, T>A, T>C, and T>G. Each type of single base substitution is further separated into 16 subtypes determined by the four possible bases 5′ and -3′ adjacent to each mutated base. The SBS-288 mutation classification extends the SBS-96 mutation classification by adding additional information for each of the 96 subtypes. Specifically, SBS-288 incorporates whether a single base substitution is in non-transcribed/intergenic DNA, on the transcribed strand of a gene, or on the untranscribed strand of the gene. *De novo* extraction was performed separately for all examined datasets. Specifically, SigProfilerExtractor was applied: *(i)* to all 2,778 whole-genome sequenced cancers from the Pan-Cancer Analysis of Whole Genomes project[43]; *(ii)* to all 1,865 whole-genome and 19,184 whole-exome sequenced cancers from the extended cohort (Table S8); *(iii)* to all samples in each of the 37 cancer types of the Pan-Cancer Analysis of Whole Genomes project[43] with each cancer type examined separately; *(iv)* to all samples in each of the 66 cancer types of the extended cohort (Table S8) with each cancer type examined separately; *(v)* to all 88 whole-genome sequenced microbiopsies of histologically normal urothelium[50]; *(vi)* to the complete set of whole-genome sequenced bladder cancers from TCGA[56]; *(vii)* to exome downsampling of all bladder whole-genome sequenced

cancers from the Pan-Cancer Analysis of Whole Genomes project[43]; *(viii)* to exome downsampling of all 88 whole-genome sequenced microbiopsies of histologically normal urothelium.[50] In all cases, the mutational catalog of each sample was taken from the respective original publications. The results from all performed *de novo* extractions can be found at: ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/. Downsampling of whole-genome sequenced samples to whole-exome was performed using SigProfilerMatrixGenerator.[15]

Additional approaches for miscellaneous analysis

Cosine similarity was used to compare the profiles of different mutational signatures. P-values can be attributed to cosine similarities based on a null hypothesis of uniform random distribution of nonnegative vectors.[48]

Briefly, the prevalence of somatic mutations in a whole-exome sample was calculated based on the identified mutations in protein coding genes and assuming that an average whole-exome has sufficient coverage of 30.0 megabase-pairs in protein coding genes. The prevalence of somatic mutations in a whole-genome sample was calculated based on all identified mutations and assuming that an average whole-genome has sufficient coverage of 3.00 gigabase-pairs.

In order to characterize the shape of the false positive signatures identified by the different signature extraction tools, we used the Shannon equitability index metric[327] for mutational signatures, defined as follows.

$$Shannon\ \ equitability\ \ index = -\frac{\sum_{i=1}^{t} p_i ln\ p_i}{ln\ t}$$

In this formula, $\boldsymbol{p}$ represents the probability of a mutation caused by a specific mutational signature to belong to a specific mutational channel, whereas $\boldsymbol{t}$ is the total number of mutational channels or rows of the input mutational matrix $\boldsymbol{M}$ (which corresponds to 96 in the case of the well-known SBS-96 classification). The range of the Shannon equitability index goes from zero, characterizing a trivial signature where only one specific mutational channel is possible (*i.e.*, null diversity of mutational channels), to one, which would correspond to a completely uniform signature where all mutational channels accumulate the same probability (for example, 1/96 in the case of the SBS-96 classification). Indeed, well-known COSMIC signatures commonly defined as sparse, such as clock-like SBS1 or APOBEC-related SBS2 show a Shannon equitability index of 0.409 and 0.267, respectively, whereas signatures usually defined as flat, including SBS3, SBS5, and SBS40 display a much higher Shannon equitability of 0.961, 0.941, and 0.949, respectively, which is closer to the maximum value.

Creation of scenarios with synthetic datasets

Benchmarking was performed on simulated datasets with and without noise. These synthetic datasets were created using a previously described method.[12] All datasets without noise were categorized as different scenarios with many of these scenarios attempting to emulate a particular set of cancer types. Specifically, 20 scenarios were created for the SBS-96 mutational classification and 12 additional scenarios were generated for the extended number of channels. For the SBS-96 classification, COSMIC signatures originally extracted from the PCAWG dataset[12] as well as signatures extracted using SignatureAnalyzer[12] and random signatures were used as ground-truth signatures. Many of the scenarios were created using a combination of tissue specific signatures. Signature profiles of extended scenarios (E) were based either on random signatures or on composite signatures extracted by SigProfiler_PCAWG or SignatureAnalyzer. Composite signatures consist of a total of 1,697 mutation types, encompassing an amalgamation of: 1536 strand-agnostic single base substitutions (SBS-1536) in a pentanucleotide context, 78 doublet-base substitutions (DBS-78), and 83 types of small insertions and deletions (ID-83).

Attributions of signatures in the different scenarios associated with a cancer type, *t*, were generated based on three parameters that were in turn based on the observed statistics for each signature, *s*, in cancer type *t*: π, the proportion of tumors of cancer type *t* with signature *s*; μ, the mean of $\log_{10}$ of the number of *s* mutations across those tumors of type *t* that have signature *s*; and σ, the standard deviation of $\log_{10}$ of the numbers of *s* mutations across those *t* tumors that have *s*.

Synthetic scenarios were labeled as easy, medium, and hard based on the number of operative signatures in each scenario. Based on our most recent analysis of mutational signatures in 82 cancer types,[12] approximately 7.4% of human cancer types have 5 or less signatures (reflected in simulations of easy scenarios), 15.9% have 11 to 21 signatures (medium scenarios), and 59.5% have 25 or more signatures (hard scenarios). Note that 17.2% of cancer types have either between 5 and 10 signatures or between 22 and 24 signatures.

Detailed description of each of the used scenarios for benchmarking is provided below. Note that some of the generated scenarios were initially created as part of Ref. [12]. The computational approach for generating the synthetic data can be found at: https://github.com/steverozen/SynSigGen. Noiseless scenarios 1 to 14 had only a single replicate while scenario 15 through 20 had 10 replicates each, and WGS and WES noise scenarios had 20 replicates each.

Scenarios 1, 2, E-1, and E-2

The scenarios were generated to emulate a subset of the pancreatic adenocarcinoma PCAWG dataset with a total 1,000 synthetic samples. Ground-truth signatures were based on COSMIC as well as on signatures extracted by SignatureAnalyzer.

Scenarios 3, 4, E-3, and E-4

Mutational spectra generated from combinations of flat, relatively featureless mutational signatures. A total of 1,000 synthetic tumors emulating a mixture of 500 synthetic renal cell carcinomas (high prevalence and mutation load from SBS5 and SBS40 signatures) and

500 synthetic ovarian adenocarcinomas (high prevalence of and mutation load from SBS3). Ground-truth signatures were based on COSMIC and signatures extracted by SignatureAnalyzer. This dataset embodies tumors with high prevalence of the main flat signatures, SBS3, SBS5, and SBS40, in a realistic context.

Scenarios 5, 6, E-5, and E-6

Mutational spectra generated from signatures with overlapping and potentially interfering profiles. A total of 1,000 synthetic tumors composed mostly from SBS2, SBS7a, and SBS7b. Mutational load distributions were drawn from bladder transitional cell carcinoma (SBS2) and skin melanoma (SBS7a, SBS7b). Ground-truth signatures were based on COSMIC and signatures extracted by SignatureAnalyzer. Most spectra contain both signatures SBS7a and SBS7b. The potential interference is between SBS2 (mainly C>T) and SBS7a, SBS7b (mainly C>T).

Scenarios 7, 8, E-7, and E-8

Mutational spectra generated from combinations of flat, relatively featureless mutational signatures. A total of 1,000 synthetic tumors emulating a mixture of 500 synthetic renal cell carcinomas (high prevalence and mutation load from SBS5 and SBS40 signatures) and 500 synthetic ovarian adenocarcinomas (high prevalence and mutation load from SBS3). Ground-truth signatures were based on COSMIC and signatures extracted by SignatureAnalyzer. This dataset embodies tumors with high prevalence of the main flat signatures, SBS3, SBS5, and SBS40, in a simplified fashion, where only these three signatures are present.

Scenarios 9, 10, E-9, and E-10

Mutational spectra generated from signatures with overlapping and potentially interfering profiles. A total of 1,000 synthetic tumors composed from SBS2, SBS7a, and SBS7b. Mutational load distributions were drawn from bladder transitional cell carcinoma (SBS2) and skin melanoma (SBS7a, SBS7b). Ground-truth signatures were based on COSMIC and signatures extracted by SignatureAnalyzer. Most spectra contain both groups of signatures. The potential interference is between SBS2 (mainly C>T) and SBS7a, SBS7b (mainly C>T). This dataset presents synthetic tumors containing these three signatures in a simplified fashion, excluding the presence of any additional mutational signature.

Scenarios 11, 12, E-11, and E-12

A set of 30 random mutational signature profiles based on SBS-96 classification and a set of 30 random 1,697 feature signature profiles (mimicking SignatureAnalyzer's composite signatures). Each of these sets of random signatures were used in two types of exposures, one with more (mean ~15.6) signatures per tumor and one with fewer (mean ~4) signatures per tumor.

Scenarios 13 and 14

A set of 2,700 synthetic whole-genome samples with mutational spectra matching the ones observed in PCAWG, including 300 spectra from each of 9 different cancer types. These spectra consist of 300 synthetic spectra from each of the following cancer types: bladder transitional cell carcinoma, esophageal adenocarcinoma, breast adenocarcinoma, lung squamous cell carcinoma, renal cell carcinoma, ovarian adenocarcinoma, osteosarcoma, cervical adenocarcinoma, and stomach adenocarcinoma. Ground-truth signatures were based on COSMIC as well as on signatures extracted by SignatureAnalyzer.

Scenarios 15 and 18

A set of 5 random mutational signature profiles based on SBS-96 mutational classification. A total of 200 synthetic tumors were generated with one scenario containing an average of 3 signatures per tumor while the other scenario had an average of 5 signatures per tumor.

Scenarios 16 and 19

A set of 15 random mutational signature profiles based on SBS-96 mutational classification. A total of 200 synthetic tumors were generated with one scenario containing an average of 10 signatures per tumor while the other scenario had an average of 5 signatures per tumor.

Scenarios 17 and 20

A set of 25 random mutational signature profiles based on SBS-96 mutational classification. A total of 200 synthetic tumors were generated with one scenario containing an average of 15 signatures per tumor while the other scenario had an average of 5 signatures per tumor.

WGS noise scenario

In addition to noiseless scenarios, we simulated 20 replicates of a WGS scenario with noise: 10 of the replicates were based on scenario 11 and another 10 replicates were based on scenario 12. In each case, white Gaussian noise was added to each replicate in order to study the performance of the tools at different amounts of noise, emulating differences in the sequencing quality of real datasets. Specifically, random noise was introduced for different noise levels (0%, 1%, 2.5%, 5%, or 10%) by resampling every data point in the mutational matrix (*i.e.*, reflecting the number of mutations of a specific mutation type in a cancer sample) using a Gaussian distribution where the mean corresponds to the value of the data point, and the standard deviation is the value of the data point multiplied by the specific noise level. Subsequently, decimal values were truncated, and negative values were replaced with zeros. Overall, 5 distinct levels of noise were generated, each repeated 20 times, with an average noise level corresponding to 0%, 1%, 2.5%, 5%, and 10% of all mutations observed in the replicate.

WES noise scenario

A WES-based scenario was generated by downsampling the WGS-based noise scenario corresponding to a 5% noise level, reflecting high-quality genomic datasets. The downsampling of synthetic cancer genomes and randomly generated ground-truth mutational signatures was done in a two-step process. Firstly, WGS-based SBS-96 mutational matrices were simulated with

SigProfilerSimulator[48] to obtain VCF files with simulated synthetic mutations spanning across the whole genome. Subsequently, exome-specific SBS-96 mutational matrices were created with SigProfilerMatrixGenerator[15] including exclusively the synthetic mutations affecting the exome portion of the human genome based on the SureSelect Human All Exon v7 protocol (Agilent, Santa Clara, CA, USA). This two-step process allows considering the differences in trinucleotide frequencies between the exome and the whole human genome, which would not be captured by direct downsampling of the original WGS mutational matrices based on the fact that the exome constitutes ~2% of the human exome but has a different trinucleotide context.

Benchmarking bioinformatic tools for *de novo* extraction of mutational signatures

The hitherto described synthetic scenarios were used to compare SigProfilerExtractor and thirteen other tools for *de novo* extraction of mutational signatures. The method and parameters used to extract signatures from the simulated datasets using each tool are described below. With the exception of EMu and SignatureAnalyzer, which support only detection of the total number of mutational signatures without a prespecified range, all other tools required specifying the range for the total number of operative mutational signatures. The ranges for benchmarking with suggested model selection, which most closely matches the analysis of a real dataset with unknown number of signatures, are provided in Table S9 for each of the scenarios. Benchmarking with forced model selection, where tools were required to extract the known number of ground-truth mutational signatures, performed *de novo* extraction based on the ground-truth number of total mutational signatures (Table S9). For the WGS noise scenario, the same ranges used for noiseless scenarios 11 and 12 were applied. In the case of the WES noise scenario, a reduced range of signatures was used to optimize running time, based on the low sensitivity observed for all tools on WES data compared to WGS. This WES-specific range was used to extract *de novo* mutational signatures in all tools except for SigneR, whose signature selection method depends on the maximum number of signatures tested (original WGS range was applied).

SigProfilerExtractor

The default settings of SigProfilerExtractor (version 1.1.4) were used to extract mutational signatures with minor modifications to reduce overall extraction time. Specifically, we utilized "NMF replicates" = 100, "minimum NMF iterations" = 1,000, "maximum NMF iterations" = 200,000, "NMF test convergence" = 1,000 and "NMF tolerance" = 1e-08 parameter settings for all scenarios without noise and all WGS and WES replicates with noise.

SigProfiler_PCAWG

The default settings of SigProExtractor[12] (version 0.0.5.48) were used to extract signatures from the benchmark scenarios with exception of "totaliteration" = 100.

SignatureAnalyzer

For the scenarios with and without noise, we used the default parameters for SignatureAnalyzer[32,34] described in https://github.com/broadinstitute/SignatureAnalyzer-GPU. For the extended scenarios, the CPU version of SignatureAnalyzer was used with 20 runs and default parameters. The mode number of signatures counts was selected for further evaluation.

MutationalPatterns

We downloaded MutationalPatterns[24] version 3.0.1 according to the instructions at https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html. To extract signatures, we used the NMF method with default parameters with the exception of the "nrun" parameter. The "nrun" parameter was set to 200 in order to increase the reliability of the extraction. To select the optimum number of signatures, as suggested by the developers of the tool, we used the RSS plot that is generated in the NMF rank survey plot.

SignatureToolsLib

We downloaded SignaturesToolsLib[35] from https://github.com/Nik-Zainal-Group/signature.tools.lib and the tool was used using the parameters recommended by the authors. Specifically, we utilized "nboots" = 20, "nrepeats" = 200 and "filterBest_RTOL" = 0.001. As suggested by the developers, we selected the optimum number of signatures from the plot illustrating the overall metrics. We mostly used the "norm.Error" and "Ave.SilWid" with Clustering with Matching (MC) to select the total number of operative mutational signatures.

SigneR

We used the SigneR[36] version 1.16.0 as described in http://bioconductor.org/packages/release/bioc/vignettes/signeR/inst/doc/signeR-vignette.html. The signatures were extracted with default parameters without using an opportunity matrix.

MutSpec

We used the command line platform of MutSpec[27] version 2.0, as described at https://github.com/IARCbioinfo/mutspec. To extract signatures, we used the default parameters. As suggested by the developers, we estimated the optimum number of signatures from the "NMF rank survey" plot generated from the "MutSpec-NMF_Estimate_Signatures" module of the package.

SomaticSignatures

We followed the instructions described at https://www.bioconductor.org/packages/release/bioc/vignettes/SomaticSignatures/inst/doc/SomaticSignatures-vignette.html to use the NMF method to assess and extract mutational signatures. SomaticSignatures[38] version 2.26.0 was used with default parameters. To access the number of signatures, we increased the value of "nmf_replicates" from 5 to 30 in order to get better reproducibility. As suggested by the developers, we selected the optimum number of signatures using the "plotNumberSignatures" function provided by the tool. In the plots, we relied on the RSS and explained variance value to choose the optimum solution.

Maftools

We followed the instructions provided at https://www.bioconductor.org/packages/release/bioc/vignettes/maftools/inst/doc/maftools.html to download and extract signatures using Maftools[22] version 2.2.0. As suggested by the developers, we estimated the goodness of fit to decide the optimal number of signatures using the "estimateSignatures" function. Then we extracted the corresponding optimal number of signatures using the "extractSignatures" function provided by the tool. All settings were kept as default, except we increased the value of "nTry" from 6 to 20 in order to increase reproducibility.

SigMiner

We installed SigMiner[31] version 1.0.0 according to the instructions provided at https://shixiangwang.github.io/sigminer-doc/. All the parameters were set as defaults to both estimate as well as extract mutational signatures. To select the optimum number of signatures, as suggested by the developers, we assessed the statistics provided in the NMF rank survey plot.

SigFit

We used the instructions provided at http://htmlpreview.github.io/?https://github.com/kgori/sigfit/blob/master/doc/sigfit_vignette.html to download and extract signatures from SigFit[28] version 2.0.0. Signature extraction was done using default parameters with the exception of the total number of iterations which was set at 100.

EMu

Benchmarking for EMu[20] was done with version 1.5.2 using default parameters for suggested extractions. Additionally, the optional "force" parameter was used for benchmarks done with a specific number of signatures. EMu was the only tool that was unable to complete *de novo* extractions from a number of synthetic datasets (Tables S1, S2 and S3) with the tool either running out of memory on instances with 256 GiB memory or running for 4+ weeks without producing any results. These scenarios were considered as failed and assigned F_1 scores of zero.

MutSignatures

MutSignatures[25] version 2.1.1 was downloaded and run according to https://cran.r-project.org/web/packages/mutSignatures/vignettes/get_sarted_with_mutSignatures.html. Signatures were extracted using 500 iterations ("num_totIterations" = 500).

TensorSignatures

TensorSignatures[40] version 0.5.0 was downloaded and run according to the instructions at https://github.com/sagar87/tensorsignatures. Input VCFs were generated from the matrices by running SigProfilerSimulator.[48] The headers of the VCF files were modified for TensorSignatures to compute the trinucleotide normalization. TensorSignatures was applied using 10,000 epochs, overdispersion of 50, and trinucleotide normalization. Each decomposition rank was run with 10 iterations. TensorSignatures was not applied to scenarios 13 and 14 as the estimated computation time, even with multiple GPUs, was expected to be more than 6 months per scenario.

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantitative and statistical analyses are described in the relevant sections of the Method details and in the figure legends.

ADDITIONAL RESOURCES

The novel mutational signatures identified in the present study were included within the reference set of mutational signatures, available at the COSMIC Mutational Signatures website (https://cancer.sanger.ac.uk/signatures/).

Cell Genomics, Volume *2*

Supplemental information

Uncovering novel mutational signatures

by *de novo* extraction with SigProfilerExtractor

S.M. Ashiqul Islam, Marcos Díaz-Gay, Yang Wu, Mark Barnes, Raviteja Vangara, Erik N. Bergstrom, Yudou He, Mike Vella, Jingwei Wang, Jon W. Teague, Peter Clapham, Sarah Moody, Sergey Senkin, Yun Rose Li, Laura Riva, Tongwu Zhang, Andreas J. Gruber, Christopher D. Steele, Burçak Otlu, Azhar Khandekar, Ammal Abbasi, Laura Humphreys, Natalia Syulyukina, Samuel W. Brady, Boian S. Alexandrov, Nischalan Pillay, Jinghui Zhang, David J. Adams, Iñigo Martincorena, David C. Wedge, Maria Teresa Landi, Paul Brennan, Michael R. Stratton, Steven G. Rozen, and Ludmil B. Alexandrov

Supplementary Figures

Figure S1. Standard set of performance metrics used for benchmarking all bioinformatics tools, Related to Figures 2 and 3. An example demonstrating the derivation of *true positive* (TP), *false positive* (FP), or *false negative* (FN) signatures for a tool applied to a synthetic dataset generated using 6 ground-truth signatures (termed, Ground-Truth Signatures 1 through 6). The tool extracts 4 signatures (termed, Extracted Signatures A through D). In this example, an extracted signature is considered a true positive if it matches one of the ground-truth signatures with a cosine similarity threshold of at least 0.90.

Simulated *dataset* using 6 ground truth signatures (Ground Truth Signatures 1 through 6). A tool extracts 4 signatures (Extracted Signatures A through D). Comparison between Ground Truth and Extracted Signatures using cosine similarity.

	Extracted Signature A	Extracted Signature B	Extracted Signature C	Extracted Signature D
Ground Truth Signature 1	0.14	**0.98**	0.56	0.36
Ground Truth Signature 2	0.35	0.29	**0.93**	0.46
Ground Truth Signature 3	0.31	0.56	0.78	0.66
Ground Truth Signature 4	0.34	0.08	0.57	0.67
Ground Truth Signature 5	**0.95**	0.15	0.81	0.39
Ground Truth Signature 6	0.23	0.74	0.48	0.26

True Positives (TP; ≥ 0.90)

Extracted Signature A
Extracted Signature B
Extracted Signature C

Signatures correctly extracted from the *dataset*

False Positives (FP)

Extracted Signature D

Signatures extracted but absent in the *dataset*

False Negatives (FN)

Ground Truth Signature 3
Ground Truth Signature 4
Ground Truth Signature 6

Signatures not extracted but used in simulating the *dataset*

Cosine similarity between Extracted Signature C and Ground Truth Signature 6

Figure S2. Comparison of the different options available in SigProfilerExtractor for matrix normalization, NMF initialization, and NMF objective function, Related to STAR Methods. Vertical axes reflect F_1 score (left plot), sensitivity (middle plot), and false discovery rate (right plot), respectively. Abbreviations: gmm: Gaussian mixture model; nndsvd_min: nonnegative double singular vector decomposition initialization where zeros are replaced by the minimum positive value.