SATS: A mutational signature analyzer of targeted sequenced tumors

Donghyuk Lee[1,2], Min Hua[1], Difei Wang[1], Lei Song[1], Tongwu Zhang[1], Kai Yu[1], Xiaohong R. Yang[1], Stephen J. Chanock[1], Jianxin Shi[1], Maria Teresa Landi[1], Bin Zhu[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America
[2]Department of Statistics, Pusan National University, Busan, Korea
Corresponding author: Bin Zhu (bin.zhu@nih.gov)

Abstract

Tumor mutational signatures are important in clinical decision-making and are typically analyzed using whole exome or genome sequencing (WES/WGS). However, targeted sequencing is more commonly used in clinical settings, posing challenges in mutational signature analysis due to sparse mutation data and non-overlapping targeted gene panels. We introduce SATS (Signature Analyzer for Targeted Sequencing), an analytical method that identifies mutational signatures in targeted sequenced tumors by analyzing tumor mutational burdens and accounting for different gene panels. We demonstrate through simulations and pseudo-targeted sequencing data (generated by down-sampling WES/WGS data) that SATS can accurately detect common mutational signatures with distinct profiles. Using SATS, we created a pan-cancer catalog of mutational signatures specifically tailored to targeted sequencing by analyzing 100,477 targeted sequenced tumors from the AACR Project GENIE. The catalog allows SATS to estimate signature activities even within a single sample, providing new opportunities for applying mutational signatures in clinical settings.

Introduction

Tumors accumulate somatic mutations that form specific patterns, known as mutational signatures[1,2], which can provide insight into the underlying mutational processes involved in carcinogenesis and inform cancer detection[3-5] and treatment[6-9]. For example, aristolochic acid-associated signatures can be used to screen for liver cancers[5], tumors with the HRD (homologous recombination deficiency)-associated signature can be treated with PARP (poly (ADP-ribose) polymerase) inhibitors[7], and ATR (ataxia telangiectasia and Rad3 related) inhibitors can be prescribed for cancers with APOBEC (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like)-associated signatures[8]. To analyze mutational signatures, multiple algorithms have been proposed[10-14] and catalogs of mutational signatures have been created[1,2] for tumors sequenced using whole exome or whole genome sequencing (WES/WGS).

In clinical practice, tumors are often sequenced using targeted gene panels that detect only a few mutations, focused on cancer driver genes with therapeutic potential. Such sparse mutation data, combined with the use of different panels across hospitals, makes it challenging to use existing tools designed for WES/WGS to analyze mutational signatures in targeted sequenced tumors. Additionally, existing catalogs of signatures identified through WES/WGS are typically based on common tumor types or subtypes and may not include signatures present in targeted sequenced tumors in clinical settings, such as tumors of rare cancer subtypes or those treated with specific therapies. Accordingly, there is a need for specialized analytical methods and a comprehensive

catalog of mutational signatures specifically tailored to targeted sequenced tumors to facilitate the practical use of mutational signatures in clinical settings.

Here, we introduce SATS (Signature Analyzer for Targeted Sequencing), designed specifically for analyzing mutational signatures in targeted sequencing data. Unlike existing methods optimized for WES/WGS, SATS considers the variability in the size and genomic context of targeted gene panels while leveraging large sample sizes of targeted sequencing studies. After discussing the limitations of existing methods and a detailed description of the SATS pipeline is provided. To investigate the factors that affect signature detection and refitting in targeted sequenced tumors, we conducted an analysis of pseudo-targeted sequencing data across various cancer types. These data include mutations that were identified in The Cancer Genome Atlas (TCGA) WES studies[1,15] or the Sanger breast cancer WGS study[16], and were located within the regions of the targeted sequencing panel. We also simulate additional breast cancer samples to examine the impact of sample size on signature detection, specifically focusing on the challenging-to-detect HRD-associated signature SBS3. Our simulation results demonstrate that with a sufficiently large sample size, SATS can identify almost all common mutational signatures of breast cancers based on targeted sequencing, including signature SBS3. Additionally, we perform simulations to validate the signatures detected by SATS across four major cancer types (lung, breast, colorectal and ovarian cancers) with sample sizes comparable to those in existing targeted sequencing studies. Furthermore, we demonstrate that SATS can accurately attribute mutations to prespecified signatures, even with limited sample sizes down to a single sample.

Finally, we apply SATS to establish a pan-cancer catalog of mutational signatures in 100,477 targeted sequenced tumors from the AACR (American Association for Cancer Research) Project GENIE (Genomics Evidence Neoplasia Information Exchange, Version 11.0-public)[17,18]. This database contains tumors collected from 18 different hospitals or cancer centers in multi-ethnic populations, representing 24 cancer types, including 14,428 lung and 11,389 breast tumors (Methods). Our analysis reveals the presence of well-established signatures in unexpected cancer types, such as the smoking-associated signature in ovarian tumors and azathioprine-induced signature in endometrial or pancreatic cancers. This comprehensive catalog of mutational signatures, tailored specifically to targeted sequenced tumors, provides an important resource for clinical applications of mutational signatures, as well as for future cohort or consortia studies that involve targeted sequenced data from multiple centers.

Results

Limitations of existing methods

Mutational signature analysis of tumors with WES/WGS data can be performed using signature extraction[10-12] or signature refitting[13,14] methods. Signature extraction aims to extract *de novo* mutational signatures, while signature refitting estimates the activity of prespecified signatures in a specific sample. However, these methods have limitations when applied to targeted sequenced tumors. Targeted sequencing detects a limited number of mutations, making it hard to distinguish correlated *de novo* signatures using signature extraction[19]. For signature refitting approaches, it is uncertain which reference signatures to use in targeted sequenced tumors. When using reference signatures from all cancer types, the signature refitting methods can incorrectly assign mutations to signatures that are not present in the targeted sequenced tumors[20]. Furthermore, most methods

assume that identical genomic regions (e.g., whole exome or genome) are sequenced across all samples, which may not hold true for targeted sequencing studies using various panels that target different genomic regions. While signature extraction and refitting methods have been successful in analyzing mutational signatures in tumors with WES/WGS, their effectiveness is poor when applied to targeted sequencing data.

Clustering methods have recently been proposed for mutational signature analysis in targeted sequenced tumors. For example, the SigMA[19] algorithm is developed to detect the HRD-associated signature SBS3, requiring WGS data from individual tumors as input to classify targeted sequenced tumors. Similarly, *MUTYH* mutation-related signatures SBS18/36 are identified in targeted sequenced colorectal cancers using WES data of individual samples as the training set for clustering[21]. However, both methods are limited in their ability to analyze multiple mutational signatures in targeted sequenced tumors since they are designed to detect specific mutational signatures. Moreover, the sample size of tumors analyzed via WGS/WES is limited, and the mutational signatures derived from these tumors might not be representative of signatures in targeted sequenced tumors. Therefore, a more flexible method that does not rely on sample-level mutations from WES/WGS is desired to detect multiple mutational signatures in targeted sequenced tumors.

Overview of SATS

We propose SATS for detecting mutational signatures and estimating their activities in targeted sequenced tumors. SATS consists of four main steps (Fig. 1): a) Apply signeR[11] to extract *de novo* signature profiles, adjusting for differences in panel sizes (see Supplementary Note for the calculation of panel sizes). b) Identify catalog signatures present in target-sequenced tumors, we use the penalized nonnegative least squares (pNNLS)[22] to map the extracted *de novo* signature profiles to the profiles of pan-cancer catalog signatures (such as ones in COSMIC signature databases[1], as described in Methods). The pNNLS identifies the catalog signatures that have a significant contribution to the *de novo* signature profiles. In addition, the detected catalog signatures can be from any cancer types included in the pan-cancer catalog, not just the cancer type of the tumors undergoing targeted sequencing. c) Refit the detected signatures to estimate the activities of signatures for individual tumors through a proposed Expectation–Maximization (EM) algorithm. d) Estimate signature expectancy which is the expected number of mutations attributed to individual signatures for each tumor (Methods).

SATS is based on a Poisson nonnegative matrix factorization (pNMF) model (Methods). While we use single base substitution (SBS) for the purpose of illustration, pNMF can be applied to other types of somatic mutations, such as double base substitutions (DBS). For SBS, we consider 96 mutation types based on 32 trinucleotide contexts, such as a C to G mutation at the trinucleotide context TCT (i.e., a T[C>G]T mutation type). The pNMF model assumes that the SBS count v_{pn} for the p^{th} mutation type in the n^{th} targeted sequenced tumor follows a Poisson distribution with mean $\ell_{pn}\sum_{k=1}^{K} w_{pk}h_{kn}$ for K signatures, where v_{pn}, ℓ_{pn}, w_{pk} and h_{kn} represent elements of the corresponding matrices **V**, **L**, **W** and **H**, respectively. The mutation type matrix **V** includes the observed number of mutations per mutation type, and the panel context matrix **L** contains the number of trinucleotide contexts at which a specific mutation type (e.g., TCT for T[C>G]T substitutions) could potentially occur in the targeted sequence. The matrices **W** (with dimension N by K) and **H** (with dimension K by 96) describe the profiles and activities of K

signatures, respectively. Here, **W** and **H** are parameters of interest and will be estimated based on the log-likelihood function of the pNMF model:

$$\log\{P(\mathbf{V}|\mathbf{L},\mathbf{W},\mathbf{H})\} = \sum_{n=1}^{N}\sum_{p=1}^{96}\{v_{pn}\log(\ell_{pn}\sum_{k=1}^{K}w_{pk}h_{kn}) - \ell_{pn}\sum_{k=1}^{K}w_{pk}h_{kn} - \log(v_{pn}!)\}. \quad (1)$$

When the genomic regions sequenced across all samples are identical, as in WES or WGS, the maximum likelihood estimate based on equation (1) is equivalent to that based on the canonical NMF (Methods). The proposed pNMF model includes the canonical NMF as a special case.

Signatures of tumor mutation burden

SATS differs from other methods for mutational signature analysis in that SATS is able to identify signatures of tumor mutation burden (TMB), rather than signatures of tumor mutation count (TMC). Traditional mutational signature analysis algorithms use the canonical NMF method to factorize a mutation type matrix **V** into a $96 \times K$ signature profile matrix $\mathbf{W}'$ and a $K \times N$ signature activity matrix $\mathbf{H}'$, such that **V** can be approximated by $\mathbf{W}'\mathbf{H}'$ as $\mathbf{V} \approx \mathbf{W}'\mathbf{H}'$, for a given number of signatures K. Thus, the estimated signature profile matrix $\mathbf{W}'$ (or its scaled version) represents the number of mutations at each mutation type for K signatures, which essentially represents the signatures of TMC. In contrast, SATS decomposes **V** as $\mathbf{V} \approx \mathbf{L} \circ \mathbf{WH}$, where $\circ$ denotes element-wise product. In this approach, SATS uses the panel context matrix **L** (measured in megabase (Mb) pairs) to estimate the signature profile matrix **W** (or its scaled version). This matrix **W** describes the number of mutations per Mb at each mutation type for K signatures, which can be interpreted as signatures of TMB.

TMB signature profiles are different from TMC signature profiles, as TMB profiles take into account the mutation context while TMC profiles do not. For example, the TMC SBS5 profile is relatively consistent across all 16 trinucleotide contexts of C to T mutations, whereas the TMB SBS5 profile shows increased C to T mutations at the NCG trinucleotides (N represents any nucleotide, Supplementary Fig. 1a), since these trinucleotides are depleted in the human genome due to frequent deamination of 5-methylcytosine to thymine[23,24] (Supplementary Fig. 1b). We compared the shape of TMB and TMC profiles by the Shannon equitability index, which measures the evenness of signature profiles across mutation types (Methods). A higher value of the index corresponds to a flatter signature profile, whereas a lower value suggests a distinct or spiker profile with significant contributions from certain mutation types as "spikes." The Shannon equitability indices of TMB and TMC profiles are highly correlated (Pearson correlation coefficient r = 0.915, Supplementary Fig. 1c), but there are a few exceptions. For example, TMB signature SBS10b and SBS15 (Shannon equitability index = 0.192 and 0.391 respectively) exhibit more pronounced spikes than their TMC counterparts (Shannon equitability index = 0.491 and 0.624 respectively, and Supplementary Fig. 1d).

One advantage of TMB signature profiles is that the results are unaffected by differences in mutation contexts between targeted sequencing and WGS. For example, targeted genome sequences often contain a higher proportion of NCG trinucleotide but a lower proportion of NTA and NTT trinucleotides compared to whole genome sequences (Supplementary Fig. 2). As a result, TMC signature profiles obtained from targeted sequencing would differ from those obtained from WGS due to different mutation contexts, while TMB signature profiles would not. Additionally, the use of TMB signatures allows for mutational signature analysis across different

targeted gene panels by normalizing the numbers of mutation contexts, which may vary across panels.

Determinants of SATS signature detection and refitting

We investigated factors that can affect the detection and refitting of mutational signatures by SATS. Specifically, we examined how the size of the targeted gene panel, the shape of the TMB signature profile (measured by the Shannon equitability index), and the cancer type influence the accuracy of detection. We hypothesized that a more common TMB signature, or one with a more distinct profile, would be more easily detected when tumors are sequenced using a large size panel. To test this hypothesis, we generated pseudo-targeted sequencing data using SBSs that were called in The Cancer Genome Atlas (TCGA)[1,15] WES studies and located in targeted sequencing panel regions (Methods). Similarly, we generated pseudo-targeted sequencing data based on 560 breast tumors[16] with WGS data (Methods and Supplementary Fig. 3a). Our analysis focused on common signatures (with respect to a cancer type) that were reported[1] to contribute more than 5% of SBSs called by WES or WGS. While the sample size could also be a determinant of signature detection, we could not assess it using limited pseudo-targeted sequencing data. Thus, we conducted *in silico* simulations in the next section to examine the impact of sample size.

First, we assessed the ability of SATS to detect common signatures through the analysis of pseudo-targeted sequencing data based on WES. Our findings show that SATS is capable of detecting common signatures, with detection probabilities that vary across the cancer types, the targeted gene panels, and the signatures being analyzed. For instance, renal transitional cell carcinoma had the highest detection probability among all cancer types, while thyroid adenocarcinoma had the lowest (Fig. 2a). Larger gene panels generally uncovered more signatures (Fig. 2a). Additionally, our analyses revealed an inverse relationship (Pearson correlation coefficient = -0.452) between the detection probability of a signature and its Shannon equitability index (Fig. 2b). This suggests that spiky signatures are more likely to be detected than flat ones, which aligns with previous observations based on WES/WGS data[9]. To simultaneously quantify the impact of these factors on the probability of detecting mutational signatures, we fit a generalized linear mixed model (GLMM) that included cancer type, panel size, signature prevalence and Shannon equitability index of signature (Methods). Our results revealed that cancer type explained 53.26% of the variance of the detection probabilities at the logit scale (Fig. 2c). This is because the specific cancer type determines which mutational signatures are present in the tumor, and some signatures are more distinct and easier to separate than others. For example, transitional cell carcinoma exhibits a spiky SBS2/13 signature and a flat SBS5 signature, which are relatively easy to detect. In contrast, breast cancer often exhibits two flat signatures, SBS3 and SBS5, which are more difficult to distinguish and detect simultaneously. Additionally, cancer types with high TMB (e.g., lung squamous cell carcinoma) have a greater probability of detecting mutational signatures than those with lower TMB (e.g., thyroid adenocarcinoma). Our GLMM analysis found that, for a given cancer type, the odds ratio (OR) of signature detection is 0.962 (95% confidence interval (CI) = 0.956-0.967) for a 0.01 increase of the Shannon equitability index, 1.12 (95% CI = 1.10-1.13) for a one percent increase of signature prevalence, and 1.21 (95% CI = 1.14-1.27) for 1Mb increase in gene panel size, respectively (Fig. 2d). These findings support our hypothesis that in a given cancer type, spikier and more prevalent signatures are more likely to be detected, particularly if using large gene

panels. This is further supported by our findings on pseudo-targeted sequencing data based on WGS (Supplementary Fig. 3b).

Next, we evaluated the signature refitting steps of SATS by estimating the mutations attributed to a given signature, namely the signature expectancy. Specifically, we compared the signature expectancy calculated based on WES[1] with that calculated using SATS based on pseudo-targeted sequencing data from the same tumors (as an example, see Supplementary Fig. 4a for SBS4 in lung cancer based on the MSK-IMPACT468 panel). A high correlation between the two would indicate that using targeted sequencing panels to evaluate the mutations attributed to a given signature would be a viable alternative to using WES. We found that the median Pearson correlation coefficient was 0.7 for panels with sizes greater than 1Mb (Fig. 2e), with higher correlation coefficients for specific signatures, such as SBS4 in lung adenocarcinoma (Pearson correlation coefficient r = 0.91) and SBS7a and SBS7b in melanoma (r = 0.98 and 0.95 respectively, Supplementary Fig. 4b, c). Our analyses show that using panels with large sizes (at least 1Mb) and focusing on signatures with a low Shannon equitability index (such as spiky signatures like SBS4 or SBS7a/b) result in a high Pearson correlation coefficient between targeted sequencing and WES for mutational signatures (Supplementary Fig. 4d). Similar results were observed for pseudo-targeted sequencing data based on WGS (Supplementary Fig. 3c).

Impact of sample sizes on SATS signature detection

To assess the effect of sample size on mutational signature detection, we conducted *in silico* simulations with varying sample sizes. We used breast cancer as the exploratory example (consisting of 12 mutational signatures with at least 1% prevalence in the TCGA breast cancer study, Fig. 3a) and simulated the mutation burden of 96 SBS mutation types for up to 1 million tumors using 21 different targeted sequencing panels with a panel size larger than 1Mb (Methods). This allowed us to use the "truly" present mutational signatures in the *in silico* simulations as benchmarks for evaluation.

We found that identifying spiky and common signatures, such as SBS1 and SBS2/13, only requires few thousand targeted sequenced tumors (Fig. 3b), while detecting less spiky or less common signatures needs more samples (e.g., SBS10a, Supplementary Fig. 5a). The flattest signatures SBS3 and SBS5 require a much larger number of samples, approximately 40,000 and 80,000 samples, respectively, to be detected by all panels (Fig. 3b). Furthermore, we found that the detection probability of signature SBS44 unexpectedly started decreasing after 10,000 samples, which coincided with an increasing detection probability of signature SBS5. This indicates that when two flat signatures, SBS3 and SBS5, are detected, another relatively flat signature, SBS44, becomes difficult to detect. This observation is consistent with previous findings on mutational signature analysis of WGS data, which showed that signatures with flat profiles are likely to be misidentified as other flat signatures[9]. The remaining signatures with a prevalence of less than 5% are unlikely to be detected even with a large number of samples (Fig. 3b), as it is for the current algorithms based on WGS and WES data[10]. Notably, the probability of detecting false positive signatures decreases from 0.35 at 10,000 samples to less than 0.01 at 200,000 samples (Supplementary Fig. 5b). Our findings indicate that with a large number of targeted sequenced tumors, SATS can detect almost all common mutational signatures (>5% prevalence) in breast cancer, including the HRD-associated signature SBS3, and effectively limit the false positive rate. Additionally, this study highlights the importance of considering sample

size when analyzing mutational signatures in targeted sequenced tumors, as the detection of certain signatures may require a large sample size.

Validation of SATS signature detection and refitting by *in silico* simulations

We conducted *in silico* simulations to validate the performance of SATS in mutational signature detection. Specifically, we used the signature profiles and the distributions of signature activities obtained from the AACR Project GENIE to simulate the mutation type matrices for lung, breast, colorectal, and ovarian cancers (see Methods). We compared signatures detected by SATS with the 'prespecified signatures' used in the simulations. When multiple flat signatures were present (e.g., SBS3/5 in lung cancer and SBS3/5/40 in ovarian cancer), we combined them into a single flat signature as the sample size of the AACR Project GENIE is insufficient to accurately distinguish between these flat signatures.

First, our simulations showed that SATS is able to accurately detect most of prespecified signatures, with the exception of few flat or rare signatures. In lung and breast cancers, SATS identified eight of the nine prespecified signatures in all 10 replicates, except for SBS89 in lung cancer, which was identified in 7 out of 10 replicates (Fig. 4a). For colorectal cancer, SATS detected five of the seven prespecified signatures, with SBS6 and SBS44 being identified in 8 and 2 out of 10 replicates, respectively. SBS44, being the second flattest signature in colorectal cancer, is challenging to distinguish from the other common flat signature, SBS5. In ovarian cancer, two signatures (SBS1: 8 of 10 replicates; SBS10c: 3 of 10 replicates) were occasionally identified, while five other signatures were detected in all replicates. SBS10c is relatively flat and rare in ovarian cancer. Notably, SATS detected only one false positive signature (SBS6 in ovarian cancer), which was no longer detected when a larger sample size was used (Supplementary Fig. 6a and b). These results suggest that SATS can effectively identify most prespecified signatures with a low rate of false positive detection.

Next, we carried out signature refitting steps of SATS and compared the estimated signature expectancies with the simulated ones which served as the 'ground truth' for our analysis. Our findings indicate that SATS can accurately estimate the expectancies of most spiky or common signatures, such as SBS2/13 in breast cancer (Pearson correlation coefficient $r = 0.96$, Fig. 4b), SBS4 in lung cancer ($r = 0.90$), SBS10a and SBS10b in colorectal cancer ($r = 0.99$ for both), and SBS92 in ovarian cancer ($r = 0.93$). However, the correlation is lower for flatter or rarer signatures, such as SBS89 in breast cancer ($r = 0.53$), SBS6 in colorectal cancer ($r = 0.44$), and SBS87 in ovarian cancer ($r = 0.61$). Overall, these results suggest that SATS is able to accurately estimate signature expectancies for the majority of prespecified signatures, but many have lower accuracy of estimated signature expectancy with flatter or rarer signatures.

Finally, we investigated the impact of including irrelevant signatures on signature refitting. We simulated breast cancer targeted sequencing data using signatures SBS1, SBS2/13, and SBS5 and performed signature refitting using 12 signatures (including the three true signatures) in the TCGA WES breast cancer study. We found that a significant proportion of mutations were incorrectly attributed to the non-existent signatures in the simulated data (Supplementary Fig. 7). This observation underscores the importance of selecting an appropriate signature list and supports the use of a targeted sequencing-based catalog of mutational signatures as a reference for refitting in targeted sequenced tumors.

Given the set of signatures present in a particular cancer type, SATS can be used for signature refitting to a small number of tumor samples or even a single tumor sample (Methods). To demonstrate this, we estimated signature activities of lung cancers in *in silico* simulations for a subset of samples at a time (Supplementary Note) and found that the signature expectancies are consistent between the estimated and simulated ones regardless of the number of samples used (Fig. 4c). Thus, SATS provides a useful tool for analyzing individual tumors in clinical settings, by reliably estimating the signature expectancy for a set or even a single sample, based on a set of known signatures from targeted sequencing data.

The pan-cancer repertoire of targeted sequencing-based mutational signatures

We used SATS to create a pan-cancer repertoire of SBS signatures based on the targeted sequenced tumors from the AACR Project GENIE. The repertoire can serve as a valuable reference set for signature refitting even in a single sample. We observed that while SBS1, caused by deamination of 5-methylcytosine to thymine, and the flat signatures (SBS3, SBS5, and SBS40 combined, given the current sample size of a cancer type being insufficient to separate them accurately) are universally present in all cancer types, other SBS signatures are specific to certain cancer types (as shown in Fig. 5a bottom panel). These include signatures associated with endogenous mutational processes, such as SBS2/13 caused by APOBEC cytosine deaminases in 7 cancer types, and signatures associated with environmental exposures such as smoking (e.g., SBS4 in lung cancer) and UV radiation (e.g., SBS7a/b in head and neck cancer, skin cancer/melanoma or soft tissue cancer). We also found signatures associated with DNA repair deficiency, such as SBS6/14/15/44 caused by mismatch repair (MMR) deficiency in seven cancer types and SBS10a/b/c resulting from polymerase epsilon (POLE) exonuclease domain mutations in eight cancer types. Additionally, we detected signatures associated with treatment, such as SBS11 caused by temozolomide in glioma and pancreatic cancers. Temozolomide is a common chemotherapeutic agent used in the treatment of glioma[25] and advanced pancreatic neuroendocrine tumors[26]. Finally, we discovered that cancers of unknown primary are enriched with the UV-induced signatures SBS7a/b and clustered with other tumors with UV-induced signatures (such as skin cancer or Melanoma, Supplementary Fig. 8), suggesting potential primary sites for these tumors.

We calculated the signature expectancies for individual cancer types (Methods). The majority of cancer types have a substantial proportion of mutations attributed to flat signatures (SBS3/5/40), except for a few (Fig. 5a top panel). For example, skin cancer or melanoma is primarily influenced by UV-induced signatures, glioma and endometrial cancers are dominated by signatures related to DNA repair deficiency and treatment, bladder cancer is dominated by APOBEC-induced signatures, and lung cancer is dominated by smoking- and APOBEC-induced signatures.

By analyzing a large number of cancer subtypes from clinics, we discovered unexpected SBS signatures in several cancer types. For example, we detected SBS92, a smoking-associated signature, in targeted sequenced ovarian cancers of AACR Project GENIE. This signature is not present in ovarian cancers from the TCGA or Pan-Cancer Analysis of Whole Genomes (PCAWG) studies that focus on the most common subtype, serous ovarian cancer (SOC). In fact, when we restricted the analysis to SOC in the AACR Project GENIE, the SBS92 signature was no longer detected. This suggests that SBS92 could be present in ovarian cancer subtypes other

than SOC. As a confirmation, an unpublished manuscript[27] reports that the signature SBS92 is depleted in low-grade SOC but enriched in Clear Cell Ovarian Cancer. We identified another signature, SBS32, in endometrial and pancreatic cancers from the AACR Project GENIE. The SBS32 signature is associated with chronic exposure to the immunosuppressive drug azathioprine, which is used after organ transplant or for treating diseases related to the immune system (e.g., multiple sclerosis). Although the SBS32 signature has been reported in skin cancers[28], it has not been previously identified in endometrial or pancreatic cancers. However, there are case reports of endometrial cancers in patients treated with azathioprine for long-term immunosuppression after organ transplant[29,30]. Moreover, a link between azathioprine use and acute pancreatitis[31,32], a known risk factor for pancreatic cancer, has also been reported[33,34].

Besides SBS signatures, we also generated a pan-cancer repertoire of DBS mutational signatures for targeted sequenced tumors. We found seven DBS signatures (Fig. 5b bottom panel) with a low mutation burden (less than one mutation per megabase, Fig. 5b top panel). We observed that the DBS1 signature, associated with UV exposure, is present in head and neck cancer, skin cancer or melanoma, soft tissue cancer, and cancers of unknown primary, which is consistent with the presence of UV exposure SBS signatures in these cancer types. Furthermore, the DBS2 signature, which is associated with smoking, was identified in bladder and lung cancer. We also observed DNA repair deficiency-associated signatures DBS3 and DBS10 in non-colorectal bowel cancer.

Discussion

We introduce SATS, a new tool to analyze mutational signatures in targeted sequenced tumors. By analyzing a large number of targeted sequenced tumors, SATS identifies mutational signatures and estimates their contributions to each sample. We validated our approach using *in silico* simulations and pseudo-targeted sequencing data. Our findings indicate that spiky signature profiles, a high signature prevalence, and large sequencing panels (> 1Mb) increase the accuracy of signature extraction and refitting. Importantly, our simulation studies on breast cancer reveal that as the number of targeted sequenced tumors increases, most common signatures can be detected with very low false discovery rates, including the therapeutically important but hard-to-detect HR deficiency-associated signature SBS3. Importantly, using a repertoire of targeted sequencing-based signatures, SATS can be used to identify signatures even in a single sample.

We utilized SATS to analyze over 100,000 targeted sequenced tumors from 24 cancer types in the AACR Project GENIE and developed a pan-cancer catalog of SBS and DBS signatures for targeted sequencing tumors. Most of the identified signatures are related to environmental exposures, treatments, or DNA repair deficiencies. Additionally, we observed unexpected occurrences of well-known mutational signatures in certain cancer types, such as the presence of a smoking-associated signature SBS92 in ovarian cancer. This highlights the importance of incorporating diverse cancer subtypes in mutational signature analysis and the potential for discovering previously unknown associations between mutational signatures and certain cancer types.

SATS and the catalog of mutational signatures generated in this study are valuable tools for analyzing mutational signatures in targeted sequenced tumors. First, SATS accounts for panel

size in its analysis, which enables the identification of signatures of tumor mutation burden that are independent of the type of targeted gene panels. This feature is particularly important in clinical settings as well as in genetic epidemiology studies where different targeted gene panels are commonly used. Second, SATS can estimate signature expectancies even with small sample sizes by running the signature refitting steps. Users can estimate signature expectancies using the catalog of mutational signatures from this study as a reference, enabling the analysis of specific mutational signatures in individual tumors. Third, unlike clustering-based methods[19,21] that require the input of WES/WGS data of individual tumors for their algorithms, SATS is more flexible and can identify multiple mutational signatures in a single analysis, providing a more comprehensive understanding of the mutational landscape of target-sequenced tumors. Finally, the established catalog of mutational signatures is based on targeted sequencing of tumors collected from multiple hospitals and cancer centers, making it more applicable to clinical settings than catalogs of mutational signatures developed in research settings based on WES/WGS.

This study has several limitations that should be considered. First, the data used in this study were collected from clinics in the United States and Western Europe as part of the AACR Project GENIE and may not be representative of targeted sequenced tumors from other regions. Second, although the identified catalog of mutational signatures for targeted sequencing tumors is extensive, it may not be comprehensive as it mainly includes signatures with spiky profiles that are easy to detect under the current sample size per cancer type in the AACR Project GENIE. This means that less frequent or flatter signatures may have been missed. Third, mutational signatures identified by SATS have been previously reported based on WGS, although not necessarily in the same cancer type. Further research is needed to develop analysis methods specific to targeted sequencing data that can uncover novel mutational signatures. Finally, the sample size per cancer type in the AACR Project GENIE is not large enough to accurately distinguish flat mutational signatures, such as separating the HRD-associated signature SBS3 from other flat signatures (SBS5 and SBS40).

To overcome these limitations, it is crucial to increase the number of targeted sequenced tumors and to share the resulting data. With the cost of targeted sequencing decreasing and its use becoming more widespread in clinics, it is feasible to achieve a large sample size. The AACR Project GENIE is already taking steps towards this goal by collecting and sharing more targeted sequencing data and inviting new participants from underrepresented and underserved populations.

In summary, we have introduced a tool for analyzing mutational signatures and created a pan-cancer catalog of mutational signatures specifically for targeted sequenced tumors. The SATS R package is publicly available on GitHub. We anticipate that SATS and the catalog will facilitate the clinical use of mutational signatures for diagnosis and treatment based on targeted sequencing as well as in epidemiological studies with targeted sequence data from different centers.

Methods

Genomic data of AACR Project GENIE

We retrieved the AACR Project GENIE dataset (version: 11.0-public) from Synapse (https://synapse.org/genie). This dataset includes 119,551 tumors that were collected as part of routine clinical practice at 18 hospitals or cancer centers and sequenced by targeted sequencing using different gene panels. The patients provided their consent, and the study was approved by an institutional review board (IRB). The dataset contains samples from diverse ethnic backgrounds, including 5,098 Asians (5.1%); 5,557 Blacks (5.5%); 71,858 Whites (71.5%); 2,725 individuals from other racial groups (2.7%); and 15,239 individuals with unknown race (15.2%). The dataset covers 107 cancer types with 741 subtypes defined by OncoTree[35] (please refer to the Supplementary Note for more information on cancer types in the AACR Project GENIE). To facilitate our analysis, we grouped the cancer types into 24 categories.

The tumors were sequenced at CLIA-/ISO-certified labs with high read depth (median: 473 reads, 1st quantile: 267, 3rd quantile: 764). Somatic mutations were called at participating centers by filtering out germline variants and artifacts using pooled external controls and databases of known germline variants, such as the Genome Aggregation Database (gnomAD)[36]. For more information on the filtering process, please refer to the "AACR GENIE 11.0-public Data Guide" (https://www.synapse.org/#!Synapse:syn26706786). The resulting dataset includes 1,065,807 somatic mutations, of which 809,429 single base substitutions (SBS), 11,688 double base substitutions (DBS), 37,152 insertions (INSs), and 102,437 deletions (DELs). We further removed somatic mutations with a read depth of less than 100 or a reference or alternative allele read count of less than 5. This resulted in a total of 737,856 SBS and 10,390 DBS from 100,477 tumors (Supplementary Table 1) for mutational signature analysis.

A Poisson NMF model for signature analysis of tumor mutation burden

We define a Poisson Non-Negative Matrix Factorization (pNMF) model for SATS, where v_{pn}, ℓ_{pn}, w_{pk} and h_{kn} denote the elements of the matrices $\mathbf{V} = [v_{pn}]$, $\mathbf{L} = [\ell_{pn}]$, $\mathbf{W} = [w_{pk}]$ and $\mathbf{H} = [h_{kn}]$ for the k^{th} mutational signature, the p^{th} SBS type, and the n^{th} targeted sequencing tumor. We assume that v_{pn} follows a Poisson distribution with the expectation $E(v_{pn}) = \sum_{k=1}^{K} e_{kpn} = \sum_{k=1}^{K} \ell_{pn} w_{pk} h_{kn}$, where e_{kpn} is the expected number of mutations attributed to the k^{th} mutational signature, $k = 1,2,\dots,K$. This model specification is equivalent to the one used in signeR[11].

With these notations, the log-likelihood function of pNMF can be written as

$$\log\{P(\mathbf{V}|\mathbf{L},\mathbf{W},\mathbf{H})\} = \sum_{n=1}^{N}\sum_{p=1}^{96} \log\left\{e^{-\ell_{pn}\sum_{k=1}^{K} w_{pk}h_{kn}} \times \frac{\left(\ell_{pn}\sum_{k=1}^{K} w_{pk}h_{kn}\right)^{v_{pn}}}{v_{pn}!}\right\}$$

$$= \sum_{n=1}^{N}\sum_{p=1}^{96}\left\{v_{pn}\log\left(\ell_{pn}\sum_{k=1}^{K} w_{pk}h_{kn}\right) - \ell_{pn}\sum_{k=1}^{K} w_{pk}h_{kn} - \log(v_{pn}!)\right\}.$$

Extraction of *de novo* TMB signatures

We utilize signeR to extract *de novo* signatures $\widehat{\mathbf{W}}$ based on pNMF model. However, because signeR is computationally demanding due to its use of the Markov Chain Monte Carlo (MCMC) method[11], we grouped samples to improve computational efficiency. Our results below reveal that grouping samples does not affect the TMB signatures profile (w_{pk}). Specifically, we define C as the set $[1,2,\dots,N]$, and C_m as the mutually exclusive set such that $C = \bigcup_{m=1}^{M} C_m$. For $v_{pm}^{\#} = \sum_{n\in C_m} v_{pn}$, the sum of the mutation count for the targeted sequencing tumors with index n belonging to the set C_m, we can show that:

$$E\left(v_{pm}^{\#}\right) = \sum_{k=1}^{K}\sum_{n\in C_m} e_{kpn} = \sum_{k=1}^{K} l_{pm}^{\#} w_{pk} h_{km}^{\#},$$

where $l_{pm}^{\#} = \sum_{n\in C_m} \ell_{pn}$ and $h_{km}^{\#} = \frac{\sum_{n\in C_m} \ell_{pn} h_{kn}}{\sum_{n\in C_m} \ell_{pn}}$. Notably, the TMB signature profile w_{pk} remains unchanged. The panel size of combined samples $l_{pm}^{\#}$ is the sum of the panel size of individual samples ℓ_{pn} , and signature activity $h_{km}^{\#}$ is the weighted sum of the signature activities of individual samples h_{kn}. The mutation count of combined samples $v_{pm}^{\#}$ follows a Poisson distribution, as the sum of independent Poisson counts is still Poisson distributed.

Grouping samples can significantly reduce computation time. For example, when analyzing ten thousand samples using SATS, analysis with grouping 100 tumors together can be completed in 28.5 minutes on a laptop with a Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz processor and 16 GB of 4267 MHz RAM. In contrast, analyzing the same set of samples without grouping tumors takes approximately 13 hours.

Mapping *de novo* TMB signatures to COSMIC catalog TMB signatures

Due to the limited number of somatic mutations detected by targeted gene panels, the detected *de novo* TMB signature profiles may be a linear combination of COSMIC catalog TMB signature profiles. To address this limitation, we map the *de novo* signature profile matrix $\widehat{\mathbf{W}} = [\widehat{w}_{pk}]$ to catalog TMB signatures $\mathbf{W}_0$ (e.g., a 96×69 COSMIC TMB signature profile matrix for 69 catalog SBS TMB signatures) by using penalized non-negative least squares[22]:

$$\min_{\boldsymbol{\beta}} \left\|\widehat{\mathbf{W}} - \mathbf{W}_0\boldsymbol{\beta}\right\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \text{ subject to } \boldsymbol{\beta} > 0 \text{ and } \lambda \geq 0,$$

where $\boldsymbol{\beta}$ is a coefficient vector and the tuning parameter λ is selected based on cross-validations. Compared with the non-negative least squares,

$$\min_{\boldsymbol{\beta}} \left\|\widehat{\mathbf{W}} - \mathbf{W}_0\boldsymbol{\beta}\right\|_2^2 \text{ subject to } \boldsymbol{\beta} > 0,$$

the penalized non-negative least squares allow us to select a smaller number of catalog signatures with the profile matrix $\mathbf{W}^*$ that have a significant contribution to the *de novo* signature profiles by shrinking small values of $\boldsymbol{\beta}$ towards zero. To reduce the randomness caused by the cross-validation step to select λ, we repeat this process 100 times, and select only catalog TMB signatures with a coefficient β greater than 0.1 in more than 80 of the iterations.

Estimation of signature activities by an expectation-maximization algorithm

We propose an expectation-maximization (EM) algorithm to estimate the signature activity matrix $\mathbf{H} = [h_{kn}]$, given the mutation type matrix $\mathbf{V} = [v_{pn}]$, the panel context matrix $\mathbf{L} = [\ell_{pn}]$, and the mapped catalog TMB signature profiles $\mathbf{W}^* = [w^*_{pk}]$. The element v_{pn} in $\mathbf{V}$ can be expressed as the sum of independent latent counts $v_{1pn}, v_{2pn}, \cdots, v_{Kpn}$ attributed to K signatures. These latent counts are treated as the missing data, following Poisson distributions with expectations $\ell_{pn}w^*_{p1}h_{1n}, \ell_{pn}w^*_{p2}h_{2n}, \cdots, \ell_{pn}w^*_{pK}h_{Kn}$, respectively. Introducing latent counts allows us to compute the complete data log-likelihood as:

$$\sum_{p=1}^{96}\sum_{n=1}^{N}\sum_{k=1}^{K}\{-\ell_{pn}w^*_{pk}h_{kn} + v_{kpn}\, log(\ell_{pn}w^*_{pk}h_{kn}) - log(v_{kpn}!)\}.$$

In addition, the conditional distribution of v_{kpn} given $\mathbf{V}$, $\mathbf{L}$, $\mathbf{W}^*$ and $\mathbf{H}^t$ (the $\mathbf{H}$ at the t'th iteration of the EM algorithm) follows a multinomial distribution with parameters v_{pn} and $p_k = w^*_{pk}h^t_{kn}/\sum_{j=1}^{K} w^*_{pj}h^t_{jn}$.

In the E-step, we compute $Q(\mathbf{H}|\mathbf{H}^t)$ as the expected complete data log-likelihood:

$$\begin{aligned} Q(\mathbf{H}|\mathbf{H}^t) &= E\left[\sum_{p=1}^{96}\sum_{n=1}^{N}\sum_{k=1}^{K}\{-\ell_{pn}w^*_{pk}h_{kn} + v_{kpn}\, log(\ell_{pn}w^*_{pk}h_{kn})\}\,|\mathbf{V}, \mathbf{L}, \mathbf{W}^*, \mathbf{H}^t\right] \\ &= \sum_{p=1}^{96}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\{-\ell_{pn}w^*_{pk}h_{kn} + \log(\ell_{pn}w^*_{pk}h_{kn})\, v_{pn}\frac{w^*_{pk}h^t_{kn}}{\sum_{j=1}^{K} w^*_{pj}h^t_{jn}}\right\}. \end{aligned}$$

In the M-step, the maximizer of $Q(\mathbf{H}|\mathbf{H}^t)$ is obtained by setting the derivative with respect to h_{kn} to 0,

$$\frac{\partial}{\partial h_{kn}} Q(\mathbf{H}|\mathbf{H}^t) = -\sum_{p=1}^{96}\ell_{pn}w^*_{pk} + \frac{1}{h_{kn}}\left(\sum_{p=1}^{96} v_{pn}\frac{w^*_{pk}h^t_{kn}}{\sum_{j=1}^{K} w^*_{pj}h^t_{jn}}\right) = 0,$$

and the updated activity value h^{t+1}_{kn} is given by:

$$h^{t+1}_{kn} = h^t_{kn}\frac{\sum_{p=1}^{96} v_{pn}\left(\frac{w^*_{pk}}{\sum_{j=1}^{K} w^*_{pj}h^t_{jn}}\right)}{\sum_{p=1}^{96}\ell_{pn}w^*_{pk}}.$$

Note that the M-step depends on the current value of activities h^t_{jn} for the n^{th} tumor only. Therefore, even though the EM algorithm updates the entire activity matrix $\mathbf{H}$ for all samples simultaneously, it is equivalent to updating the activity of one tumor at a time. In other words, the EM algorithm of SATS for signature refitting estimates signature activities independently of other tumors, enabling signature activities to be estimated accurately for a single tumor or small subset of samples.

To complete the EM algorithm, the E-step and the M-step are iterated until convergence and output the estimated activity matrix $\widehat{\mathbf{H}}$.

Calculation of signature expectancy

To calculate the expected number of mutations attributed to a signature (referred to as the signature expectancy) $\boldsymbol{E} = [E_{kn}]$, we use the estimated activity matrix $\widehat{\mathbf{H}} = [\hat{h}_{kn}]$ from the EM algorithm. The signature expectancy E_{kn} of the k^{th} signature in the n^{th} tumor is then calculated as the sum of the product of the panel size ℓ_{pn}, the catalog signature profile w^*_{pk} and the estimated signature activity $\hat{h}_{kn}$ across 96 SBS types as $E_{kn} = \sum_{p=1}^{96} \ell_{pn} w^*_{pk} \hat{h}_{kn}$.

Relationship to the canonical NMF

The pNMF model proposed here incorporates the panel context matrix **L**, extending the canonical NMF, the standard method for identifying mutational signatures in tumors sequenced by WES or WGS. When all samples are sequenced using the same genomic regions (i.e., $\ell_{pn} = \ell_p$), the log-likelihood function in equation (1) is simplified as

$$\log\{P(\mathbf{V}|\mathbf{L},\mathbf{W},\mathbf{H})\} = -D_{KL}\left(\mathbf{V}\middle|\mathbf{W}',\mathbf{H}\right) + C,$$

where C is a constant irrelevant to **W** and **H**. It is worth noting that

$$D_{KL}\left(\mathbf{V}\middle|\mathbf{W}',\mathbf{H}\right) = \sum_{n=1}^{N}\sum_{p=1}^{96}\left\{v_{pn}\log\left(\frac{v_{pn}}{\sum_{k=1}^{K} w'_{pk}h_{kn}}\right) + \sum_{k=1}^{K} w'_{pk}h_{kn} - v_{pn}\right\} \qquad (2)$$

is equivalent to the objective function of the canonical NMF[37] with $w'_{pk} = \ell_p w_{pk}$.

The $D_{KL}(\mathbf{V}|\mathbf{W}',\mathbf{H})$ in equation (2) with $w'_{pk} = \ell_p w_{pk}$ highlights the relationship between signatures of TMB and TMC: TMB signature profile w_{pk} normalizes TMC signature profile w'_{pk}, by the number of mutation context ℓ_p (i.e., $w_{pk} = w'_{pk}/\ell_p$). This means that we can create a catalog of TMB signatures based on WGS, dividing the catalog of TMC signatures (e.g., COSMIC WGS catalog TMC signatures[1]) by the number of trinucleotide contexts from which the mutation type could occur in the human reference whole genome.

Signatures of tumor mutation burden

To create a catalog of signature profiles of tumor mutation burden (TMB) in Supplementary Table 2, we normalize COSMIC signature profiles of tumor mutation count (TMC) by the size of mutation contexts in the whole genome. This is done by following these steps:

1. Download the COSMIC SBS and DBS signature profiles (version 3.2) from https://cancer.sanger.ac.uk/signatures/
2. For an SBS signature profile, divide the level of each mutation type (e.g., A[C > G]G) by the size of the corresponding mutation context (e.g., ACG for A[C > G]G) that can occur in the whole genome (Supplementary Table 3)
3. Rescale 96 mutation types to sum to one.
4. Similarly, create DBS signature profiles of TMB (Supplementary Table 4) based on COSMIC DBS signature profiles of TMC and the number of genomic contexts (Supplementary Table 5) for which DBS can occur.

Shannon equitability index of TMB mutational signatures

We use Shannon equitability index to measure the diversity or "flatness" of a signature profile[12,38]. The index is calculated as

$$Shannon\ equitability\ index\ = -\frac{\sum_{i=1}^{n} p_i log(p_i)}{\log(n)},$$

where p_i is the level of signature profile at the ith mutation type and the sum across all mutation types (i=1 to n) is equal to 1.

A higher value of the index indicates a more even distribution of mutation types. The index ranges from 0 to 1, with a value of 1 indicating a completely flat signature profile where all mutation types are represented equally and a value of 0 indicating a signature profile with a single dominant "spike" where a single mutation type has a proportion of 1 and all other mutation types have a proportion of 0. Among SBS TMB signatures with $n = 96$, some profiles are characterized by a few specific mutation types at high levels, referred to as "spikes" (e.g., SBS1 with the C>T substitution at the NCG trinucleotide has a Shannon equitability index of 0.317, and SBS10a with the T[C>A]T substitution has a Shannon equitability index of 0.192). Other signature profiles are more evenly distributed across all mutation types, referred to as "flat" (e.g., SBS3 has a Shannon equitability index of 0.974, SBS5 has a Shannon equitability index of 0.903, and SBS40 has a Shannon equitability index of 0.969)

Generation of pseudo-targeted sequencing data

To investigate the impact of various factors on mutational signature detection, we create pseudo-targeted sequencing datasets using sequencing data from two sources respectively: the TCGA WES studies[1,15] and the Sanger breast cancer (BRCA) 560 WGS study[16]. We assume that targeted sequencing would identify SBS that are identified in the WES or WGS studies, as long as SBSs are located within the targeted genomic regions of the panels. This assumption is reasonable since targeted sequencing typically provides much higher coverage than WES or WGS. The steps to generate the simulated data are outlined below:

1. Download the TCGA WES data (mc3.v0.2.8.PUBLIC) from the Cancer Genome Data Portal (https://gdc.cancer.gov/about-data/publications/mc3-2017) and WGS data of Sanger BRCA560 study (Caveman_560_20Nov14_clean) from ftp://ftp.sanger.ac.uk/pub/cancer/Nik-ZainalEtAl-560BreastGenomes.
2. Download the genomic information file of AACR Project GENIE (https://www.synapse.org/#!Synapse:syn26706790) which specifies the chromosome, start position, and end position of genomic regions for each targeted sequencing panel.
3. For SBS in WES or WGS studies, select those located in the genomic regions of a targeted sequencing panel to create the SBS mutation type matrix as pseudo-targeted sequencing data. We generated 648 pseudo-targeted sequencing datasets, encompassing 18 TCGA WES cancer types and 36 targeted sequencing panels (panel size: 0.05 Mb to 9.95 Mb). In addition, we generated 36 pseudo-targeted sequencing datasets based on 560 breast tumors with WGS data, using 36 targeted gene panels.

Analysis of pseudo-targeted sequencing data

We calculate the signature detection probability, which represents the percentage of common signatures detected by 36 targeted sequencing panels within a cancer type. Next, we employ a generalized linear mixed model (GLMM) to analyze the factors that influence the detection probability of TMB mutational signatures across 648 pseudo-targeted sequencing datasets (by 18 TCGA cancer types and 36 targeted sequencing panels). The GLMM incorporates several fixed effects, including the flatness of the signature profile (quantified using the Shannon equitability index), the prevalence of the mutational signature in the TCGA WES study (as a percentage of SBS attributed to the signature), and the panel size (per megabase). To account for any variation in the results due to the different cancer types under investigation, we include cancer type as a random intercept in the model.

Evaluation of the impact of sample sizes

We conducted an *in silico* simulation to investigate the effect of sample size on the ability to detect mutational signatures in breast cancer. The simulation was executed using varying sample sizes, ranging from one thousand to up to one million samples, based on the panel context matrix, signatures profile matrix (consisting of 12 mutational signatures with at least 1% prevalence in the TCGA breast cancer study), and signature activity matrix (following the distributions of the signature activity matrix of the TCGA breast cancer study).

1. We run signature refitting on the TCGA breast cancer dataset (accessible at https://www.synapse.org/#!Synapse:syn11726618), using12 known mutational signatures (SBS1, 2, 3, 5, 7a, 10a, 10b, 13, 15, 29, 30, 44 and 58) that have a prevalence greater than 1% (based on https://www.synapse.org/#!Synapse:syn11801497). Specifically, we applied the EM algorithm to estimate the signature activity matrix $\mathbf{H}_B^*$, from the mutation type matrix $\mathbf{V}_B$, the panel context matrix $\mathbf{L}_{WES}$ of the whole exome sequencing, and the pre-defined TMB signature matrix $\mathbf{W}_B^*$.
2. We simulated mutation type matrix $\mathbf{V}_B^{sim}$ for 21 targeted sequencing panels with a panel size larger than 1Mb, using a range of sample sizes from 1000 tumors to 1 million tumors. Specifically, we simulated mutation type matrix $\mathbf{V}_B^{sim}$ from a Poisson distribution with the mean $\mathbf{L}_S \circ \mathbf{W}_B^* \mathbf{H}_B^b$, where $\mathbf{L}_S$ represents the panel size matrix for a given targeted sequencing panel (S), $\mathbf{H}_B^b$ is sampled from the estimated signature activity matrix $\mathbf{H}_B^*$. As the activities of APOBEC signatures SBS2 and SBS13 are highly correlated, their activities were jointly sampled. Finally, we excluded any tumors with zero mutation count.
3. We applied signeR to extract *de novo* signatures $\widehat{\mathbf{W}}_B^S$ from the simulated mutation type matrix $\mathbf{V}_B^{sim}$. Then, we employed penalized non-negative least squares to select the mapped catalog TMB signatures, $\mathbf{W}_B^{S*}$. Finally, we estimated signature activities and expectancies using the EM algorithm.
4. To evaluate the ability to detect the pre-specified signatures, we analyzed the proportion of 21 panels that were able to rediscover the prespecified 12 mutational signatures using SATS. We also tracked the probability of detecting false positive signatures that were not used to simulate mutation counts.

Validation of SATS by *in silico* simulations

We conduct *in silico* simulations to evaluate whether SATS can detect and estimate the prespecified signatures in simulated datasets.

1. We first calculate the expectation matrix $\mathbf{E}_c^*$ as $\mathbf{E}_c^* = \mathbf{L}_c \circ \mathbf{W}_c^* \mathbf{H}_c^*$, where $\circ$ denotes element-wise product, $\mathbf{L}_c$ a panel size matrix, $\mathbf{W}_c^*$ a signature profile matrix and $\mathbf{H}_c^*$ a signature activity matrix of a cancer type (c). The matrices $\mathbf{W}_c^*$ and $\mathbf{H}_c^*$ are estimated from the AACR Project GENIE, allowing us to generate simulated data that accurately reflects actual observations.
2. We generate ten replicates of the mutation type matrix $\mathbf{V}_c^{sim}$ for lung cancer, breast cancer, colorectal cancer, and ovarian cancer respectively by simulating data from the Poisson distribution using expectation matrix $\mathbf{E}_c^*$. The number of simulated samples is the same as in the corresponding AACR Project GENIE studies. In addition, for ovarian cancer, we perform bootstrapping on the $\mathbf{H}_c^*$ to simulate more samples.
3. We apply signeR and penalized non-negative least squares to estimate TMB signatures $\mathbf{W}_c^{est}$ for each simulated mutation type matrix $\mathbf{V}_c^{sim}$. We then compare these estimated signatures with the ground truth signatures $\mathbf{W}_c^*$.
4. Using the simulated mutation type matrices ($\mathbf{V}_c^{sim}$), panel size matrices ($\mathbf{L}_c$), and estimated TMB signatures ($\mathbf{W}_c^{est}$), we estimate the signature activity matrix ($\boldsymbol{H}_c^{est}$) for all tumors using the EM algorithm. We then calculate the signature expectancy based on $\boldsymbol{H}_c^{est}$, which is compared with the simulated signature expectancy as the ground truth. For lung cancer, we also estimate $\boldsymbol{H}_c^{est}$ for a subset of samples or even for one sample, as detailed in the Supplementary Note.

Software

The R package SATS is publicly available at https://github.com/binzhulab/SATS.

Acknowledgment

This research was conducted with support from Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics (DCEG). Additionally, Dr. Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No. RS-2023-00213625). The authors would like to acknowledge the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of study authors. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD: https://biowulf.nih.gov. We would like to thank Bill Wheeler (Information Management Services) for computation support.

Figure legends:

Fig. 1 A schematic workflow of SATS. The workflow starts with summarizing somatic mutations, such as single base substitutions (SBS), into a mutation type matrix **V**. This illustrative example shows 16 of 96 SBS types with C to T mutations of tumors from 6 subjects (Sub1, Sub2, …, Sub6) are shown. SATS then performs the following four steps: a) signeR is utilized to identify *de novo* tumor mutation burden (TMB) signatures while adjusting for panel

sizes. The profiles of SBS TMB signatures are displayed, with the 16 C to T SBS types highlighted in red; b) The *de novo* TMB signatures are mapped to TMB signatures from a WES or WGS catalog (namely catalog TMB signatures) through penalized non-negative least squares (pNNLS); c) Signature activities are estimated using the proposed Expectation-Maximization (EM) algorithm with input from the mutation type matrix **V**, panel context matrix **L**, and signature profile matrix **W** of catalog TMB signatures; d) The expected number of mutations attributed to each catalog TMB signature, known as signature expectancy, is calculated for each targeted sequenced tumor.

Fig. 2 Determinants of tumor mutation burden signature detection. a. Percentage of common signatures detected by 36 targeted sequencing panels (i.e., detection probability) across 18 TCGA cancer types. Common signatures contribute at least 5% of single base substitutions for a cancer type in the TCGA WES study. The number in parentheses on the y-axis indicates the median sample size of pseudo-panel data across 36 panels. The number in parentheses on the x-axis shows the size of the genomic regions targeted by a sequencing panel. **b.** The percentage of common TCGA WES signatures detected by 36 targeted sequencing panels versus the Shannon equitability index of the signature profile. The blue line refers to the linear regression line. The black dots refer to the undetected signatures. **c.** The proportion of variance in detection probabilities of common TCGA WES signatures that can be explained by determinant factors, including the Shannon equitability index of the signature profile (flatness), the frequency of TCGA WES signatures (prevalence), panel size, and cancer type. **d.** The odds ratio of determinant factors. Bars represent the 95% confidence intervals (CIs). **e.** The median Pearson correlation coefficient for all detected common WES signatures across 18 TCGA cancer types was compared to panel size. The Pearson correlation coefficient is a measure of the correlation between the number of single base substitutions (SBS) attributed to a signature, estimated from the pseudo-panel data, and the number of SBS attributed to the same signature previously reported in the TCGA WES study. The blue curve represents LOWESS (Locally Weighted Scatterplot Smoothing) curve, and the shaded area is 95% CIs. CA: cervical cancer; HCC: hepatocellular carcinoma; RCC: renal cell carcinoma; CNS: central nervous system; GBM: glioblastoma; AdenoCa: adenocarcinoma; SCC: squamous cell carcinoma; Thy: thyroid; Prost: prostate; Colorect: colon or rectum; DLBC: diffuse large B cell lymphoma.

Fig. 3 Impact of sample sizes on mutational signature detection. a. The scatterplot of the flatness (measured by Shannon equitability index) of signature profiles and percentage of single base substitutions (SBS) attributed to the signatures in the TCGA breast cancer (BRCA) WES study. The reference line with Shannon equitability index one refers to the theoretical maxima (by a completely flat signature). **b.** The probability of signature detection. Each dot represents the probability of signature detection, measuring the proportion of 21 targeted sequencing panels (with panel size larger than 1Mb) that can identify the signature at a given sample size. The blue line is the LOWESS (Locally Weighted Scatterplot Smoothing) curve, and the shaded area represents 95% confidence intervals (CIs).

Fig. 4 Validation of TMB signature analysis. a. Frequency of signatures detected in 10 replicates for lung, breast, colorectal, and ovarian cancers, respectively. These signatures are used to simulate mutation counts and are considered as ground truth. The x-axis represents the proportion of SBS attributed to a signature in the simulations. The y-axis describes the flatness of

the signature profile (measured by its Shannon equitability index). The dot size is proportional to the detection rate of mutational signatures. **b.** Pearson correlation coefficient between the simulated SBS signature expectancies (as the benchmark) and the estimated ones for four cancer types. The bars represent the mean of the Pearson correlation coefficient, and the intervals are the mean plus or minus one standard deviation. **c.** The Pearson correlation coefficient between simulated and estimated SBS signature expectancies in lung cancer with various sample sizes for signature refitting. The bars in the figure represent the mean Pearson correlation coefficient for simulation replicates, while the x-axis indicates the number of samples used for signature refitting, including 100 samples, 10 samples, or even one sample at a time. The error bars in the figure represent the mean plus or minus one standard deviation.

Fig. 5 Repertoire of mutational signatures in the AACR Project GENIE. a. Single base substitution (SBS) signatures. The top bar chart displays the stacked tumor mutation burden (TMB) attributed to specific signatures, with the colors indicating different mutational signatures. The bottom panel illustrates the presence of SBS signatures for individual cancer types, with dot sizes representing the proportion of tumors in which an SBS signature is present. The sample size of targeted sequenced tumors is indicated between the two panels, and the proposed etiology of the mutational signature is included in parentheses. **b.** Double base substitution (DBS) signatures. The top stacked bar chart shows the TMB of DBS signatures, and the bottom panel shows the proportion of tumors for which a DBS signature is present. 5meC: 5-Methylcytosine; APOBEC: apolipoprotein B mRNA-editing enzyme, catalytic polypeptide, MMR: mismatch repair; UV: ultraviolet radiation; POLE-exo*: mutations in polymerase epsilon exonuclease domain; TMZ: temozolomide; BER: base excision repair; AZA: azathioprine; AID: activation-induced deaminase; TP: thiopurine.

Supplementary figure legends:

Supplementary Fig. 1 Mutation profiles and contexts in the human whole genome. **a.** Profiles of SBS5 signature based on tumor mutation count (TMC) and tumor mutation burden (TMB), respectively, with 96 mutation types on the x-axis and contributions on the y-axis. **b.** 32 mutation contexts for single base substitutions (SBS) in the human whole genome, with significantly depleted ACG, TCG, GCG, and CCG trinucleotides. **c.** Scatterplot of the Shannon equitability index of tumor mutation count (TMC) and tumor mutation burden (TMB) signature profiles. The black line represents the diagonal line, the blue line the linear regression line, and the shaded area is 95% confidence intervals. The dots are annotated when the differences of Shannon equitability index between TMC and TMB signature profiles are more than 0.1. **d.** Profiles of SBS10b and SBS15 signatures based on TMC and TMB, respectively.

Supplementary Fig. 2 Boxplots of mutation context ratios between targeted sequences vs. whole genome sequence. Each dot represents the ratio of the proportion (by the corresponding sequence size) of a mutation context (e.g., CCG) in the genomic regions targeted by a sequencing panel, compared to the proportion of the same mutation context in the whole genome.

Supplementary Fig. 3 Results of pseudo-targeted sequencing data based on Sanger breast cancer 560 WGS study. a. The entropy of signature profiles and the percentage of single base

substitutions (SBS) attributed by a signature for the Sanger breast cancer (BRCA) 560 WGS study. The signatures with > 5% prevalence are highlighted in color, while others are shown in gray. **b.** The odds ratio of the Shannon equitability index of the signature profile, frequency of Sanger BRCA 560 WGS study signatures, and panel size. The bars represent 95% confidence intervals (CIs). **c.** The median of the Pearson correlation coefficient for all common WGS signatures detected. The blue curve represents LOWESS (Locally Weighted Scatterplot Smoothing) curve with the shaded area for 95% CIs.

Supplementary Fig. 4 Results of pseudo-targeted sequencing data based on TCGA WES study. a. Scatterplot of the number of mutations attributed by the smoking-related signature SBS4 (on the y-axis) in the MSK-IMPACT 468 panel compared to the number of mutations attributed by the same signature in the TCGA WES study (on the x-axis). The black line represents the ratio of the MSK-IMPACT 468 panel size to the WES size. The blue line represents the linear regression line and the shaded area represents 95% confidence intervals (CIs). **b.** and **c.** Medians of Pearson correlation coefficients for signatures SBS4 (**b**) and SBS7a/b (**c**) are shown. Pearson correlation coefficient measures the correlation between the number of mutations attributed by a signature using the pseudo-targeted sequencing data and the number of mutations attributed by the same signature reported previously in the TCGA WES study. The blue curve represents LOWESS (Locally Weighted Scatterplot Smoothing) curve, with the shaded area representing 95% CIs. **d.** Scatterplot of Pearson correlation coefficient (larger than 0.1 and highlighted in black) versus Shannon equitability index of TCGA WES breast cancer signature profiles.

Supplementary Fig. 5 Detection probability of TCGA breast cancer signatures in simulations. The probability of signature detection is shown for increasing sample sizes (**a.** up to 10,000 tumors, **b.** up to 1,000,000 tumors). Each dot represents the proportion of targeted sequencing panels (with a panel size larger than 1Mb) that are able to identify the signature at the corresponding sample size. Blue curves are LOWESS (Locally Weighted Scatterplot Smoothing) curves, and shaded areas represent 95% confidence intervals.

Supplementary Fig. 6 Validation of TMB signature analysis for ovarian cancers with 10,190 bootstrapped samples. a. Frequencies of signatures detected in 10 replicates. These signatures simulate mutation counts and are considered as ground truth. The x-axis represents the proportion of SBS attributed by a signature in the simulations. The y-axis describes the flatness of the signature profile (measured by its Shannon equitability index). The dot size is proportional to the detection rate of mutational signatures. **b.** Pearson correlation coefficient between the simulated SBS signature expectancies (as the benchmark) and the estimated ones for ovarian cancers with bootstrapped samples. The bars represent the mean of the Pearson correlation coefficient, and the intervals are the mean plus or minus one standard deviation.

Supplementary Fig. 7 The proportion of tumors in which the mutation signatures were detected in simulated breast cancer data. Mapped signatures SBS1, SBS2/13, and SBS5 were used for simulation. Mapped signatures or WES-based signatures were used for signature refitting, respectively. WES-based signatures include SBS1, SBS2/13, SBS5, SBS3, SBS7a, SBS10a, SBS10b, SBS15, SBS29, SBS30, SBS44 and SBS58, which are present in more than 1% of mutations in TCGA WES breast cancer study. The horizontal lines represent the actual

proportions in the simulated data. The error bars in the figure represent the mean plus or minus one standard deviation.

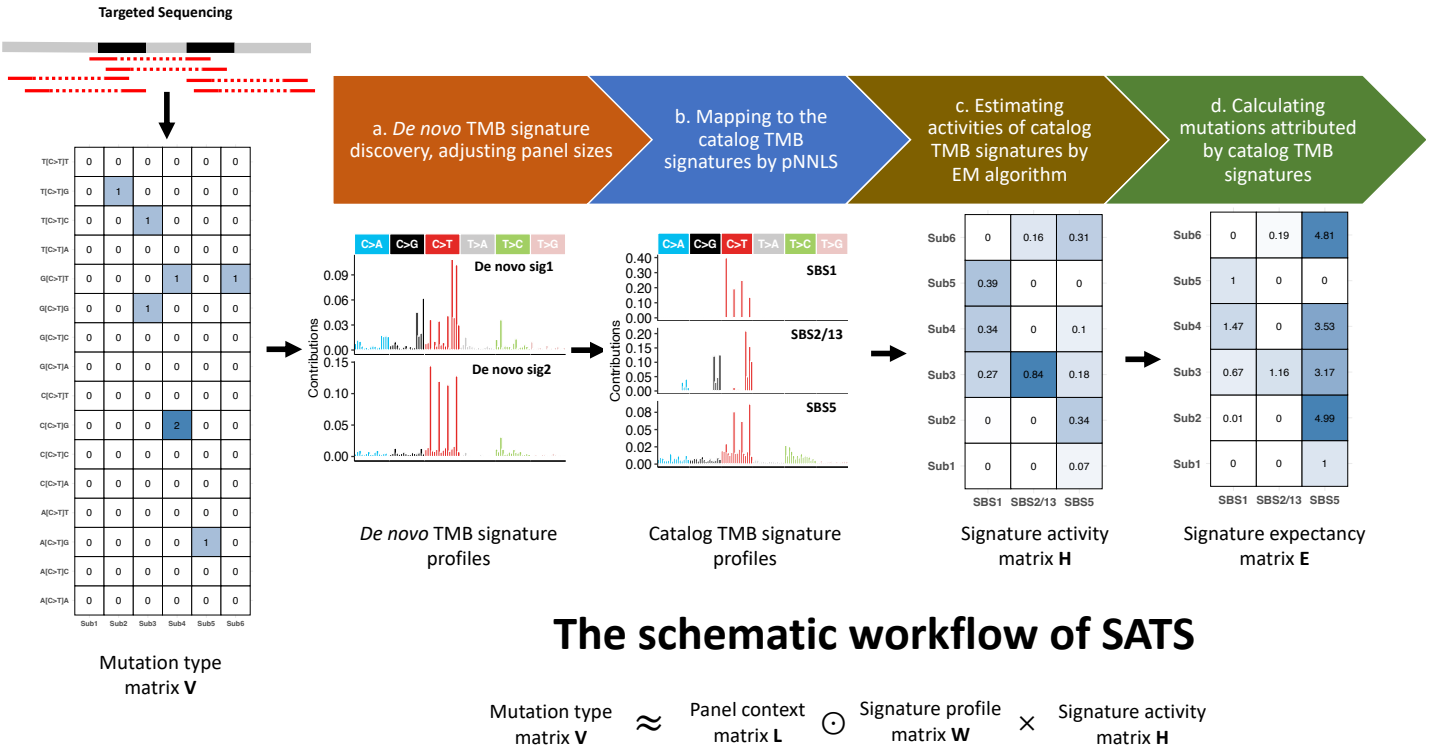
Supplementary Fig. 8 UMAP visualization of estimated signature expectancy. The scatterplots show 2-dimensional projections of the estimated signature expectancy for tumors from Memorial Sloan Kettering Cancer Center (**a**) and Dana-Farber Cancer Institute (**b**) using UMAP (Uniform Manifold Approximation and Projection). Due to the large sample size, the UMAP projections are computed for the medians of clustered tumors. The signature expectancies have been scaled into proportions within each tumor, and tumors have been clustered using hierarchical clustering with Ward's minimum variance method, with a cut-off value of 0.05. Each dot represents a cluster of tumors with similar signature expectancies.

References

1. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
2. Degasperi, A. *et al.* Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science* **376**, abl9283 (2022).
3. Poon, S.L., McPherson, J.R., Tan, P., Teh, B.T. & Rozen, S.G. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome Med* **6**, 24 (2014).
4. Wan, J.C.M. *et al.* Genome-wide mutational signatures in low-coverage whole genome sequencing of cell-free DNA. *Nat Commun* **13**, 4953 (2022).
5. Poon, S.L. *et al.* Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* **5**, 197ra101 (2013).
6. Van Hoeck, A., Tjoonk, N.H., van Boxtel, R. & Cuppen, E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* **19**, 457 (2019).
7. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* **23**, 517-525 (2017).
8. Buisson, R., Lawrence, M.S., Benes, C.H. & Zou, L. APOBEC3A and APOBEC3B Activities Render Cancer Cells Susceptible to ATR Inhibition. *Cancer Res* **77**, 4567-4578 (2017).
9. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer* **21**, 619-637 (2021).
10. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-59 (2013).
11. Rosales, R.A., Drummond, R.D., Valieris, R., Dias-Neto, E. & da Silva, I.T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8-16 (2017).
12. Islam, S.M.A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom* **2**, 100179 (2022).

13. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).
14. Huang, X., Wojtowicz, D. & Przytycka, T.M. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* **34**, 330-337 (2018).
15. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034-1035 (2018).
16. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).
17. Consortium, A.P.G. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov* **7**, 818-831 (2017).
18. Pugh, T.J. *et al.* AACR Project GENIE: 100,000 Cases and Beyond. *Cancer Discov* **12**, 2044-2057 (2022).
19. Gulhan, D.C., Lee, J.J., Melloni, G.E.M., Cortes-Ciriano, I. & Park, P.J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat Genet* **51**, 912-919 (2019).
20. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat Commun* **10**, 2969 (2019).
21. Georgeson, P. *et al.* Identifying colorectal cancer caused by biallelic MUTYH pathogenic variants using tumor mutational signatures. *Nature Communications* **13**, 3254 (2022).
22. Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **73**, 273-282 (2011).
23. Russell, G.J., Walker, P.M., Elton, R.A. & Subak-Sharpe, J.H. Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J Mol Biol* **108**, 1-23 (1976).
24. Bird, A.P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8**, 1499-504 (1980).
25. Stupp, R. *et al.* Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol* **10**, 459-66 (2009).
26. Kunz, P.L. *et al.* Randomized Study of Temozolomide or Temozolomide and Capecitabine in Patients With Advanced Pancreatic Neuroendocrine Tumors (ECOG-ACRIN E2211). *J Clin Oncol* **41**, 1359-1369 (2023).
27. Lee, D. *et al.* The heterogeneity of mutational signatures in 100,477 targeted sequenced tumors. *Unpublished Manuscript* (2023).
28. Inman, G.J. *et al.* The genomic landscape of cutaneous SCC reveals drivers and a novel azathioprine associated mutational signature. *Nat Commun* **9**, 3667 (2018).
29. Kobayashi, T. *et al.* Endometrial Cancer After Pancreas-After-Kidney Transplantation: A Case Report and Review of the Literature. *Transplant Proc* **54**, 560-564 (2022).
30. Hodgkinson, D.J. & Williams, T.J. Endometrial carcinoma associated with azathioprine and cortisone therapy. A case report. *Gynecol Oncol* **5**, 308-12 (1977).
31. Floyd, A., Pedersen, L., Nielsen, G.L., Thorlacius-Ussing, O. & Sorensen, H.T. Risk of acute pancreatitis in users of azathioprine: a population-based case-control study. *Am J Gastroenterol* **98**, 1305-8 (2003).

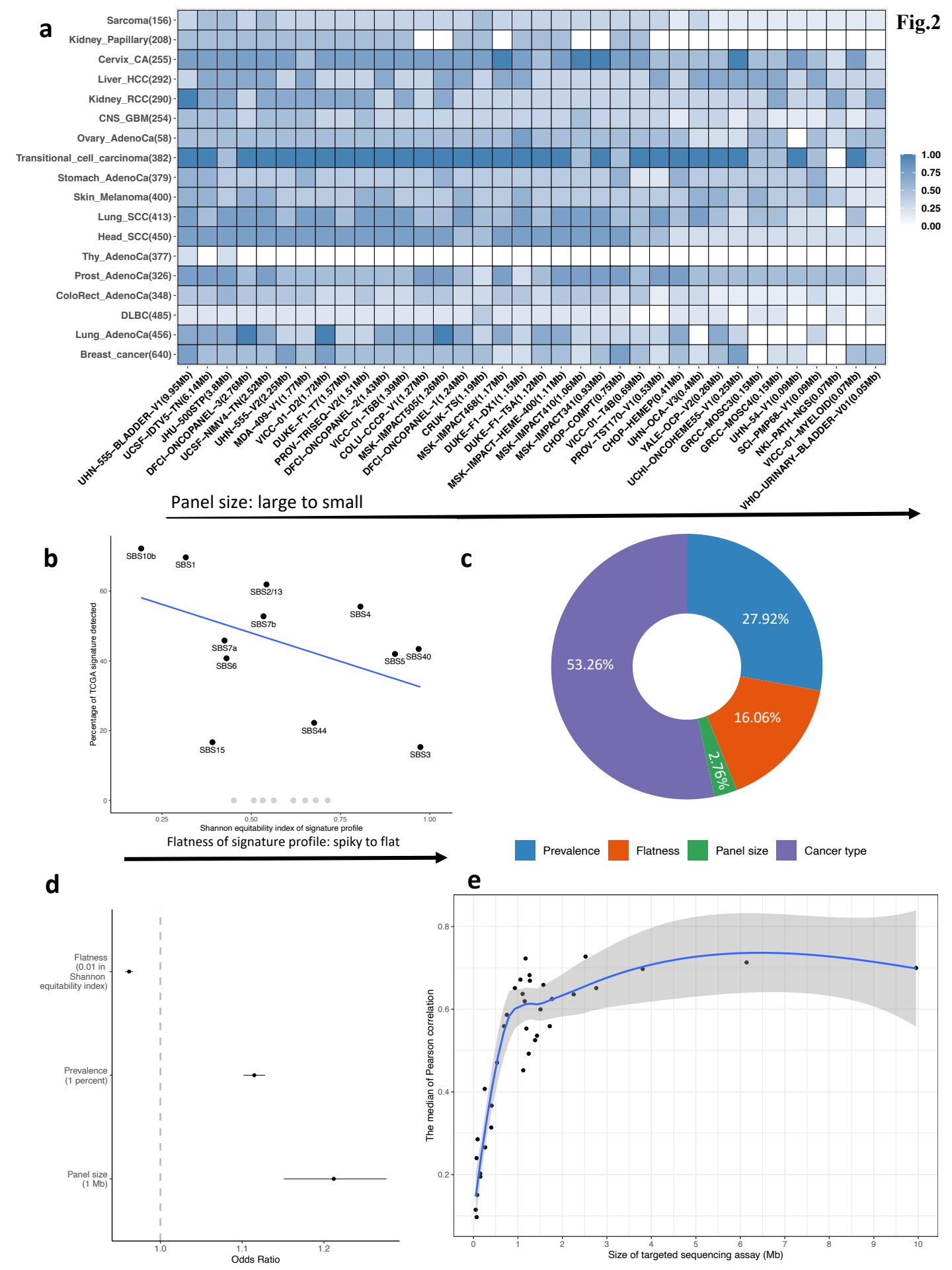
32. Weersma, R.K. *et al.* Increased incidence of azathioprine-induced pancreatitis in Crohn's disease compared with other diseases. *Aliment Pharmacol Ther* **20**, 843-50 (2004).
33. Kirkegard, J., Cronin-Fenton, D., Heide-Jorgensen, U. & Mortensen, F.V. Acute Pancreatitis and Pancreatic Cancer Risk: A Nationwide Matched-Cohort Study in Denmark. *Gastroenterology* **154**, 1729-1736 (2018).
34. Sadr-Azodi, O. *et al.* Pancreatic Cancer Following Acute Pancreatitis: A Population-based Matched Cohort Study. *Am J Gastroenterol* **113**, 1711-1719 (2018).
35. Kundra, R. *et al.* OncoTree: A Cancer Classification System for Precision Oncology. *JCO Clin Cancer Inform* **5**, 221-230 (2021).
36. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
37. Févotte, C. & Cemgil, A.T. Nonnegative matrix factorizations as probabilistic inference in composite models. in *2009 17th European Signal Processing Conference* 1913-1917 (2009).
38. Shannon, C.E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379-423 (1948).



The schematic workflow of SATS

Mutation type matrix **V** ≈ Panel context matrix **L** ⊙ Signature profile matrix **W** × Signature activity matrix **H**

Fig.2


a
Sarcoma(156)
Kidney_Papillary(208)
Cervix_CA(255)
Liver_HCC(292)
Kidney_RCC(290)
CNS_GBM(254)
Ovary_AdenoCa(58)
Transitional_cell_carcinoma(382)
Stomach_AdenoCa(379)
Skin_Melanoma(400)
Lung_SCC(413)
Head_SCC(450)
Thy_AdenoCa(377)
Prost_AdenoCa(326)
ColoRect_AdenoCa(348)
DLBC(485)
Lung_AdenoCa(456)
Breast_cancer(640)
1.00
0.75
0.50
0.25
0.00
UHN–555–BLADDER–V1(9.95Mb)
UCSF–IDTV5–TN(6.14Mb)
JHU–500STP(3.8Mb)
DFCI–ONCOPANEL–3(2.76Mb)
UCSF–NIMV4–TN(2.52Mb)
UHN–555–V2(2.25Mb)
MDA–409–V1(1.77Mb)
VICC–01–D2(1.72Mb)
DUKE–F1–T7(1.57Mb)
PROV–TRISEQ–V2(1.51Mb)
DFCI–ONCOPANEL–2(1.43Mb)
VICC–01–T6B(1.39Mb)
COLU–CCCP–V1(1.27Mb)
MSK–IMPACT505(1.26Mb)
DFCI–ONCOPANEL–1(1.24Mb)
CRUK–TSI(1.19Mb)
MSK–IMPACT468(1.17Mb)
DUKE–F1–DX1(1.15Mb)
DUKE–F1–T5A(1.12Mb)
MSK–IMPACT–HEME–400(1.11Mb)
MSK–IMPACT410(1.06Mb)
MSK–IMPACT341(0.93Mb)
CHOP–COMPT(0.75Mb)
VICC–01–T4B(0.69Mb)
PROV–TST170–V1(0.53Mb)
CHOP–HEMEP(0.41Mb)
UHN–OCA–V3(0.4Mb)
YALE–OCP–V2(0.26Mb)
UCHI–ONCOHEME55–V1(0.25Mb)
GRCC–MOSC3(0.15Mb)
GRCC–MOSC4(0.15Mb)
UHN–54–V1(0.09Mb)
SCI–PMP68–V1(0.09Mb)
NKI–PATH–NGS(0.07Mb)
VICC–01–MYELOID(0.07Mb)
VHIO–URINARY_BLADDER–V01(0.05Mb)
Panel size: large to small

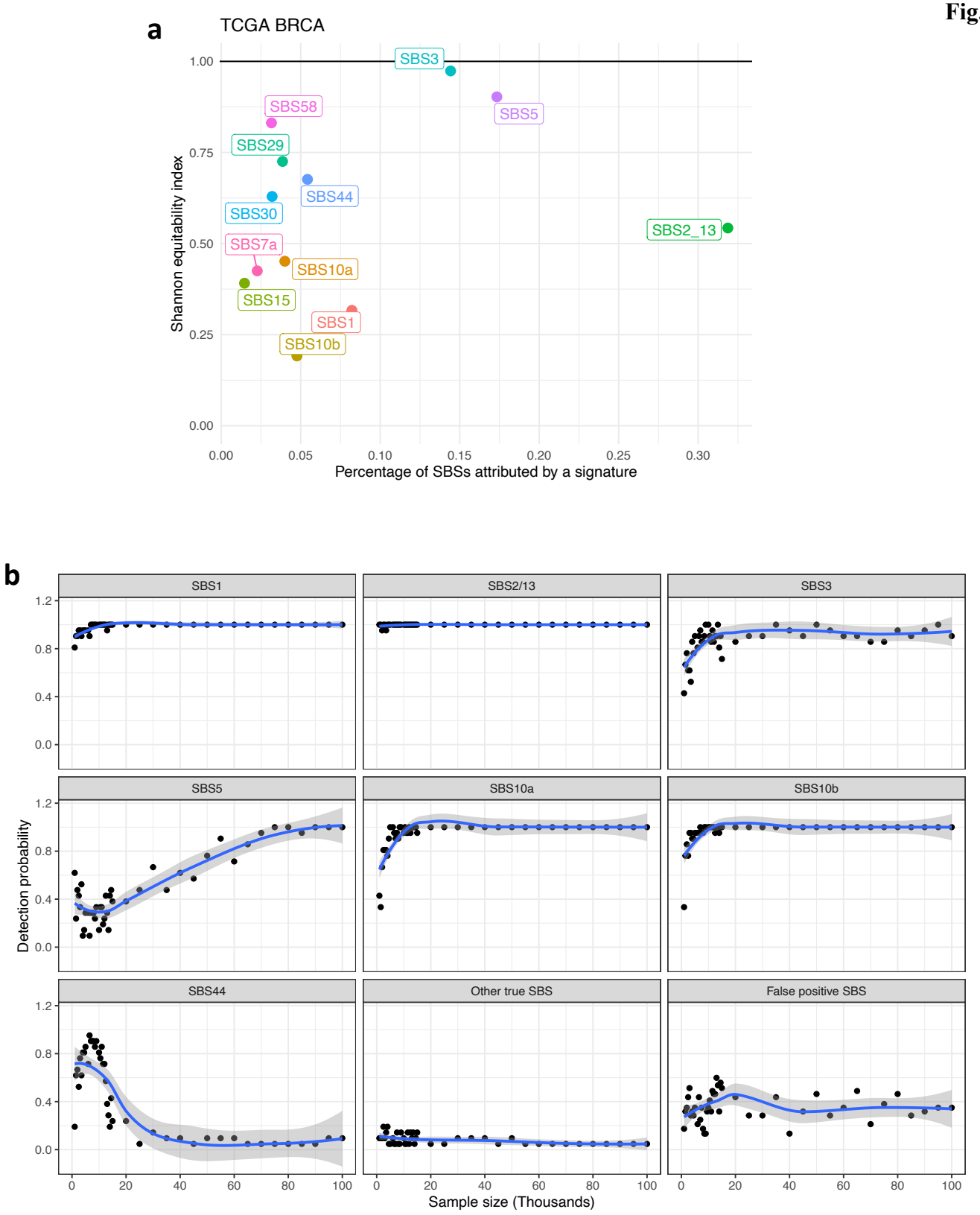
b
SBS10b
SBS1
SBS2/13
SBS4
SBS7b
SBS7a
SBS6
SBS5
SBS40
SBS44
SBS15
SBS3
Percentage of TCGA signature detected
Shannon equitability index of signature profile
Flatness of signature profile: spiky to flat

c
27.92%
16.06%
2.76%
53.26%
Prevalence
Flatness
Panel size
Cancer type

d
Flatness (0.01 in Shannon equitability index)
Prevalence (1 percent)
Panel size (1 Mb)
Odds Ratio

e
The median of Pearson correlation
Size of targeted sequencing assay (Mb)

Fig.3


a
TCGA BRCA
Shannon equitability index
Percentage of SBSs attributed by a signature
SBS3
SBS5
SBS58
SBS29
SBS44
SBS30
SBS2_13
SBS7a
SBS10a
SBS15
SBS1
SBS10b
b
SBS1
SBS2/13
SBS3
SBS5
SBS10a
SBS10b
SBS44
Other true SBS
False positive SBS
Detection probability
Sample size (Thousands)

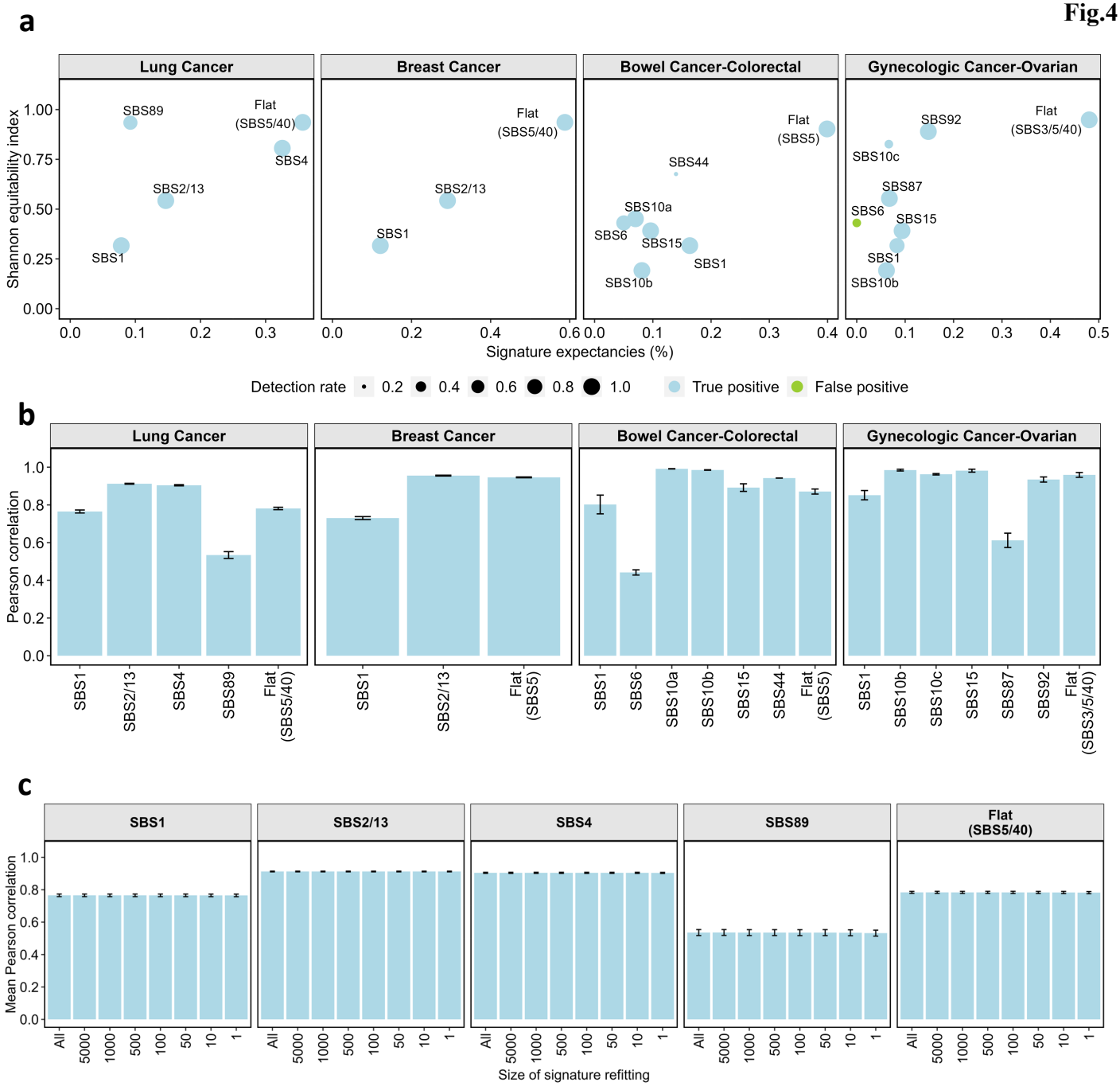
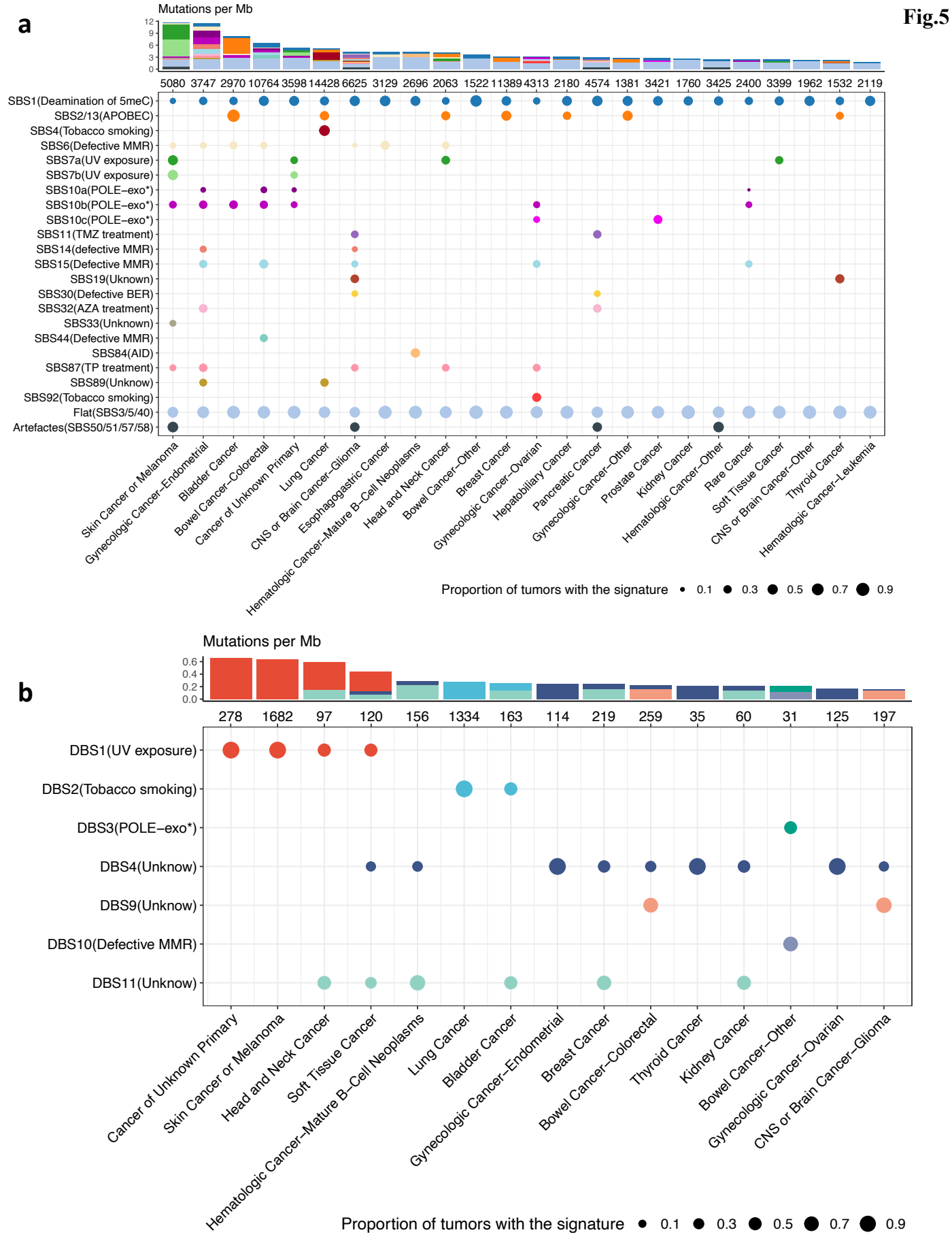

a
Lung Cancer
Breast Cancer
Bowel Cancer-Colorectal
Gynecologic Cancer-Ovarian
Shannon equitability index
Signature expectancies (%)
SBS89
Flat (SBS5/40)
SBS4
SBS2/13
SBS1
SBS10a
SBS6
SBS15
SBS10b
Flat (SBS5)
SBS44
SBS92
Flat (SBS3/5/40)
SBS10c
SBS87
Detection rate
0.2
0.4
0.6
0.8
1.0
True positive
False positive
b
Pearson correlation
SBS1
SBS2/13
SBS4
SBS89
Flat (SBS5/40)
SBS6
SBS10a
SBS10b
SBS15
SBS44
Flat (SBS5)
SBS10b
SBS10c
SBS15
SBS87
SBS92
Flat (SBS3/5/40)
c
SBS1
SBS2/13
SBS4
SBS89
Flat (SBS5/40)
Mean Pearson correlation
All
5000
1000
500
100
50
10
1
Size of signature refitting

Fig.5


a
Mutations per Mb
12
9
6
3
0
5080 3747 2970 10764 3598 14428 6625 3129 2696 2063 1522 11389 4313 2180 4574 1381 3421 1760 3425 2400 3399 1962 1532 2119
SBS1(Deamination of 5meC)
SBS2/13(APOBEC)
SBS4(Tobacco smoking)
SBS6(Defective MMR)
SBS7a(UV exposure)
SBS7b(UV exposure)
SBS10a(POLE−exo*)
SBS10b(POLE−exo*)
SBS10c(POLE−exo*)
SBS11(TMZ treatment)
SBS14(defective MMR)
SBS15(Defective MMR)
SBS19(Uknown)
SBS30(Defective BER)
SBS32(AZA treatment)
SBS33(Unknown)
SBS44(Defective MMR)
SBS84(AID)
SBS87(TP treatment)
SBS89(Unknow)
SBS92(Tobacco smoking)
Flat(SBS3/5/40)
Artefactes(SBS50/51/57/58)
Skin Cancer or Melanoma
Gynecologic Cancer−Endometrial
Bladder Cancer
Bowel Cancer−Colorectal
Cancer of Unknown Primary
Lung Cancer
CNS or Brain Cancer−Glioma
Esophagogastric Cancer
Hematologic Cancer−Mature B−Cell Neoplasms
Head and Neck Cancer
Bowel Cancer−Other
Breast Cancer
Gynecologic Cancer−Ovarian
Hepatobiliary Cancer
Pancreatic Cancer
Gynecologic Cancer−Other
Prostate Cancer
Kidney Cancer
Hematologic Cancer−Other
Rare Cancer
Soft Tissue Cancer
CNS or Brain Cancer−Other
Thyroid Cancer
Hematologic Cancer−Leukemia
Proportion of tumors with the signature 0.1 0.3 0.5 0.7 0.9
b
Mutations per Mb
0.6
0.4
0.2
0.0
278 1682 97 120 156 1334 163 114 219 259 35 60 31 125 197
DBS1(UV exposure)
DBS2(Tobacco smoking)
DBS3(POLE−exo*)
DBS4(Unknow)
DBS9(Unknow)
DBS10(Defective MMR)
DBS11(Unknow)
Cancer of Unknown Primary
Skin Cancer or Melanoma
Head and Neck Cancer
Soft Tissue Cancer
Hematologic Cancer−Mature B−Cell Neoplasms
Lung Cancer
Bladder Cancer
Gynecologic Cancer−Endometrial
Breast Cancer
Bowel Cancer−Colorectal
Thyroid Cancer
Kidney Cancer
Bowel Cancer−Other
Gynecologic Cancer−Ovarian
CNS or Brain Cancer−Glioma
Proportion of tumors with the signature 0.1 0.3 0.5 0.7 0.9