Supplementary Methods

1. Validating the background signature.

We obtained whole-genome sequencing data for 10 randomly selected normal breast tissue samples from Nik-Zainal et al.[1] and ran Sentieon Genomics[2] (default parameters, minimum quality of 30 for calls) on the BAM files to obtain variant calls. We compiled a list of unique high-confidence SNPs across all 10 samples. For all positions where the ancestral allele (obtained from Phase 3 data from the 1000 Genomes project) differed from the reference allele, the ancestral allele was taken in place of the reference allele. We then counted the number of SNPs in each of the 96 trinucleotide-based categories and divided them by the total number of SNPs to obtain the germline mutation spectrum. We found a cosine similarity of 0.98 between this background and the one based on Rahbari et al[3].

2. Simulations to choose the percentage of cells in the matrix M to be held out for cross-validation.

We selected a set of 10 mutational signatures from the ones listed in COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures) with a known etiology. These are: Signatures 1, 2, 7, 10, 11, 13, 15, 22, 24, 26 in COSMIC.

We then generated random configurations of signatures by selecting the germline background, the methylation signature (Signature 1 from the list above) and 4 additional signatures randomly chosen among the 9 remaining signatures of the list above.

For each selected signature, we randomly generated the magnitude of its exposure per patient; this was done by sampling the number of point mutations per signature using a negative binomial distribution with mean 6000 and dispersion parameter 1.5, with the constraints of a minimum number of 1000 mutations per tumor and a maximum of 20000. These parameters were estimated from real data[1]. For each configuration, we generated simulated data from 500 and 1000 tumors.

This procedure was repeated to obtain 100 simulated datasets with 500 tumors and 100 simulated datasets with with 1000 tumors.

For cross-validation, values of K ranging from 3 to 7 were tested. λ was set to be equal to 0.05.

For each dataset, the initial values of the signature matrix β were computed by 5 repetitions of NMF. Cross-validation was performed by holding out (replacing with 0) 0.1%, 1% and 10% of the entries in the mutation count matrix M. Each cross-validation was repeated 5 times with 5 restarts each.

In these simulations, using 1% of the entries for cross-validation resulted in the most accurate prediction of K, both in datasets containing 500 samples and 1000 samples. Detailed results are reported in Supplementary Table 1.

3. Methods and parameters used for signature discovery on the pan-cancer dataset.

SparseSignatures:

K (number of signatures): 2 to 16
λ (sparsity): ranging from 0.01 to 0.1
Number of repetitions of NMF to calculate initial values: 100
Number of iterations to fit signatures using the alternating method with sparsity: 20
Number of repetitions of bi-cross-validation: 500
Number of restarts per repetition of cross-validation: 5

SignatureAnalyzer:

Maximum value of K: 60.
Repetitions of complete pipeline: 10

nsNMF:

Number of repetitions: 100

4. Comparison with signatures generated by SignatureAnalyzer and nsNMF.

Using its inbuilt method, SignatureAnalyzer selected 49 as the best number of signatures to fit the pan-cancer dataset. For nsNMF, using the common criterion of minimal residual error and high reproducibility of signatures, we selected 20 as the number of signatures.

Both sets of signatures (49 and 20) did not include any signature resembling our replication error background.

For direct comparison with the 10 signatures generated by SparseSignatures, we also set 10 as the number of signatures for both methods and obtained the fitted signatures. These are shown in supplementary figures 5 and 6.

References:

1. Nik-Zainal, Serena, et al. "Landscape of somatic mutations in 560 breast cancer whole-genome sequences." Nature 534.7605 (2016): 47.
2. Freed, Donald N., et al. "The Sentieon Genomics Tools-A fast and accurate solution to variant calling from next-generation sequence data." bioRxiv (2017): 115717.
3. Rahbari, Raheleh, et al. "Timing, rates and spectra of human germline mutation." Nature genetics 48.2 (2016): 126.