Appendix

A.1 Proof of a property mentioned in the Methods section

Theorem 1. *Consider any mutation count vector c and a signature vector q. The magnitude of the least-loss exposure vector in the direction of q is given by the scalar projection of c onto q, written* $\text{proj}_q(c) = \frac{1}{||q||}\langle q, c\rangle$.

Proof. Given count vector c and signature q, we wish to find the least-loss exposure of c onto q. We seek the exposure a such that the residual between c and aq (the exposure *vector*) is minimized. We write $||\cdot||$ to mean the 2-norm. We wish to solve $\min_a ||c - aq||$, or equivalently:

$$\min_a ||c - aq||^2. \tag{6}$$

Writing Eq. 6 as an inner product, yields:

$$||c - aq||^2 = \langle c - aq, c - aq\rangle = a^2||q||^2 - 2a\langle q, c\rangle + ||c||^2 \tag{7}$$

Eq. 7 is quadratic in a. Thus, we solve for a by taking the derivative with respect to a and setting it to be 0, which gives $0 = 2a||q||^2 - 2\langle q, c\rangle$. Rearranging, we find the solution, $a = \frac{\langle q,c\rangle}{||q||^2}$.

Notice that a is indeed the least-loss exposure as any solution is non-negative because since q and c are non-negative. The exposure *vector* aq is given precisely by the projection of c onto q. Writing the unit vector in the direction of q as $\hat{q} = \frac{q}{||q||}$, we get,

$$aq = \frac{\langle q, c\rangle}{||q||^2}q = \frac{\langle q, c\rangle}{||q||}\frac{q}{||q||} = \text{proj}_q(c)\,\hat{q}. \tag{8}$$

Thus the magnitude of aq, the least-loss exposure *vector*, of mutation counts c in the presence of a single signature q, is scalar projection of c onto q, $\text{proj}_q(c) = \frac{\langle q,c\rangle}{||q||}$. More specifically, *magnitude* of the least-loss exposure *vector* and relates to the *value* of the least-loss exposure, a, by a constant factor $\frac{1}{||q||}$, with $a = \frac{1}{||q||}\text{proj}_q(c)$. ■

A.2 Active samples for each signature

Signature	Active Samples	Active Samples
2	232	41.4 %
3	278	49.6 %
8	494	88.2 %
13	262	46.8 %
18	64	11.4 %
30	135	24.1 %

Table A1. The number of *active* samples in the 560 breast cancers samples for the mutational signatures known to be present in breast cancer. A signature is active if that signature is responsible for 5% or more of the total mutations in the sample. Signatures 6, 10, 17, 20, and 26 are active in fewer than 5% of samples and thus are not considered in this study. Signatures 1 and 5 are also not considered because they are expected to be active in *every* sample [1].

A.3 Spearman correlations between panel and whole genome exposures at various activity thresholds

Signature	Activity Threshold	Active Samples	SCALPELSIG ($\alpha = 1$)	SCALPELSIG ($\alpha = 0.5$)	MSK-IMPACT
2	1%	88.4%	0.2889	0.403	0.3013
2	5%	41.4%	0.3643	0.3754	0.2774
2	10%	23.6%	0.3766	0.4584	0.3509
2	20%	10.7%	0.3327	0.4516	0.3318
3	1%	66.4%	0.3556	0.3809	0.1823
3	5%	49.6%	0.2605	0.44	0.2579
3	10%	38.4%	0.3747	0.3776	0.2666
3	20%	28.4%	0.415	0.3882	0.2278
8	1%	95.5%	–	–	–
8	5%	88.2%	0.3613	0.3875	0.0262
8	10%	73.6%	0.2962	0.3036	0.0853
8	20%	27%	0.3576	0.3023	0.0982
13	1%	92%	0.4836	0.5367	0.3955
13	5%	46.8%	0.5176	0.582	0.475
13	10%	25.4%	0.5296	0.4992	0.4611
13	20%	12.1%	0.4924	0.4972	0.4514
18	1%	32.9%	0.1346	0.0882	0.0067
18	5%	11.4%	0.0717	0.1364	0.0854
18	10%	4.6%	–	–	–
18	20%	0.5%	–	–	–
30	1%	63%	-0.0954	0.1169	0.1154
30	5%	24.1%	0.0442	0.0371	0.0534
30	10%	3.5%	–	–	–
30	20%	0.2%	–	–	–

Table A2. Spearman correlation between exposures computed using panel regions and exposures computed from whole genome mutation counts, for panels constructed using various thresholds for activity classification. The Activity Threshold column gives the required percentage of mutations contributed by a signature for a sample to be considered 'active.' That is, in the first row samples were classified as 'active' if at least 1% of their mutations were contributed by Signature 2; in the second row samples were considered 'active' if at least 5% (the default for this study) came from Signature 2, etc. For the SCALPELSIG columns, and the MSK-IMPACT column, values shown are median Spearman correlation coefficients across 15 randomized test and train sets. Notably, different random test/train splits were sampled for each row. This was a necessary step since we use stratified sampling to guarantee that the class balance of test sets is equivalent to the overall cohort (and the class balance changes as the Activity Threshold is varied). As a consequence, the values for the MSK-IMPACT panel vary across experiments despite the fact that the observed genome regions remain the same. Cases where either the active or inactive class comprised fewer than 5% of the samples were discarded, since these cases reduce the possible variation in the test set.

A.4 Evaluation of SCALPELSIG with varying amounts of training data

Signature	Training Data (% of cohort)	Training samples (count)	Spearman	AUPR
2	90	504	0.3819	0.633
2	80	448	0.4006	0.6462
2	60	336	0.385	0.6199
2	40	224	0.3975	0.6302
2	20	112	0.3639	0.6197
3	90	504	0.3883	0.7191
3	80	448	0.383	0.6984
3	60	336	0.3635	0.6888
3	40	224	0.3572	0.7002
3	20	112	0.3722	0.7025
8	90	504	0.3895	0.8929
8	80	448	0.3479	0.8834
8	60	336	0.3600	0.891
8	40	224	0.3434	0.8891
8	20	112	0.3400	0.8978
13	90	504	0.5749	0.7562
13	80	448	0.5467	0.7433
13	60	336	0.5393	0.7405
13	40	224	0.5596	0.7372
13	20	112	0.5416	0.7409
18	90	504	0.1482*	0.2111
18	80	448	0.0792*	0.158
18	60	336	0.0749*	0.1612
18	40	224	0.0792*	0.141
18	20	112	0.0965*	0.163
30	90	504	0.0569*	0.2499
30	80	448	0.1032*	0.2493
30	60	336	0.0787*	0.2567
30	40	224	0.06884*	0.2479
30	20	112	0.0888	0.254

Table A3. Spearman correlation between exposures computed using panel regions and exposures computed from whole genome mutation counts, as well as AUPR for signature activity prediction, for panels constructed by ScalpelSig using various amounts of training data. For each row, 15 randomized training sets were obtained using stratified sampling as detailed in the Methods section. The size of the training sets for each row is indicated in the Training Data column. The SCALPELSIG algorithm was run separately on each of the training sets, and evaluated using all of the held out samples. Values shown are mean Spearman correlation coefficients and mean AUPR across the randomized trials. Spearman values where fewer than half of the trials yielded p-value < 0.05 are marked with an asterisk.

A.5 Preliminary results for the combined-signature panel

We report the results of an experiment evaluating a simple approach for constructing a panel for multiple signatures simultaneously in Table A4.

	Spearman correlation		AUPR	
Signature	**ScalpelSig** ($\alpha = 0.5$)	**MSK-IMPACT**	**ScalpelSig** ($\alpha = 0.5$)	**MSK-IMPACT**
2	0.3865	0.3162	0.6232	0.6091
3	0.3700	0.1698	0.6894	0.5790
8	0.2948	0.0507	0.8918	0.8900
13	0.4796	0.4401	0.7270	0.6962
18	0.1102	0.0675	0.1723	0.1193

Table A4. Preliminary results from assessing genomic panels designed to detect multiple signatures simultaneously. The panels were constructed by combining windows from SCALPELSIG panels for individual signature detection. No windows for detection of Signature 30 were incorporated into the panel and we did not evaluate the combined panel's performance at detecting Signature 30, since the SCALPELSIG panels that were individually optimized for Signature 30 detection did not outperform baselines. Values shown are median AUPR and Spearman correlation across 15 randomized test and train sets. Note that when testing for the activity of multiple signatures simultaneously, samples can no longer be neatly categorized into binary classes (i.e. a sample that is active for some signatures may be inactive for others). As a result, we could not use the stratified sampling protocol that was used to split the data into test and train sets in the individual signature panel experiments. We instead used uniform random sampling to split the data into test and train sets. This means that there was more variance in test sets of the combined panel experiments, which likely accounts for the differences in the MSK-IMPACT scores in comparison to the individual signature panel experiments.