**Supplementary Note**

To explore how consistently signatures are extracted from different datasets of the same tissue type, and whether they were robust and stable, we used two breast cancer cohorts:

* one comprising 522 samples available and downloaded from COSMIC
* one comprising 560 samples from an ICGC Breast Cancer publication (Nik-Zainal et al, 2016)

We deliberately permitted some shared samples to exist in the two cohorts expecting to therefore see similar signatures extracted in two independent runs.

Much of the work that underpins the consensus set of 30 signatures that are currently displayed on the COSMIC webpage <http://cancer.sanger.ac.uk/cosmic/signatures>, were derived using an algorithm that had been previously published (Alexandrov et al, 2013). We thus used an alternative algorithm (SomaticSignatures (Gehring et al, 2015])) though not dissimilar in its underlying principle, so that we could contrast the results from these different algorithms.

We ran SomaticSignatures with default input parameters. In order to find the optimal solution, that is, the optimal number of signatures, we followed the authors’ suggestion of assessing the trend of the explained variance produced by the algorithm and to focus the decision on the first inflection point. This approach suggested that the optimal number of signatures for the ICGC dataset was 13 while that for the COSMIC dataset was 14. Supplementary Figures 1 and 2 show the profiles of the extracted signatures for the ICGC and COSMIC cohorts respectively.

First, although there are differences observed between the signatures extracted by SomaticSignatures from the two datasets, there is a lot of similarity between them. Ten signatures are nearly identical in the two analyses (Supplementary Figure 3). When the signatures extracted from the two datasets were compared to the consensus set of 30 signatures (Supplementary Figures 4 and 5), seven of the ten signatures that were very similar to each other, shared strong concordance with COSMIC consensus signatures. In decreasing order of similarity between the two extractions – Signatures 2, 13, 1, 18, 8, 3 and 17 – all had very high cosine similarity (> 0.9). There was some difficulty in stably extracting signatures associated with mismatch repair deficiency (Signatures 6, 20 and 26) possibly because these signatures are rare in breast tumors, present in only 1-2% of the cohort. In both datasets, the algorithm struggled to disentangle a set of signatures that were previously called Signatures 5 and 30 in the consensus set of 30 signatures.

In the independent analyses, there were signatures that were “unique” to specific datasets. Some of this may be true biological signal. For example, it is true that there is a single hypermutated sample that harbours a high number of mutations associated with (what we believe is) Signature 30 in the ICGC dataset. There is also a POLE hypermutated sample in the breast cancers downloaded from COSMIC, which does not exist in the other dataset, yet the algorithm was able to detect this extremely rare tumor and read-out the associated Signature 10 correctly. It may also be that some of the “unique” signatures are simply artefactual – a product of mathematically choosing the wrong solution for biological signatures.

We also used SomaticSignatures to extract the signatures in sixteen other cancer types (Supplementary Table 1). In this way, we were able to compare the twelve signatures extracted in the breast datasets with what we believe to be the same signatures extracted in the other cancer types to get an idea of variance across tissues. Figure 2 shows the bar plots of the twelve substitution signatures extracted from the breast data, and the error bars show the variability observed across other cancer types that also have this signature.

**Supplementary Figure 1**: Thirteen mutational signatures extracted using SomaticSignatures from the cohort of 560 ICGC breast cancers.

**Supplementary Figure 2**: Fourteen mutational signatures extracted using SomaticSignatures from an independent analysis of a cohort of 522 COSMIC breast cancers.

**Supplementary Figure 3**: Comparing the mutational signatures extracted from the two independent analyses. Although there are differences observed between the signatures extracted by SomaticSignatures from the two datasets, there is a lot of similarity between them (as shown in the matrix reporting the cosine similarities that report the similarity between the signatures from the two extractions – a higher score indicates a closer likeness; in red font are the signatures extracted from the ICGC 560 cancers which are on the x-axis (S1-S13) and signatures extracted from the 522 COSMIC cohort are on the y-axis (S1-S14)). Ten signatures are very similar in the two analyses (highlighted by a green box) – seven are nearly identical and also share extraordinarily close similarity to the consensus set of 30 signatures (depicted in green font on either axis as CS 1-30)). There was some difficulty in stably extracting signatures associated with mismatch repair deficiency (consensus Signatures 6, 20 and 26 (S9 and S11 on x-axis and S9 on y-axis) possibly because these signatures are rare in breast tumors, present in only 1-2% of the cohort. In both datasets, the algorithm struggled to disentangle a set of signatures that were previously called Signatures 5 and 30 in the consensus set of 30 signatures (S10 and S13 on the y-axis, S12 and S14 on the x-axis highlighted in yellow font). In the independent analyses, there were signatures that were “unique” to specific datasets (S7 and S8 on y-axis and S4 and S11 on the y-axis).

**Supplementary Figure 4**: Comparing the extracted mutational signatures using SomaticSignatures on the 560 ICGC breast cancers (S1-S13 on the y-axis), to the consensus set of thirty substitution signatures that are present at <http://cancer.sanger.ac.uk/cosmic/signatures> (extracted using Alexandrov’s method on the x-axis). Values in the matrix reflect cosine similarity between the extracted signature and each consensus signature. There are signatures that appear to be unique to this dataset. For example, the presumptive extracted signature S10 is closely correlated to Signature 30 of the consensus signatures. It is present in a single hypermutated sample that harbours a high number of mutations associated with this signature. However, other unique and artefactual signatures may also be present.

**Supplementary Figure 5**: Comparing the extracted mutational signatures using SomaticSignatures on the 522 COSMIC breast cancers (S1-S14 on the y-axis), to the consensus set of thirty substitution signatures that are present at <http://cancer.sanger.ac.uk/cosmic/signatures> (extracted using Alexandrov’s method on the x-axis). Values in the matrix reflect cosine similarity between the extracted signature and each consensus signature. A unique signature here is the extracted signature S5 which has a very high cosine similarity to that of consensus Signature 10. There is indeed a sample with this particular signature that has been previously associated with mutations in POLE. Thus, this appears to be true biological signal.

**Supplementary Table 1**: List of tissue types, number of samples for each tissue-type and number of signatures identified in each of those tissue types. BRCA\_BASIS = ICGC 560 breast cancer cohort, BRCA\_COSMIC = COSMIC 522 breast cancers, BTCA = biliary tree cancer, CESC = cervical cancer, COAD = colon adenocarcinoma, ENDO = , ESAD = esophageal cancer, KICA = kidney cancer, LIVC = liver cancer, LUSC = lung squamous cancer, LYMP = lymphoma, MELA = melanoma, OVCA = ovarian cancer, PACA = pancreatic cancer, PNET = , PRAD = prostate cancer, STAD = stomach cancer, THCA = thyroid cancer.