Science
AAAS

Supplementary Materials for

Substitution mutational signatures in whole-genome-sequenced cancers in the UK population

Andrea Degasperi[1,2], Xueqing Zou[1,2], Tauanne Dias Amarante[1,2], Andrea Martinez-Martinez[1,2], Gene Ching Chiek Koh[1,2], João M. L. Dias[1,2], Laura Heskin[1,2], Lucia Chmelova[1,2], Giuseppe Rinaldi[1,2], Valerie Ya Wen Wang[1,2], Arjun S. Nanda[1,2], Aaron Bernstein[1,2], Sophie E. Momen[1,2], Jamie Young[1,2], Daniel Perez-Gil[1,2], Yasin Memari[1,2], Cherif Badja[1,2], Scott Shooter[1,2], Jan Czarnecki[1,2], Matthew A. Brown[3,4], Helen R. Davies[1,2], Genomics England Research Consortium[3,*], Serena Nik-Zainal[1,2]

Correspondence to: sn206@cam.ac.uk

This PDF file includes:

Materials and Methods
Figs. S1 to S53
Captions for Tables S1 to S33

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S33

Materials and Methods

Datasets

We used three large pan-cancer whole genome cohorts: the Genomics England Limited (GEL) version 8 cohort of the 100,000 Genomes Project (*7*), the ICGC cohort (*9, 11*) and the Hartwig cohort (*12*).

The GEL cohort contains 15,838 cancer whole genomes involving 23 tumor-types. We performed two steps of quality control: an automated check of sequencing and mapping quality parameters (table S2), and a visual curation (e.g., missing data and evidence of contamination from other samples). We included 12,222 samples across 19 organs for mutational signature analysis. We excluded formalin-fixed paraffin embedded samples (FFPE) and samples with short fragment size and low mapping rate. Around 6.5% of samples had a few cycles of PCR (table S1). The GEL version 8 dataset can be accessed via https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access.

The ICGC cohort contains 3,001 cancer whole genomes across 19 organs, comprising 2471 samples from PCAWG (EGAS00001001692) and 530 additional breast cancers (450 from EGAS00001001178 and 80 from EGAD00001002740).

The Hartwig cohort contains 3,417 metastatic cancer whole genomes across 18 organs. Data can be accessed via at www.hartwigmedicalfoundation.nl/en.

The count of single nucleotide variants, double nucleotide variants, indels and rearrangements in the three cohorts can be found in table S1.

Each cohort uses a different nomenclature for classifying cancer samples into tissues-of-origin. In order to perform comparisons across the three cohorts, we used a common organ name for all cohorts (table S5).

The number of samples for each organ of each cohort can be found in tables S3 and S4, and the full list of samples is available in table S6.

Mutational Signature Extraction

For each sample in each cohort, we constructed single and double base substitution (SBS and DBS) mutational catalogs (tables S16 and S17) as previously described (*5, 14*). We performed independent signature extractions per organ for each cohort.

We used a signature extraction framework that improves upon our recent work (*17*), which permits distinguishing common from rare signatures (Fig. 1C).

First, we clustered the mutational catalogs in each organ for each of the three cohorts, in order to seek out samples that had distinctive, unusual mutational profiles. We used hierarchical clustering with average linkage and used 1 – cosine similarity as distance (fig. S1A). We then followed the dendrogram, and manually split the catalogs into two sets, one with the unusual/rare profiles and one with common/recurrent profiles, which was used as the initial set for the signature extraction (fig. S1B). From this initial set, we excluded GEL samples that were created using PCR library preparations, and samples with unusual/rare profiles. By

excluding these samples, the number of mutational signatures in the initial set was limited to common patterns, reducing the mixing of signatures in the extraction process.

Second, we performed signature extraction (*17*) on this initial set (fig. S1C). In table S6 we indicate which samples were used. In brief, given a matrix of catalogs C, we applied non-negative matrix factorization (NMF) to 20 matrices C', bootstrapped from C. To solve NMF we used the Lee and Seung multiplicative algorithm that optimizes the Kullback-Leibler divergence (KLD) (*46*), producing a matrix of signatures S and a matrix of exposures E for each NMF run, such that $C' \approx SE$. We repeated NMF at least 300 times for each bootstrap matrix, using random initializations, and selected only the solutions that had a final KLD within 0.1% of the best solution found (the solution with the lowest KLD). Then, we clustered all the selected solutions using clustering with matching and computed the data-model error as the average KLD, the goodness of clustering as the average silhouette width (ASW), and the consensus signatures as the medoid of each cluster. Finally, we repeated the above procedure for different values of number of signatures k and manually selected k as the trade-off between data-model error and the ASW. Thus, for each organ in each cohort we reported a set of signatures, that we term common signatures.

Third, for each organ in each cohort, we took all samples into consideration and attempted to identify additional signatures that could be present in the samples (Fig. 1C). To do so we fitted the common signatures S to all the samples (KLD optimization), and identified the samples with a high normalized error, calculated as the sum of absolute deviations between the original catalog c and reconstructed catalog Se, divided by the total mutations in the catalog (fig. S1D). In addition to the high normalized error, we also required that the samples had a residual error above a minimum number of mutations, which was manually tuned for each extraction (3-400 mutations for SBS and 40-50 mutations for DBS). A sample residual error (fig. S1E) was calculated after estimating the sample exposures using least squares (limSolve R package) with the constraint that the difference between original and reconstructed catalogs should be mostly positive ($c - Se > -\tau \cdot \sum_i c_i$), $\tau = 0.003$. We then clustered the residual errors using hierarchical clustering with average linkage and 1 – cosine similarity as distance (fig. S1, F and G). Finally, for each cluster of residual errors, we extracted one signature using a version of NMF where the signature matrix S contained the common signatures as constants, and one additional column that was estimated to contain the new signature, using the NNLM R package, (fig. S1H). Thus, for each organ in each cohort we reported an additional set of signatures, that we term rare signatures.

The total number of common and rare SBS and DBS signatures found is 757 and 301 respectively (tables S9 and S10). The number of common and rare signatures found in each organ in each cohort can be found in tables S11 and S12. It should be noted that the terms common and rare refer to the step at which the signature was identified in a specific organ. In practice, a specific mutational pattern could be considered rare in one per organ extraction of one cohort and be a common pattern in another.

Mutational signature exposures

We fitted common and rare signatures to each sample catalog independently. Rare signatures were fitted only into the samples where the signatures were identified.

In the case of SBS signatures, as described previously (*17*), for each sample we performed 200 signature fits using bootstrapped catalogs and KLD optimization, obtaining an ensemble of 200

exposure estimates for each signature, and chose as a point estimate the median of the exposures. Finally, to increase specificity and reduce the false positive assignment we set to zero the point estimate exposure of a signature, if the proportion of exposures below a certain threshold (5% of the total number of mutations) was higher than 5% (empirical p-value of 0.05). Exposures of fewer than 50 mutations were also set to zero.

In the case of DBSs, the number of mutations were too low to perform the bootstrap-based fit described above, so we performed a single signature fit instead. To increase specificity, exposures were set to zero if they contributed to less than 25% of the total number of mutations and if they were less than 25 mutations.

Reference signatures

To be able to describe and discuss signatures across organs and cohorts, we determined a set of reference signatures to denote unifying processes (Fig. 1G). Each signature extracted in each cohort-organ combination could then be mapped to one or more reference signatures.

We clustered all organ signatures from all three cohorts (757 in the case of SBS, 301 DBS) using hierarchical clustering with average linkage and 1 – cosine similarity as distance. We then manually identified clusters by following the hierarchical clustering dendrogram (tables S13 and S14). Manual clustering was necessary because it was not possible to use a single threshold for the dendrogram that would be appropriate for all recurrent patterns. The clusters were selected so that all the signatures within each cluster were highly similar. We then computed the average of each cluster and termed these 'distinct patterns' (Fig. 1G and tables S15 and S16).

Next, we considered that each distinct pattern (187 for SBS and 60 for DBS) was either: i) a reliably recurrent distinct pattern that we could observe in multiple independent extractions; ii) a mix of two or more distinct patterns; iii) a singleton pattern found only in one organ in one cohort (Tables S17 and S18).

We clustered the recurrent distinct patterns to determine whether some distinct patterns could be a variant of the same pattern. Cluster means were then reported as a first set of highly reliable reference signatures.

To identify mixed distinct patterns, we performed a signature fit (KLD optimization) of each possible combination of two distinct patterns into each distinct pattern. Mixed patterns were not considered reference signatures, but rather a combination of reference signatures obtained from recurrent distinct patterns.

We investigated the singleton distinct patterns to determine if patterns were likely variants of the reference signatures, and if not, they were reported as additional reference signatures, some of which may have been reported in other studies.

A total of 120 SBS and 39 DBS reference signatures were identified (tables S19, S20, S21 and S22).

A conversion matrix was constructed to map the cohort-organ signatures to the reference signatures (tables S25 and S26). Most signatures can be mapped exactly to one reference signature (entry 1 in the conversion matrix) based on the distinct patterns clustering. Cohort-

organ signatures that clustered into mixed distinct patterns were mapped to multiple reference signatures using the coefficients determined at the identification of mixed distinct patterns.

We used the conversion matrix and information about common/rare signatures to rename the cohort-organ signatures in a meaningful way. For example, "GEL-Ovary_common_SBS1+18" indicates that the signature is from the GEL cohort, Ovary organ, was identified among the common signatures, it is an SBS signature and according to the conversion matrix it is a mix of reference signatures SBS1 and SBS18.

Finally, we used the conversion matrix to convert the cohort-organ signature exposures into reference signature exposures (tables S23 and S24).

Quality control of reference signatures

Each reference signature was given a QC status of "green", "amber" or "red", according to additional evidence. For example, signatures observed in multiple cohorts and multiple organs, or observed in previous studies, were considered "green", while patterns that were only observed in a single extraction were usually considered "amber" denoting some uncertainty. The "red" status was given to patterns considered mathematical or alignment artefacts.

After quality control, 82/120 SBS and 27/39 DBS reference signatures had QC "green" status (tables S19 and S20).

When seeking etiologies and/or potential artefacts for the signatures, we performed the following additional QC:

- Genetically:
    - we check relatedness of samples (because some patients do have more than one sample in the 100,000 Genomes Project)
    - we seek potential germline variants as a contributing cause for a signature and
    - we go through somatic driver mutations
- In many cases, medical records were searched for:
    - past medical histories
    - past occupational exposures and
    - past treatment histories.

Organ-specificity of signatures

For all common signatures in 16 organs that were mutually present across GEL, ICGC and Hartwig, we sought the most similar signature in another cohort (minimum cosine similarity of 0.85) and checked whether it belonged to the same organ. For each organ in each cohort, this resulted in a proportion of signatures that best matched signatures of the same organ in a different cohort (fig. S2A). These proportions could be calculated in all cohort directions: from ICGC to GEL, GEL to ICGC, GEL to Hartwig, Hartwig to GEL, ICGC to Hartwig and Hartwig to ICGC, resulting in six proportions per organ (fig. S2B).

We calculated the proportion of signatures that matched different organs as well, for example the proportion of GEL-Breast common signatures that best matched ICGC-Ovarian signatures, resulting in 12 proportion values for each match of different organs (fig. S2C).

Finally, we used a Tukey test (confidence level 0.95, p-value threshold 0.05) to determine whether the proportion of signatures matching the same organ was significantly higher than the proportion of signatures matching different organs (fig. S2D). The total number of organ comparisons was 16, thus we counted how many times a comparison with the same organ was found to be significantly higher than the different organs comparisons, according to the Tukey test. In fig. S2E, we simply reported *** for significantly higher than all other 15 organ comparisons, ** for higher than 11 other organ comparisons and * for higher than 7 other organ comparisons.

Notice that these proportions do not simply indicate organ signatures similarity across cohorts, but rather that looking across an entire other cohort the most similar signatures are in a given organ. If the largest proportions are consistently found in same organ comparisons, this in turn implies organ-specificity of signatures. In some organs, such as biliary and stomach, while there were similar signatures in the same organ across cohorts, these were not consistently the most similar when considering all organs in a cohort, and organ-specificty was not detected.

Additional evaluations of DBS reference signatures

We performed three additional evaluations of the DBS signatures.

First, for each DBS reference signature we selected representative samples that had a high number of mutations (exposures) associated with that signature. Then we manually checked aligned reads at DNV locations to determine if the two substitutions that composed each DNV were in *cis*, i.e., on the same DNA molecule.

Second, for each SBS reference signature that had an associated DBS reference signature (high correlation of SBS and DBS exposures), we performed an *in-silico* analysis, to determine whether the DBS could be explained simply by SNVs of that signature falling adjacent to each other by chance. For each SBS, we sampled 1 million SNV mutations randomly across the genome with the same trinucleotide context and proportion of mutation types defined by the SBS. We then constructed the *in-silico* DBS using the SNVs that fell next to each other (fig. S8, D to I).

Third, for each DBS reference signature we selected representative samples that had a high number of mutations (exposures) associated with that signature. Then, we inspected the mutational context of DNVs, up to 10 bp 5-prime and 3-prime of each DNV (Fig. 3B and fig. S7, B to F).

Replication and transcription strand bias calculation

All single base substitutions were converted to a pyrimidine reference and annotated with respect to a replication and/or transcriptional strand. Leading and lagging strands were determined using replication-sequencing data from the breast cancer cell line MCF-7. Transcribed and non-transcribed strands, associated with gene orientation, were defined for the regions of the genome with transcribed genes. The mutations were further stratified into the respective substitution class (C>A, C>G, C>T, T>A, T>C, T>G). Mutations in each sample were assigned to mutational signatures based on the maximum likelihood (methodology can be found in (*42*)).

Various metrics were used to determine whether strand bias was occurring for each substitution class in each signature: the p-value of the paired two-tailed Student's t-test applied to the proportion of mutations in each strand across samples, using the "natural" bias as the true mean for the test; the log2 ratio between the mutations in each strand, summing all mutations across samples, corrected for the "natural" log2 ratio bias; the contribution of each mutation class in each signature. We thus determined that bias in a substitution class in a signature was present if the p-value was below 0.1 (ignored if only one or two samples have the signature), the absolute value of the log2 ratio was above 0.2 and the contribution of the mutation class to the signature was at least 15%. All metrics are available in table S32.

HRDetect bootstrap scores

HRDetect is a logistic regression classifier that uses whole genome sequencing data to compute the probability of a sample as being Homologous Recombination deficient. To compute the HRDetect score we determined the following input features: exposures of signatures SBS3 and SBS8, as well as rearrangement signatures 3 and 5, the proportion of short deletions at microhomology, and the HRD-LOH index. In particular, the rearrangement signatures exposures were obtained by signature fit of previously published organ specific rearrangement signatures (*17*). We computed both a single score, using the median of bootstrap fits for substitution and rearrangement exposures, as well a distribution of bootstrap scores, perturbing the input features as previously described, and reporting 5^{th}, 50^{th} and 95^{th} scores from a total of 1000 bootstrapped scores. We considered a sample to be classified as HR deficient if the 5^{th} percentile score was above 0.5, i.e., if 95% of the bootstrapped scores were above 0.5, which corresponds to the empirical p-value of 0.05 of score>0.5 (table S31).

FitMS and simulation study

We provide a signature fitting algorithm called signature Fit Multi-Step (FitMS), which allows users to fit our mutational signatures into their own samples. FitMS is written in R and is available in our *signature.tools.lib* package (*45*).

In general, given a mutational catalog c, a signature fit algorithm attempts to find a set of non-negative exposures e that indicate the number of mutations associated with each signature in a given signature matrix S, such that $c \approx Se$.

FitMS is organized in two main steps. In the first step, a set of common signatures is fitted into a sample, while in the second step, the algorithm attempts to improve the fit by adding a small number of rare signatures (one by default).

We implemented two strategies in FitMS:

1. *constrainedFit*: common signatures are fitted using a non-negative least squares algorithm with the additional constraint that the difference between the original catalog c and the reconstructed catalog Se should be mostly positive, $c - Se > -\tau \cdot \sum_i c_i$, with $\tau = 0.003$ (limSolve R package). The residual $R = c - Se$ is then compared to the rare signatures, and if there are rare signatures with cosine similarity of at least 0.8 to R, then the rare signature with the highest cosine similarity is chosen. Finally, the common signatures and the selected rare signature are fitted into the catalog using a non-negative KLD optimization (NNLM R package);

2. *errorReduction*: common signatures are fitted using a non-negative KLD optimization. All rare signatures are then fitted one at a time along with the common signatures. Rare signatures that caused the mean absolute deviation between c and Se to reduce at least 15% with respect to using the common signatures alone, are considered. Finally, if more than one rare signature is considered, the rare signature that induced the highest cosine similarity between the catalog c and the model Pe is selected.

We determine the set of common and rare signatures to be fitted in a sample in an organ-specific way.

- For common signatures, we use the GEL organ-specific common signatures, with the exception of Esophagus and Head_neck, where ICGC signatures are used, because these organs were not available in GEL.
- For rare signatures, we instead chose high-quality reference signatures observed as rare signatures across the various organs and cohorts at least twice, and that did not already belong to the set common signatures.

The list of common and rare signatures that can be used in the 21 organs is available in table S33.

To evaluate the performance of FitMS, we used a simulation study. We simulated 100 genomes so that each genome contained 5 of 9 random GEL-Breast common SBS signatures. In addition, 25 of the 100 samples had one additional rare signature, randomly selected from 54 SBS reference signatures (Table S33). The minimum number of mutations in a sample was 5000 and maximum 50000, sampled uniformly in log scale, so that the values close to 5000 were more likely.

We compared the two FitMS strategies against a "fit all" algorithm, where all 9+54 signatures were fitted into a catalog at the same time using a non-negative KLD optimization.

Each signature fit strategy estimated the exposures of the given signatures in each sample. This first estimate of the exposures usually tends to overfit the signatures into the samples, resulting in false positive exposures consisting of very small number of mutations. Thus, we set an exposure to zero if the number of mutations was lower than a certain threshold, given as a percentage of the total number of mutations in a sample. To assess how the performance of the algorithms changed according to the threshold used, we used different threshold values: 0, 1, 2, 5 and 10% (fig. S53, D to I).

Criteria for calling potential driver variants in GEL data

Potential driver mutations were sought in specific cancer genes associated with mutational signatures. For all genes investigated, germline variants which were called as pathogenic or likely pathogenic in ClinVar were included as potential drivers. For tumor suppressor genes any germline or somatic variant which was predicted to inactivate the gene was included as a potential driver variant. These included both substitutions and small insertions and deletions resulting in; stop gain, frameshift, splice donor and splice acceptor variants and structural rearrangement mutations (deletions, inversion, tandem duplications or translocations) which disrupted the footprint of the gene. In addition, for both tumor suppressor genes and oncogenes,

somatic missense mutations which had been previously reported recurrently in cancer were also considered as potential drivers, including those variants recorded as pathogenic or likely pathogenic in ClinVar and those present in COSMIC database greater than four times (https://cancer.sanger.ac.uk/cosmic). Additional published data was also used to assist driver assignment for the following genes, *POLE* (*47*), *POLD1* (*29*) and *MBD4* (*21*).

Evidence to indicate all wild type alleles of tumor suppressor genes were inactivated in the tumor was provided by either the existence of two or more inactivating mutations or by Loss of Heterozygosity (LOH) of the alternate allele. LOH was indicated by a combination of copy number estimates provided by Canvas, tumor content and estimates of the Variant Allele Fraction (VAF) in the tumor. VAF was used to determine whether LOH of germline variants was in favor of the wildtype or mutant allele and in identifying variants with high VAF where LOH may have been missed by copy number analysis.

Per sample mutations are available in tables S28 to S30.