Published in final edited form as:
Nat Genet. 2015 December ; 47(12): 1402–1407. doi:10.1038/ng.3441.

Clock-like mutational processes in human somatic cells

Ludmil B. Alexandrov[1,2,3], **Philip H. Jones**[1,4], **David C. Wedge**[1], **Julian E. Sale**[5], **Peter J. Campbell**[1,6], **Serena Nik-Zainal**[1,7], and **Michael R. Stratton**[1]

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridgeshire, United Kingdom

[2]Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States of America

[3]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States of America

[4]MRC Cancer Unit, Hutchison-MRC Research Centre, University of Cambridge CB2 0XZ, Cambridge, United Kingdom

[5]Medical Research Council Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, United Kingdom

[6]Department of Haematology, University of Cambridge, Cambridge CB2 0XY, United Kingdom

[7]Department of Medical Genetics, Addenbrooke's Hospital National Health Service (NHS) Trust, Cambridge CB2 0QQ, United Kingdom

Abstract

During the course of a lifetime somatic cells acquire mutations. Different mutational processes may contribute to the mutations accumulated in a cell, with each imprinting a mutational signature on the cell's genome. Some processes generate mutations throughout life at a constant rate in all individuals and the number of mutations in a cell attributable to these processes will be proportional to the chronological age of the person. Using mutations from 10,250 cancer genomes across 36 cancer types, we investigated clock-like mutational processes that have been operating in normal human cells. Two mutational signatures show clock-like properties. Both exhibit different mutation rates in different tissues. However, their mutation rates are not correlated indicating that the underlying processes are subject to different biological influences. For one signature, the rate of cell division may influence its mutation rate. This study provides the first survey of clock-like mutational processes operative in human somatic cells.

Correspondence and requests for materials should be addressed to L.B.A (lba@lanl.gov) and M.R.S. (mrs@sanger.ac.uk).
Author Contributions: L.B.A. and M.R.S. conceived the overall approach and wrote the manuscript. L.B.A., P.H.J., S.N.-Z. and M.R.S. carried out signatures and/or statistical analyses with assistance from D.C.W., J.E.S. and P.J.C.

URLs
COSMIC signatures website: http://cancer.sanger.ac.uk/cosmic/signatures
MathWorks mutational signatures framework: http://www.mathworks.com/matlabcentral/fileexchange/38724

Conflict of interests: M.R.S. and P.J.C. are founders, stock holders, and consultants for 14M Genomics Ltd. The remaining authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

INTRODUCTION

The mutational processes generating somatic mutations in normal cells are not well understood and quantification of their *in vivo* mutation rates is lacking for almost all human cell types. These metrics are likely to be fundamental to understanding of cancer development and ageing. Comprehensive investigation of *in vivo* somatic mutation rates will ultimately depend on accurate, single cell, whole genome sequencing of normal somatic cells. However, all cancers are clonal cell populations expanded from normal single cells. To a first approximation the catalogue of somatic mutations shared by most members of a cancer cell population is the set that was present in the progenitor cell of the final dominant clonal expansion of the cancer. This catalogue informs on the mutational processes to which the lineage of cells from the fertilised egg to that progenitor cell has been exposed[1]. Under a simple model this lineage has three phases: embryonic and foetal development; postnatal life in normally functioning differentiated cells; and after neoplastic transformation (Fig. 1).

During this lineage, some mutational processes may have acted in an episodic manner, generating mutations in bursts over short time periods. Others may have operated continuously, in a clock-like manner, generating mutations at a steady rate. For such clock-like mutational processes, the number of mutations acquired during embryonic and foetal development will be similar in cancers of the same type from different individuals, as this phase is of a fixed duration. Conversely, the same process operating in normally functioning cells during postnatal life will result in a mutation load that is proportional to the age of the person at the time the cancer is sampled, with more mutations present in older individuals (Fig. 1). The number of mutations acquired after initiation of neoplastic change will be unrelated to age of diagnosis but will depend upon the duration of the period between the first cancer driver mutation and initiation of the final dominant clonal expansion and, potentially, also upon changes to the mutation rate contingent upon acquiring the neoplastic phenotype. The latter features may be highly variable within and between cancer types.

Under this simple model, mutations with clock-like features in cancer genomes predominantly derive from the normal, post-natal part of the lineage. However, mutations from the developmental and/or neoplastic phases could obscure their clock-like features and affect estimation of the mutation rate during the normal postnatal phase. To evaluate this possibility we performed simulations which showed that the clock-like mutational processes can be detected, and the estimated mutation rates are relatively unaffected, unless the mutations generated during the developmental and/or neoplastic phases constitute the large majority of the total number of mutations in the cancer. Therefore, analysis of the several thousand cancer genomes thus far sequenced can provide a first survey of clock-like mutational processes operating in a wide range of normal human cell types.

Different mutational processes generate distinct combinations of mutation types in cancer genomes[2,3]. These characteristic imprints of mutational processes have been termed “mutational signatures”. We previously reported a mathematical approach and computational framework to extract mutational signatures from catalogues of somatic mutations from human cancers[4-6]. Using a 96 category classification of base substitutions based on the type of substitution and the bases immediately 5′ and 3′ to the mutated base,

we identified 21 mutational signatures operating over 30 cancer types[4]. Of these, the numbers of mutations associated with signature 1 correlated with age of cancer diagnosis for some cancer types[4].

Our previous analysis extracted mutational signatures separately from each cancer type and then quantified the mutations contributed by these signatures to each case of that cancer type. Many mutational signatures are found in multiple different cancer types and a central scientific question to address is the comparison of the contributions of such signatures across cancer types. However, a particular mutational signature found in multiple cancer types will be contaminated to differing extents by other signatures and by noise in each of the different cancer types. Hence, our previous approach did not allow accurate quantification of mutation rates for direct comparisons between cancer types. We have, therefore, reformulated the approach to derive a single consensus version of each signature and used these consensus signatures to estimate the number of mutations contributed to each cancer sample across all cancer types (Methods). Our refined approach was applied to a larger dataset of 7,329,860 somatic mutations from 10,250 cancer genomes (Supplementary Data 1 and 2) derived from diverse epithelial, mesenchymal, glial, haematopoietic and lymphoid cells that collectively constitute an extensive, albeit incomplete, sampling of normal cell types in the human body. This analysis has then allowed us to estimate mutation contributions to individual cancer cases across cancer types and hence enabled comparison of the clock-like mutation rates that reflect mutations in normal tissues.

RESULTS

Applying our refined approach to 10,250 cancer samples revealed the patterns of 33 distinct mutational signatures. We were able to perform validation for 29 of these 33 mutational signatures using our established methodology for validating mutational signatures[4]. This new analysis confirmed the patterns of the 21 previously identified mutational signatures[4] demonstrating the robustness of the computational approach. Additionally, examining this significantly larger dataset allowed us to disentangle the patterns of another eight distinct mutational signatures. A curated list of the validated mutational signatures and the cancer types in which they are present can be found at our COSMIC signatures website. Note that signatures 25, 29, and 30 are not part of the analysis presented here since the relevant samples were either cancer cell lines or lacked information about age of diagnosis. Further, the list of mutational signatures on our website does not include signatures of sequencing artefacts and signatures for which validation has not been performed. We have, however, included these mutational signatures in the current analysis and their patterns are shown in Supplementary Figure 1.

To identify mutational signatures showing clock-like behaviour, we first combined mutations and samples from all cancer types. Of the 33 signatures examined, signatures 1 and 5 showed a correlation between numbers of mutations and age of diagnosis and, for both, the numbers of mutations increased with age (signature 1, Spearman rank correlation 0.34, false discovery rate (FDR)–corrected for all 33 signatures q-value=4.7×10^{-162}; signature 5, Spearman rank correlation 0.13, false discovery rate (FDR)–corrected for all 33 signatures q-value=2.1×10^{-46}; combining the numbers of mutations attributed to signatures 1 and 5

resulted in Spearman rank correlation 0.37 and p-value=8.2×10^{-254}). No other mutational signature exhibited a statistically significant correlation (q-value < 0.05) with age of cancer diagnosis. The total number of somatic mutations in each sample (Fig. 1) also exhibited a correlation with age of diagnosis across all samples (Spearman rank correlation 0.37 and p-value=3.1×10^{-215}). However, after subtracting the numbers of signature 1 and 5 mutations, which in aggregate only account for 23% of the total number of mutations, no correlation was found (p-value=0.21) indicating that this is predominantly explained by signatures 1 and 5. C>T mutations at NpCpG trinucleotides (often termed CpG dinucleotides) constitute the major component of signature 1 and also showed a correlation with age (p-value=1.0×10^{-189}) (Fig. 2). Subtracting the numbers of C>T at NpCpG mutations from the numbers attributed to signature 1 left a residual correlation with age of cancer diagnosis (p-value= 1.4×10^{-19}) indicating that, in addition to C>T at NpCpG mutations, other components of this signature also behave in a clock-like manner.

26 out of 36 cancer types individually showed correlations with age (p-value < 0.05) for signature 1 and/or signature 5 mutations (Fig. 2, Table 1, and Supplementary Figure 2). Mutations associated with signature 1 were correlated with age of diagnosis in 17 of the cancer types, while mutations associated with signature 5 in 12 of the cancer types. In three cancer types (*viz.*, breast, low grade glioma, and glioblastoma) the mutational burdens of both signatures correlated with age of cancer diagnosis. Although some cancer types exhibited negative correlations, in all such cases the correlations were statistically not significant (Table 1). Similar to the analysis of all samples, no other mutational signature showed a correlation with age of diagnosis in any individual cancer type, while there was some correlation with total mutations and C>T mutations at NpCpG (Supplementary Data 3 through 5).

We then compared the signature 1 and 5 mutation rates between different tissue types. Signature 1 mutation rates showed substantial variation, being high in stomach (23.7 mutations per gigabase per year), colorectum (23.4), glioblastoma (19.8), oesophagus (19.6), medulloblastoma (16.1) and pancreas (14.7) compared to ovary (4.0 mutations per gigabase per year), breast (3.7), melanoma (3.2), myeloma (3.1) and pilocytic astrocytoma (0.65) (Fig. 3 and Supplementary Figure 3). In breast the rates were similar for oestrogen receptor positive (3.9 mutations per gigabase per year) and oestrogen receptor negative (3.1) cancers (Supplementary Figure 4).

Based on similarities of mutational signature, the mutational process underlying signature 1 is likely to be deamination of 5-methylcytosine at CpG dinucleotides leading to T:G mismatches which are not repaired prior to DNA replication[7]. It seems unlikely that the observed variation in signature 1 mutation rate between cell types is simply due to differences in the extent of CpG methylation, because this is similar in most cell types[3,8], although it could be due to differences in rates of cytosine deamination and/or T excision at T/G mismatches by thymine DNA glycosylase or mismatch repair.

It is notable, however, that many cancer types with high signature 1 mutation rates are derived from normal epithelia with high turnover, *e.g.*, stomach and colorectum (Supplementary Table 1; Supplementary Data 6; Supplementary Figure 5; p-value=0.0033).

Since DNA replication without prior repair will convert T:G mismatches arising from deamination of 5′-methylcytosine into C>T mutations, it is plausible that cell types with high mitotic rates exhibit higher mutation rates due to this mutational process. If correct, this interpretation indicates that the signature 1 mutation rate can serve as a clock registering the number of mitoses a cell has experienced during the lineage of cell divisions from the fertilised egg.

The signature 5 mutation rate also showed substantial variation between cancer types. It was high in kidney papillary cell (31.8 mutations per gigabase per year), neuroblastoma (25.8) and kidney clear cell (22.7) compared to breast (5.3 mutations per gigabase per year), kidney chromophobe (5.1), medulloblastoma (3.0) and acute myeloid leukaemia (2.8).

The mutational process underlying signature 5 is not well understood. It primarily features C>T and T>C transitions. Such mutations can be explained by replication of deaminated cytosine (uracil, which is read as thymine) and adenine (hypoxanthine, which is read as guanine resulting in A>G/T>C transition). However, in addition, the T>C mutations exhibit transcriptional strand bias, potentially indicating that some of these mutations arise from adducts subject to transcription coupled repair[9]. The signature 5 mutation rate is high in kidney clear cell and papillary cancers, which are thought to originate from kidney proximal tubular epithelium which absorbs metabolites, but low in kidney chromophobe tumours, which may arise from cells of the cortical collecting duct[10]. This raises the possibility that continuous exposure to a ubiquitous metabolic mutagen, which is actively reabsorbed in the kidney proximal tubule resulting in an elevated exposure in these cells, may underlie signature 5.

In some tumour types a correlation with age was not observed, despite substantial numbers of signature 5 mutations (*e.g.*, head and neck, colorectal and lung squamous cancers; Fig. 3), and thus the absence of correlation is unlikely to be due to limitations of statistical power. One possible explanation is that the mutational process underlying signature 5 is substantially activated by other factors during life or as part of the neoplastic phenotype in some tumour classes, thus obscuring the correlation between signature 5 mutations generated by the clock-like process and age.

Across tumour types, signature 1 and 5 mutation rates do not closely correlate with each other (Spearman rank correlation −0.08; p-value = 0.63). For example, in medulloblastoma the signature 1 rate is 16.1 and the signature 5 rate is 3.0 while in kidney papillary cancer the rates for signature 1 and 5 are −0.3 and 31.9 respectively (Table 1 and Fig. 3). Thus, the biological determinants of the mutation rates of the two processes may be different and cell proliferation rate may not be a major factor for signature 5 compared to its influence on signature 1.

DISCUSSION

Peering through the “cracked lens” of cancer genomes may obscure or distort the estimates of clock-like mutation rates of normal cells that are progenitors of the cancers. The data originate from dozens of laboratories, multiple sequencing platforms and many mutation-

calling algorithms. They include subclonal mutations, which occur after the last dominant clonal expansion, to different extents and are from samples with varying amounts of normal tissue contamination. The signature 1 and 5 mutation numbers have been estimated from mutational catalogues to which multiple other mutational processes have often contributed and may still be affected by their presence, despite extraction by our method. The simple, pragmatic classification of cancer types used is likely, in many instances, to hide greater complexity of biological subclass and each subclass could derive from a distinct type of non-neoplastic cell characterised by different signature 1 and 5 mutation rates. The mutation rate estimates are based on age of cancer diagnosis as a surrogate for the age of the driver mutation initiating the last clonal expansion and several years may intervene between these two points in time (and this period may differ between tumour types). As shown in the simulations, if substantial numbers of signature 1 and 5 mutations occur after neoplastic transformation they could obscure clock-like processes and affect the estimated mutation rates. Finally, the profiles of signatures 1 and 5 may be further refined in future and this may also affect estimates of mutation rate.

Signatures 1 and 5 demonstrate significant variability in the numbers of mutations per megabase even for samples of the same cancer type and age of diagnosis (Supplementary Data 2). While some of this variability may be attributable to limitations of the data and analysis described above it is also plausible that some reflects biological variation. For example, the rates of the clock-like mutational processes may vary between individuals depending on environment or lifestyle and inherited predisposition, and ancestor cells of some tumours may acquire mutator phenotypes for signatures 1 or 5 decades before the last clonal expansion. Future studies will be needed to evaluate the effects of these factors on the rates of clock-like mutational processes.

Remarkably, despite these multiple muddying influences, the clock-like nature of signatures 1 and/or 5 is visible in most cancer types. The proposition that these clock-like mutations derive from normal cells is supported by the observation that the profile of signatures 1 and 5 combined is very similar to the somatic mutational patterns observed in the small set of non-neoplastic human somatic cells thus far sequenced[11]. Moreover, a combination of signatures 1 and 5 also recapitulates the pattern of *de novo* mutations found in the human germline (data from refs.[12-14]), and this *de novo* germline pattern cannot be parsimoniously generated by other combinations of known mutational signatures (Supplementary Figure 6).

The results therefore provide the first survey of clock-like somatic mutational processes over a broad range of normal human cell types and quantification of the mutation rates exhibited by these mutational processes. They indicate that there are two clock-like mutational signatures, that both the signature 1 and signature 5 mutation rates differ widely between cell types, that the biological factors which determine these rates are different for signatures 1 and 5, that cell proliferation rate may be one of the dominant factors influencing the mutation rate of signature 1, and that signature 5 may be activated by non-clock-like influences and/or as part of the neoplastic process. Despite the ubiquity of both signatures in normal somatic cells, and their likely presence in the germline generating the sequence variation underlying human healthy and disease phenotypes, we have hardly any understanding of the biological processes underlying at least one of them, signature 5. These

signature 1 and 5 mutation rates will be refined over the next several years by the direct deployment of large-scale normal single cell sequencing and will provide the basis for future exploration of the range of mutational processes and their rates in human cells affected by mutagenic exposures, in precancerous neoplastic cells and in cells involved in disease processes other than cancer in which mutation rates may be affected.

ONLINE METHODS

Curation of freely available cancer samples

No data were generated for this study. Rather, data curation was performed to annotate freely available cancer genomes. Somatic mutations from 10,250 genome pairs (consisting of a cancer genome and the genome of a matched-normal tissue) were curated. 607 of the 10,250 matched-normal pairs had their whole-genome sequenced, while 9,643 were whole-exome sequenced. Data were retrieved from three sources: (i) the data portal of The Cancer Genome Atlas (TCGA), (ii) the data portal of the International Cancer Genome Consortium (ICGC), and (iii) previously published data. Information for each sample is provided in Supplementary Data 1. The somatic mutations for all samples are freely available and can be retrieved based on the information provided in Supplementary Data 1.

Filtering mutations, generating mutational catalogues and displaying signatures

This study relies on previously sequenced cancer genomes and on the subsequently used bioinformatics to identify cancer specific somatic mutations. The data were filtered prior to analysis as previously described in ref. [4] and mutational catalogues were generated using ENSEMBL Core APIs for human genome build GRCh37.

The prevalence of somatic mutations in each sample was estimated based on a haploid human genome after filtering as previously done in ref [4].

Mutational signatures are displayed based on the observed trinucleotide frequency of the human genome.

Refined approach for deciphering mutational signatures

The mutational catalogues of all samples were examined following two steps. Initially, *de novo* extraction was performed to derive the set of novel consensus mutational signatures. Briefly, mutational signatures were deciphered independently for each of the 36 cancer types using our MATLAB framework[4]. The computational framework for deciphering mutational signatures is freely available for download from MathWorks. The algorithm deciphers the minimal set of mutational signatures that optimally explains the proportion of each mutation type found in each catalogue and then estimates the contribution of each signature to each mutational catalogue. Mutational signatures were extracted separately for genomes and exomes. Mutational signatures extracted from exomes were normalized to the trinucleotide frequency of the human genome. All mutational signatures were clustered using unsupervised agglomerative hierarchical clustering and a threshold was selected to identify the set of consensus mutational signatures. Misclustering of signatures was avoided as previously described in ref [4]. Overall, we identified 33 consensus mutational signatures.

Signatures 1 through 28 (note, signature 25 is not found in this dataset) were validated and, thus, these processes most likely reflect biological processes. Signatures R1 through R3 were previously found in ref. [4] and attributed to sequencing artefacts. We were not able to perform validation for signatures U2 through U4 since we did not have access to the respective biological samples or BAM files.

The *de novo* extraction was used to identify the set of consensus mutational signatures across the examined 10,250 samples. This first step of extracting mutational signatures and generating consensus patterns follows the previously introduced approach in ref [4]. However, our prior methodology did not use consensus mutational signatures to evaluate their contributions in each sample, thus not allowing accurate comparison of mutation rates between different cancer types. To address this limitation, we refined our approach by introducing another step of analysis, which is focused on accurately estimating the numbers of mutations associated with each consensus signature in each sample. We usually refer to this number of somatic mutations as either the "contribution" of a mutational signature or the "exposure" to a mutational signature. Calculating the contributions of all mutational signatures was performed by estimating the number of mutations associated to the *consensus* patterns of the signatures of all operative mutational processes in each cancer sample. This approach allows direct comparison between cancer types because identical signatures were used to estimate the contributions in each cancer type. More specifically, all consensus signatures were examined as a set P containing 33 vectors

$$P=\left\{\begin{bmatrix} p_1^1 \\ \vdots \\ p_1^{96} \end{bmatrix}, \begin{bmatrix} p_2^1 \\ \vdots \\ p_2^{96} \end{bmatrix} \cdots \begin{bmatrix} p_{32}^1 \\ \vdots \\ p_{32}^{96} \end{bmatrix}, \begin{bmatrix} p_{33}^1 \\ \vdots \\ p_{33}^{96} \end{bmatrix}\right\},$$

where each of the vectors is a discrete probability density function reflecting a consensus mutational signature. The 96 nonnegative components of each vector correspond to mutation types (*i.e.,* substitutions and their immediate sequencing context) of the signatures. The contributions of the signatures were estimated independently for each of the 10,250 samples with a subset of consensus mutational signatures. For each sample, the estimation algorithm consists of finding the minimum of the Frobenius norm of a constrained linear function (see below for constraints) for a set of vectors $S_{1..q}, q \leq 33$, belonging to the subset Q, where $Q \subseteq P$.

$$\min \left\| \overrightarrow{M} - \sum_{i=1}^{q} \left(\overrightarrow{S_i} \times E_i \right) \right\|_2^F \quad (1)$$

Q is determined based on the known operative mutational processes in the cancer type of the examined sample from the signature extraction process described above. For example, for any neuroblastoma sample, Q will contain signatures 1, 5 and 18 since these are the only known signatures of mutational processes operative in neuroblastoma (Supplementary Data 2). In equation (1), $\overrightarrow{S_i}$ and $\overrightarrow{M}$ represent vectors with 96 nonnegative components reflecting, respectively, a consensus mutational signature and the mutational catalogue of a sample. Hence, $\overrightarrow{S_i} \in \Re_+^{96}$ while $\overrightarrow{M} \in \mathrm{N}_0^{96}$. Further, both vectors have known numerical values either from the *de novo* extraction (*i.e.,* $\overrightarrow{S_i}$) or from generating the original mutational catalogue of

the sample (*i.e.*, $\overrightarrow{M}$). In contrast, E_i corresponds to an unknown scalar reflecting the number of mutations contributed by signature $\overrightarrow{S_i}$ in the mutational catalogue $\overrightarrow{M}$.

Minimization of equation (1) is performed under several biologically meaningful constraints. The set of vectors in the examined set Q is constrained based on previously identified biological features of the consensus mutational signatures. For example, consensus signature 6 causes high levels of indels at mono/polynucleotide repeats[4]. Thus, this mutational signature will be excluded from the set Q when the mutational catalogue of an examined sample has only a few such indels. Similarly, there are signatures associated with other types of indels, transcriptional strand bias, dinucleotide mutations, hypermutator phenotypes, *etc.* and these signatures are included in the set Q only when the sample in question exhibits one or more of these features. Lists of features associated with mutational signatures can be found in ref [4]. In addition to sample specific constraints to the set Q, equation (1) was universally constrained in regards to the parameter E_i. These constraints can be mathematically expressed as $0 \leq E_i \leq \left\|\overrightarrow{S_i}\right\|_1, i=1..q$, and $\sum_{i=1}^{q} E_i = \left\|\overrightarrow{S_i}\right\|_1$. All results for the contributions of all operative signatures in all samples from the hitherto described approach are provided in Supplementary Data 2, while the original somatic mutations can be found using Supplementary Data 1.

Factors influencing signature extraction

We have previously simulated data to describe a plethora of factors influencing the accuracy mutational signatures extraction[6]. Such factors include the number of available samples, the number of somatic mutations in a sample, the number of mutations contributed by different mutational signatures, the similarity between the patterns of the signatures of mutational processes operative in cancer samples, as well as the computational limitations of our framework. Nevertheless, in the past three years, our framework has proven robust and has described multiple similar and validated signatures across the spectrum of human cancer[3-5,15-22].

Patterns of signatures 1 and 5

In a previous analysis[4], we extracted 21 mutational signatures and noted that signatures 1A and 1B correlate with age of diagnosis for some cancer types. Further, we noted that since signatures 1A and 1B "are almost mutually exclusive among tumour types they probably represent the same underlying process, with signature 1B representing less efficient separation from other signatures in some cancer types"[4]. In our previous report, we referred to these two signatures as a single signature termed signature 1A/B. In the current analysis, encompassing ~50% more data and a refined algorithm, signature 1A is found in more cancer types, including some in which signature 1B was seen previously. A detailed examination of the pattern of signature 1B revealed that this mutational signature is a linear combination of signatures 1A and 5. More specifically, a combination of signatures 1A and 5 can be used to account for 0.97 of the pattern of signature 1B (where 1.00 is perfect correlation), and no other combination of signature can be used to explain signature 1B. Thus, in the current manuscript signature 1B is no longer referred to and, in cancer types

Europe PMC Funders Author Manuscripts

from which it was extracted, signatures 1A and 5 have been reintroduced to assess mutation contributions.

Robustness and reproducibility of mutational signatures

In this analysis, we use an elaborated version of our framework for extracting mutational signatures and apply it to a much larger dataset. Comparison between the set of previously extracted mutational signatures[4] and the set of mutational signatures found here reveals both stability and reproducibility of mutational signatures. This reproducibility can be observed by comparing the mutational signatures on the COSMIC signatures website with the ones from ref. 4. Further, the similarity can be also quantified using a cosine similarity as previously done in ref. [6]. The cosine similarity between any combination of signatures that were derived in ref. 4 and also found in this analysis (*i.e.,* signatures 1 through 21) is more than 0.97, where a similarity of 1.00 is an exact match.

Statistical analysis of relationships between age and mutations

Global analysis was performed for the identified 33 mutational signatures across all samples in all cancer types. Zero mutations were attributed to all signatures that were not found in a sample. The data heteroscedasticity and presence of outliers mandates the use of an appropriate statistical approach[23]. We leveraged robust linear regression to evaluate linear dependencies between the numbers of mutations associated with each mutational signature across all examined samples and the ages of cancer diagnosis of these samples. The calculated p-values from the applied robust regression were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure. Only signatures 1 and 5 exhibited statistically significant correlation (q-value < 0.05) with age of cancer diagnosis.

Each cancer type was examined independently for a linear dependence between the ages of cancer diagnosis for the curated samples in that cancer type and the numbers of mutations attributed to each of the signatures of the operative mutational processes in that cancer type. Since most traditional or generalized linear regression approaches are very sensitive to outliers[23] and since many of the examined cancer samples are hypermutators (*i.e.*, outliers), we leveraged a robust regression model. The robust regression iteratively reweights least squares with a bi-square weighting function and overcomes some (if not most) of the limitations of traditional approaches[24-26]. Similarly, we report results using Spearman's rank correlation coefficient since it is more robust to data outliers when compared to Pearson's product-moment correlation coefficient[27]. It should be noted that, while samples with missing information about their age of cancer diagnosis were excluded from this analysis, these samples were used in the *de novo* extraction of mutational signatures and the subsequent estimation of the signatures' contributions.

Each signature was examined separately in each cancer type in which that signature was identified. The examination was based on: a robust linear regression model that estimates the slope of the line and whether this slope is significantly different from a horizontal line with a slope of zero (F-test; p-value<0.05) as well as by calculating the Spearman's rank correlation coefficient. While robust linear regression models provide confidence intervals and p-values, we decided to take a more conservative approach and report results after

bootstrapping the data. Bootstrapping (*i.e.,* random sampling with replacement) was used to derive measures of accuracies: the best fit for the slope and the slope's 95% confidence intervals. In total, we performed 100,000 bootstrapping iterations per signature per cancer type (total of ~2×10^7 iterations). Each of the iterations for which the robust regression returned a p-value<0.05 was considered statistically significant, while iterations with p-value>=0.05 were considered not statistically significant. The overall p-value per signature per cancer type was calculated as follows: $\frac{NumberOfNotSignificantIterations+1}{100,000+1}$. It should be noted that the number of iterations limits the minimum possible p-value, in this case 9.99E-06, and p-values reported to be equal to 9.99E-06 are most likely lower. The reported p-values and confidence intervals are the ones after applying the bootstrapping procedure. MATLAB code for calculating the p-values across individual cancer types is provided in Supplementary Materials.

The results of estimating the line's slope by robust regression and the Spearman's rank correlation coefficient for each of the signatures in each of the cancer types can be found in Supplementary Data 3. As before, we have used a p-value<0.05 to identify statistically significant dependencies. However, this cut-off is arbitrary and summarized results using different cut-off values are shown in Supplementary Figure 2. Examination of individual cancer types is based on the hypothesis that signatures 1 and 5 are the only signatures reflecting the activity of clock-like mutational processes. This hypothesis was constructed by examining the activity of all signatures across all cancer types (signature 1, false discovery rate (FDR)–corrected for all 33 signatures q-value=4.7×10^{-162}; signature 5, false discovery rate (FDR)–corrected for all 33 signatures q-value=2.1×10^{-46}). Thus, for our analysis of individual cancer types, the p-values reported in the main manuscript have not been corrected for multiple hypothesis testing. Nevertheless, for consistency, we have provided p-values corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure for each cancer type in Supplementary Data 3. It should be noted that using FDR-corrected p-values to evaluate the significance of the analysis does not affect the overall message of the manuscript.

Lastly, we also evaluated (following the hitherto described approach) whether there is a linear dependency between the total numbers of somatic mutations and/or the C>T mutations at CpG dinucleotides and the ages of cancer diagnosis. Similarly, this examination was done separately for each of the cancer types and the results from the analysis can be found in Supplementary Data 5.

Evaluating the robustness and limitations of the performed analysis with simulated data

A myriad of known and unknown processes may be affecting the performed analyses. Some of these include: data generation by different institutes and laboratories, contamination of subclonal mutations, endogenous or exogenous factors affecting the rates of signatures 1 and 5, inaccuracies of the patterns of signatures 1 and 5, mutations generated during the developmental and/or neoplastic phases, limitations of the signatures extraction algorithm, small numbers of samples and/or somatic mutations, misannotation of samples, *etc.* In

principle, quantifying the overall error introduced by even a subset of these processes is impractical.

To evaluate the robustness and limitations of our analysis, we simulated data with two types of mixture noise: (i) noise affecting the *bona fide* somatic mutations associated with a clock-like signature of a mutational process operative in a sample, and (ii) noise affecting the age of cancer diagnosis of a sample. It was assumed that the mixture of all factors affecting the *bona fide* number of mutations associated with a clock-like signature in a sample reflects a mixture of random processes and, thus, it can be approximated by white additive Gaussian noise. Further, folded normal Gaussian noise (*i.e.,* half-normal distribution) with mean value of 2 years and standard deviation of 4 years was added to the age of cancer diagnosis of a sample. This noise reflects average cancer detection within 2 years of neoplastic initiation with cancers detected in 84% and 98% of patients within 6 and 10 years, respectively. The distribution is half-bounded since a cancer cannot be detected before it has occurred.

Clock-like mutational signatures were simulated in 100 cancers. The ages of diagnosis of the cancers were sampled with replacement from the data in Supplementary Data 1, while the mutational rates per year per gigabase (*i.e.,* slope) were taken from a uniform distribution between the minimum and maximum statistically significant rates detected by the performed analysis (Supplementary Data 3). In total, 17 simulation scenarios were performed each with different percentages of white additive Gaussian noise (Supplementary Figure 7). The noise to *bona fide* somatic mutations was varied between 0% and 200%, where 0% reflects no noise and 200% corresponds to twice as much noise compared to genuine somatic mutations. Note that most cancer genomics papers report sensitivity and specificity rates of more than 90% and thus the false positive rates derived from our simulations are probably over-pessimistic. In all scenarios, the noise added to the ages of diagnosis followed the hitherto described folded normal distribution. Each simulation scenario was repeated 1,000 times and the simulated data were analysed to identify clock-like mutational signatures in exactly the same way as the experimental data used in this study. Any iteration with a statistically significant p-value for a slope (p-value<0.05) in which the simulated slope was within +/− 10% of the derived slope or within the 95% confidence intervals of the derived slope was considered a genuine detection and, thus, a true positive result. In contrast, any other iteration with a statistically significant p-value for a slope (p-value<0.05) that did not satisfy the abovementioned conditions was considered a false positive result.

The results from the 17 performed scenarios revealed that when the noise levels are less than 35%, our analysis is able to find the genuine slopes in ~90% of the iterations while yielding no more than 0.55% false-positives (Supplementary Figure 7). Increasing the noise levels does not increase the number of false positives but rather reduces the number of genuinely detected slopes (*i.e.*, true positives). Our simulations indicate that the confidence intervals of the majority of detected slopes include the genuine slope of a clock-like mutational signature, while the approach used yields few false positives.

Displaying age of diagnosis and clock-like mutational signatures

Linear relationships between the ages of cancer diagnosis and the mutations attributed to mutational signatures are displayed only for signatures 1 and 5 as no other mutational

signature displayed statistically significant correlations (Supplementary Data 3). These linear relationships are displayed both for the average mutational burden attributed to a signature (Fig. 3) as well as for all individual mutational burdens attributed to a signature (Supplementary Figure 3). In both cases, the displayed slopes and their confidence intervals are those derived by the hitherto described analysis and do not depend on the choice of depiction. For brevity, linear relationships are displayed for only 27 of the 36 analysed cancer types. Nevertheless, the data provided in Supplementary Data 2 and 3 can be used to display all linear relationships.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust (grant number 098051). S.N.-Z. is a Wellcome-Beit Prize Fellow and is supported through a Wellcome Trust Intermediate Fellowship (grant WT100183MA). P.J.C. is personally funded through a Wellcome Trust Senior Clinical Research Fellowship (grant WT088340MA). J.E.S. is supported by an MRC grant to the Laboratory of Molecular Biology (MC_U105178808). L.B.A. is supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. P.H.J is supported by the Wellcome Trust, a Medical Research Council Grant-in-Aid and Cancer Research UK (Programme grant C609/A17257). This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396. Research performed at Los Alamos National Laboratory was carried out under the auspices of the National Nuclear Security Administration of the United States Department of Energy. We would like to thank Matthew E. Hurles and Richard Durbin for early discussions about the performed analyses. We would like to thank The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), and the authors of all previous studies cited in Supplementary Data 1 for providing free access to their somatic mutational data.

REFERENCES

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–24. [PubMed: 19360079]
2. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014; 24:52–60. [PubMed: 24657537]
3. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014; 15:585–98. [PubMed: 24981601]
4. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21. [PubMed: 23945592]
5. Nik-Zainal S, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 149:979–93. [PubMed: 22608084]
6. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013; 3:246–59. [PubMed: 23318258]
7. Bell SP, Dutta A. DNA replication in eukaryotic cells. Annu Rev Biochem. 2002; 71:333–74. [PubMed: 12045100]
8. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013; 14:R115. [PubMed: 24138928]
9. Fousteri M, Mullenders LH. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. Cell Res. 2008; 18:73–84. [PubMed: 18166977]
10. Davis CF, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014; 26:319–30. [PubMed: 25155756]
11. Welch JS, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012; 150:264–78. [PubMed: 22817890]

12. Kong A, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; 488:471–5. [PubMed: 22914163]
13. Michaelson JJ, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012; 151:1431–42. [PubMed: 23260136]
14. Conrad DF, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011; 43:712–4. [PubMed: 21666693]

REFERENCES FOR ONLINE METHODS

15. Behjati S, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature. 2014; 513:422–5. [PubMed: 25043003]
16. Bolli N, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nat Commun. 2014; 5:2997. [PubMed: 24429703]
17. Ju YS, et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife. 2014; 3 [PubMed: 25271376]
18. Murchison EP, et al. Transmissible [corrected] dog cancer genome reveals the origin and history of an ancient cell lineage. Science. 2014; 343:437–40. [PubMed: 24458646]
19. Nik-Zainal S, et al. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nat Genet. 2014; 46:487–91. [PubMed: 24728294]
20. Gerlinger M, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Genet. 2014; 46:225–33. [PubMed: 24487277]
21. Yates LR, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015; 21:751–9. [PubMed: 26099045]
22. Wagener R, et al. Analysis of mutational signatures in exomes from B-cell lymphoma cell lines suggest APOBEC3 family members to be involved in the pathogenesis of primary effusion lymphoma. Leukemia. 2015; 29:1612–5. [PubMed: 25650088]
23. Barnett, V.; Lewis, T. Outliers in statistical data. Vol. xvii. Wiley, Chichester; New York: 1994. p. 584
24. Holland PW, Welsch RE. Robust Regression Using Iteratively Reweighted Least-Squares. Communications in Statistics: Theory and Methods. 1977; A6:813–827.
25. Huber, PJ.; Ronchetti, E. Robust statistics. Vol. xvi. Wiley; Hoboken, N.J.: 2009. p. 354
26. Street J, Carroll R, Ruppert D. A Note on Computing Robust Regression Estimates Via Iteratively Reweighted Least Squares. The American Statistician. 1988; 42
27. Abdullah MB. On a Robust Correlation Coefficient. The Statistician. 1990; 39:455–460.

Europe PMC Funders Author Manuscripts

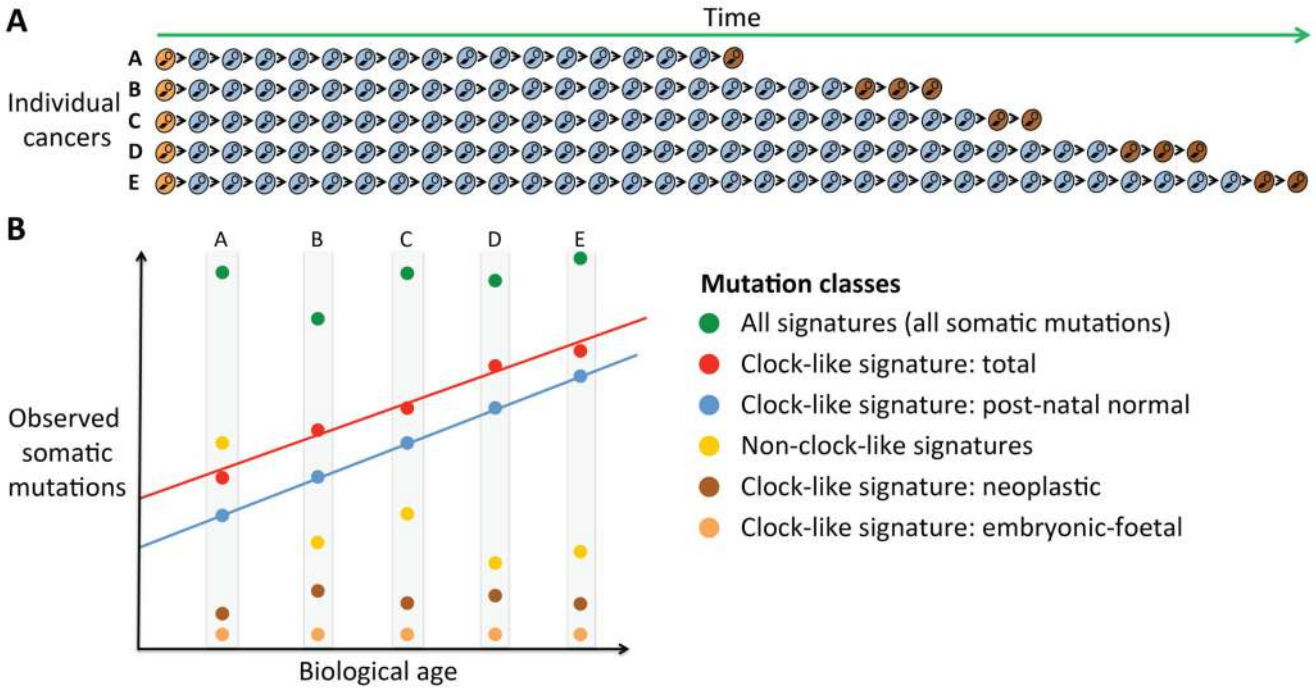

Figure 1. A model for the accumulation of somatic mutations in cancers

(a) Cell lineages leading from fertilised egg to cancer cell in five different individuals with cancer; A, B, C, D and E. Orange: embryonic/foetal cell divisions; blue, postnatal divisions of normal cells; brown, cell divisions post-neoplastic change. ***(b)*** Accumulation of somatic mutations due to clock-like and non-clock-like mutational signatures in the same five patients. The correlation between age and somatic mutations due to a clock-like mutational process operating in normal postnatal cells is detectable using the mutations found in cancers, with the rate relatively unaffected, if the number of mutations acquired during the embryonic/foetal and neoplastic phases is limited. Note that this figure is provided as a simple illustration of the activity of clock-like mutational processes and it is not intended to be a realistic representation of actual cancer samples. In reality, the numbers of cellular divisions will be tissue dependent and the numbers of neoplastic mutations may be many folds of magnitude higher.

Europe PMC Funders Author Manuscripts

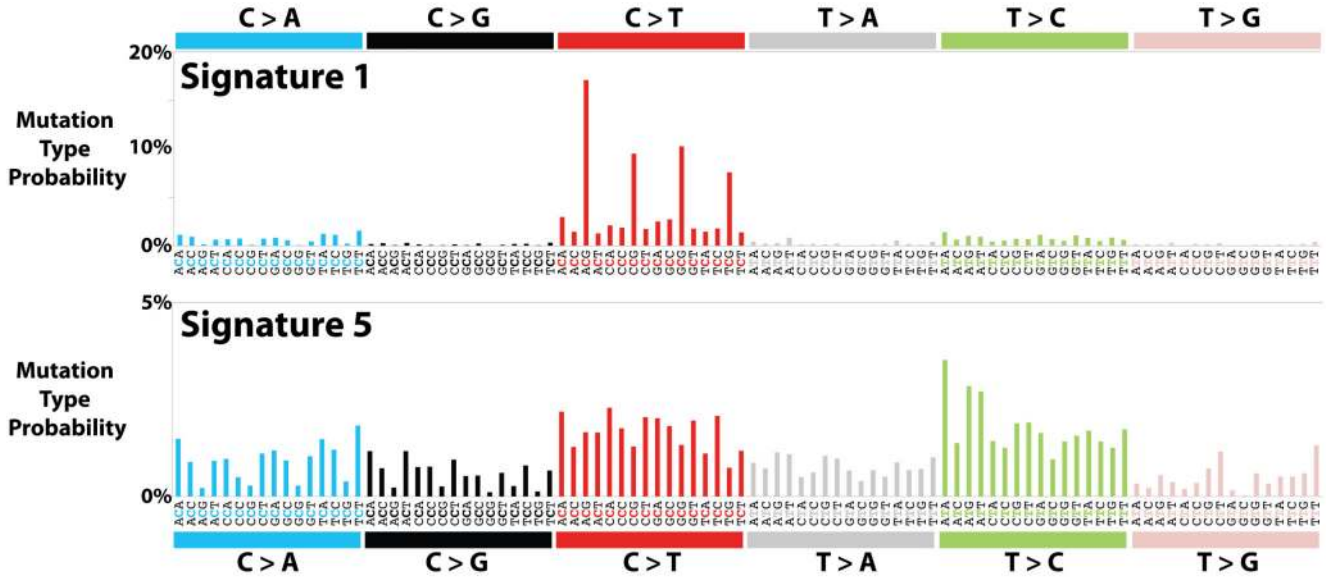

Figure 2. Patterns of mutational signatures 1 and 5

The signatures are displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 5′ and 3′ to the mutated base. The probability bars for the six substitution classes are displayed in different colours. The mutation types are on the X-axes, whereas Y-axes show the percentage of mutations in the signature attributed to each mutation type. Signatures are displayed on the basis of the trinucleotide frequencies of the whole human genome.

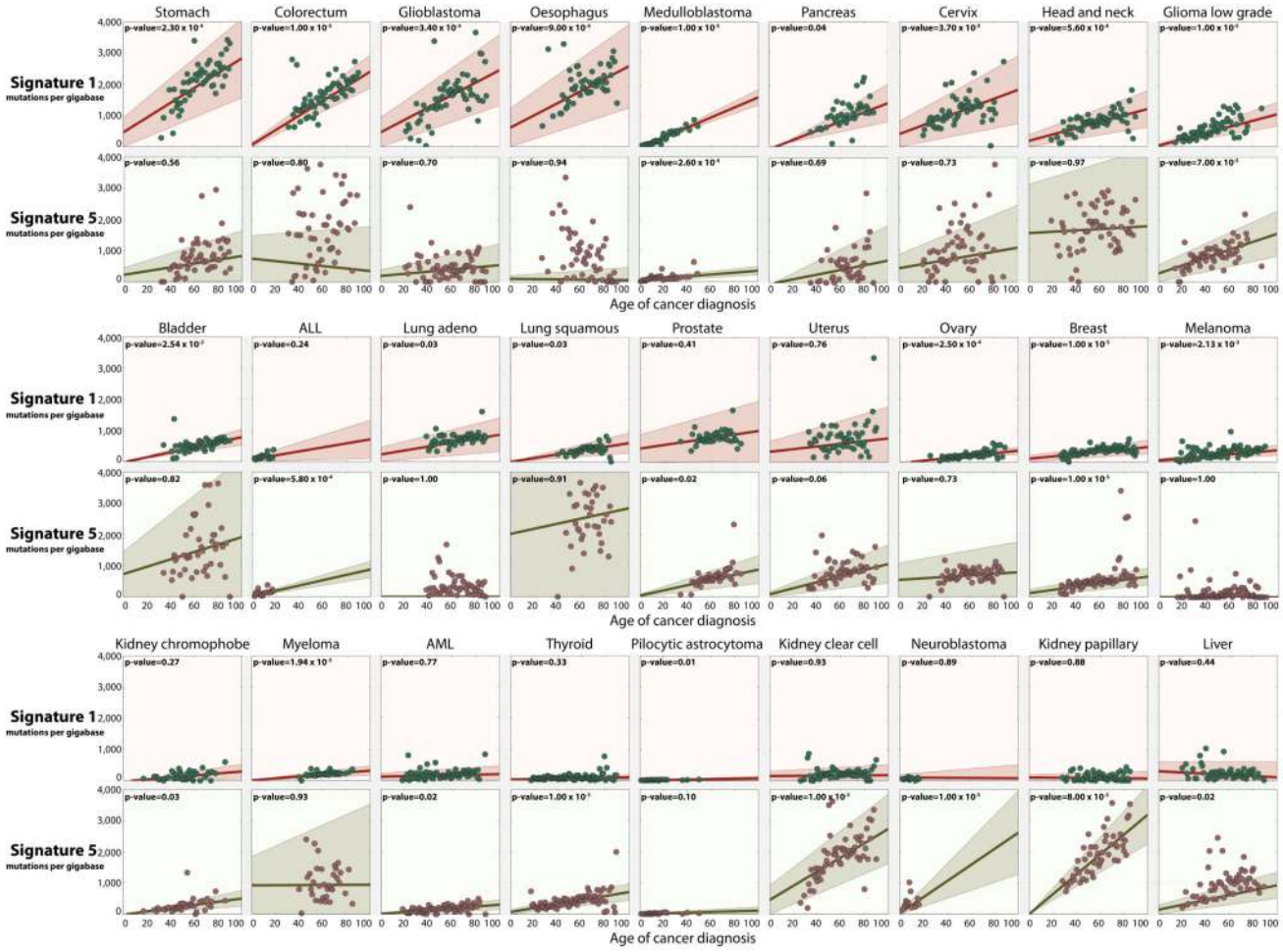

Figure 3. Correlations between ages of cancer diagnosis and mutations attributed to signatures 1 and 5

Y-axes show numbers of somatic substitutions per gigabase attributed to either signature 1 or signature 5, while X-axes show ages of cancer diagnosis. Each panel corresponds to a cancer type and panels are sorted in a decreasing order of the estimated slopes for signature 1. Each dot represents the median number of somatic mutations for all cancers of a given age. Red and green lines show best estimates for the slopes, *i.e.*, mutation rates, of signature 1 and 5 respectively. 95% confidence intervals for the slopes are shown in lighter green and lighter red shading for signatures 1 and 5 respectively. Note that for several cancer types slopes extend far beyond the available data points; this representation is not intended to be a prediction but rather it is done for consistent presentation across all panels in the figure. Slopes and p-values are also provided in Table 1. Panels showing mutational burdens in individual samples in each of the cancer types are provided in Supplementary Figure 3. Furthermore, a supplementary figure depicting the slopes for each cancer type is provided (Supplementary Figures 8 through 43).

Europe PMC Funders Author Manuscripts

Table 1

Rates of somatic substitution accumulation for clock-like mutational signatures

Somatic substitutions per gigabase per year for signatures 1 and 5 for all examined cancer types including their p-values and the number of examined samples in each cancer type. Rates of mutation accumulation and p-values for all mutational signatures in all cancer types are provided in Supplementary Data 3.

Cancer Type	Number of samples	Signature 1		Signature 5	
		Slope	P-value	Slope	P-value
Acute lymphoid leukaemia (ALL)	141	6.45	**0.24**	8.55	**5.80×10^{-4}**
Acute myeloid leukaemia (AML)	202	0.80	**0.77**	2.89	**0.02**
Adrenocortical carcinoma	92	2.56	**0.90**	3.94	**0.78**
Bladder cancer	238	8.07	**2.54×10^{-3}**	11.87	**0.82**
Brain adult lower grade glioma	465	10.02	**1.00×10^{-5}**	12.70	**7.00×10^{-5}**
Breast cancer	1,170	3.71	**1.00×10^{-5}**	5.31	**1.00×10^{-5}**
Cervical cancer	198	14.14	**3.70×10^{-3}**	6.57	**0.73**
Chronic lymphocytic leukaemia (CLL)	131	−1.45	**0.50**	5.52	**0.07**
Colorectal cancer	559	23.43	**1.00×10^{-5}**	−3.97	**0.80**
Glioblastoma multiforme	332	19.85	**3.40×10^{-4}**	3.44	**0.70**
Head and neck cancer	591	10.20	**5.60×10^{-4}**	2.20	**0.97**
Kidney chromophobe	65	3.18	**0.27**	5.16	**0.03**
Kidney renal clear cell carcinoma	468	0.26	**0.93**	22.75	**1.00×10^{-5}**
Kidney renal papillary cell carcinoma	169	−0.29	**0.88**	31.86	**8.00×10^{-5}**
Liver cancer	290	−1.93	**0.44**	7.81	**0.02**
Lung adenocarcinoma	795	6.30	**0.03**	0.00	**1.00**
Lung small cell carcinoma	69	0.6	**0.99**	5.58	**0.81**
Lung squamous cell carcinoma	176	6.00	**0.03**	8.22	**0.91**
Lymphoma B-cell	24	0.90	**0.34**	5.46	**0.05**
Medulloblastoma	100	16.16	**1.00×10^{-5}**	3.06	**2.60×10^{-4}**
Melanoma	514	3.25	**2.13×10^{-3}**	0.00	**1.00**
Multiple myeloma	69	3.11	**1.94×10^{-3}**	0.17	**0.93**
Nasopharyngeal carcinoma	55	2.62	**0.87**	−4.44	**0.71**
Neuroblastoma	231	−0.23	**0.89**	25.80	**1.00×10^{-5}**
Oesophageal cancer	329	19.66	**9.00×10^{-4}**	−0.42	**0.94**
Ovarian cancer	466	4.01	**2.50×10^{-4}**	2.42	**0.84**
Pancreatic cancer	231	14.73	**0.04**	7.47	**0.69**
Paraganglioma	179	1.85	**0.08**	2.49	**0.06**
Pilocytic astrocytoma	101	0.65	**0.01**	1.05	**0.10**
Prostate cancer	520	5.62	**0.41**	8.31	**0.02**
Stomach cancer	472	23.73	**2.30×10^{-4}**	6.04	**0.56**
Thyroid cancer	404	0.66	**0.33**	6.39	**1.00×10^{-5}**

Europe PMC Funders Author Manuscripts

Cancer Type	Number of samples	Signature 1		Signature 5	
		Slope	P-value	Slope	P-value
Urothelial carcinoma	26	−4.85	**0.94**	−15.75	**0.75**
Uterine carcinoma	241	4.28	**0.76**	9.68	**0.06**
Uterine carcinosarcoma	57	4.51	**0.82**	5.53	**0.84**
Uveal melanoma	80	1.97	**0.55**	2.26	**0.77**