Protein structural consequences of DNA mutational signatures: A meta-analysis of somatic variants and deep mutational scanning data

Joseph Chi-Fung Ng
Randall Centre for Cell and Molecular Biophysics
King's College London
London SE1 1UL
United Kingdom
joseph.ng@kcl.ac.uk

Franca Fraternali
Randall Centre for Cell and Molecular Biophysics
King's College London
London SE1 1UL
United Kingdom
franca.fraternali@kcl.ac.uk

May 27, 2021

Abstract

Signatures of DNA motifs associated with distinct mutagenic exposures have been defined for somatic variants, but little is known about the consequences different mutational processes pose to the cancer cell, particularly the distribution of the resulting variants in the implied proteins and their structural regions (surface, core, interacting interface). Here we first compare the protein-level consequences of six mutational signatures (Aging, APOBEC, POLE, UV, 5-FU and Platinum) characterised by clear DNA motif preferences. By mapping individual substitution events observed in tumours to three-dimensional protein structures, we show that these common somatic mutational signatures are biased against the protein core, consistent with the lower tolerability of substitutions at such structurally important regions. On the other hand, deep mutational scanning (DMS) data allow us to probe the 'dark matter' of somatic mutational landscape, exploring variants which are otherwise removed in purifying selection. A computational DMS analysis identifies mutational contexts (5'-G/C[T>G]A/G-3') which are associated with damaging mutations, by altering physicochemical characteristics of amino acids at the protein core. We argue that comprehensive DMS analysis can contribute to classification of variants according to their true impact to the stability/activity of the affected protein, decoupling this from pathogenicity prediction offered by conventional variant impact classifiers.

***K*eywords** Somatic mutations · Mutational signatures · Protein structures · Structural bioinformatics

Introduction

All somatic cells are vulnerable to genetic variants contributed by various endogenous and exogenous processes. The concept of "mutational signatures" associates each of these mutagenic processes to the DNA alterations it generates, by specifying the sequence contexts immediately 5' and 3' to base substitution events [1, 2]. Algorithms to extract such signatures [1, 2], together with sequencing studies of *in vitro* systems exposed to specific mutagens [3, 4], allow for cataloguing the footprints different mutagenic processes leave on the somatic genome [5]. These studies link various epidemiological risk factors of cancer to their molecular effects at the single-base resolution [2, 6]. Mutational signatures also highlight hitherto under-appreciated endogenous processes which contribute somatic mutations, most notably the APOBEC3 (apolipoprotein B mRNA editing catalytic polypeptide-like 3) enzymes. A large body of work has established the expression pattern of these enzymes [7] and the genomic contexts of APOBEC3 mutagenesis [8], including their preferences in terms of 3D DNA conformational features [9].

While much research has focused on the aetiology of various mutational signatures, little is known about the consequences of the genetic variants generated by different mutagenic processes. This issue remains challenging to study due to the difficulty in attributing a given DNA substitution event to a specific mutagenic process: existing signature discovery methods do not explicitly assign causal mutagenic process to individual substitutions [2]; moreover, a mutational signature is often collation of several DNA motifs; for motifs which are associated with numerous signatures, attribution of causal mutagenic process becomes conceivably more challenging [1, 2]. As a result, existing methods which annotate consequences of genetic variants and predict cancer driver genes have yet to consider the responsible "culprit" of specific damaging mutations. This, however, could have great significance and benefit to advance our understanding of cancer biology, in rationalising the mutational landscape of tumours based on known lifestyle and clinicopathological features, and monitoring the acquisition of mutations attributable to drug regimens that could present as side-effects in patients who undertake these treatments.

There are important considerations in the approach to comprehensively survey the impact of variants. Notably, biological samples are constantly under evolutionary selection pressure which removes genetic variants that are deemed harmful to cell viability [10]. In somatic evolution, those variants which enhance viability are retained and positively selected; such signals have been exploited to identify cancer driver genes [11]. Sequencing experiments performed on biological samples therefore present a biased picture in terms of the consequences genetic variants bring to the cells; conceivably, mutational processes which generate the most damaging variants - should such processes be active in the cell in the first place - would be difficult to sample. Such variants could be understood as part of what we term as the "dark matter" of the somatic mutational landscape, arising from the under-sampling of such variants in observational studies. One way to address this issue is to perform "Deep Mutational Scanning" (DMS) [12] where every position in the protein is mutated to any other

19 amino acids either *in vitro* or *in silico*, followed by assays to measure the stability and/or activity of the mutants. Experimental DMS have been applied on several cancer-related proteins and domains [13, 14, 15]; computationally, variant impact predictors which incorporate sequence conservation (large protein/domain multiple sequence alignments), structural/physicochemical features (protein secondary/tertiary structures) and protein dynamics (imposing network models onto three-dimensional structures to account for molecular vibrations) have been developed and applied in a DMS context [16, 17]. To link impact assessment from DMS data to DNA mutational signatures one would require mapping of proteins to their corresponding coding sequences; to this date, such analyses which identify damaging mutational signatures have yet to be performed.

Here we capitalise upon our recent database (ZoomVar [18]) on mapping genetic variations to protein structures, and explore the protein-level consequences of variants attributable to common somatic mutational signatures. We examine the accumulation of somatic mutations in tumours attributable to these signatures across protein surface, core and interacting interface, and contrast the impact of these mutations with experimental and computational DMS profiles, comparing the consequences of mutations occurring in other DNA motifs which are otherwise under-represented in the observed mutational profiles.

Results

Mapping protein structural impact of mutagenic processes

In this study we focus on six mutagenic processes characterised by clear and specific patterns of DNA sequence motifs (Figure 1A). These cover well-studied endogenous and exogenous mutagenic processes (Aging, APOBEC3, ultraviolet light [UV], and hypermutation caused by *POLE* mutations [herafter "POLE"]) [1] as well as common therapeutic strategies applied in cancer treatment (5'-fluorouracil [5-FU] and Platinum-based therapies) [5]. Since the 20 amino acids differ in terms of their preference in different protein structural regions, we reason that in coding sequences the distribution of such short DNA motifs would vary across these 3D-defined regions, and therefore provide different amount of "substrates" for these mutagenic processes to act upon.

We first investigated whether there are individual positions harbouring recurrent variants observed in tumour samples which could be attributed to specific mutagenic processes. As introduced earlier, the attribution of individual mutations to specific mutagenic processes remains a challenge in mutational signature extraction tools. Indeed, in mutation data from The Cancer Genome Atlas (TCGA) database [22], we do not find recurrent hotspots occurring in DNA motifs attributable to the UV, 5-FU and Platinum signatures (Supplementary Figure S1); there were well-characterised mutational hotspots occuring in the aging signature context, however these mutations do not appear to be associated with age of cancer onset (Supplementary Figure S2), suggesting these could also be coincidental associations. However, we find recurrent mutations for

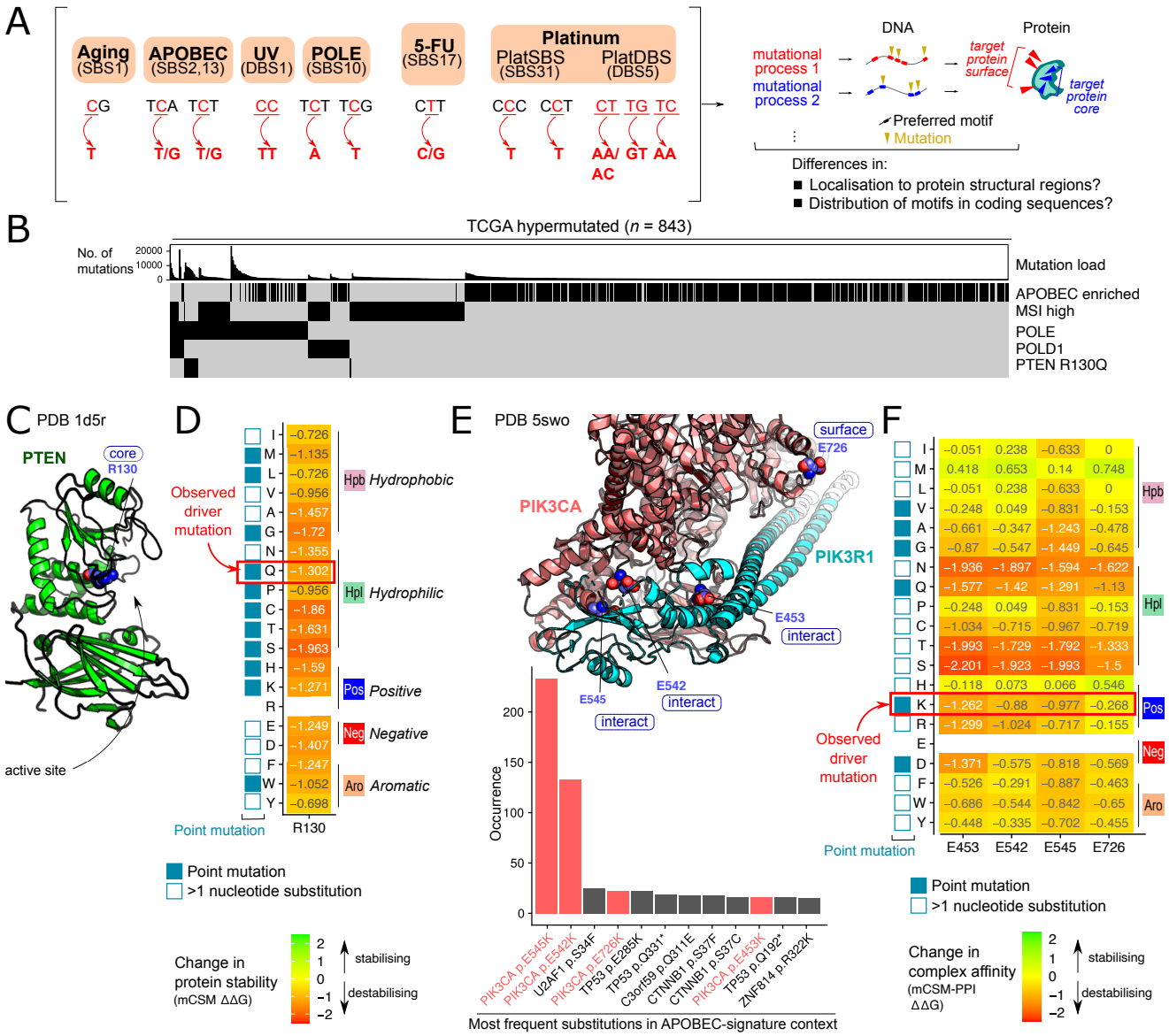


Figure 1: Mapping somatic mutagenic processes to protein structures. (A) Mutational signatures considered in this study. SBS, single-base substitution; DBS, doublet-base substitution. The nomenclature of mutational signatures according to the COSMIC catalogue (v3.1, https://cancer.sanger.ac.uk/cosmic/signatures/) are noted. (B) Occurrence of molecular alterations across $n = 843$ TCGA hypermutated cases (arranged in columns). APOBEC signature enrichment was calculated as described [19]. MSI (Microsatellite instability) status was taken directly from TCGA clinical annotations. Sequence alterations in POLE, POLD1 and the PTEN p.R130Q mutations are indicated. (C) p.R130Q mutation mapped onto the PTEN structure (Protein Data bank [PDB] 1d5r). (D) Change in protein stability predicted using mCSM [20] for the PTEN Arg130 position. Points mutations and amino acid (AA) changes arising from more than 1 nucleotide substitution are indicated next to every assayed AA change. (E) (bottom) Most frequent substitutions in the APOBEC signature. PIK3CA mutations are highlighted in red. (top) Structural view of PIK3CA (red) in complex with PIK3R1 (cyan) (PDB 5swo). Positions with recurrent mutations are highlighted. (F) Change in complex affinity predicted using mCSM-PPI [21].

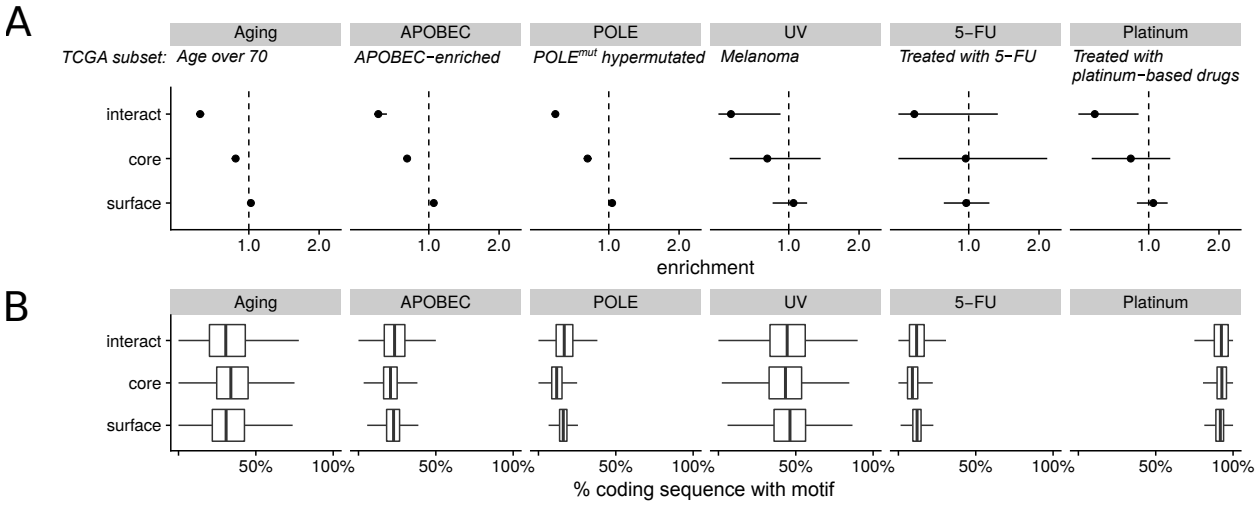


Figure 2: Distribution of mutational signature motifs across protein structural regions. (A) Variant enrichment in protein structural regions for different mutational signatures in different defined subsets of the TCGA data (indicated in italics). Error bars depict 95% confidence interval estimated by bootstrapping. (B) Distributions of the abundance of motifs (in terms of the percentage of human coding sequences with such motifs) in native, wild-type coding sequences associated with each signature in protein surface, core and interacting interface.

the POLE and APOBEC signatures which are conceivably associated with these aetiologies. Firstly, we observe the mutation PTEN p.R130Q which is present almost exclusively in POLE-mutated hypermutated tumours (Figure 1B). Arg130 is a buried position close to the active site of the PTEN phosphatase domain (Figure 1C) [23]; the R→Q mutation is predicted to be destabilising to the protein (Figure 1D), suggesting its possible loss-of-function effect, although its effect is not extreme relative to other amino acid substitutions arising from point mutations at this position. For the APOBEC signature, we notice that four mutations in PIK3CA are amongst the most frequent substitutions in the APOBEC signature context (Figure 1E); two of these (p.E545K and p.E542K) are well-documented mutational hotspots which are linked to elevated APOBEC mutagenesis [24]. These variants reside at the interface between PIK3CA and its regulatory partner PIK3R1 (Figure 1E) [25], and are predicted to be mildly destabilising to this protein complex (Figure 1F) [21]. In summary, these analyses map the protein structural impact of different mutagenic processes, and link these mutations to the biophysical basis of their possible alteration towards function. Interestingly, the observed recurrent mutations are often not the substitution which could theoretically confer the most dramatic impact; the most damaging amino acid changes often require the rarer mutational event of tandem nucleotide substitutions (Figure 1D,F). This suggests a degree of protection from the rule-book of codon usage and sequence preferences of mutagenic processes (see Discussion).

We then compare the general trends of TCGA mutations that occur in contexts attributable to each of the signature in terms of their localisation to protein surface, core and interface, defined by examining solvent accessibility using three-dimensional structures of individual protomers and protein complexes (see Methods). This analysis shows for all six mutational signatures, an enrichment for mutations on protein surface and depletion in the protein core and interacting interface can be observed (Figure 2A). However, these regions do not show large differences in terms of the amount of mutable DNA motifs in the native coding sequences

(Figure 2B). These results suggest that TCGA mutations show a bias towards protein surface for all of the six examined mutational signatures, probably a reflection of the difference in tolerability of mutations at different protein structural regions (see below).

Protein structural localisation is indicative of driver status, clonality and selection pressure

The bias of all examined mutational signatures towards generating mutations on protein surface could be an indication of the different impact to the affected protein upon mutation in different regions of protein structure. Protein cores are typically hydrophobic [26] and therefore amino acid substitutions that significantly alter hydrophobicity would likely render the protein unstable [27] (Figure 3A); such restrictions are less severe in protein surfaces, although a subset of such positions would map to protein interaction sites, and mutations that reside in these positions could block formation of complexes with specific protein partners, thereby rewiring molecular interaction networks [28] (Figure 3A).

Here we characterise the features associated with variants observed in tumour samples falling into protein surfaces, cores and interacting interfaces, both at the molecular (the types of physicochemical changes impart on the affected protein) and cellular (frequencies within the tumour and selection pressure) levels. First, we observe that variants in protein core or interacting interfaces are more likely to reside in cancer drivers if they are more clonal (Figure 3B), i.e. occur earlier in tumour evolution and therefore are more frequent across the population of tumour cells. This indicates that not only protein structural localisation of variants indicates driver status (as previously reported, e.g. [29]), but also a reflection of the clonality of mutations. We further observe, using Gene Set Enrichment Analysis (GSEA) [30], that clonal mutations at interacting interfaces perturb important signaling pathways and cell adhesion, while those at the core perturb proteins such as transporters and molecules responsible for binding ions and nucleotides (Supplementary Figure S3). Considering amino acid changes observed in tumours, substitution profiles in protein surface, core and interface reflect differences in solvent accessibility, with substitutions amongst hydrophobic amino acids much more common in the protein core (Figure 3C). On the other hand, changes from positive to polar amino acids are prominent in protein surfaces and interacting interfaces. The similarity between data from protein surfaces and interacting interfaces implies that the likelihood of transition between amino acid types are indications of solvent accessibility. The more restrictive nature of substitutions in the core is applicable both to a set of essential proteins whose removal would render cancer cells nonviable as assayed in publicly available gene knockdown screens [31, 32] (Supplementary Figures S4 and S5), and generally to whole TCGA and COSMIC cohorts (Supplementary Figure S6).

We investigate further the hypothesis that the core is subjected to greater purifying selection pressure in somatic evolution by explicitly calculating the dN/dS metric [11] which quantifies selection pressure, using datasets representing three different states in somatic evolution: (a) *in vitro*-cultured induced pluripotent stem cells (from the HipSci project [33]; see Supplementary Figure S7 and Methods for details) (b) TCGA

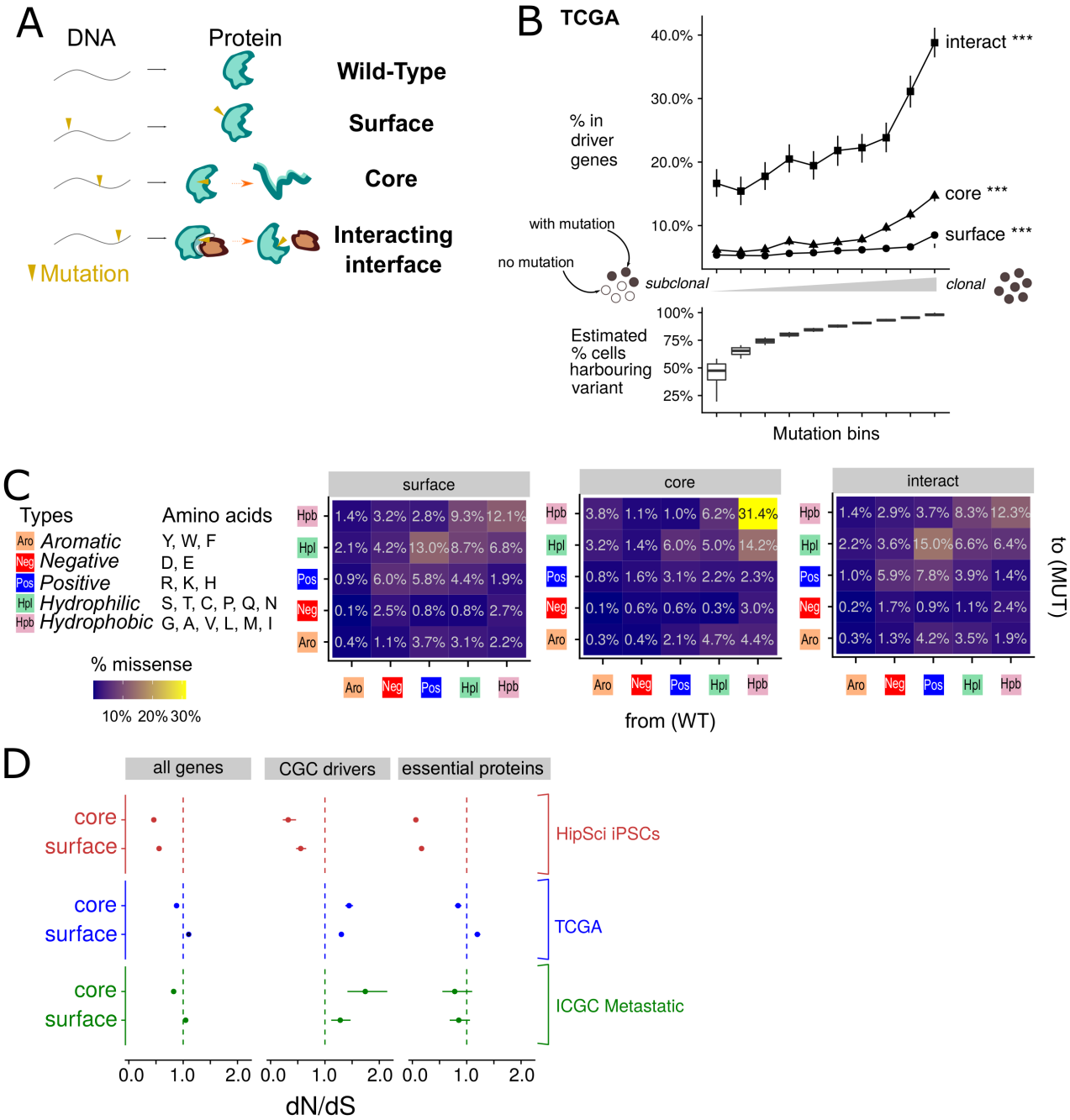


Figure 3: Protein structural localisation is indicative of driver status, clonality and selection pressure. (A) Schematic of possible consequences of missense mutations localised to different regions of protein. (B) The percentage of mutations that localise to a COSMIC Cancer Gene Census (CGC) driver gene across the spectrum of clonality (measured by the estimated proportion of cells harbouring such mutations, see Methods), considered separately protein surface, core and interface. Error bars depict 95% confidence interval estimated by bootstrapping. Trends are evaluated using the Jonckheere-Terpstra Test; ***, $p = 0.001$. (C) Distribution of variants in essential proteins (see Methods) which change from one amino acid (AA) type to another in surface, core and interacting interface. The wild-type (WT) residue type is depicted on the horizontal axis and the mutant (MUT) residue on the vertical axis. The proportion of variants for each transition type is noted. (D) dN/dS comparison of protein surface and core, calculated using the dNdScv package [11]. Error bars show 95% confidence intervals. As controls the same calculation was performed on two gene sets: Cancer Gene Census (CGC) drivers and essential proteins. Supplementary Table S3 contains the values visualised here.

(mainly primary tumours) and (c) a dataset of metastatic cancers from the International Cancer Genome Consortium (ICGC) [34]. Comparing protein surface and core, we observe that protein core has a much more negative selection pressure compared to protein surfaces (Figure 3D); this trend is consistent across all three examined datasets. As controls, we perform the same dN/dS calculations specifically on essential proteins; expectedly, for HipSci iPSCs which are non-cancerous, we observe negative selection at both protein surfaces and cores (Figure 3D). Signal for purifying selection is weaker in the TCGA and ICGC metastatic samples (dN/dS closer to 1, indicating neutral selection pressure), suggesting that such proteins are perturbed with variants as cells become cancerous. Importantly, dN/dS at the protein surface is higher than the core in all cases for such essential proteins. As a positive control, cancer drivers (COSMIC Cancer Gene Census [35]) show a positive selection pressure in cancerous states. Interestingly, this is the only scenario where dN/dS at the core is higher than that at protein surface in the cancerous states, consistent with positively selected mutations in tumour suppressor genes which preferentially reside in the protein core [18]. These results show a robust difference in selection pressure and mutability of protein surfaces and cores, which resembles previous data observed in species evolution [36]; in the case of somatic variants the same difference also applies, linking protein structural localisation to selection. The only exception is in cases where core mutations contribute proliferative advantage which overrides purifying selection arising from damaging protein activity and stability. The clonality analysis supports the importance of 3D structural properties in modulating the intratumoural frequency of somatic mutations.

Deep mutational scanning data reveals DNA mutational signatures associated with damaging variants

We next reasoned the bias towards protein surface for all six examined mutational signatures could be an indication that the observed mutational profile is under-sampling mutations in protein cores and interface which confer more significant damage to the implicated protein and the cell. This difference in mutational "tolerability" affects their abundance in the tumour population; mutations which are less tolerable would be rarer and therefore more difficult to be sampled in conventional exome/genome sequencing experiments (Figure 4A). An unbiased comparison of the impact of variants that occur in different mutational contexts would require comprehensive measurements which consider every amino acid positions and survey substitutions to any other 19 amino acids. Such deep mutational scanning (DMS) data [12, 13, 37] would be crucial as a control to assess selection against intolerable variants present in datasets obtained from observed tumour samples. We first obtained experimental DMS data from a public database (MaveDB [37]) on two cancer-related proteins, PTEN [13, 14] and MSH2 [15]; these are two human proteins harbouring known, missense, cancer driver variants with DMS data available in the database. Overlaying ClinVar and TCGA variants onto these DMS profiles, we see that TCGA variants have less severe mutational impact than known pathogenic variants (Figure 4B), suggesting signals from analysing TCGA data could be masked by dominance of

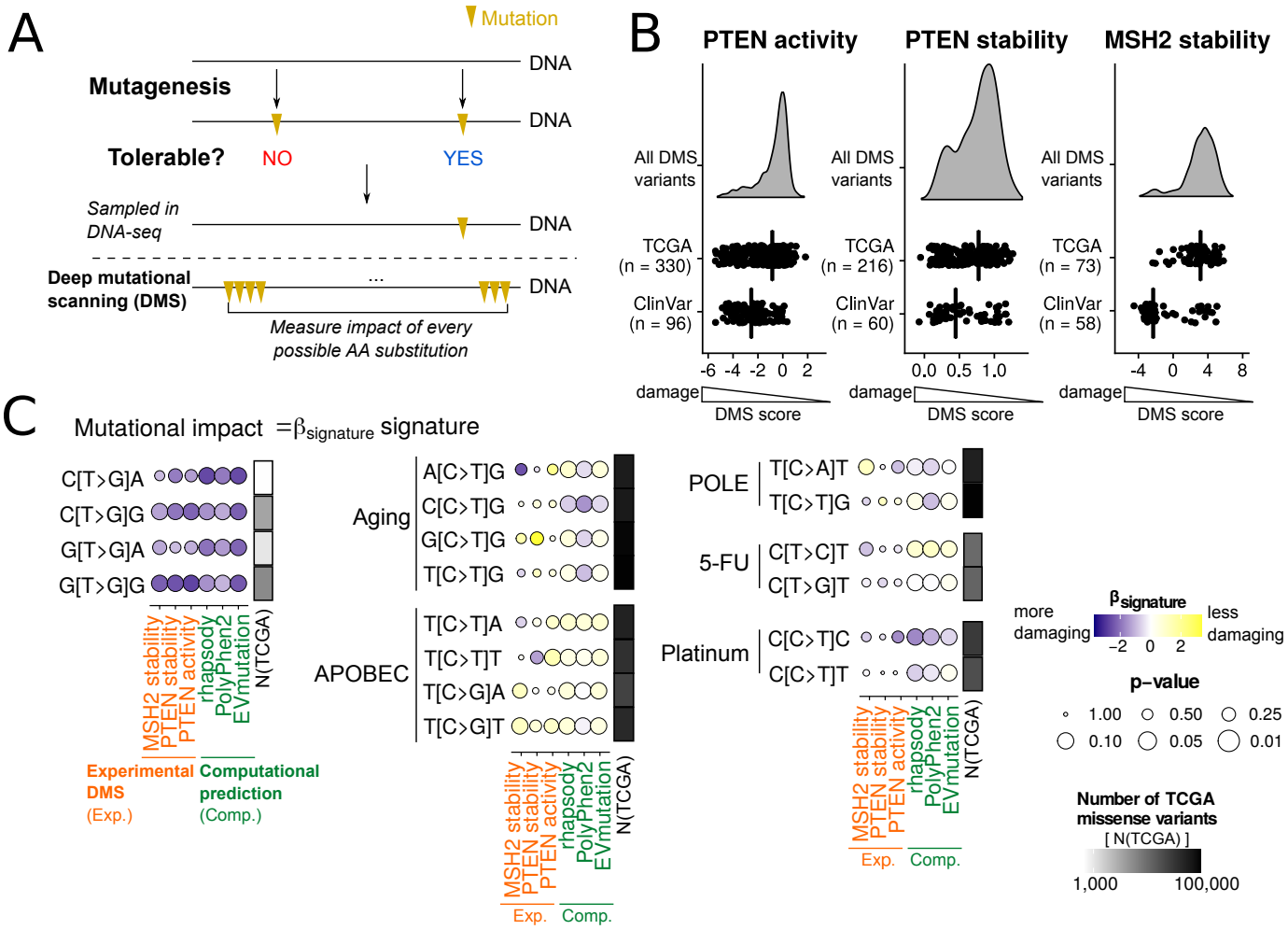


Figure 4: Deep mutagenesis scanning (DMS) reveals DNA sequence contexts associated with damaging mutational impact on proteins. (A) Schematic of the impact of mutational tolerability on the ability to sample mutations in conventional DNA sequencing experiments. DMS addresses this problem by exhaustively measuring impact of every possible amino acid substitutions on the protein. (B) Experimental DMS profiles of two cancer-related proteins, PTEN and MSH2, obtained from MaveDB. The impact scores of ClinVar and TCGA variants are depicted as dotplots. Thickened lines indicate medians. (C) Mutational impact assessment was obtained through either experimental (of specific proteins; coloured orange, horizontal axis) or computational (over the entire human proteome; coloured green) DMS, and regression models were fitted against the 96 mutational contexts. The magnitude and statistical significance of the coefficients of each context ($\beta_{\text{signature}}$) were visualised. The number of TCGA variants occuring in each context is depicted in greyscale (rectangles, right column). Selected contexts which show consistent association with damaging impact (left panel) or are attributed to well-defined mutational signatures (middle and right panels) are depicted. See Supplementary Figure S8 for the complete plot, and Supplementary Table S4 for data underlying this analysis.

passenger variants with little variant impact. We further obtained analogous computational DMS data by applying variant impact predictors (PolyPhen-2 [38], EVmutation [39] and rhapsody [16]) over full-length coding sequences of a total of 6,918 human proteins (see Methods). Comparing to the experimental DMS profiles above, this computational dataset greatly expands the extent of the proteome surveyed; moreover, the predictors we used consider different (sequence conservation, static protein secondary/tertiary structures, and dynamics models defined on structures) features [16, 38, 39, 40], taking into account different aspects of the molecular profile in variant impact annotation. We developed a R package (CDSMutSig) with utilities to map any amino acid substitution to the corresponding coding sequence change and DNA mutational contexts (see Methods). We took into consideration the fact that a missense amino acid change could be caused by nucleotide substitution at any position of the codon, depending on the exact missense change in question. Therefore, the trinucleotide mutational contexts could be defined "in-frame" within the same codon, or span tandem codons. Utilities in the CDSMutSig package considered any possible nucleotide (and therefore amino acid) substitution, and allow for filtering to remove amino acid changes which requires tandem nucleotide changes. We evaluate the association of each possible three-base DNA mutational contexts with mutational impact by fitting linear models of the obtained DMS profiles using the trinucleotide mutational signatures as the explanatory variable (Figure 4C). Comparing the regression coefficients, we find that four sequence contexts, namely `G[T>G]G`, `G[T>G]A`, `C[T>G]G` and `C[T>G]A` (these motifs are written in the direction from 5' to 3') are associated with more damaging variants in all of the experimental and computational DMS dataset considered (Figure 4C). In contrast, sequence contexts attributable to the commonly observed mutational signatures examined above either do not exhibit consistent association with variant impact, or confer less damaging impact in certain cases. Importantly, we observe much fewer TCGA mutations occurring in all four damaging sequence contexts in comparison with contexts associated with common mutational signatures (Figure 4C). Taken together, we demonstrate that the use of DMS data can offer a more unbiased view of protein impact of variants occurring in different DNA sequence contexts, in comparison to analysis of large molecular profiles obtained from tumour samples. The data support the dominance of tolerable mutations in observed mutational profiles, and suggest that common mutagenic processes avoid DNA motifs which could give rise to damaging amino acid substitutions; this has important implications in designing therapeutic strategies to exploit vulnerability of tumours towards damaging mutations, and in building mutational impact predictors which decouple variants pathogenicity from their impact on protein stability and/or activity (see Discussion).

Physicochemical impact of damaging mutational contexts

Our analyses on both (observed TCGA mutations and DMS) datasets illustrate constraints exerted on the damaging impact of mutations: recurrent hotspots attributable to specific mutagenic processes are not necessarily the substitution which could bring the most extreme biophysical impact (Figure 1); compared

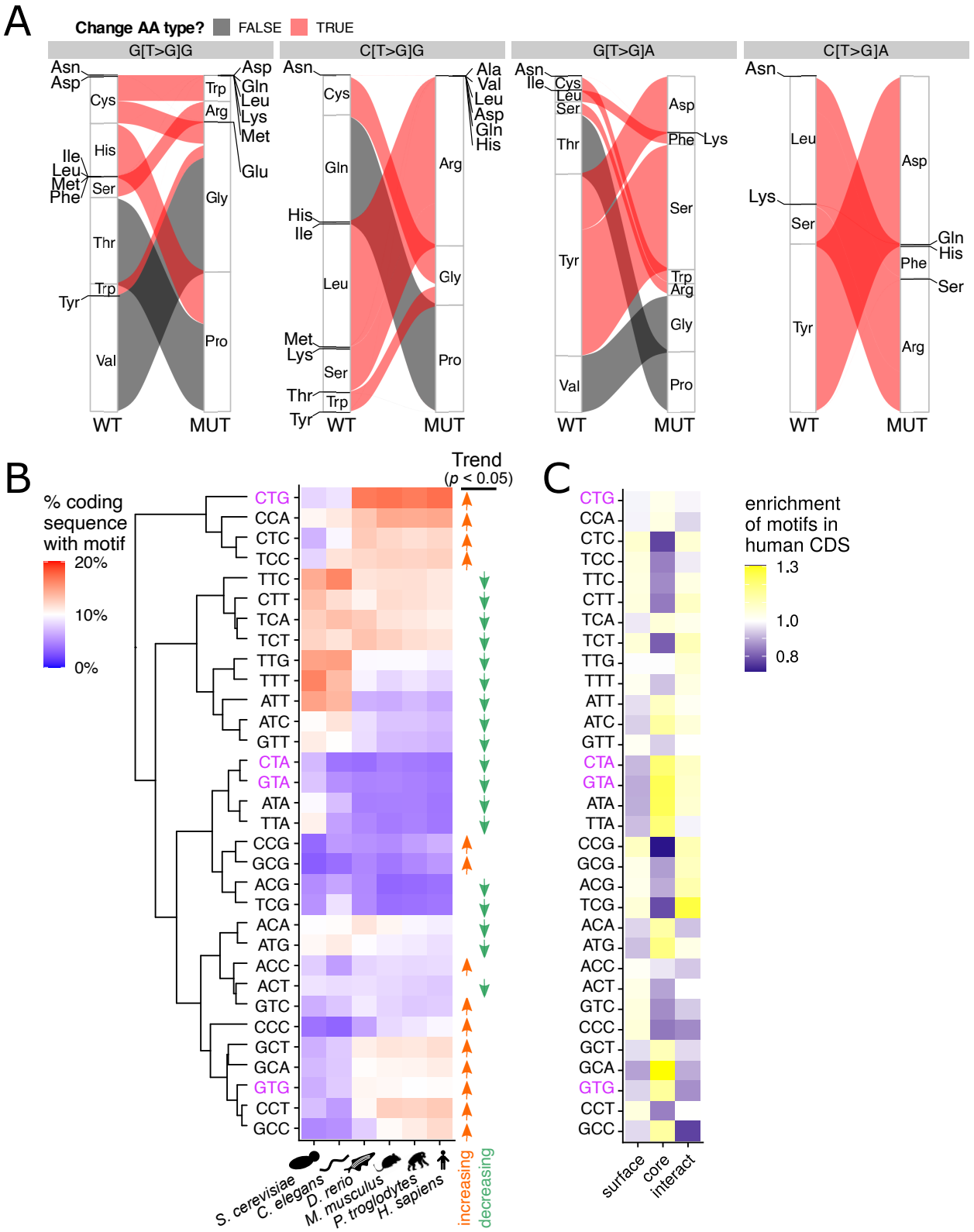


Figure 5: Physicochemical, protein structural and evolutionary characterisation of damaging mutational contexts. (A) Amino acid (AA) substitutions attributed to each damaging mutational contexts (Figure 4C). The width of ribbons correspond to the amount of AA substitutions assayed in computational DMS. Ribbons are coloured to indicate whether the AA change constitute changes in physicochemical features. (B) Proportion of the coding sequence harbouring each mutational context compared across six different species. Trends were evaluated using a Jonckheere-Terpstra Test. Mutational contexts depicted in panel (A) are highlighted in magenta. (C) The enrichment of mutational motifs in protein surface, core or interface calculated over all human CDS with structural coverage. Order of motifs on the vertical axis follows that in panel (B).

with commonly observed mutational signatures, far fewer observed variants occur within the most damaging mutational contexts (Figure 4C). We ask whether there is a biophysical basis behind the damaging impact scores as revealed in experimental and computational DMS. Figure 5A shows the amino acid substitutions attributable to each of the damaging mutational contexts, had such mutations occurred and been sampled in sequencing experiments. We observe that these mutational contexts bring about three main alterations to the physicochemical properties of the mutated amino acid positions (Figure 5A): (i) change in charge (mutations to arginine/aspartate); (ii) change in polarity (e.g. tyrosine → serine); (iii) change in structural flexibility (mutations to glycine/proline). We next ask whether the abundance of these sequence motifs within coding sequences change across species evolution. By enumerating these sequence contexts in homologous proteins across different species, we find distinct trends in the abundance of motifs across evolution (Figure 5B). On the damaging mutational contexts, two of them (`CTG` and `GTG`) increase in abundance, especially in vertebrates, while `CTA` and `GTA` are rarer in coding sequences. We further find that all such contexts tend to localise to protein cores relative to surface and interface (Figure 5C). Together, these results provide biophysical explanations, at the protein level, behind damaging mutational impact associated with specific DNA sequence contexts, consistent with negative selection in the protein core observed in the TCGA data. The abundance of mutable sequence contexts give theoretical indications of the possible extent of damage a mutational process could exert on the coding genome. While some of the damaging mutational contexts are rare within coding sequences, other DNA motifs which could potentially bring about damaging variants are found more abundant in human and other vertebrate coding sequences, probably a cost of maintaining hydrophobicity and/or polarity to preserve the stability of the protein core (see Discussion).

Discussion

In this study we map protein-level annotations to somatic variants attributable to differnet mutagenic processes, to explore specific impact in protein structural and physicochemical features that these processes are capable to impart. Exploration of residue solvent accessibility, as well as other protein structural features, have been reported in studies of protein evolution to explain differences in evolutionary rates across different protein regions [41]. To our knowledge this is the first work addressing the consequences of specific somatic DNA mutational signatures at the protein level. The problem of attributing mutations to likely contributory mutagenic processes is to date intractable as there are no technical solutions which directly address this problem [1, 2]; here we side-step this problem by considering signatures with clear patterns of DNA sequence motif preferences. The exact genomic locations mutated upon exposure of a given mutagenic process are known to vary across different biological samples and experimental conditions [42], therefore it is impossible to attribute, with certainty, mutagenic processes with individual genetic variants. Nevertheless, since the signatures we examined are each over-represented with variants occurring within specific DNA motifs, these data illustrate the likely impact these signatures could cause to individual proteins. We were able to observe,

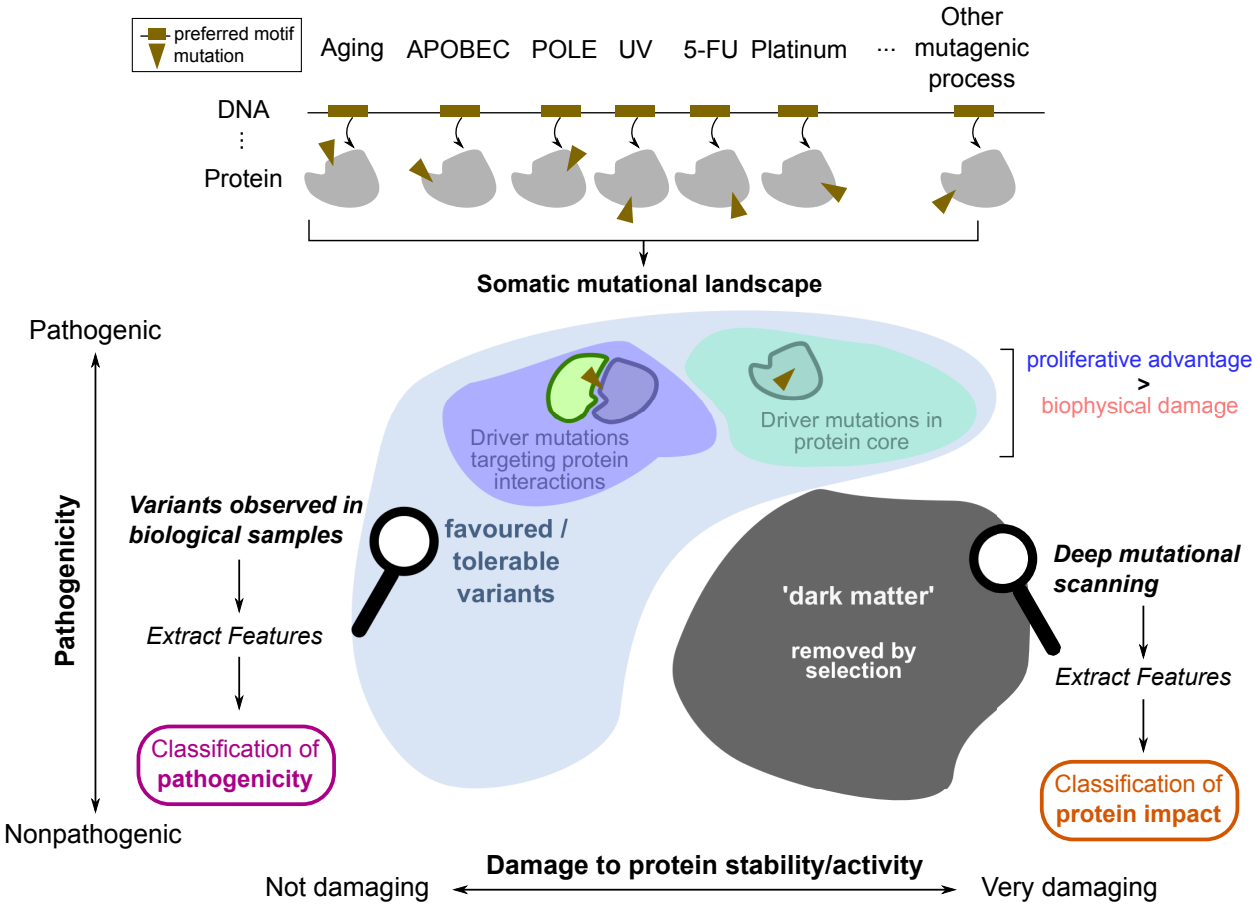


Figure 6: Overview of this analysis. Mutagenic processes generate mutations that can be mapped to different locations in protein structures. These mutations can be contextualised in terms of their pathogenicity and their impact to the stability/activity of the affected protein. A regime of damaging variant impact constitutes what we term the "dark matter" of somatic mutational landscape which is under-sampled in sequencing experiments of biological samples, due to the fact that they are likely to be removed in selection. DMS data allow for probing of this space to extract features that explain damage to protein stability and/or activity. Outside of this "dark matter" regime variants are either favoured or allowed, and map to different protein structural regions. Those variants which bring about greater proliferative advantage which offset any biophysical damage will be retained and/or positively selected, as in the case of driver variants targeting protein core and interactions. Such variants can readily be sampled by sequencing, allowing for the discovery of features which explain variant pathogenicity. See Discussion for further details.

from sequenced tumour samples, recurrent destabilising variants (such as the PTEN and PIK3CA variants respectively associated with POLE and APOBEC signatures) likely caused by specific mutagenic processes (Figure 1): PTEN mutations have been said to occur concurrent with or subsequent to acquiring POLE mutations [43]; the said PIK3CA variants have been shown experimentally attributable to APOBEC3 activation [24].

The most important finding is that by studying mutational profiles obtained from tumour samples, all six common mutational signatures examined here show enrichment in the surface and depletion in the protein core and interacting interface (Figure 2A), while these regions show similar amounts of DNA motifs attributable to these signatures (Figure 2B). This is consistent with the idea that amino acid preference is subject to greater constraints in the protein core in comparison with the protein surface [27]: we show specifically that the core mainly accepts substitutions between hydrophobic amino acids (Figure 3C) and accordingly has lower dN/dS (Figure 3D). These are perhaps unsurprising findings given the physicochemical nature local to the protein core [26]. The significance of this finding, however, is that observed mutational profiles are naturally biased towards tolerable mutations (and, in the case of somatic mutations, favourable mutations which bring about proliferative advantage to the cell), while truly damaging mutations (which are more likely removed through tumour evolution) are under-represented. This represents the "dark matter" of the somatic mutational profile, where variants which are highly damaging to the affected protein are likely under-sampled, rendering comprehensive characterisations of their true impact on protein activity/stability difficult using solely sequencing experiments performed on biological samples (Figure 6). The only exception where damaging somatic variants could avoid the under-sampling problem is when these variants bring greater proliferative advantage to the cell, that overrides any costs imparted on the stability of the affected protein (Figure 6), e.g. loss-of-function mutations that target the protein cores of tumour suppressor genes [18], which are likely to be positively selected during tumour evolution. The distribution of driver mutations in protein surface, core and interface has been widely discussed [29]; our data additionally suggest that driver mutations located in protein core and interface are more clonal (Figure 3B), supporting the link between mutational frequency and such cost-benefit balance. Experimentally, to override the under-sampling problem of damaging variants, one potential solution is to use dedicated methods such as ultra-deep sequencing [44] or emerging single-nucleus sequencing approaches [45] to generate mutational profiles from biological samples. Another potential solution is to generate and analyse DMS data which explicitly measure the impact of every possible amino acid substitution. Here we show that using such data we have been able to find specific DNA sequence contexts that could have generated damaging mutations had they been observed in biological samples (Figure 4-5). The finding that very few mutations that occur in such contexts can be observed in TCGA (Figure 4C) suggests that these mutations bring excessive damage to protein stability/activity, without sufficient compensatory proliferative advantage to tumour evolution. This presents a therapeutic opportunity: mutators which target these DNA motifs associated with a damaging protein-level impact can

orchestrate an increase in tumour mutational burden to accumulate a large amount of damaging genetic variants [46]. If this could be directed specifically towards tumour cells, this represents a "killing strategy" for tumours which otherwise do not harbour any molecular alterations targetable using existing therapies.

The bias towards tolerable variants in observational studies has important implications in understanding mutational impact and developing variant impact annotation tools. A large number of sequence- and structure-based variant impact annotation tools have been shown to be valuable in offering insights of variant pathogenicity [18, 47]. These tools typically utilise information from established variant datasets observed from biological samples of different extents of pathogenicity, to build classifiers using features which distinguishes pathogenic from nonpathogenic variants. While existing tools may demonstrate a good level of sensitivity and specificity in classifying pathogenicity, this does not necessarily mean these predictions offer accurate indications of protein-level impact of variants, partly a consequence of limited availability of variants which truly damage the stability and/or activity of the affected protein for predictors to be trained on. DMS data can provide a comprehensive training dataset to develop predictors sensitive enough to distinguish between intolerable versus tolerable variants. This offers a route to provide annotation of the impact of variants on protein stability and/or activity, which may not be directly associated with pathogenicity. Existing variant pathogenicity predictions are effective in assisting routine genetic diagnoses; however, given the advancement of sensitive detection of ultra-rare variants in the genome [45], DMS data can potentially offer an additional filter to identify truly damaging variants ignored by conventional pathogenicity predictors. DMS by its nature provides position-specific evaluation of the vulnerability of the protein towards mutations at any given amino acid residue. Exploiting these dataset to identify epistatic effects between multiple residues could aid in predicting the three-dimensional folds of protein without resolved structures [48] and designing compensatory mutations which negate effects of a pathogenic variant [49], both of which are of immense interests for molecular biologists interested in the effects of mutations on protein stability and activity.

The landscape of somatic variants is shaped by the mutagenic (available armoury of mutational processes), biophysical (physicochemical nature of proteins) and evolutionary (positive and negative selection) constraints on the genome. We are partly protected by the molecular rule-book, as we see in the cases of PTEN and PIK3CA variants, that the observed recurrent variants do not represent the most damaging possibility (Figure 1). Different DNA sequence contexts show either increasing or decreasing trends in terms of their abundance in coding sequences from invertebrate and vertebrate species (Figure 5B), suggesting changes in terms of preferences in codon usage. A well-established body of research in evolutionary biology has established that residue solvent accessibility and protein stability are major factors to explain differences in evolutionary rates within and between proteins [41]. We show that the observation where protein surfaces are more mutable than the buried core in species evolution [36] also applies to somatic variants. We have found damaging somatic mutational contexts which primarily target protein cores, and alter the polarity, charge and flexibility of these positions (Figure 5A,C). The conservation of a stable protein core necessitates the maintenance of polarity,

hydrophobicity and packing arrangements of side chains in this region [50]. Retaining and/or accumulating amino acids which contribute to core stability could come with a potential cost to cellular fitness, if such evolutionary constraint implies accumulation of DNA motifs which could act as "substrates" for mutagenesis. The biological relevance of this potential damage depends on whether mutagenic processes which target these motifs are active; although we observe fewer TCGA variants that occur in such contexts (Figure 4C), these mutational contexts fall into the "dark matter" of somatic substitutions: their depletion could either be due to purifying evolution, or that there are no endogenous and exogenous mutagenic processes which exist to target these DNA motifs in human cancers. A systematic investigation of the relationship between different mutagenic exposures and evolutionary conservation of functional sites warrant careful sequence and structural homology mapping. This could be a further important step in understanding the potential risks different mutagenic exposures could pose towards protein function and therefore their contribution in disease processes, across species of varying lifespan and living conditions.

Methods

Variant datasets

Cancer. In this analysis two sources of cancer variant data were considered. Firstly, we analysed a total of 27 cancer types from The Cancer Genome Atlas (TCGA; [22]) database, taking as input the annotated TCGA Mutation Annotation Files (MAFs) available on Broad GDAC Firehose (https://gdac.broadinstitute.org/). Secondly, somatic mutation data of all metastatic tumours from the International Cancer Genome Consortium (ICGC; [34]) database were downloaded from the ICGC data portal on 03/06/2019, by selecting only the tumour samples with the annotation of relapse type "`Distant recurrence/metastasis`". A total of 785 ICGC metastatic tumour samples were analysed.

Induced pluripotent stem cells (iPSCs). Whole-exome sequencing (WES) variant data were acquired from the HipSci project [33], a collection of induced pluripotent stem cell (iPSC) and skin fibroblast cell lines from donors, as a dataset representing cancer-free but proliferative condition. Only donors without disease annotations were considered. The Variant Call Format (VCF) files generated from mpileup variant calls for all such iPSC and skin fibroblast lines were downloaded, where access is publicly available from the European Nucleotide Archive (ENA, Project accession PRJEB7243). For cell lines where multiple vcfs were available, data generated from Illumina HiSeq 2000 was taken. To filter for somatic variants acquired during the course of cell line derivation and passage (see below), we only considered iPSC lines from donors for which WES data on fibroblasts (progenitors of the iPSC lines) were also available. Only variants marked "PASS" in the vcf FILTER field was included for downstream annotation and analysis. A total of 300 iPSC lines from 192 donors were analysed. Selection was quantified for this dataset to compare with the two (ICGC metastatic and TCGA) tumour datasets.

Variant processing

Variant filtering and data mapping. For consistency, only missense variants lying in exon regions were considered in the ICGC dataset, since TCGA and iPSC variants were called from WES data. These were extracted by intersecting (using bedtools [v2.25.0]) each vcf file with entries marked "exons" in the Ensembl (release 93, genome build GRCh37) Gene Annotation File. For the iPSC data, variant data from each iPSC line was compared with that profiled on its fibroblast progenitor, and filter for only variants unique to the iPSC lines but not found in the fibroblast (Supplementary Figure S7). For both cell types, variants localising to HLA genes were filtered away owing to their intrinsic highly polymorphic nature. TCGA variants are also cleaned to group together variants mapped to consecutive genomic positions in order to distinguish variants occurring in doublet-base substitution (DBS) signatures.

We calculated the mutational density per exome by normalising the total count of mutations for a given exome by its size (38 Mb was taken as an estimate [51]). Tumours with density > 10 mutations per Mb were taken as hypermutated [52]. Finally, a density-based enrichment of mutational signatures was calculated following the method in [19] for calculating APOBEC-signature enrichment.

Variant annotation. To unify with the TCGA annotated MAFs, variant data from both the ICGC metastatic cohort and the HipSci iPSCs were annotated with gene and protein information using Oncotator (v1.9.9.0; [53]). Oncotator prioritises a set of transcripts with reading frames identical to UniProt sequences; these are directly taken forward for mapping these mutations to structure using ZoomVar, which were mapped, using UniProt identifiers as keys, to protein structural regions. See below (section "Mapping of coding sequences and protein structural data") for a detailed description of parameters used to define structural regions. Amino acids were classified by their physicochemical properties into any one of the following 5 groups: (i) aromatic (Y, W,F); (ii) negative (D, E); (iii) positive (R, K, H); (iv) hydrophilic (S, T, C, P, Q, N), and; (v) hydrophobic (G, A, V, L, M, I). The sequence contexts were also annotated for substitution variants, using which variants are grouped as attributable to different signatures (Figure 4C and Supplementary Table S1). Additionally, for TCGA variants the frequency of variants (in terms of the percentage of tumour cells harbouring variants, also known as "cancer cell fraction") was estimated following the method described by McGranahan and colleagues [54].

Annotation of protein stability change. Using the ZoomVar database we annotated, for each TCGA variant with 3D structural coverage, the PDB entry with the highest sequence identity and lowest E-value match with the affected protein. We considered only those structures which harbours the same amino acid at the position in question as the wild-type. These variants were subject to variant impact prediction using the following tools: (i) the mCSM server [20] to predict the change in protein stability upon mutation; (ii) the rhapsody package [16] which predicts the probability that a given mutation is pathogenic, using results from

PolyPhen-2 [38] and EVmutation [39], as well as coarse-grained network dynamics models defined on the input 3D protein structure [40].

Gene sets

Cancer drivers. The Cancer Gene Census ("CGC", v86) was downloaded from the COSMIC database [35]. A gene set of cancer drivers was derived from the CGC by selecting only the "Tier 1" (documented relevance to cancer) genes, totalling to 574 genes.

Essential proteins. We utilised two data sources to obtain a list of essential proteins. First, a list of 109 essential proteins (see main text) was defined based on gene knockdown assays performed on cancer cell lines across cancer types. Both the CRISPR screen (AVANA project; [31]), and the RNAi screen ("Combined RNAi dataset" from the depmap project, based on short hairpin RNA [shRNA] libraries; see [55]), were downloaded from depmap (`https://depmap.org/`) and analysed. Both assays screened the impact of knocking down individual genes (CRISPR: 17,634 genes; RNAi: 17,309 genes) on the viability of cancer cell lines (CRISPR: 558 cell lines; RNAi: 712 cell lines) covering a total of 21 cancer types. This impact is represented by a "dependency score" akin to a z-score, where the more negative this score, the more dependent the cell line is to this particular gene (since its knockdown resulted in a significant decrease in viability, see Figure S4). We selected for genes demonstrating essentiality in all 21 cancer types, first calculating the mean dependency score for each gene in each cancer type, and then filtering for genes whose dependency score is in the top (most negative) 5% across all genes in all of the cancer types examined. Second, we collected another list of essential proteins defined using a different CRISPR knockdown dataset from Hart and colleagues [56]. This gene set was taken directly from `https://github.com/macarthur-lab/gene_lists`. We overlapped it with our own curated set in Figure 3.

General pathways and molecular functions. KEGG pathways and GO molecular function gene sets were acquired from MSigDB (v7.1) [30].

Mapping of coding sequences and protein structural data

Protein structural localisation. For each protein we obtained residue-level mapping of its core, surface and protein-protein interaction interface from the ZoomVar database [18]. The generation of ZoomVar was described elsewhere. Protein positions were classified into protein core, surface and interacting interfaces following [18]. Briefly, core residues are defined to have a relative solvent accessible surface area (SASA) of lower than 0.15 [57, 58]; otherwise they are labelled as surface. Interfaces were annotated using POPSCOMP [59] by considering the change in solvent accessible surface area upon complex formation; residues were annotated to be in the interacting interface if such change $> 15\%$ [57].

Enumerating coding sequences containing DNA mutational contexts. We obtained from UniProt its curated set of coding sequences (CDS) for proteomes examined in this work (Supplementary Table S2), and mapped, for each protein, the corresponding CDS sequence on a per-residue level. We then enumerated the percentage of residues within each CDS which could be altered by mutation within every possible trinucleotide motif. Since every position of the codon could be substituted to cause an amino acid change, trinucleotide mutational contexts could involve the codon immediately before (if codon position 1 was substituted) or after (if codon position 3 was substituted) the given amino acid position. We therefore considered, for each amino acid position, a 5-mer consisting of the triplet code for the position, plus one base 5' and one base 3' to the triplet code. Using these 5-mers we classified and counted whether each residue contains the named sequence motif.

Quantification of variant/motif enrichment

For variant enrichment, we quantified such enrichment in a given protein structural region by calculating the following density metric ω:

$$\omega_{\text{region}} = \frac{X_{\text{region}}/\text{Size}_{\text{region}}}{X_{\text{all regions}}/\text{Size}_{\text{all regions}}} \tag{1}$$

where

X is the number of mutations, and

"Size" refers to the number of amino acids belonging to the 3D structural region.

Similarly, the enrichment of motifs (Figure 5C) were calculated using the same formula as above, except that here X_{region} refers to the number of amino acids in the given 3D region that harbour the motif in question, and $X_{\text{all regions}}$ refers to the analogous amino acid counts in all regions together.

dN/dS calculations

dN/dS calculations were performed using the dNdScv package [11] was used here to quantify the dN/dS value for the iPSCs, TCGA and ICGC Metastatic datasets. Calculations were considered for protein surface and core by generating subsets of variants locating in the two regions and providing these two lists separately as input to dNdScv. Additional calculations on specific gene sets (cancer drivers, essential proteins) were performed by further filtering the dNdScv input by relevant gene symbols.

Deep mutational scanning (DMS)

Experimental DMS data. Three experimental DMS datasets were downloaded from MaveDB [37]. (i) DMS on MSH2 (MaveDB accession 00000050-a-1) with protein stability as readout [15]; (ii) DMS on PTEN

with activity measured by a lipid phosphatase assay (MaveDB 00000054-a-1) [14]; (iii) DMS on PTEN with protein stability as readout (MaveDB 00000013-a-1) [13]. In the stability DMS, protein abundance is taken as a proxy of protein stability; we previously reported that abundance and stability are broadly correlated with one another [18]. PTEN and MSH2 are the two examples human proteins which (a) have available DMS data deposited on MaveDB (accessed 5 January 2021) and (b) contain known missense cancer driver variants.

Computational DMS data. We sought to obtain an analogous DMS dataset computationally, by applying mutational impact predictors over full-length coding sequences of a comprehensive set of human proteins, with annotation of the mutational contexts of each assayed AA substitution. We considered all proteins with mutations observed in the TCGA dataset and took the best PDB match (see above) as the corresponding 3D structure as input to the mutational impact predictors. Due to potential one-to-many mappings between protein and transcripts, we prioritised transcripts with matched annotation between NCBI and EBI (from RefSeq MANE; https://www.ncbi.nlm.nih.gov/refseq/MANE/) in order to obtain unambiguous mapping of mutational contexts for each substitution. This leaves 6,918 unique Ensembl protein entries to be subjected to *in silico* saturation mutagenesis using the rhapsody package [16]. Since the run also involves results from EVmutation [39] and PolyPhen-2 [38], these 3 predictors were considered alongside with rhapsody output.

Mapping DNA mutational contexts to amino acid substitutions. DMS data contain readout/predictions for every amino acid position within a protein, mutating to every other 19 possible amino acids. A set of R functions, made available in the CDSMutSig package which accompanies this manuscript (see section "Code Availability"), have been developed to map DNA mutational contexts to DMS measurements. We mapped Ensembl (version 86) proteins to their corresponding transcripts, obtain the coding sequence (CDS; genome build GRCh38), and extracting the DNA substitution as well as three-base sequence context around the substitution that brings about each amino acid mutation. This provides a filter with which amino acid substitutions that do not arise from a single-nucleotide variant (SNV) can be removed.

Testing association between mutational impact and sequence context. All mutational impact scores are harmonised such that more negative/smaller values indicate more damaging variant impact. To examine the association between mutational impact and sequence contexts, the following linear model is considered:

$$\text{Mutational impact} = \beta_{\text{signature}}\text{signature} \tag{2}$$

where

"Mutational impact" refers to the impact scores obtained from computational/experimental DMS profiles, and "signature" corresponds to each of the 96 possible three-base mutational contexts.

Such model is fit onto each of the 6 DMS datasets (3 [MSH2 stability, PTEN activity, PTEN stability] experimental DMS profiles and 3 [rhapsody, PolyPhen-2 and EVmutation] computational DMS predictions), using the R glm function. The magnitude and p-value of $\beta_{\text{signature}}$ from the linear model are taken for analysis.

Statistics and data visualisation

All analyses and visualisations of data have been performed in the R statistical computing environment (v3.6.2). Bootstrapping statistics were generated using the R package boot [60, 61]. The Jonckheere-Terpstra test (R package DescTools [62]) was used to test for significance differences across samples binned by a given explanatory variable. Plots were generated using plotting functionalities in base R and the ggplot2 package [63]. Heatmaps were produced using the gplots package [64].

Code availability

The CDSMutSig R package provides functionalities to map between protein and DNA coding sequences and obtain trinucleotide mutational contexts for any amino acid mutations. This package is available at https://github.com/josef0731/CDSMutSig. All analysis code can be found at https://bitbucket.org/josef0731/zoomvarsomaticmutsig.

Data availability

All data files generated for this analysis is available at this zenodo repository: https://doi.org/10.5281/zenodo.4726168.

References

[1] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell reports* **3**, 246–59 (2013).

[2] Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

[3] Zavadil, J. & Rozen, S. G. Experimental delineation of mutational signatures is an essential tool in cancer epidemiology and prevention. *Chemical research in toxicology* **32**, 2153–2155 (2019).

[4] Boot, A. *et al.* In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome research* **28**, 654–665 (2018).

[5] Pich, O. *et al.* The mutational footprints of cancer therapies. *Nature genetics* **51**, 1732–1740 (2019).

[6] Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature reviews genetics* **15**, 585–98 (2014).

[7] Ng, J. C. F., Quist, J., Grigoriadis, A., Malim, M. H. & Fraternali, F. Pan-cancer transcriptomic analysis dissects immune and proliferative functions of APOBEC3 cytidine deaminases. *Nucleic acids research* **47**, 1178–1194 (2019).

[8] Hoopes, J. I. *et al.* APOBEC3A and APOBEC3B preferentially deaminate the lagging strand template during DNA replication. *Cell reports* **14**, 1273–1282 (2016).

[9] Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364** (2019).

[10] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–8 (1976).

[11] Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).

[12] Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods* **11**, 801–7 (2014).

[13] Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature genetics* **50**, 874–882 (2018).

[14] Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *American journal of human genetics* **102**, 943–955 (2018).

[15] Jia, X. *et al.* Massively parallel functional testing of MSH2 missense variants conferring lynch syndrome risk. *American journal of human genetics* **108**, 163–175 (2021).

[16] Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N. & Bahar, I. Rhapsody: predicting the pathogenicity of human missense variants. *Bioinformatics* **36**, 3084–3092 (2020).

[17] Muiños, F., Martinez-Jimenez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. *bioRxiv* (2020). URL https://www.biorxiv.org/content/early/2020/06/09/2020.06.03.130211.

[18] Laddach, A., Ng, J. C. F. & Fraternali, F. Pathogenic missense protein variants affect different functional pathways and proteomic features than healthy population variants. *PLoS Biology* **19**, 1–33 (2021).

[19] Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nature genetics* **47**, 1067–72 (2015).

[20] Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335–42 (2014).

[21] Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic acids research* **47**, W338–W344 (2019).

[22] Hutter, C. & Zenklusen, J. C. The cancer genome atlas: Creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).

[23] Lee, J. O. *et al.* Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. *Cell* **99**, 323–34 (1999).

[24] Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell reports* **7**, 1833–41 (2014).

[25] Miller, M. S. *et al.* Identification of allosteric binding sites for PI3Kα oncogenic mutant specific inhibitor design. *Bioorganic and medicinal chemistry* **25**, 1481–1486 (2017).

[26] Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB Journal* **10**, 75–83 (1996).

[27] Pál, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature reviews genetics* **7**, 337–48 (2006).

[28] Laddach, A., Ng, J. C.-F., Chung, S. S. & Fraternali, F. Genetic variants and protein-protein interactions: a multidimensional network-centric view. *Current opinion in structural biology* **50**, 82–90 (2018).

[29] Ozturk, K. & Carter, H. Identifying driver interfaces enriched for somatic missense mutations in tumors. *Methods in molecular biology* **1907**, 51–72 (2019).

[30] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–50 (2005).

[31] Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature genetics* **49**, 1779–1784 (2017).

[32] Dempster, J. M. *et al.* Extracting biological insights from the project achilles genome-scale CRISPR screens in cancer cell lines. *bioRxiv* (2019). URL https://www.biorxiv.org/content/early/2019/07/31/720243.

[33] Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).

[34] Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–8 (2010).

[35] Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**, D805–11 (2015).

[36] Goldman, N., Thorne, J. L. & Jones, D. T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–58 (1998).

[37] Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome biology* **20**, 223 (2019).

[38] Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248–9 (2010).

[39] Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature biotechnology* **35**, 128–135 (2017).

[40] Ponzoni, L. & Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 4164–4169 (2018).

[41] Moutinho, A. F., Trancoso, F. F. & Dutheil, J. Y. The impact of protein architecture on adaptive evolution. *Molecular biology and evolution* **36**, 2013–2028 (2019).

[42] Koh, G., Zou, X. & Nik-Zainal, S. Mutational signatures: experimental design and analytical framework. *Genome biology* **21**, 37 (2020).

[43] Hatakeyama, K. *et al.* Molecular profiling and sequential somatic mutation shift in hypermutator tumours harbouring POLE mutations. *Scientific reports* **8**, 8700 (2018).

[44] Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 14508–13 (2012).

[45] Bohrson, C. L. *et al.* Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nature genetics* **51**, 749–754 (2019).

[46] Fox, E. J. & Loeb, L. A. Lethal mutagenesis: targeting the mutator phenotype in cancer. *Seminars in cancer biology* **20**, 353–9 (2010).

[47] Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).

[48] Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nature genetics* **51**, 1170–1176 (2019).

[49] Gupta, K. & Varadarajan, R. Insights into protein structure, stability and function from saturation mutagenesis. *Current opinion in structural biology* **50**, 117–125 (2018).

[50] Chen, J. & Stites, W. E. Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry* **40**, 15280–9 (2001).

[51] Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome medicine* **9**, 34 (2017).

[52] Campbell, B. B. *et al.* Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).

[53] Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Human mutation* **36**, E2423–9 (2015).

[54] McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science translational medicine* **7**, 283ra54 (2015).

[55] Tsherniak, A. *et al.* Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).

[56] Hart, T. *et al.* Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda, Md.)* **7**, 2719–2727 (2017).

[57] Fornili, A., Pandini, A., Lu, H.-C. & Fraternali, F. Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. *Journal of chemical theory and computation* **9**, 5127–5147 (2013).

[58] Tuncbag, N., Gursoy, A. & Keskin, O. Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. *Physical biology* **8**, 035006 (2011).

[59] Kleinjung, J. & Fraternali, F. POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic acids research* **33**, W342–6 (2005).

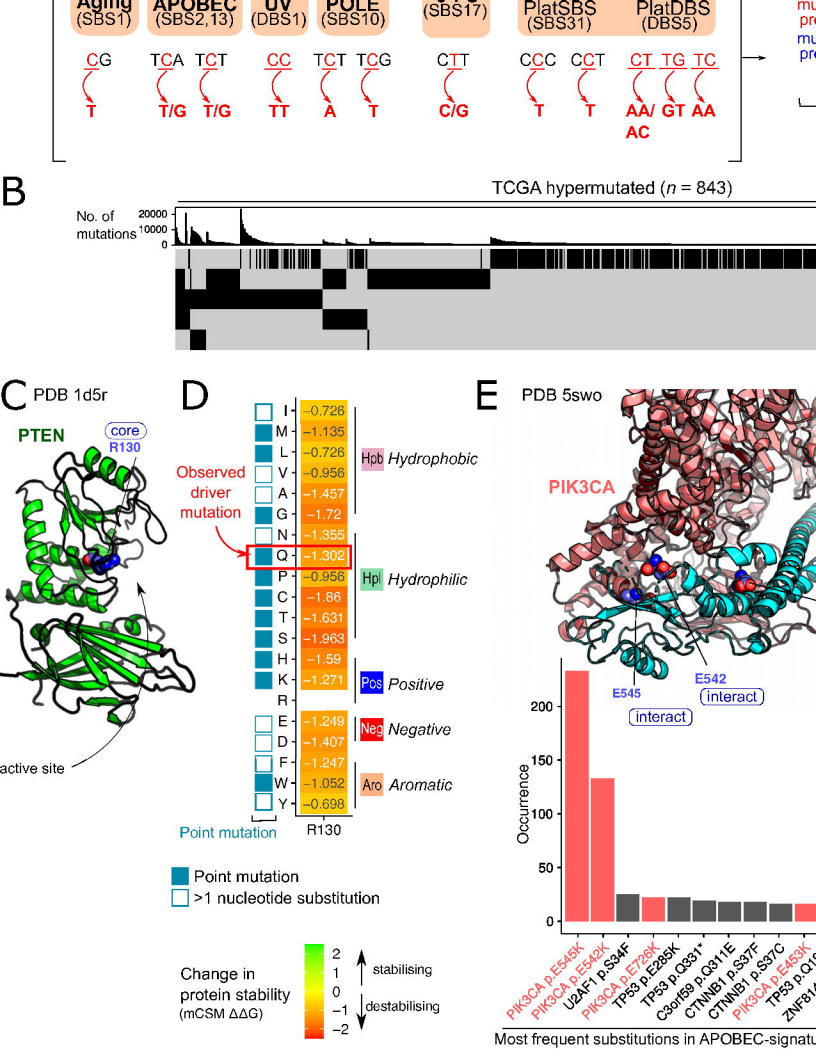
[60] Canty, A. & Ripley, B. D. *boot: Bootstrap R (S-Plus) Functions* (2017). R package version 1.3-20.

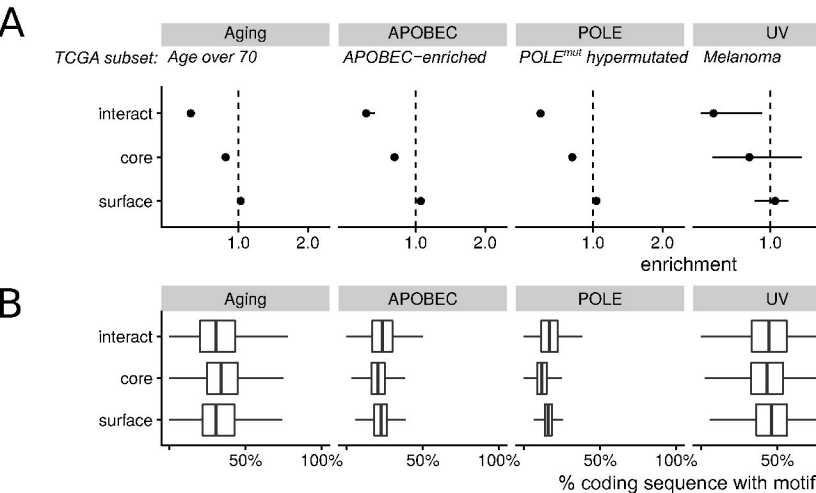
[61] Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Applications* (Cambridge University Press, Cambridge, 1997). URL http://statwww.epfl.ch/davison/BMA/. ISBN 0-521-57391-2.

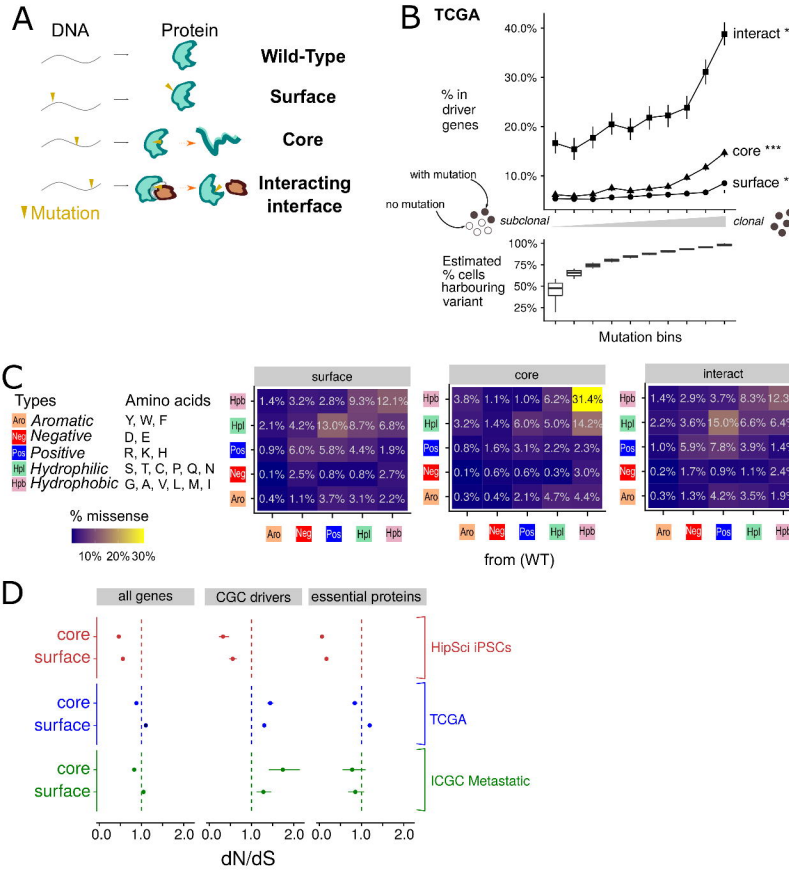
[62] Signorell, A. *et al.* *DescTools: Tools for Descriptive Statistics* (2019). URL https://cran.r-project.org/package=DescTools. R package version 0.99.31.

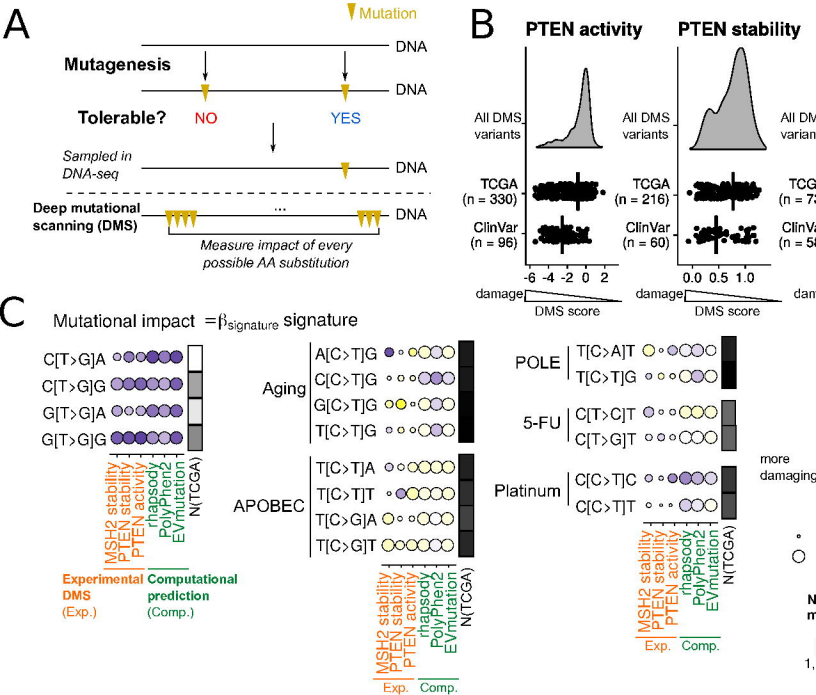
[63] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016). URL https://ggplot2.tidyverse.org.

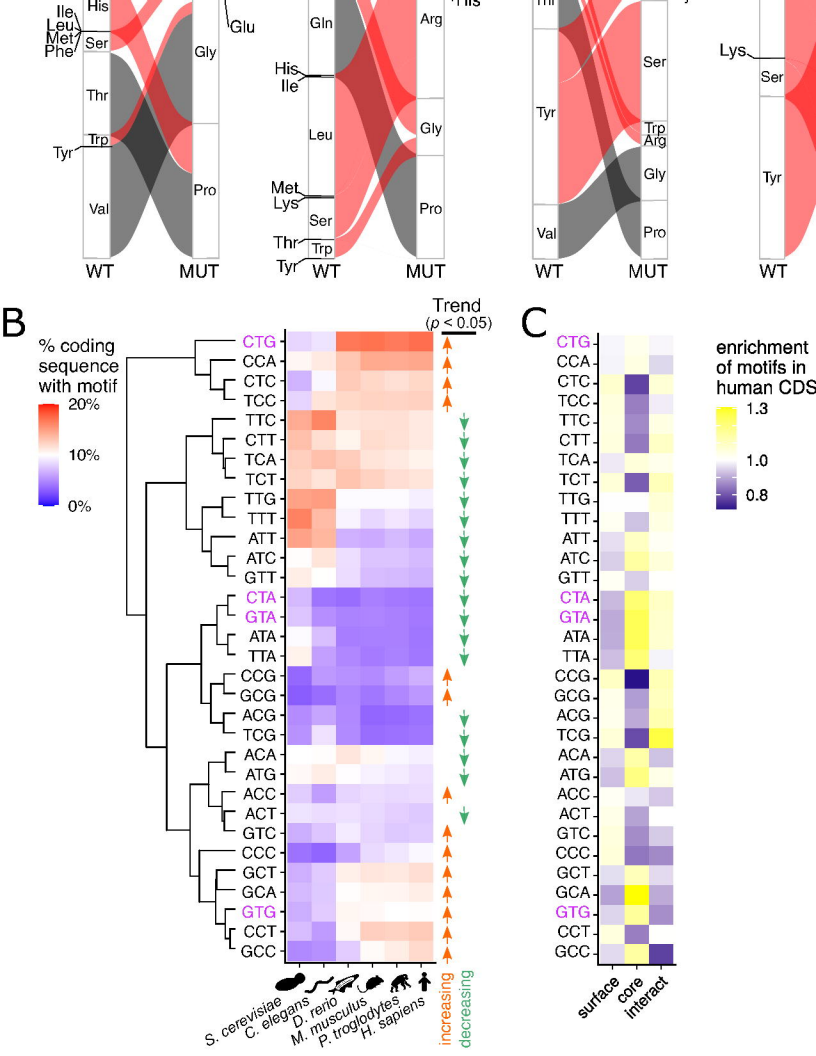
[64] Warnes, G. R. *et al.* *gplots: Various R Programming Tools for Plotting Data* (2019). URL https://CRAN.R-project.org/package=gplots. R package version 3.0.1.1.

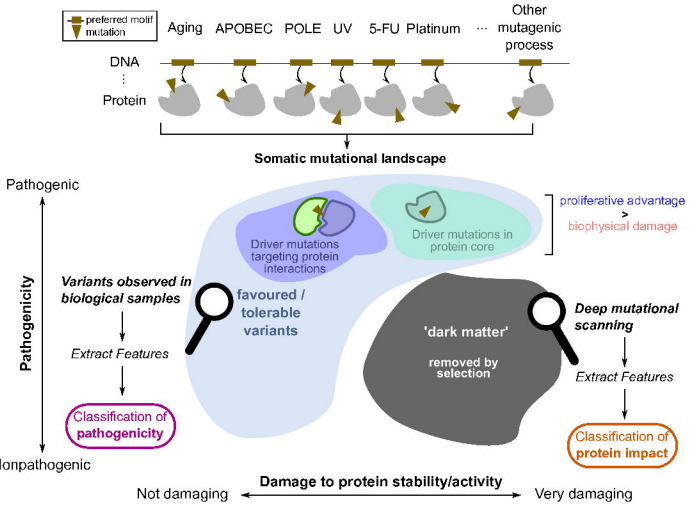

Aging (SBS1)
APOBEC (SBS2,13)
UV (DBS1)
POLE (SBS10)
(SBS17)
PlatSBS (SBS31)
PlatDBS (DBS5)
CG
TCA TCT
CC
TCT TCG
CTT
CCC CCT
CT TG TC
T
T/G T/G
TT
A T
C/G
T T
AA/ AC GT AA
B
TCGA hypermutated (n = 843)
No. of mutations
20000
10000
0
C
PDB 1d5r
PTEN
core
R130
active site
D
Observed driver mutation
I -0.726
M -1.135
L -0.726
V -0.956
A -1.457
G -1.72
N -1.355
Q -1.302
P -0.956
C -1.86
T -1.631
S -1.963
H -1.59
K -1.271
R
E -1.249
D -1.407
F -1.247
W -1.052
Y -0.698
Hpb Hydrophobic
Hpl Hydrophilic
Pos Positive
Neg Negative
Aro Aromatic
Point mutation
R130
Point mutation
>1 nucleotide substitution
Change in protein stability (mCSM ΔΔG)
2
1
0
−1
−2
stabilising
destabilising
E
PDB 5swo
PIK3CA
E545
interact
E542
interact
Occurrence
200
150
100
50
0
PIK3CA p.E545K
PIK3CA p.E542K
U2AF1 p.S34F
PIK3CA p.E726K
TP53 p.E285K
TP53 p.G331*
C3orf69 p.Q311E
CTNNB1 p.S37F
CTNNB1 p.S37C
PIK3CA p.E450K
TP53 p.Q1
ZNF81
Most frequent substitutions in APOBEC-signatu


A
Aging
APOBEC
POLE
UV
TCGA subset:
Age over 70
APOBEC−enriched
POLE^{mut} hypermutated
Melanoma
interact
core
surface
1.0
2.0
enrichment
B
50%
100%
% coding sequence with motif


A
DNA
Protein
Wild-Type
Surface
Core
Interacting interface
Mutation
B
TCGA
% in driver genes
40.0%
30.0%
20.0%
10.0%
interact ***
core ***
surface ***
with mutation
no mutation
subclonal
clonal
Estimated % cells harbouring variant
100%
75%
50%
25%
Mutation bins
C
Types
Amino acids
Aro Aromatic Y, W, F
Neg Negative D, E
Pos Positive R, K, H
Hpl Hydrophilic S, T, C, P, Q, N
Hpb Hydrophobic G, A, V, L, M, I
% missense
10% 20% 30%
surface
core
interact
Hpb 1.4% 3.2% 2.8% 9.3% 12.1%
Hpl 2.1% 4.2% 13.0% 8.7% 6.8%
Pos 0.9% 6.0% 5.8% 4.4% 1.9%
Neg 0.1% 2.5% 0.8% 0.8% 2.7%
Aro 0.4% 1.1% 3.7% 3.1% 2.2%
Hpb 3.8% 1.1% 1.0% 6.2% 31.4%
Hpl 3.2% 1.4% 6.0% 5.0% 14.2%
Pos 0.8% 1.6% 3.1% 2.2% 2.3%
Neg 0.1% 0.6% 0.6% 0.3% 3.0%
Aro 0.3% 0.4% 2.1% 4.7% 4.4%
Hpb 1.4% 2.9% 3.7% 8.3% 12.3%
Hpl 2.2% 3.6% 15.0% 6.6% 6.4%
Pos 1.0% 5.9% 7.8% 3.9% 1.4%
Neg 0.2% 1.7% 0.9% 1.1% 2.4%
Aro 0.3% 1.3% 4.2% 3.5% 1.9%
Aro Neg Pos Hpl Hpb
from (WT)
D
all genes
CGC drivers
essential proteins
core
surface
HipSci iPSCs
TCGA
ICGC Metastatic
0.0 1.0 2.0
dN/dS


A
Mutation
DNA
Mutagenesis
DNA
Tolerable?
NO
YES
Sampled in DNA-seq
DNA
Deep mutational scanning (DMS)
...
DNA
Measure impact of every possible AA substitution
B
PTEN activity
PTEN stability
All DMS variants
TCGA (n = 330)
ClinVar (n = 96)
TCGA (n = 216)
ClinVar (n = 60)
-6 -4 -2 0 2
0.0 0.5 1.0
damage
DMS score
C
Mutational impact = β signature signature
C[T>G]A
C[T>G]G
G[T>G]A
G[T>G]G
Aging
A[C>T]G
C[C>T]G
G[C>T]G
T[C>T]G
APOBEC
T[C>T]A
T[C>T]T
T[C>G]A
T[C>G]T
POLE
T[C>A]T
T[C>T]G
5-FU
C[T>C]T
C[T>G]T
Platinum
C[C>T]C
C[C>T]T
MSH2 stability
PTEN stability
PTEN activity
rhapsody
PolyPhen2
EVmutation1
N(TCGA)
Experimental DMS (Exp.)
Computational prediction (Comp.)
Exp.
Comp.
more damaging


His
Ile
Leu
Met
Phe
Ser
Thr
Trp
Tyr
Val
Glu
Gly
Pro
WT
MUT
Gln
His
Ile
Leu
Met
Lys
Ser
Thr
Trp
Tyr
Arg
Gly
Pro
Ser
Tyr
Trp
Arg
Gly
Val
Pro
Lys
Ser
Tyr
B
% coding sequence with motif
20%
10%
0%
Trend
(p < 0.05)
CTG
CCA
CTC
TCC
TTC
CTT
TCA
TCT
TTG
TTT
ATT
ATC
GTT
CTA
GTA
ATA
TTA
CCG
GCG
ACG
TCG
ACA
ATG
ACC
ACT
GTC
CCC
GCT
GCA
GTG
CCT
GCC
S. cerevisiae
C. elegans
D. rerio
M. musculus
P. troglodytes
H. sapiens
increasing
decreasing
C
enrichment of motifs in human CDS
1.3
1.0
0.8
surface
core
interact


preferred motif
mutation
Aging
APOBEC
POLE
UV
5-FU
Platinum
...
Other mutagenic process
DNA
Protein
Somatic mutational landscape
Pathogenic
Pathogenicity
Nonpathogenic
Variants observed in biological samples
Extract Features
Classification of pathogenicity
Driver mutations targeting protein interactions
Driver mutations in protein core
proliferative advantage
>
biophysical damage
favoured / tolerable variants
'dark matter'
removed by selection
Deep mutational scanning
Extract Features
Classification of protein impact
Damage to protein stability/activity
Not damaging
Very damaging