A framework for mutational signature analysis based on DNA shape parameters

Aleksandra Karolak [1,2], Fran Supek [2,3,*]

[1] Department of Population Sciences and Department of Computational and Quantitative Medicine, Division of Mathematical Oncology, Beckman Research Institute, City of Hope, 91010 Duarte, CA, USA.

[2] Genome Data Science, Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, 08028 Barcelona, Spain.

[3] Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain.

* correspondence to: fran.supek@irbbarcelona.org

Abstract

The propensity to acquire mutations depends on the oligonucleotide context of a DNA locus. In turn, this differential mutability of oligonucleotides varies across individuals due to exposure to mutagenic agents or due to variable efficiency of DNA repair pathways. Such variability is captured by mutational signatures, mathematical constructs resulting from a deconvolution of mutation frequency spectra across individuals. There is a need to enhance methods for inferring mutational signatures to make better use of sparse mutation frequency data that results from genome sequencing, and additionally to facilitate insight into underlying biological mechanisms. In cancer genomics, novel approaches to analyze somatic mutation patterns may help explain the etiology of various tumor types, as well as provide a more accurate baseline to infer positive and negative selection on somatic changes that drive tumor evolution. We propose a conceptualization of mutational signatures that represents oligonucleotides *via* descriptors of DNA conformation: base pair, base pair step, and minor groove width parameters. We demonstrate how such DNA structural parameters can accurately predict mutation occurrence due to DNA repair failures or due to exposure to diverse mutagens, including radiation, chemical exposure and the APOBEC cytosine deaminase enzymes. Furthermore, the mutation frequency of DNA oligomers classed by structural features can

accurately capture systematic variability in mutational spectra of >1,000 tumors originating from diverse human tissues. Overall, we suggest that the power of DNA sequence-based mutational signature analysis can be enhanced by drawing on DNA shape features.

Introduction

Advances in analysis of mutation profiles are transforming cancer genomics.[1–5] Tumor evolution is characterized by distinctive somatic mutational processes resulting from environmental mutagen exposures or defects in DNA repair mechanisms.[6–8] Identification of these processes can add to our knowledge of DNA damage and repair mechanisms that operate in human cells;[9,10] it can contribute to the understanding of etiology of various tumor types, with implications for predicting cancer risk;[11,12] it can improve statistical methodologies for detecting cancer driver genes by refining baseline estimates of mutation rates;[13,14] and finally, it has the potential to identify mutational biomarkers that can aid diagnostics[15,16] and personalized treatment of tumors.[17,18]

The genomic landscapes of individual cancers result from multiple overlapping mutational processes, making the deconvolution of the mutagenic processes from genomic data a challenge. There are many existing approaches to this (a recent overview of implementations can be found in [19]), which typically apply a factorization technique to a frequency table of occurrences of mutations in various trinucleotide DNA contexts. The resulting mutational signatures are able to recover known examples of mutagenic mechanisms present in certain cancer types.[1,2,17] However, many of the signatures were not able to be matched with a mechanism or a clear biological covariate,[1] and so might represent incompletely resolved mixtures of mutational processes, or also sequencing/alignment/mutation calling artefacts or also mathematical artefacts resulting from signature extraction. Additionally, the existing statistical approaches to infer mutational signatures – despite generally relying on similar methodologies – do not necessarily extract consistent sets of mutational signatures. Moreover, even with the same method, minor perturbations to the input data can result in different solutions. One reason for the lack of robustness is that the somatic mutation frequency data tend to be

sparse; Poisson noise due to low mutation counts can overwhelm biological signal. This is aggravated when changing the tabulation of oligonucleotides from the default trinucleotide (3 nt) DNA sequence representation, to longer, more descriptive representations – pentanucleotides (5 nt) or heptanucleotides (7 nt) – where the combinatorially increasing number of possible oligonucleotides increases sparseness. In addition to these statistical considerations, DNA sequence is usually not in obvious ways related to the biochemical aspects of the DNA damage and repair processes that generate the mutations, and so the sequence-based mutational signatures do not readily facilitate insight into underlying mechanisms.

Overall, there is a need to enhance the DNA oligomer representation for mutation signature analysis, in order to improve robustness of the methodologies for signature inference and also to improve interpretability of the signatures. Here, we propose a framework to integrate information about DNA structure[20,21] of the mutated DNA oligomers, providing substantial power to predict mutation occurrence at a locus and also to deconvolve underlying mutagenic mechanisms. We were inspired by the known individual examples of DNA structural features that appear susceptible to certain mutagens, such as DNA hairpin structures vulnerable to the APOBEC3A cytosine deaminase,[22] various types of DNA repeats with tendency to form non-B-DNA conformations associated with higher mutability,[23] curvature of longer DNA segments that associates with mutation rates,[24] or DNA structural features resulting from transcription factor binding that sensitizes to UV damage and consecutively mutation.[25,26] Our framework aims to generalize over these examples by associating mutational signatures with a descriptive representation of the DNA shape of neighborhoods of mutated loci in human cancer. Our implementation utilizes base-pair, base-pair step, and minor groove parameters of DNA oligomers;[27,28] such structural parameters have proven to be an accurate description of DNA conformation, summarizing atomic coordinates.[29–31] Because DNA predictably acquires a sequence-dependent local conformation, DNA susceptibility to mutagenic agents or recognition by DNA repair enzymes might be enhanced or disrupted by genetic differences in the regions flanking the mutation site.[32,33] This provides a rationale for implementing DNA shape parameters into a framework to predict and classify mutagenesis.

Materials and methods

Obtaining somatic mutations from cancer genomic data

We extracted somatic single-nucleotide variants (SNVs; henceforth: mutations) from the whole-genome sequences (WGS) of tumors from 30 cancer types; we did not consider indels or structural variation. The called mutations from cancer WGS were collected from: (i) The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga), studies: BLCA, BRCA, CESC, COAD, DLBC, GBM, HNSC, KICH, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, OV, PRAD, READ, SARC, SKCM, STAD, THCA, UCEC; (ii) International Cancer Genome Consortium data portal (ICGC, https://icgc.org), including somatic mutations from studies CLLE-ES, ESAD-UK, LIRI-JP, MALY-DE, MELA-AU, PACA-IT, RECA-EU; and (iii) samples downloaded from the websites of individual WGS study of MDBA.[34] The final list included over 1,600 tumor samples, for which the chromosome number and coordinate of each somatic mutation were extracted using R v3.6.[35] Next, DNA motifs up to +/- three nucleotides flanking each mutation site were retrieved using human GRCh37/hg19 as reference genome. This resulted in tri, penta, and heptanucleotide sequences ("3nt", "5nt", "7nt", respectively) with the mutation placed in the central position of the oligonucleotide. For the Poisson regression analyses (see below), all three types of DNA sequence motifs were further processed to count their mutation frequency in each tumor sample, and, as a baseline, their overall occurrence in the human genome. For the Principal Component (PC) analysis, each set of mutations (sorted by tri, penta or heptanucleotide) in a given tumor sample was further separated into six mutation outcomes: C>A, C>G, C>T, T>A, T>C and T>G (these are equivalent to, and were considered together with G>T, G>C, G>A, A>T, A>G and A>C, respectively, due to DNA strand symmetry).

DNA representation using structural features

Within pentanucleotide DNA motifs, we extracted various structural parameters thereof. This produced: one minor groove width parameter defined for the central nucleotide

(*mgw*0); six base pair (bp) parameters: propeller (*prop*0), opening (*open*0), buckle (*buck*0), stretch (*stre*0), stagger (*stag*0), shear (*shea*0); and additionally six base-pair step parameters, each preceding and following the central nucleotide (thus times two), resulting in total of twelve base-pair step parameters: twist (*tw*-1, *tw*+1), roll (*ro*-1, *ro*+1), tilt (*ti*-1, *ti*+1), slide (*sl*-1, *sl*+1), rise (*ri*-1, *ri*+1), and shift (*sh*-1, *sh*+1).[27] DNA shape parameters extraction was performed using DNAshape R routine,[36,37] which uses the Curves+ algorithm.[38] The total number of parameters increased from nineteen (1+6+12) for the pentanucleotide-based DNA structure representation ("5nt-str") to forty five for the heptanucleotide-based DNA structure representation (3+18+24; "7nt-str"). An illustration of the relationship between nucleotide position and the DNA structural parameters is shown for minor grove width, base-pair parameter propeller, and the base-pair step parameter roll in Figure 1. After extraction of the most informative parameters: minor groove width, propeller, twist, and roll, values of each parameter were normalized to range from 0 to 1, and discretized into three equally distributed bins: high ("H"), medium ("M"), and low ("L"), based on the known, nearly symmetrical distributions of the shape parameters around their equilibrium values.[28]

Figure 1. Mapping of the structural DNA parameters on DNA sequence, where N can be any of the four DNA bases: A (adenine), C (cytosine), G (guanine), or T (thymine). Structural features: minor groove width (mgw), base-pair parameters (bp, shown for propeller), and base-pair step parameters (bp step, shown for roll) and their position within the (A) 5nt-long and (B) 7nt-long DNA motifs.

Modelling mutation counts in oligonucleotides by Poisson regression

The DNA sequence features on the one hand, and DNA shape parameters on the other hand, were applied to the task of predicting the mutation propensity of each oligonucleotide, for six individual hypermutated tumor samples (described below in the Results section), using count modelling and in particular the Poisson regression analysis

as implemented in the R environment via the *glm* function.[39] For DNA sequence features, particular each trinucleotide or pentanucleotide was a data point in the regression and was described by the DNA sequence motif features (henceforth referred to as “3nt-seq” or “5nt-seq”, respectively; these are simply variables indicating occurrences of A, G, C or T at each position), the mutation count of the tri/pentanucleotide in that tumor sample, and its occurrence within the human reference genome GRCh37/hg19 (the latter was introduced as an offset value into the regression, thus adjusting for the differential occurrence of oligonucleotides in the genome sequence). For DNA structural features, the sets of all structural parameters extracted for pentanucleotide and heptanucleotides (henceforth, “5nt-str” and “7nt-str”), each divided into L, M, and H bins, was used in place of the 3nt-seq or 5nt-seq in the DNA sequence analysis. The McFadden pseudo-R^2 (henceforth, pR^2) statistic was used to evaluate model performance, for each set of features and for each of the six tumors. All Poisson regression analyses and visualization were performed with R v3.6.[35]

Estimating systematic variation in mutational patterns by principal component analysis (PCA)

PCA was performed jointly for blocks of features describing the DNA sequence, and features describing DNA structure. On the DNA sequence side, 96 mutation contexts were considered within each tumor sample. In order to keep the number of structural features the same, thus making the estimations of relative contributions of sequence and shape parts toward each principal component balanced, we prepared 96 mutation contexts on the DNA shape side. This was achieved by focusing only on the L/M/H binned (see above) features of the six DNA structure parameters with the highest potential for interpretation: *mgw*0, *prop*0, *ro-1*, *ro*+1, *tw*-1, and *tw*+1. Six parameters could sample a conformational space assigned to one of the three bins: L, M, and H, generating 18 features, which were examined separately for six mutation types (C>A, C>G, C>T, T>A, T>C, T>G; considered DNA strand symmetrically) and thereby a total of one hundred eight contexts were obtained. After removing energetically forbidden (not sampled), conformational parameters: *prop*0*(L)* and *ro*-1*(H)* on the C sites, and *prop*0*(H)* and

tw+1*(L)* on the T sites, the number 108 was reduced by 12 and resulted in 96 mutation contexts on the DNA structural side. Next, mutation counts in the structural block of each tumor sample were adjusted to match the sequence contribution to mutation burden in the corresponding sample, i.e. sampled to sum to the same values for the DNA sequence feature block and the DNA structure feature block. Additionally, the DNA structural features block was adjusted to have equal counts to the DNA sequence feature block, using a sampling function with the *UPmultinomial* function in R package *sampling*.[35] In this way, the noise due to low mutation counts will be equal in the two feature blocks and will not bias the PC analysis toward one block of features.

Results

Based on the known individual examples of DNA structural features associated with mutation rates,[22–25] we hypothesized that a set of diverse DNA structural features, when considered jointly, would be able to predict mutation rates due to diverse mutagenic exposures. To test this hypothesis, we employed a statistical method to predict mutational frequency of DNA segments using Poisson regression. In brief, this approach predicts mutation counts in a certain DNA oligonucleotide, which can be defined either by the DNA sequence features (occurrence of A, G, C or T in a certain position of the flanks of a locus) or by the DNA structural features (see Methods), while adjusting for the total number of occurrences of that oligonucleotide in the human genome sequence. Out of a total of ~1.6 thousand tumor samples covering 30 cancer types that were considered (see Methods), we selected six hypermutated tumor samples as representatives of important mutagenic processes: (i) microsatellite instability (MSI) caused by defective DNA mismatch repair in a representative colorectal tumor sample; (ii) activity of the mutant replicative polymerase epsilon (POLE) in another colorectal cancer sample; (iii) a bladder tumor sample bearing the mutational signature of the APOBEC cytosine deaminases;[20] (iv) a lung adenocarcinoma sample highly enriched with the tobacco smoking mutational signature; (v) ultraviolet (UV) light-induced mutagenesis in a melanoma sample; and (vi) the hypermutation induced by therapy by the DNA methylating drug temozolomide (TMZ) in a glioblastoma sample. Such hypermutated tumors derive the overwhelming majority of

their mutations from a single mechanism, thus allowing to examine these diverse mutagenic processes individually and to identify the sequence and/or structural features of highly mutable DNA oligomers.

Sequence and structural features of oligonucleotides that hypermutate upon DNA repair failures

The Poisson regression coefficients corresponding to DNA sequence features – derived from trinucleotide (Figure 2A, “3nt-seq”) and pentanucleotide (Figure 2B, “5nt-seq”) neighborhoods – quantify the impact of flanking DNA sites on mutational frequency of the central site. For example, in the MSI tumor (Figure 2A), the central cytosine (or, equivalently, the guanine it pairs with) is more mutable than the thymine (equivalently, adenine). Furthermore, a guanine in the immediate 3’ flanking position (“+1” in Figure 2A, MSI tumor) is further associated with a higher mutation rate of the central nucleotide in this tumor, consistent with very high mutation rates of the (commonly methylated) CpG dinucleotide in MMR-deficient cancers,[40] and consistent with the mutational signature SBS6 previously detected for many colorectal cancers with MSI.[1] The extended pentanucleotide neighborhood appears to have more subtle associations with the MSI mutagenesis, at least in this particular MSI tumor; nonetheless some enrichment with C at -2 and +2 positions were observed (Figure 2B)

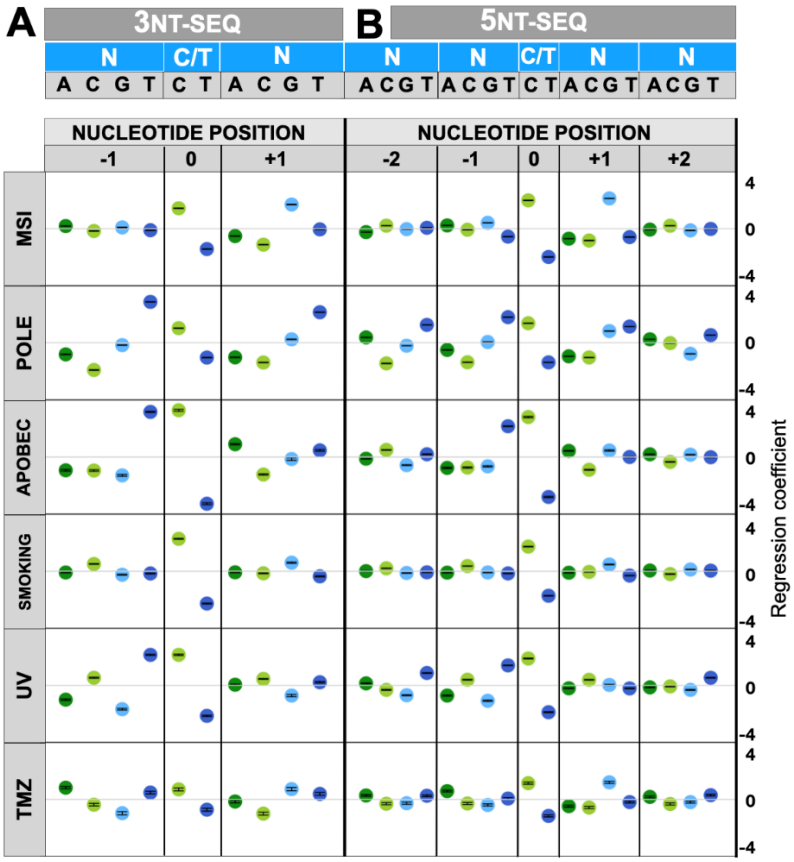


Figure 2. DNA sequence as predictor of mutational burden in tumor samples by mutagens: MSI, POLE, APOBEC, smoking, UV, TMZ. N can be any of the four DNA bases: A, C, G, or T (color coded respectively: dark green, light green, light blue, or dark blue). Poisson regression coefficients for (A) 3nt-seq, and (B) 5nt-seq. Error bars superimposed on each symbol show 95% C.I.

Next, we turned to examine the DNA structural features that associate with mutation occurrence in various loci of the MSI sample (top row, Figure 3A, "5nt-str") by examining Poisson regression coefficients for the high "(H)", medium "(M)" and low "(L)" bins of each structural feature, as derived from the pentanucleotide neighborhoods (see Methods). Among other features, we observed a slight positive association with the *mgw*0*(H)* bin (pentanucleotides with high minor groove width around the central position), as well as with the *tw*+1*(L)* bin (a low value of twist parameter for the +1 position; see Figure 1 for schematic). A further analysis using structural features derived from an extended, heptanucleotide neighborhood ("7nt-str"; top row, Figure 3B) confirm the above and further suggest a narrowing of the minor groove at the position +1. To quantify the overall utility of DNA structural features for predicting mutation rates, we examined the overall fit

of the model *via* the McFadden pseudo-R^2 statistic (pR^2; see Methods). The DNA structural features exhibited a higher predictive ability (pR^2=0.60 for the *7nt-str*) compared to the composition DNA sequence features (highest pR^2=0.56 for the *5nt-seq*). We note that our shape features were computed from the DNA sequences by employing data resulting from previous simulations of DNA structures of various oligonucleotides (see Methods). This increase in model fit by using the DNA structural features indicates that they capture those statistical interactions between sites in a DNA motif that are relevant for predicting mutability of the oligomer. In other words, this representation of DNA shape preserves important information for describing mutation rates, suggesting that DNA structural descriptors may be an appropriate representation for mutational signature analysis. In addition, the particular structural features highlighted by the regression (see above) may have potential for interpretation of mutagenic mechanisms in MMR-deficient, MSI tumor samples.

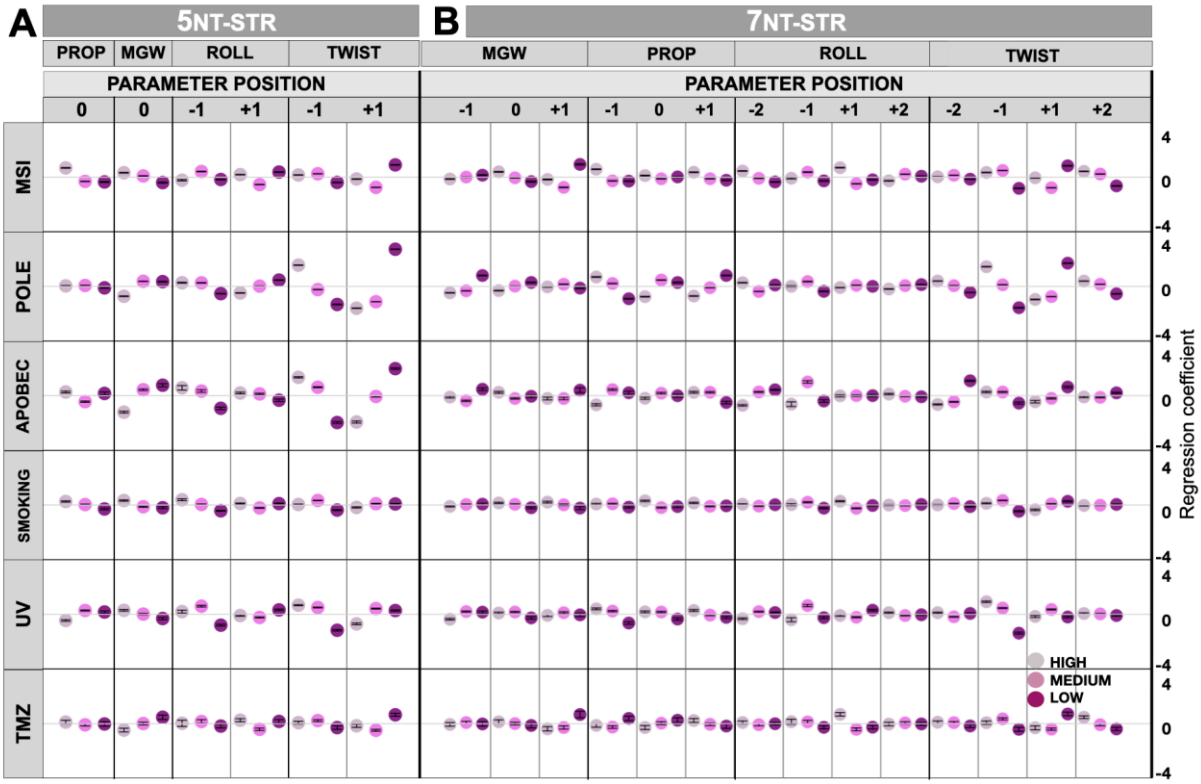


Figure 3. Poisson regression coefficients describing mutation rates resulting from six mutagenic processes (MSI, POLE, APOBEC, SMOKING, UV and TMZ) evaluated from (A) 5nt-str, and (B) 7nt-str representations. Parameters were normalized and divided into 3 equally spaced bins: high, medium, and low. Error bars are 95% C.I.

Next, we turned to examine mutations resulting from another sort of DNA repair failure common in tumors – deficient proofreading activity of the replicative DNA polymerase epsilon due to a mutation in the POLE gene.[41] In the POLE-signature enriched colorectal tumor sample (POLE row, Figure 2), the landscapes of regression coefficients from DNA sequence features contain many noticeable signals located further from the central nucleotide, consistent with previous work that suggested that up to nonanucleotide sized DNA motifs are informative for POLE mutagenesis.[13] We detected enrichments of thymine at -1 and at +1 position next to the mutated central cytosine, as well as the enrichments of thymine, and to some degree adenine, located at -2 and +2 positions (Figure 2). The observed pattern of regression coefficients resembles the mutational signatures SBS10a and SBS10b detected in colon and uterus cancers,[1] further supporting the use of our methodology based on count models (here, Poisson regression). More interestingly for the matter at hand, the DNA structural analyses of the mutated loci in the POLE tumor (Figure 3) show strong signal in the twist at -1 and +1 positions, where DNA is over-twisted and under-twisted, respectively. This suggests there is a local deformation at the POLE hypermutable sites, tapering off towards a more regular shape further away from the center. As with the MSI tumor, the overall fit of the predictive model in the POLE tumor, the structural DNA features compared favorably (pR^2=0.81 for *7nt-str*, Table 1) to the sequence DNA features (pR^2=0.81 for *5nt-seq*), further supporting that the structural features we examined are appropriate to describe the propensity of DNA sites to mutate in DNA repair-deficient tumors. The overall predictive power was higher for the POLE tumor than for MSI tumor, suggesting that mutational hotspots can be predicted from DNA shape to a variable extent across different mutational processes.

Table 1. Performance of four models: 3nt-seq, 5nt-seq, 5nt-str, and 7nt-str. Representative of distinct mutational processes, each row is a tumor sample with one strong mutational exposure.

Predominant signature	Cancer type	Hypermutation type	McFadden R^2			
			3nt-seq	5nt-seq	5nt-str	7nt-str
SBS6	COAD	MSI	0.57	0.56	0.55	0.60
SBS10	COAD	POLE	0.86	0.81	0.76	0.81
NA	BLCA	APOBEC	0.94	0.92	0.77	0.85
SBS4	LUAD	SMOKING	0.98	0.83	0.85	0.50
SBS7	SKCM	UV	0.89	0.82	0.66	0.74
NA	GBM	TMZ	0.78	0.58	0.52	0.28

Structural features associated with mutagenesis due to an endogenous DNA damaging activity of the APOBEC3A enzyme

In contrast to examples of DNA repair deficiencies in cancer (MSI, POLE), next we turned to examining DNA structural features of sites mutated by DNA damaging agents, both endogenous (APOBEC cytosine deaminases) or exogenous agents; the latter group includes radiation (UV) and chemicals (tobacco smoke, the DNA methylating drug TMZ). Regression on trinucleotide contexts in the APOBEC-enriched bladder cancer sample shows a very strong association of thymine at -1 position with the mutation frequency (Figure 2A, APOBEC row), as expected from known mutational signatures SBS2 and SBS13.[2] The pentanucleotide regression (Figure 2B) finds enrichment of pyrimidines at position -2, supporting the role of the APOBEC3A enzyme[42] rather than its APOBEC3B paralog in mutagenizing this particular tumor; this is consistent with high propensity towards APOBEC3A mutagenesis in bladder cancer.[43] With respect to the DNA shape analysis, we note that APOBEC binding and deamination occurs on single stranded DNA, where DNA shape parameters have altered interpretation or are not well defined. However, the steps subsequent to the APOBEC-mediated DNA damage (i.e. cytosine to uracil conversion), involving repair by the base excision repair pathway (e.g. the UNG protein), can involve double-stranded DNA, making the structural features we examined also pertinent to APOBEC mutagenesis. The local tendency towards over-twisting at -1 (upstream of the mutated C) and then under-twisting at +1 (downstream of the mutated C) is evident from the DNA structure coefficients in the flanking sites (Figure 3A). Analysis of structural features in a broader DNA context (Figure 3B, *7nt-str*) additionally shows that position -2 exhibits under-twisting at APOBEC3A mutated sites. Considered together with over- and under-twisting at positions -1 and +1, respectively, this suggests that the following base pair may be more prone to have flipped out bases. There are no notable changes in the roll parameter across the DNA motif, suggesting it is not DNA bending that affects APOBEC mutation frequency but rather the tendency toward exposing a single (here, central) nucleobase. Further supporting that a broader DNA oligomer context is predictive for APOBEC mutagenesis, the model fit of the *7nt-str* descriptors is higher than for *5nt-str* descriptors (Table 1). We note that the simple trinucleotide representation

of DNA sequence is highly predictive in itself (Table 1), probably reflecting the strict requirement for 3' T in APOBEC mutagenesis.

DNA structural features confer risk of mutation resulting from exogenous DNA damaging agents

Turning towards mutations resulting from exogenous agents, we examined a tobacco smoking signature-enriched lung adenocarcinoma sample. Such tumors are predominantly associated with the signature SBS4, consisting of C>A mutations in various trinucleotide contexts.[1,2] In accordance with SBS4, the trinucleotide and pentanucleotide DNA sequence coefficients indicate strong impact of the central C nucleotide, without particular preferences towards certain nucleotides in the flanking sequence (Figure 2, "smoking" row; of note, there is a slight preference towards upstream C). This is also reflected through the DNA shape, where the associations with mutational burden are likewise subtle, when considered individually (Figure 3, "smoking" row); there are, for instance, positive associations with higher values of propeller and roll parameters at positions 0 and +1, respectively. When considered jointly, the structural features do convey much information with predictive potential, which easily matches other considered mutational signatures (pR^2=0.85 for the *5nt-str* approach, Table 1).

In addition to the chemical-induced mutagenesis in lung cancer, we also examined radiation (UV light) induced mutagenesis in a melanoma skin cancer using the Poisson regression analysis (Figure 2, "UV" row). It is known that pyrimidine base pair steps define the hotspots for electron excitation by UV light, leading to the formation of cyclobutane dimers.[44–46] The condition of having TC or CC steps for the excitation process is reflected in the known mutational signatures SBS7a and SBS7b, as well as in our regression analysis using trinucleotides (*3nt-seq*, Figure 2A). Moreover, we observe the association with T as a flanking nucleotide at positions -2 and +2 (*5nt-seq*, Figure 2B), consistent with previously reported pentanucleotide contexts for the C>T UV-associated mutations.[47]

The DNA shape analysis also reflects a stronger preference of UV-mediated mutagenesis at C sites which are surrounded by T nucleotides, presumably due to electron excitation (Figure 3, UV row). For example, a strong signal is observed for "*tw*-1*(H)*" (high twist at the -1 position), which is known to reach its highest values at TC base pair steps.[29,48] Overall, the UV radiation induced mutation propensity of a site was highly predicable from the heptanucleotide-derived DNA shape features (*7nt-str*, pR^2=0.74, Table 1), similarly as for the DNA repair-related POLE (0.81) and APOBEC (0.85) mutagenesis. Thus, DNA shape features are useful to predict occurrence of mutations resulting from various causes, although this will not necessarily be the case for every mutagen to the same extent.

We examined a further chemical mutagenic agent: we applied the regression analysis to the temozolomide (TMZ)-treated glioblastoma tumor sample (Figures 2 and 3, TMZ row). TMZ is a DNA alkylating agent and its signature SBS11 has been detected in copious amounts in TMZ pretreated tumor samples.[1] We note subtle associations with individual descriptors in all four groups of DNA shape features (*mgw*, *prop*, *roll* and *twist*; Figure 3). Compared to the other mutagenic agents considered, the DNA shape features considered jointly appeared less predictive for TMZ mutations (Table 1). This suggests that either DNA shape is less important for activity of this particular chemical mutagen, or that our set of shape features does not incorporate those features that are relevant for TMZ mutagenesis. However, an overall pR^2 of 0.52 (at *5nt-str*) implies there is still some signal relevant for predicting TMZ mutations contained in the DNA shape.

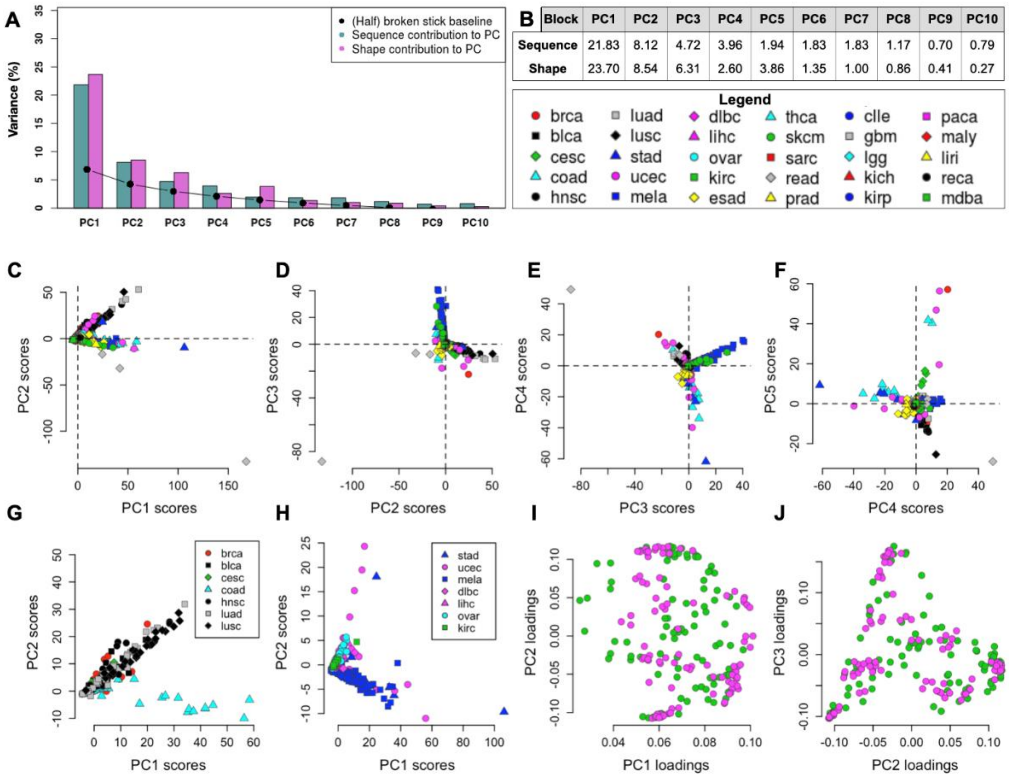
A comparison of performance of the Poisson regression models in predicting mutability of oligonucleotides suggests that the *7nt-str* model tends to be more predictive (higher pR^2) than *5nt-str*, at least for the six considered types of hypermutation. This indicates that the shape of DNA in broader, heptanucleotide neighborhoods is relevant for the intensity of many mutational processes. The pentanucleotide-based DNA structure model has overall similar accuracy to the sequence-composition based (*3nt-seq* and *5nt-seq*) predictors, although their relative ranking is variable across different mutagens.

Overall, the set of DNA shape features we examined is broadly reflective of oligonucleotide properties relevant for mutagenesis. Nonetheless we do not rule out that this set may be able to be further refined, thus possibly further improving its accuracy in predicting DNA mutational hotspots for a broader range of mutagens.

DNA shape features capture the systematic variability in mutational exposures across individuals

Above, we have shown that DNA structure-based descriptors were associated with mutations resulting from exposure to diverse mutagens. Some features appeared to commonly predict higher mutation rates resulting from different mutagens (e.g. under-twisting at +1 position, Figure 3). However, many other features appeared associated with some mutagens but not with others (e.g. high roll at +1, low minor groove width at +1, high propeller at -1, etc.; see Figure 3). Thus, we hypothesized that DNA shape descriptions of mutated loci in a genome sequence could be used to quantitate mutagenic exposures that cell has undergone previously. In other words, our DNA shape features would enable a different conceptualization of mutational signatures.[1,2] Mutational signatures are usually defined via trinucleotide neighborhoods of mutated loci, but broader neighborhoods were also considered previously (which however means that statistical interactions between nucleotides may need to be ignored).[49] To derive mutational signatures, various forms of factor analysis can be applied to the mutation frequency data; commonly, non-negative matrix factorization (NMF) was used[2] although many other methods were also applied.[19] Here, we evaluated the potential of DNA shape descriptors to generate informative mutational signatures. To this end we employed a principal components (PC) analysis to generate mutational signature PCs and measure the amount of systematic variability in the data, here implying the differential exposures to mutagenic effects across tumor samples. In the PC analysis, we used either the mutability of various DNA trinucleotides (sequence descriptors, *3nt-seq*) or mutability of various DNA shapes (*5nt-str* features; for the PC analyses, adjusted to match the number of the DNA sequence features, see Methods). Both sets of features were included simultaneously in a PC analysis of somatic mutations from over 1.6 thousand whole

genome sequences of tumors. This showed that DNA structure features can explain more systematic variance than the DNA sequence features in four out of the five dominant PCs: PC1, PC2, PC3, and PC5 (the PC analysis scree plot, broken down into the DNA sequence-feature part and the DNA shape-feature part, is shown in Figure 4A). Overall, the top five PCs explain 85.6 % variance in the mutation rates of DNA oligonucleotides across tumors, with the sequence features covering 40.6%, and structural features 45.0% variance (Figure 4B). The consistent excess of variance explained the structural features amongst the top PCs (Figure 4B) suggests they are more descriptive markers of the variability in mutagenic processes between human tumors, compared to a standard DNA sequence representation. Thus, DNA structural features are well-suited for inferring mutational signatures.



Block	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Sequence	21.83	8.12	4.72	3.96	1.94	1.83	1.83	1.17	0.70	0.79
Shape	23.70	8.54	6.31	2.60	3.86	1.35	1.00	0.86	0.41	0.27

Figure 4. PC analysis of DNA sequence and structural features in 1,637 cancer genomes in 30 cancer types. (A) Top 10 PCs separated into sequence and structural contribution (baseline is the half of a 'broken stick' estimate). (B) Numerical representation of each PC's contribution to the overall variance in the sequence or structure block. (C) PC1 vs. PC2 scores, (D) PC2 vs. PC3 scores, (E) PC3 vs. PC4 scores, (F) PC4 vs.PC5 scores, (G-H) Examples of PC1 and PC2 scores of cancer samples separated into smaller groups of tissues, (K-L) Examples of loadings between PC1 – PC2 and PC2 – PC3 corresponding to the (C) and (D) PC scores.

Over 45% of the variation that is explained by the dominant, first PC contains contributions from both the DNA shape features (23.7% variance) and also from the DNA sequence features (21.8% variance). All mutation features had positive loadings on the PC1 (Supplementary Table S1), indicating that the PC1 reflects the overall mutation burden across tumors, rather than the differential mutational signatures. We noted a consistently high PC1 loading on the various types of mutations at loci with low DNA roll at -1 position ("ro1L"; Supplementary Table S1) suggesting that this structural feature may be associated with higher DNA mutability generally i.e. in a manner not related to a particular mutagen. Furthermore, high-*mgw* and low-*twist* loci mutation frequencies had high PC1 loadings across multiple mutation types (Supplementary Table S1), suggesting additional structural features that characterize mutation-prone DNA loci.

In PC2 and the following PCs (Figure 4C-F), both positive and negative loadings of different DNA sequence features and structural features were observed. This means that these PCs are able to distinguish relative contributions of mutational processes across individual tumors i.e. the PCs constitute mutational signatures in a broad sense. Expectedly, these PCs distinguish between abundance of six different mutation types across individuals, for instance PC2 contrasts genomes rich in C>G and or C>A changes, from those rich in A>C or A>G changes, PC3 has a high positive weight on C>T changes and so on (see Supplementary Table S1). In addition to contrasting mutation types, the PC mutational signatures further distinguish between DNA structure and/or sequence contexts within each mutation type (see Supplementary Table S1). For instance, the PC2 – the dominant direction of differential mutability in our analysis – has high loadings (absolute value) for DNA under-twisting at the +1 position ("tw+1L" feature; Supplementary Table 1), mostly across the C>G and A>C mutation type. The PC3 similarly highlights importance of the twist parameter at +1 position but additionally for the C>T changes; this mutation type also shows differential PC3 loadings on the *mgw* and the *roll* parameters (Supplementary Table 1). Thus, diverse types of structural features appear to describe the differential mutation rates of DNA oligomers across individuals. We note that DNA sequence features may also have high contributions to the top PCs

(particularly PC4, PC7; Figure 4A), suggesting that the DNA sequence and the structural features are best considered jointly when inferring mutational signatures.

To further focus on how the dominant mutational signatures (here: first two PCs) vary across different types of cancer, we divided the full dataset (from Figure 4C) into smaller groups containing diverse cancer types (Figure 4G, H). The PC loadings of each original variable are plotted in Figure 4I, J. Again, the PC1 components have only positive loadings, describing overall mutation burden, while the PC2 and following describe differential mutation rates (Figure 4I, J). In the various groupings of cancer types tested, the PC2 separates between cancer types: in the particular examples of colorectal cancer *versus* others, or uterus cancer *versus* others (Figure 4G, H), suggesting the ability of the DNA shape descriptors to capture tissue-specific mutational processes.

Discussion

Our work highlights the ability of DNA shape features to predict mutational propensity of individual loci in human cancers that have undergone various mutagenic processes, ranging from DNA repair failures to exposures to mutagenic chemicals or radiation. Our predictive models further showcase the ability to identify DNA structural determinants associated with each particular mutagenic process, potentially furthering our understanding of the underlying mechanisms of DNA damage and repair. Furthermore, we demonstrate that the DNA shape features of the mutated loci can capture a higher amount of systematic variability in mutational processes across cancer samples than a plain representation of the DNA sequence composition. We note that these two groups of features (DNA shape and sequence) have a degree of redundancy – expectedly so, given that the shape features were derived from oligonucleotide sequences. Importantly, however, the shape features additionally capture the interactions across neighboring nucleotides within a pentanucleotide or longer neighborhood, while keeping the overall representation relatively simple, and thus useful for analysis of (relatively sparse) somatic mutation count data. If the full set of statistical interactions were to be represented for

DNA sequence representations, the number of features would rise combinatorially, making it very difficult to analyze heptanucleotide and longer contexts (for the exome sequencing data, which is most abundant, even the pentanucleotides are likely out of reach). Use of structural features however makes such analyses of broader DNA contexts feasible; this is clearly of interest, given that heptanucleotide and even nonanucleotide contexts appear important for some mutagenic processes, such as UV mutagenesis,[47] POLE mutagenesis in cancers,[13] and also some mutational processes in the human germline.[50,51] We suggest that use of structural DNA features may help overcome hurdles for analyzing the roles of longer oligonucleotide neighborhoods as determinants of DNA mutability.

Additionally, there is a growing awareness that analysis of mutational patterns may be able to guide personalized therapy for cancer patients. While current focus is on refining drug selection or delivery guided by the cancer-type of origin or synergistic potential between drugs, rather than the underlying genetic profile of tumors, there is much potential in using mutational profiles to better stratify patients.[52,53] This might entail ascertaining existence of certain driver mutations that can be matched to targeted drugs (classically EGFR or BRAF mutations or HER2 amplifications, but many more are investigated).[54] However another genetic marker of rising importance are genomic profiles of mutational signatures. Because these reflect ongoing genomic instability, which is common in tumors, they could help better stratify patients for targeted therapies, as exemplified in signatures of deficiency in DNA mismatch repair (MMR) nominating patients for immunotherapy.[55,56] The mutational Signature SBS3 and also a pattern of deletions with microhomology signal failures in the homologous recombination repair pathway (e.g. via loss-of-function in the BRCA1 gene), which is a vulnerability targetable by PARP inhibitors.[17,18] In addition, mutational analysis might conceivably predict when resistance to therapy arises; for example, given that the error-prone DNA polymerase eta can overcome treatment by cisplatin,[57] identification of its mutational signatures – previously via DNA sequence features[43] – but possibly also by the affinities of this error-prone DNA polymerase(s) towards DNA shape features – could inform treatment decisions. In summary, there is great promise for clinical use genomic markers in tumors

even though many mechanisms remain elusive,[54] and among such genomic markers, the utility of mutational processes in particular may merit more attention.

With a joint DNA sequence and structure representation as a basis for mutation signature inference, we posit that future work will be able to disentangle the mutagenic mechanisms at increased accuracy compared to current, sequence-only representations. Such future analyses would not necessarily focus only on hypermutator tumor samples – as we have done here to establish a proof-of-principle – but would also be able to look at DNA shape determinants of the more subtle mutational processes (which are more challenging to unmix). Also, future methods based on DNA structure-based encoding principles we described here would be able to more accurately deconvolute mutational processes in tumors where more than one process occurs simultaneously at high levels. Next, the current implementation of the multi-sample analysis, based on a simple principal components analysis approach, suggests that follow-up work using other, more refined mutational signature extraction methodologies would be fruitful. In particular, NMF approaches are popular for analysis of mutational compendia, because the resulting signatures have attractive properties for interpretation (non-negativity and sparseness) and also because of empirical evidence: the NMF-based signatures have been shown to sometimes map to *bona fide* mutagenic mechanisms. Our representation of mutation rates of DNA oligomers using shape properties could be adapted to provide data for analyses using NMF[2] or other upcoming approaches for mutational signature deconvolution that use e.g. topic models.[58,59]

In conclusion, we propose a framework for quantifying mutational signatures via use of DNA shape descriptors, which may advance mechanistic understanding of mutagenesis and also investigation of somatic mutational landscapes.

Acknowledgements

This work has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014-2020), under the Marie Skłodowska-Curie

PROBIST grant agreement No. 754510 (to A.K., PROBIST co-fund fellowship of the Barcelona Institute of Science and Technology) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 757700, to F.S). F.S. is funded by the ICREA Research Professor program. A.K. and F.S. acknowledge support of the Severo Ochoa Centres of Excellence program of the Spanish Ministry of Economy and Competitiveness to the IRB Barcelona. Work in the laboratory of F.S. was supported by the ERDF/Spanish Ministry of Science, Innovation and Universities-Spanish Research State Agency/RegioMut project (grant agreement No. BFU2017-89833-P).

References

1. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* (2020) doi:10.1038/s41586-020-1943-3.
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* (2013) doi:10.1038/nature12477.
3. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology* (2018) doi:10.1038/s41571-018-0002-6.
4. Glockzin, G., Schlitt, H. J. & Piso, P. Therapeutic options for peritoneal metastasis arising from colorectal cancer. *World J. Gastrointest. Pharmacol. Ther.* (2016) doi:10.4292/wjgpt.v7.i3.343.
5. Nakagawa, H. & Fujita, M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Science* (2018) doi:10.1111/cas.13505.
6. Lim, B. *et al.* Genetic alterations and their clinical implications in gastric cancer peritoneal carcinomatosis revealed by whole-exome sequencing of malignant ascites. *Oncotarget* (2016) doi:10.18632/oncotarget.6977.
7. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* (2012) doi:10.1038/nature11252.
8. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics* (2014) doi:10.1038/nrg3729.
9. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* (2019) doi:10.1016/j.cell.2019.03.001.
10. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* (2019) doi:10.1016/j.cell.2019.02.012.
11. Behjati, S. *et al.* Mutational signatures of ionizing radiation in second malignancies. *Nat. Commun.* (2016) doi:10.1038/ncomms12605.
12. Poon, L. S. *et al.* Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med.* (2015) doi:10.1186/s13073-015-0161-3.
13. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic

Tissues. *Cell* (2017) doi:10.1016/j.cell.2017.09.042.
14. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* (2020) doi:10.1038/s41588-019-0572-y.
15. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLoS Comput. Biol.* (2019) doi:10.1371/journal.pcbi.1006953.
16. Marquard, A. M. *et al.* TumorTracer: A method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Med. Genomics* (2015) doi:10.1186/s12920-015-0130-0.
17. Staaf, J. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* (2019) doi:10.1038/s41591-019-0582-4.
18. Gulhan, D. C., Lee, J. J. K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0390-2.
19. Omichessan, H., Severi, G. & Perduca, V. Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. *PLoS One* (2019) doi:10.1371/journal.pone.0221235.
20. Morgan, M. A. & Shilatifard, A. Chromatin signatures of cancer. *Genes and Development* (2015) doi:10.1101/gad.255182.114.
21. Madakashira, B. P. & Sadler, K. C. DNA methylation, nuclear organization, and cancer. *Frontiers in Genetics* (2017) doi:10.3389/fgene.2017.00076.
22. Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science (80-.).* (2019) doi:10.1126/science.aaw2872.
23. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* (2018) doi:10.1101/gr.231688.117.
24. Duan, C. *et al.* Reduced intrinsic DNA curvature leads to increased mutation rate. *Genome Biol.* (2018) doi:10.1186/s13059-018-1525-y.
25. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.* (2018) doi:10.1038/s41467-018-05064-0.
26. Fredriksson, N. J. *et al.* Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLoS Genet.* (2017) doi:10.1371/journal.pgen.1006773.
27. Lu, X. J. & Olson, W. K. 3DNA: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* (2008) doi:10.1038/nprot.2008.104.
28. Karolak, A. & van der Vaart, A. Enhanced sampling simulations of DNA step parameters. *J. Comput. Chem.* (2014) doi:10.1002/jcc.23751.
29. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.* (1998) doi:10.1073/pnas.95.19.11163.
30. Abe, N. *et al.* Deconvolving the recognition of DNA shape from sequence. *Cell* (2015) doi:10.1016/j.cell.2015.02.008.

31. Lankaš, F., Šponer, J., Langowski, J. & Cheatham, T. E. DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.* (2003) doi:10.1016/S0006-3495(03)74710-9.
32. Karolak, A. & Van Der Vaart, A. BII stability and base step flexibility of N6-adenine methylated GATC motifs. *Biophys. Chem.* (2015) doi:10.1016/j.bpc.2015.05.001.
33. Karolak, A. & Van Der Vaart, A. Molecular Dynamics Simulations of 5-Hydroxycytosine Damaged DNA. *J. Phys. Chem. B* (2016) doi:10.1021/acs.jpcb.5b09250.
34. Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* (2017) doi:10.1038/nature22973.
35. R Core Computing Team. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput. Vienna, Austria* (2017).
36. Chiu, T. P. *et al.* DNAshapeR: An R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btv735.
37. Zhou, T. *et al.* DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.* (2013) doi:10.1093/nar/gkt437.
38. Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D. & Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* (2009) doi:10.1093/nar/gkp608.
39. Zeileis, A., Kleiber, C. & Jackman, S. Regression models for count data in R. *J. Stat. Softw.* (2008) doi:10.18637/jss.v027.i08.
40. Poulos, R. C., Olivier, J. & Wong, J. W. H. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkx463.
41. Shinbrot, E. *et al.* Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.* (2014) doi:10.1101/gr.174789.114.
42. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* (2015) doi:10.1038/ng.3378.
43. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* (2017) doi:10.1016/j.cell.2017.07.003.
44. Goto, N. *et al.* Detection of UV-induced cyclobutane pyrimidine dimers by near-infrared spectroscopy and aquaphotomics. *Sci. Rep.* (2015) doi:10.1038/srep11808.
45. Martinez-Fernandez, L., Banyasz, A., Esposito, L., Markovitsi, D. & Improta, R. UV-induced damage to DNA: Effect of cytosine methylation on pyrimidine dimerization. *Signal Transduct. Target. Ther.* (2017) doi:10.1038/sigtrans.2017.21.
46. Cadet, J. & Douki, T. Formation of UV-induced DNA damage contributing to skin cancer development. *Photochemical and Photobiological Sciences* (2018) doi:10.1039/c7pp00395a.

47. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* (2010) doi:10.1038/nature08658.
48. Fujii, S., Kono, H., Takenaka, S., Go, N. & Sarai, A. Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.* (2007) doi:10.1093/nar/gkm627.
49. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLoS Genet.* (2015) doi:10.1371/journal.pgen.1005657.
50. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* (2016) doi:10.1038/ng.3511.
51. Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* (2018) doi:10.1038/s41467-018-05936-5.
52. Harmon, R. L. & Sugarbaker, P. H. Prognostic indicators in peritoneal carcinomatosis from gastrointestinal cancer. *International Seminars in Surgical Oncology* (2005) doi:10.1186/1477-7800-2-3.
53. McMullen, J. R. W., Selleck, M., Wall, N. R. & Senthil, M. Peritoneal carcinomatosis: Limits of diagnosis and the case for liquid biopsy. *Oncotarget* (2017) doi:10.18632/oncotarget.16480.
54. Middleton, G. *et al.* The National Lung Matrix Trial of personalized therapy in lung cancer. *Nature* (2020) doi:10.1038/s41586-020-2481-8.
55. Schrock, A. B. *et al.* Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer. *Ann. Oncol.* (2019) doi:10.1093/annonc/mdz134.
56. Sahin, I. H. *et al.* Immune checkpoint inhibitors for the treatment of MSI-H/MMR-D colorectal cancer and a perspective on resistance mechanisms. *British Journal of Cancer* (2019) doi:10.1038/s41416-019-0599-y.
57. Srivastava, A. K. *et al.* Enhanced expression of DNA polymerase eta contributes to cisplatin resistance of ovarian cancer stem cells. *Proc. Natl. Acad. Sci. U. S. A.* (2015) doi:10.1073/pnas.1421365112.
58. Funnell, T. *et al.* Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput. Biol.* (2019) doi:10.1371/journal.pcbi.1006799.
59. Matsutani, T., Ueno, Y., Fukunaga, T. & Hamada, M. Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz266.