Mutation rate variations in the human genome are encoded in DNA shape

Zian Liu[1]

Md Abul Hassan Samee[1, 2]

1: Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX 77030

2: Corresponding author

Email: zian.liu@bcm.edu, samee@bcm.edu

Author Contributions: ZL and MAHS conceived the study, ZL performed data processing and modeling, ZL and MAHS analyzed and interpreted the results, ZL and MAHS wrote the manuscript.

Competing Interest Statement: We have no competing interests to disclose.

Keywords: single nucleotide mutation rate, machine learning, population genetics, DNA shape

This PDF file includes:

Main Text
Figures 1 to 4
Tables 1 to 5

Abstract

Single nucleotide mutation rates have critical implications for human evolution and genetic diseases. Accurate modeling of these mutation rates has long remained an open problem since the rates vary substantially across the human genome. A recent model, however, explained much of the variation by considering higher order nucleotide interactions in the local (7-mer) sequence context around mutated nucleotides. Despite this model's predictive value, we still lack a clear understanding of the biophysical mechanisms underlying the variations in genome-wide mutation rates. DNA shape features are geometric measurements of DNA structural properties, such as helical twist and tilt, and are known to capture information on interactions between neighboring nucleotides within a local context. Motivated by this characteristic of DNA shape features, we used them to model mutation rates in the human genome. These DNA shape feature based models improved both the accuracy (up to 14%) and the interpretability over the current nucleotide sequence-based models. The models also discovered the specific shape features that capture the most variability in mutation rates, and distinguished between the most and the least mutated sequence contexts, thus characterizing mutation promoting properties of the genomic DNA. To our knowledge, this is the first attempt that demonstrates the structural underpinnings of nucleotide mutations in the human genome and lays the groundwork for future studies to incorporate DNA shape information in modeling genetic variations.

Main Text

Introduction

Nucleotide mutation is an important genetic process showing potentially causal association with diseases like cancer, autism spectrum disorder, and Alzheimer's disease (1-7). Single nucleotide mutation rate ("mutation rate") represents the probability of a single nucleotide mutating in an individual genome. Human and other mammalian genomes show extremely high variations in mutation rates (8). Predictive modeling of mutation rates and explaining the source of its genome-wide variations is an essential goal in human genetics. This understanding is the key to studying evolutionary divergence between species, inferring ancestral states, detecting adaptive evolution (8-10), identifying functional elements in the genome and predicting deleterious single nucleotide variants in the human genome (1-3,8), and identifying disease subtypes (4-7).

Predicting the high variations in genome-wide mutation rates is known to be difficult. Multiple molecular factors could influence these rates, yet none of these factors provide a satisfactory explanation of mutation rate variations in the human genome (8,10-13). Sequence context, i.e., the DNA sequence flanking a mutated position, has been hypothesized to be a strong genetic factor influencing mutation rates. The "CpG context", for example, explains the 14-fold increase in mutation rates in the context of CpG dinucleotides (8,11-12). However, it was not clear whether this idea could be extended to k-mer contexts, i.e., the k nucleotide long DNA sequence flanking a mutated position (8,11,13-16). Aggarwala and Voight recently showed that mutation rates estimated from 7-mer sequence contexts (three nucleotides upstream and downstream from the mutated positions) could optimally explain mutation rate variations in the human non-coding genome (9). They also showed that these estimates are (a) not influenced by rates of recombination, (b) strongly correlated with rates of species divergence, (c) are consistent for both rare and common genetic variants, and (d) are also reflected in de novo mutational events (9). Remarkably, they were able to model these estimates using linear regression allowing up to fourth-order interactions between nucleotides and explain ~81% of mutation rate variations in the human noncoding genome (9). Although Aggarwala and Voight's model is a breakthrough, it suffers from a lack of interpretability. In particular, the higher order interactions between nucleotides make it challenging to understand the biophysical basis of mutation rate variations (9). In this study, we questioned whether it is possible to build an accurate mutation rate prediction model that is also biophysically interpretable.

DNA shape features represent k-mer-dependent biophysical features that describe the local three-dimensional structures of the DNA molecule (16-24). DNA shapes are highly interpretable as they represent biophysical properties and essentially capture the interactions between neighboring nucleotides. However, the use of DNA shape features has been mostly limited to predicting transcription factor (TF) binding (25-34), with only one study associating DNA curvature with mutation rates in the URA3 gene in yeast (35). Taking advantage of population-scale whole genome sequence data from the 1000 Genomes project (1KG) (14,36) and the accurate predictions of DNA shape features by the recent *DNAShapeR* model (21-24), here we explore the relationships between DNA shape features and mutation rate variations in the human genome. We modeled Aggarwala and Voight's estimated mutation rates using DNA shape features and show that DNA shape features can provide more accurate and biophysically interpretable models than those with higher order interactions between nucleotide sequence features. Interpreting our models, we find that DNA roll, shift, tilt, stretch, and helical twist encode the most information related to mutation rate changes. We also find that DNA shape distinguishes between the most and the least mutated sequence contexts with DNA helical twist and roll being the most predictive features. Overall, our study presents the strong relationships between DNA shape and mutation rate variation in the human genome and lays the groundwork for future genetic studies to exploit the structural underpinnings of the DNA molecule.

Results

DNA shape features improve the modeling of mutation rates estimated from an expanded sequence context

Aggarwala and Voight recently showed that mutation rates estimated from 7-mer sequence contexts could optimally explain mutation rate variations in the human genome (9). Briefly, for each class of mutation, say A-to-C, they partitioned its occurrences according to the k nucleotides flanking the mutated position ($\lfloor k/2 \rfloor$ nucleotides on both sides). Then, instead of studying the genome-wide A-to-C mutations, Aggarwala and Voight studied k_1-to-k_2 mutations for k-mers k_1 and k_2 that differ only in the mutated (middle) position, where k_1 has an A and k_2 has a C. In this setup, k=1 represents the conventional approach for studying mutation rates. However, by increasing the value of k from 1 to 7, Aggarwala and Voight showed that, the k_1-to-k_2 mutation rates estimated from a training set of human chromosomes correlates significantly better with the mutation rates estimated from a validation set.

Aggarwala and Voight used these mutation rates estimated from 7-mer sequence contexts as the data for modeling mutation rate variation. In particular, for a pair k_1 and k_2 of 7-mers that differ at the central (mutated) position, they modeled $k_1 \rightarrow k_2$ mutation rates from the nucleotides in k_1 and k_2. They found a linear regression model to be optimal for this data, explaining 81% of the variance in mutation rates (9). Although this performance is remarkable, the model suffers from a lack of interpretability due to dependence on higher order (up to fourth-order) interaction terms between nucleotides flanking the mutated position. Since DNA shape is a biophysical representation of interactions between neighboring nucleotides, we hypothesized that DNA shape features could improve the performance and interpretability of the above model of genome-wide mutation rates. We tested this hypothesis using a series of increasingly complex models from DNA nucleotide and shape features, and selected the optimal models using an 8-fold CV strategy.

Like Aggarwala and Voight, we separately modeled each mutation class after combining mutation rates from reverse-complement sequences. If the mutated nucleotide was a cytosine (C), we considered two mutation classes based on whether a guanine (G) follows the cytosine (CpG context) or not (non-CpG). Thus, we built models for nine mutation classes (Figure 1A). We fit the models with nucleotide and shape features (Figure 1B-C), either alone or in combination, while systematically increasing model complexities by including higher order interactions (Figure 1D). In particular, we allowed for up to fourth-order interactions between nucleotide features following Aggarwala and Voight, and up to second-order interactions between DNA shape features since their interactions beyond second-order are not well-understood (21). We trained each model by optimizing L1-penalized mean squared error loss, and selected the optimal model based on 8-fold cross validation (Methods). We then tested each model on a separately held-out test data and confirmed the model performances are robust across training and testing data (Table S1-4). Model coefficients were also documented for two of our models (Table S5-6).

Across all nine mutation classes, the models containing second-order shape interactions or combining shape and nucleotide features outperformed the current best model of Aggarwala and Voight (Figure 2A-C, Table 1, Figure S3-4). The models showed a median improvement of 3% (2.48% additional explained variance) with the highest improvement of 14.68% improvement (8.12% additional explained variance) in C-to-A mutations in CpG context. The other classes that benefitted the most from shape features are A-to-T, C-to-A (both CpG and non-CpG), and C-to-G. Below we further tease out the interplay between nucleotide and shape features and interpret our models in terms of DNA shape.

Models using DNA shape features alone, with at most second-order interactions, improve over the current best models that use nucleotide features and fourth-order interactions

Given the above improvements in model performance through introducing DNA shape features, we asked whether and to what extent shape features could replace nucleotide features. Thus, we focused on the models that use DNA shape features alone. As noted above, these models were allowed at most second-order interactions between shape features, and the interactions were limited to feature values only from adjacent positions. In the following, we refer to these models

as shape-only models. We use the term sequence-only models to refer to Aggarwala and Voight's models utilizing nucleotide features with up to fourth-order interactions, and the term combined models to refer to our models described above that use both sequence and shape features.

Remarkably, our shape-only models with second-order interactions outperformed Aggarwala and Voight's sequence-only models by over 1.7% for four of the nine mutation classes and had similar amounts of variance explained for the other five classes (Table 2, Figure S5-6). The improvements were largest in the mutation classes A-to-T and CpG C-to-A, where shape-only models represented 11.18% and 14.68% improvements, or additional 6.8% and 8.1% variance explained (Table 2). Overall, we found that DNA shape features could fully replace the nucleotide features, as well as mostly replace the combined features with only a median of 1.2% change in performance (1% drop in variance explained) as compared to the combined models. (Table 1-2, Table S2)

Single nucleotide mutation rates depend on both generic and mutation class specific shape features

Since our shape-only models with up to second-order interactions were consistently on par with nucleotide sequence-based models (Aggarwala and Voight's sequence-only models and our own combined models), we reasoned that interpreting the shape-only models could elicit the DNA structure-based underpinnings of mutation rates. Thus, we asked a series of questions about the importance of different shape features in the models for different mutation classes. We also asked if the importance of a shape feature changes with its distance from the mutated position and whether the shape features in the reference 7-mer or that in the alternative 7-mer are more important.

We first normalized the model coefficients to make them comparable across the nine classes (Methods). We defined the importance of a shape feature as the absolute value of its coefficient in the model and its utilization as the number of times it was selected in a model divided by the total number of shape features used in that model.

We noted that across all mutation classes, second-order features are more commonly utilized than first-order features, where they have utilizations of 70-86% (Figure 3A). This is expected since the shape features do not show high correlations and a second-order feature would capture more information than a first order feature. However, the utilization of second-order features systematically decreases in the more important bins (Figure 3A). Importantly, the most utilized shape feature, regardless of its appearance as a first-order term or a second-order term, are consistent between classes. We noticed that Roll, Shift, Tilt, and Stretch are consistently the most utilized features across all nine mutation classes (utilization of ~10%) (Figure 3B-C). The features ProT, Shear, Stagger, MGW, and Buckle had utilizations between 5% to 10%; and the least utilized (<5%) features were HelT, EP, Opening, and Rise (Figure 3B-C).

We noticed that HelT is highly enriched in the top 5-10% features despite being overall underutilized, suggesting it as primarily useful for predicting large variations in mutation rates (Table 3, Table S7-9). In contrary, the aforementioned Roll, Shift, Tilt, and Stretch all have low enrichment values (Table 3, Table S7) with Roll, Shift, and Tilt often enriched in bins of lower feature importance (Table S8-9); their abundant usage in the model coupled with low enrichment suggest that these features explain subtle changes in mutation rates beyond those explained by changes in HelT.

Since DNA shape features are location-dependent and we have modeled DNA shapes using both the reference and alternative 7-mers, we then asked whether the location of a shape feature and the feature being originated from the reference and the alternate 7-mers could influence mutation rate variations. We found that shape features close to the central position are more utilized than those on the adjacent positions (Figure 3D). Along with the previously described phenomenon where second-order shape interaction features are highly utilized, we noticed that the most utilized features are interaction features incorporating both reference and alternative 7-mers instead of features that incorporate shape features from only one of these two 7-mers (Figure 3E). We also compared shape feature specific information, where it appears that different shape

features tend to have slightly different utilization of locations as well as the 7-mer (reference or alternative) (Figure S7-8).
Further analyses into exact feature coefficients show that there are multiple incidences of the same shape features, on both reference and alternate 7-mers, occurring with similar importance but opposite signs (Figure S9). For example, the most important features for mutation class CpG C-to-T are three instances of the Rise feature at a particular location, but on both reference and alternative 7-mers; in mutation class C-to-G, the first and the seventh features are both HelT * Shift, albeit the first feature on the reference 7-mer while the seventh feature on the alternative 7-mer (Figure S9). These are the cases where the direction in shape change is consistent from reference to alternative 7-mers and is also consistent with mutation rate variations.

Decision-tree based mutation rate comparisons show importance of helical twist and roll as predictors

The 7-mer context-based mutation rates could change by as much as 10-folds between the most and the least likely to mutate 7-mer contexts (Figure S10-11). Characterizing the 7-mer contexts that are the most and the least likely to mutate could help characterize mutation-promoting and mutation-averting contexts (9). A regression model may not discover this information. Thus, to explore whether and how DNA shape may cause certain sequence contexts to have very high or very low mutation rates, we built 2-layer decision tree models using first-order shape features to distinguish between the most and the least likely to mutate 7-mer contexts. In particular, we built a set of models to distinguish between the sequence contexts of the highest N% and the lowest N% mutation rates, where $N \in \{50,25,10,5\}$. We used 2-layer decision trees to ensure that the models show simple rules between particular DNA shapes and mutation rate variations and remain interpretable.
Given that the mutation rates in each class fall in a continuum, we expected that the decision tree models may not perform very well for large values of N such as 25 or 50. Indeed, decision tree performances across mutation classes have an inverse correlation with the size of N (Figure 4A), suggesting that the highest and the lowest 10% and 5% sequences hold unique shape characteristics that are linked to their extreme mutation rates. We have included the AUROC values to show performances (Figure 4A, Table 4, Table S10) as well as the ROC curves (Figure S12) and confusion matrices (Table S11-12).
To understand whether certain shape features contribute more to distinguishing mutation rates, we have plotted the coefficient values of each shape feature across all nine mutation classes for $N \in \{5,10\}$ (Figure 4B-C). We have also plotted decision tree outputs of the nine mutation classes to demonstrate how features were utilized in making decisions (Figure 4D, Figure S13). Interestingly, unlike the regression models, the most important features in the classification models corresponded to completely different shape features across the nine mutation classes (Figure 4B-C). In a few classes, such as in A-to-G, A-to-T, and C-to-T in the CpG context, the root feature (feature selected in the first layer of the decision tree) alone could make highly accurate predictions (Figure 4B-D, Figure S13). Features from all possible locations, as well as both reference and alternative 7-mers, were roughly equally represented across the root features (Figure 4D, Figure S13). Overall, this objective analysis showed DNA shape can characterize the mutation promoting and mutation averting sequence contexts with high accuracy. To our knowledge this is the first such thorough characterization of mutation promoting and averting sequence contexts; the previous analysis in this realm was based on enrichment in the highest and the lowest 1% mutation rates (9).

DNA shape features capture the effect of non-neighboring nucleotides' interactions on mutation rate variation

The above analyses established the suitability of DNA shape features in modeling genome-wide mutation rate variations in an accurate and interpretable manner. We finally asked if, in a reciprocal manner, these mutation rate models have provided us with any novel insight on DNA shape. As discussed above, Aggarwala and Voight's sequence-only model allows nucleotide interactions between any combination of up to four nucleotides within the 7-mer context of a

mutated position (9). However, for DNA shape features, even when we limited their interactions to only adjacent positions, the shape-only models with up to second-order interactions were consistently as good as the sequence-only models (Table S3-4). This difference between the two models has critical mechanistic implications on the ability of DNA shape in capturing the effects of nucleotide interactions on mutation rates. It has been suggested that DNA shape features essentially capture the interactions between di-nucleotides, i.e., second-order interactions between neighboring positions (29). This implies that a sequence-only model with up to fourth-order nucleotide interactions, but the interactions limited to only adjacent positions, would be able to capture the effects of DNA shape. Thus, limiting nucleotide interactions to adjacent positions should not lower the performance of Aggarwala and Voight's sequence-only model compared to our shape-only model. A contrary would suggest that some DNA shape features capture nucleotide interactions between non-neighboring nucleotides.

We found that limiting nucleotide interactions to only neighboring positions caused up to 23% drops in explained variance of the sequence-only models across different mutation classes (Table 5). This suggests that some DNA shape features indeed capture nucleotide interactions beyond non-neighboring nucleotides. To identify the cases where DNA shape features capture nucleotide interactions beyond di-nucleotides, we asked if combining DNA shape features with the above sequence-only model where interactions are limited to adjacent positions could raise the model's performance back to Aggarwala and Voight's sequence-only model. Incorporating first-order DNA shape features to above model indeed recovered the performance to a large extent for all classes (Table 5, Table S3-4). Furthermore, incorporating second-order shape features almost entirely recovered the performances, although we limited the shape features' interactions to only neighboring positions (Table 5, Table S3-4). The notable exceptions here were C-to-G mutations (both within and outside the CpG dinucleotide context). Overall, this analysis suggested that some DNA shape features capture nucleotide interactions beyond di-nucleotides and second-order DNA shape feature interactions, even if limited to adjacent positions, are generally sufficient to capture complex higher order nucleotide interactions.

Discussion

We have shown here the efficacy of DNA shape in explaining the genome-wide single nucleotide mutation rate variations in the human genome. Our models not only outcompeted the current best models, but also offered novel insights into the structural underpinnings of single-nucleotide mutations. We have also built simple decision tree-based models that use 1-3 shape features to categorize the sequences that are more likely to mutate, and showed the importance of HelT and Roll in predicting mutation rates. To our knowledge, this the first detailed investigation relating DNA shape and single nucleotide mutation rates. A previous paper has commented on DNA curvature having a negative correlation with mutation rate in the *URA3* gene (35). The components of DNA curvature, Roll and Tilt, are available in *DNAshapeR* (17,22), Both Roll and Tilt have played interesting roles in the linear regression models, and Roll has appeared as one of the most important shape predictors for our classification setting. This is an encouraging piece of evidence that our finding has echoed with previous literature.

Our models revealed two critical points on the role of DNA shape and the necessity of considering the exact nucleotide's identity in modeling mutation rates. First, we found that using only DNA shape and without any nucleotide sequence features, we were able to outperform the current base sequence feature-based models. Furthermore, when we interpreted the models, we found the same features to be the most important and the most utilized for all mutation classes. This demonstrates that DNA shape features encode essential information to explain mutation rate variations. On the other hand, like the previous works in this realm, we needed to limit our modeling to individual mutation classes. This necessity to maintain the information of the mutated nucleotides' identities implies that there could be additional shape features or other properties of the DNA molecule that we still lack in the models.

We have excluded non-neighboring DNA shape interactions for all our reported models. It is

worth noting that we have compared the performances of models that include or exclude non-neighboring shape interactions. Removing non-neighboring interactions in DNA shape features had little adverse effect on model performance (Table S3-4) but reduces the number of input features from 4158 to 2089 (see Methods). This implies that the neighboring shape interactions capture all necessary information for predicting mutation rate variations, thus prompting us to remove non-neighboring shape interactions.

Ideally, we would expect that going beyond 7-mers would capture more shape-specific information and improve the models further. Although we have attempted using up to 9-mer local sequence context to explain more variance in mutation rates, we found many 9-mer mutation patterns were not observed within the Phase 1 1KG and resulted in data sparsity (36). We also note that at the level of k >= 11, "nullomers", DNA k-mer sequences expected to occur in the human genome but did not, will occur (37-38). This leads us to a hypothetical maximum k of 9 for our local sequence context models should we expand our choice of k to increase the amount of variance explained.

It has been well established by literature that CpG sites significantly elevate mutation rates (8,11-13). It is also known that 70-80% of all CpG sites in adult human tissue are permanently methylated with another ~20% dynamically regulated (39). As such, we built our models under the assumption all instances of CpGs in our input 7-mers are methylated. Hence, we have used methylated shapes in *DNAshapeR* for the four shapes (Roll, ProT, HelT, MGW) that have it available (see Methods). Interestingly, HelT proved to be one of the most important features in both our regression and classification models (Figure 4B-C, Table 3, Table S7-9). In the future, we look forward to include information from the dynamic DNA methylation landscape into our mutation rate models.

With the latest update from the Genome Aggregation Database (15) and a recently published method for estimating DNA elasticity (40), we believe it is possible to build an increasingly accurate mutation rate model using these updated data. In future studies, we may also be able to use up to local 9-mer contexts for predicting mutation rates, which we failed to do in our current study due to limitations from the 1KG dataset. We also envision using our models and DNA shape features to build a whole-genome variant prioritization framework by using the noncoding mutation rate variations as a prior for regulatory region-specific mutation rate models.

From a different point of view, we believe proper interpretations of a mutation rate model may also shed light on what processes cause certain 7-mers to have drastically different rates of mutation (Figure S10-11). It is generally believed that mutations occur and retain in the population through a process of 1) a mutation-promoting event causes the mutation, 2) the DNA repair mechanism fails to repair the mutation or repairs it incorrectly, and 3) the individual with the mutation manages to reproduce and retain the mutation in the population. Although step 3 has been a recurring theme in variant prioritization and pathogenicity prediction frameworks, there is limited understanding of how the interplay of steps 1 and 2 affect mutation rates of k-mers, which we have plans to explore in future studies.

Materials and Methods

Mutation rate data

We used Aggarwala and Voight's estimates of single nucleotide mutation rates in the human genome (9). Considering the 7-mer context around each mutated nucleotide, they estimated the rates from the African population in Phase 1 of 1KG (36). Briefly, they first filtered the 1KG variants by population and excluded all annotated genes, centromeres, telomeres, repetitive regions, and regions not annotated in the accessibility mask of 1KG. Then, they calculated the count of each 7-mer, as well as the count of each unique mutation pattern along with their 7-mer context, separately for each human autosome. Finally, they computed the training and test data of mutation rates by combining the counts from all even-numbered autosomes (training data) and from all odd-numbered autosomes (testing data). See Figure 1A for an overview of the steps and Figure S1 for an overview of how we modeled the data.

DNA shape data

We used the R package *DNAshapeR* to obtain DNA shape features of the mutated nucleotides and their flanking sequences (21-24). *DNAshapeR* provides DNA shape data of 14 physio-chemical features estimated from all-atom Monte Carlo simulations (21-24). Given an input DNA sequence, *DNAshapeR* scans the sequence in a sliding window of length five and outputs the DNA shape features of each 5-mer window in the sequence. The 14 features classify into three types: inter-base pair shapes including Shift, Slide, Rise, Tilt, Roll, and HelT (helical twist); intra-base pair shapes including Shear, Stretch, Stagger, Buckle, ProT (propeller twist), and Opening; and minor groove shapes including MGW (minor groove width) and EP (electrostatic potential). The intra-base pair and minor groove features generate one output per sliding 5-mer, while the inter-base pair features generate two outputs per sliding 5-mer.

DNAshapeR also provides the estimates of four shape features (HelT, Roll, ProT, MGW) for methylated CpG dinucleotides. Since over 70% of all CpG sites are permanently methylated in the human genome (39), we have assumed all CpGs to be methylated and estimated their shape features as such. Thus, for 7-mers that include the "CG" dinucleotide, we would use the methylated shape data for HelT, Roll, ProT, and MGW, and use the normal shape data for all other features since *DNAshapeR* does not provide their estimates in the methylated state.

Categorizing mutation rates into mutation classes

The mutation class of a given single nucleotide mutation represents the identities of its reference and alternative alleles. Following Aggarwala and Voight (9), we modeled each mutation class separately. Since we have folded complementary sequences, the central reference allele could only be A or C; thus, we have characterized a total of nine classes: A-to-C, A-to-G, A-to-T, C-to-A, C-to-G, C-to-T, CpG C-to-A, CpG C-to-G, CpG C-to-T. See Figure 1D (top) for an example of mutation class generation for a 5-mer mutation.

Preparing model inputs from DNA sequence features

For the nucleotide sequence features, we have used a modified one-hot encoding scheme where we encode A as [0,0,0] and C, G, and T as [1,0,0], [0,1,0], and [0,0,1], respectively. We do not explicitly encode the mutated nucleotide (the middle position in the 7-mer), since this information does not change within a given mutation class. This results in 18 first-order features. Like Aggarwala and Voight, we allowed up to fourth-order interactions between nucleotide features. For the higher order interactions, we only considered nucleotides from different locations; same-location interactions would result in duplicated feature values. See Figure 1D (middle) for an example of how first- and second-order nucleotide features are calculated for a 5-mer mutation.

Preparing model inputs from DNA shape features of a mutated nucleotide's sequence context

We first included the measurements of all 14 DNA shape features available from *DNAshapeR* for all positions in the 7-mer context of a mutated nucleotide. This corresponds to 48 features for the reference 7-mer and the alternative 7-mer, resulting in a total of 96 first-order features for our linear regression model. Then, we computed second-order features by 1) using the PolynomialFeatures function from *sklearn* with degree=2 and include_bias=False, resulting in a total of 4752 features (4158 after removing low variance features); alternatively, 2) we designed a custom function that computes polynomial terms but only included neighboring interactions, resulting in a total of 2282 features (2089 after removing low variance features). See Figure 1B and Figure 1D (bottom) for more information.

Prior to making predictions, we have conducted min-max scaling of all shape features so that the maximum and minimum possible values for any first or second-order shape feature are (1, 0), while the geometric scale is preserved. This is done to 1) make linear regression models converge faster, and 2) ensure equal comparison of feature importance. Then, we removed features which had variances of less than 0.01 after min-max scaling as mentioned previously to ensure there are no low-variance features that may contribute minimally to model prediction.

Since adjacent sliding 5-mer windows may be intercorrelated, we have conducted a preliminary analysis of the shape features by calculating Pearson correlations between them (Figure S2).

There is no observed strong correlation between the 14 different types of shapes, thus justifying our use of conventional regression model frameworks.

Regression model building

Using the nucleotide sequence and shape features described above, we fit a separate regression model for each mutation class. For each model, we partitioned the data into training and test sets, and fit the model on the training data using 8-fold cross validation (CV). We then report the model's performance on the separately held out test set. To eliminate redundant features, we used Lasso regularization, a technique that attempts to minimize the coefficients of redundant features to zero. Thus, given the feature matrix X (each row corresponds to a mutation $k_1 \rightarrow k_2$ where k_1 and k_2 are the two 7-mers, and each column corresponds to a sequence or shape feature) and data vector Y (the mutation rates), we optimize the following objective function on the training data:

$$\frac{1}{2N}\|Y - XW\|_2^2 + \alpha\|W\|_1$$

where W is the vector of model coefficients and α represents the Lasso regularization parameter: higher values of α implies stronger regularization, i.e., optimizing the objective function will require minimizing more coefficients toward zero. We searched for the optimal value of α from the following set of plausible values using 8-fold CV. We used a stepwise decreasing array of alpha values until decreasing the value of alpha increases the 8-fold CV validation MSE of our model; the set of values for α is determined by the following:

$$\alpha = A \times 10^{-B}, A \in \{9,8,\ldots,1\}, B \in \{6,7,8,9\}$$

We used the "random" feature selection with a fixed seed to speed up our selections. We did not change the tolerance setting, but instead increased the maximum allowed iteration number so that the functions will eventually converge.

We built our training and test data by computing mutation rates from all even- and all odd-numbered chromosomes, respectively. We used 8-fold CV in fitting the model on training data, as the number of mutations in a 7-mer context is divisible by eight.

Classification model building

Similar to the regression models, we first built the models using only the training data. We used decision tree classifiers, which iteratively identifies the most informative features for a given classification task. For each individual model, we used 8-fold CV as mentioned above to search for the optimal hyperparameter values from a set of hyperparameter combinations (minimum sample per split "min_samples_split", minimum samples per leaf "min_samples_leaf", the maximum number of features used "max_features") using grid search (sklearn.GridSearchCV). These functions are available through scikit-learn (sklearn), a Python library of machine learning and statistical methods (41). We limited the depth of our decision trees to 2 to ensure easy interpretability. The best models were selected based on maximizing balanced accuracy, and model performances were evaluated based on ROC curves (false positive rate – true positive rate) and AUROC values based on predictions made on the held-out testing data. The optimization uses reductions in Gini impurity. For a binary classification, Gini impurity is calculated by:

$$I_G(p) = \sum_{i=0}^{J=1} p_i(1 - p_i) = 1 - p_0^2 - p_1^2$$

Intuitively, Gini impurity describes the imbalance in the distribution of labels in the subset. For binary classification with no class imbalance, a null model would produce a Gini impurity of 0.5, a 100% accurate model would produce a Gini impurity of 0, while a 0% accurate model would produce a Gini impurity of 1.

Model interpretation

To compare the coefficient of each DNA shape feature across the regression models of different mutation classes, we first divided the value of each feature's coefficient (W) by the range of

values of the dependent variable (Y, the mutation rates) in that mutation class:

$$W_{norm} = \frac{W}{\max(Y) - \min(Y)}$$

To compare the coefficient of each DNA shape feature across classification models, each feature is matched with a "feature importance" value, which is calculated by decreases in Gini impurities as mentioned above. For both situations, the scales of coefficient values are correlated with feature importance, thus we defined the absolute value of a feature coefficient as the importance of a feature.

For the regression models, we also considered how often a DNA shape feature was included in the optimized model. We computed two metrics. The first metric, utilization, is the count of a type of feature divided by the total number of features within the model. The second metric, enrichment, is the ratio of utilization in different importance bins of the model. By importance bins, we separated the regression model features into five or 10 equal-sized bins based on the percentile of feature importance of each feature; for example, the least important features would be assigned to bin 1, while the most important features would be assigned to bin 10. We classified features based on 1) which DNA shape they belong to, 2) their relative location in the 7-mer, and 3) their location on the reference or the alternative 7-mer. We also included the exact model descriptions of the classification models including the identities of the nodes as well as decision boundaries (Figure 4D, Supplemental figure 13).

Model implementation

We have used Python and Jupyterlab for coding tasks; the library *scikit-learn* (*sklearn*) (41) and *scipy* were used for our model building exercises, with certain tasks run on an Ubuntu terminal. The R package *DNAshapeR* was used to obtain DNA shape features.

Acknowledgments

We thank Benjamin F. Voight, Ph.D. for kindly providing the input data and guidance on data processing for this manuscript. We thank the Samee Lab members for sharing their comments during the development of this project.

References

1. di Iulio, J., Bartha, I., Wong, E.H.M., Yu, H.-C., Lavrenko, V., Yang, D., Jung, I., Hicks, M.A., Shah, N., Kirkness, E.F., et al. (2018). The human noncoding genome defined by genetic diversity. *Nature Genetics 50*, 333–337.
2. Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., and di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature Communications 10*, 5241.
3. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods 12*, 931–934.
4. Temiz, N.A., Donohue, D.E., Bacolla, A., Vasquez, K.M., Cooper, D.N., Mudunuri, U., Ivanic, J., Cer, R.Z., Yi, M., Stephens, R.M., et al. (2015). The somatic autosomal mutation matrix in cancer genomes. *Hum Genet 134*, 851–864.
5. Kuijjer, M.L., Paulson, J.N., Salzman, P., Ding, W., and Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. *British Journal of Cancer 118*, 1492–1501.
6. Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., Ridder, J. de, Herpen, C. van, Lolkema, M.P., Steeghs, N., et al. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications 11*, 728.
7. Zhang, Y., Xiao, Y., Yang, M., and Ma, J. (2020). Cancer mutational signatures representation by large-scale context embedding. *Bioinformatics 36*, i309–i316.
8. Hodgkinson, A., and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics 12*, 756–766.
9. Aggarwala, V., and Voight, B.F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics 48*, 349–355.
10. Scally, A., and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics 13*, 745–753.
11. Rideout, W.M., Coetzee, G.A., Olumi, A.F., and Jones, P.A. (1990). 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science 249*, 1288–1290.
12. Blake, R.D., Hess, S.T., and Nicholson-Tuell, J. (1992). The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J Mol Evol 34*, 189–200.
13. Duret, L. (2009). Mutation patterns in the human genome: more variable than expected. *PLOS Biology 7*, e1000028.
14. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. *Nature 526*, 68–74.
15. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 581*, 434–443.
16. Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M., and Zhurkin, V.B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *PNAS 95*, 11163–11168.
17. Bolshoy, A., McNamara, P., Harrington, R.E., and Trifonov, E.N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *PNAS 88*, 2312–2316.
18. Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and Mann, R.S. (2015). Deconvolving the recognition of DNA shape from sequence. *Cell 161*, 307–318.
19. Wang, X., Zhou, T., Wunderlich, Z., Maurano, M.T., DePace, A.H., Nuzhdin, S.V., and Rohs, R. (2018). Analysis of genetic variation indicates DNA shape involvement in purifying selection. *Mol Biol Evol 35*, 1958–1967.
20. Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res 41*, W56–W62.
21. Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (2016). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics 32*, 1211–1213.

22. Li, J., Sagendorf, J.M., Chiu, T.-P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res 45*, 12877–12887.
23. Chiu, T.-P., Rao, S., Mann, R.S., Honig, B., and Rohs, R. (2017). Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. *Nucleic Acids Res 45*, 12565–12576.
24. Rao, S., Chiu, T.-P., Kribelbauer, J.F., Mann, R.S., Bussemaker, H.J., and Rohs, R. (2018). Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding. *Epigenetics & Chromatin 11*, 6.
25. Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein–DNA recognition. *Nature 461*, 1248–1253.
26. Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-Box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports 3*, 1093–1104.
27. Harteis, S., and Schneider, S. (2014). Making the bend: DNA tertiary structure and protein-DNA interactions. *Int J Mol Sci 15*, 12335–12363.
28. Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., and Wasserman, W.W. (2016). DNA shape features improve transcription factor binding site predictions in vivo. *Cell Systems 3*, 278-286.e4.
29. Rube, H.T., Rastogi, C., Kribelbauer, J.F., and Bussemaker, H.J. (2018). A unified approach for quantifying and interpreting DNA shape readout by transcription factors. *Mol Syst Biol 14*, e7902.
30. Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol 13*, 910.
31. Samee, M.A.H., Bruneau, B.G., and Pollard, K.S. (2019). A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst 8*, 27-42.e6.
32. Afek, A., Shi, H., Rangadurai, A., Sahay, H., Senitzki, A., Xhani, S., Fang, M., Salinas, R., Mielko, Z., Pufall, M.A., et al. (2020). DNA mismatches reveal conformational penalties in protein-DNA recognition. *Nature 587*, 291–296.
33. Dantas Machado, A.C., Cooper, B.H., Lei, X., Di Felice, R., Chen, L., and Rohs, R. (2020). Landscape of DNA binding signatures of myocyte enhancer factor-2B reveals a unique interplay of base and shape readout. *Nucleic Acids Res 48*, 8529–8544.
34. Schnepf, M., von Reutern, M., Ludwig, C., Jung, C., and Gaul, U. (2020). Transcription factor binding affinities and DNA shape readout. *iScience 23*, 101694.
35. Duan, C., Huan, Q., Chen, X., Wu, S., Carey, L.B., He, X., and Qian, W. (2018). Reduced intrinsic DNA curvature leads to increased mutation rate. *Genome Biology 19*, 132.
36. McVean, G.A., Altshuler (Co-Chair), D.M., Durbin (Co-Chair), R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature 491*, 56–65.
37. Acquisti, C., Poste, G., Curtiss, D., and Kumar, S. (2007). Nullomers: really a matter of natural selection? *PLoS One 2*, e1022.
38. Vergni, D., and Santoni, D. (2016). Nullomers and high order nullomers in genomic sequences. *PLoS One 11*, e0164540.
39. Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature 500*, 477–481.
40. Velasco-Berrelleza, V., Burman, M., Shepherd, J.W., Leake, M.C., Golestanian, R., and Noy, A. (2020). SerraNA: a program to determine nucleic acids elasticity from simulation data. *Phys. Chem. Chem. Phys. 22*, 19254–19266.
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Figures and Tables

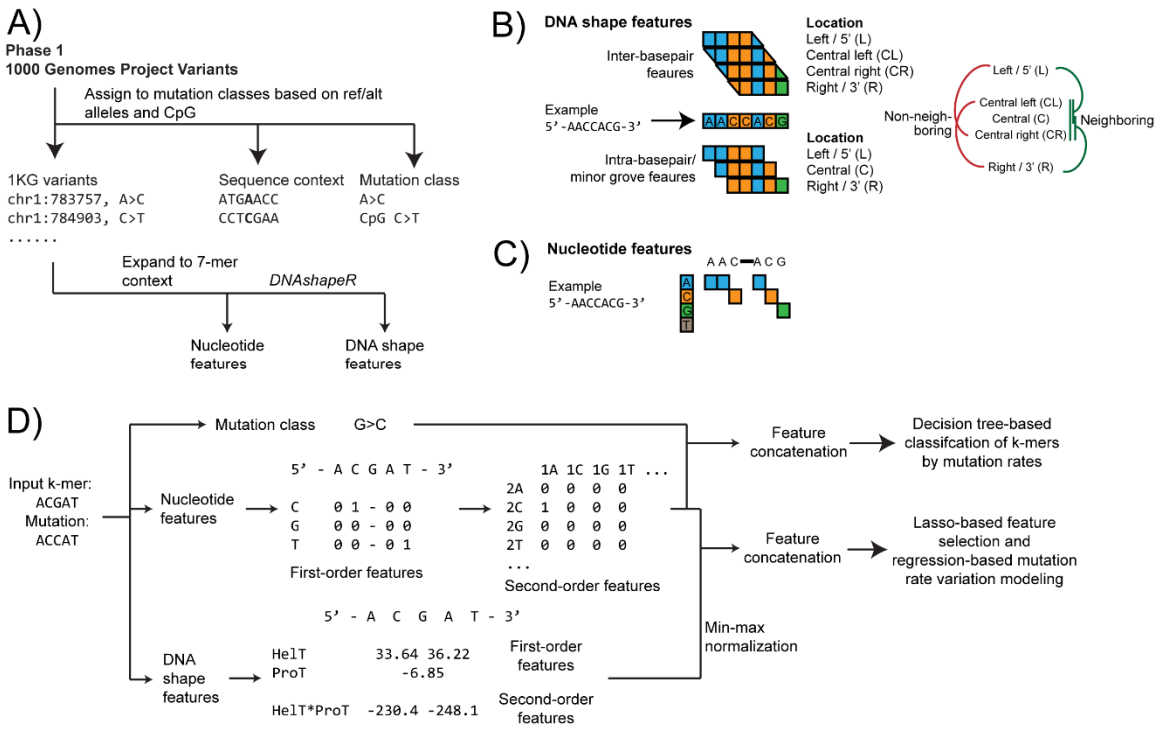

Figure 1. Overview of our study. A) Our pipeline for the generation of predictive features from mutation data. Mutation data were obtained from phase 1 of the 1000 Genomes Project (1KG) as described in the Methods section. Mutations were then extended into 7-mers based on the identities of the three flanking nucleotides; within each 7-mer context, mutation rate will be calculated by dividing the total number of mutations with that particular 7-mer context with the total number of the corresponding 7-mer context observed in the human noncoding genome. B-C) Illustrations of the generation of first-order B) DNA shape and C) nucleotide features, as well as the criteria for neighboring versus non-neighboring interactions for DNA shape features. Nucleotide and shape features were generated based on the various 7-mer contexts, with the nucleotide features generated by one-hot encoding, and the shape features generated by the *DNAshapeR* package. See the Methods section for details. D) Example of our machine learning pipeline given an input 5-mer of ACGAT, note that we listed a 5-mer example here for simplicity, but we only used 7-mers in the study. For nucleotide features, we will generate three corresponding first-order features corresponding to C, G, and T for each location on the k-mer; we have excluded the central location as the central location will become a uniform feature due to us modeling each mutation class separately. For shape features, the *DNAshape* method generates 14 unique types of shapes, for which we show two examples (helical twist, HelT; propeller twist, ProT) here. Since HelT is an inter-basepair feature and ProT is an intra-basepair feature, there are two HelT and one ProT values for each 5-mer context, see the methods section for more information. We also show one example of interaction feature HelT × ProT, for which we will assume to be inter-basepair if any one of the two interacting features is inter-basepair.

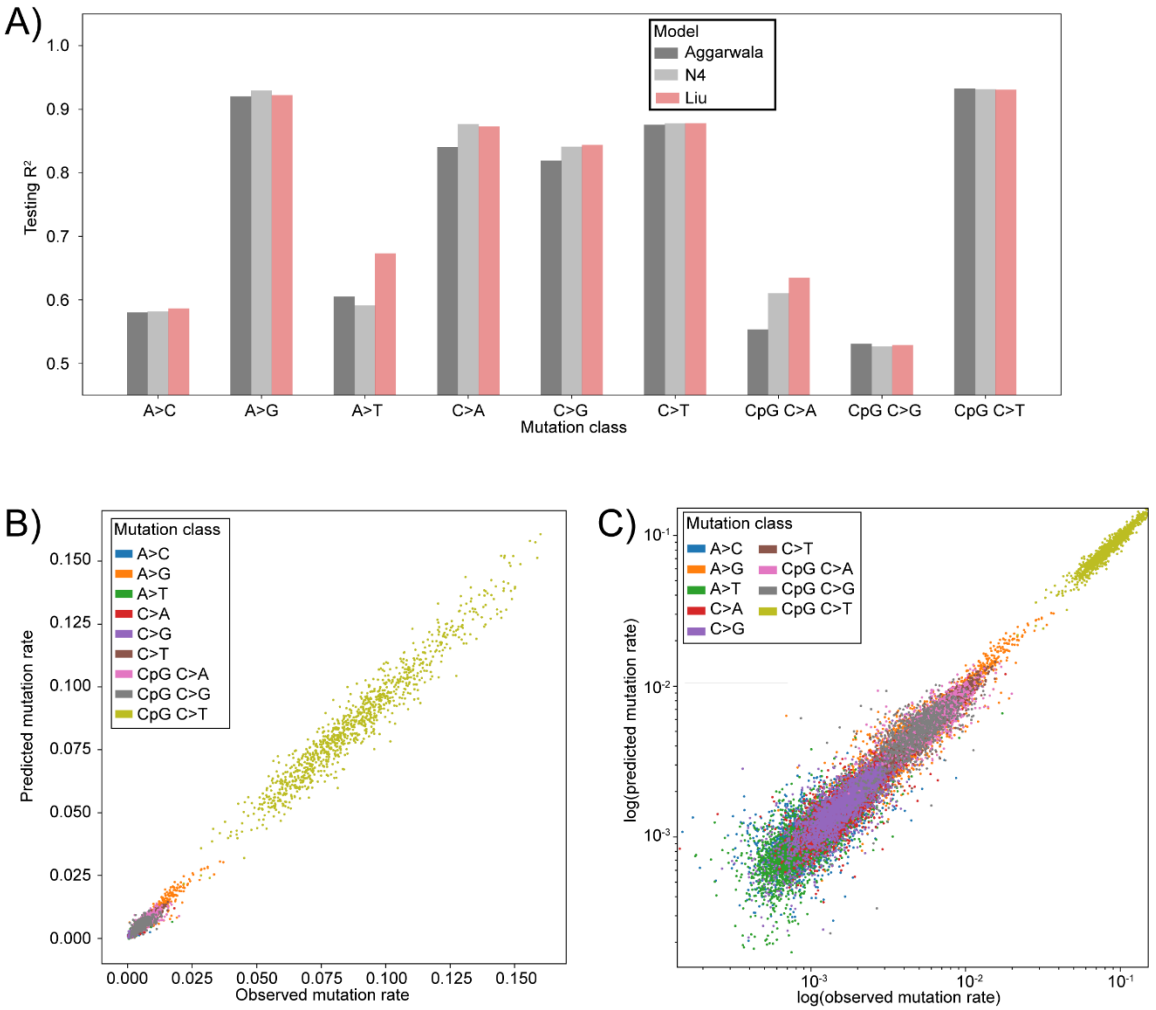

Figure 2. Our best performing model outperforms the Aggarwala and Voight model. A) Histogram comparison of R^2 values of the Aggarwala and Voight model (Aggarwala) (9), a Lasso feature selection-based fourth-order nucleotide features model (N4), and our best performing model (Liu) on the independent testing data. B-C) Scatterplots showing comparison of predicted mutation rates and observed mutation rates from the 1KG data, in B) linear scale and C) logarithmic scale. The x-axes show observed mutation rates, while the y-axes show model-predicted mutation rates.

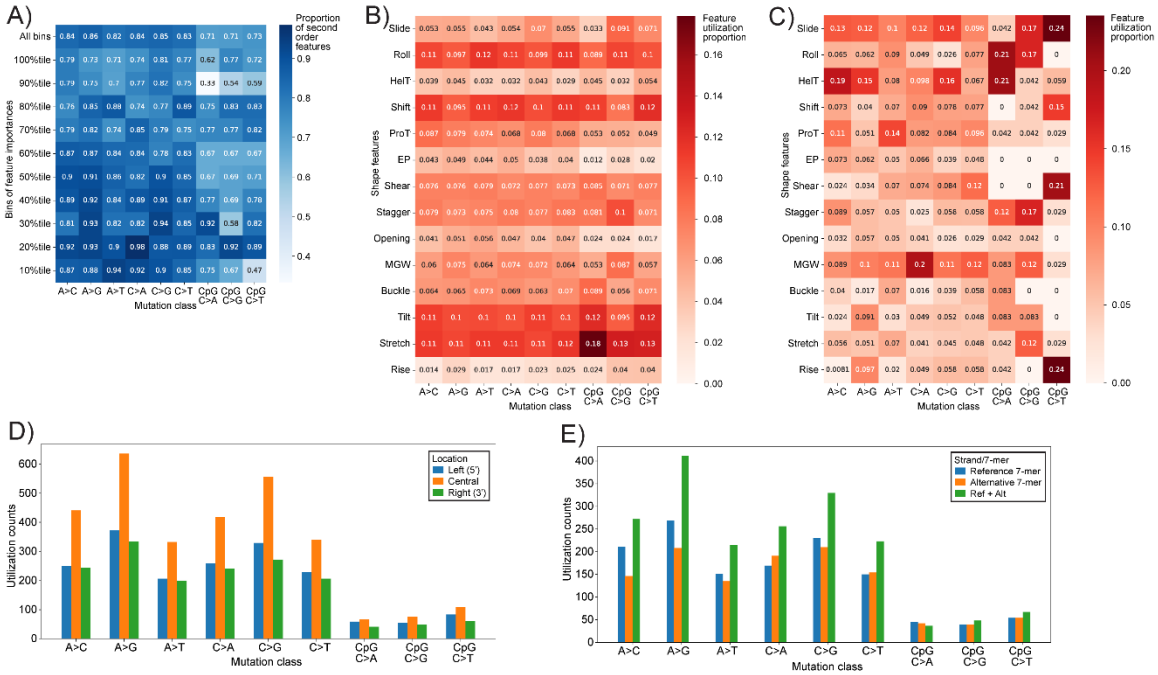

Figure 3. Coefficient analyses of DNA shape-based regression model suggest feature utilization trends. A) Ratio of second-order shape features among all features used in each of the nine mutation classes as well as in 10% bin increments of importance. The first row describes ratios in the entire models, while the subsequent rows describe ratios in each bin of importance. See Methods for how the bins are defined. B-C) Relative utilization of 14 shape features across models in all nine mutation classes using B) all features or C) top 10% features with the highest importance values. D-E) Bar charts of feature utilization based on D) feature location or E) feature strand (reference or alternative 7-mer) across models in all nine mutation classes. Feature utilization scores were calculated by counting the number of times a feature is used, and relative feature utilization scores were calculated by dividing the feature utilization scores by the total number of features in a particular model or a particular subset of features.

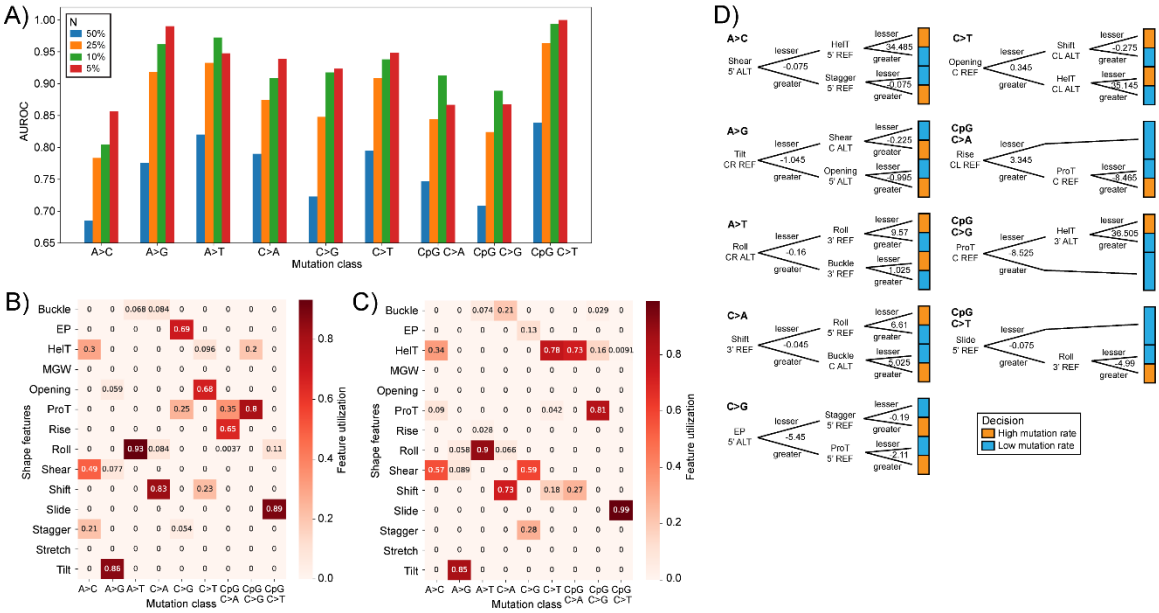

Figure 4. Decision tree-based classification models show high accuracy and interpretability for classifying k-mers with the highest or lowest mutation rates. A) Bar chart of area under receiver-operating curve (AUROC) values for classification models for all nine mutation classes and all four values of N. The ROC curves were constructed using false positive rates and true positive rates. B-C) Heatmaps of feature importance, as calculated by the total reduction in Gini coefficient by a particular feature, across all shape features (y-axis), all nine mutation classes (x-axis), and values of N equal to B) 5% and C) 10%. See Methods section for details on feature importance calculations. D) Detailed decision tree architectures for N=5 in all nine mutation classes. Each decision tree feature is labeled by the type of shape feature, location, and strand; possible locations include 5', central-left (CL) and central-right (CL) for inter-basepair features, central (C) for intra-basepair features, and 3'; possible strands include reference (REF) and alternative (ALT). Note that for all subplots, N represents the percentage of k-mers with the highest or lowest mutation rates and were used as inputs to the classification models, see the methods section for more information.

Table 1. Model performance of our best-performing model in the testing data. The “maximum achievable R^2” column values were calculated by directly comparing testing data with the training data and represent “maximum” model performances. The “Aggarwala” column describes performances of the current state of the art model. Performance percentage changes were calculated by $(R^2 - R_0^2)/R_0^2$ comparing our model to the state of the art model. The “parameters” column describes the identities of the input predictors for each mutation class. Regarding the abbreviations in the parameters column, numbers represent degrees of polynomial transformations, "sh" and "sc" represent shape and nucleotide (sequence context) features, and the "neibr" suffix represents that only neighboring interactions were included (see Methods). The mean and median values were computed based on R^2 values of the nine individual models and were not computed for performance changes to avoid ambiguity.

Mutation class	**Maximum achievable R^2**	**Aggarwala**	**Liu**	**Performance change**	**Parameters**
A>C	0.666	0.580	0.586	1.02%	sh1_sc4
A>G	0.961	0.920	0.922	0.24%	sh2neibr
A>T	0.750	0.605	0.673	**11.18%**	sh2neibr
C>A	0.905	0.840	0.873	3.88%	sh2neibr_sc1
C>G	0.867	0.819	0.844	3.03%	sh2neibr_sc1
C>T	0.864	0.875	0.878	0.29%	sh1_sc4
CpG_C>A	0.534	0.553	0.635	**14.68%**	sh2neibr
CpG_C>G	0.417	0.531	0.529	-0.34%	sh1_sc3
CpG_C>T	0.934	0.932	0.930	-0.20%	sh1_sc3
Mean	0.766	0.740	0.763		
Median	0.864	0.819	0.844		

Table 2. Model performance of our second-order shape model in the testing data. The "maximum achievable R^2" column values were calculated by directly comparing testing data with the training data and represent "maximum" model performances. The "Aggarwala" column describes performances of the current state of the art model. The "sh2neibr" column describes performances of our second-order DNA shape model. Performance percentage changes were calculated by $(R^2 - R_0^2)/R_0^2$ comparing our model to the state of the art model. The mean and median values were computed based on R^2 values of the nine individual models and were not computed for performance changes to avoid ambiguity.

Mutation class	**Maximum achievable R^2**	**Aggarwala**	**sh2neibr**	**Performance change**
A>C	0.666	0.580	0.579	-0.26%
A>G	0.961	0.920	0.922	0.24%
A>T	0.750	0.605	0.673	**11.18%**
C>A	0.905	0.840	0.868	3.31%
C>G	0.867	0.819	0.834	1.79%
C>T	0.864	0.875	0.870	-0.63%
CpG_C>A	0.534	0.553	0.635	**14.68%**
CpG_C>G	0.417	0.531	0.517	-2.46%
CpG_C>T	0.934	0.932	0.925	-0.78%
Mean	0.766	0.740	0.758	
Median	0.864	0.819	0.834	

Table 3. Fisher's exact tests with Benjamin-Hochberg false discovery rate (FDR-BH)-adjusted p-values of feature utilization between top 10% features and all features. The top 10% features were defined based on feature importance, or absolute coefficient values.

Shape features	A>C	A>G	A>T	C>A	C>G	C>T	CpG C>A	CpG C>G	CpG C>T
Slide	0.152	0.106	0.468	0.297	0.166	0.839	1	1	0.220
Roll	1	1	1	1	1	1	1	1	1
HelT	**0.000349**	**0.00132**	0.498	0.114	**0.00132**	0.857	0.498	1	1
Shift	1	1	1	1	1	1	1	1	1
ProT	1	1	0.498	1	1	1	1	1	1
EP	1	1	1	1	1	1	1	1	1
Shear	1	1	1	1	1	0.857	1	1	0.650
Stagger	1	1	1	1	1	1	1	1	1
Opening	1	1	1	1	1	1	1	1	1
MGW	1	0.971	0.860	0.021	0.857	0.583	1	1	1
Buckle	1	1	1	1	1	1	1	1	1
Tilt	1	1	1	1	1	1	1	1	1
Stretch	1	1	1	1	1	1	1	1	1
Rise	1	**0.0384**	1	0.498	0.545	0.766	1	1	**0.0384**

Table 4. Area under receiver operating curve (AUROC) values for decision tree models with N in {5, 10, 25, 50} and all mutation classes. The horizontal axis denotes the values of N, which corresponds to the proportion of k-mers with the highest and lowest N percent of mutation rates subsequently used for the classification task. See supplemental tables for performance metrics on the training data or the confusion matrices.

Mutation class	50%_test	25%_test	10%_test	5%_test
A>C	0.685	0.783	0.804	0.857
A>G	0.776	0.918	0.962	0.990
A>T	0.820	0.933	0.972	0.948
C>A	0.790	0.874	0.908	0.939
C>G	0.723	0.848	0.917	0.923
C>T	0.795	0.908	0.938	0.949
CpG_C>A	0.747	0.844	0.913	0.866
CpG_C>G	0.708	0.824	0.889	0.868
CpG_C>T	0.839	0.963	0.993	1.000

Table 5. Model performance between models that include or exclude non-neighboring nucleotide interactions, rescued by shape, in the testing data. We measured reductions in R^2 when we removed fourth-order nucleotide interaction terms, as well as how this reduction is partially mitigated by including DNA shape features. For comprehensive R^2 values, please see the supplemental tables. Performance percentage changes were calculated by $(R^2 - R_0^2)/R_0^2$. Regarding the abbreviations in the horizontal axis, numbers represent degrees of polynomial transformations, "sh" and "sc" represent shape and nucleotide (sequence context) features, and the "neibr" suffix represents that only neighboring interactions were included (see Methods). The mean and median values were computed based on R^2 values of the nine individual models and were not computed for performance changes to avoid ambiguity.

Mutation class	**sc4**	**sc4neibr**	**Performance change**	**sh1_sc4**	**sh1_sc4neibr**	**Performance change**	**sh2neibr_sc4**	**sh2neibr_sc4neibr**	**Performance change**
A>C	0.582	0.446	-23.23%	0.586	0.502	-14.35%	0.592	0.555	-6.25%
A>G	0.929	0.751	-19.21%	0.930	0.854	-8.16%	0.931	0.914	-1.85%
A>T	0.591	0.573	-3.12%	0.610	0.610	-0.01%	0.631	0.590	-6.59%
C>A	0.877	0.765	-12.71%	0.878	0.801	-8.75%	0.882	0.850	-3.70%
C>G	0.841	0.718	-14.64%	0.842	0.754	-10.54%	0.844	0.787	-6.71%
C>T	0.878	0.819	-6.71%	0.878	0.825	-5.98%	0.882	0.870	-1.34%
CpG_C>A	0.610	0.595	-2.55%	0.609	0.616	1.17%	0.611	0.606	-0.72%
CpG_C>G	0.527	0.497	-5.61%	0.526	0.495	-5.99%	0.524	0.490	-6.49%
CpG_C>T	0.931	0.912	-2.07%	0.932	0.917	-1.61%	0.934	0.922	-1.32%
Mean	0.752	0.675		0.755	0.708		0.759	0.732	
Median	0.841	0.718		0.842	0.754		0.844	0.787	