HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2024 February 01.

Published in final edited form as:

Cell Rep. 2024 January 23; 43(1): 113655. doi:10.1016/j.celrep.2023.113655.

DNA polymerase ε and δ variants drive mutagenesis in polypurine tracts in human tumors

Daria Ostroverkhova[1,5], **Kathrin Tyryshkin**[1,5], **Annette K. Beach**[2], **Elizabeth A. Moore**[2], **Yosef Masoudi-Sobhanzadeh**[1], **Stephanie R. Barbari**[2,6], **Igor B. Rogozin**[3], **Konstantin V. Shaitan**[4], **Anna R. Panchenko**[1,*], **Polina V. Shcherbakova**[2,7,*]

[1]Department of Pathology and Molecular Medicine, School of Medicine, Queen's University, Kingston, ON, Canada

[2]Eppley Institute for Research in Cancer and Allied Diseases, Fred & Pamela Buffett Cancer Center, University of Nebraska Medical Center, Omaha, NE, USA

[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

[4]Lomonosov Moscow State University, Moscow, Russia

[5]These authors contributed equally

[6]Present address: GT Molecular, 221 E Lincoln Avenue, Fort Collins, CO 80524, USA

[7]Lead contact

SUMMARY

Alterations in the exonuclease domain of DNA polymerase ε cause ultramutated cancers. These cancers accumulate AGA>ATA transversions; however, their genomic features beyond the trinucleotide motifs are obscure. We analyze the extended DNA context of ultramutation using whole-exome sequencing data from 524 endometrial and 395 colorectal tumors. We find that G>T transversions in POLE-mutant tumors predominantly affect sequences containing at least six consecutive purines, with a striking preference for certain positions within polypurine tracts. Using this signature, we develop a machine-learning classifier to identify tumors with hitherto unknown *POLE* drivers and validate two drivers, *POLE-E978G* and *POLE-S461L*, by functional assays in yeast. Unlike other pathogenic variants, the E978G substitution affects the polymerase domain of Pol ε. We further show that tumors with *POLD1* drivers share the extended signature of

*Correspondence: anna.panchenko@queensu.ca (A.R.P.), pshcherb@unmc.edu (P.V.S.).

AUTHOR CONTRIBUTIONS

Conceptualization, funding acquisition, supervision, and project administration, P.V.S. and A.R.P.; data curation and software, D.O. and K.T.; formal analysis, D.O., K.T., A.K.B., Y.M.-S., S.R.B., and P.V.S.; investigation, D.O., K.T., A.K.B., E.A.M., S.R.B., and A.R.P.; methodology, D.O., K.T., I.B.R., and K.V.S.; resources, D.O., A.K.B., and S.R.B.; validation, D.O., K.T., A.K.B., E.A.M., and S.R.B.; visualization, D.O., K.T., and P.V.S.; writing, D.O., K.T., A.R.P., and P.V.S.

SUPPLEMENTAL INFORMATION

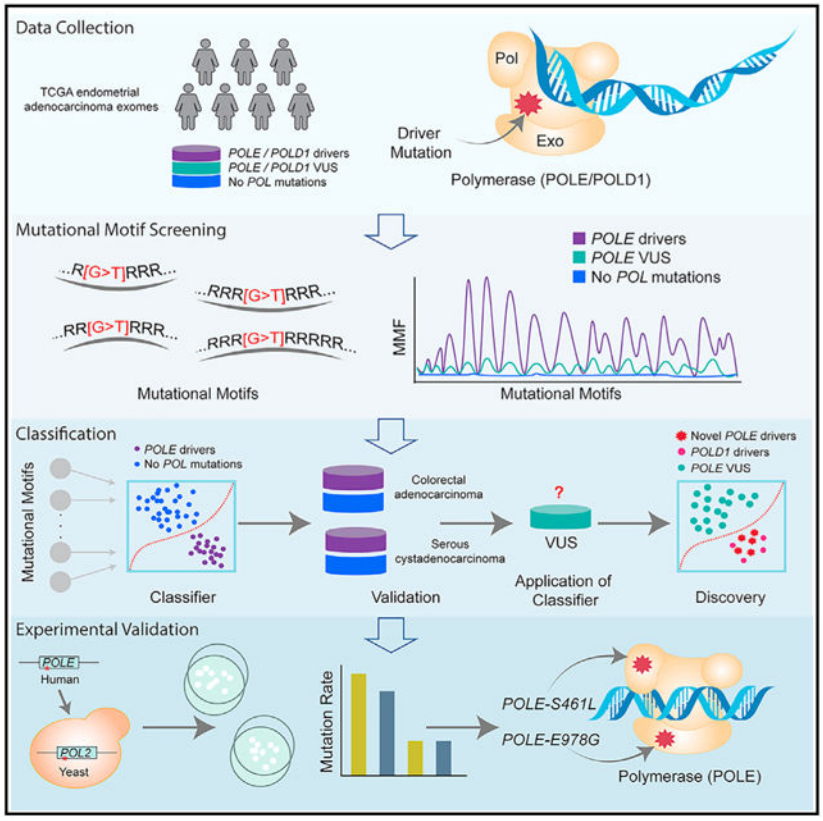
Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113655.

DECLARATION OF INTERESTS

The authors declare no competing interests.

POLE ultra-mutation. These findings expand the understanding of ultramutation mechanisms and highlight peculiar mutagenic properties of polypurine tracts in the human genome.

Graphical Abstract



In brief

Using a computational approach, Ostroverkhova et al. show that *POLE*- and *POLD1*-mutant tumors accumulate G>T mutations at defined positions of polypurine tracts, creating a characteristic quantifiable pattern. A classifier based on this signature facilitates the discovery of *POLE* driver alleles, including a driver affecting the DNA polymerase domain of the protein.

INTRODUCTION

Cancer results from the accumulation of mutations in critical genomic regions.[1,2] The sources of mutations include the infidelity of DNA replication machinery, exposure to endogenous or exogenous DNA-damaging agents, defective DNA repair, and enzymatic modification of DNA.[1] Each mutational process may generate different types of mutations in specific DNA sequence contexts.[3] Imbalance between the burden of DNA damage or replication errors and the restorative capacity of DNA repair pathways determines the ultimate mutational load of cancer genomes.[4] In the past decade, it became apparent that tumors with the highest mutational load carry defects in the replication error avoidance pathways. DNA replication fidelity in all organisms relies on accurate nucleotide selection

by replicative DNA polymerases, exonucleolytic proofreading activity of the polymerases, and post-replicative DNA mismatch repair (MMR).[5] In humans, inherited MMR defects cause cancer predisposition in Lynch syndrome, and MMR-deficient sporadic tumors show hypermutation and microsatellite instability (MSI).[6,7] However, the highest mutational burdens (>100 mutations per Mb) occur in tumors with heterozygous mutations in the *POLE* and *POLD1* genes encoding the catalytic subunits of the replicative DNA polymerases Pol ε and Pol δ.[8–11]

Pol ε and Pol δ are responsible for the bulk of chromosomal DNA replication and proofreading of errors on leading and lagging strands.[12] They are B-family DNA polymerases, with the polymerase domains catalyzing processive synthesis and the exonuclease domains required to proofread errors. Both germline and somatic alterations of these enzymes have been linked to tumorigenesis. Germline *POLE* and *POLD1* mutations cause predisposition to colorectal cancer and may also increase the risk of other malignancies.[13–16] Somatic *POLE* mutations occur in most cancer types but are particularly common in ultramutated colorectal and endometrial tumors.[17–20] The connection between somatic *POLD1* variants and ultramutation is less understood. *POLD1* variants are less frequent in sporadic tumors and always coincide with MMR deficiency, which alone is sufficient to elevate the mutation rate. Many *POLE* and *POLD1* variants inevitably arise as passenger mutations in tumors with high mutation load. However, experimental evidence has identified 14 *POLE* and at least six somatic *POLD1* variants as *bona fide* drivers of hypermutability.[21–34] These pathogenic variants impact the exonuclease domain of Pol ε and both exonuclease and DNA polymerase domains of Pol δ. Patients with *POLE*-driven ultramutation have excellent prognosis and respond well to immunotherapy.[35,36] Thus, it is important to understand the molecular features of these tumors.

Previous analyses showed that *POLE*-mutant tumors preferentially accumulate G>T transversions in AGA trinucleotide context and G>A transitions in CGA context.[1,13,33,37] Multiple studies have since used this signature to quantify Pol ε errors, detect known and new pathogenic *POLE* variants, explore the evolution of mutagenic processes in tumors, infer the origin of mutations in tumor-suppressor genes, and assess DNA replication mechanisms.[17,33,34,38–49] Others have used the trinucleotide *POLE* signature for clinical classification of tumors.[50] In contrast, the features of mutational sites beyond the trinucleotide context escaped proper attention. Replicative DNA polymerases contact at least ten nucleotides of duplex DNA upstream of the 3′ terminus.[51,52] Accordingly, the extended DNA neighborhood around the mutated site is an important determinant of the individual site's mutability.[3,4,53,54] Previous pilot analysis of a 26-gene panel in one *POLE*-mutant and one *POLD1*-mutant cell line detected a strong preference for G>T transversions in polypurine tracts of varying length and composition.[31] We could also recapitulate the high frequency of G>T transversions in polypurine tracts by ectopically expressing a mutator *POLD1* allele in cells lacking DNA polymerase mutations.[31]

Inspired by these initial findings, here we used the whole-exome sequencing data for 524 endometrial and 395 colorectal tumors reported by The Cancer Genome Atlas (TCGA) network to comprehensively test a hypothesis that *POLE*-driven G>T transversions preferentially occur in extended polypurine tracts and to determine specific features of the

most mutable tracts. We deduced a genomic signature of *POLE* driver variants that considers the extended DNA sequence context of mutations. Using this signature, we developed a machine-learning classifier to identify tumors with hitherto unknown *POLE* drivers and validated two drivers, *POLE-E978G* and *POLE-S461L*, by functional assays in yeast. Unlike previously known driver alleles, *POLE-E978G* impacts the DNA polymerase rather than the exonuclease domain of Pol ε. Analysis of the structural effects of E978G substitution suggested the existence of a distinct class of ultramutation drivers mapped to the polymerase domain. We also found that the extended genomic signature of *POLE* ultramutation is shared by tumors with *POLD1* driver alleles. Finally, our study provided an unexpected insight into the peculiar mutagenic properties of polypurine tracts in the human genome.

RESULTS

POLE driver status is associated with increased frequency of G>T transversions in polypurine tracts

The endometrial adenocarcinoma (EAC) dataset from the Uterine Corpus Endometrial Carcinoma (UCEC) cohort used in the initial analysis consisted of 391 tumors (Tables S1 and S2; Figure S1A). These included 44 tumors with *POLE* driver mutations, five tumors with *POLD1* driver mutations, one tumor with both *POLE* and *POLD1* drivers, 49 tumors with *POLE* and/or *POLD1* variants of unknown significance (VUSs), and 292 tumors with no *POLE* or *POLD1* mutations. The majority (82%) of samples with *POLE* driver mutations were microsatellite stable (MSS), while all six samples with *POLD1* drivers had MSI. Because of the small number of tumors with *POLD1* driver alleles, we first focused on understanding the mutational footprint of *POLE*-driven ultramutation.

G>T transversions in AGA context and, to a lesser degree, G>A mutations in CGA context dominate the mutational spectra of tumors with *POLE* drivers.[1,13,33,37] We used mutational motif frequency (MMF) analysis to test the hypothesis that G>T transversions in these tumors primarily occur in longer-than-trinucleotide polypurine tracts. A mutational motif is defined here as a specific mutation type at a specific position within a specific nucleotide sequence. We examined 104 motifs comprising polypurine tracts of different lengths (2–14 nt) with all possible positions of the G>T mutation site within the tract ($_{Y}(R)_{n}[G>T](R)_{m}{}_{Y}$; see STAR Methods and Table S3). Hereafter, we refer to these motifs as “polypurine motifs.” We found that a high MMF for all 104 motifs combined was significantly associated with the presence of *POLE* driver mutations (Figure 1B). Comparative analysis with respect to motif lengths and positions of the mutated guanine showed a peculiar pattern for samples with *POLE* drivers and, to some extent, for samples with *POLE* VUSs (Figure 1A). Overall, MMF values were modest for shorter tracts (<6 purines) and particularly low for 2- and 3-nt-long tracts. However, MMF increased notably once the tract length reached six purines.

Furthermore, mutations in tumors with *POLE* drivers showed a striking preference for certain positions within the polypurine tracts (Figure 1A). Distance to the beginning and the end of the tract appeared to define MMF of individual sites. High MMF values were associated with the locations of mutated guanine 2–3 purines and, to a progressively lesser degree, 7–8 purines, and 10 purines away from the end of the tract (Figure 1C). At the same

time, the locations 3–4 purines and, to a lesser degree, 9–10 purines from the beginning of the tract resulted in high MMF values (Figure 1C). Accordingly, sites within 6- and 7-nt-long tracts located simultaneously 2–3 purines from the end and 3–4 purines from the start were the most mutated among the 104 tested sites (Figure 1A). Tracts of R10 purines showed two separate MMF spikes: a larger spike 2–3 purines from the end and a smaller spike 3–4 purines from the beginning. The existence of two separate context factors with apparently additive effects might reflect differences in Pol ε synthesis when the tracts are presented as polypurine vs. polypyrimidine sequences in the leading strand template. The 14-nt-long tracts where the left and right MMF spikes were the most distant showed a potential third spike in the middle. However, the scarcity of these long tracts in the genome resulted in a large standard error and precluded definitive conclusions. As expected, MMF values were the lowest for motifs with no purines before or after the G>T mutation. These patterns persisted if we normalized MMF by the mutational burden (Figure S2) and were similar in nucleosomal and nucleosome-free DNA (Figure S3). For comparison, we performed the same analysis using a set of 104 motifs with G>T mutations in polypyrimidine context ($_{R}(Y)_{n}[G>T](Y)_{m}{}_{R}$; Table S3). We observed no statistically significant association of MMF with driver mutations except for cases with G>T at the end of the polypyrimidine run, i.e., next to a purine (Figure S4), further supporting the importance of a purine environment.

The $_{Y}(R)_{n}[G>T](R)_{m}{}_{Y}$ motifs comprise a complex mixture, with all possible combinations of adenines and guanines in the $(R)_{n}$ and $(R)_{m}$ parts. The analysis of individual sequences is computationally challenging due to the large number of purine combinations in our motif set (n = 196,610) and would also require much larger tumor mutation datasets to evaluate statistical significance. However, the quantification of $_{Y}(A)_{n}[G>T](A)_{m}{}_{Y}$ and $_{Y}(G)_{n}[G>T](G)_{m}{}_{Y}$ motifs showed that tracts with the polyA context may be primarily responsible for the observed mutagenesis pattern. MMF reaches maximum values at the same positions of mutated G for $_{Y}(A)_{n}[G>T](A)_{m}{}_{Y}$ and $_{Y}(R)_{n}[G>T](R)_{m}{}_{Y}$ motifs, and the increase in mutability at R6-nt tract lengths is also similar (Figures S5A and 1A). Moreover, maximum MMF values for $_{Y}(A)_{n}[G>T](A)_{m}{}_{Y}$ motifs were almost an order of magnitude higher than for $_{Y}(R)_{n}[G>T](R)_{m}{}_{Y}$, indicating that the latter group included a substantial number of less mutable sites. In contrast, $_{Y}(G)_{n}[G>T](G)_{m}{}_{Y}$ motifs showed low MMF values regardless of the mutated G position (Figure S5B). Thus, polyguanine tracts contribute minimally, if at all, to the observed mutagenesis pattern. We also quantified G>T transversions in dinucleotide $(AG)_{n}$ repeats of different lengths (Table S3). The mutability of $(AG)_{n}$ repeats was similar to that of $_{Y}(R)_{n}[G>T](R)_{m}{}_{Y}$ motifs: increased MMF in R6-nt tracts and spikes at positions with two purines after the mutation and three purines before the mutation. The mean MMF values were also similar for $(AG)_{n}$-repeat motifs and $_{Y}(R)_{n}[G>T](R)_{m}{}_{Y}$ motifs (Figure S6 and 1A). We conclude that $(AG)_{n}$ repeats contribute to the pattern of polypurine tract mutability in *POLE* tumors along with other polypurine sequences. However, they are not major contributors like $(A)_{n}[G>T](A)_{m}$ tracts, which show significantly higher MMF. Since previous studies linked adenine context to A>C transversions in *POLE*-mutant tumors, we also assessed the impact of polyA tract length and position within the tract on the frequency of A>C mutations. Tumors with *POLE* drivers showed a moderate increase in MMF for some $_{Y}(A)_{n}[A>C](A)_{m}{}_{Y}$ motifs as compared to tumors with *POLE* VUSs or no

POLE or *POLD1* mutations. However, maximum MMF values for A>C mutations were approximately 20-fold lower than for G>T in polyA context and showed no dependence on the length of polyA tract (Figures S7 and S5A). The positions with the highest MMF in the tracts were different for A>C and G>T. Thus, the propensity of tumors with *POLE* drivers to accumulate A>C mutations in polyA context is modest as compared to G>T and likely involves a different mechanism.

We next investigated whether tumors with different *POLE* driver alleles share the pattern of polypurine tract mutability. One tumor in the UCEC EAC dataset was excluded from this analysis because it harbored two *POLE* driver mutations, P286R and L424I. Among the remaining *POLE* tumors, 18 and 11 carried the most common driver mutations, P286R and V411L, respectively. Fourteen tumors carried less prevalent variants (S459F, P436R, A456P, L424I, L424V, S297F, M295R, M444K, or D368Y; Table S2). When grouped together, tumors with less common alleles showed a mutation pattern similar to those of P286R and V411L tumors (Figure S8A), although absolute MMF values were different (P286R > V411L > other variants; Figures S8A and S8C). The similarity in the mutation patterns became even more apparent after normalizing MMF by the number of G>T transversions in the tumors (Figure S8B), but quantitative differences persisted (Figure S8D). Separating individual rare alleles revealed significant variation in their propensity to generate mutations at hotspots 2–3 nt from the end of the polypurine tracts (Figure 1D; Table S4). Curiously, alleles that are considered more pathogenic showed the highest preference for these hotspot contexts. While there is insufficient information to definitively establish the degree of pathogenicity, indirect evidence suggests that the five alleles with the highest MMF are all strongly pathogenic. P286R and V411L variants are both highly recurrent in sporadic tumors, followed by S459F.[22] Germline V411L, A456P, and M444K variants are associated with pediatric rather than adult-onset cancers typical of L424V carriers.[44,55,56] Lastly, in contrast to the alleles in the low-MMF group, P286R, V411L, S459F, M444K, and A456P extremely rarely or never occur in combination with MMR defects, suggesting a stronger mutator effect incompatible with MMR inactivation due to a cata-strophic increase in the mutation rate. Taken together, these observations suggest that the preference for errors at positions 2–3 nt away from the end of polypurine tracts may be intimately linked to the pathogenicity of Pol ε variants. Tumors with presumed less-pathogenic variants (S297F, L424I, P436R, D368Y) showed a lower contribution of polypurine tract hotspots to the total mutation burden than tumors carrying presumed strong drivers, particularly P286R, S459F, and M444K (Figure S9A). Of note, about half of the tumors with the highly recurrent V411L variant showed a similar low contribution of polypurine hotspots to the overall mutation load (Figure S9A). For tumors with the same *POLE* driver (P286R or V411L), the preference for the hotspot contexts, quantified as MMF for the hotspot motifs normalized by the total number of G>T transversions, decreased with increasing mutation burden (Figure S9B), indicating that mutagenic processes involving other DNA sequence contexts are more prominent at high ultramutation levels.

MMR deficiency does not affect the association of *POLE* driver status with G>T transversions in polypurine tracts

Many tumors in the UCEC EAC dataset are MMR deficient (Table S2), which could impact mutagenesis in polypurine tracts regardless of the *POLE* status. To assess the individual contributions of MMR deficiency and *POLE* mutations, we used MMF for motifs with 2–3 purines after the G>T transversion as a dependent variable and four categorical independent variables defined according to *POLE* and MSI status of tumors: "*POLE* drivers, MSS," "*POLE* drivers, MSI," "no *POLE*/*POLD1* mutations, MSS," and "no *POLE*/*POLD1*mutations, MSI." The Kruskal-Wallis test showed a statistically significant difference in MMF between the four categories ($p << 0.01$). *POLE* wild-type MSI tumors showed only a minor, although statistically significant, increase in MMF as compared to *POLE* wild-type MSS tumors (effect size = 0.141; $p << 0.01$; Figure 1E; Table S5). In contrast, MMF was greatly elevated in *POLE*-mutant MSS tumors compared to *POLE* wild-type tumors. Notably, tumors carrying simultaneous polymerase and MMR defects showed no further increase in MMF. Instead, there was a modest decrease compared to *POLE*-mutant MSS tumors ($p < 0.05$; Figure 1E; Table S5). Although this decrease may appear surprising, we believe it is due to the different nature of *POLE* mutations in MSS and MSI tumors. Most MSS tumors carry highly pathogenic P286R and V411L variants likely incompatible with MMR deficiency. With one exception, MSI tumors carry less common *POLE* alleles (Table S2) that are weaker drivers of polypurine tract mutability (Figure 1D). Accordingly, G>T transversions in polypurine tracts contribute less to the overall mutation burden of tumors with MSI compared to MSS tumors (Figures S9C and S9D). Taken together, these results suggest that *POLE* mutations promote mutagenesis at defined positions within the polypurine tracts, while MMR defects contribute minimally, if at all, to this mutational specificity.

Classification of tumors by *POLE* status using polypurine mutational motifs

The relationship between the presence of driver alleles and polypurine motif frequency suggests the utility of MMF as a feature for tumor classification. Therefore, we developed the MMF-based polypurine motif classifier to distinguish samples/patients with *POLE* driver mutations from those without *POLE* drivers (Figure 2A). We trained the classifier using EAC samples with *POLE* driver mutations and samples with no *POLE* or *POLD1*mutations.The feature selection method identified $Y(R)_4[G>T](R)_2Y$, $Y(R)_3[G>T](R)_3Y$, and $Y(R)_5[G>T](R)_2Y$ motifs as the most discriminating between the two groups (Figures 2B and S10). Using these selected motifs, the trained Bagging classifier achieved almost perfect classification accuracy with the 5-fold cross-validation (Figure 2C; area under the curve [AUC] = 0.99 and Matthews correlation coefficient [MCC] = 0.99).

To test the classifier on an independent data, we applied it to 133 UCEC serous cystadenocarcinoma (SCAC) samples (Table S6; Figure S1B) and 395 colorectal adenocarcinoma (COAD) samples (Table S7; Figure S1C). The MMF classifier accurately predicted the *POLE* driver mutation status for all samples in both UCEC SCAC and COAD datasets (Figure 2C; Table S8). Moreover, the model considerably outperformed classification using other features such as mutational burden, trinucleotide mutational signatures, or frequency of G>T mutation with no specified nucleotide context (Figure

2C; Table S8). For the classification with the mutational burden, which is widely used in tumor triage, the training accuracy was comparatively high (MCC = 0.94) but dropped considerably in validation on independent UCEC SCAC and COAD datasets (MCCs = 0.81 and 0.58, respectively). The frequency of G>T mutation feature (without accounting for context) showed low prediction accuracy for training (MCC = 0.69) and for both validation datasets (MCCs = 0.39 and 0.37, respectively). Classification using trinucleotide signature sets associated with *POLE* and *POLD1* mutations (COSMIC signatures SBS10a/b/c/d) showed accuracy similar to that of the MMF classifier for the COAD dataset but considerably lower accuracy for UCEC SCAC (MCC = 0.70 compared to MCC = 1.0 for polypurine motifs; Figure 2C; Table S8). Using all trinucleotide COSMIC signatures yielded better performance than COSMIC *POLE*-related signatures, indicating that additional mutational processes operate in these tumors. However, a classifier with only one polypurine motif produced comparable results to the all-COSMIC-signature classifier, which is remarkable given that the number of features (67 signatures vs. one motif) was considerably higher in the mutational signature classifier. These observations suggest that polypurine motifs represent *POLE*-related processes far more accurately than the trinucleotide signatures.

Disambiguation of *POLE* VUSs using the polypurine motif classifier

The UCEC EAC dataset contained 34 samples with *POLE* VUSs but no known driver mutations (Table S2; Figure S11A). Polypurine motif frequency in this group was higher than in samples with no *POLE* or *POLD1* mutations (Figures 1A and 1B), suggesting that unidentified drivers may be present among *POLE* VUSs. To find samples with possible *POLE* driver mutations, we applied our trained classifier to all tumors with *POLE* VUSs in all datasets (UCEC EAC [n = 34], UCEC SCAC [n = 3]; Table S6; Figure S11B) and COAD (n = 24; Table S7; Figure S11C). We predicted that six UCEC EAC samples contained driver mutations (Table 1). These six samples clustered with the known *POLE* driver samples on the t-distributed stochastic neighbor embedding (t-SNE) plot (Figure 2B). None of the SCAC samples with *POLE* VUSs were predicted to have *POLE* driver mutations, but the classifier tagged one COAD tumor with a very high mutation burden (TCGA-CA-6717-01A-11D-1835-10; Table S7; Figure S11C) as carrying an unknown driver. The two *POLE* VUSs present in this tumor (L1235I and R1371Ter) are outside of the catalytic domains of Pol ε and are unlikely to produce a mutator polymerase. A thorough analysis of published studies determined that this tumor carried two additional *POLE* variants, P286R (a known driver) and P286C, which were missed by mutation-calling algorithms (Figure S12). This case further illustrates the consistent predicting power of our classifier.

Interestingly, among six EAC samples predicted to carry driver mutations, three contained known *POLD1* drivers even though samples with *POLD1* drivers were deliberately excluded from the training dataset. Thus, tumors with *POLD1* driver alleles share the propensity to accumulate G>T transversions in polypurine tracts. The remaining three samples contained a total of 11 *POLE* and five *POLD1* VUSs (Table 1). After eliminating silent and nonsense mutations, the list of candidate driver mutations comprised five *POLE* and four *POLD1* missense variants (Table 1). For tumors with multiple missense mutations in *POLE*/*POLD1*,

we identified the *bona fide* drivers of ultramutability by functional analysis of individual variants in yeast. We introduced mutations mimicking human *POLE-R742H*, *POLE-P916L*, *POLE-E978G*, *POLE-S461L*, *POLD1-G724R*, *POLD1-R549C*, and *POLD1-T954M* into the *POL2* and *POL3* yeast genes, respectively (Figures 3A and 3B). The *POLD1-A769D* mutation affected a non-conserved amino acid residue and could not be modeled. We then determined the effects of these alleles on the rate of spontaneous mutation conferring canavanine resistance (Can^r) and reversion of the *his7-2* reporter. The *pol2-E991G* and *pol2-S476L* mutations analogous to human *POLE-E978G* and *POLE-S461L* increased the mutation rate, while other variants present in the same tumors were not mutators (Figure 3C; Tables S9 and S10). This analysis validated E978G and S461L substitutions as drivers of ultramutation. Notably, unlike previously known drivers, the E978G variant affects the DNA polymerase rather than the exonuclease domain of Pol ε (Figure 4A). The sole variant, *POLE-R705W*, present in tumor TCGA-AX-A2HD-01A-21D-A17D-09 affects a non-conserved amino acid residue. While the polypurine motif signature and the lack of other polymerase variants suggests that *POLE-R705W* is the driver mutation, its validation awaits future studies with human cells and proteins.

To further test the robustness of the polypurine motif classifier, we assessed the functional significance of four *POLE* variants and one *POLD1* variant from tumors that had a high mutation burden but were not flagged by the classifier as carrying driver mutations. Four of these, *POLE-P102L*, *POLE-T278M*, *POLE-C402R*, and *POLD1-V477M*, yielded wild-type levels of mutagenesis when modeled in yeast (Figure 3D; Tables S9 and S10). These results reinforce our conclusion that classification based on mutation burden alone does not accurately identify tumors with functionally significant *POLE* variants. The yeast mimic of the fifth variant, *POLE-A465V*, was a mild mutator. However, the increase in the mutation rate was smaller than typically observed for pathogenic variants at the protein-DNA interface in the exonuclease domain, where Ala465 is located.[22] The significance of this mild mutator activity for mutation accumulation in the context of tumorigenesis may be limited. Tumors are heterozygous for *POLE* mutations, and, at least in yeast models, the mutator effects are reduced by half in heterozygous cells.[22] Thus, increases smaller than 3-fold become negligible in the presence of the wild-type allele. Overall, the classifier efficiently discriminates between driver and passenger mutations, but we cannot exclude that very weak drivers could escape detection.

Structural consequences of E978G and S461L variants

Proofreading by replicative polymerases entails a transfer of the primer terminus from the polymerase to the exonuclease active site and a subsequent rapid and processive return of the trimmed primer to the polymerase active site. In the wild-type enzymes, the two activities are optimally balanced for efficient and accurate replication.[58] All previously known *POLE* driver mutations affect amino acid residues in the exonuclease domain of Pol ε. Studies in yeast suggested that these mutations produce hyperactive polymerases, in which the balance of DNA synthesis and proofreading is shifted toward synthesis.[21,59] The increased polymerase activity, which we previously proposed to be key to the pathogenicity of Pol ε variants, results from the inability to move the 3′ DNA end to the exonuclease active site.[21,60] Similar to the previously known drivers, the S461L variant affects the exonuclease

domain of Pol ε (Figures 3A and 4A). The side chain of the analogous S476 in yeast Pol ε does not directly contact DNA, and the leucine substitution may indirectly impair DNA binding at the exonuclease site by altering the position of other amino acid residues critical for proofreading.

The E978G variant, however, is intriguing, as it maps to the DNA polymerase domain (Figures 3A and 4A). In the crystal structure of yeast Pol ε, the side chain of E991 (analogous to E978 in human Pol ε) faces the DNA minor groove near the −4 and −7 positions of the primer and template strands, respectively.[52] Minor groove interactions in eukaryotic and bacteriophage B-family polymerases generally involve contacts between the DNA backbone and multiple positively charged amino acid residues in the protein.[51,52,61] E991 in yeast Pol ε, together with E985 and E1048, form a cluster of acidic residues facing the DNA upstream of the primer terminus (Figure 4B), which reduces the overall positive charge of the DNA-binding interface. Our calculations of electrostatic potential show that the E991G substitution in yeast Pol ε (and presumably human E978G) increases the overall positive charge of the protein at the DNA interface (Figure 4C), which could enhance DNA binding by the polymerase domain. The net outcome of this substitution would be the same shift in the proofreading-polymerization balance that exonuclease domain mutations achieve by disrupting DNA binding in the exonuclease site.

Besides affecting surface charge, the E978G substitution may stabilize the primer terminus in the polymerase domain by altering the position of critical DNA-binding residues. In the yeast Pol ε structure, the Oε2 atom of E991 forms a hydrogen bond with Nε of a conserved R988 (Figure 4D). E991 also interacts with R989 and R1149. These three arginine residues make multiple contacts with the DNA backbone and bases in the minor groove (Figure 4D). The contact between E991 and R988 seems particularly important. R988 and equivalent residues in other B-family polymerases interact with the DNA bases at positions −4 and −5 (Figures 4D and 4E), which was proposed to be a mechanism for sensing mismatches at these positions.[51,52,61,62] According to this model, nucleotide misincorporations disrupt the interaction with R988 and destabilize the primer terminus in the polymerase active site, resulting in switching to the exonuclease site.[62] The hydrogen bond between E991 and R988 could be essential to properly position R988 (Figure 4E). The mutator effect of E991G substitution could then result from a failure to discriminate against mismatches at the −4/−5 position.

DISCUSSION

Polypurine tracts as mutation hotspots in tumors

The analysis presented here identified peculiar mutagenic properties of polypurine sequences in the human genome. We found that G>T transversions, a hallmark of *POLE*-mutant tumors, accumulate preferentially at specific positions of polypurine tracts. The mutation frequency is modest for sequences containing only three consecutive purines but increases as the tract length reaches six or more purines. Thus, the AGA trinucleotides, commonly known as hotspots of *POLE*-driven mutagenesis, are not highly mutable unless present as part of longer polypurine tracts. Perhaps the most intriguing observation is the pattern of mutability with spikes every 3 to 4 nt within the tracts.

It is not immediately obvious what properties of polypurine tracts define this unique mutagenesis pattern. We determined that tracts with the polyA context may be the primary contributor. polyA/polyT tracts (A-tracts) have distinctive structural features that may affect their mutagenic properties. Compared to B-form DNA (B-DNA), they have a narrower minor groove,[63] and minor groove interactions are critical for DNA polymerase fidelity.[64] A-tracts can also cause bending of the DNA helical axis at the junction with B-DNA segments.[65] The degree of bending increases with the divalent cation concentration and depends on the lengths of tracts. A-tracts shorter than 10 bp promote bending and those longer than ~10 bp stiffen the DNA and prevent bending. It is possible that the interruption of an A-tract by a G in $Y(A)_nG(A)_mY$ sequences creates a distortion that promotes DNA polymerase errors. Indeed, a recent study suggested that the sites of *POLE*-driven mutations in tumors have overtwisting of DNA at the −1 position and undertwisting at the +1 position in regard to the mutation site.[66] Guanines at certain positions within the $Y(A)_nG(A)_nY$ sequences could also be more prone to damage, as shown for guanines in G-tracts.[67] Polypurine tracts were also linked to recombination hotspots,[68,69] further illustrating their special features.

Notably, we observed a strong propensity for errors in polypurine tracts in tumors with *POLE*/*POLD1* mutations (this work) and in cultured cells expressing mutator polymerase variants[31] but not during synthesis by the corresponding purified polymerases *in vitro*.[59,70] Thus, DNA structure alone does not define the mutational pattern of *POLE*/*POLD1* tumors. Various cellular factors, including binding of transcription factors at polyA sequences,[71] may affect the probability of a mutation. Many polyA tracts are located in regulatory nucleosome-depleted regions near the transcriptional start sites or at replication origins.[72] However, the pattern of *POLE*-driven G>T transversions was similar in nucleosomal and nucleosome-free DNA (Figure S3), arguing against a major role of nucleosome exclusion. Furthermore, guanines in polyT context (polyA on the opposite strand) rarely mutate (Figure S4), although polyT sequences can similarly repel nucleosomes, and $(T)_nG(T)_m$ motifs could readily produce G>T transversions via slippage-mediated polymerase errors.[73] Overall, the highly reproducible pattern of polypurine tract mutability in *POLE*/*POLD1*-mutant cells and tumors suggests that cell-specific factors define the susceptibility of certain positions to mutation. The exact nature of these factors is yet to be discovered.

Applications of polypurine motif classifier

Extreme hypermutation in *POLE*-mutant tumors is associated with high neoantigen production, an enhanced anti-tumor immune response, improved patient outcomes, and responsiveness to immune checkpoint blockade.[35,36] Thus, the identification of tumors with *POLE* driver mutations is of high clinical importance. We show that the polypurine motif classifier performs better than the current approaches based on the presence of *POLE* exonuclease domain mutations, high tumor mutational burden, and contribution of *POLE*-related trinucleotide mutational signatures. We also show that the classifier can identify tumors with *POLD1* driver mutations, for which no reliable classification methods exist. Due to a low number of *POLD1*-mutant cases, the progression-free survival has not been systematically evaluated for these patients. However, it is worth noting that the statuses of all patients with the *POLD1* mutations are shown as alive in TCGA database. It is likely

that the *POLD1* mutations will define prognosis and therapy outcomes similarly to *POLE* mutations. The comparable level of hypermutation (Figure S11) and the shared polypurine motif signature further supports this view. We also note that the polypurine motif classifier is highly efficient at distinguishing tumors with driver and non-driver polymerase mutations regardless of the MSI status of the tumor.

POLE driver mutations identified in this work

Using the polypurine motif classifier and functional studies in yeast, we identified the *POLE-E978G* and *POLE-S461L* alleles as driver mutations. This expands the list of 14 previously validated somatic *POLE* drivers.[21,22,24–29,33,34] The E978G variant was reported in three pediatric brain tumors[17,34,74] and a uterine carcinosarcoma,[17] in addition to the UCEC EAC case analyzed here. Others previously predicted it to be a driver allele based on the consistent association with a high mutation burden.[17] However, another amino acid substitution (E978K) and a nonsense mutation at codon 978 also occurred in tumors,[17] suggesting that this site could be a mutagenesis hotspot. The changes at codon 978 could then be a consequence, rather than the cause, of hypermutation. The S461L variant was reported previously in a sporadic colorectal tumor[75] and another tumor of unspecified origin,[17] in addition to the UCEC EAC case. A proline substitution for S461 has previously been shown to reduce the fidelity of Pol ε *in vitro*.[34] To our knowledge, no studies have addressed the functional significance of S461L or E978G variants. Our analysis of polypurine motif signature and the mutator effects *in vivo* unequivocally verifies their driver status.

The location of E978G in the DNA polymerase domain is note-worthy. All previously known pathogenic Pol ε mutations affect amino acid residues in the exonuclease domain. Studies of yeast analogs suggested that these mutations increase the DNA synthesis capacity by preventing the primer transfer to the exonuclease active site.[21,59,60] However, changes not only in exonuclease but also in polymerase domains of B-family DNA polymerases can shift the proofreading-synthesis balance and generate mutator enzymes. Early studies of T4 DNA polymerase identified multiple mutator variants with altered proofreading-polymerization balance.[76–78] These variants occurred in both exonuclease and polymerase domains, as well as in the N-terminal domain also present in Pol ε. Biochemical characterization of the mutator polymerase domain variants suggested an increased stability of the replicating complex, which reduces the opportunity for proofreading despite the intact exonuclease domain.[79] Similarly, the L612M substitution in the polymerase domain of yeast Pol d (analogous to the L606M mutation in the UCEC EAC dataset) impairs the polymerase-to-exonuclease switching.[80] In light of these studies, it is surprising that driver mutations affecting the DNA polymerase domain of Pol ε have not been previously found. We hypothesize that the E978G substitution in human Pol ε improves DNA binding by the polymerase domain by increasing the positive charge at the DNA interface or by altering the orientation of the key DNA-binding arginine residues (Figures 4B–4E). Of note, many mutator T4 DNA polymerase variants deficient in polymerase-to-exonuclease switching increase the overall positive charge of the polymerase domain.[76,78] Indeed, Reha-Krantz and co-authors proposed that changes to basic amino acid residues produce mutator enzymes by

stabilizing the primer terminus in the polymerase active site.[78] Future biochemical studies will help clarify the mechanism of Pol ε-E978G infidelity.

In summary, the discovery and experimental validation of the E978G variant as pathogenic point to the existence of a distinct class of driver mutations that impact the DNA polymerase rather than the exonuclease domain of Pol ε. Our structural analysis and comparison to prior findings with homologous polymerases suggest that E978G may act by increasing DNA binding in the DNA polymerase domain and shifting the proofreading-polymerization balance toward synthesis. Additional driver mutations of this class will likely be discovered. Besides the classic minor groove interface, driver mutations could affect the P-domain, a DNA-binding moiety unique to Pol ε.[52] Because amino acid substitutions in different sites could produce the same effect on the polymerase/exonuclease balance, the recurrence of mutations in tumors is not an essential criterion of pathogenicity. Each individual substitution could be a rare driver. Our findings highlight the limitations of the current focus on exonuclease domain mutations when stratifying cancer patients.

Limitations of the study

This study established the association of *POLE*-mutant tumor status with G>T transversions at defined positions of polypurine tracts. We could not unequivocally identify the exact polypurine sequences responsible for this mutagenesis pattern due to the large number of possible purine combinations and the limited size of TCGA dataset. Larger datasets would also provide statistical power to assess the mutational properties of longer polypurine tracts (>14 nt) that do exist in the human genome but remain unexplored. The understanding of mutator effects of rare *POLE* alleles also remains limited because very few sequenced tumors with these alleles are available. Accurate evaluation of the impact of MMR deficiency in *POLE*-mutant tumors has been challenging in the present and other studies, as current datasets lack a sufficient number of tumors with the same *POLE* allele and different MMR statuses. Finally, the mechanism of polypurine tract hypermutability awaits further exploration.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Polina Shcherbakova (pshcherb@unmc.edu).

Materials availability—Yeast strains and plasmids generated in this work are available upon request.

Data and code availability

- This paper analyzes existing, publicly available data. The combined MAF files for the UCEC and COAD cohorts are available in the Zenodo repository at https://zenodo.org/records/10022634. The original TCGA somatic mutation calls

are available to download from the Genomic Data Commons repository of the National Institutes of Health.

- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Dataset description—TCGA somatic mutation data for UCEC and COAD cohorts were downloaded from the GDC Data Portal repository (accessed on Aug 4, 2021). The data analysis and classification models were constructed using 391 UCEC EAC samples with diagnoses of adenocarcinoma, endometrioid adenocarcinoma, clear cell adenocarcinoma, or endometrioid adenocarcinoma, secretory variant. Samples with prior treatment, undifferentiated carcinoma and serous surface papillary carcinoma were excluded from the analysis. The two independent test sets contained 133 UCEC SCAC and 395 COAD samples. The SCAC set consisted of four tumors with known *POLE* drivers, five tumors with *POLE* and/or *POLD1* VUSs, and 124 tumors with no *POLE* or *POLD1* mutations (Table S6). The COAD samples included seven tumors with known *POLE* driver mutations, 42 samples with *POLE* and/or *POLD1* VUSs, and 346 tumors with no *POLE* or *POLD1* mutations (Table S7). No known *POLD1* driver mutations were present in the SCAC or COAD datasets.

Yeast strains—Haploid *Saccharomyces cerevisiae* strains with *pol2* and *pol3* mutations mimicking those present in human tumors were constructed as described previously.[22] Briefly, the mutations were created by site-directed mutagenesis in *URA3*-based integrative plasmids containing N- or C-terminal fragments of *POL2* or *POL3* genes (Tables S11 and S12). The diploid strain PSD93 (*MAT*a/*MAT*α *ade5-1*/*ade5-1 lys2::InsE*$_{A14}$/*lys2::InsE*$_{A14}$ *trp1-289*/*trp1-289 his7-2*/*his7-2 leu2-3,112*/*leu2-3,112 ura3-52*/*ura3-52*) was transformed with YIpDK1-pol2-x, p173-pol2-x, p170-pol3-x, or pAB0068-pol3-x digested with an appropriate restriction enzyme (Table S11) to achieve integration of the plasmid into one chromosomal copy of *POL2* or *POL3*. The resulting modified loci contained a full-length mutant *pol* allele and a truncated wild-type allele separated by the *URA3* marker. The presence of heterozygous mutations was confirmed by PCR and Sanger sequencing. The diploids were then sporulated and *pol2* and *pol3* mutant haploids were obtained by tetrad dissection. Derivatives of these haploids that underwent spontaneous recombination to lose the *URA3* vector sequences and retain a single, full-length mutant or wild-type *POL2* or *POL3* allele were selected for on media containing 5-fluoroorotic acid. Clones containing the mutant *pol2* or *pol3* alleles were identified by Sanger sequencing. If the tetrad dissection revealed that the polymerase mutation was lethal in the absence of the wild-type allele (*pol3-G731R* and *pol2-C417R* mutations), the lethality was further confirmed as follows. The haploid strain E134 (*MAT*α *ade5-1 lys2-InsE*$_{A14}$ *trp1-289 his7-2 leu2-3,112 ura3-52*) was transformed with p170-pol3-G731R or YIpDK1-pol2-C417R (Table S11) digested with BseRI and BglII, respectively, which places the *URA3* marker between a full-length wild-type *POL* allele and a truncated mutant allele at the chromosomal *POL3* or *POL2* locus. These strains were subsequently transformed with pPOL3[23] (2mm ori *TRP1 POL3*)

or YEpPOL2-TRP[22] (2mm ori *TRP1 POL2*) to introduce extrachromosomal full-length wild-type polymerase genes. Maintaining selection for the *TRP1* plasmid, we then applied 5-fluoroorotic acid selection to obtain clones that underwent spontaneous recombination within the integrated construct, have lost the *URA3* vector sequences, and retained a single, full-length mutant or wild-type chromosomal *POL* allele. Clones containing the chromosomal *pol3-G731R* or *pol2-C417R* alleles were identified by Sanger sequencing, and lethality was confirmed by the inability of these clones to lose the plasmid with the wild-type *POL* gene in the absence of selection for the *TRP1* marker.

METHOD DETAILS

Mutation calling and tumor classification—To minimize errors from different mutation calling algorithms,[88] results from mutect2, muse, somaticsniper and varscan were combined using the consensus approach.[89] Namely, a mutation was included in the analysis if it was called by two or more mutation calling algorithms. All samples were labeled based on the presence or absence of known *POLE* and *POLD1* driver mutations and MSI status. Driver mutations in *POLE* and *POLD1* were annotated based on the existing functional analysis data (mutator or tumor phenotype in model systems and/or demonstrated exonuclease defect), as well as co-segregation with the cancer phenotype for variants that have also been reported as germline mutations (Table S13). The mutator effect of the *POLD1-E318K* allele altering the catalytic glutamate in the exonuclease active site of Pol δ and, therefore, suspected to be a driver mutation, was validated by modeling in yeast prior to the analysis (Table S14). Mutations not annotated as drivers were categorized as VUSs. The MSI status for all samples was extracted from the consensus-combined MAF files using the TCGA biolinks R/Bioconductor package v. 2.18.0 and categorized into MSI-high (MSI), MSI-low (MSI-L) and microsatellite-stable (MSS) status as previously described.[84] For samples where the MSI status was unknown, it was predicted using the MSIpred python package.[83] Microsatellite-stable (MSS) tumors and tumors with low MSI (MSI-L) were grouped together and referred to as “MSS”.

Mutation rate measurements—The rate of spontaneous Can^r mutation, His^+ reversion, and Lys^+ reversion in haploid (if viable) or heterozygous diploid yeast *pol* mutants was measured by fluctuation analysis as described previously.[22]

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of mutational motifs—We have generated 104 $(R)_n[G > T](R)_m$, $\{n, m \in \mathbb{N} \mid 0 \leq n, m \leq 13\}$ mutational motifs (Table S3). These motifs are equivalent to C>A mutations in a polypyrimidine context on the opposite strand, (Y)n[C>A](Y)m. To distinguish between polypurine tracts of different lengths, pyrimidine (Y) “terminators” were added at the beginning and the end of each motif, y(R)n[G>T](R)my. We also generated 104 analogous y(A)n[G>T](A)my, $\{n, m \in \mathbb{N} \mid 0 \leq n, m \leq 13\}$, $Y(G)_n[G > T](G)_m Y$, $\{n, m \in \mathbb{N} \mid 0 \leq n, m \leq 13\}$, and y(A)n[A>C](A)my, $\{n, m \in \mathbb{N} \mid 0 \leq n, m \leq 13\}$ motifs for the analysis of mutations in polyA and polyG context, 28 y(AG)nA[G>T](AG)my, $\{n, m \in \mathbb{N} \mid 0 \leq n, m \leq 6\}$

motifs for the analysis of mutations in (AG)n repeats, and additional 104 $R(Y)_n[G > T](Y)_mR, \{n, m \in \mathbb{N} \mid 0 \leq n, m \leq 13\}$ mutational motifs as a control.

For each query motif, the whole exome of a tumor was scanned and a total number of occurrences of the motif in a sample was calculated using the MutaGene[82,90] python package, "motif" module (parameters: -g hg38 -w 100, -t 1.0), which can be found at https://github.com/Panchenko-Lab/mutagene). Both DNA strands were treated equally. Whole exome sequences were extracted using BioMart (v. 2.46.3).[86] For each sample and each motif, MMF was calculated by dividing the number of mutational motifs by the total number of occurrences of that nucleotide context in the exome (Table S15), MMF = $M_{mut\_motif}/N_{context}$. For example, for $_Y$R[G>T]R$_Y$ motif, MMF = $M_{YR[G>T]RY}/N_{YRGRY}$.

For the trinucleotide mutational signature analysis, we applied the "identify" module of the MutaGene package with the default parameters.[90] The MutaGene package uses maximum likelihood estimation procedure for decomposition of the mutational profile into a predefined set of mutational signatures and computes their exposures. We utilized a set of 67 signatures from the Catalog of Somatic Mutations in Cancer (COSMICv3). The set contained 49 mutational signatures associated with certain mutational processes whereas other 18 signatures were indicated as signatures associated with sequencing artifacts. Furthermore, the number of mutations explained by a given mutational signature was used as a feature to develop a machine learning model to discriminate between samples with *POLE* drivers and samples with no *POLE*/*POLD1* mutations for comparison with the classification based on the polypurine motifs. In addition to the primary analysis, we utilized the MutSigProfiler python package to perform a trinucleotide mutational signature analysis using a set of 67 mutational signatures from COSMICv3.

Data analysis and machine learning modeling—Data processing, statistical tests, and visualization were performed in MATLAB (Mathworks, Inc., MA, USA, v. R2020a), Python (v. 3.8.5), and R (v. 4.0.3) platforms. The statistical significance is displayed as following: *, p value <0.05; ***, p value <0.001; ns, no significance, as indicated in figure legends.

Since our variables are not normally distributed, we applied non-parametric ANOVA test (Kruskal-Wallis) to analyze the association between MSI status, *POLE* driver mutation status and MMF. Dunn's multiple comparison test was used to perform post-hoc analyses. The FDR correction was applied to correct for multiple comparison. The tSNE was applied for high-dimensional data visualization. Horizontal lines and stars indicate statistical associations between groups. To select mutational motifs most predictive of the *POLE* driver status, feature selection was performed using *MFeaST*.[85] Classification model development was carried out with the scikit-learn python package (v. 1.1.1) and the MATLAB Classification Learner App with 5-fold cross-validation and default tuning hyper-parameters. To identify the most accurate classification model,[91] several machine learning classifiers were applied, namely linear discriminant, support vector machine, gradient boosting, AdaBoostClassifier, Bagging classifier, KNeighborsClassifier, Decision Tree, Random Forest, Gaussian Naive Bayes. We used the following metrics for model evaluation: Receiver Operating Characteristics (ROC), precision, recall, and F1 score. As

our training and test datasets are imbalanced, Matthew's correlation coefficient (MCC) was also calculated.

Detecting mutational motifs in nucleosomal and non-nucleosomal DNA— To determine whether there is a statistically significant difference between MMF in nucleosomal and non-nucleosomal DNA, we constructed the contingency table based on the following three steps. In the first step, 147-bp paired-end MNase-seq DNA fragments associated with seven human lymphoblastoid cell lines were downloaded from NCBI (GSE36979). This data had higher genomic coverage than other available MNase-seq data, and its 147-bp DNA fragments can specify nucleosome positioning with up to a single nucleotide resolution.[81] The MNase-seq fragments were mapped to the reference human genome (GRCh38), and the nucleosome positioning scoring profile was obtained. Since the generated profile included several local peaks, a kernel smoother function was used to determine the exact location of nucleosomes.[92] In the second step, the exome sequences were extracted using the BioMart software library. Both the plus and minus strands of the exome sequences were mapped to whole genome (GRChr38), and their genomic coordinates were specified. Then, the locations of the acquired G>T mutations were mapped to the coordinates of exome sequences and nucleosomal and non-nucleosomal regions. In the third step, every sequence motif was searched in all exome sequences, and the total number of G>T mutations and non-mutated G nucleotides were counted for both the nucleosomal and non-nucleosomal DNA. Finally, we performed the Fisher's exact test to determine if there is an association between mutation in the motif and the presence/absence of the motif in the nucleosomal region.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Chelsea Bulock for the YIpAM26 derivative with wild-type *POL3*, Wyatt Petersen for help with the preparation of Figure 3, and Linda Reha-Krantz for critical comments on the manuscript. This work was supported by the National Institutes of Health grants ES015869 and CA239688 to P.V.S. D.O., Y.M.-S., and A.R.P. were supported by the Department of Pathology and Molecular Medicine, Queen's University, Canada. A.R.P. acknowledges the support of the New Frontier in Research Fund (NFRF) and the Natural Sciences and Engineering Research Council of Canada (NSERC). A.R.P. is the recipient of a Senior Canada Research Chair in Computational Biology and Biophysics and a Senior Investigator Award from the Ontario Institute of Cancer Research, Canada. K.V.S. was supported by the Russian Science Foundation (grant RSF 19-74-30003).

REFERENCES

1. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, et al. (2013). Signatures of mutational processes in human cancer. Nature 500, 415–421. [PubMed: 23945592]
2. Stratton MR, Campbell PJ, and Futreal PA (2009). The cancer genome. Nature 458, 719–724. [PubMed: 19360079]
3. Rogozin IB, Pavlov YI, Goncearenco A, De S, Lada AG, Poliakov E, Panchenko AR, and Cooper DN (2018). Mutational signatures and mutable motifs in cancer genomes. Briefings Bioinf. 19, 1085–1101.

4. Ostroverkhova D, Przytycka TM, and Panchenko AR (2023). Cancer driver mutations: predictions and reality. Trends Mol. Med 29, 554–566. [PubMed: 37076339]
5. Ganai RA, and Johansson E (2016). DNA replication—a matter of fidelity. Mol. Cell 62, 745–755. [PubMed: 27259205]
6. Lynch HT, Snyder CL, Shaw TG, Heinen CD, and Hitchins MP (2015). Milestones of Lynch syndrome: 1895–2015. Nat. Rev. Cancer 15, 181–194. [PubMed: 25673086]
7. Yamamoto H, and Imai K (2019). An updated review of microsatellite instability in the era of next-generation sequencing and precision medicine. Semin. Oncol 46, 261–270. [PubMed: 31537299]
8. Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337. [PubMed: 22810696]
9. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73. [PubMed: 23636398]
10. Barbari SR, and Shcherbakova PV (2017). Replicative DNA polymerase defects in human cancers: consequences, mechanisms, and implications for therapy. DNA Repair 56, 16–25. [PubMed: 28687338]
11. Rayner E, van Gool IC, Palles C, Kearsey SE, Bosse T, Tomlinson I, and Church DN (2016). A panoply of errors: polymerase proofreading domain mutations in cancer. Nat. Rev. Cancer 16, 71–81. [PubMed: 26822575]
12. Burgers PMJ, and Kunkel TA (2017). Eukaryotic DNA replication fork. Annu. Rev. Biochem 86, 417–438. [PubMed: 28301743]
13. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I, et al. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat. Genet 45, 136–144. [PubMed: 23263490]
14. Palles C, Martin L, Domingo E, Chegwidden L, McGuire J, Cuthill V, Heitzer E, CORGI Consortium; Kerr R, Kerr D, et al. (2022). The clinical features of polymerase proof-reading associated polyposis (PPAP) and recommendations for patient management. Fam. Cancer 21, 197–209. [PubMed: 33948826]
15. Rohlin A, Zagoras T, Nilsson S, Lundstam U, Wahlström J, Hultén L, Martinsson T, Karlsson GB, and Nordling M (2014). A mutation in *POLE* predisposing to a multi-tumour phenotype. Int. J. Oncol 45, 77–81. [PubMed: 24788313]
16. Valle L, Hernández-Illán E, Bellido F, Aiza G, Castillejo A, Castillejo MI, Navarro M, Seguí N, Vargas G, Guarinos C, et al. (2014). New insights into *POLE* and *POLD1* germline mutations in familial colorectal cancer and polyposis. Hum. Mol. Genet 23, 3506–3512. [PubMed: 24501277]
17. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, Davidson S, Edwards M, Elvin JA, Hodel KP, et al. (2017). Comprehensive analysis of hypermutation in human cancer. Cell 171, 1042–1056.e10. [PubMed: 29056344]
18. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2, 401–404. [PubMed: 22588877]
19. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, D805–D811. [PubMed: 25355519]
20. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, and Staudt LM (2016). Toward a shared vision for cancer genomic data. N. Engl. J. Med 375, 1109–1112. [PubMed: 27653561]
21. Barbari SR, Beach AK, Markgren JG, Parkash V, Moore EA, Johansson E, and Shcherbakova PV (2022). Enhanced polymerase activity permits efficient synthesis by cancer-associated DNA polymerase ε variants at low dNTP levels. Nucleic Acids Res. 50, 8023–8040. [PubMed: 35822874]
22. Barbari SR, Kane DP, Moore EA, and Shcherbakova PV (2018). Functional analysis of cancer-associated DNA polymerase ε variants in *Saccharomyces cerevisiae*. G3 (Bethesda) 8, 1019–1029. [PubMed: 29352080]

23. Daee DL, Mertz TM, and Shcherbakova PV (2010). A cancer-associated DNA polymerase δ variant modeled in yeast causes a catastrophic increase in genomic instability. Proc. Natl. Acad. Sci. USA 107, 157–162. [PubMed: 19966286]
24. Galati MA, Hodel KP, Gams MS, Sudhaman S, Bridge T, Zahurancik WJ, Ungerleider NA, Park VS, Ercan AB, Joksimovic L, et al. (2020). Cancers from novel Pole-mutant mouse models provide insights into polymerase-mediated hypermutagenesis and immune checkpoint blockade. Cancer Res. 80, 5606–5618. [PubMed: 32938641]
25. Herzog M, Alonso-Perez E, Salguero I, Warringer J, Adams DJ, Jackson SP, and Puddu F (2021). Mutagenic mechanisms of cancer-associated DNA polymerase ε alleles. Nucleic Acids Res. 49, 3919–3931. [PubMed: 33764464]
26. Hodel KP, Sun MJS, Ungerleider N, Park VS, Williams LG, Bauer DL, Immethun VE, Wang J, Suo Z, Lu H, et al. (2020). POLE mutation spectra are shaped by the mutant allele identity, Its abundance, and mismatch repair status. Mol. Cell 78, 1166–1177.e6. [PubMed: 32497495]
27. Kane DP, and Shcherbakova PV (2014). A common cancer-associated DNA polymerase ε mutation causes an exceptionally strong mutator phenotype, indicating fidelity defects distinct from loss of proofreading. Cancer Res. 74, 1895–1901. [PubMed: 24525744]
28. Lee M, Eng G, Barbari SR, Deshpande V, Shcherbakova PV, and Gala MK (2020). Homologous recombination repair truncations predict hypermutation in microsatellite stable colorectal and endometrial tumors. Clin. Transl. Gastroenterol 11, e00149. [PubMed: 32352724]
29. Li HD, Cuevas I, Zhang M, Lu C, Alam MM, Fu YX, You MJ, Akbay EA, Zhang H, and Castrillon DH (2018). Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load. J. Clin. Invest 128, 4179–4191. [PubMed: 30124468]
30. Li L, Murphy KM, Kanevets U, and Reha-Krantz LJ (2005). Sensitivity to phosphonoacetic acid: a new phenotype to probe DNA polymerase d in *Saccharomyces cerevisiae*. Genetics 170, 569–580. [PubMed: 15802517]
31. Mertz TM, Baranovskiy AG, Wang J, Tahirov TH, and Shcherbakova PV (2017). Nucleotide selectivity defect and mutator phenotype conferred by a colon cancer-associated DNA polymerase δ mutation in human cells. Oncogene 36, 4427–4433. [PubMed: 28368425]
32. Murphy K, Darmawan H, Schultz A, Fidalgo da Silva E, and Reha-Krantz LJ (2006). A method to select for mutator DNA polymerase ds in *Saccharomyces cerevisiae*. Genome 49, 403–410. [PubMed: 16699561]
33. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, Chao H, Doddapaneni H, Muzny DM, Gibbs RA, et al. (2014). Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. Genome Res. 24, 1740–1750. [PubMed: 25228659]
34. Shlien A, Campbell BB, de Borja R, Alexandrov LB, Merico D, Wedge D, Van Loo P, Tarpey PS, Coupland P, Behjati S, et al. (2015). Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. Nat. Genet 47, 257–262. [PubMed: 25642631]
35. Church DN, Stelloo E, Nout RA, Valtcheva N, Depreeuw J, ter Haar N, Noske A, Amant F, Tomlinson IPM, Wild PJ, et al. (2015). Prognostic significance of *POLE* proofreading mutations in endometrial cancer. J. Natl. Cancer Inst 107, 402. [PubMed: 25505230]
36. Ma X, Riaz N, Samstein RM, Lee M, Makarov V, Valero C, Chowell D, Kuo F, Hoen D, Fitzgerald CWR, et al. (2022). Functional landscapes of *POLE* and *POLD1* mutations in checkpoint blockade-dependent antitumor immunity. Nat. Genet 54, 996–1012. [PubMed: 35817971]
37. Church DN, Briggs SEW, Palles C, Domingo E, Kearsey SJ, Grimes JM, Gorman M, Martin L, Howarth KM, Hodgson SV, et al. (2013). DNA polymerase ε and δ exonuclease domain mutations in endometrial cancer. Hum. Mol. Genet 22, 2820–2828. [PubMed: 23528559]
38. Cui J, Chen X, Zhai Q, Chen N, Li X, Zhang Y, Wang H, Bian X, Gao N, Chen D, et al. (2023). A novel somatic mutation in POLE exonuclease domain associated with ultra-mutational signature and MMR deficiency in endometrial cancer: a case report. Diagn. Pathol 18, 19. [PubMed: 36765365]

39. Haradhvala NJ, Kim J, Maruvka YE, Polak P, Rosebrock D, Livitz D, Hess JM, Leshchiner I, Kamburov A, Mouw KW, et al. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat. Commun 9, 1746. [PubMed: 29717118]
40. Hatakeyama K, Ohshima K, Nagashima T, Ohnami S, Ohnami S, Serizawa M, Shimoda Y, Maruyama K, Akiyama Y, Urakami K, et al. (2018). Molecular profiling and sequential somatic mutation shift in hypermutator tumours harbouring POLE mutations. Sci. Rep 8, 8700. [PubMed: 29880869]
41. Hodel KP, de Borja R, Henninger EE, Campbell BB, Ungerleider N, Light N, Wu T, LeCompte KG, Goksenin AY, Bunnell BA, et al. (2018). Explosive mutation accumulation triggered by heterozygous human Pol ε proofreading-deficiency is driven by suppression of mismatch repair. Elife 7:e32692. [PubMed: 29488881]
42. Jaksik R, Wheeler DA, and Kimmel M (2023). Detection and characterization of constitutive replication origins defined by DNA polymerase ε. BMC Biol. 21, 41. [PubMed: 36829160]
43. Johansen AFB, Kassentoft CG, Knudsen M, Laursen MB, Madsen AH, Iversen LH, Sunesen KG, Rasmussen MH, and Andersen CL (2019). Validation of computational determination of microsatellite status using whole exome sequencing data from colorectal cancer patients. BMC Cancer 19, 971. [PubMed: 31638937]
44. Lindsay H, Scollon S, Reuther J, Voicu H, Rednam SP, Lin FY, Fisher KE, Chintagumpala M, Adesina AM, Parsons DW, et al. (2019). Germline *POLE* mutation in a child with hypermutated medullo-blastoma and features of constitutional mismatch repair deficiency. Cold Spring Harb. Mol. Case Stud 5, a004499. [PubMed: 31624068]
45. Park VS, Sun MJS, Frey WD, Williams LG, Hodel KP, Strauss JD, Wellens SJ, Jackson JG, and Pursell ZF (2022). Mouse model and human patient data reveal critical roles for Pten and p53 in suppressing POLE mutant tumor development. NAR Cancer 4, zcac004. [PubMed: 35252866]
46. Poetsch AR, Boulton SJ, and Luscombe NM (2018). Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. Genome Biol. 19, 215. [PubMed: 30526646]
47. Temko D, Van Gool IC, Rayner E, Glaire M, Makino S, Brown M, Chegwidden L, Palles C, Depreeuw J, Beggs A, et al. (2018). Somatic POLE exonuclease domain mutations are early events in sporadic endometrial and colorectal carcinogenesis, determining driver mutational landscape, clonal neoantigen burden and immune response. J. Pathol 245, 283–296. [PubMed: 29604063]
48. Tomkova M, Tomek J, Kriaucionis S, and Schuster-Böckler B (2018). Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 19, 129. [PubMed: 30201020]
49. Yu S, Sun Z, Zong L, Yan J, Yu M, Chen J, and Lu Z (2022). Clinicopathological and molecular characterization of high-grade endometrial carcinoma with POLE mutation: a single center study. J. Gynecol. Oncol 33, e38. [PubMed: 35320887]
50. Momen S, Fassihi H, Davies HR, Nikolaou C, Degasperi A, Stefanato CM, Dias JML, Dasgupta D, Craythorne E, Sarkany R, et al. (2019). Dramatic response of metastatic cutaneous angiosarcoma to an immune checkpoint inhibitor in a patient with xeroderma pigmentosum: whole-genome sequencing aids treatment decision in end-stage disease. Cold Spring Harb. Mol. Case Stud 5, a004408. [PubMed: 31645345]
51. Franklin MC, Wang J, and Steitz TA (2001). Structure of the replicating complex of a Pol α family DNA polymerase. Cell 105, 657–667. [PubMed: 11389835]
52. Hogg M, Osterman P, Bylund GO, Ganai RA, Lundström EB, Sauer-Eriksson AE, and Johansson E (2014). Structural basis for processive DNA synthesis by yeast DNA polymerase ε. Nat. Struct. Mol. Biol 21, 49–55. [PubMed: 24292646]
53. Maki H (2002). Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. Annu. Rev. Genet 36, 279–303. [PubMed: 12429694]
54. Nesta AV, Tafur D, and Beck CR (2021). Hotspots of human mutation. Trends Genet. 37, 717–729. [PubMed: 33199048]
55. Sehested A, Meade J, Scheie D, Østrup O, Bertelsen B, Misiakou MA, Sarosiek T, Kessler E, Melchior LC, Munch-Petersen HF, et al. (2022). Constitutional *POLE* variants causing a

phenotype reminiscent of constitutional mismatch repair deficiency. Hum. Mutat 43, 85–96. [PubMed: 34816535]

56. Wimmer K, Beilken A, Nustede R, Ripperger T, Lamottke B, Ure B, Steinmann D, Reineke-Plaass T, Lehmann U, Zschocke J, et al. (2017). A novel germline *POLE* mutation causes an early onset cancer prone syndrome mimicking constitutional mismatch repair deficiency. Fam. Cancer 16, 67–71. [PubMed: 27573199]
57. Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, Brookes DH, Wilson L, Chen J, Liles K, et al. (2018). Improvements to the APBS biomolecular solvation software suite. Protein Sci. 27, 112–128. [PubMed: 28836357]
58. Reha-Krantz LJ (2010). DNA polymerase proofreading: Multiple roles maintain genome stability. Biochim. Biophys. Acta 1804, 1049–1063. [PubMed: 19545649]
59. Xing X, Kane DP, Bulock CR, Moore EA, Sharma S, Chabes A, and Shcherbakova PV (2019). A recurrent cancer-associated substitution in DNA polymerase ε produces a hyperactive enzyme. Nat. Commun 10, 374. [PubMed: 30670691]
60. Parkash V, Kulkarni Y, Ter Beek J, Shcherbakova PV, Kamerlin SCL, and Johansson E (2019). Structural consequence of the most frequently recurring cancer-associated substitution in DNA polymerase ε. Nat. Commun 10, 373. [PubMed: 30670696]
61. Swan MK, Johnson RE, Prakash L, Prakash S, and Aggarwal AK (2009). Structural basis of high-fidelity DNA synthesis by yeast DNA polymerase δ. Nat. Struct. Mol. Biol 16, 979–986. [PubMed: 19718023]
62. Ganai RA, Bylund GO, and Johansson E (2015). Switching between polymerase and exonuclease sites in DNA polymerase ε. Nucleic Acids Res. 43, 932–942. [PubMed: 25550436]
63. Alexeev DG, Lipanov AA, and Skuratovskii I (1987). Poly(dA).poly(dT) is a B-type double helix with a distinctively narrow minor groove. Nature 325, 821–823. [PubMed: 3821870]
64. Kunkel TA, and Bebenek K (2000). DNA replication fidelity. Annu. Rev. Biochem 69, 497–529. [PubMed: 10966467]
65. Hud NV, and Plavec J (2003). A unified model for the origin of DNA sequence-directed curvature. Biopolymers 69, 144–158. [PubMed: 12717729]
66. Karolak A, Levatic J, and Supek F (2022). A framework for mutational signature analysis based on DNA shape parameters. PLoS One 17, e0262495. [PubMed: 35015788]
67. Morikawa M, Kino K, Oyoshi T, Suzuki M, Kobayashi T, and Miyazawa H (2014). Analysis of guanine oxidation products in double-stranded DNA and proposed guanine oxidation pathways in single-stranded, double-stranded or quadruplex DNA. Biomolecules 4, 140–159. [PubMed: 24970209]
68. Bagshaw ATM, Pitt JPW, and Gemmell NJ (2006). Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots. BMC Genom. 7, 179.
69. Myers S, Bottolo L, Freeman C, McVean G, and Donnelly P (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science 310, 321–324. [PubMed: 16224025]
70. Mertz TM, Sharma S, Chabes A, and Shcherbakova PV (2015). Colon cancer-associated mutator DNA polymerase d variant causes expansion of dNTP pools increasing its own infidelity. Proc. Natl. Acad. Sci. USA 112, E2467–E2476. [PubMed: 25827231]
71. Solomon MJ, Strauss F, and Varshavsky A (1986). A mammalian high mobility group protein recognizes any stretch of six A.T base pairs in duplex DNA. Proc. Natl. Acad. Sci. USA 83, 1276–1280. [PubMed: 3456586]
72. Tubbs A, Sridharan S, van Wietmarschen N, Maman Y, Callen E, Stanlie A, Wu W, Wu X, Day A, Wong N, et al. (2018). Dual roles of poly(dA:dT) tracts in replication initiation and fork collapse. Cell 174, 1127–1142.e19. [PubMed: 30078706]
73. Kunkel TA, and Soni A (1988). Mutagenesis by transient misalignment. J. Biol. Chem 263, 14784–14789. [PubMed: 3049589]
74. Johnson A, Severson E, Gay L, Vergilio JA, Elvin J, Suh J, Daniel S, Covert M, Frampton GM, Hsu S, et al. (2017). Comprehensive genomic profiling of 282 pediatric low- and high-grade gliomas reveals genomic drivers, tumor mutational burden, and hypermutation signatures. Oncol. 22, 1478–1490.

75. Stenzinger A, Pfarr N, Endris V, Penzel R, Jansen L, Wolf T, Herpel E, Warth A, Klauschen F, Kloor M, et al. (2014). Mutations in *POLE* and survival of colorectal cancer patients - link to disease stage and treatment. Cancer Med. 3, 1527–1538. [PubMed: 25124163]
76. Li V, Hogg M, and Reha-Krantz LJ (2010). Identification of a new motif in family B DNA polymerases by mutational analyses of the bacteriophage T4 DNA polymerase. J. Mol. Biol 400, 295–308. [PubMed: 20493878]
77. Reha-Krantz LJ, and Bessman MJ (1981). Studies on the biochemical basis of mutation VI. Selection and characterization of a new bacteriophage T4 mutator DNA polymerase. J. Mol. Biol 145, 677–695. [PubMed: 6267293]
78. Stocki SA, Nonay RL, and Reha-Krantz LJ (1995). Dynamics of bacteriophage T4 DNA polymerase function: identification of amino acid residues that affect switching between polymerase and 3'–>5' exonuclease activities. J. Mol. Biol 254, 15–28. [PubMed: 7473755]
79. Reha-Krantz LJ, Woodgate S, and Goodman MF (2014). Engineering processive DNA polymerases with maximum benefit at minimum cost. Front. Microbiol 5, 380. [PubMed: 25136334]
80. Nick McElhinny SA, Stith CM, Burgers PMJ, and Kunkel TA (2007). Inefficient proofreading and biased error rates during inaccurate DNA synthesis by a mutant derivative of *Saccharomyces cerevisiae* DNA polymerase δ. J. Biol. Chem 282, 2324–2332. [PubMed: 17121822]
81. Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, and Pritchard JK (2012). Controls of nucleosome positioning in the human genome. PLoS Genet. 8, e1003036. [PubMed: 23166509]
82. Brown AL, Li M, Goncearenco A, and Panchenko AR (2019). Finding driver mutations in cancer: Elucidating the role of background mutational processes. PLoS Comput. Biol 15, e1006981. [PubMed: 31034466]
83. Wang C, and Liang C (2018). MSIpred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. Sci. Rep 8, 17546. [PubMed: 30510242]
84. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 44, e71. [PubMed: 26704973]
85. Gerolami J, Wong JJM, Zhang R, Chen T, Imtiaz T, Smith M, Jamaspishvili T, Koti M, Glasgow JI, Mousavi P, et al. (2022). A computational approach to identification of candidate biomarkers in high-dimensional molecular data. Diagnostics 12, 1997. [PubMed: 36010347]
86. Durinck S, Spellman PT, Birney E, and Huber W (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc 4, 1184–1191. [PubMed: 19617889]
87. Diaz-Gay M, Vangara R, Barnes M, Wang X, Islam SMA, Vermes I, Narasimman NB, Yang T, Jiang Z, Moody S, et al. (2023). Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. Preprint at bioRxiv.
88. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY, et al. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat. Methods 12, 623–630. [PubMed: 25984700]
89. Wang M, Luo W, Jones K, Bian X, Williams R, Higson H, Wu D, Hicks B, Yeager M, and Zhu B (2020). SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. Sci. Rep 10, 12898. [PubMed: 32732891]
90. Goncearenco A, Rager SL, Li M, Sang QX, Rogozin IB, and Panchenko AR (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic Acids Res. 45, W514–W522. [PubMed: 28472504]
91. Duda RO, Hart PE, and Stork DG (2001). Pattern Classification, 2nd edn (Wiley).
92. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, and Sidow A (2011). Determinants of nucleosome organization in primary human cells. Nature 474, 516–520. [PubMed: 21602827]

Highlights

- POLE variants promote genome-wide mutations at defined positions of polypurine tracts
- Polypurine-motif-based classifier identifies tumors with POLE and POLD1 drivers
- Two POLE drivers, S461L and E978G, are validated experimentally
- E978G driver mutation affects the DNA polymerase domain of Pol ε

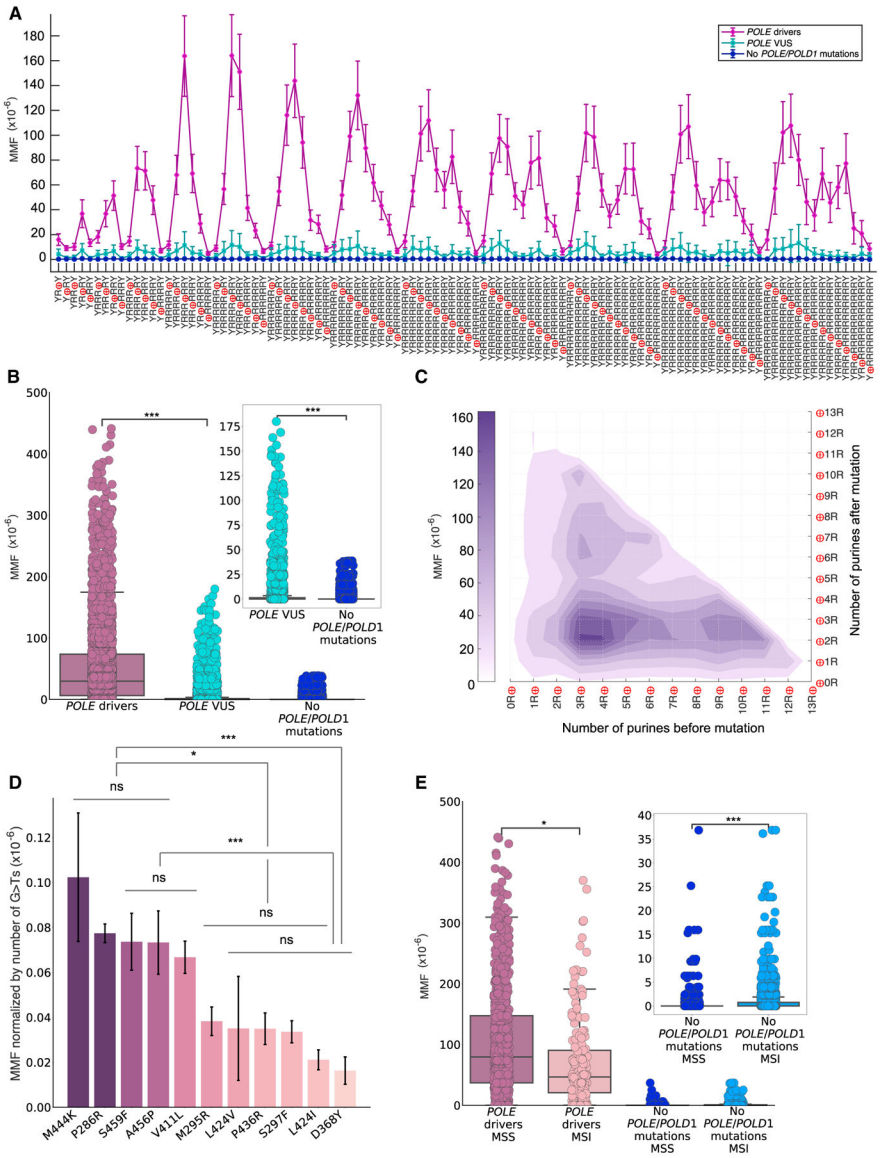


Figure 1. Tumors with POLE driver mutations accumulate G>T transversions in polypurine tracts

(A) The mean value of MMF for each of 104 polypurine motifs in tumors with *POLE* driver mutations (n = 44), *POLE* VUSs (n = 30), or no *POLE* or *POLD1* mutations (n = 292). Error bars indicate 95% confidence interval (CI). The red 4 symbols indicate mutation positions within the motifs. The lines were drawn to emphasize the differences between motifs with varying positions of the mutated G and to not assume a continuum of MMF values.

(B) Comparison of MMF (all 104 motifs combined) between samples with *POLE* driver mutations (n = 44 samples × 104 motifs = 4,576), *POLE* VUSs (n = 3,120), and no *POLE* or *POLD1* mutations (n = 30,368) (Kruskal-Wallis, $p < 0.001$). Each dot displays MMF value for an individual mutational motif. Mean MMF was 51.9×10^{-6} for samples with *POLE* drivers, 5.14×10^{-6} for samples with *POLE* VUSs, and 0.30×10^{-6} for samples with no *POLE* or *POLD1* mutations.

(C) Contour plot of the mean MMF among samples with *POLE* drivers showing MMF dependence on the position of mutated site within a motif.
(D) The mean MMF for motifs with two or three purines after the mutation in tumors with different *POLE* driver alleles. MMF values were normalized by the number of G>T transversions in each tumor. Error bars indicate 95% CI.
(E) Comparison of MMF for motifs with two or three purines after the mutation in different categories of tumors in regard to *POLE* and MSI statuses: "*POLE* drivers, MSS" (n = 851), "*POLE* drivers, MSI" (n = 161), "no *POLE/POLD1* mutations, MSS" (n = 4,508), and "no *POLE/POLD1* mutations, MSI" (n = 2,208). Each dot displays the MMF value for an individual mutational motif. Horizontal lines and stars indicate statistical associations between groups. *$p < 0.05$; ***$p < 0.001$; ns, no significance.

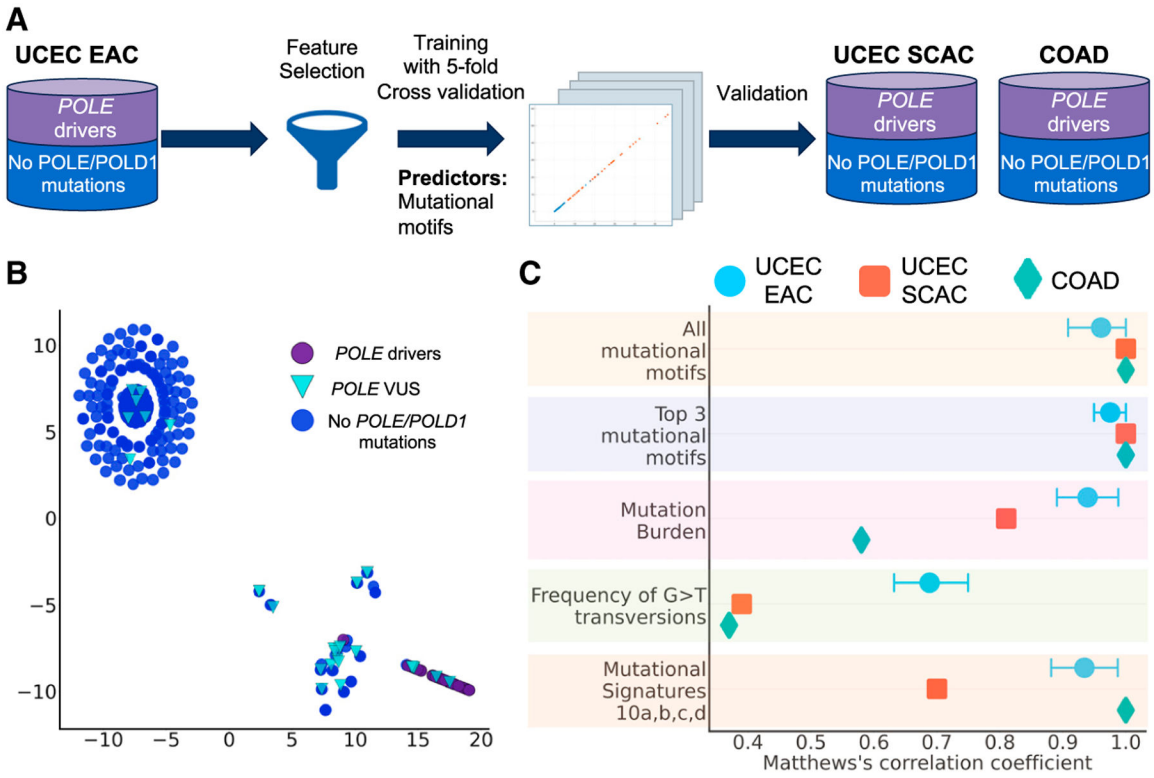


Figure 2. Classification workflow

(A) Workflow for the development of tumor classification model.

(B) t-SNE plot visualizing different clusters of tumors with *POLE* drivers and tumors with no *POLE*/*POLD1* mutations from the UCEC EAC cohort using three top-ranked motifs.

(C) Classification accuracy based on different features for the training dataset (UCEC EAC) and for two independent datasets (UCEC SCAC and COAD). Error bars show 95% CIs.

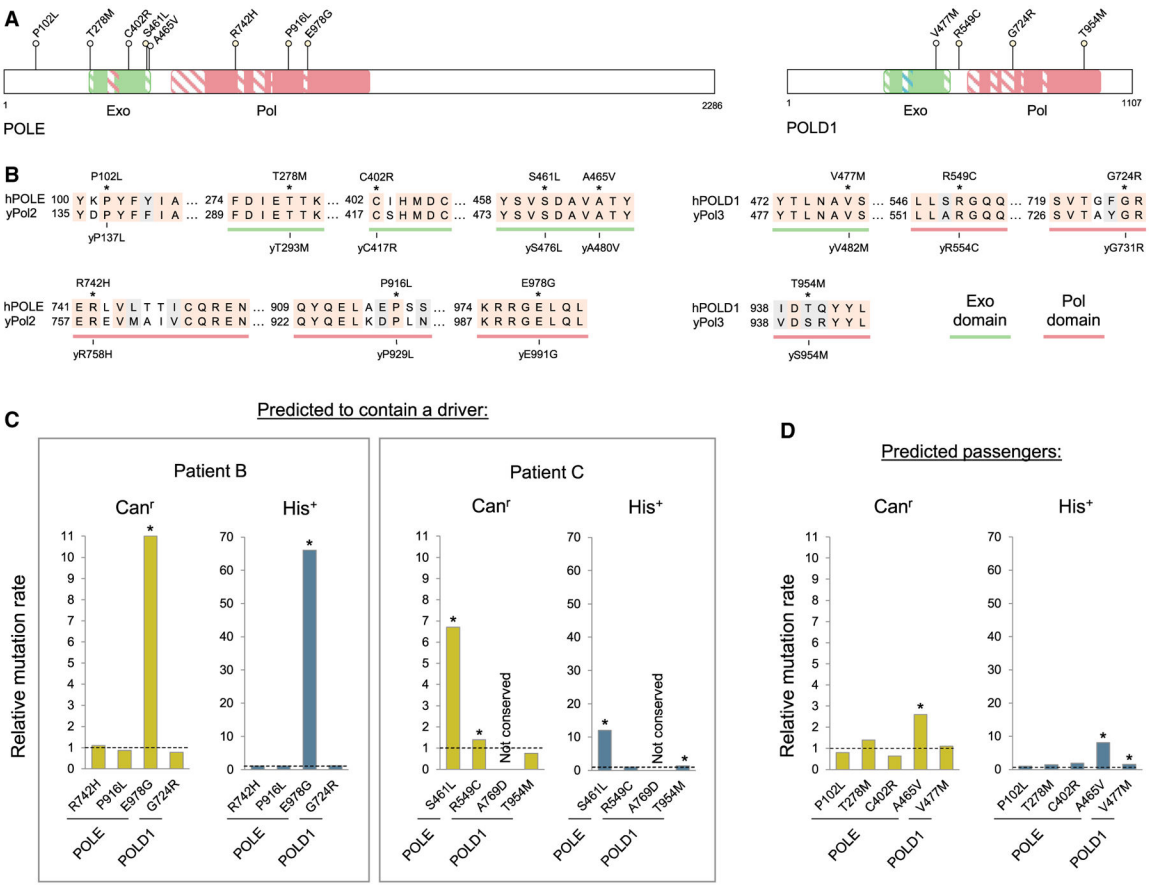


Figure 3. Validation of novel POLE drivers by functional analysis in yeast

(A) Schematic representation of human POLE and POLD1 showing the location of VUSs tested in this work. Exo, exonuclease domain; Pol, DNA polymerase domain. Striped boxes indicate conserved exonuclease and DNA polymerase motifs.

(B) Alignment of amino acid sequences of human and yeast proteins around the mutation sites. Human VUSs and analogous yeast variants are shown above and below the alignments, respectively.

(C) Identification of mutator variants in tumors with multiple VUSs predicted to carry driver mutations. Mutation rate relative to wild type is shown for yeast strains with *pol2* or *pol3* mutations mimicking the indicated human variants.

(D) VUSs from tumors not predicted to contain driver mutation confer mild or no mutator effects in yeast.

In (C) and (D), the *pol3-G731R* and *pol2-C417R* haploid strains were inviable, and the mutator effect is shown for heterozygous diploids. All other data are for haploid strains carrying the indicated alleles as the sole source of Pol ε or Pol δ. Data are from Tables S9 and S10. Dashed lines show wild-type mutagenesis levels. Asterisks indicate $p < 0.05$ by Wilcoxon-Mann-Whitney compared to the wild-type strain (null hypothesis: $MR_{mutant} > MR_{wild\text{-}type}$).

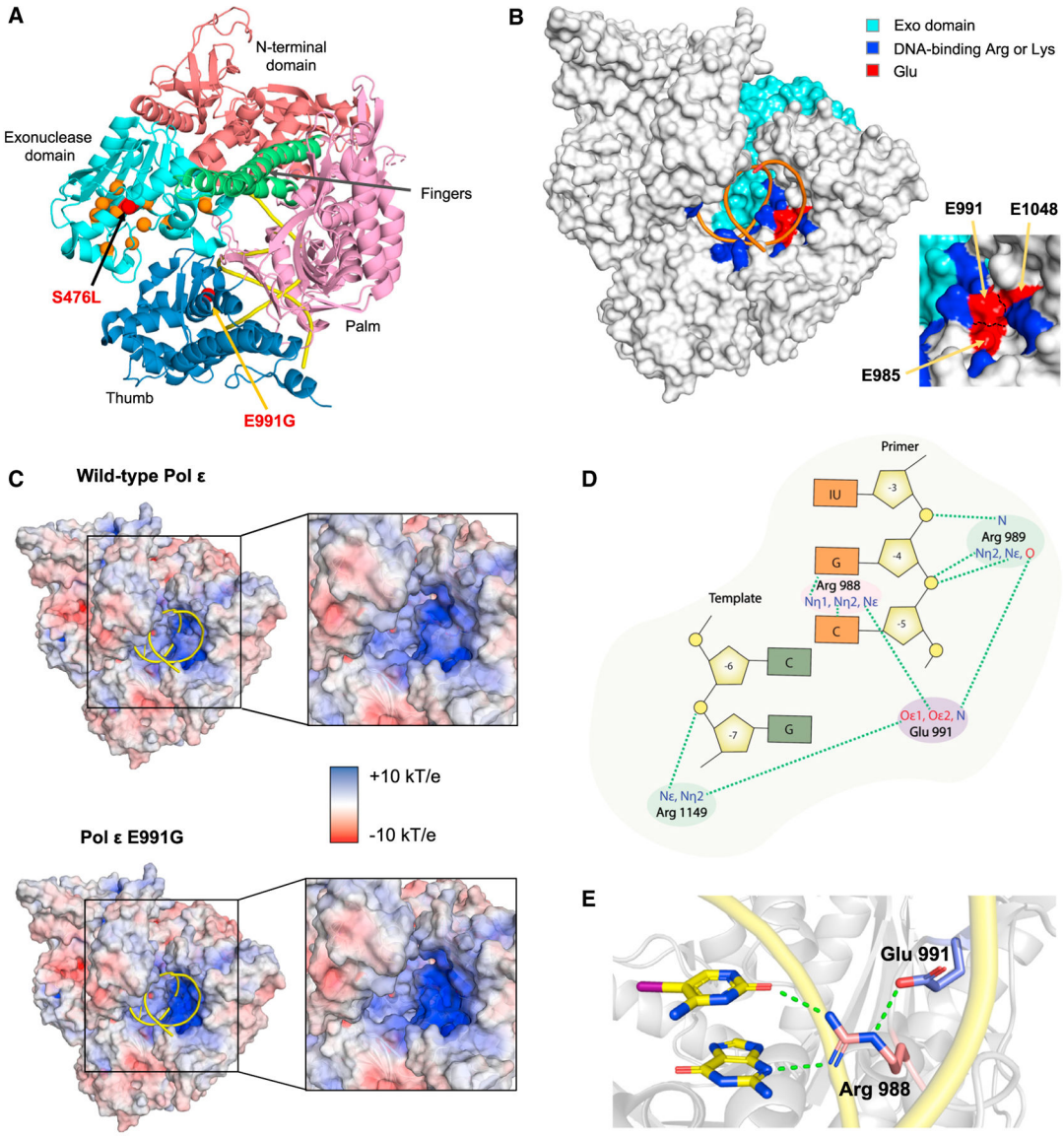


Figure 4. Structural implications of novel POLE driver mutations

(A) Locations of the previously known (orange balls) and new (red balls) driver mutations on the structure of the catalytic core of yeast Pol ε (PDB: 4M8O[52]). Note the location of the E991G substitution (human E978G) in the thumb subdomain of the DNA polymerase domain. DNA with the primer terminus in the polymerase active site is in yellow.

(B) E991, E985, and E1048 form a cluster of negatively charged residues at the DNA-binding interface of yeast Pol ε. Surface representation of the catalytic core is shown with the DNA cartoon in orange. Exonuclease domain (cyan) is in the back, partially visible through the cleft in the polymerase domain. DNA-binding arginine and lysine residues are in blue, and the three glutamate residues are in red. A close-up view of the glutamate cluster is shown on the right.

(C) The E991G substitution is predicted to increase the positive charge of the DNA-binding interface in the polymerase domain. Red and blue colors display negative and

positive electrostatic potentials, respectively. Electrostatic potentials were calculated with the Adaptive Poisson-Boltzmann Solver method[57] and mapped onto the molecular surfaces of wild-type Pol ε (top) and Pol ε E991G (bottom). The mutagenesis wizard of PyMOL was applied to generate the E991G variant. The color intensity was scaled with electrostatic potential values of the surface.

(D) Schematic representation of contacts between E991 and three DNA-binding arginine residues in yeast Pol ε. A −1 nucleotide position corresponds to the last base pair in the polymerase active site. See the main text for details.

(E) A close-up view of the interaction between E991, R988, and DNA bases at −4 and −5 positions of the primer. Yellow rods show the position of the DNA backbone.

Table 1.

Endometrial adenocarcinoma samples predicted by polypurine motif classifier to contain novel *POLE* driver mutations

Patient	Sample name	*POLE* mutations	*POLD1* mutations
A	TCGA-AP-A1DK-01A-11D-A135-09	T1228N	L606M - driver mutation
		Q1335Ter	
B	TCGA-AX-A1CE-01A-11D-A135-09	P338P	G724R
		G388G	S758S
		R742H	
		P916L	
		E978G	
		D2013D	
C	TCGA-AX-A2HD-01A-21D-A17D-09	S461L	R549C
		T998T	A769D
		N1971N	T954M
		A2006A	
D	TCGA-B5-A1MR-01A-31D-A14G-09	R705W	
E	TCGA-B5-A1MX-01A-11D-A142-09	V1227D	Q51H
			S478N - driver mutation
F	TCGA-DF-A2KN-01A-11D-A17W-09	K1070N	D316G - driver mutation

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA somatic mutation calls	NIH Genomic Data Commons	https://portal.gdc.cancer.gov [portal.gdc.cancer.gov]
TCGA clinical and survival data	NIH Genomic Data Commons	https://portal.gdc.cancer.gov [portal.gdc.cancer.gov]
Genome-wide maps of nucleosome occupancy in human lymphoblastoid cell lines (GSE36979)	Gaffney et al., 2012[81] Gene Expression Omnibus NCBI database	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36979 [ncbi.nlm.nih.gov])
Experimental models: Organisms/strains		
Saccharomyces cerevisiae strain PSD93	Daee et al., 2010[23]	PSD93
Oligonucleotides		
Oligonucleotides for site-directed mutagenesis	IDT	See Table S12
Software and algorithms		
MutaGene python package	Brown et al., 2019[82]	https://github.com/Panchenko-Lab/mutagene
MSIpred python package	Wang and Liang, 2018[83]	https://github.com/wangc29/MSIpred
TCGA biolinks R package	Colaprico et al., 2016[84]	https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
MFeaST MATLAB package	Gerolami et al., 2022[85]	https://www.renwicklab.com/molecular-feast/
BioMart R package	Durinck et al., 2009[86]	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
SigProfilerAssignment python package	Diaz-Gay et al., 2023[87]	https://github.com/AlexandrovLab/SigProfilerAssignment