TARGET: IJMS (MDPI), https://www.mdpi.com/journal/ijms/special_issues/genome_cancer

DNA methylation, deamination and translesion synthesis combine to generate footprint mutations in cancer driver genes in B-cell derived lymphomas and other cancers

Igor B. Rogozin[1], Abiel Roche-Lima[2], Kathrin Tyrishkin[3], Kelvin Carrasquillo-Carrión[4], Artem G. Lada[5], Lennard Y. Polikov[6], Vyacheslav Yurchenko[6], David N. Cooper[7], Anna R. Panchenko[3], Youri I. Pavlov[8,9]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; rogozin@ncbi.nlm.nih.gov
[2]Center for Collaborative Research in Health Disparities – RCMI Program, University of Puerto Rico, San Juan, Puerto Rico; abiel.roche@upr.edu
[3]Department of Pathology and Molecular Medicine, School of Medicine, Queen's University, ON, Canada; kt40@queensu.ca; anna.panchenko@queensu.ca
[4]Integrated Informatics Services core – RCMI, University of Puerto Rico, San Juan, Puerto Rico; kelvin.carrasquillo@upr.edu
[5]Department Microbiology and Molecular Genetics, University of California, Davis, CA, USA;
[6]Life Science Research Centre, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic; lypoliakov@gmail.com, vyacheslav.yurchenko@osu.cz;
[7]Institute of Medical Genetics, Cardiff University, Cardiff, UK; CooperDN@cardiff.ac.uk
[8]Eppley Institute for Research in Cancer and Allied Diseases, Omaha, NE, USA;
[9]Departments of Microbiology and Pathology; Biochemistry and Molecular Biology; Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA; ypavlov@unmc.edu

Abstract

Cancer genomes harbor numerous ~~of~~ genomic alterations. Indeed, m~~M~~any cancers accumulate thousands of nucleotide sequence variations ~~number~~. A prominent fraction of these mutations arise as a consequence of the off-target activity of DNA/RNA editing cytosine deaminases ~~and~~ followed by the replication/repair of edited sites by DNA polymerases, as deduced by the analysis of the DNA sequence context of the mutations ~~in cancers~~. Here, we have used the weight matrix (sequence profile) approach ~~for the~~to analyse~~is of~~ the mutagenesis caused by Activation Induced Deaminase~~AID~~ and two error-prone DNA polymerases. Control experiments using shuffled weight matrices and somatic mutations in immunoglobulin genes confirmed ~~a~~ the~~high~~ power of the weight matrix method. Analysis of somatic mutations in various cancers suggested that AID and DNA polymerases η and θ [IGOR: both? together?] generate mutations in contexts that almost universally ~~correlate with context of~~ [IGOR: implicate? point to? match?] somatic mutations in A:T and C:G sites. Analysis of methylation data ~~in~~ from malignant lymphomas (the MALY-DE dataset) suggested that driver genes are ~~likely to have properties of~~subject to a different (de)methylation process~~es different from~~ passenger genes [IGOR: 'driver and passenger mutations' are OK. 'Driver genes' is OK. But I really don't like 'passenger genes' How about 'non-driver genes'?]. This may reflect functional ~~importance of~~ interplay between somatic mutagenesis and (de)methylation processes.

Keyword: tumor cells, frequency matrices, database, computational biology, somatic hypermutation, immunoglobulin genes

Introduction

Epigenetic reprogramming in cancer genomes creates a distinct DNA methylation landscape encompassing clustered sites of hypermethylation at regulatory regions and protein-coding genes separated by large intergenic tracks of hypomethylated regions. This DNA methylation landscape is displayed by most cancer types, ~~thus~~ and hence may ~~even~~ serve as a universal cancer biomarker (PMID: 30514834). Most previous research has focused on the biological consequences of DNA methylation changes whereas its impact on DNA physicochemical properties [IGOR: what does this mean?]remains unexplored (PMID: 30514834). Sina *et al.* examine the effect of levels and genomic distribution of methylcytosines on the physicochemical properties of DNA to detect methylation landscape biomarkers[IGOR: what does this mean?]. Quick and selective electrochemical or colorimetric assays for the detection of cancer were developed (PMID: 30514834).

Another prominent feature of cancer initiation and progression are genomic alterations. Cancer genomes harbor numerous ~~of~~ genomic alterations (PMID: 28498882). Many cancers accumulate ~~hundreds/~~thousands of nucleotide sequence variations. A prominent fraction of these mutations arises as a consequence of the off-target activity of DNA/RNA editing cytosine deaminases and the replication/repair of edited sites by DNA polymerases, as deduced by the analysis of the DNA sequence context of mutations in cancers. Analyses of various types of cancer~~s~~ using classification approaches produced many mutation signatures and suggested that there are many mechanisms of hypermutation in cancer cells (PMID: 28472504,28498882).

There are associations between DNA methylation and genomic alterations. CpG sites are known to be hypermutable in both cancer and normal cells (PMID: 3338800,28472504,28498882). For example, recently we observed a substantial excess of mutations within a novel hybrid nucleotide motif [IGOR: what does this mean?]: the signature of somatic hypermutation (SHM) enzyme, Activation Induced Deaminase (AID), which overlaps the CpG methylation site (PMID: 27924834). This finding implies that in many cancers the SHM-like machinery acts at genomic sites containing methylated cytosine (PMID: 27924834). We identified the prevalence of this hybrid mutational signature[IGOR: what does this mean? Is this a novel concept?] in many other types of human cancer, suggesting that AID-mediated, CpG-methylation dependent mutagenesis is a common feature of tumorigenesis connecting methylation and hypermutation (PMID: 27924834).

Another prominent feature of carcinogenesis is the presence of cancer driver and passenger mutations.

A driver ~~is a~~ mutation ~~that~~ directly or indirectly confers a selective advantage on the cell in which it occurs, while a passenger ~~is a~~ mutation ~~that~~ does not exert~~s no~~ any selective growth advantage on the cell in which it occurs (PMID: 19360079). There is a difference between a driver gene and a driver gene mutation: a driver gene harbors recurrent driver mutations but may also harbor recurrent passenger gene [IGOR: Please try to avoid the term ‘passenger gene’] mutations (PMID: 19360079). In addition, some genes contain only recurrent passenger mutations with frequencies comparable to driver genes (PMID: 28498882). In this study we operationally defined a passenger gene[IGOR: Please try to avoid the term ‘passenger gene’] as a gene that contains numerous mutations that are classified as passenger mutations according to various computational tools.

We attempted to study an association of mutable motifs ~~produced~~ generated by the combined action of AID and two error-prone DNA polymerases and the methylation status in sets of driver and passenger genes. The conventional method used for the analysis of mutable DNA motifs is the consensus approach, for example, 5’WRC for the AID enzyme (W=A or T, R = A or G, the mutable position is underlined; PMID: 28498882) or 5’WA for DNA pol eta (PMID: 29139326). Here, we applied the frequently used weight matrix (sequence profile) approach (PMID: 30759888) to the analysis of methylation profiles and mutagenesis caused by AID and two error-prone DNA polymerases in CpG dinucleotides. Control experiments using shuffled sites and somatic mutations in immunoglobulin genes suggested that the weight matrix method is a useful approach to study mutagenesis. Analysis of somatic mutations in various cancers suggested that AID and DNA polymerase~~s~~ η mutable motifs ~~are~~ almost universally correlate with somatic mutations in C:G sites. Analysis of mutations and motifs in A:T sites produced similar results for pol η. Analysis of methylation data in malignant lymphomas (the MALY-DE dataset) suggested that driver genes are likely to have properties of (de)methylation processes different from passenger genes.

Results

1. Weight matrices are powerful descriptors of mutable motifs

Application of weight matrices is a novel technique to describe mutable motifs [IGOR: what do you mean? The frequency and distribution? Sequence context?] (PMID: 30759888). It was shown to be a robust and precise technique to describe AID/APOBEC mutable motifs in cancer cells. Briefly, weight matrices include information on a frequency of A, T, G, C bases in each of the ten positions surrounding detected sites of mutation (5 bases downstream and 5 bases upstream). Weight matrices were shown to be good descriptors of so-called mutable motifs, we studied AID/APOBEC enzymes using this technique (PMID: 30759888). AID and DNA pol η are involved in somatic hypermutation (SHM) in immunoglobulin (Ig) genes) (PMID: 11554790). It was also suggested that pol θ is involved in SHM (PMID: 18503084). Thus, we decided to derive weight matrices for both DNA polymerases. It should be noted that previously we derived weight matrices using collections of mutations in yeast genomes (PMID: 30759888). For human DNA polymerases eta and theta such collections are not available. Thus, we used a collection of mutations obtained by means of *in vitro* experiments for human pol eta and theta (PMID: 11554790, 11376340,) (Supplementary Figures S1 and S2).

Matrices of nucleotide frequencies are shown in the Figure 1. DNA polymerases eta and theta exhibit substantial variability in terms of their mutable motifs (Figure 1). W (A or T) or A in position –1 (Figure 1) was the most prominent feature of A:T mutations produced by pol eta and theta, accordingly. This is consistent with previous studies (PMID: 11554790, 18503084). An interesting feature of DNA polymerase theta is an elevated frequency of C in the position –1 for mutations in C:G positions (Figure 1). Thus, pol theta tends to produce more errors in CpG dinucleotides. This may indicate this DNA polymerase is involved in methylation/demethylation of CpG dinucleotide although this hypothesis requires further analyses. Although the pol eta tends to produce less number of mutations in the CpG context (Figure 1A), it is hard to demarcate the mutational signature of this DNA polymerase using the consensus approach due to the high variability of information content across sites (Figure 1). Thus, the weight matrix approach is likely to be more objective way to describe mutable motifs.

(A)

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	50	58	58	35	52	0	42	32	56	58	59
T	59	73	39	70	55	0	58	52	86	56	41
G	47	54	52	43	41	224	**28**	58	49	70	60
C	68	39	75	76	76	0	96	82	33	40	64
						G	*H(?)*				

(B)

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	108	121	108	80	**125**	**388**	111	63	122	103	87
T	91	65	54	113	**107**	0	131	101	85	67	89
G	70	91	116	65	83	0	41	101	89	72	104
C	119	111	110	130	73	0	105	123	92	146	108
					W	**A**					

(C)

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	16	23	15	19	7	0	11	8	23	16	12
T	14	15	25	24	8	0	14	26	16	18	12
G	19	21	20	20	14	**69**	15	15	11	18	22
C	20	10	9	6	**40**	0	29	20	19	17	23
				B(?)	*C*	*G*					

(D)

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	16	40	29	33	**65**	**139**	35	17	21	22	30
T	40	27	26	43	12	0	38	35	44	39	25
G	43	49	63	41	31	0	36	42	36	45	48
C	40	23	21	22	31	0	30	45	38	33	36
					A	**A**					

Figure 1. Nucleotide frequency matrices for DNA polymerases eta (A) G:C sites; B) - A:T sites) and theta (C - G:C sites; D - A:T sites). Raw numbers of nucleotides are shown. Known mutable motifs (consensus sequences) are shown below each matrix in bold, mutable positions are underlined. Putative mutable motifs are italicized, W = A or T, B = A, T or G, H = A, T or C. [IGOR: Unclear what these data are or where they are from]

Next, we compared the nucleotide composition of mutation sites (±5 nucleotides, Figure 1) for DNA polymerases eta and theta using the χ^2 test. We found that these DNA polymerases were significantly different with respect to the DNA sequence context of mutation sites expressed in the form of nucleotide frequency matrices (A:T sites: χ^2 = 155.0, df = 40, P =1.9 x 10^{-15}; G:C sites: χ^2 = 82.2, df = 40, P = 0.00007). Thus, DNA polymerases eta and theta have different properties of the DNA sequence context of mutations and can be used as informative descriptors of pol eta/theta mutable motifs.

2. Pol eta and pol theta weight matrices across various cancers

Previously we demonstrated using the consensus approach that AID is likely to be involved in demethylation of CpG dinucleotides in follicular lymphomas and many other cancers (PMID: 27924834). In another paper we put forward a hypothesis that pol eta may be also involved in methylation/demethylation of CpG dinucleotides in cancer cells (PMID: 29139326). [IGOR: would it be worth introducing this in the general context of the different mechanisms for demethylation of eukaryotic genomes?]The weight matrix approach and the MALY-DE datasets (CpG methylation spectra and somatic mutations, see Materials and Methods) allow us to test these hypotheses.

We examined the correlation between the nucleotide context of somatic mutations in cancers and two studied [IGOR: in vitro mutational spectra?] DNA polymerases mutable motifs. A correlation between a mutable motif and the DNA context of somatic mutations from the COSMIC database was stated [IGOR: adduced? assumed? inferred?] when the results of two statistical tests (Monte Carlo test and *t*-test, see Materials and Methods) were both significant. AID was already studied (PMID: 30759888), it was shown that it is the most ubiquitous enzyme according to its characteristic signature (the AID weight matrix) in various cancer types (PMID: 30759888).

Analysis of DNA polymerase~~s~~-induced mutations in C:G sites suggested that both mutable signatures ~~are~~ almost universally correlate with the nucleotide context of somatic mutations in C:G sites (Figure 2). However, analysis of mutations and motifs in A:T sites revealed correlation for pol eta only (Figure 2). Only for a few cancers a significant correlation with pol theta was found (Figure 2). Such discrepancy (why we assume that pol theta should behave as pol eta?) for pol theta is likely to be explained by the presence of the CG motif that we noticed before the explanation is not very clear [IGOR: meaning unclear!] (Figure 1). The CG motif is known to be a prominent feature of somatic mutations in cancer, and this may be the reason for the discrepancy. Most likely, pol theta may be active [IGOR: isn’t it active in all cells?] in some types of cancer. Pol eta is likely to be one of the most ubiquitous enzymes according to its characteristic signature (the weight matrix) in various cancer types, this is consistent with our previous study (PMID: 29139326).

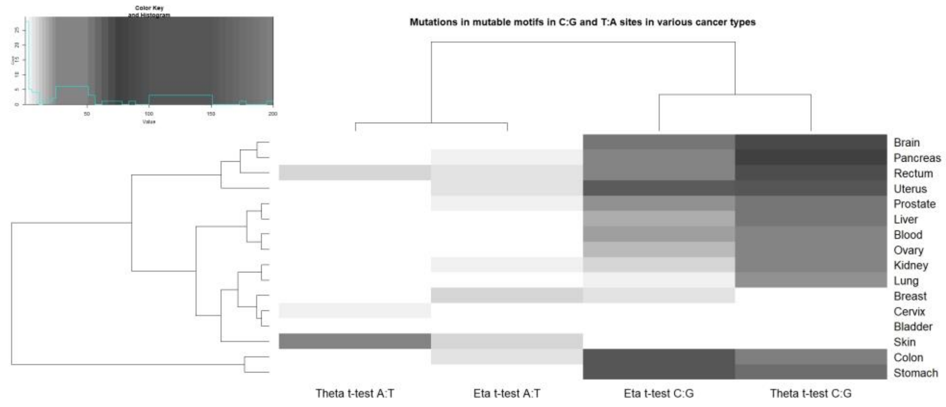

Figure 2. Correlation pol eta/theta mutable motifs and the sequence context of somatic mutations. For the actual data, see Supplementary Tables S1 and S2. The intensities of the gray color correspond to the *t*-test values (the ratio being the *t*-test value of the mutated sites divided by the mean weight of the non-mutated sites). The unweighted pair group method with arithmetic mean (UPGMA) clustering of ratio values for the pol eta/theta footprints and tissues is shown[IGOR: origin of data?]

3. Control experiments

In vitro collections of mutations that were used to reconstruct weight matrices for DNA polymerases eta and theta (PMID: 11554790, 18503084) are relatively small (Figure 1); thus control experiments are important for derived weight matrices. Previously we demonstrated that analysing the~~s of~~ correlation between the matrices of shuffled sites of mutations and the nucleotide context of somatic

Formatted: Font: Italic

mutation in various cancer cell types is a reliable approach to estimate the impact of false positives (PMID: 30759888). Analysis of 16 types of cancer (Supplementary Table S3) suggested that the AID weight matrix is less prone to false positives compared to pol eta / pol theta (Supplementary Table S3). Only a few types of cancer~~s~~ have a low level of false negatives. Fortunately, for our study of MALY-DE sets, “Blood” tissue, GCB lymphomas (from the COSMIC database) and MALY_DE malignant lymphomas have extremely low rate of false positives (Supplementary Table S3). Therefore, we decided to use the derived matrices for further analysis of the MALY-DE datasets.

Analysis of somatic mutations in ~~Ig~~ immunoglobulin (Ig) genes can be used to estimate the rate of ~~of~~ false negatives because mutations in human Ig genes are known to be associated with AID and pol eta mutable motifs (PMID: 11554790). Thus, these mutations can be used as a control set. Indeed, a significant association between the AID mutable motif and mutations was found in all three studied sets of somatic mutations (PMID: 9671757, 15944281) (Table 1), confirming that the AID weight matrix is a reliable descriptor of AID-induced mutagenesis. The Pol eta weight matrices revealed a significant association for all studied cases except XPV [IGOR: Xeroderma pigmentosum?] patients where pol eta is inactive (Table 1) (PMID: 15944281). Pol theta did not yield significant results for some studied cases (Table 1), ~~this~~ which is consistent with the hypothesis that pol theta is also involved in SHM (PMID: 18503084). The results of both control experiments suggested that the weight matrix technique approach is adequate to study~~ied~~ DNA polymerase~~s~~ mutational spectra.

Table 1. Correlation between the sequence context of somatic mutations and mutable motifs in fragments of human immunoglobulin genes.

Locus	Test	Number of Mutations	AID / G:C	Pol η / G:C	Pol θ / G:C	Number of Mutations	Pol η / A:T	Pol θ / A:T
	Ratio	583	1.208	1.027	1.091		1.082	0.979
VH26	*t*-test		**13.1***	NSE	**5.9***	351	**5.3***	NSE
	MC test		<0.001	0.004	<0.001		<0.001	0.699
JH4 intron,	Ratio	177	1.341	1.050	1.029		1.041	1.032
control	*t*-test		**12.3***	**2.8***	NSE	95	**2.4***	**2.2***
individuals	MC test		<0.001	0.002	0.106		0.004	0.011
JH4 intron,	Ratio	227	1.278	1.009	1.011		0.957	0.980
XP-V	*t*-test		**9.9***	NSE	NSE	25	NSE	NSE
patients	MC test		<0.001	0.329	0.061		0.776	0.670

NSE (no significant excess) indicates the absence of a significant excess of mutations in mutable motifs suggesting there to be no association between mutagenesis and motifs. The significance of any excess was measured using the Student *t* and Monte Carlo (MC) tests. The asterisk (*) denotes that the corresponding $P < 0.01$; this is a conservative estimate of the critical overall value of the *t*-test having allowed for multiple testing by means of the Bonferroni correction (5 comparisons). “Ratio” is the mean weight of mutated sites divided by the mean weight of non-mutated sites.

4. Analysis of driver/passenger genes

Analyses of driver/passenger mutations and genes are known to be powerful approach in cancer genomics and even can be diagnostic~~s~~ of various cancers (PMID: 32015527, 28472504; 31034466; 31202631). We derived lists of driver and passenger genes [IGOR: please try to avoid!]using three approaches (for details see Materials and Methods). Final lists of genes are shown in the Supplementary Tables S4 and S5 (we used the ENSEMBL IDs as recommended by the DAVID web site, https://david.ncifcrf.gov/). The total number of driver [IGOR: how do you define your 'driver genes' in practical terms? Do you use the COSMIC list? https://cancer.sanger.ac.uk/census#cl_search] and passenger genes is 134 genes and 210 genes, accordingly. We performed pathway/keywords enrichment analyses (PMID: 24480647, 25243088) using the DAVID web site. Results are shown in the Supplementary Table S4. Keywords "methylation", "nuclear chromatin" and numerous pathways/terms associated with various types of cancer are consistent with properties of GCB lymphomas (PMID: 27924834). The KEGG pathway "pathways in cancer" (P = 0.025) is another important descriptor of the driver gene list (Supplementary Table S6.) In general, the driver gene set appears to be highly informative and contains many features expected for cancer-related genes. ~~In~~ By contrast, analysis of passenger genes did not produce many significant results (Supplementary Table S6).

Analysis of association between mutable motifs and somatic mutations detected an interesting difference between driver and passenger genes: mutable motifs of pol eta and theta do not correlate with somatic mutations in driver genes whereas mutable motifs of pol eta and theta correlate with somatic mutations in passenger genes (Table 2). Correlation of the pol theta mutable motif with mutations in G:C sites of driver genes can be explained to some extent by the presence of CpG consensus sequence in the pol theta mutable motif (Figure 1), this dinucleotide is known be mutable in many cancers (PMID: 28472504,28498882). An important feature of driver and passenger genes is substantially higher frequency of mutations in G:C nucleotides compared to all genes [IGOR: meaning unclear! Are there genes which are not driver or passenger genes?] (Table 2), this may be explained by an important role of AID in somatic mutagenesis of driver and passenger genes.

Table 2. Correlation between mutable motifs and the sequence context of somatic mutations in driver and passenger genes.

Group of genes	**Test**	**Number of Mutations**	**AID / G:C**	**Pol η / G:C**	**Pol θ / G:C**	**Number of Mutations**	**Pol η / A:T**	**Pol θ / A:T**
All genes	Ratio	137775	1.021	1.005	1.091	145768	0.992	1.011
	t-test		**23.4***	**7.2***	**23.0***		NSE	**15.8***
	MC test		<0.001	0.055	<0.001		1.000	<0.001
Drivers	Ratio	4246	1.107	1.001	1.007	3918	0.980	1.032
	t-test		**20.0***	NSE	NSE		NSE	**7.8***
	MC test		<0.001	0.346	0.037		1.000	<0.001
Passengers	Ratio	3553	1.079	1.059	1.057	2793	0.995	1.045
	t-test		**14.2***	**13.8***	**11.7***		NSE	**8.9***
	MC test		<0.001	<0.001	<0.001		0.874	<0.001

NSE (no significant excess) indicates the absence of a significant excess of mutations in mutable motifs suggesting there to be no association between mutagenesis and motifs. The significance of any excess was measured using the Student *t* and Monte Carlo (MC) tests. The asterisk (*) denotes that the corresponding $P < 0.01$; this is a conservative estimate of the critical overall value of the *t*-test having allowed for multiple testing by means of the Bonferroni correction (5 comparisons). "Ratio" is the mean weight of mutated sites divided by the mean weight of non-mutated sites.

5. Analysis of DNA methylation patterns of driver/passenger genes using weight matrices

The a~~A~~verage methylation level of driver and passenger genes was found to be approximately the same: ~78% for both sets of genes. Analysis of methylation in mutable motifs was performed using the threshold methylation values 25 and 75. For the threshold value 25%, average weights of AID mutable motifs for driver genes smaller and greater than 25% are 57.1 and 56.7, accordingly. The ratio is 1.025 (57.8/56.4 = 1.025) (Table 4). This difference is statistically significant, albeit ~~is ~~subtle (Table 4). Average weights of AID mutable motifs for passenger genes below and above the threshold 25% are 57.8 and 56.2, accordingly. The ratio is 1.027, this difference is also statistically significant (Table 4). These results suggest that “stronger” AID mutable motifs are associated with lower methylation levels in driver and passenger genes. Somewhat different results were obtained for pol eta and theta, differences are not significant for both driver and passenger genes (Table 4). These results suggest that these DNA polymerases are unlikely to influence the global level of methylation in driver and passenger genes for the threshold level = 25%.

For the threshold value 75% we observed the opposite trend. For example, average weights of AID mutable motifs for driver genes greater and smaller than 75% are 56.9 and 56.7, accordingly. The ratio is 1.004 (56.9/56.7 = 1.004) (Table 4). This difference is not statistically significant (Table 4). The ratio is also low for the passenger gene set although it is significant (Table 4). However, mutable motifs for both studied DNA polymerases seem to be associated with the methylation level for this threshold. These results suggest that these DNA polymerases may influence the global level of methylation in driver and passenger genes for the threshold level [IGOR: Is it not more likely the other way around?]= 75% (heavily methylated positions).

Table 3. Levels of methylation in CpG sites associated with mutable motifs, the threshold value = 25%.

Group of genes	Number of CpG sites below and above the threshold	Tests	AID	Pol η	Pol θ
Driver	2867 149480	Ratio *t*-test MC test	1.025 **3.2*** <0.001	0.997 NSE 0.772	0.994 NSE 0.950
Passenger	5558 239220	Ratio *t*-test MC test	1.027 **5.4*** <0.001	0.993 NSE 0.989	0.985 NSE 0.989

Table 4. Levels of methylation in CpG sites associated with mutable motifs, the threshold value = 75%.

Group of genes	Number of CpG sites above and below the threshold	Tests	AID	Pol η	Pol θ
Driver	96917 51290	Ratio *t*-test MC test	1.004 NSE 0.433	1.009 **7.9*** <0.001	1.021 **20.4*** <0.001

Passenger	155205 89573	Ratio *t*-test MC test	1.007 **4.5*** <0.001	1.009 **9.8*** <0.001	1.023 **28.6*** <0.001

6. Analysis of somatic mutations in CpG sites of driver/passenger genes

We analyzed the level of methylation in CpG sites that coincide with positions of somatic mutations. It should be noted that the studied sets are small; however, they are still amendable to statistical analysis using the threshold =75% (Table 6). Unfortunately, the number of mutations for the threshold = 25% wa~~i~~s too small for statistical analyses, the number of sites with methylation levels below 25% is 0 and 3 for driver and passenger genes accordingly.

The first result is that the fraction of CpG sites below the threshold 75% (0.35, Table 5) and the fraction of mutation sites with the methylation level below the threshold 75% (0.17, Table 6) is dramatically different for driver genes. Thus, sites with somatic mutations in driver genes tend to have higher methylation values, this difference is statistically significant ($P < 0.001$ according to the Fisher exact test). The fraction of mutation sites with the methylation level below the threshold 75% is also different for driver and passenger genes (0.17 and 0.40, accordingly, Table 6), this suggests some differences in methylation/demethylation processes in driver and passenger genes.

The second interesting result is the significant correlation of AID, pol eta and pol theta with mutation positions having low methylation level (below 75%) (Table 6). For AID this is more pronounced for driver genes (Table 6). Pol eta seems to be involved in CpG mutagenesis for both sets of genes (Table 6). Pol theta is likely to be involved in mutagenesis as well (Table 6).

Table 5. Levels of methylation in positions of somatic mutations in CpG sites, the threshold value = 75.

Group of genes	**Number of mutations in CpGs sites <u>below</u> and <u>above</u> the threshold**	**Tests**	**AID**	**Pol η**	**Pol θ**
Driver	52 249	Ratio *t*-test MC test	1.111 **2.9*** 0.004	1.136 **7.8*** <0.001	1.046 NSE 0.035
Passenger	264 390	Ratio *t*-test MC test	1.015 NSE 0.222	1.125 **7.3*** <0.001	1.061 **3.7*** <0.001

Discussion

The advantage of the weight matrix approach is that it is a unified computational technique that allowed an objective and accurate comparison of the mutational contribution of various mutable enzymes under the same experimental conditions and for the same datasets. We confirm that while the mutational footprints of DNA polymerase eta and theta are prominent in some cancers, mutable motifs characteristic of the humoral immune response somatic hypermutation machine, AID, is likely to be the most widespread feature of somatic mutation spectra attributed to any enzyme in cancer genomes (PMID: 29139326,30759888). It is important to note that the suggested technique does not depend on expert opinion as to the exact consensus sequences, and therefore objectively represents mutable motifs.

A high rate of false positives for many types of cancer (Supplementary Table S3) is likely to be due to small datasets for DNA polymerase eta and theta (Figure 1). Larger sets of mutations are likely to improve the power of prediction. Still we can infer that some types of cancer including GCB lymphomas do not have a noticeable rate of false positives (Supplementary Table S3). We applied all weight matrices to study mutable motifs and methylation in the MALY-DE datasets. We demonstrated that mutable motifs are associated with CpG dinucleotides and their methylation status. Another problem is a small number of MALY-DE samples (26 samples), this may cause problems for prediction of driver and passenger mutations. These problems one of possible explanations why differences between driver and passenger genes are subtle (albeit significant) (Tables 2-5).

Sophisticated classification approaches have been developed to extract the most prominent signatures from a complex mix of mutational targets resulting from the action of a variety of mutagens, both exogenous and endogenous, operating during tumor evolution (PMID: 28472504,28498882). Both driver and passenger mutations have been used in the analysis without any attempt to separate them. In this study we analyzed these two sets separately. We detected significant differences in methylation/demythelation processes in driver and passenger genes (Tables 4-6). It is not that easy to interpret those differences because the role of methylated CpG dinucleotides in exons is not well understood (PMID: 28225755). It was suggested that changes in intragenic DNA methylation is important in several human diseases including syndromic and sporadic forms of autism that involve methylation defects, including Rett syndrome, Prader–Willi and Angelman syndromes, and others, suggested that differential methylation of genes may underlie one aspect of autism pathogenesis (PMID: 27974215; 29986017). Moreover, several studies of likely deleterious mutations and pathway enrichment have observed that genes controlling chromatin accessibility or remodeling (and hence gene expression) are enriched for genes with recurrent mutations (PMID: 25891009; 26402605; 28628100). The observed differences between driver and passenger genes may reflect such effects in gene expression triggered by cancer progression.

Methods

Mutable motif construction using weight matrices

Several approaches have been developed for the analysis of a set of mutated sequences (PMID: 6364039, 29139326,30759888). A mononucleotide weight matrix is a simple and straightforward way to present the structure of a functional signal and to calculate weights for the signal sequence. Each matrix includes information on a normalized frequency of A, T, G, C bases in each of the ten positions surrounding detected sites of mutation (5 bases downstream and 5 bases upstream). We calculated the weight matrices for the two studied DNA polymerases (Supplementary Figures S1 and S2).

A simple formula for W(b,j) was used for data analysis: W(b,j) = log2[f(b,j)/e(b)], where f(b,j) is the observed frequency of the nucleotide b in position j and e(b,j) is the expected frequency of the nucleotide b in position j calculated as the mean nucleotide frequencies of positions –5,-4, +4, +5 for sites of mutations in the yeast genome; the resulting W(b,i) matrices are shown in the Figure 1.

The matching score S(b1,...,bL) of a sequence b1,...,bL is:

$$S(b1,...,bL) = \underset{j=1,L}{e}\ W(b,j) \qquad (1)$$

The matching score between sequence b1,...,bL and a weight matrix can be further expressed as a percentage:

$$\%\ \text{matching score} = 100\ r\ (S(b1,...,bL) - Smin) / (Smax - Smin) \qquad (2)$$

$$Smin = \overset{L}{\underset{j=1}{e}}\ \underset{b}{\text{MIN}}\ W(b,j) \qquad Smax = \overset{L}{\underset{j=1}{e}}\ \underset{b}{\text{MAX}}\ W(b,j) \qquad (3)$$

Hereafter, we use the term “weight” instead of “% matching score”. We used the positions –3:+3 to estimate the weights of sites.

ICGC/TCGA Mutation datasets

Somatic mutation data from the ICGC and TCGA cancer genome projects were extracted from the Sanger COSMIC Whole Genome Project v75 (http://cancer.sanger.ac.uk/wgs). The ICGC/TCGA datasets are almost exclusively passenger mutations and they are unlikely to be subject to selection to promote cellular proliferation. Thus, they are more likely to reflect the unselected mutational spectra (PMID: 28472504,28498882). The tissues and cancer types were defined according to the primary tumor site and the cancer project in question (PMID: 28472504,28498882). We used collections of mutations obtained by means of *in vitro* experiments for human pol eta (PMID: 11554790) and pol theta (PMID: 185030 (Supplementary Figures S1 and S2) to build weight matrices.

Analysis of mutations

DNA sequences surrounding the mutated nucleotide represent the mutation context. We compared the frequency of known mutable motifs for somatic mutations with the frequency of these motifs in the vicinity of the mutated nucleotide. Specifically, for each base substitution, the 121 bp sequence centered at the mutation was extracted (the DNA neighborhood). We used only the nucleotides immediately flanking mutations because repair/replication enzymes are thought to scan a very limited region of DNA (PMID: 28472504,28498882). This approach does not exclude any specific area of the genome, but rather uses the areas within each sample where mutagenesis has occurred (taking into account the variability in mutation rates across the human genome), and then evaluates whether the mutagenesis in these samples were enriched for AID/APOBEC motifs (PMID: 29139326). This approach was thoroughly tested, and the high accuracy of the analysis was demonstrated (PMID: 29139326) . The mean weight of mutable motifs (Supplementary Figure S1) in the positions of somatic mutations was compared to the mean weight of the same motifs in the

DNA neighborhood using the t-test (2-tail test) and Monte Carlo test (MC, 1-tail test) similar to the consensus method as previously described (PMID: 29139326).

In addition to analyses of the derived mutational signatures in cancer genomes, we performed a control experiment: we randomly shuffled a dataset of sequences surrounding mutations in the studied target sequences (Supplementary Figure S1 and S2) keeping position 6 (the position of mutations) intact. Each sequence was shuffled separately; thus, the overall base composition and the base compositions of each sequence were the same. Weight matrices were derived from these shuffled sequences, the sampling procedure was repeated 1000 times.

Detection of driver/passenger genes

In this study we used two independent methods to predict the driver status of cancer mutations: MutaGene online package (PMID:28472504; 31034466) and Chasmplus (31202631). These methods showed the top performance on a recent benchmarking set (PMID: 31034466). MutaGene is a probabilistic approach which adjusts the number of mutation recurrences in patients by cancer-type specific background mutation model. The MutaGene driver mutation prediction method is not explicitly trained on any sets of mutations. The background models estimate the probability to obtain a nucleotide or codon substitution from the underlying processes of mutagenesis and repair that are devoid of cancer selection component affecting a specific genomic (or protein) site. We used two MutaGene background models: one was derived from the pan-cancer mutational data (“Pancancer” model in MutaGene) and another one was constructed directly from the MALY-DE mutational data since this cancer-~~-~~specific model was not present in the MutaGene database of background models. As a result, two ranking lists of driver mutations were produced for three types of mutations: missense, nonsense and silent. Chasmplus is a machine learning method which was trained using somatic mutations from TCGA. Since no cancer specific model was available for MALY-DE, we used pan-cancer predictions while running Chasmplus. Then we merged the predictions produced by the three different models/methods and reported only those mutations as drivers (highlighted in red) which were predicted as “drivers” or “potential drivers” by MutaGene and had a Chasmplus score cutoff larger than 0.5. In orange we highlighted those mutations which satisfied two of the above-mentioned criteria. Since Chasmplus does not produce predictions for nonsense and silent mutations, only predictions for missense mutations were reported. In addition, some mutations/genes were not reported by Chasmplus since it excluded them from the list of potential cancer driver genes.

Methylation data

For the analysis of the association between somatic mutations, mutable motifs and methylation, datasets for 26 patients with malignant lymphoma (https://dcc.icgc.org/projects/MALY-DE) were used. In the analyzed datasets, the data for all patients were pooled together. Each position is characterized by the methylated/unmethylated read count and the methylation ratio (the number of methylated reads divided by the total number of reads overlapping this position and multiplied by 100). Only positions with more than nine associated reads were included in the analysis.

Acknowledgements

This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health [to I.B.R.]; RCMI grant U54 MD007600 (National Institute on Minority Health and Health Disparities) from the National Institutes of Health [to A.R.-L.], the

European Regional Funds (CZ.02.1.01/16_019/0000759) and the Grant Agency of Czech Republic (18-15962S) (to V.Y.); NE DHHS LB506, grant 2017-48 [to Y.I.P.]; the Fred & Pamela Buffett Cancer Center Support Grant from the National Cancer Institute under award number P30 CA072720 [to Y.I.P.] and Qiagen Inc through a License Agreement with Cardiff University [involving~~to~~ D.N.C.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. ARP and KT were supported by the Department of Pathology and Molecular Medicine, Queen’s University, Canada. ARP 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.