BMC Genomics

RESEARCH

Open Access

Linking somatic mutations in cancer to the electronic properties of DNA

Benoît de Witte[1,2,3], Cyril Karamaoun[1,2], Pauline Hermans[1,2], Maxime Tarabichi[2,3], Fabrizio Pucci[1,2] and Marianne Rooman[1,2*]

Abstract

Oxidative stress, generated by both endogenous and exogenous agents, can cause DNA lesions that, if not repaired, accumulate as somatic mutations and can contribute to cancer initiation. Here, we explored this problem through the lens of DNA electronic properties, quantified by the vertical ionization potential (vIP) of nucleobase motifs, which reflects their susceptibility to oxidation. We analyzed genome-wide experimental data on oxidative DNA damage and found that the highest damage levels occur in regions with low vIP values, suggesting a causal link between them. The analysis of cancer mutational signatures and their annotated aetiologies revealed strong anticorrelations between mutation frequency and vIP values, particularly in cancers driven by oxidative DNA damage, such as lung cancer. We further computed anticorrelations between vIP values and the frequencies of mutated motifs across coding and non-coding regions and across different mutation types, observing the strongest anticorrelations for silent mutations, consistent with their reduced selective pressure. Moreover, similar anticorrelations were observed for somatic mutations in cancer and normal tissues, as well as for germline mutations, suggesting that they arise from similar mutagenesis processes. This work clarifies how oxidative damage, DNA electronic properties and carcinogenesis are related and help identify genomic regions more prone to mutations.

Keywords DNA nucleobase motifs, Vertical ionization potentials, Genome mutability, Carcinogenesis, Somatic mutations, Germline mutations, Cancer mutational signatures

Introduction

Cancer initiation and progression are driven by cellular alterations that primarily result from the progressive accumulation of somatic mutations in the genomes of normal tissues throughout an individual's lifetime. These genetic changes can disrupt normal cellular functions, ultimately leading to uncontrolled cell growth and division and, consequently, to tumor development.

But only some, called driver mutations, provide a selective advantage to the cells in which they occur, often by impairing the function of some key proteins that, for example, control the cell cycle or apoptosis [1]. These events facilitate the transition from a benign to a malignant lesion and promote genome instability. In contrast, passenger mutations also accumulate but do not confer any selective advantage to their host cells [2] and can in some cases even have slightly deleterious effects [3, 4].

Among the various types of mutations, single nucleotide substitutions (SBSs), in which one nucleobase is replaced by another, play a major role and we will focus

*Correspondence:
Marianne Rooman
Marianne.Rooman@ulb.be
[1]Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium
[2]Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium
[3]Institute of Interdisciplinary Research in human and molecular Biology, Université Libre de Bruxelles, Brussels, Belgium

on them in this work. SBSs are primarily caused by cellular exposure to a wide range of endogenous and exogenous agents and lie at the heart of carcinogenesis [5]. There are multiple mechanisms of SBS formation, which depend on a wide variety of mutagenesis pathways. One example is the oxidation of DNA bases by reactive oxygen species (ROS), which are generated as byproducts of cellular metabolism [6–8]. Guanine is oxidized to 8-oxo-7,8-dihydroguanine (8-oxoG) at a high rate in the genome, estimated at around 3,000 lesions per cell per day in unstressed conditions [9, 10]. If not detected in time by cellular DNA repair mechanisms, that are often inactivated in cancer cells [11], the G→8-oxoG oxidation can lead, after replication, to a G→T base transversion [12], which can then be fixed in the cell's genome. Other well-known examples of mutational processes are those related to cell exposure to tobacco smoke, which leads to an overrepresentation of C→A mutations in lung cancer genomes [13, 14], and to ultraviolet (UV) light, which results in an overrepresentation of C→T mutations in melanoma [15].

An attempt at a systematic classification of SBSs and their aetiologies has been made in the COSMIC database [16], where 95 mutational signatures from different cancer types have been identified and linked to distinct mutational processes. Nevertheless, the precise identification of their underlying mechanisms remains highly challenging, and about one-fourth of the COSMIC signatures still have an unknown aetiology [17].

The formation of SBSs has been linked to the presence of electron holes migrating along the nucleobase stacks of the double-stranded DNA molecules that make up the genome [18–21]. Indeed, the nucleobases, which are the aromatic moieties of the nucleotides arranged in parallel π–π stacks along each of the two strands, can be easily ionized by various physical and chemical agents. The resulting electron holes can migrate along the base stack, generally in the 5' to 3' direction without interstrand hopping. They remain localized in regions of low ionization potential, sometimes far from their original site of formation [18], and can ultimately trigger SBSs [20–24]. This mechanism is illustrated in Fig. 1.

The DNA ionization propensity and charge migration depend on the nucleobase sequence, with electron holes preferentially accumulating in guanine-rich regions [25, 26]. This preference arises because guanine has the lowest vertical ionization potential (vIP) among the DNA bases [19, 21, 27, 28]. In addition, DNA ionization propensity has been shown to depend, to a lesser extent, on the nucleobase stack conformation [27, 29]. Note that electronic properties of DNA play a fundamental role not only in mutagenesis processes but also in a wide range of physiological processes within the cell. For example, charge transfer appears to be essential in signaling mechanisms to identify and repair DNA damage [30–32]. Oxidative DNA damage has also been suggested to be involved in gene regulation by functioning as a kind of epigenetic mark [33].

To analyze the relationship between DNA electronic properties and genome mutability, we previously used quantum chemistry approaches to compute the vertical ionization potential (vIP) of stacked nucleobase motifs [21, 27, 29], defined as the energy required to remove an electron from a DNA molecule without taking into account possible structural rearrangements. We showed [21] that DNA motifs with low vIP values are more frequently mutated, whereas those with higher vIP tend to be more protected against mutations.

In this study, we extended the analysis of DNA electronic properties to gain deeper insight into their relationship with cancer mutagenesis processes. Specifically, we compared these properties with experimental data on DNA mutability across both local and genome-wide scales. We further examined the sequence-dependent tendency of DNA motifs to undergo oxidation and its association with the frequency of SBSs across different sequence contexts in somatic mutations from normal and cancer tissues, as well as in germline mutations. Finally, we explored how DNA electronic characteristics correlate with known cancer mutational signatures [16], and identify those in which they play a prominent role.

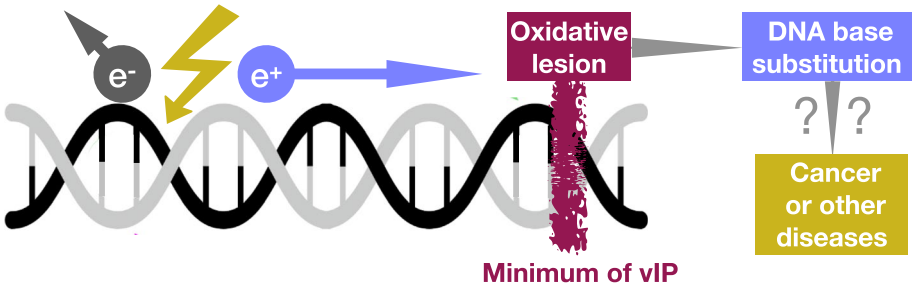

Fig. 1 Illustration of the mechanism of electron hole formation and migration along the DNA stack toward low-vIP regions, where the holes can cause oxidative damage and trigger SBSs, possibly leading to diseases such as cancer

Materials and methods

Datasets

We constructed two datasets of annotated somatic mutations, one in normal tissues and another in cancer tissues, and a third dataset of germline variants, among which benign and disease-causing variants. We only considered SBSs, and overlooked insertions and deletions. We provide below a brief description of the dataset characteristics; additional information is available in Supplementary Sections 1 and 2:

- S_{SC} is a dataset of somatic SBSs in cancer tissues that we collected from the COSMIC database [16], limited to nuclear DNA. It contains about 12.5 million SBSs; about 40% are situated in coding regions and the remaining 60% in non-coding regions.
- S_{SN} is a dataset of somatic SBSs in normal tissues that we collected and curated from SomaMutDB [34]. It contains about 8 million SBSs, of which only 2.9% are located in coding regions, with the rest occurring in non-coding regions.
- S_G is a collection of germline SBSs from ClinVar [35], which have been carefully annotated with genomic and clinical annotations. After curation, we obtained information for about 3 million germline SBSs, among which 80 % are in coding regions.

We downloaded the variant data in June 2025 and filtered them to retain only SBSs. These variants were then annotated using snpEff (v5.2e) [36] based on the GRCh38 reference genome. They were classified into different categories: coding, separated into nonsense, missense and silent; untranslated region (UTR); splice region found at the beginning and end of the introns and spanning the intron–exon junction; introns; and intergenic (see Supplementary Table S1). Only canonical transcripts were considered.

Sequence context of the variants

For each SBS in each of the three datasets S_{SC}, S_{SN} and S_G, we collected the flanking sequence of the mutated base 'X' in the reference genome, both upstream and downstream. More precisely, we considered motifs with n nucleobases upstream and n downstream ($n \leq 5$), noted 5'-N..NXN..N-3', or simply N..NXN..N, with the mutated base X at the middle position and N one of the four nucleobases. The frequency $f_{\text{N..NXN..N}}$ of occurrence of these motifs was computed separately in the three mutation datasets S_{SC}, S_{SN} and S_G and, when required, separately for each mutation category (e.g., missense, silent). In addition, we computed the frequencies $\hat{f}_{\text{N..NXN..N}}$ of the same base motifs, irrespective of mutation status, in the reference human genome GRCh38. The natural logarithm of the normalized frequency of each mutated motif, called F, is defined as:

$$F_{\text{N..NXN..N}} = \log \frac{f_{\text{N..NXN..N}}}{\hat{f}_{\text{N..NXN..N}}}. \quad (1)$$

The more negative the F value, the less frequently the variant is observed in the N..NXN..N context compared to a random mutation in the genome. In contrast, positive F values indicate that the variant is more frequent in that specific nucleobase motif. Note that we took the normalizing frequencies $\hat{f}$ independent of the DNA region, e.g., coding and non-coding, given that the electron hole migration is essentially blind to it. This choice is discussed in the Discussion and conclusion section.

Most sequencing datasets do not retain information about the original strand on which a mutation first occurred, since sequencing is performed after the variant has been fixed. Consequently, we calculated the $\overline{F}$ values from the sum of the frequencies of the motifs N..NXN..N and their reverse complement $\overline{\text{N..NXN..N}}$:

$$\overline{F}_{\text{N..NXN..N}} = \log \frac{f_{\text{N..NXN..N}} + f_{\overline{\text{N..NXN..N}}}}{\hat{f}_{\text{N..NXN..N}} + \hat{f}_{\overline{\text{N..NXN..N}}}} \quad (2)$$

When the structure of the genetic code does not allow an SBS to give rise to a particular type of mutation, it is excluded from the frequency counting. This situation often arises for nonsense mutations, given that only three stop codons exist. Particular care must also be taken with base-context changes at exon–intron boundaries, which may create a stop codon after mutation; such cases are, of course, excluded.

Vertical ionization potentials of nucleobase motifs

The vIP is the energy needed to remove an electron from a molecule, without a change in geometry. Thus, vIP values are always positive, and the lower they are, the easier it is to extract an electron.

In a previous paper [29], we have introduced the freely available program vIPer (github.com/3BioCompBio/vIPer) that computes the vIP value of all possible single-stranded DNA stacks in the commonly observed B conformation. More precisely, we have calculated the vIP values for all single nucleobases N, doublets NN, triplets NNN, and quadruplets NNNN using *ab initio* quantum chemistry methods. For longer nucleobase sequences, we have developed an iterative computational model, called vIPer, which is based on the vIP value of the submotifs included in the motif, and is able to return the vIP values of a motif of any length.

We also calculated the vIP of double-strand motifs, noted $\overline{\text{vIP}}$, as the mean of the vIP values of both the forward and reverse strand motifs:

$$\overline{\text{vIP}}_{\text{N..NXN..N}} = \frac{1}{2}(\text{vIP}_{\text{N..NXN..N}} + \text{vIP}_{\overline{\text{N..NXN..N}}}). \quad (3)$$

Both vIP and $\overline{\text{vIP}}$ values were obtained using the vIPer software, version v1.0 [29].

Among all four nucleobases, guanine has been experimentally shown to be the nucleobase from which electrons are most easily extracted, and therefore it has the lowest vIP value, followed by adenine, cytosine, and thymine [37]. On the average, longer motifs exhibit lower vIP values; however, the magnitude of this decrease diminishes as motif length increases. For example, the mean vIP computed by vIPer is equal to 8.37, 7.90, 7.68, 7.54, 7.43 eV for singlets, doublets, triplet, quadruplets and quintuplets, respectively [29] (Supplementary Figure S1a).

The vIP values are obviously related to the guanine content of the motifs. This relation is illustrated for quintuplet motifs in Supplementary Figure S1b. As expected, the correlation is strong but imperfect: two different motifs with the same guanine content can differ in vIP by up to 1 eV. To remove the influence of the guanine content on the correlation R between vIP and F, we computed the corrected partial correlation coefficient defined as:

$$R(X, Y \mid Z) = \frac{R(X,Y) - R(X,Z)R(Y,Z)}{\sqrt{1 - R(X,Z)^2}\sqrt{1 - R(Y,Z)^2}}, \quad (4)$$

where X, Y and Z are the vIP, F and guanine content, respectively. This expression estimates the true relationship between vIP and F by controlling for the Gua content. In the absence of information about the DNA strand on which the mutation occurs, X, Y and Z are the $\overline{\text{vIP}}$, $\overline{F}$ and Gua/Cyt content. R is either the Pearson (r) or Spearman (ρ) correlation coefficient.

DNA oxidative damage and mutational signature

We collected the 95 mutational signatures contained in the COSMIC database [16], each corresponding to a different aetiology. Each signature is characterized by the probability of observing a triplet motif NXN mutated into NYN. By convention, the triplets reported in the signatures are those on the DNA strand in which the mutated nucleobase is a pyrimidine (C or T) [16]. There are thus six possible substitution types: C→ (A, G, T), and T→ (A, G, C), and a total of 96 double-stranded triplet motif substitutions NXN→NYN for each mutational signature. In our analysis, we essentially considered the wild-type motifs and not the specific substitutions. We thus took into account the frequency of the 32 wild type triplet motifs and their reverse complements in each signature, divided it by the frequencies $\hat{f}$ of the same motifs in the reference human genome, and took the logarithm of this ratio, thus obtaining $\overline{F}_{\text{NXN}}$ defined in Eq. (2).

DNA oxidative damage data

We collected and analyzed published experimental data on DNA oxidative damage at the genome scale, generated by two high-throughput sequencing techniques, each designed to detect distinct chemical signatures of oxidative stress in DNA [38, 39].

The first technique is repair-assisted damage detection (RADD) sequencing, which locates oxidative DNA lesions at about 150 base pair (bp) resolution [39]. DNA was first exposed to the oxidizing agent potassium bromate. Repair enzymes were then used to excise the DNA lesions, and to repair them *in vitro* with biotinylated nucleotides. The positions of the lesions were identified through standard DNA fragmentation, immunoprecipitation, and sequencing [39]. The data were reported on consecutive 200 bp segments of DNA.

The second technique is the apurinic-site (AP) sequencing approach, which is designed to detect apurinic sites induced by X-ray treatment [38]. AP sites are first labeled with an aldehyde-reactive probe, forming a biotin-tagged DNA adduct. This is followed by DNA fragmentation, immunoprecipitation, and sequencing, allowing for genome-wide mapping of DNA lesions at approximately 250 bp resolution. Three replicas of this experiment were performed in [38] and the data was provided for genome segments of heterogeneous length.

To enable direct comparison between the AP and RADD data over identical genomic regions, we first computed per-nucleotide AP values for each of the three replicates. For each segment, the strength of the AP signal was divided by the segment length, and the resulting normalized value was assigned to all nucleotides within that segment. The per-nucleotide values were then averaged across replicates. To obtain an AP value for each genome region of 200 bp defined by the RADD data, we simply summed the per-nucleotide values within that region.

The electronic properties of these genome regions were evaluated in two ways. First, we computed the $\overline{\text{vIP}}$ values for all the 200 bp genome segments using the vIPer program [29]. Second, we counted the number of overlapping triplets within these 200 bp segments whose $\overline{\text{vIP}}$ value is below a threshold value $T = 7.6$ eV, as discussed in Supplementary Section 4.1. This number, noted $M_{\overline{\text{vIP}}}$, is an estimation of the number of low-$\overline{\text{vIP}}$ traps likely to host electron holes and cause oxidative damage.

We considered the AP and RADD data and the $\overline{\text{vIP}}$ and $M_{\overline{\text{vIP}}}$ values at multiple genomic resolutions: we first considered these data at the finest resolution of 200 bp, and then combined several successive genome regions to form regions of up to 100,000 bp, which is the size proposed in [38] to reach reasonable AP data accuracy. The

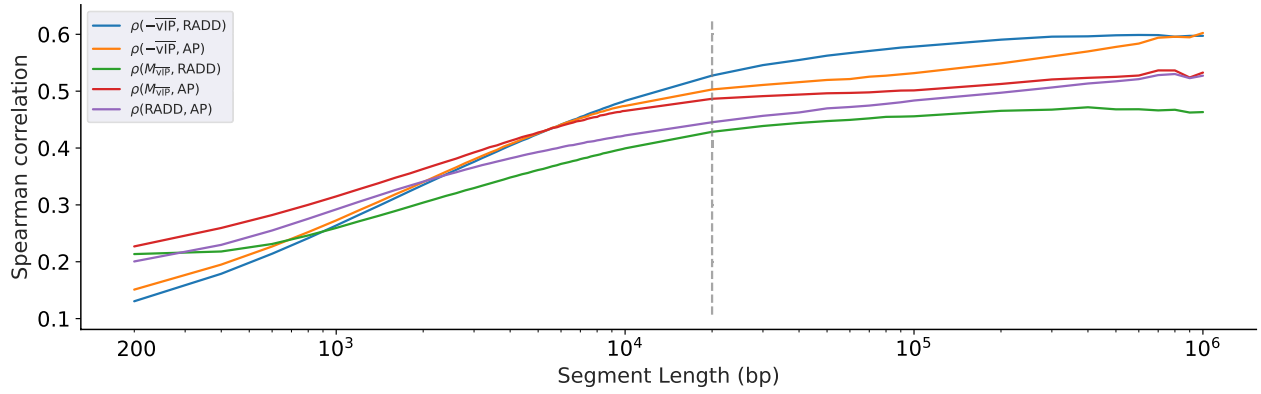

Fig. 2 Spearman correlation coefficient ρ between RADD, AP, $-\overline{\text{vIP}}$ and $M_{\overline{\text{vIP}}}$ scores (with threshold $T = 7.4$ eV) as a function of the genome segment length (actual values on logarithmic scale). The vertical dashed line indicates 20,000 bp, beyond which the correlations remain essentially constant

Table 1 Spearman correlation coefficients ρ between the values of RADD, AP, $-\overline{\text{vIP}}$, and $M_{\overline{\text{vIP}}}$ (with threshold $T = 7.4$ eV) for genome segments of various lengths

	$(-\overline{\text{vIP}},\ \boldsymbol{RADD})$	$(-\overline{\text{vIP}}, \mathbf{AP})$	$(M_{\overline{\text{vIP}}}, \mathbf{RADD})$	$(M_{\overline{\text{vIP}}}, \mathbf{AP})$	**(RADD,AP)**	$(-\overline{\mathbf{vIP}}, M_{\overline{\mathbf{vIP}}})$
200 bp	0.13	0.15	0.21	0.23	0.20	0.56
20 kbp	0.53	0.50	0.43	0.49	0.44	0.90
100 kbp	0.58	0.53	0.46	0.50	0.48	0.92
1 Mbp	0.60	0.60	0.46	0.53	0.52	0.89

RADD, AP, $\overline{\text{vIP}}$, and $M_{\overline{\text{vIP}}}$ values in these larger DNA segments were taken as the average of the scores of the combined segments.

Results

Genome-scale oxidation damage data and vIP

Over the past few decades, many experimental studies have been devoted to quantifying DNA oxidative damages, from single-site to genome-wide levels, and to improve our understanding of the biophysical processes that cause them. Here, we collected a series of data to more deeply investigate the DNA sequence dependence of oxidative damage and to compare them with calculated $\overline{\text{vIP}}$ and $M_{\overline{\text{vIP}}}$ values, which estimate how easily a DNA motif oxidizes. This is fundamental for understanding the genetic stability of the genome and for potentially identifying mutational hotspots.

In a previous study [29], we have analyzed single-site–level data on the photoinduced oxidation of twelve TNGNT motifs [40], and demonstrated a strong anticorrelation between measured relative reactivity and calculated vIP values. Here we considered data obtained through two high-throughput techniques developed to study oxidative damage on the genome scale: RADD sequencing of DNA oxidized using potassium bromate [39] and AP sequencing after X-ray treatment [38]. As detailed in the Methods section, RADD scores were available on successive 200 bp genomic segments, and we processed the AP data to yield AP scores for the same DNA segments.

We computed the Spearman correlation ρ between the two experimental oxidation damage scores with the vIP values of these segments on the whole human genome. As we do not have information on which strand the mutation primarily occurs, we rather considered the mean of the vIP of both strands, i.e. $\overline{\text{vIP}}$ as defined in Eq. (3). We also counted the number of low-$\overline{\text{vIP}}$ traps $M_{\overline{\text{vIP}}}$ in the segments, estimating the number of bases where the electron holes can localize and cause DNA damage. As described in the Methods section, we tested several threshold values T to define $M_{\overline{\text{vIP}}}$ traps. The threshold yielding the highest Spearman correlations with the RADD and AP scores for 200 bp segments, while maintaining good correlations at larger segment lengths, is T=7.4 eV, as shown in Supplementary Figure S2. In what follows, we only consider $M_{\overline{\text{vIP}}}$ scores with this threshold value.

We computed the correlation between RADD, AP, $\overline{\text{vIP}}$, and $M_{\overline{\text{vIP}}}$ on the original 200 bp segments as well as on aggregated consecutive segments of up to 1,000,000 bp. $\overline{\text{vIP}}$ and oxidative damage scores are anticorrelated, as lower vIPs indicate easier electron extraction; we thus correlated the experimental scores with $-\overline{\text{vIP}}$ instead of $\overline{\text{vIP}}$. The results are shown in Fig. 2 and Table 1. The correlation improves as a function of segment length, and reaches good values in the 20 kbp to 1,000 kbp range, as already noted for AP replicas [38]. Indeed, a certain level of variability is inherent to these experiments, as the formation of electron holes arises from intrinsically

probabilistic processes, which hinder accurate estimation of mutability at small scales in such genome-wide analyses.

For the smallest genome segment length of 200 bp, the best correlation values with experimental oxidation scores are obtained with $M_{\overline{\text{vIP}}}$ values: 0.21 for RADD scores and 0.23 for AP, which are close to the RADD-AP correlation of 0.20. For larger segment lengths, $-\overline{\text{vIP}}$ performs slightly better than $M_{\overline{\text{vIP}}}$, reaching 0.60 for 1 Mbp segments. This result highlights the close agreement between our estimates of the electronic oxidative properties of DNA and the experimental mutagenesis data. Notably, the correlation between experimental RADD and AP scores is significantly lower than that with $-\overline{\text{vIP}}$, i.e., 0.52 for 1 Mbp segments. This likely reflects differences in the mutagenic agents and cell types used in each experimental assay, which result in distinct patterns of DNA accessibility. Specifically, RADD exposes an osteosarcoma cell line to potassium bromate [39] whereas AP exposes the hepatocyte-derived HepG2 cell line to X-ray irradiation [38].

To further investigate the relationship between experimental and computed oxidation DNA lesions, we plotted RADD, AP, $(-\overline{\text{vIP}})$ and $M_{\overline{\text{vIP}}}$ scores along the reference human genome. For comparison purposes, we normalized these scores based on their ranks. The results are shown in Fig. 3 for chromosome 1 and in Supplementary Figure S3 for all other chromosomes. We plotted the results using 100 kbp segments to obtain a smooth representation. A zoomed-in view is shown in Supplementary Figure S4. We observe that the three curves are consistent and follow each other closely, in all chromosomes, identifying regions with low $\overline{\text{vIP}}$ that are highly mutated.

We observe a high variation in mutability across the entire genome, resulting from a complex interplay of genetics and epigenetics [41]. DNA electronic properties play a key role in the mutability process, as shown by the agreement of $(-\overline{\text{vIP}})$ and $M_{\overline{\text{vIP}}}$ values with experimental mutability scores. However, DNA accessibility is another important factor as already observed in [39]: heterochromatic regions, identified as dark bands in the karyogram, are highly condensed and less accessible to ROS; they generally appear as minima in Fig. 3 and Supplementary Figure S3. In contrast, light bands, corresponding to open euchromatic regions, are more prone to mutation.

In summary, the $(-\overline{\text{vIP}})$ and $M_{\overline{\text{vIP}}}$ values constitute very good proxies of mutability, with overall high correlation with the experimental mutability scores. The correlation strength varies within each chromosome, with certain regions displaying almost perfect correlations and others much lower ones. This may stem from local genomic and epigenomic features that are expected to modulate DNA damage, among which DNA methylation, DNA accessibility or alternative DNA structures, which are tissue-dependent and not accounted for in our current model (see Discussion and conclusion).

Cancer mutational signatures and vIPs

To understand the relationship between DNA electronic properties and carcinogenesis, we analyzed the 95 cancer mutational signatures from the COSMIC database [16] and correlated them with the calculated $\overline{\text{vIP}}$ values, as described in the Methods section. Each signature corresponds to a different cancer aetiology. Some are directly related to DNA oxidation processes, such as SBS18, which is associated with ROS, whereas others, such as SBS2 linked to APOBEC-mediated deamination, are not.

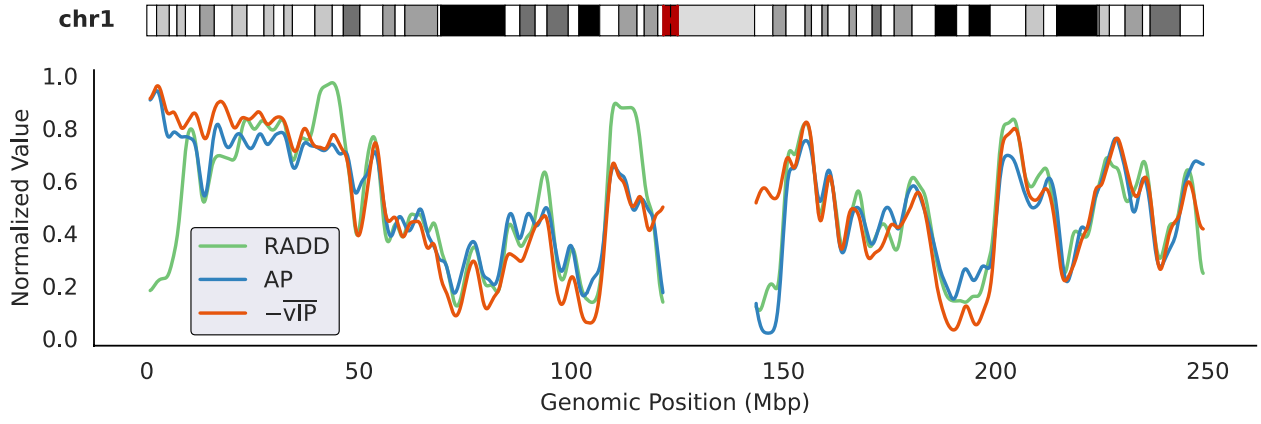

Fig. 3 Experimental RADD and AP scores and calculated $(-\overline{\text{vIP}})$ scores along human chromosome 1, averaged over 100 kbp genome segments. $M_{\overline{\text{vIP}}}$ values are not shown as they are basically superimposed with $(-\overline{\text{vIP}})$ values. For visualization purposes, the scores were rank-normalized and subsequently smoothed using a Gaussian filter spanning ten bins, as implemented in the `ndimage.gaussian_filter1d` function from the *SciPy* Python library. A karyotype of the chromosome is displayed at the top; the darker the regions, the fewer genes they contain and the more compacted they are; the red region corresponds to the centromere. The results on all the other human chromosomes are available in Supplementary Figure S3, and a zoom-in representation in Supplementary Figure S4

In a first step, we classified all signatures into six categories, following [42]: mutations resulting from exposure to electrophilic agents, exposure to UV light, defects in DNA repair mechanisms, other mutational processes, experimental artifacts, or unknown causes. Details on this classification is provided in Supplementary Table S4. Note that the identification of mutational signatures in which DNA oxidation plays a major role is a non-trivial issue and clearly depend on the flanking sequence of the mutated bases [33, 43].

To explore this further, we computed, for each signature, the correlation between F_{NXN}, the logarithm of the normalized mutation frequency of the base triplets NXN defined in Eq. (1), with their calculated vIP values (see Methods). As we do not have information on which strand the mutation primarily occurs, we rather considered $\overline{\mathrm{vIP}}$ (Eq. (3)) and $\overline{F}$ (Eq. (2)). Note that using the vIP of the reference strand or of the strand of minimal vIP yields slightly weaker anticorrelations (see Supplementary Table S5). We focused here on the wild-type triplet motifs and aggregated the data over all substitution types, as our primary aim is to characterize the global effect of DNA electronic properties on mutagenesis. For a detailed analysis of the relationship between $\overline{\mathrm{vIP}}$ and mutational signatures across substitution types and a discussion on the specific mechanisms underlying each mutation type, see Supplementary Section 5.2.

We considered here the Pearson correlation as the relationship between an energy (vIP) and the logarithm of a frequency (F) is expected to be linear in the context of the Boltzmann law; we added for comparison the Spearman correlation in Supplementary Figure S5 and Table S4. Both correlation types show the same trends and lead to the same conclusions.

The results of the $\overline{F} - \overline{\mathrm{vIP}}$ Pearson correlations for the 68 signatures that are not annotated as experimental artifacts are shown in Fig. 4. We observe that the majority of signatures (72%) show a negative correlation. This proportion is even larger for the signatures categorized as being due to electrophilic substances (80%) or to defects in repair mechanisms (80%).

The SBS4 signature shows the strongest anticorrelation: $r = -0.79$. It is attributed to tobacco smoking and annotated as likely resulting from DNA lesions directly induced by tobacco carcinogens [42]. For more details on this signature and the underlying mutational process, see the analysis presented in Supplementary Section 5.2. Related signatures are SBS92 ($r = -0.60$), linked to tobacco smoking, and SBS29 ($r = -0.69$), associated with tobacco chewing. The SBS5 signature ($r = -0.65$) has an unknown aetiology, yet its mutational burden is annotated to be consistently elevated in cancers related to tobacco exposure. These four signatures show a strong anticorrelation with $\overline{\mathrm{vIP}}$, consistent with the presence of a large variety of electrophilic molecules in tobacco, such as the polycyclic aromatic hydrocarbons, that are capable of inducing oxidative DNA damage [45].

Another signature with a strong anticorrelation of -0.65 is SBS18. This is expected, as SBS18 is linked to ROS, which react with DNA to generate a variety of oxidative lesions (e.g., 8-oxoG). ROS arise both from normal cellular metabolism, especially mitochondrial activity, and from external stressors such as radiation or pollutants, making oxidative DNA damage a burden for the cell [45, 46].

It is also noteworthy that the signature that exhibits the strongest $\overline{\mathrm{vIP}} - \overline{F}$ Pearson anticorrelation ($r = -0.81$) is SBS45 (Supplementary Table S4); it is not shown in Fig. 4 as it is annotated as an artifact. Notably, these mutations arise from the oxidation of guanine to 8-oxoG during sequencing sample preparation, a process driven by

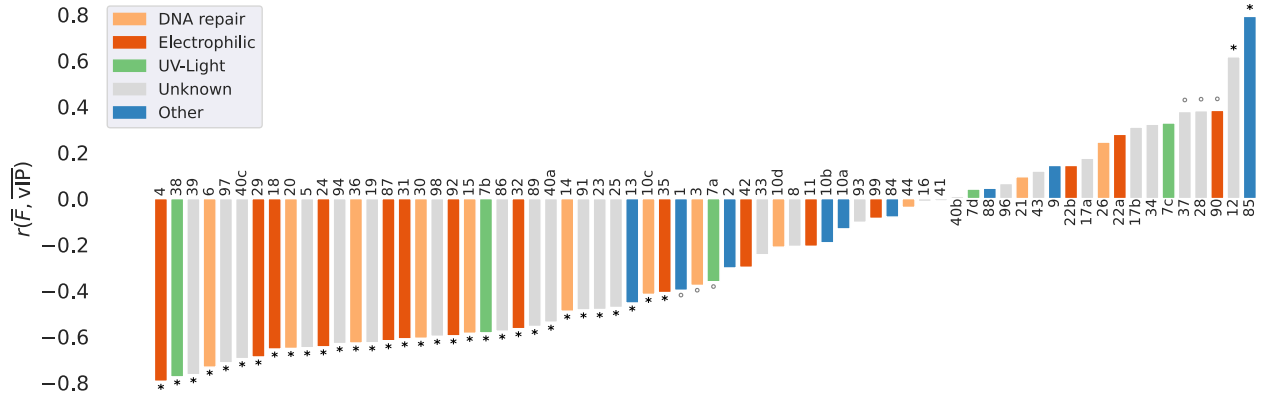

Fig. 4 Pearson correlation coefficients between $\overline{F}$ and $\overline{\mathrm{vIP}}$ for all cancer mutational signatures that are not reported to be experimental artifacts. The classification is performed according to [42]: exposure to electrophilic agents, exposure to UV light, defects in DNA repair mechanisms, other mutational processes or unknown mechanisms (see Supplementary Table S4). The '*' symbol indicates that the adjusted p-value, computed using the Benjamini–Hochberg correction for multiple testing [44], is ≤ 0.05. Signatures with a nominal p-value ≤ 0.05 (but not meeting the adjusted p-value threshold) are denoted by the 'o' symbol. A similar figure based on Spearman correlations is depicted in Supplementary Figure S5

oxidative stress [47]. This explains why SBS45 shows such a pronounced $\overline{\text{vIP}} - \overline{F}$ anticorrelation.

The reason why DNA repair–associated mutational signatures are anticorrelated with $\overline{\text{vIP}}$ is that several of them, such as the base excision repair–related signatures SBS36 ($r = -0.63$) and SBS30 ($r = -0.61$), are directly linked to the repair of oxidative DNA damage. When the repair machinery is impaired, oxidative lesions remain unrepaired and consequently accumulate as mutations of oxidative origin. Mutational signature SBS6 also shows a strong anticorrelation with $\overline{\text{vIP}}$ ($r = -0.73$). SBS6 is primarily associated with mismatch repair deficiency, often resulting in microsatellite instability, and is responsible for correcting DNA polymerase errors. Despite this, it also correlates strongly with vIP, suggesting that it may additionally be related to oxygen-induced genomic mismatches, as recent evidence appears to indicate [48].

Surprisingly, SBS90 shows a positive correlation with $\overline{\text{vIP}}$, although it is associated with duocarmycin exposure, an electrophilic substance. However, its specific mode of action explains the positive rather than negative correlation: it binds within the minor groove of DNA, where it induces adenine alkylation [49]. Therefore, its specificity for adenine bases severs potential correlation with vIP.

UV radiation induces diverse forms of DNA damage [50]. For example, it can cause oxidative damage either directly, through ionizing interactions with DNA, or indirectly, via UV-induced free radicals and ROS. Conversely, it can also generate double-strand breaks that are not directly linked to oxidative reactivity. Annotations generally do not explicitly specify the underlying mutational processes, but we can reasonably assume that certain UV-related signatures may primarily be due to electron-extracting processes, whereas others may arise from distinct, non-oxidative pathways. The correlation pattern observed with vIP values may therefore help disentangle which UV signatures are predominantly associated with specific types of photoproducts. Although the aetiology of the SBS38 signature is annotated as unknown, it is detected exclusively in UV-associated melanomas and shows a statistically significant correlation of r = -0.57, which suggests that it could be caused by indirect DNA oxidation damage caused by UV light. Signatures SBS7b and SBS7a are also due to UV exposure and have statistically significant anticorrelations of $r = -0.58$ and -0.36, respectively. In contrast, SBS7d and SBS7c display a positive, but not statistically significant correlation, suggesting they might not be due to UV-caused oxidative damage.

Comparison between somatic and germline mutations

Here we analyzed the correlation between the $\overline{\text{vIP}}$ of mutated motifs and their normalized frequency of observation $\overline{F}$ in the datasets of somatic cancer mutations (S_{SC}), somatic normal tissue mutations (S_{SN}), and germline mutations (S_G) (see Methods). The Pearson correlations are shown for both triplets (Table 2.a)

Table 2 Pearson correlation coefficients r and r_{GC} between $\overline{F}$ and $\overline{\text{vIP}}$ values of (a) triplets and (b) quintuplets in the three datasets S_{SC}, S_{SN}, S_G

Triplets - NXN								
	S_{SC}		S_{SN}		S_G		Total	
	r	r_{GC}	r	r_{GC}	r	r_{GC}	r	r_{GC}
silent	-0.85	(-0.70)	-0.87	(-0.74)	-0.86	(-0.72)	-0.86	(-0.72)
missense	-0.77	(-0.52)	-0.76	(-0.52)	-0.72	(-0.45)	-0.76	(-0.52)
nonsense	-0.70	(-0.55)	-0.62	(-0.50)	-0.61	(-0.45)	-0.69	(-0.54)
intronic	-0.67	(-0.53)	-0.71	(-0.57)	-0.74	(-0.53)	-0.69	(-0.54)
UTR	-0.75	(-0.58)	-0.75	(-0.59)	-0.75	(-0.56)	-0.76	(-0.59)
splice region	-0.61	(-0.41)	-0.68	(-0.49)	-0.64	(-0.41)	-0.64	(-0.43)
intergenic	-0.72	(-0.55)	-0.72	(-0.57)	-0.80	(-0.56)	-0.72	(-0.57)

(a)

Quintuplets - NNXNN								
	S_{SC}		S_{SN}		S_G		Total	
	r	r_{GC}	r	r_{GC}	r	r_{GC}	r	r_{GC}
silent	-0.77	(-0.58)	-0.79	(-0.61)	-0.77	(-0.57)	-0.78	(-0.59)
missense	-0.73	(-0.49)	-0.76	(-0.53)	-0.71	(-0.45)	-0.73	(-0.49)
nonsense	-0.63	(-0.47)	-0.59	(-0.47)$^{\times}$	-0.60	(-0.44)	-0.59	(-0.43)
intronic	-0.59	(-0.51)	-0.64	(-0.57)	-0.69	(-0.51)	-0.62	(-0.53)
UTR	-0.70	(-0.53)	-0.70	(-0.57)	-0.75	(-0.56)	-0.71	(-0.55)
splice region	-0.47	(-0.28)	-0.52	(-0.37)$^{\times}$	-0.45	(-0.24)	-0.47	(-0.28)
intergenic	-0.66	(-0.53)	-0.61	(-0.56)	-0.78	(-0.53)$^{\times}$	-0.63	(-0.57)

(b)

"Total" means the ensemble of the three datasets. r_{GC} is the Gua-Cyt-corrected partial correlation coefficient $r(\overline{\text{vIP}}, \overline{F} \mid \text{G-C})$ defined in Eq. (4). The p-values are all lower than 0.001. Darker shades indicate stronger anticorrelation (r). A $\times$ superscript indicates that the dataset contains less than 50 times the number of triplets (which is 64; in (a)) or of quintuplets (1024; in (b)); the dataset content is specified in Supplementary Table S2. We disregard these values with low statistical significance in our analyses. Similar Tables with Spearman correlations are found in Supplementary Table S6

and quintuplets (Table 2.b); Spearman correlations are given for comparison in Supplementary Tables S6a-b. We added in these Tables the partial Pearson and Spearman correlation coefficients between $\overline{\text{vIP}}$ and $\overline{F}$ while controlling for the effects of the G–C content, as defined in Eq. (4), in order to remove its contribution to the observed correlation. Note that the analysis of germline mutations further elaborates on our earlier studies [21, 29].

First, we observe strong Pearson anticorrelations of -0.75 for triplets and -0.71 for quintuplets when averaged across all datasets and substitution types. These highly significant anticorrelations indicate that lower vIP values are associated with higher mutability. This trend is consistently observed across somatic mutations in both normal and cancer tissues, as well as in germline mutations, with comparable correlation values. Interestingly, this suggests that these mutational processes may arise through similar mechanisms, with oxidative damage playing a central role.

One might suspect that these strong anticorrelations are driven primarily by the Watson–Crick G–C base pair content rather than by the intrinsic vIP values themselves. This is not the case: the partial correlation coefficient $r(\overline{\text{vIP}}, \overline{F} \mid \text{G-C})$ for given G-C content remain substantial: -0.58 for triplets and -0.57 for quintuplets. This result demonstrates that vIP-derived electronic properties contribute intrinsically to DNA mutability beyond simple base composition effects.

Within coding regions, we classified mutations into silent, missense, and nonsense. Silent mutations show the strongest $\overline{\text{vIP}} - \overline{F}$ anticorrelations across all three datasets: -0.85 to -0.87 for triplets and -0.77 to -0.79 for quintuplets. Missense mutations also display strong anticorrelations, though to a lesser extent (-0.72 to -0.77 for triplets and -0.71 to -0.76 for quintuplets).

The stronger anticorrelations observed for silent compared with missense mutations might be attributed to their lack of effect on the encoded amino acid. As these mutations are typically under weaker selective constraints, the resulting correlation pattern might more faithfully mirror the intrinsic mutational process. It should be noted that silent mutations are not totally neutral: they can influence translation efficiency and mRNA stability, and some have been implicated in cancer pathogenesis [51–53]. However, missense mutations are generally under stronger purifying selection, which removes deleterious variants from the observable spectrum and consequently distorts the underlying $\overline{\text{vIP}} - \overline{F}$ relationship. Nonsense mutations, which introduce premature stop codons, are typically highly deleterious and therefore subject to even stronger negative selection, which may account for the weaker anticorrelations observed. Overall, the decline in anticorrelation from silent to missense to nonsense mutations can be interpreted as reflecting their progressively greater functional impact and the corresponding intensification of negative selection. Further investigations are needed to support this hypothesis.

In non-coding regions, the strongest $\overline{F} - \overline{\text{vIP}}$ anticorrelation occur in UTRs, which show anticorrelations comparable to missense mutations (-0.75 to -0.76 for triplets and -0.70 to -0.74 for quintuplets). Indeed, UTRs regulate mRNA stability, localization, and translation, so mutations in these regions can significantly disrupt gene expression [54].

Splice regions show weaker anticorrelations (-0.61 to -0.68 for triplets and -0.45 to -0.47 for quintuplets), consistent with their high functional impact. Mutations in these sites often produce aberrant transcripts and are well known to contribute to inherited diseases and cancer [55, 56].

Introns display intermediate anticorrelations (-0.67 to -0.74 for triplets and -0.59 to -0.69 for quintuplets). Some intronic motifs have been reported to play a role of limiting the impact of oxidative damage. It has indeed be observed that poly-Gua regions are frequent near the 5' termini of introns, where they have been suggested to serve as sacrificial anodes to protect the protein-coding part of the genes from oxidative damage [57]. Further analyses are needed for interpreting the mechanisms and impact of mutations in these regions.

Finally, the anticorrelations in intergenic regions are in the -0.72 to -0.80 range for triplets and between 0.61 and -0.66 for quintuplets. The observed variability between datasets, with slightly stronger anticorrelations in the S_G set, may stem from differences in DNA sequencing methods. Specifically, capture sequencing and whole-exome sequencing tend to focus on intergenic regions adjacent to genes which are enriched in regulatory motifs, whereas whole-genome sequencing targets the entire genome. S_G essentially contain variants collected with the former methods and S_{SN} with the latter; S_{SC} contains both types.

We emphasize that, across all mutation categories considered, not only the $\overline{F}$–$\overline{\text{vIP}}$ anticorrelations but also the corresponding partial anticorrelations controlling for G–C content are strong. This further supports the conclusion that vIP is intrinsically linked to DNA mutability.

When interpreting differences in anticorrelation values across genomic regions, it is important to recall that F is obtained by normalizing the mutated motif frequency by the motif frequency in the whole genome (see Eq. (1)). We made this choice because the electron hole migration towards the DNA region of lowest vIP is blind to the type of DNA region it crosses. However, although the electron holes can migrate over long distances, of at least 100 Å [58], they do not migrate across the whole genome. Moreover, some regions may adopt conformations or interact with proteins which either block or facilitate

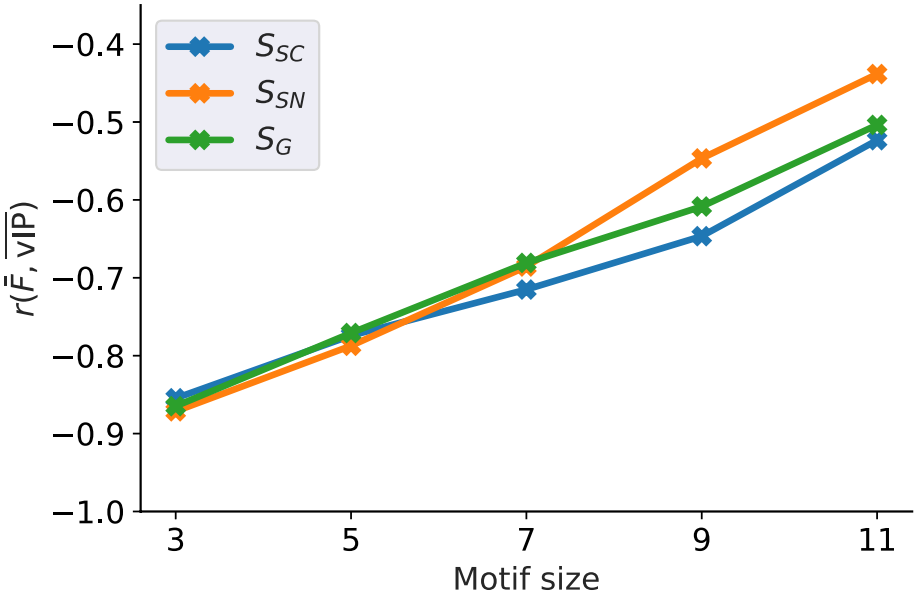

Fig. 5 Pearson correlation r between $\overline{F}$ and $\overline{\text{vIP}}$ as a function of motif length, for silent mutations in the datasets S_{SC}, S_{SN}, S_G

the migration [59]. So, the chosen normalization is an approximation, and differences in the levels of anticorrelation values in different regions may partly be related to this choice. Note, however, that altering the normalization scheme by computing motif frequencies within specific genome regions, does not substantially affect the results, as mentioned in Supplementary Section 8.

Influence of flanking sequences. We observed a decrease in $\overline{\text{vIP}} - \overline{F}$ anticorrelation strength between triplets and quintuplets, accompanied by a strong increase in statistical significance, as measured by p-values of about 10^{-10} and 10^{-100} for triplets and quintuplets, respectively. Here, we investigated whether this anticorrelation persists for longer motifs, assessing whether flanking sequence context beyond quintuplets and its associated $\overline{\text{vIP}}$ continues to modulate DNA mutability. We specifically focused on silent mutations, where the negative selection is lowest. As shown in Fig. 5, the strength of anticorrelations slightly decreases with motif length to approximately the same extent for the three datasets S_{SC}, S_{SN}, S_G, but remain strong: about -0.5 for undecuplet motifs. In parallel, the statistical significance increases sharply, reaching an undetectable p-value for undecuplets. The decrease in the strength of the $\overline{\text{vIP}} - \overline{F}$ anticorrelation may be attributed to the increasing uncertainty in vIP calculations for longer motifs and the limited availability of mutational data to adequately sample all motifs.

However, the persistence of significant anticorrelation values underscores the important influence of flanking sequences on both vIP and mutation frequency. These results point out towards a causative link between vIP and oxidative, mutation-inducing phenomena, which are both non-local in nature. Indeed, oxidative damage has been experimentally shown to occur sometimes far from the eventual mutation site, and electron holes can migrate along the DNA stack over long distances to reach regions of lowest vIP [58].

Dependence on chromosome. We tested whether the strength of the anticorrelation between $\overline{\text{vIP}}$ and $\overline{F}$ depends on the chromosome. For this purpose, we considered together all the silent mutations contained in the datasets S_{SC}, S_{SN}, S_G. As shown in Supplementary Table S7, there are only minor variations in Pearson anticorrelation across the different chromosomes: from -0.73 for chromosome 13 to -0.79 for chromosomes 16, 20, and 22. The only exception is the Y chromosome, for which $r = -0.64$. As the most compact and densely packed chromosome, it may be partially "protected" from oxidative damage. However, the limited number of silent mutations on the Y chromosome (446) precludes this result from reaching statistical significance.

Dependence on the tissue. As various tissues encounter diverse levels and types of mutagens and mutation processes, face unique functional constraints, and develop cancers through different molecular pathways, we investigated the $\overline{\text{vIP}} - \overline{F}$ anticorrelation strength of cancer mutations in S_{SC} according to the tissue; we focused again on silent mutations.

We found the anticorrelations for triplet motifs to lie between $r = -0.87$ for mutations in the lung and in the kidney to $r = -0.72$ for mutations in the nervous system (Supplementary Table S8). These correlation differences are related to the mutational signatures discussed in the previous section. For example, lung cancer is predominantly caused by oxidative substances present in tobacco,

and thus both tobacco-related signatures and mutations in the lung anticorrelate strongly with $\overline{\text{vIP}}$. This reflects the dominant contribution of oxidative agents present in tobacco to lung carcinogenesis. In contrast, only a subset of DNA lesions caused by UV light are of oxidative nature. As a consequence, different UV-related signatures have different anticorrelation strengths, which can be related to the observation that skin cancer mutations anticorrelate less strongly ($r = 0.76$).

Discussion and conclusion

Oxidative DNA damage, generated by endogenous metabolic by-products or exogenous environmental stressors, produces lesions that are key drivers of carcinogenesis and determinants of chemotherapy efficacy. However, gaining a precise understanding of how oxidative agents give rise to SBSs remains challenging, due to the complexity and diversity of oxidative lesions and their repair pathways.

In this study, we advanced this question by studying the DNA electronic properties quantified through vIP values. Low-vIP genomic regions are particularly susceptible to oxidative lesions, as they have a greater propensity for electron extraction leading to electron-hole formation. Such lesions are either repaired by the cellular repair machinery or cause sequence alterations such as SBSs. Moreover, when electron holes are generated in higher vIP regions, they tend to migrate along the nucleobase stack in the 5'–3' direction until they reach a low-vIP site [60, 61], where a lesion and possibly a mutation may be induced. Note that, in the absence of information on the strand on which the mutation occurs, we considered here the mean vIP of both strands, i.e., $\overline{\text{vIP}}$ rather than vIP.

In our study, we compared $\overline{\text{vIP}}$ profiles with cancer mutational signatures and demonstrated that such profiles reliably highlight signatures linked to oxidative damage aetiologies. Therefore, our $\overline{\text{vIP}}$-based approach could be used to predict the oxidative nature of cancer mutational signatures of unknown origin, thereby providing new insights into their aetiology. A key perspective of this study involves designing a tool to enable such $\overline{\text{vIP}}$ -based predictions.

Furthermore, we observed that selective pressure shapes the mutation spectrum observed in sequencing data. Highly deleterious SBSs, such as those disrupting essential regulatory elements or introducing nonsense mutations, are often removed from the cell population and therefore go undetected. This depletion of highly deleterious variants might explain why severe mutation types (e.g., nonsense and, to a lesser extent, missense) tend to show weaker anticorrelations with $\overline{\text{vIP}}$ than more benign ones such as silent mutations. However, this interpretation requires further investigation to be substantiated.

An interesting finding is that the $\overline{\text{vIP}} - \overline{F}$ anticorrelations we observed are very similar for somatic and germline mutations, as well as in cancer and normal tissues. This suggests that these mutations arise through similar mechanisms, in which oxidative damage plays a central role. Some small differences are nevertheless observed, which could be attributed to varying levels of negative selection [62, 63].

Although our vIP model already describes DNA electronic properties and their relationship to mutational processes very well, its ability to identify genomic regions prone to mutation could still be improved. Obtaining more precise vIP values using advanced quantum chemistry techniques and considering alternative DNA conformations, thereby dropping the assumption that chromosomes adopt a uniform B-form DNA structure, could further refine the model and provide a more accurate representation of sequence-dependent mutational susceptibility.

While oxidative damage is strand specific, this information is often unavailable in mutation databases, leading us to use the vIP and mutation frequency averaged over both strands. Using the vIP of the exact strand where the oxidative damage occurred could further improve our understanding of these processes. Interestingly, recent results suggest a strand bias in oxidative damage, with transcribed strands experiencing less oxidative damage than their non-transcribed counterparts [43].

It would also be interesting to incorporate chromatin state and tissue-specific genomic data in our model. Chromatin accessibility, nucleosome positioning, and tissue-dependent epigenetic landscapes all influence how DNA is exposed to damaging agents and how efficiently lesions are repaired. Possible important protein-DNA interactions likely to affect electron hole migration [59, 64] could also be integrated. Including these layers of information could therefore improve the ability of our vIP model to describe mutational processes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12879-5.

Supplementary Material 1.

Authors' contributions

B.d.W., C.K., P.H. and M.R. developed the computational methods. B.d.W., C.K. and M.R. analyzed and interpreted the data. F.P. and M.R. conceived the study. M.T, F.P. and M.R. supervised the study. B.d.W., F.P. and M.R. wrote the initial draft of the manuscript. All authors reviewed, edited, and approved the final manuscript.

Funding
This work was supported by a Télévie grant and a PDR project from the F.R.S.-FNRS Fund for scientific research.

Data availability
Code used to generate the figures and analyses are available on GitHub at https://github.com/3BioCompBio/MutationvIP Preprocessed data available for public redistribution are deposited on Zenodo at https://doi.org/10.5281/zenodo.17870586.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 January 2026 / Accepted: 17 April 2026
Published online: 02 May 2026

References

1. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A Compendium of Mutational Cancer Driver Genes. Nat Rev Cancer. 2020;20(10):555–72. https://doi.org/10.1038/s41568-020-0290-x.
2. McFarland CD, Mirny LA, Korolev KS. Tug-of-War between Driver and Passenger Mutations in Cancer and Other Adaptive Processes. Proc Natl Acad Sci USA. 2014;111(42):15138–43. https://doi.org/10.1073/pnas.1404341111.
3. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of Deleterious Passenger Mutations on Cancer Progression. Proc Natl Acad Sci USA. 2013;110(8):2910–5. https://doi.org/10.1073/pnas.1213968110.
4. Tilk S, Tkachenko S, Curtis C, Petrov DA, McFarland CD. Most Cancers Carry a Substantial Deleterious Load Due to Hill-Robertson Interference. Elife. 2022;11:e67790.
5. Mendelsohn J, Howley PM, Israel MA, Gray JW, Thompson C, editors. The Molecular Basis of Cancer. 4th ed. Philadelphia, PA: Saunders/Elsevier; 2015.
6. Cooke MS, Evans MD, Dizdaroglu M, Lunec J. Oxidative DNA damage: mechanisms, mutation, and disease. FASEB J. 2003;17(10):1195–214. https://doi.org/10.1096/fj.02-0752rev.
7. Klaunig JE, Kamendulis LM. The Role of Oxidative Stress in Carcinogenesis. Annu Rev Pharmacol Toxicol. 2004;44(1):239–67. https://doi.org/10.1146/annurev.pharmtox.44.101802.121851.
8. Liou GY, Storz P. Reactive oxygen species in cancer. Free Radical Res. 2010;44(5):479–96.
9. Tubbs A, Nussenzweig A. Endogenous DNA damage as a source of genomic instability in cancer. Cell. 2017;168(4):644–56.
10. Hahm JY, Park J, Jang ES, Chi SW. 8-Oxoguanine: from oxidative damage to epigenetic and epitranscriptional modification. Exp Mol Med. 2022;54(10):1626–42.
11. Alhmoud JF, Woolley JF, Al Moustafa AE, Malki MI. DNA Damage/Repair Management in Cancers. Cancers. 2020;12(4):1050. https://doi.org/10.3390/cancers12041050.
12. David SS, O'Shea VL, Kundu S. Base-excision repair of oxidative DNA damage. Nature. 2007;447(7147):941–50. https://doi.org/10.1038/nature05978.
13. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8. https://doi.org/10.1038/nature12213.
14. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016;354(6312):618–22.
15. Jin SG, Pettinga D, Johnson J, Li P, Pfeifer GP. The major mechanism of melanoma mutations is based on deamination of cytosine in pyrimidine dimers as determined by circle damage sequencing. Sci Adv. 2021;7(31):eabi6508.
16. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941–7. https://doi.org/10.1093/nar/gky1015.
17. Hu X, Xu Z, De S. Characteristics of mutational signatures of unknown etiology. NAR cancer. 2020;2(3):zcaa026.
18. Hall DB, Holmlin RE, Barton JK. Oxidative DNA damage through long-range electron transfer. Nature. 1996;382(6593):731–5.
19. Bacolla A, Temiz NA, Yi M, Ivanic J, Cer RZ, Donohue DE, et al. Guanine Holes Are Prominent Targets for Mutation in Cancer and Inherited Disease. PLoS Genet. 2013;9(9):e1003816. https://doi.org/10.1371/journal.pgen.1003816.
20. Suárez-Villagrán MY, Azevedo RB, Miller JH Jr. Influence of electron-holes on DNA sequence-specific mutation rates. Genome Biol Evol. 2018;10(4):1039–47.
21. Pucci F, Rooman M. Relation between DNA ionization potentials, single base substitutions and pathogenic variants. BMC Genomics. 2019;20(S8):551. https://doi.org/10.1186/s12864-019-5867-y.
22. Suárez MY, Villagrán, Miller JH. Computational DNA hole spectroscopy: A new tool to predict mutation hotspots, critical base pairs and disease 'driver' mutations. Sci Rep. 2015;5(1):13571.
23. Bacolla A, Cooper DN, Vasquez KM. Mechanisms of base substitution mutagenesis in cancer genomes. Genes. 2014;5(1):108–46.
24. Suárez-Villagrán M, Mitsakos N, Miller Jr J. Beyond sequence: a physics-informed machine learning framework for predicting DNA mutations. Comput Struct Biotechnol J. 2025;27:3985–92.
25. Giese B. Long-distance electron transfer through DNA. Annu Rev Biochem. 2002;71:51.
26. Chakraborty T. Charge migration in DNA: perspectives from physics, chemistry, and biology. Berlin Heidelberg: Springer Science & Business Media; 2007.
27. Rooman M, Wintjens R. Sequence and conformation effects on ionization potential and charge distribution of homo-nucleobase stacks using M06–2X hybrid density functional theory calculations. J Biomol Struct Dyn. 2014;32(4):532–45.
28. Peluso A, Caruso T, Landi A, Capobianco A. The Dynamics of Hole Transfer in DNA. Molecules. 2019;24(22):4044. https://doi.org/10.3390/molecules24224044.
29. Rooman M, Pucci F. Estimating the Vertical Ionization Potential of Single-Stranded DNA Molecules. J Chem Inf Model. 2023;63(6):1766–75.
30. Boal AK, Genereux JC, Sontz PA, Gralnick JA, Newman DK, Barton JK. Redox signaling between DNA repair proteins for efficient lesion detection. Proc Natl Acad Sci. 2009;106(36):15237–42.
31. Zwang TJ, Tse EC, Barton JK. Sensing DNA through DNA charge transport. ACS Chem Biol. 2018;13(7):1799–809.
32. Sontz PA, Muren NB, Barton JK. DNA charge transport for sensing and signaling. Acc Chem Res. 2012;45(10):1792–800.
33. Poetsch AR. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. Comput Struct Biotechnol J. 2020;18:207–19.
34. Sun S, Wang Y, Maslov AY, Dong X, Vijg J. SomaMutDB: a database of somatic mutations in normal human tissues. Nucleic Acids Res. 2022;50(D1):D1100–8.
35. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res. 2020;48(D1):D835–44.
36. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92.
37. Hush N, Cheung AS. Ionization potentials and donor properties of nucleic acid bases and related compounds. Chem Phys Lett. 1975;34(1):11–3.
38. Poetsch AR, Boulton SJ, Luscombe NM. Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. Genome Biol. 2018;19(1):1–23.
39. Gilat N, Fridman D, Sharim H, Margalit S, Gassman NR, Michaeli Y, et al. From Single-molecule to genome-wide mapping of DNA lesions: Repair-assisted damage detection sequencing. Biophys Rep. 2021;1(2):100017.
40. Saito I, Nakamura T, Nakatani K, Yoshioka Y, Yamaguchi K, Sugiyama H. Mapping of the hot spots for DNA damage by one-electron oxidation: efficacy of GG doublets and GGG triplets as a trap in long-range hole migration. J Am Chem Soc. 1998;120(48):12686–7.
41. Oman M, Ness RW. Comparing the predictors of mutability among healthy human tissues inferred from mutations in single-cell genome data. Genetics. 2025;229(3):iyae215.
42. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94–101.
43. Takhaveev V, Püllen NJ, Singh NK, Lefevre L, Aghajani EA, Huber SM, et al. Click-code-seq reveals strand biases of DNA oxidation and depurination in human genome. Nat Chem Biol. 2025;22(5):716–27.

44. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol). 1995;57(1):289–300.
45. Barnes JL, Zubair M, John K, Poirier MC, Martin FL. Carcinogens and DNA damage. Biochem Soc Trans. 2018;46(5):1213–24.
46. Cheung EC, Vousden KH. The role of ROS in tumour development and progression. Nat Rev Cancer. 2022;22(5):280–97.
47. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41(6):e67.
48. Lózsa R, Németh E, Gervai JZ, Márkus BG, Kollarics S, Gyüre Z, et al. DNA mismatch repair protects the genome from oxygen-induced replicative mutagenesis. Nucleic Acids Res. 2023;51(20):11040–55.
49. Morcos A, Jung Y, Galvan Bustillos J, Fuller RN, Caba Molina D, Bertucci A, et al. A Comprehensive Review of the Antitumor Properties and Mechanistic Insights of Duocarmycin Analogs. Cancers. 2024;16(19):3293.
50. Rastogi RP, Richa N, Kumar A, Tyagi MB, Sinha RP. Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. J Nucl Acids. 2010;2010(1):592980.
51. Dhindsa RS, Copeland BR, Mustoe AM, Goldstein DB. Natural selection shapes codon usage in the human genome. Am J Hum Genet. 2020;107(1):83–95.
52. Gutman T, Goren G, Efroni O, Tuller T. Estimating the predictive power of silent mutations on cancer classification and prognosis. NPJ Genom Med. 2021;6(1):67.
53. Kikutake C, Suyama M. Possible involvement of silent mutations in cancer pathogenesis and evolution. Sci Rep. 2023;13(1):7593.
54. Su JY, Wang YL, Hsieh YT, Chang YC, Yang CH, Kang Y, et al. Multiplexed assays of human disease-relevant mutations reveal UTR dinucleotide composition as a major determinant of RNA stability. Elife. 2025;13:RP97682.
55. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet. 2018;59(3):253–68.
56. Ouyang J, Zhang Y, Xiong F, Zhang S, Gong Z, Yan Q, et al. The role of alternative splicing in human cancer progression. Am J Cancer Res. 2021;11(10):4642.
57. Friedman KA, Heller A. On the non-uniform distribution of guanine in introns of human genes: possible protection of exons against oxidation by proximal intron poly-G sequences. J Phys Chem B. 2001;105(47):11859–65.
58. Wagenknecht HA. Remote Photodamaging of DNA by Photoinduced Energy Transport. ChemBioChem. 2022;23(2):e202100265.
59. Rooman M, Cauët E, Liévin J, Wintjens R. Conformations consistent with charge migration observed in DNA and RNA X-ray structures. J Biomol Struct Dyn. 2011;28(6):949–54.
60. Boon EM, Barton JK. Charge transport in DNA. Curr Opin Struct Biol. 2002;12(3):320–9.
61. Grodick MA, Muren NB, Barton JK. DNA charge transport within the cell. Biochemistry. 2015;54(4):962–73.
62. Chu D, Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC Cancer. 2019;19(1):359.
63. Vitale E, Ricci D, Corrao F, Fiduccia I, Cruciata I, Carollo PS, et al. Nonsense Mutations in Rare and Ultra-Rare Human Disorders: An Overview. IUBMB Life. 2025;77(6):e70031.
64. Rooman M, Liévin J, Buisine E, Wintjens R. Cation-π/H-bond stair motifs at protein-DNA interfaces. J Mol Biol. 2002;319(1):67–76.

Publisher's Note