Purine stretches are avoided by cancer mutations

Aleksandr Vikhorev[1]

Ivan Savelev[1]

Oksana Polesskaya[1]

Richard Alan Miller[2]

Max Myakishev-Rempel[1]

1. DNA Resonance Research Foundation, San Diego, CA, USA

2. OAK, Inc., Grants Pass, OR, USA

Max Rempel email: max@dnaresonance.org

Keywords: Purine Stretches, Cancer Development, Aromaticity, Electron Delocalization, DNA-triplex, Enol-Amine Forms, Quantum-Chemical Models, Flanking Sequence Context, Mutation Susceptibility

Abstract

Purine stretches, sequences of adenine (A) and guanine (G) in DNA, play critical roles in binding regulatory protein factors and influence gene expression by affecting DNA folding. This study investigates the relationship between purine stretches and cancer development, considering the aromaticity of purines, quantified by methods like Hückel's rule and NICS calculations, and the importance of the flanking sequence context, affecting mutations up to 2000 nucleotides away. A pronounced influence of cancer mutations on purine stretch lengths was observed, particularly in intergenic regions, pointing to the role of intergenes in chromatin reorganization. A statistically significant shortening in cancerous tissue (p-value < 0.0001) was found, along with varied changes in purine stretch length within gene promoters across different cancer types. This reflects complex interactions between purine stretches and cancer. The study also revealed that common cancer mutations avoid purine stretches, especially in shorter and more prevalent mutations. The insights into the aromatic nature of purines and their stacking energies explain the role of purine stretches in DNA structure, contributing to the complex relationship with cancer. This research lays the groundwork for understanding the multifaceted nature of purine stretches, emphasizing their importance in gene regulation and chromatin restructuring, and offers potential avenues for novel cancer therapies and insights into cancer biology.

Introduction

Purine stretches are sequences of purine nucleotides (adenine and guanine) in DNA that serve as binding sites for regulatory protein factors involved in transcription, replication, and recombination. These stretches can also play a role in regulating gene expression by affecting the folding of the DNA molecule and determining the accessibility of specific regions to the cellular machinery involved in gene regulation.

The flanking sequence context of cancer mutations refers to the nucleotide sequence that surrounds a specific mutation within a gene. These flanking sequences influence where the mutations happen and the frequencies of substitutions of one nucleotide by the others. The flanking sequence context can affect the stability and conformation of the DNA molecule and changes in the distribution of positive and negative charges along the flanking sequence can lead to changes in genomic function. Not only does the immediate flanking sequence influence the position and the nature of mutations, but the effects reach up to 2000 nucleotides. Such effects fade with distance and are called "distance-decaying" relationships (Elango et al., 2008)

To understand the fusion and delocalization of pi-electrons in purine stretches, it is important to consider the

aromaticity of purines. Purines (nucleotides A and G) have a fused double ring consisting of a hexagon fused with a pentagon. Aromaticity occurs via the fusion of pi-electrons into a ring, such as in benzene, and results in their delocalization within a ring. Purines A and G having two aromatic rings (hexagon and pentagon) are more aromatic than pyrimidines C and T, which contain only a single hexagon. The order of aromaticity of the four nucleotides in DNA (adenine, guanine, cytosine, and thymine) is A>G≥C>T (Cysewski, 2005; Cysewski and Szefler, 2010; Krygowski et al., 2014).

Quantifying the extent of aromaticity in nucleotides is a challenging task, as there is no universally accepted definition or method for measuring aromaticity. However, here are a few potential surrogate measures that have been proposed or used in the literature to assess the aromatic character of nucleotides:

1. Hückel's rule: This rule, developed by Erich Hückel, states that an aromatic molecule must have a planar ring with 4n + 2 π electrons, where n is a non-negative integer. Based on this rule, A and G are considered more aromatic than C and T because they have more π electrons in their ring structures.
2. Nucleus-independent chemical shift (NICS) calculations: The NICS values of nucleotides can be calculated using quantum chemical methods, and can be used as a surrogate measure of their aromaticity. A and G are expected to have lower NICS values compared to C and T, indicating their more aromatic nature.

It is well known that DNA bases stick to each other in the base stack in the double helix. The stacking of DNA bases is based on their aromaticity and is responsible for the helical structure of DNA: the sugar-phosphate backbone is negatively charged and, accordingly, aims to stretch in a straight line due to the repulsion of negative charges of phosphates. These phosphates are hydrophilic and well-hydrated. The inner part of the base stack, on the other hand, is hydrophobic and aims at minimizing its bordering surface with water. Without the sugar-phosphate backbone, it would form a globule. But due to the covalent bonding of the base stack, trying to avoid water, and the backbone that is hydrated and stretches long, a firm and geometrically perfect double-helical structure is formed.

The effect of stacking as the longitudinal attraction of the nucleotide occurs due to the interaction of pi-electron rings of aromatic carbon-nitrogen rings in the bases (Sponer et al., 2013). Stacking energy between bases is approximately and respectively: purine-purine =2-3 kcal/mol, purine-pyrimidine = 1-2 kcal/mol, and pyrimidine-pyrimidine = 0.5-1 kcal/mol (Punnoose et al., 2022). DNA melting and other spectroscopic studies suggest that the aromatic electrons of purines are delocalized by extended conjugation over the stacked bases (Cantor and Schimmel, 1980). Charge transfer experiments also suggest aromatic electron delocalization along stacked purines in purine stretches (Murphy et al., 1993; Venkatramani et al., 2011).

There is no consensus in the literature on the electron behavior in purine stretches. To shed light on the matter, it's crucial to consider that electrons are best explained by quantum-chemical models. The Schrodinger wave function, for instance, determines the probability of an electron's location in space - even a highly localized electron has a small chance of being found elsewhere. Hence, the question of electron delocalization should be approached quantitatively. It's well-established that the electrons in the aromatic rings of purines are delocalized within each pi-ring. And due to stacking, these pi-rings fuse, resulting in merged electron clouds. The electrons of adjacent stacked purine bases are likely delocalized across at least these two bases. However, the extent of this delocalization and whether electrons can freely move or tunnel along the purine stretch spanning multiple bases is still unclear.

The role of electron charge transfer in DNA repair and protection is also of practical importance, as the delocalization of electrons in purine stretches protects against chemical and radiation damage. Indole structures possess special aromaticity due to their unique molecular geometry and bonding pattern. The atoms in the ring are arranged in such a way that they follow Hückel's rule for aromaticity, which states that a planar, fully conjugated ring system with 4n + 2 pi electrons (where n is a non-negative integer) exhibits aromatic properties. The nitrogen atom in the indole ring also contributes to its stability by creating a conjugated system

with the adjacent aromatic ring. This results in a molecule with high stability, making it an important component in many biologically active compounds.

Purines (A, G) can exist in the enol-amine form (often referred to as the imidazole form) and keto-imine forms. The enol-amine form is more stable and biologically significant than the keto-imine form. This enhanced stability is attributed to the fully conjugated ring system in the enol-amine form, which adheres to Hückel's rule and becomes aromatic. The presence of a delocalized pi-electron cloud within this fully conjugated ring system results in an aromatic molecule. In contrast, the keto-imine form lacks full conjugation in its ring system due to a broken double bond between the nitrogen and carbon atoms, rendering it non-aromatic. The pi-electron cloud in this form is more localized.

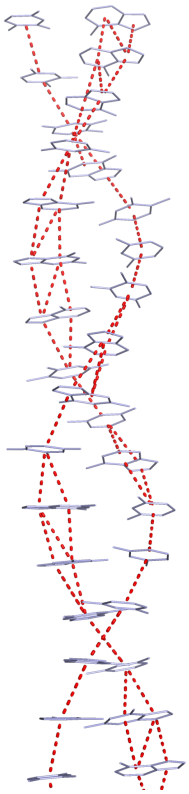


Fig.1 [Wires] Basestack with electron wires.

The DNA structure enables long-distance electron transfer via stacked pi electron rings, primarily in purines, acting as "electron wires". This is due to pi-pi interactions among the overlapping, delocalized electrons in these aromatic rings. Complementarity in DNA ensures that a stretch of purines (each with two pi-rings) in one strand always pairs with a stretch of pyrimidines (single pi-ring) in the complementary strand. This arrangement inherently forms a "double wire" in the purine strand and a "single wire" in the corresponding pyrimidine stretch, facilitating electron transport.

Potential roles for stretches of purines in DNA are intriguing to consider, albeit they currently remain under-explored. The inherent "double wire" electron transport system in these purine stretches, due to their stacked pi electron rings, might theoretically enable long-range electron transfer, which could be implicated in processes such as DNA repair. Similarly, the positioning and presence of these purine sequences could hypothetically influence the three-dimensional structure of DNA, potentially affecting gene regulation. They might also play a part in moderating the interactions between DNA and various proteins or even be implicated in cellular redox reactions. These are conjectural roles, and the precise functions of purine stretches in terms of electron transport within DNA remain open for exploration.

The concept that purine stretches or "wires" in intergenic regions could encode functional information presents a fascinating hypothesis in the context of gene regulation and chromatin dynamics, thereby potentially influencing cellular responses to environmental cues.

Intergenic regions, often considered "junk" DNA, are increasingly recognized as functionally important, harboring elements such as enhancers, silencers, and insulators that regulate gene expression. If purine stretches within these regions were acting as electron transport highways, it could add a new dimension to our understanding of these areas.

Specifically, purine stretches might influence chromatin structure and dynamics in a sequence-specific manner. Chromatin, the complex of DNA and proteins that make up chromosomes, is not static; it's continually remodeled, affecting which genes are accessible for transcription. The electron transport capability of purine wires might contribute to this dynamic process. For example, they could potentially help drive the energy-dependent reactions that reposition nucleosomes, influencing gene accessibility and, therefore, expression.

Further, if the purine wires can sense changes in the cellular environment, which could affect electron transport - this could modulate their impact on chromatin structure and gene expression. Thus, the purine wires could play a role in the cell's response to environmental changes, perhaps even forming a part of what could be considered the cell's thinking process, dynamically adjusting gene expression in response to its environment.

The intriguing idea of purine-based "electron wires" in DNA, and the potential interplay of their induced magnetic fields, offers a novel perspective on the dynamic mechanisms governing chromatin structure and gene regulation.

Purine stretches in DNA, given their stacked pi electron rings, may theoretically create an "electron wire" system, potentially driving energy-dependent reactions within chromatin remodeling. By affecting the positioning of nucleosomes, these electron wires could dictate gene accessibility and thus, their expression.

Simultaneously, the spinning of these electron wires around their internal axes, attributed to circulating pi electrons in aromatic purine rings, could generate local magnetic fields, a phenomenon observed in aromatic systems. These magnetic fields could interact with the cellular environment and magnetic fields from other DNA sequences. This interaction could lead to magnetic attractions or repulsions, influencing the spatial organization and dynamics of DNA.

This interplay of electron transport and magnetic fields could provide a novel layer of gene regulation. For instance, magnetic fields could bring distant enhancer regions closer to their target genes or segregate repressed chromatin areas, adding to the dynamic nature of gene expression.

The spinning of aromatic electrons in DNA in the magnetic field leading to DNA magnetism was discovered by Bliumenfeld et al (Bliumenfeld and Benderskii, 1960) and recently reviewed by Lee et al. (Lee et al., 2011). One of the magnetic field effects on an aromatic molecule is the aromatic ring current, which can be thought of as the induction of a circular current of π-electrons upon the application of a magnetic field perpendicular to the π-system of the molecule. The properties of organic molecules can be influenced by magnetic fields, and these magnetic field effects are diverse. They range from inducing nuclear Zeeman splitting for structural determination in NMR spectroscopy to polaron Zeeman splitting organic spintronics and organic magnetoresistance (Kudisch et al., 2020).

Consider that DNA packed in chromatin is a living perpetually self-organizing liquid crystal. It is a living self-organizing gel, the approximate chemical content of which is summarised in Fig.2 [Nucleoplasm]. The role of this perpetually self-organizing chromatin is to receive outside information, interpret it using the genome as a program, and regulate gene expression accordingly. Therefore, chromatin functions as a living computer, or the brain of the cell. As is seen from Fig.2 [Nucleoplasm], DNA and proteins comprise nearly 20% of nucleoplasm making it a thick perpetually self-structuring living gel. The DNA sequence defines electron wire patterns through which the genomic sequence guides the perpetual self-organization of chromatin as electrically and

magnetically active, perpetually self-organizing chromatin liquid crystal. This way the genome thinks and reacts to the external stimuli.

Component	w/w %	mM
Water	78%	55000
Histones	**8%**	2
Soluble nucleoplasmic proteins	**7%**	2
Non-histone chromosomal proteins	**4%**	1
DNA	**1%**	0.0
Other small molecules and ions	1%	-
K+	0.4%	140
RNA	0.2%	0.2
Na+	0.02%	10
Glutamate	0.02%	5
NAD+	0.02%	0.3
Cl-	0.01%	4
ATP	0.01%	3
Glycine	0.01%	2
Inorganic Phosphate, Aspartate, Glutamine, Alanine, ea	0.01%	1
Glucose, Mg2+, GTP, Arginine, Serine, Proline, ea	0.01%	0.5
Proteins and DNA subtotal:	**19%**	

Fig. 2 [Nucleoplasm] Estimated composition of the nucleoplasm in fibroblasts. *Fibroblasts are taken as typical widely studied somatic cells. The table is an original estimate based on literature and machine intelligence tools.*

To verify this hypothesis, one of the most accessible ways is to look at the function of purine stretches in the genomic sequence and functional genomic data. Here we ventured to explore computationally purine stretches as containers for contiguous electron clouds that work as oscillators or resonators. The size of these oscillators and resonators would define their oscillation frequency. The oscillations in these resonators could be stretching or twisting, Fig.3 [Twist]. The frequency and therefore the function of these oscillations would depend on the length of the purine stretch. In this work, we ventured to explore the function of the length of purine stretches in the public functional genomic data.

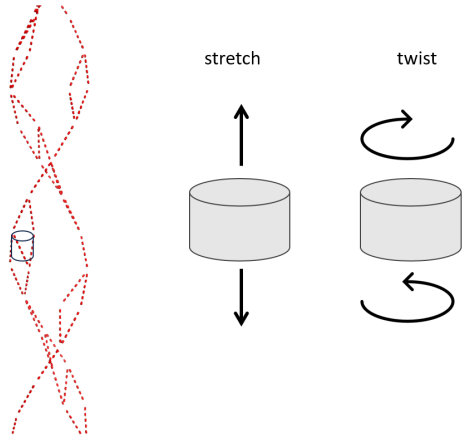
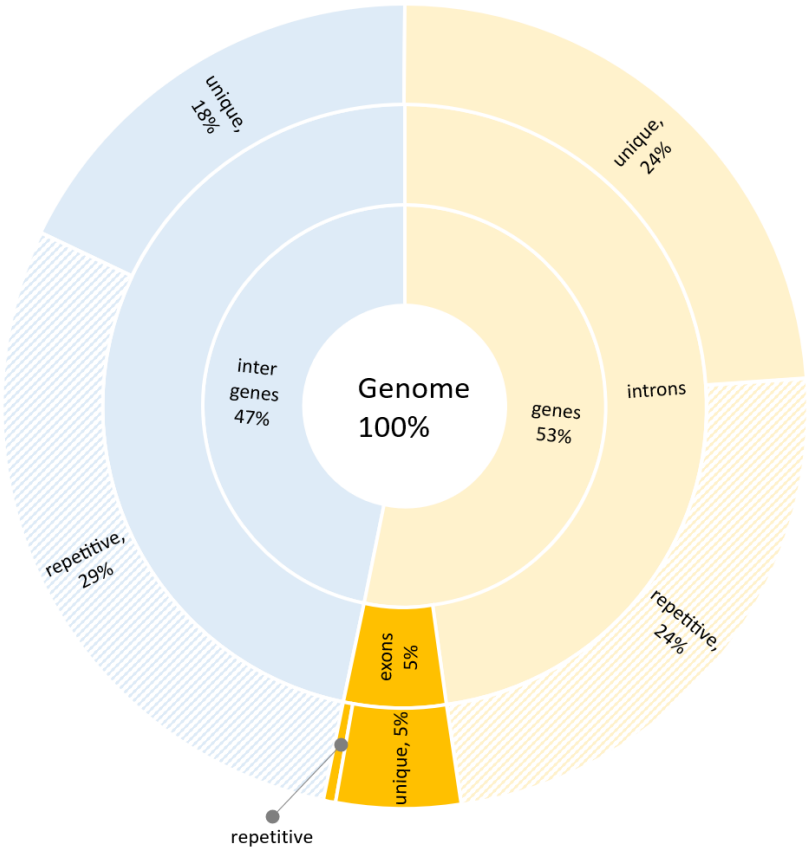


Fig.3[Twist] Possible stretching and twisting oscillations in aromatic electron wires in purine stretches.

To produce better computational signal and reduce noise we focused our analysis on the areas of the genome where we expected the function of purine stretches to be most revealed and unhindered by other functions. Since protein-coding exons are under strong pressure to produce correct proteins, we don't expect these sequences to reveal the function of purine stretches in dynamic chromatin self-organization. Therefore we explored this in noncoding parts of the genome (to be more exact in non-protein-coding parts of the genome). Initially, we searched and couldn't find a comprehensive summary of noncoding areas of the genome, so we calculated the coverage of the genome by coding and noncoding parts based on the current genome assembly Fig.4[Intergenes].



	% of the genome	% of repetitive in each
intergenes	47%	61%
genes: introns	48%	50%
genes: protein-coding exons	5%	10%

Fig.4 [Intergenes] Proportion of exons, introns, and intergenic regions in the human genome summarized from Ensembl Annotation (release 108, Oct. 2022)

As shown in Fig.4 [Intergenes], we found that protein-coding sequences (protein-coding exons) occupied 5% of the genome. The remaining 95% are not coding for proteins. 10% of protein-coding exons are repetitive (as judged by Repeat Masker). Genes occupied 53% of the genome. In addition to the above-mentioned protein-coding exons (5% of the genome) genes contained introns (48% of the genome). Half of the intronic sequence is repetitive (24% of the genome). Intergenic sequences (intergenes) comprised 47% of the genome, over half (61%) of which were repetitive (29% of the genome). Note that only an estimated 50% of genes are expressed in the same somatic cell at the same time.

Currently known functions of introns include enhancing gene expression by increasing the efficiency of transcription and enabling alternative splicing, leading to the production of multiple protein isoforms from a single gene. Introns can also serve as regulators of gene expression at various levels and influence the positioning of nucleosomes, thereby affecting chromatin structure and accessibility in a sequence-specific manner. They play a role in maintaining genome integrity through their involvement in recombination and repair processes, and some even exhibit ribozyme activity. Introns may host sequences that give rise to microRNAs (miRNAs), contributing to post-transcriptional regulation, and facilitating evolutionary innovation by allowing greater flexibility in exon shuffling. Moreover, they can serve as a source for the creation of new genes through

evolutionary mechanisms and may influence translation efficiency through their impact on the splicing and translation process. We suggest that the main functions of introns are still undiscovered. We find it very unlikely that known functions of introns justify their length. Since the cells spend much of their resources on maintaining introns that occupy 48% of the genome, they must be playing yet unknown important functions. Since introns are transcribed and this RNA is likely playing important known and yet unknown functions, the sequences of introns are under evolutionary pressure in a different way than intergenic sequences. In our various analyses, we noticed very pronounced differences in DNA patterns between introns and intergenes. Our initial analyses demonstrated signals from purine stretches were much stronger in intergenic sequences than in introns, so the bulk of our analyses was done in intergenes.

Materials and Methods

Dataset of mutations in cancer

The data source used in this work was the CosmicNCV.tsv dataset, which was downloaded from the COSMIC website (https://cancer.sanger.ac.uk/cosmic/download). This dataset contains mutations in the non-coding part of the genome in cancer patients. For each mutation, information is given about the patient (identifier, organ where the cancer occurred, type of cancer, cancer histology), and position of the mutation in the genome. The mutations are paired per patient: for each patient a reference allele from the normal tissue and the mutant allele from the cancer tumor) are available in the dataset. The dataset contains a total of 18466909 mutations and 158422 patients (the average number of mutations per patient is 117).

Total lengths of genes, introns, exons, and repetitive DNA.

The human genome version GRCh38 (release 108) was downloaded from the Ensembl site (https://ftp.ensembl.org/pub/release-108/fasta/homo_sapiens/dna/). The annotation of the human genome in **gtf** format was also downloaded from the Ensembl website (https://ftp.ensembl.org/pub/release-108/gtf/homo_sapiens/). Genomic regions were analyzed using the following algorithm. In the first step, the file with the annotation of the genome was parsed, and the coordinates of the beginning and end of all genes, all exons, and all UTRs for chromosomes 1-22 + X, Y were written out in separate dictionaries. The coordinates of introns were obtained by finding the gaps between exons, and the coordinates of intergenic sites were obtained by finding the gaps between genes. Since genes in the genome overlap, it was necessary to combine all overlapping coordinates and remove nested coordinates so that the same coordinate was not counted twice when calculating site lengths. This was done using the merge_regions() and drop_nested() functions from Genome_Regions_Analysis.ipynb notebook. Then, after merging and clearing of nesting, the sum of coordinates within all the lists was calculated. The fraction N of nucleotides was calculated by finding the coordinates of all features in the genome and counting the fraction N within these sites. The type of exon in which N is present was also taken from the **gtf** annotation.

Getting the type of mutation and length of the purine stretch

In the first stage of the work, information was obtained on the effect of the mutation on the purine chain, which includes the mutated nucleotide. To do this, for each mutation, its position on the chromosome was found. Then, the chromosome segment flanking this mutation was taken with a shoulder length of 60 nucleotides (60 on the left, 60 on the right = total length of the segment of 120 nucleotide pairs). All sequences were converted to the purine code (purines A, G = R; pyrimidines T, C = Y). If the mutated nucleotide was a purine, the analysis continued on this DNA strand. If the mutated nucleotide was a pyrimidine, the analysis continued on the complementary DNA strand (R was replaced by Y and Y by R). If both the reference and mutant alleles were of the same type (purines or pyrimidines), such a mutation was labeled as synonymous. Otherwise, the type of mutation was determined as follows: if the mutation occurred within the purine chain (purines to the left

and right of the mutated nucleotide - RRR), then the type of mutation was designated Break. If there were pyrimidines to the left and right of the mutation, such a mutation restored the purine chain, and its type was designated Join (YRY). If there was a pyrimidine on the left of the mutation and a purine on the right, such a mutation switched the purine chain and its type was designated LeftFlip (YRR). If there was a purine to the left of the mutation and a pyrimidine to the right, such a mutation was designated as RightFlip (RRY). After determining the type of mutation, the length of the purine chain, which includes the mutated nucleotide, was determined before and after the mutation (Ref_Chain_Length and Alt_Chain_Length, respectively). For this purpose, the length of the left and right arms (purine sequences to the left and right of the mutated nucleotide: left_shoulder and right_shoulder) was first calculated. Then, the lengths of the reference and mutant chains were calculated using the following formulas. For Break type (RRR): Ref_Chain_Length = left_shoulder + right_shoulder - 1, Alt_Chain_Length = left_shoulder - 1 if left_shoulder >= right_shoulder else right_shoulder - 1. For Join type (YRY): Ref_Chain_Length = left_shoulder - 1 if left_shoulder >= right_shoulder else right_shoulder - 1, Alt_Chain_Length = left_shoulder + right_shoulder - 1. For LeftFlip type: Ref_Chain_Length = left_shoulder - 1 if left_shoulder - 1 >= right_shoulder else right_shoulder, Alt_Chain_Length = left_shoulder if left_shoulder + 1 >= right_shoulder else right_shoulder - 1. For the RightFlip type: Ref_Chain_Length = left_shoulder if left_shoulder + 1 >= right_shoulder else right_shoulder - 1, Alt_Chain_Length = left_shoulder - 1 if left_shoulder >= right_shoulder else right_shoulder. Then, for each mutation, the characteristic of the relative change in the purine chain (Rel_Change_Pu) was calculated using the following formula: (x-y)/(max(x,y)), where x is Ref_Chain_Length and y is Alt_Chain_Length. Then, all found characteristics: mutation type, Ref_Chain_Length, Alt_Chain_Length, Rel_Change_Pu, as well as the primary purine chain sequence in which the mutation is included (Primary_Sequence) were recorded in the dataframe under analysis (CosmicNCV).
Also, it was additionally determined whether the mutation is included in the transcription start site (TSS). For this purpose, the origin coordinates of all genes were listed from the annotation file (Homo_sapiens.GRCh38.108.gtf). The site 400 nucleotides before the start of the gene was taken as the transcription start site (TSS). Then, for each mutation, it was determined whether it belongs to one of the transcription start sites, and this information was also added to the dataframe being analyzed (as well as information about which gene that transcription start site belongs to, if the mutation belongs to it).
All the above procedures were done for each chromosome individually using the SNPU_w_syn.py script. The results for all chromosomes were then combined into one common dataset.

Relative change of the Purine stretch length analysis

The analysis of the obtained dataframe was performed in Jupyter Notebook, using pandas, numpy, scipy, matplotlib, and seaborn libraries. The basic statistics of the dataframe (mean, standard deviation, median, and quartiles) were obtained using the pandas describe() method. The statistical significance of differences in Purine_stretch_length was calculated using the t-test (ttest_ind() from scipy.stats).

Purine stretch break vs join analysis, BRETORA.

The Break/Total ratio (BRETORA) as a function of Purine_stretch_length (Wire_Length) and mutation frequency (Prevalence) was measured using the following algorithm. First, the number of times each mutation occurs in the dataset was counted (using value_counts() and the unique mutation identifier HGVSG). Then the following analysis was performed for each Prevalence from 1 to 8. From the dataframe containing mutations with a given Prevalence, mutations in the purine chain of a certain length (3 to 17) were filtered out. Then, for these mutations, the number of breaks and the ratio of breaks to all mutation types (BRETORA) were counted. Thus, for each combination of Prevalence and Wire_length values, the ratio of breaks to all other mutations (BRETORA) was obtained. Based on this, a summary table was made and graphs (heatmap, lineplot) were plotted using the Seaborn library.

Randomized control

In order to check the significance of the result, a random control was performed. To do this, the coordinates of all genes in the human genome were first written out from the annotation file. Then, 1000000 random numbers ranging from 1 to the chromosome length (mutation position) were simulated for each chromosome. It was then checked to see if this mutation fell into the position of one of the genes that were found in the previous step. If it does not, then it is an intergenic mutation and can continue to work with it. Next, a random nucleotide was chosen where the substitution occurred. Then, all the simulated mutations were combined into one dataframe, and in this dataframe, all the steps described in the previous paragraphs were performed, until a summary table containing the BRETORA values depending on Prevalence and Wire Length was obtained.

Cancer Prevalence analysis, PrevRa.

The BRETORA ratio of Prevalence 2+ mutations to BRETORA of Prevalence 1 mutations was analyzed. This characteristic was called PrevRa. This analysis was performed for the 10 most frequent cancer types in the dataset. For each cancer type, the analysis described in the BRETORA analysis item was first performed. Then, the PrevRa ratio was calculated from the summary table and a heatmap was constructed reflecting this PrevRa value for each Purine_stretch_length in each cancer type.

Results

Study design

In the background section, we explained that purine stretches harbord a collectively shared aromatic electron ring stack that possesses unique electromagnetic properties. We hypothesized that purine stretches in intergenic regions play a significant function in the process of continuous sequence-specific restructuring of chromatin and this way perform complex gene regulation functions. Here we ventured to test this in public functional genomic data.
The Cosmic dataset (https://cancer.sanger.ac.uk/cosmic)(Tate et al., 2019) is an ideal dataset for such analysis. It is curated and of high quality. It contains 4.7M mutations and each mutation is represented by a pair of alleles for each patient - an allele from the normal healthy tissue (Normal, Ref_Chain, reference chain) and an allele from a cancer tumor (Cancer, Alt_Chain, alternative chain). The fact that cancer and normal alleles are present in the same patient and the reference allele serves as a healthy control within the same patient, makes these data very information-rich and reduces the signal-to-noise ratio. Also very helpful is that the cancer tumors were rigorously characterized. We used the sequence surrounding the mutations to measure the length of the purine stretches, Fig. 5 [Stretch].

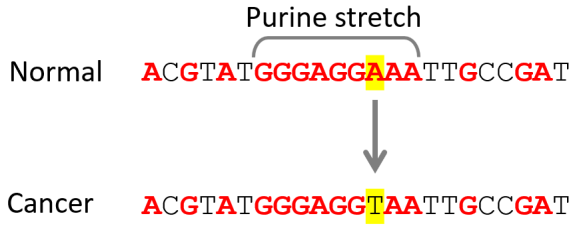


Fig. 5 [Stretch] An example of shortening of Purine_stretch_length in cancer. A purine stretch is defined as a string of purines (A, G) in either DNA strand framed on both sides by at least one pyrimidine (C, T). A mutation that substitutes a purine with a pyrimidine shortens a purine stretch.

Since many of the cancer mutations are typical and essential for the clonal evolution of cancer, if they have a preference for shortening or lengthening purine stretches, this would provide evidence for the functional

importance of their length. Therefore, we analyzed the influence of cancer mutations on the length of purine stretches. In the initial analyses, we separated all genomic sequences into exonic (for protein-coding exons), intronic, and intergenic. Our initial analyses demonstrated that all influences of cancer mutations on purine stretch lengths were substantially more pronounced in intergenic regions than in genes (introns and exons). We interpreted this in such a way that genes that are transcribed into RNA are under evolutionary pressure to perform functions related to RNA while intergenic regions (intergenes) likely function to guide continuous reorganization of chromatin in a sequence-specific manner. Therefore, we limited our consequent analyzes to intergenic regions.

On average, cancer mutations avoid purine stretches

The dataset contained a pair of alleles for each patient - an allele from the normal healthy tissue and an allele from a cancer tumor. The Purine_stretch_length was averaged for all SNPs and plotted as Y axis. "Purine_stretch_length" was defined as the length of the purine stretch encoding the mutation in either strand of DNA and limited by pyrimidines. A total of 4.7M mutations were analyzed. Of these, the most mutations are of the Break type (1.5M), followed by Join type (1.1M), and mutations that don't change the Purine_stretch_length (2.1 M). For each mutation, the length of the purine stretch, which includes the mutated nucleotide, was measured. The average Purine_stretch_length was observed to be shorter in cancer than in normal tissue (p-value < 0.0001 by paired t-test), **Fig. 6**.

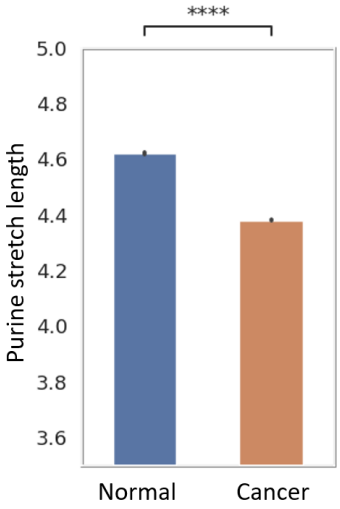


Fig. 6 Intergenic purine stretches are shortened in cancer. Mutations in the dataset were represented by a pair of alleles for each patient - an allele from the normal healthy tissue and from a cancer tumor. Purine_stretch_length was averaged for all mutations and plotted as Y axis. Purine_stretch_length was defined as the length of the purine stretch containing the mutation in either strand of DNA and framed (interrupted) by pyrimidines. The average Purine_stretch_length was observed to be shorter in cancer than in normal tissue (p-value < 0.0001).

Mutations of the purine stretches in promoters

Among intergenic regions gene promoters upstream of the transcription start site are known to be functional and conserved in evolution since they contain a substantial part of the information of when and where the gene is expressed. We looked at the intergenic part of the promoters (400 bp upstream of the transcription start site). For the outcome metric, we used Rel_Change_Pu (relative change in Purine_stretch_length), which was calculated using the following formula: (normal-cancer)/(max(normal, cancer)), where "normal" and "cancer" are corresponding Purine_chain_lengths. Rel_Change_Pu was measured separately for each type of cancer (**Fig. 7**).

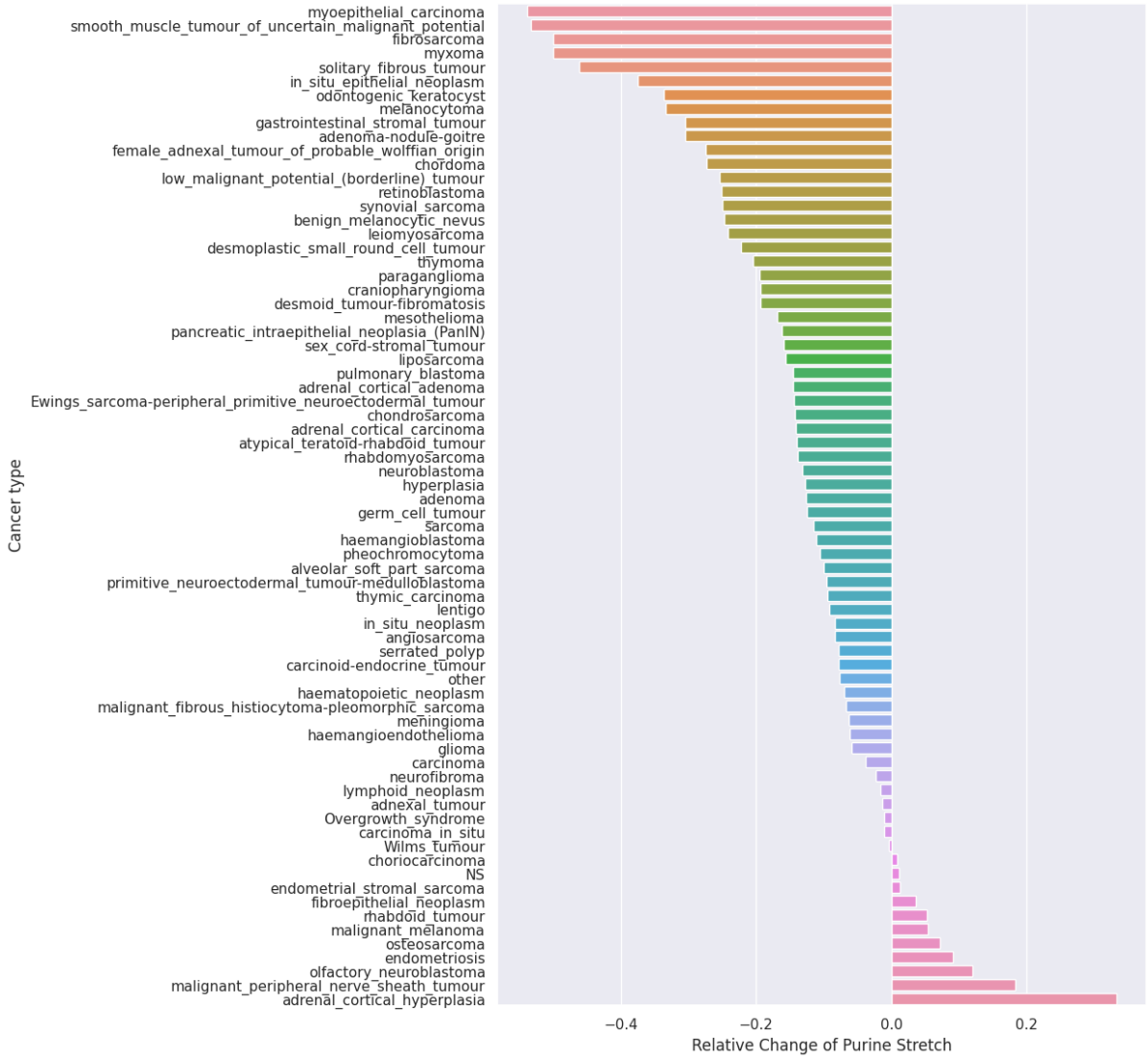


Fig. 7. Distribution of the relative change in the purine chain in different types of cancer. The x-axis is the relative change in the purine chain (Rel_Change_Pu), and the y-axis is the cancer type. Rel_Change_Pu was calculated using the following formula: (normal-cancer)/(max(normal, cancer)), where "normal" and "cancer" are corresponding Purine_chain_lengths. A positive value indicates a shortening of the purine chain in cancer compared to normal tissue, while a negative value indicates an extension.

In this analysis, Rel_Change_Pu varied greatly depending on the type of cancer. In most cancer types, Rel_Change_Pu was negative, which means that the purine chains in the cancer were elongated. However, there are also types (e.g., melanoma) in which purine chains are more frequently shortened.

The comparison of Purine_stretch_length with the prevalence of each mutation in cancer

Next, we explored two metrics: Purine_stretch_length, the primary measure of the purine stretch, and the prevalence of each mutation in cancer, defined by the number of patients within the dataset where the specific mutation occurred. For context, many mutations in the dataset were unique, occurring only once, while others appeared twice (i.e., in 2 out of 1,233 patients) or in a greater number of patients. Mutations with a prevalence of 1 largely consist of random occurrences, likely resulting from cancer rather than causative. Mutations with prevalences of 2, 3, or 4 typically correspond to causative cancer factors but are not under robust positive

selection by clonal cancer evolution. In contrast, mutations with a prevalence greater than 4 are typical cancer mutations, subject to strong selection.
In purine stretches point substitution mutations can result in three types of alterations: breaks, joins, and synonymous substitutions. We implemented the BRETORA metric to quantify the frequency of breaks, with "BRETORA" standing for "BREaks TO all RAtio." This metric is defined as the ratio of the number of breaks to the total count of all mutations.
In examining purine chains, we aimed to measure the frequency of breaks in relation to chain length, specifically focusing on the functional influence we believe is associated with electron fusion in the aromatic rings of purines. However, longer purine stretches tend to be more frequently affected by random mutations, a factor we needed to exclude to obtain an accurate measurement. To normalize the data and exclude the random influence of length, we performed a random control by simulating 1,000,000 mutations. A summary table was then produced, presenting BRETORA values as functions of both Prevalence and Purine_stretch_length. These values were normalized to the random ones, creating a new metric called BRETORANRA, short for "BRETORA Normalized to RAndom." Figure 8 shows a graph depicting the dependence of this parameter on chain length for Prevalence 1 and 2+ mutations, reflecting our targeted analysis of chain length's influence.

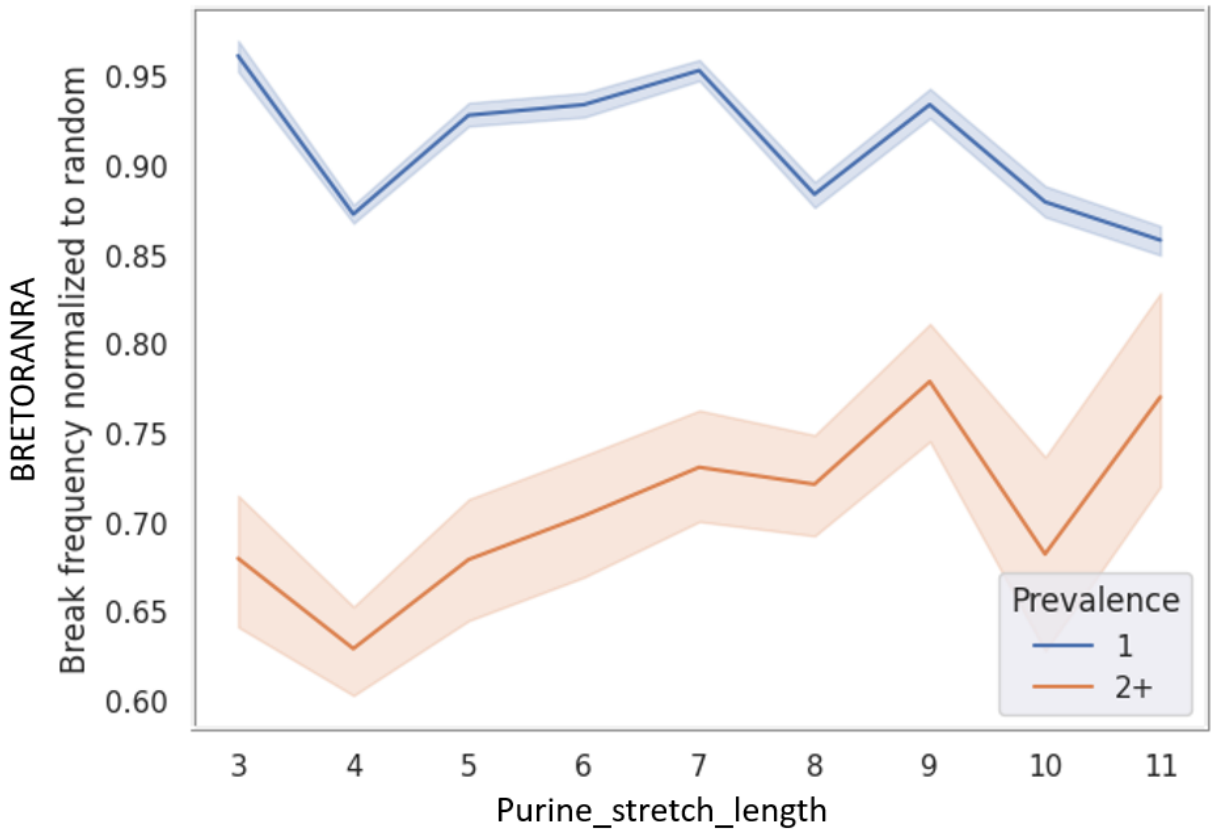


Fig. 8. Typical cancer mutations avoid purine stretches. Plotted is Break frequency normalized to random (BRETORANRA) versus mutation Prevalence. The dotted blue line represents mutations with Prevalence 1; the dotted purple line represents mutations with Prevalence 2+. Shades represent a 95% percentile interval.

The findings presented in Fig. 8 reveal that common cancer mutations with a prevalence of 2+ exhibit a notably lower normalized frequency of breaks when contrasted with primarily random mutations characterized by a prevalence of 1. This difference is particularly pronounced in shorter purine stretches.

Discussion

Purine stretches, composed of adenine (A) and guanine (G), serve crucial roles in DNA binding and gene

regulation, and any mutations or alterations within these regions could potentially influence cancer development. This is underpinned by the aromatic nature of purines and their electron delocalization, factors that influence DNA structure, and their specific stacking energies that could make these stretches more susceptible to specific mutations or alterations relevant to cancer.

The delocalization of electrons in purine stretches also plays a protective role against damage, and any disruption to this could make DNA more susceptible to damage, leading to mutations commonly found in cancer. Additionally, the flanking sequences' influence and the distance-decaying relationships extending up to 2000 nucleotides might impact how a mutation in or near a purine stretch affects adjacent regions, possibly relating to the proliferation or suppression of cancerous cells.

This complexity is further highlighted by the presence of more stable enol-amine forms compared to keto-imine forms due to aromaticity, which might have biological implications in cancer. Understanding the electron behavior in purine stretches through quantum-chemical models could provide novel insights or therapeutic strategies, emphasizing the multifaceted nature of purine stretches in cancer.

Building upon these concepts, the investigation into the role of purine stretches in intergenic regions using the Cosmic dataset has revealed a pronounced influence of cancer mutations on purine stretch lengths, notably more observed in intergenic regions. This discovery indicates that intergenes likely guide the continuous reorganization of chromatin in a sequence-specific manner, a complex phenomenon underpinning diverse gene regulation mechanisms.

The study also found a distinct shortening of purine stretches in cancerous tissue compared to normal tissue, reflecting potentially different strategies that cancer cells might use to modulate gene expression. A compelling finding was the tendency for common cancer mutations to avoid purine stretches, hinting at a possible protective mechanism or structural significance of these regions.

These findings collectively support the hypothesis that purine stretches in intergenic regions have functional importance, especially in complex gene regulation functions and chromatin restructuring. The variations in purine stretches associated with different cancer types may contribute to unique disease pathways, adding to the complexity of cancer as a multifactorial disease.

Overall, the results provide compelling evidence for the importance of purine stretches in gene regulation and suggest a complex relationship with different types of cancer. This research seems crucial in unraveling these complex interactions and finding novel avenues for treatment or prevention, laying the groundwork for future studies that may offer novel insights into cancer biology and therapeutics.

Acknowledgments

We thank Michael Rempel for the discussion. The work was funded exclusively by Max Myakishev-Rempel.

REFERENCES

Bliumenfeld, L.A., Benderskii, V.A., 1960. Magnetic and dielectric properties of high-ordered macromolecular structures. Nauk SSSR.

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry: Part I: The Conformation of Biological Macromolecules. W. H. Freeman.

Cysewski, P., 2005. An ab initio study on nucleic acid bases aromaticities. Journal of Molecular Structure: THEOCHEM 714, 29–34. https://doi.org/10.1016/j.theochem.2004.10.030

Cysewski, P., Szefler, B., 2010. Environment influences on the aromatic character of nucleobases and amino acids. J. Mol. Model. 16, 1709–1720. https://doi.org/10.1007/s00894-010-0806-5

Elango, N., Kim, S.-H., Vigoda, E., Yi, S.V., 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput. Biol. 4, e1000015. https://doi.org/10.1371/journal.pcbi.1000015

Krygowski, T.M., Szatylowicz, H., Stasyuk, O.A., Dominikowska, J., Palusiak, M., 2014. Aromaticity from the viewpoint of molecular geometry: application to planar systems. Chem. Rev. 114, 6383–6422.

https://doi.org/10.1021/cr400252h
Kudisch, B., Maiuri, M., Moretti, L., Oviedo, M.B., Wang, L., Oblinsky, D.G., Prud'homme, R.K., Wong, B.M., McGill, S.A., Scholes, G.D., 2020. Ring currents modulate optoelectronic properties of aromatic chromophores at 25 T. Proc. Natl. Acad. Sci. U. S. A. 117, 11289–11298. https://doi.org/10.1073/pnas.1918148117
Lee, C.H., Kwon, Y.-W., Jin, J.-I., 2011. Electrical and Magnetic Properties of DNA, in: Materials Science of DNA. unknown, pp. 121–162. https://doi.org/10.1201/b11290-6
Murphy, C.J., Arkin, M.R., Jenkins, Y., Ghatlia, N.D., Bossmann, S.H., Turro, N.J., Barton, J.K., 1993. Long-range photoinduced electron transfer through a DNA helix. Science 262, 1025–1029. https://doi.org/10.1126/science.7802858
Punnoose, J.A., Thomas, K., Hayden, A., Banco, T., Halvorsen, K., 2022. Single-molecule quantification of individual base-stacking energies using a centrifuge force microscope. Biophys. J. 121, 283a–284a.
Sponer, J., Sponer, J.E., Mládek, A., Jurečka, P., Banáš, P., Otyepka, M., 2013. Nature and magnitude of aromatic base stacking in DNA and RNA: Quantum chemistry, molecular mechanics, and experiment. Biopolymers 99, 978–988. https://doi.org/10.1002/bip.22322
Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E., Speedy, H.E., Stefancsik, R., Thompson, S.L., Wang, S., Ward, S., Campbell, P.J., Forbes, S.A., 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 47, D941–D947. https://doi.org/10.1093/nar/gky1015
Venkatramani, R., Keinan, S., Balaeff, A., Beratan, D.N., 2011. Nucleic Acid Charge Transfer: Black, White and Gray. Coord. Chem. Rev. 255, 635–648. https://doi.org/10.1016/j.ccr.2010.12.010