Mutagenesis at non-B DNA motifs in the human genome: a course correction

McGinty RJ[1,2] and Sunyaev SR[1,2]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA.
[2] Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.

Abstract

Non-B DNA structures formed by repetitive sequence motifs are known instigators of mutagenesis in experimental systems. Analyzing this phenomenon computationally in the human genome requires careful disentangling of intrinsic confounding factors, including overlapping and interrupted motifs, and recurrent sequencing errors. Accounting for these factors eliminates all signals of repeat-induced mutagenesis that extend beyond the motif boundary, and eliminates or dramatically shrinks the magnitude of mutagenesis within motifs, contradicting previous reports. Non-B DNA makes only a modest contribution to mutation in the human germline compared with other intrinsic sources of mutation. G-quadruplex motifs are potent instigators of recurrent sequencing errors, but are only mildly mutagenic. Polymerase slippage is the dominant mutagenic mechanism affecting STRs, while polymerase template-switch is exclusively responsible for mutations in palindromic repeats, which perfect and extend inverted symmetry.

Main text

By nature of the genetic code, some repetitive DNA sequences are capable of intra-strand base pairing, while other motifs can allow non-Watson-Crick base pairing. This non-canonical base-pairing can stabilize secondary structures, including hairpins, triplex, Z-DNA and quadruplex structures. Non-B-form DNA structures are disruptive to DNA replication, with

possible consequences including slippage, template-switching, replication fork stalling and collapse, and double-strand breaks. Subsequently, repetitive sequences can cause misalignment during replication restart and homology-directed DNA repair. This can lead to any combination of genomic rearrangements, repeat tract-length changes and single nucleotide mutations within and beyond the motif[1]. Much of this was uncovered in experimental systems, where long repeat tracts model the genetic component of several neurodegenerative disorders.

While the human genome contains numerous repetitive elements, vanishingly few reach long lengths in healthy individuals. Nonetheless, the number and variety of repeats in the genome creates an opportunity to investigate mutational mechanisms affecting short repeats, by starting with a collection of mutations from large-scale sequencing efforts and measuring their enrichment surrounding repeat motifs. While several groups have performed similar analyses[2,3,4,5], we find strikingly different results after controlling for several key factors: namely, that there is abundant overlap between repeat categories, that motifs are clustered together, that repeats and their flanking regions have highly non-random nucleotide composition, and that repeats are responsible for recurrent sequencing errors. We find that mutagenesis induced by short repeats is limited to a subset of motifs and does not extend beyond the motif boundaries. Additionally, we find polymerase template-switch and slippage during replication are the dominant mechanisms contributors to this mutagenesis.

Results

Confounders of non-B motif analysis:

Sequence symmetry exists along three axes: direct symmetry (a sequence followed by itself), mirror symmetry (a sequence followed by itself in reverse) and inverted symmetry (a

sequence followed by its reverse complement). Notably, hairpin structures require inverted symmetry. Certain sequences may contain multiple symmetries. For example, AT di-nucleotide repeats contain direct, mirror and inverted symmetry. Other short tandem repeats (STRs) can also satisfy the requirements of G-quadruplex and Z-DNA motifs. We therefore generated a database of unique repeats by first separating out STRs by motif, and then ensuring that direct, mirror, inverted, Z-DNA, and G4 motifs did not contain overlapping coordinates with any other motif in the database **(Fig. 1, Fig. S1)**.

Given the large number of repeat motifs in the genome, many motifs lie in close proximity to one another by chance. In order to avoid mistaking mutations surrounding one motif for mutations that are actually within another motif, we ensured that flanking regions were free of any other discernible repeat motif, including transposable elements **(Fig. 1, Fig. S1)**. Additionally, we searched for the continuation of a motif past short interruptions **(Fig. 1)**. The advantage is two-fold: interruptions are not mistaken for flanking sequences, and mutations can be categorized by whether or not they perfect the motif.

Repetitive regions are defined by their unusual nucleotide content, but nucleotide composition alone has a large effect on mutation rate. There is a >80-fold difference between the mutation frequency of the lowest and highest trinucleotides due to 5-methylcytosine deamination, and a >10-fold difference between the lowest and highest non-CpG trinucleotide mutations[6]. By normalizing to the observed trinucleotide mutation frequency across the entire genome, we ensure that any observed increase in mutation frequency is not due to the effect of nucleotide composition alone.

Since repeats are known to cause replication errors *in vivo*, we surmised that repeats would likely cause sequencing errors *in vitro.* This is especially likely because the template is single-stranded and free to form secondary structures, and because *in vitro* synthesis lacks a

helicase to unwind those structures, as well as other fork-protecting factors. G-quadruplex structures have been shown to induce sequencing errors in both short-read Illumina sequencing and long-read Pac-Bio sequencing, which both employ sequencing-by-synthesis[7,8]. Increasing sequencing depth alone is not a cure-all for this issue, as repeats may induce highly-recurrent sequencing errors[7,9,10]. Furthermore, repetitive regions may also suffer from sequence alignment artifacts. In light of this confounding factor, we employed sequencing quality control metrics to identify mutagenesis signals likely to be affected by sequencing errors. In particular, we see that G-quadruplex motifs and STRs are heavily prone to errors **(Fig 2, Fig S2)**.

Elimination of repeat-induced mutagenesis at a distance:

Experimental studies have shown that long repeats can trigger mutagenesis beyond the motifs, due to the action of break-induced replication (BIR)[1]. Thus, we first examined mutation frequencies surrounding repeat motifs. We find no mutagenesis at a distance for motifs containing direct, mirror or inverted symmetry, in stark contrast to the appearance of widespread mutation prior to our corrective efforts **(Fig 2, Fig. S2)**. Likewise, there is no mutagenic signal attributable to Z-DNA motifs following the removal of TG dinucleotide repeats **(Fig 2, Fig. S2)**. Signals of mutagenesis surrounding G4 motifs are entirely driven by sequencing errors. Mutagenesis surrounding STRs that remains after quality filtering does not extend more than 3 nt beyond the motif, and rarely exceeds a two-fold increase in mutation rate **(Fig. 2, Fig. S2)**. This signal is modest compared with other endogenous mutational processes[11] and not due to undercounting recurrent mutations **(Fig. S3)**. The short range of this signal is explained by the difficulty of classifying every potential configuration of an interrupted motif **(Fig. 3) (see Methods)**, as we observe an excess of mutations and/or errors in this region that would correct interruptions, implicating polymerase slippage **(Fig. S4)**. In

essence, these regions are not true flanking regions, but rather should be considered part of the interrupted motif. The only long-range mutagenic signal that we observed, surrounding CG dinucleotide repeats, is due entirely to their presence within CpG islands, and is not related to the repeat itself **(Fig S2)**.

Mutagenesis within non-B motifs:

Having eliminated any signals of long-range mutagenesis, we next examined mutagenesis within repeat motifs. Experimental evidence suggests that a variety of mechanisms contribute to mutations within non-B motifs[1]. We find that only a subset of motifs are mutagenic within the motif, while sequencing errors are frequent. Mono- and di-nucleotide STRs appear prone to both mutations and sequencing errors, including interruption-perfecting errors, at longer motif lengths **(Fig 3, Fig. S5)**. This suggests polymerase slippage is a likely mechanism to explain mutations within STRs, whether *in vivo* or *in vitro*. This is further supported by the observation that mutations within STRs are highly enriched for 'mixed' loci, which contain both SNVs and indels in the population **(Fig S5)**, as polymerase slippage is known to generate frequent indels. Deletion of an interruption, along with a length-compensating insertion elsewhere in the motif, would be indistinguishable from a motif-perfecting mutation **(Fig 3)**, similar to an effect previously reported at compound STRs[12]. A high rate of interruption-perfecting errors would cause undercounting of novel true mutations in perfect STRs. This effect should be proportional to the observed rate of errors at known STR interruptions in the reference genome, and the resulting signal should remain modest.

Interrupted positions within inverted repeats mutate to restore uninterrupted inverted symmetry **(Fig 4)**. This cannot be explained by polymerase slippage, but is consistent with template switch during replication. This mechanism is well-studied in bacterial model systems, but thus far has only been observed in humans in a limited capacity[13,14]. We find that this

mechanism is exclusively responsible for mutations within inverted repeats, not only at interrupted positions within stems, but surprisingly also in the hairpin spacer **(Fig 4)**. Specifically, we see perfecting mutations of the first and last spacer positions, while the middle positions of the spacer are not mutable. This further supports a template-switch model rather than a mismatch repair model, as the spacer loops do not resemble typical DNA mismatches. In contrast, the immediate flanking positions are not mutable **(Fig 4)**. Thus, inverted symmetry would tend to extend inward towards the spacer, generating stronger inverted repeats, but not outward into the genome, limiting growth of symmetry. The frequency of motif perfection is dependent on stem length, and thus on the stability of the hairpin structure. In contrast, direct and mirror repeats show no signs of perfecting or extending symmetry **(Fig S6)**. However, it should be noted that homopurine mirror repeats, which may form H-DNA structures, are very rare in the genome following exclusion of STRs, precluding their separate analysis.

Finally, we observe that G-quadruplex motifs are instigators of recurrent sequencing errors, especially T>G mutations on the 5' end of spacers (sequences between G-runs) **(Fig 5)**. After accounting for errors, G-runs appear mildly mutagenic, while spacers do not **(Fig 5)**. This refutes earlier reports that single-strandedness predisposes spacers to mutation.

Discussion

Ultimately, after carefully avoiding issues confounding the analysis, we find that short repeat motifs in the human reference genome pose little mutagenic threat to the genome. All signals of mutagenesis surrounding short repeat motifs are eliminated after carefully removing confounding factors. The frequency of mutagenesis within motifs is modest in comparison with other mutagenic processes that are active in the human germline, including those driven

by sequence context[6] and regional variation[11]. This suggests that DNA replication and repair systems are generally capable of handling short repetitive sequences, as opposed to the simplistic *in-vitro* synthesis during sequencing. This resolves some tension in the notion that non-B DNA structures may serve functional roles in biology[15,16,17].

Furthermore, these results are actually in line with experimental systems, which demonstrate a repeat length threshold for instability.[1] Short repeats are mostly stable, with replication slippage causing small-scale expansions, contractions and occasional mutations. Completely derailing replication requires very long repeats, leading to extreme outcomes such as BIR. While experimental work demonstrates that long-range mutagenesis is a consequence of BIR, the more detrimental consequence is that of frequent genomic rearrangements[18]. For this reason, repeats capable of triggering BIR cannot be sustained for long in the germline, and thus are extremely rare in the genome and in the population.

Predicting rates of mutagenesis is crucial to characterizing causative mutations in congenital disease and driver mutations in cancer. The sequence context of adjacent nucleotides is broadly predictive of the rate and type of mutations, and increasing the length of the context window further improves the model's prediction[6]. Polymerase slippage mutations at STRs are amenable to this approach to an extent, though STR lengths can easily extend beyond the context window of such models. Interestingly, template-switch mutations at inverted repeats are invisible to simple context-based models. This is not only due to the length of the motifs, but also because the mechanism requires symmetry rather than nucleotide content for both the context and the outcome of the mutation. With tens of millions of short inverted repeats in the human genome, as much as 1% of the genome may subject to template-switch mutations. Thus, our study characterizes a widespread source of previously unpredictable mutations in the human germline.

Methods

Variants:

Human germline variants were downloaded from the gnomAD consortium[19], with version 3.1 consisting of 759,302,267 variants derived from 76,156 whole genome sequences. We filtered this set to single nucleotide variants (SNVs) with a maximum allele frequency of 0.01%. Germline variants of sufficient rarity have previously been shown to serve as a good proxy for mutations, since their composition has not been detectably altered by selective pressure[6]. gnomAD provides a recommended quality cutoff based on allelic imbalance and allele-specific VQSLOD, which is a machine-learning-derived sequencing quality metric based on a combination of strand bias, read position, mapping quality, and read depth[20]. gnomAD's passing VQSLOD score is -2.7 and above, which excludes 12.2% of SNVs, and we applied additional filters above VQSLOD 0 and 4, where indicated. Rather than deciding on a singular quality cutoff, which would necessitate a trade-off between over-filtering some sites and under-filtering others, we present mutation frequencies without quality control, and across this range of quality filter regimes. This highlights regions prone to potential sequencing quality artifacts (where the filtered and unfiltered frequencies are distinct), and demonstrates that mutagenesis in other regions is invariant to sequencing quality (where the filtered and unfiltered frequencies overlap).

Because of the large number of genomes in the gnomAD database, we considered that some loci with an allele count >1 could be produced by independent mutations in multiple individuals. Because the gnomAD database is de-identified, we could not check for saturation effects by downsampling to remove individuals. Instead, we proportionally downsampled the

database per each allele count bin, using the downsampling fraction multiplied by the allele count.

Repeat motif database generation:

Where indicated, repeats motif coordinates were gathered from Non-B DB[21]. However, because this database does not account for interrupted motifs, we generated our own database of repeat motifs. Starting from the hg38 human reference genome assembly, we first masked transposable elements based on Repeatmasker 'non-simple repeat' regions, both downloaded from the UCSC Genome Browser[22]. We searched autosomal chromosomes for all non-redundant STR motifs from 1-9 nt, using custom Python scripts employing 'regex' matching. Coordinates (consisting of chromosome, start position and end position) were then extended upstream and downstream to search for continuation of the motif, including partial motif extensions. For each motif, coordinates were then reduced to a minimal set of non-overlapping coordinates, allowing for up to one interrupted position on either side. The complete set of STRs was then used to mask the reference genome prior to all additional 'regex'-based searches.

Importantly, the initial motif search required a choice in minimum length requirements. We used a minimum of 5 units for mononucleotides, 3 units for dinucleotides, and two units for trinucleotides and above. Longer length requirements were used, where noted, following the incorporation of interruptions. We required that interruptions must link two segments that each meet the minimum length requirements. Thus, longer length requirements make the motif more easily recognizable., while shorter lengths maximize detection of interrupted motifs. However, choosing length requirements too short would undermine the classification of motifs. For an A-mononucleotide repeat, for example, it is not clear *a priori* whether 1-4 As

following an interruption is relevant to mutagenesis. Similarly, it is not clear whether interruptions of 2 or more nucleotides should be allowed for any given motif. Thus, while our method captures many of the possible motif configurations, capturing all is an impossibility, and some flanking regions may still be misassigned. Based on a 5 nt minimum length requirement, most of this uncertainty should exist within 5 nt surrounding the motif.

Pairs of motifs containing 5 nt of mirror or inverted symmetry separated by no more than 100 nt were found via ‘regex’ searches, and then coordinates for each pair were extended both outwards and inwards until symmetry was lost, or a masked region was reached on either side. 1 nt interruptions were then allowed on either side, and coordinates were again extended both outwards and inwards until symmetry was lost. For direct repeat motifs, the search proceeded left from the 5’ ends of both motifs, and right from the 3’ ends, again allowing for a single interruption on either side. Because the left/right extension necessarily shrinks the spacer distance, the initial search parameters allowed pairs to be separated by up to 1000 nt, and spacers were filtered down to 100 nt after the extension step. For inverted, mirror and direct motifs, the repeat length was filtered to a minimum of 10 nt after extension.

Potential Z-DNA-forming regions were identified based on two widely-used definitions. The first definition specifies that 10 nt or more must consist of alternating purines and pyrimidines, with the exclusion of any AT dinucleotides. This search was accomplished by replacing all purines and pyrimidines in the reference genome sequence with the symbols ‘R’ and ‘Y’ respectively, and using ‘regex’ to find coordinates matching ‘RYRYRYRY’ motifs. Next, the ‘ATGC’ reference sequence was retrieved for each set of coordinates, in order to filter out motifs containing ‘AT’ dinucleotides. Then, each entry was extended in the 5’ and 3’ directions until each entry no longer contained alternating purines and pyrimidines, contained an AT

245 dinucleotide, or reached a masked region. Finally, entries shorter than 10 nt were filtered out. The second definition specifies G followed by Y for at least 10 nucleotides. The search was performed as above with this additional restriction, as well as for the opposite strand (C followed by R). All entries from the second, more restrictive Z-DNA definition are contained within the first Z-DNA definition.

250 For G-quadruplex motifs, we began with a set of coordinates derived from experimental mapping of quadruplex-forming structures under two different quadruplex-stabilizing conditions[23]. PDS is a stronger stabilizer, detecting weaker G4 motifs, while K^+ is a weaker stabilizer, detecting stronger G4 motifs. Data is shown for K^+ conditions unless otherwise specified. The coordinates were lifted over from hg19 to hg38, reduced to a set of 255 minimal overlapping coordinates, then filtered for those containing at least 4 GGG-tracts (or CCC tracts, depending on the strand). Those coordinates were then trimmed to start and end with GGG-tracts, and filtered to remove any tracts longer than 1kb, as well as any tracts overlapping masked regions.

STR, inverted, mirror, direct, Z-DNA and G4 motif coordinates were combined into a 260 single database, in order to uncover overlaps between motifs. Because STRs were masked prior to other motif searches, STRs with multiple axes of symmetry exist only in the STR category. Inverted, mirror, direct, Z-DNA and G4 motifs fully or partially overlapping any other motif were removed, creating a database of unique repeat categories. We additionally set a buffer zone of 20 nt for repeat motifs and 5 nt for masked regions (Repeatmasker 265 transposons, centromeres, etc.) to avoid any near-overlaps. The buffer zone was initially set to 150 nt (data not shown), related to the typical Illumina read length, but was reduced after we observed no mutational effects extending beyond 20 nt, in order to exclude fewer repeats and gain additional power.

Flanking mutation frequency calculation and progressive filtering:

For each repeat motif, we consider the start and end position of the motif, as well as the distance from the start position to the nearest 5' motif or masked region, and the distance from the end position to the nearest 3' motif or masked region. Moving 1 nt to the 3' side of the motif, we filter the list of motifs for any where the flanking position overlaps another motif, with an additional buffer zone as described above. For any remaining motifs, the coordinates are used to generate a count of trinucleotides at these positions, as well as a count of any overlapping mutations in the gnomAD rare variant data set. The count of variants is split according to trinucleotide and mutation type (AAA>ACA, AAA>ATA, etc.). This process is repeated for positions 2 nt to the 3' side of the motif, and so on, until less than 1000 motifs remain in the list. This process is then repeated on the 5' side. To combine reverse-complementary STRs, or G-strand with C-strand G4 motifs, 5' and 3' coordinates were reversed (ie. position +1 for G strand and position -1 for C strand are combined).

For each position, the trinucleotide mutation counts are divided by the trinucleotide counts multiplied by the number of genomes in the gnomAD database to determine mutation frequency per trinucleotide per genome. This frequency is then normalized to the mutation frequency per trinucleotide per genome at all positions in the genome. By accounting for the expected genome-wide trinucleotide mutation frequency at every position, this eliminates potential artifacts from a non-random composition of nucleotides.

The trinucleotide-normalized values were then used to create a weighted average mutation frequency per position, with the weights being the trinucleotide mutation counts. To account for possible regional effects of variant QC filtering, we further normalized the weighted average mutation frequency to that derived from a set of ~100,000 random genomic coordinates (not containing any previously masked regions). Random non-repetitive regions

have some baseline mutation frequency under a given QC-filtering regime, and repeat motifs may induce (or protect from) additional mutations beyond this baseline.

For each position, 95% binomial proportion confidence intervals were calculated, with the number of successes being the total number of mutations across all trinucleotides, and the number of trials being the total number of trinucleotides multiplied by the number of genomes in gnomAD. The upper and lower frequencies were then adjusted based on the ratio of the trinucleotide-normalized and weighted mutation frequency to a weighted average of unnormalized trinucleotide frequencies, resulting in normalized weighted confidence intervals.

For mutations surrounding motifs, for clarity we examine perfect motifs, i.e. motifs with no detected interruptions. Additionally, we exclude the bottom 80% of motif lengths for each category. Because shorter motifs are exponentially more common, excluding the bottom 80% by length typically excludes the very shortest motifs in the database, which could dilute a true signal from longer repeats.

Mutation frequency calculation within motifs, and motif-perfecting mutations:

After generating the database of repeat motif coordinates, we searched within each motif for interruptions allowed by the algorithm. For STRs, we separated in-frame and out-of-frame interruptions. In-frame interruptions, which we examine here, can be corrected by mutations, while out-of-frame interruptions can be corrected by indels. For other motifs, we only searched for in-frame interruptions. We measured mutation frequency separately at interrupted sites, and generated separate counts for mutations that restore the perfect motif, and mutations which do not. Mutation frequencies were normalized and weighted as above, with the additional correction that one of three possible mutations can perfect the motif, while two of three cannot.

For inverted and mirror repeats, we separately counted mutations that convert the first and last position of the spacer into an extended symmetry, and the same for the first 5' and 3' flanking positions. For direct repeats, extending symmetry requires mutation of the position 5' of the left repeat to match the position 5' of the right repeat and *vice-versa*, or mutation of the position 3' of the left repeat to match the position 3' of the right repeat and *vice-versa* (labeled 'flank' and 'spacer', respectively, for their position relative to the left repeat).

For G4 motifs, we distinguished positions within G-runs (3 or more Gs in a row) and within spacers. We further distinguished 5', middle and 3' positions by their trinucleotide sequence: HGG, GGG, GGH, respectively, for G-runs, and GHH, NNN, HHG, respectively, for spacers, with a further designation for 1 nucleotide GHG spacers (H being A, C or T).

References

1 Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J Biol Chem*. 2020;295(13):4134-4170. doi:10.1074/jbc.REV119.007678

2 Du X, Gertz EM, Wojtowicz D, et al. Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic Acids Res*. 2014;42(20):12367-12379. doi:10.1093/nar/gku921

3 Zou X, Morganella S, Glodzik D, et al. Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res*. 2017;45(19):11213-11221. doi:10.1093/nar/gkx731

4 Georgakopoulos-Soares I, Morganella S, Jain N, et al. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res*. 2018;28(9):1264-1271. doi:10.1101/gr.231688.117

5 Guiblet WM, Cremona MA, Harris RS, et al. Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res*. 2021;49(3):1497-1516. doi:10.1093/nar/gkaa1269

6 Carlson J, Locke AE, Flickinger M, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun*. 2018;9:3753.

7 Chambers VS, Marsico G, Boutell JM, Di Antonio M, Smith GP, Balasubramanian S. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol*. 2015;33(8):877-881. doi:10.1038/nbt.3295

8 Guiblet WM, Cremona MA, Cechova M, et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res*. 2018;28(12):1767-1778. doi:10.1101/gr.241257.118

9 Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*. 2014;30(20):2843-2851. doi:10.1093/bioinformatics/btu356
10 Muyas F, Bosio M, Puig A, et al. Allele balance bias identifies systematic genotyping errors and false disease associations. *Hum Mutat*. 2019;40(1):115-126. doi:10.1002/humu.23674
11 Seplyarskiy VB, Soldatov RA, McGinty RJ, et al. Population sequencing data reveal a compendium of mutational processes in human germline. Preprint at https://www.biorxiv.org/content/10.1101/2020.01.10.893024v1 (2020)
12 Wang Q, Pierce-Hoffman E, Cummings BB, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat Commun*. 2020;11(1):2539. Published 2020 May 27. doi:10.1038/s41467-019-12438-5
13 Lovett ST. Template-switching during replication fork repair in bacteria. *DNA Repair (Amst)*. 2017;56:118-128. doi:10.1016/j.dnarep.2017.06.014
14 Löytynoja A, Goldman N. Short template switch events explain mutation clusters in the human genome. *Genome Res*. 2017;27(6):1039-1049. doi:10.1101/gr.214973.116
15 Wang G, Vasquez KM. Z-DNA, an active element in the genome. *Front Biosci*. 2007 May 1;12:4424-38. doi: 10.2741/2399.
16 Brázda, V., Laister, R.C., Jagelská, E.B. et al. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Molecular Biol* 12, 33 (2011). https://doi.org/10.1186/1471-2199-12-33
17 Guiblet WM, DeGiorgio M, Cheng X, et al. Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. *Genome Res*. 2021;31(7):1136-1149. doi:10.1101/gr.269589.120
18 Kockler ZW, Osia B, Lee R, Musmaker K, Malkova A. Repair of DNA Breaks by Break-Induced Replication. *Annu Rev Biochem*. 2021;90:165-191. doi:10.1146/annurev-biochem-081420-095551
19 Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7
20 Tiao G and Goodrich J. gnomAD v3.1 New Content, Methods, Annotations, and Data Availability. https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability/ (2020)
21 Cer RZ, Donohue DE, Mudunuri US, et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res*. 2013;41(Database issue):D94-D100. doi:10.1093/nar/gks955
22 Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006. doi:10.1101/gr.229102
23 Marsico G, Chambers VS, Sahakyan AB, et al. Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res*. 2019;47(8):3862-3874. doi:10.1093/nar/gkz179

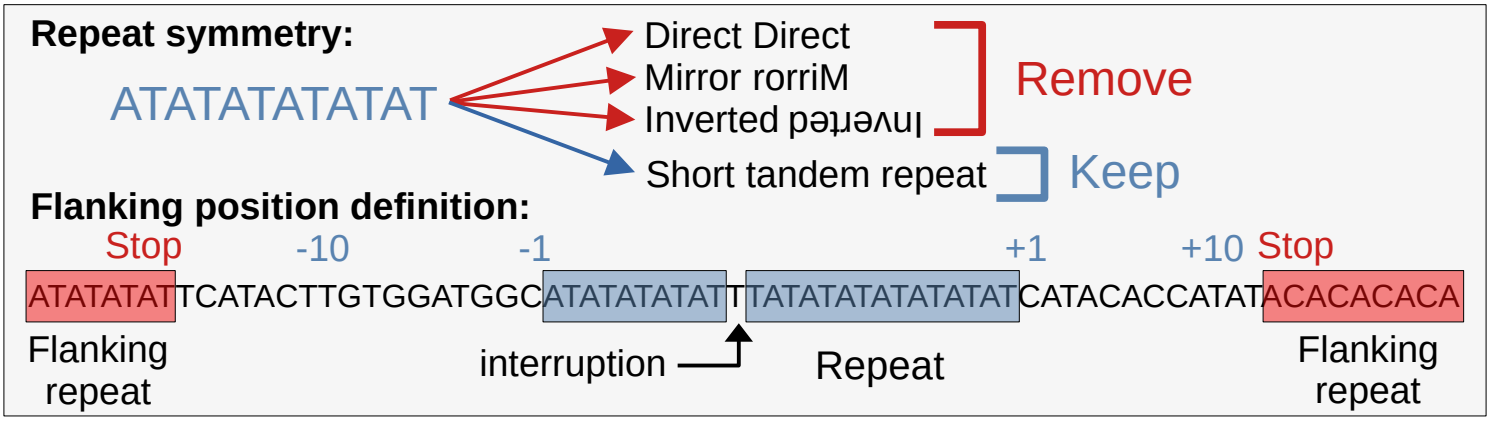

Fig. 1: Defining repeat motifs and their flanking positions.
Top: example repeat motif containing multiple symmetries, but classified only as an STR. Bottom: Relative flanking positions begin after any interruptions, and end when another motif is encountered.

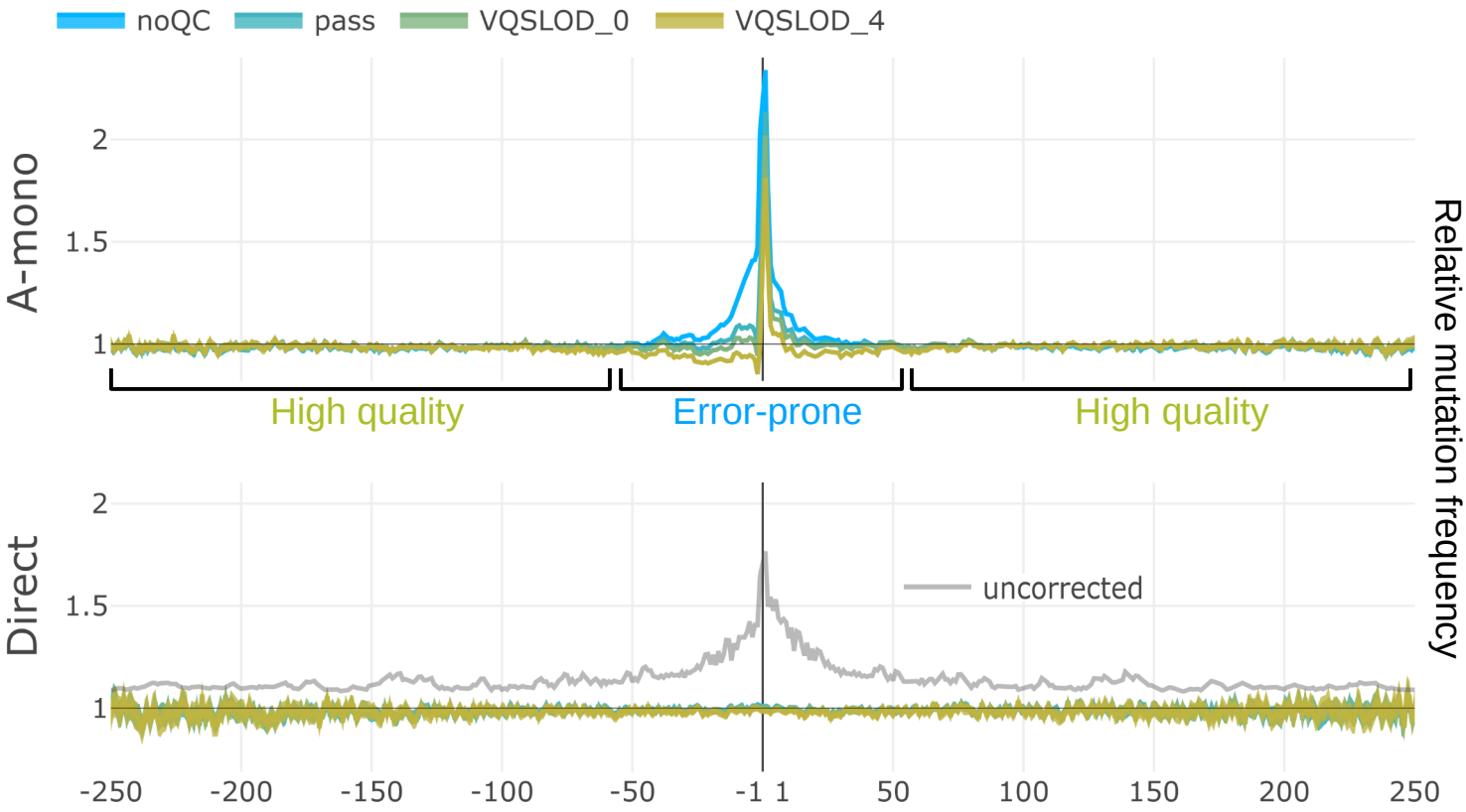

Fig. 2: Mutagenesis flanking non-B motifs.
X-axes are coordinates relative to central motif (mutation within motif not shown). Y-axes are relative mutation frequency (compared to the gnomAD average, normalized by trinucleotide mutation type). Blue lines: no sequencing quality filters, green and yellow lines: increasingly stringent filters. 95% binomial confidence intervals indicated in transparency. Top: A-mononucleotide STR motifs, demonstrating poor sequencing quality. Bottom: Direct repeat motifs, demonstrating comparison to analysis without corrective measures (STRs not excluded, repeats and transposons not removed from flanking positions, no trinucleotide normalization).

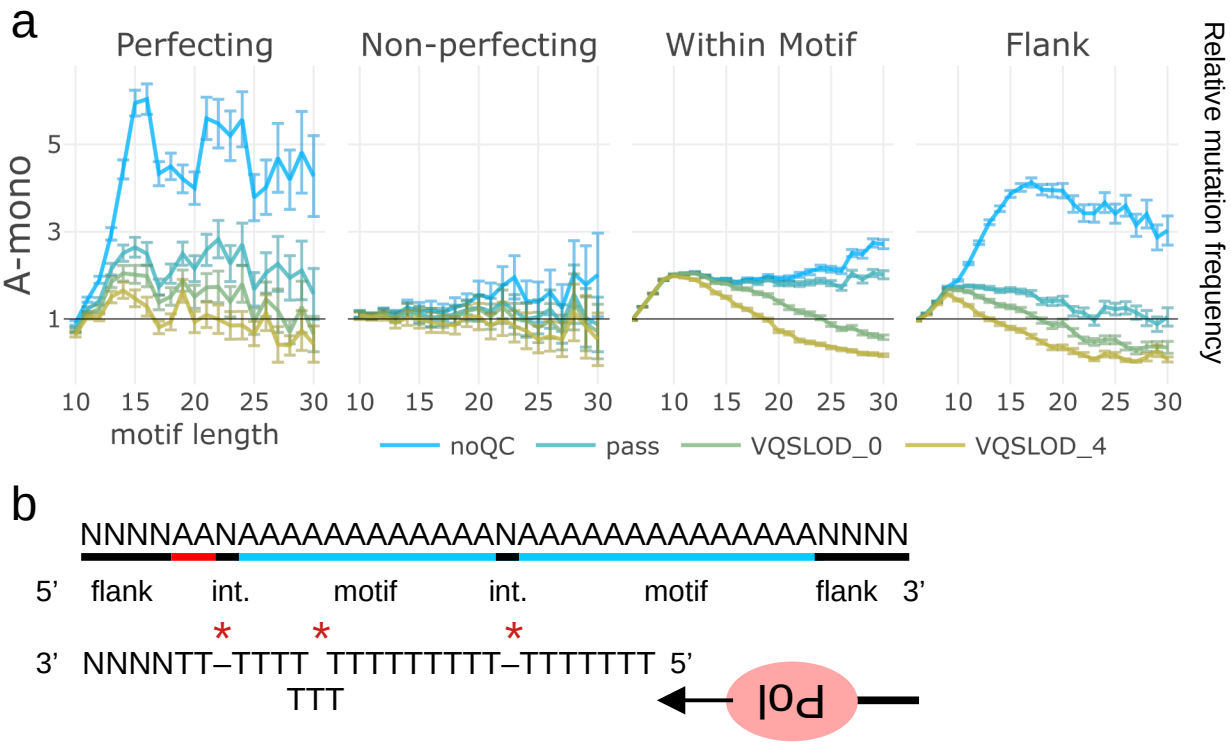

Fig. 3: Polymerase slippage generates mutations and sequencing errors at STRs.
a) Mutations that perfect or don't perfect interruptions within A-mononucleotide repeats. Also, mutations within the motif, and immediately flanking the motif. Y-axes are relative mutation frequency. Error bars indicate 95% binomial confidence intervals. X-axes are motif lengths. Blue: no sequencing quality filters, green and yellow: increasingly stringent filters. **b)** Polymerase slippage along STRs. Blue regions contain the motif, while the red region shows a sequence consistent with the motif, but not detected due to algorithmic constraints in recognizing interruptions. Polymerase slippage, especially during sequencing, skips over interruptions and inserts additional bases within the motif, resulting in the appearance of length-neutral mutations.

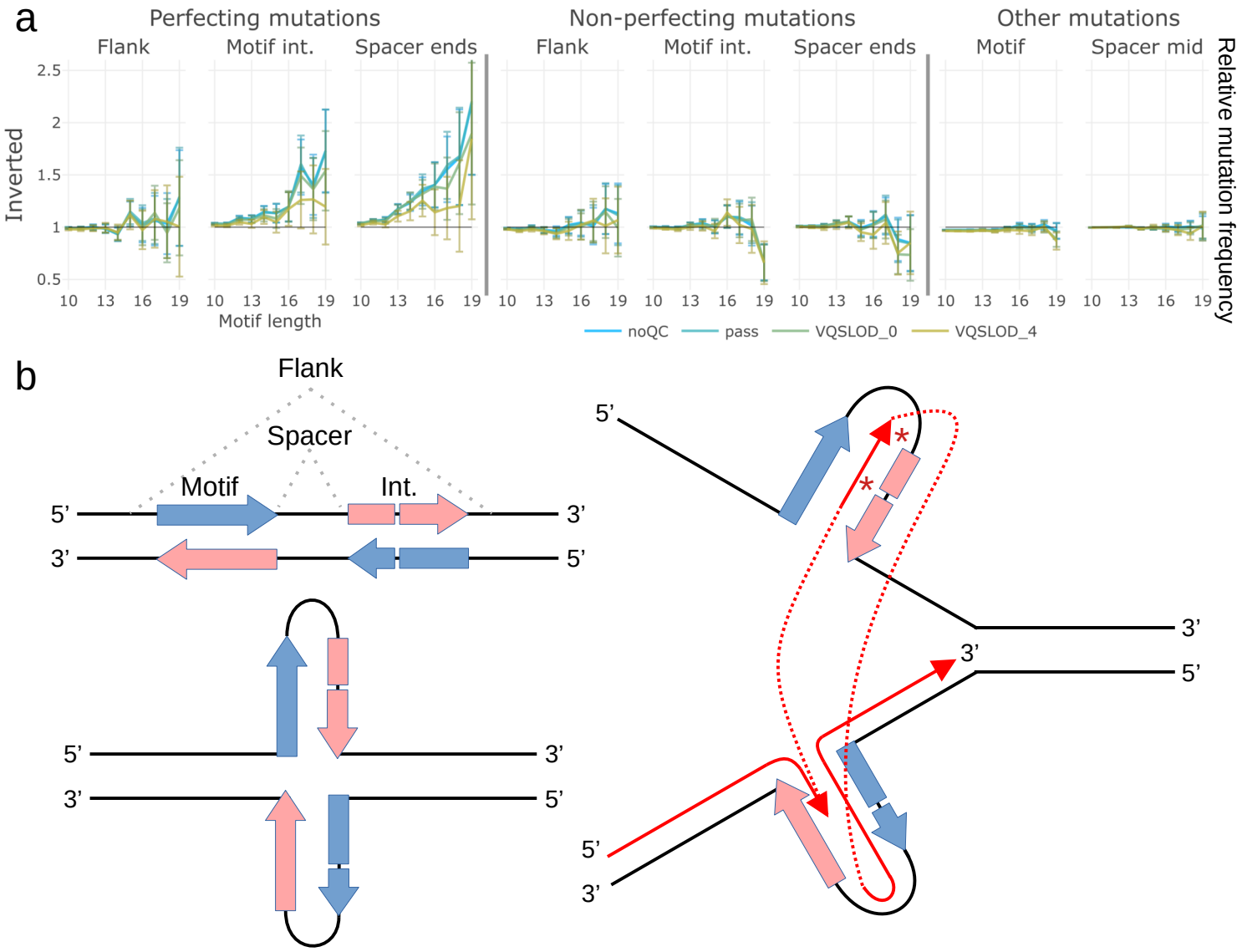

Fig. 4: Mutagenesis within inverted repeat motifs undergoing template-switch.
a) Left: Mutations that perfect/extend symmetry within inverted repeats, at the immediate flanking positions, at interruptions within the motif, and at the first and last spacer positions between the motifs. Middle: Mutations that don't perfect/extend symmetry. Right: Mutations within the motif that disrupt symmetry, and mutations in the middle of the spacer region. X-axes are motif lengths. Y-axes are relative mutation frequency. Error bars indicate 95% binomial confidence intervals. Blue: no sequencing quality filters, green and yellow: increasingly stringent filters. **b)** Top left: Diagram illustrating inverted repeats. Bottom left: Folding into non-B hairpin/cruciform structures. Right: Template-switch during replication produces motif-perfecting and inward-extending mutations. The red line follows the path of the polymerase as it proceeds 5' to 3'. Dotted lines indicate template-switches. * indicates positions of frequent perfecting/extending mutations.

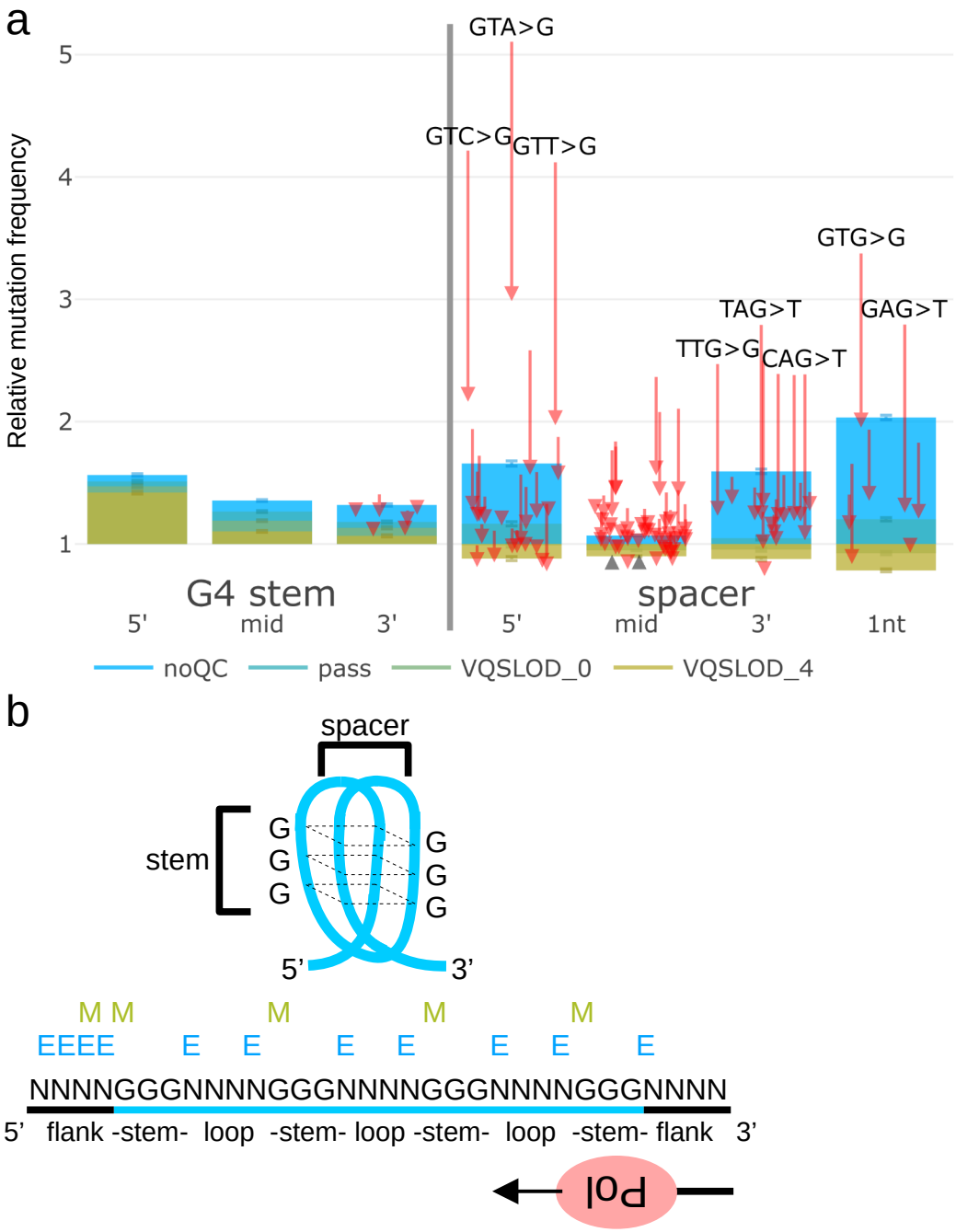

Fig. 5: G4 motifs are prone to recurrent sequencing errors.
a) Mutations within G4 motifs. X-axis indicates positions within G-runs, or spacers between G-runs. Spacers of 1 nt in length are separated. Y-axis is relative mutation frequency. Blue: no sequencing quality filters, green and yellow: increasingly stringent filters. Error bars indicate 95% binomial confidence intervals. Arrows indicate magnitude and direction of change for individual trinucleotide mutation frequencies before and after sequencing quality filters. Mutations with large magnitude changes are highlighted in text. **b)** Top: Diagram of a G-quadruplex structure. Bottom: Replication through a G4 motif produces a distinct pattern of mutations (M) and errors (E).