Received: 10 September 2021 | Revised: 2 November 2021 | Accepted: 14 December 2021

DOI: 10.1002/humu.24314

RESEARCH ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Distinct sequence features underlie microdeletions and gross deletions in the human genome

Mengling Qi[1] | Peter D. Stenson[2] | Edward V. Ball[2] | John A. Tainer[3] | Albino Bacolla[3] | Hildegard Kehrer-Sawatzki[4] | David N. Cooper[2] | Huiying Zhao[1]

[1]Department of Medical Research Center, Sun Yat-sen Memorial Hospital, Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangzhou, China

[2]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK

[3]Departments of Cancer Biology and of Molecular and Cellular Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA

[4]Institute of Human Genetics, University of Ulm, Ulm, Germany

Correspondence

Huiying Zhao, Department of Medical Research Center, Sun Yat-Sen Memorial Hospital, 107 Yan Jiang West Rd, Guangzhou, 500001, China.
Email: Zhaohy8@mail.sysu.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 81801132, 81971190; National Institutes of Health, Grant/Award Numbers: P01 CA092584, R35 CA220430

Abstract

Microdeletions and gross deletions are important causes (~20%) of human inherited disease and their genomic locations are strongly influenced by the local DNA sequence environment. This notwithstanding, no study has systematically examined their underlying generative mechanisms. Here, we obtained 42,098 pathogenic microdeletions and gross deletions from the Human Gene Mutation Database (HGMD) that together form a continuum of germline deletions ranging in size from 1 to 28,394,429 bp. We analyzed the DNA sequence within 1 kb of the breakpoint junctions and found that the frequencies of non-B DNA-forming repeats, GC-content, and the presence of seven of 78 specific sequence motifs in the vicinity of pathogenic deletions correlated with deletion length for deletions of length ≤30 bp. Further, we found that the presence of DR, GQ, and STR repeats is important for the formation of longer deletions (>30 bp) but not for the formation of shorter deletions (≤30 bp) while significantly (χ^2, $p < 2E{-}16$) more microhomologies were identified flanking short deletions than long deletions (length >30 bp). We provide evidence to support a functional distinction between microdeletions and gross deletions. Finally, we propose that a deletion length cut-off of 25–30 bp may serve as an objective means to functionally distinguish microdeletions from gross deletions.

KEYWORDS

DNA sequence motifs, DNA structure, GC content, gross deletions, HGMD, microdeletions, non-B DNA-forming repeats

1 | BACKGROUND

Deletions are responsible for many human genetic diseases and together constitute about 20% of all mutations known to cause human inherited disease (Stenson et al., 2020). Deletions are associated not only with common disorders, such as Alzheimer's disease (Cukier et al., 2016; Prihar et al., 1999), Parkinson's disease (Tan, 2016), intellectual disability (Sharp et al., 2006), autistic spectrum disorders (Sato et al., 2012; Vaags et al., 2012), and heritable cancers (Guo et al., 2018; Xu et al., 2012) but also rare or low-frequency diseases (Nambot et al., 2018). Disease-associated deletions in humans may range in size between 1 bp up to many thousands or even millions of base-pairs (bp). Historically, the

Human Gene Mutation Database (HGMD) has subdivided genomic deletions into microdeletions (1–20 bp) and gross deletions (>20 bp) (Stenson et al., 2020), but this distinction was originally made fairly arbitrarily for reasons of practical utility rather than for any cogent biological reason. Many studies (Carvalho & Lupski, 2016; Keute et al., 2020; Maranchie et al., 2004; Sahoo et al., 2006) have suggested the involvement of different mechanisms in the formation of microdeletions and gross deletions including nonhomologous end-joining (NHEJ), microhomology-mediated end-joining (MMEJ), non-allelic homologous recombination (NAHR), retrotransposon-mediated mechanisms, and replication-based errors including fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) (Abelleyro et al., 2020; Eckelmann et al., 2020; Bauters et al., 2008; Carvalho et al., 2009; Férec et al., 2006; Gadgil et al., 2020; Hastings, Ira, et al., 2009; Hastings, Lupski, et al., 2009; Hu et al., 2019; Lee et al., 2007; Marey et al., 2016; Summerer et al., 2018; J. Vogt et al., 2014; Zhang et al., 2009, 2010). Jahic et al. (2017) have presented doublet-mediated DNA rearrangements as a mechanism for the formation of recurrent pathogenic deletions of exon 10 in the *SPAST* gene. The operation of these different mutational mechanisms may be inferred by the presence of different breakpoint sequence features (Kidd et al., 2010).

Both gross deletions and microdeletions are non-randomly distributed in the human genome and are known to be strongly influenced by the local DNA sequence environment (Cooper et al., 2011; Del Mundo et al., 2017; Georgakopoulos-Soares et al., 2018). Previous studies have found that both gross deletions and microdeletions originate through the formation and subsequent resolution of aberrant DNA secondary structures, and we now know that the process of secondary structure formation is strongly sequence-mediated (Férec et al., 2006; Kouzine et al., 2017; Krawczak & Cooper, 1991; Wu et al., 2014). Previous studies have found that the breakpoints of deletions often possess a significant number of identical nucleotides, indicating the involvement of direct repeats (Kato et al., 2008), while replication slippage is recognized as a common cause of microdeletions (MacLean et al., 2006). Recent studies have revealed that replication-based mechanisms are frequently involved in gross duplications and deletions (Ankala et al., 2012; Carvalho & Lupski, 2016; Geng et al., 2021; Marey et al., 2016; Tsutakawa et al., 2017; Seo et al., 2020). Analyzing 8399 microdeletions in 940 genes from HGMD, one early study found that 81% of microdeletions (<21 bp) were located in the vicinity of direct, inverted, or mirror repeats (Ball et al., 2005). Another study attempted to relate the occurrence of microdeletions to the presence of non-B DNA structures by employing a set of 17,208 microdeletions (defined as being of length <21 bp), and found that 56% of microdeletions harbored either direct repeats or mirror repeats near the breakpoints (Kamat et al., 2016). An analysis of 11 gross deletions associated with autosomal dominant polycystic kidney disease, early-onset Parkinsonism, Menkes disease, α^+ thalassemia, adrenoleukodystrophy, and hydrocephalus, respectively, concluded that these large deletions were mediated by negative supercoiling-dependent non-B DNA conformations (Bacolla et al., 2004). Sequence motifs capable of forming non-B DNA structures contribute to the genome-wide instability responsible for both small- and large-scale copy number variants (Brown & Freudenreich, 2021; Guiblet et al., 2021). Arlt et al. (2009) reported that replication stress induces genome-wide copy number changes resembling pathogenic deletions and duplications. Most deletion breakpoint junctions were characterized by microhomologies suggesting that the deletion breakpoint junctions were formed by MMEJ, NHEJ or a replication-coupled process (Seo et al., 2020; Eckelmann et al., 2020; Dutta et al., 2017). Marey et al. (2016) illustrated the important role of NHEJ in the formation of *DMD* gene deletions.

Different forms of sequence capable of forming non-B DNA structures predispose certain genomic regions to instability causing pathogenic rearrangements (Zhao et al., 2010). The relationship between deletions and non-B DNA structures has been investigated in terms of the molecular properties of the deletion breakpoints (the breakpoints being defined as the junctions between the normal and rearranged DNA sequences) (Bacolla et al., 2006; Damas et al., 2014; Keegan et al., 2019). Verdin et al. (2013) identified various genomic architectural features, including sequence motifs, putative sites of non-B DNA conformations, and repetitive elements in breakpoint regions. Recurrent gross chromosomal rearrangements, including large deletions of several hundred kb are mediated by non-allelic homologous recombination (NAHR) (Demaerel et al., 2019; Dittwald et al., 2013; Harel & Lupski, 2018; Hillmer et al., 2016; Inoue & Lupski, 2002; Liu et al., 2012; P. H. Vogt et al., 2021). Finally, Abyzov et al. (2015) analyzed a total of 8943 non-pathogenic deletion breakpoints from 1092 healthy humans, revealing that NAHR-mediated breakpoints are associated with open chromatin. To our knowledge, however, no study has been performed that systematically explores the range of structural features associated with, and the mechanisms underlying, the full spectrum of human pathogenic gene deletions of different lengths, extending from the smallest of microdeletions to gross deletions. Such a study is needed to determine how microdeletions differ from gross deletions in terms of their underlying generative mechanisms, and whether there is a natural threshold or cut-off between these two entities or if they simply form the discrete ends of a continuum.

Besides a relationship between non-B DNA structure-forming motifs and deletion mutagenesis, several studies show that increasing GC content is associated with elevated rates of mutation and recombination (Kiktev et al., 2018; Romiguier et al., 2010). Deletion rates also vary between species in relation to genomic GC content (Hardison et al., 2003; Lindsay et al., 2019). A study of eutherian genomes found that increased GC content was associated with an increase in germline deletion frequency (Hardison et al., 2003). In similar vein, an analysis of 33 mammalian genomes found that GC-rich sequences were especially prone to deletion (Romiguier et al., 2010). These discoveries have indicated the importance of GC content in the formation of deletions in several different contexts. However, all these studies have either been inter-species comparisons or intra-genome comparisons in healthy humans and did not investigate pathogenic deletions. Importantly, to our knowledge, no study has yet investigated the relationship between GC content and deletion length in a disease context. Thus, here we formally investigate the relationship between GC content and pathogenic deletion length.

Various sequence motifs have been reported to be over-represented in the vicinity of microdeletion breakpoints (Ball et al., 2005). For example, purine-pyrimidine sequences and

polypurine tracts are significantly enriched in the vicinity of gross gene deletions (Abeysinghe et al., 2003). Recurrent large deletion of 1.11-Mb in 14q32.2 is catalyzed by large (TGG)*n* tandem repeats (Béna et al., 2010). One study reporting the breakpoint junctions of 30 rare deletions spanning between 91 bp and 14 kb found that most breakpoints exhibited microhomologies and were associated with specific sequence motifs (Vissers et al., 2009). Currently, we estimate that at least 78 sequence motifs have been found to occur at elevated frequencies in the vicinity of deletion, recombination, or translocation breakpoints (Abeysinghe et al., 2003; Ball et al., 2005; Chuzhanova et al., 2009). Ball et al. (2005) reported 30 motifs, including the heptanucleotide CCCCCTG, DNA polymerase pause sites, and topoisomerase cleavage sites that occurred frequently near deletion breakpoints. Chuzhanova et al. (2009) showed that DNA sequence motifs, known to be associated with site-specific cleavage/recombination, gene mutations, and various "super-hotspot motifs," were over-represented in the vicinity of microdeletions. However, to our knowledge, no attempt has as yet been made to analyze a large set of pathogenic deletions, including both microdeletions and gross deletions, to systematically explore the relationship between deletion length and occurrence frequency for the different types of sequence motif residing in the vicinity of breakpoints.

Here, we have performed an analysis of pathogenic gene deletions on two originally distinct microdeletion and gross deletion datasets from the HGMD (Stenson et al., 2020). Together, these comprise 42,098 breakpoints in a total of 3685 genes. We used simulated "deletions" matched by length and genomic position as controls. The purpose of this analysis was to assess the combined datasets in terms of the frequencies of six types of non-B DNA-forming repeat, GC content, the frequencies of specific sequence motifs, and microhomologies adjacent to the breakpoints. We propose several possible mechanisms for the formation of microdeletions and gross deletions. In addition, we compare generative mechanisms of microdeletions and gross deletions and suggest a new working definition with which to discriminate between microdeletions and gross deletions in terms of their size and underlying mechanisms of formation.

2 | MATERIALS AND METHODS

2.1 | Mutation and control datasets

In December 2019, the HGMD (Stenson et al., 2014, 2020) Professional release 2019.4 [http://www.hgmd.org] contained 38,725 microdeletions of ≤20 bp and 3373 gross (>20 bp) deletions, all characterized at single base-pair resolution, then constituting about 20% of all sequence-characterized mutations causing human inherited disease. These two deletion datasets were collected from the primary literature in precisely the same way; the 20 bp cut-off employed historically between microdeletions and gross deletions was entirely arbitrary and did not influence collation efficiency in any way. For the purposes of this study, these datasets were merged and were together termed the "HGMD-deletion data set." In total, 42,098 deletions were included in the HGMD-deletion data set. Of these deletions, 40,037 (95.1%) have a length ≤106 bp while 2061 (4.9%) deletions have a length between 107 and 28,394,429 bp. Figure S1 displays the log values of deletion numbers (length <107 bp) along deletion lengths. Table S1 includes the number of deletions with a specific length.

To assess the nonrandomness of the HGMD-deletion data set, we generated 100 simulated breakpoints for each deletion; these were randomly sampled within 3000 bp of the upstream region of each pathogenic deletion breakpoint. This process yielded 4,209,800 random breakpoints for the HGMD-deletion data set. Then, according to the coordinates of the 100 simulated breakpoints, we generated random deletions that matched each pathogenic deletion in terms of its length. By centering each simulated breakpoint around a 1 kb bin, we generated a sequence around the breakpoint and included it in the control0 data set. In total, the control0 data set includes 4,209,800 × 2 breakpoints and 4,209,800 × 2 flanking sequences. By randomly sampling 10 deletions for each pathogenic deletion from control 0, we generated the simulated data set, termed control1 that contained 420,980 deletions. If the simulated sequences contained undefined bases (*N*), these sequences were excluded from the analysis, and new random breakpoints and flanking sequences were generated by resampling. The coordinates of the simulated sequences were retrieved from a genome sequence file in version hg19 that was downloaded from https://www.gencodegenes.org/human/. Table S2 shows the coordinates of the control1 data set.

2.2 | Searching for non-B DNA-forming repeats in flanking sequences

Non-B DNA-forming repeats within each flanking sequence were obtained from the non-B DB database (Cer et al., 2011, 2013) with custom filters for mirror repeats (Table S3). As shown in Table S3, the mirror repeats were filtered by triplex-motif that is predicted by non-B DB as subset = 1. In this study, six types of non-B DNA-forming repeat were considered, specifically direct repeats (DR), inverted repeats (IR), mirror repeats (MR), G-quartets (GQ), short tandem repeats (STR, and Z-DNA (Z) (Ghosh & Bansal, 2003; Kondrashov & Rogozin, 2004; Wells, 2007). More detailed information on each type of non-B DNA-forming repeat is to be found in Table S3. The frequencies of the non-B DNA-forming repeats in the flanking sequences of the pathogenic deletions were compared with the frequencies of these repeats in the simulated data, the control1 data set. Statistical significance was assessed by means of the Student *t* test, and a Bonferroni correction was applied to allow for multiple testing.

For GQs, we divided all GQs into C-rich GQs and G-rich GQs. In DNA replication, the leading strand is elongated continuously in the direction of fork opening, whereas the lagging strand is made discontinuously in the opposite direction producing Okazaki fragments.

Okazaki fragments in eukaryotes are 150–250 bp in length. When a G-rich region is located at the end of a lagging strand in an Okazaki fragment, G-rich motifs will occur more often than C-rich motifs around deletion breakpoints.

2.3 | Specific sequence motifs in deletion flanking sequences

From previous publications (Abeysinghe et al., 2003; Ball et al., 2005; Chuzhanova et al., 2009), we collected a total of 78 sequence motifs (Table S4) that have been reported to occur in the vicinity of deletion/rearrangement breakpoints and are thought to play a role in the breakage and rejoining of DNA molecules. Briefly, Abeysinghe et al. (2003) listed 36 sequence motifs known to be associated with site-specific recombination, mutation, and DNA cleavage. In their later study, Ball et al. (2005) collected an additional 24 sequence motifs thought to be involved in site-specific recombination and putative deletion/insertion hotspots. Finally, Chuzhanova et al. (2009) reported 18 further motifs associated with deletions and recombination. We computed the frequency for each of the 78 motifs in the 1 kb-long sequences flanking the pathogenic deletions from the HGMD-deletion data set and in the control0 data set using the R package Biostrings (Pagès et al., 2020). We utilized the simulated deletions to determine whether the number of any type of motif in the vicinity of each breakpoint was higher than expected by computing an "experience hit" (eH-value), that is, the number of times the number of the motifs in the vicinity of the simulated breakpoints of the control data set was larger than the number of motifs in the vicinity of the pathogenic deletion breakpoints, divided by 100. The relationship between deletion length and motif frequency was then explored by calculating the average motif frequency for each deletion length.

2.4 | GC content

GC content was calculated for sequences in 1 kb bins centered at the breakpoints of the pathogenic deletions and simulated deletions using custom R codes. GC content was calculated for each deletion and each location from breakpoints, respectively. We explored the relationship between GC content and deletion length by considering average GC content centered around the deletion breakpoint for each deletion length.

2.5 | Mfold and SNP analyses

We installed mfold v.3.6 (http://www.unafold.org/mfold/software/download-mfold.php) on a Red Hat Enterprise Linux Server release 7.9 and launched the program in C on a parallel environment using the Message Passing Interface. We partitioned the human genome sequence version hg38 into 6,176,502 nonoverlapping 500-base bins and used mfold to compute the strongest ΔG value for the global folding into hairpin-loop structures for each of these 500-base sequences. Sequences with gaps and "N," were removed from further analyses. To determine the relationships between GC content and ΔG, we first computed the GC content of the 500-base sequences, ranked them by ΔG values, binned the ranked data into 100 bins, each containing ~54,000 sequences, and computed the average GC content for each of the 100 bins. The genomic coordinates of single nucleotide polymorphisms (SNPs) were taken from dbSNP build 151 (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/); these were used to assess the number of SNPs within ±500 bases flanking the junctions of 3373 gross deletions >20 bp and 1000 1 kb random sequences (control2). The sequences for control2 were obtained with bedtools and twoBitToFa (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) from an initial set of 1500 sequences from which the Y chromosome and records with gaps and "N" were discarded. In contrast to control1, control2 was intended to represent random sampling genome-wide. Statistical significance was assessed using the non-parametric Mann–Whitney rank-sum test.

2.6 | Microhomology analysis

To determine the extent to which microhomologies are associated with deletion variants, we used MHcut (Grajcarek et al., 2019) to search for homologous sequences at the junction sequences of deletion variants, thereby yielding a score with which to evaluate any microhomology present. For each deletion entry, microhomology was tested for both flanking configurations (5′ flanking region with 3′ variant sequence and 3′ flanking region with 5′ variant sequence), from which we selected the one with the highest score. The enrichment of microhomologies in the flanking sequence of deletions was assessed by means of the χ^2 test.

3 | RESULTS

3.1 | Non-B DNA-forming repeats and deletion breakpoints

A major goal of this study was to ascertain whether gene deletions causing human inherited disease occur disproportionately at sites that are capable of adopting non-B DNA structures, including hairpin and looped-out bases (direct repeats [DR] and short tandem repeats [STR]), cruciform (inverted repeats [IR]), mirror repeats (MR), G4 DNA (G-quartets [GQ]), and left-handed Z-DNA (Z-DNA [Z]). We show two typical examples of these repeats around both the longer deletions and shorter deletions in Figure S2. Using criteria defined in previous studies (Cer et al., 2011, 2013) and in Table S3, we searched for uninterrupted versions of each type of repeat within a 1 kb window centered at each deletion breakpoint. Then, we analyzed the length distribution of each type of repeat, and found that most of the identified repeat sequences were less than 50 bp in length (Figure 1).

As shown in Figure 1, more IR and STR were found in the deletion-flanking sequences than other types of repeats. The distribution (on a log scale) of the number of deletions against the deletion length is shown in Figure S3.

We compared the total numbers of repeats within 1 kb bins centered at the breakpoints for the HGMD-deletion data and the simulated deletion data set. All repeats occurred with a higher frequency in the vicinity of the gross deletions (length >20 bp) than in the control1 data set (Table 1). However, when we combined the gross deletions and microdeletions, we found that the numbers of repeats in the individual DR, IR, MR, STR, and Z DNA categories around the pathogenic deletion breakpoints were lower than those around the simulated data (Table 1 and Figure 2). Table S5 shows the detailed comparison of frequencies of different types of non-B DNA-forming repeat in the vicinity of breakpoints of deletions of different lengths. The frequencies of GQ around the pathogenic deletion breakpoints were higher than around the simulated data when the GQ was about 150 bp away from the deletion breakpoints (Figure 2d). However, when the GQ was close to the deletion breakpoints, the frequency of this repeat around the pathogenic deletion breakpoints was lower than around the simulated data (Figure 2d). We also partitioned the GQs around the breakpoints of deletions into G-rich GQs (15,931/32,067, 49.68%) and C-rich GQs (16,136/32,067, 50.32%), and compared their frequencies around pathogenic deletion breakpoints with the simulated data, control1. We found that the frequencies of C- and G-rich GQs around breakpoints of pathogenic deletions were rather similar and generally higher than around the simulated deletion breakpoints of control1 (Figure S4A and B). Then, we identified GQs in the human genome by means of the non-B DB. In total, 360,575 GQs were identified genome-wide. Among them, 11,450 were observed within 150 bp of the breakpoints of the deletions. A χ^2 test indicated that the deletion breakpoints were significantly enriched in GQs within 150 bp compared to genomic regions without deletions ($p < 2.2e{-}16$). When we used the χ^2 test to examine the enrichment of GQs between 150 and 500 bp to the deletion breakpoints, we also found GQs to be significantly enriched in this region ($p = 8.237e{-}06$). Thus, the positioning of deletion breakpoints is likely related to the presence of GQs.

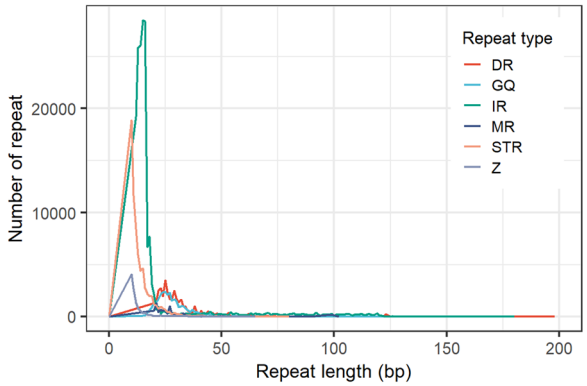

FIGURE 1 Repeat length distribution in all 1 kb bins centered at the breakpoints of the HGMD-deletion data. DR, direct repeats; GQ, G-quartets; IR, inverted repeats; MR, mirror repeats; STR, short tandem repeats; Z, Z-DNA

To ascertain whether we could identify a cut-off that would help to functionally distinguish gross deletions from microdeletions based on the occurrence of non-B DNA-forming motifs, we determined the average frequency of all types of non-B DNA-forming repeat in the 1 kb bins centered at the deletion breakpoints. As shown in Figure 3a, as the length of the pathogenic deletions increased, so too did the average frequency of non-B DNA-forming repeats around the deletion breakpoints. When the deletion length was ≤8 bp, the frequency of occurrence of non-B DNA-forming repeats in the vicinity of deletion breakpoints was lower than random expectation. Here, only 40,037 deletions shorter than 106 bp in length were analyzed because beyond this length the number of deletions of each length was less than 4 and the number of deletions was only 4.9% of the total. When we used a 10 bp sliding window to separate the deletions into bins and computed the average frequency of non-B DNA-forming repeats around the deletion breakpoints for the deletions in each bin,

TABLE 1 The density of non-B DNA-forming motifs in 1 kb sequences lefted at breakpoints as presented by average numbers of repeats per kb

Repeat type	Deletion (*n*/kb)	Microdeletion (*n*/kb)	Gross deletion (deletions >20 bp) (*n*/kb)	Control (*n*/kb)
ALL	83.22 (0–1095)	81.871 (0–1095)	98.713 (0–735)	89.176 (1.2–1148.6)
DR	12.441 (0–881.5)	12.086 (0–881.5)	16.518 (0–621)	13.797 (0–616.2)
IR	43.32 (0–394.5)	43.045 (0–394.5)	46.48 (0–313.5)	45.137 (0–575.85)
MR	2.352 (0–219)	2.252 (0–219)	3.504 (0–83)	2.935 (0–152.45)
GQ	11.118 (0–511)	11.075 (0–511)	11.618 (0–328)	9.355 (0–524.5)
STR	11.992 (0–257)	11.46 (0–257)	18.108 (0–238)	15.477 (0–396.9)
Z	1.996 (0–206)	1.954 (0–206)	2.485 (0–92.5)	2.476 (0–231.1)

Abbreviations: DR, direct repeats; GQ, G-quartets; IR, inverted repeats; MR, mirror repeats; STR, short tandem repeats; Z, Z-DNA.

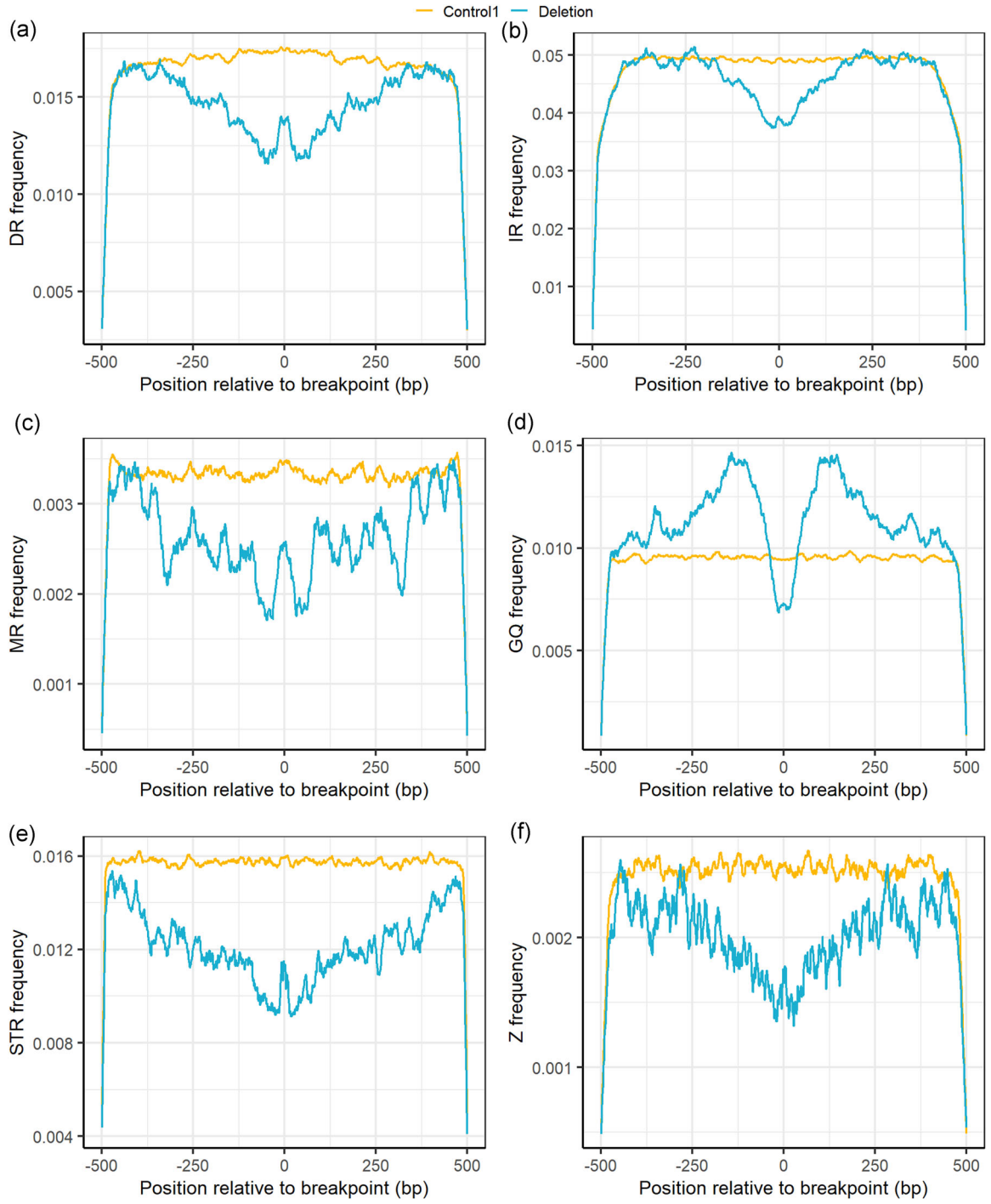

FIGURE 2 Frequency of non-B DNA forming repeats occurring near the breakpoints of the HGMD-deletion data set. The *x*-axis represents the position relative to the deletion breakpoint whilst the y-axis is the repeat frequency. (a–f) Depict the frequencies of direct repeats (DR), inverted repeats (IR), mirror repeats (MR), G-quadruplex-forming (GQ), short tandem repeats (STR), and Z DNA sequences, respectively. These frequencies refer to the proportion of sequences with repeats at each location

we found that deletion length was positively correlated with the frequency of non-B DNA-forming repeats although this was not statistically significant (Pearson Correlation Coefficient [PCC] = 0.33, $p = 0.32$) (Figure S5).

We then tested the correlation between deletion length and the frequency of non-B DNA-forming repeats. When the deletion length was ≤9 bp, the PCC of deletion length and average non-B DNA-forming repeat frequency was 0.79 ($p = 1.10E{-}2$). When the deletion length was less than ≤27 bp, the PCC attained its maximal value, 0.91 ($p = 3.39E{-}11$), whereas when the deletion length was less than ≤30 bp, the PCC was 0.80 ($p = 9.06E{-}8$) (Figure 3c). There was however no significant correlation between deletion length and

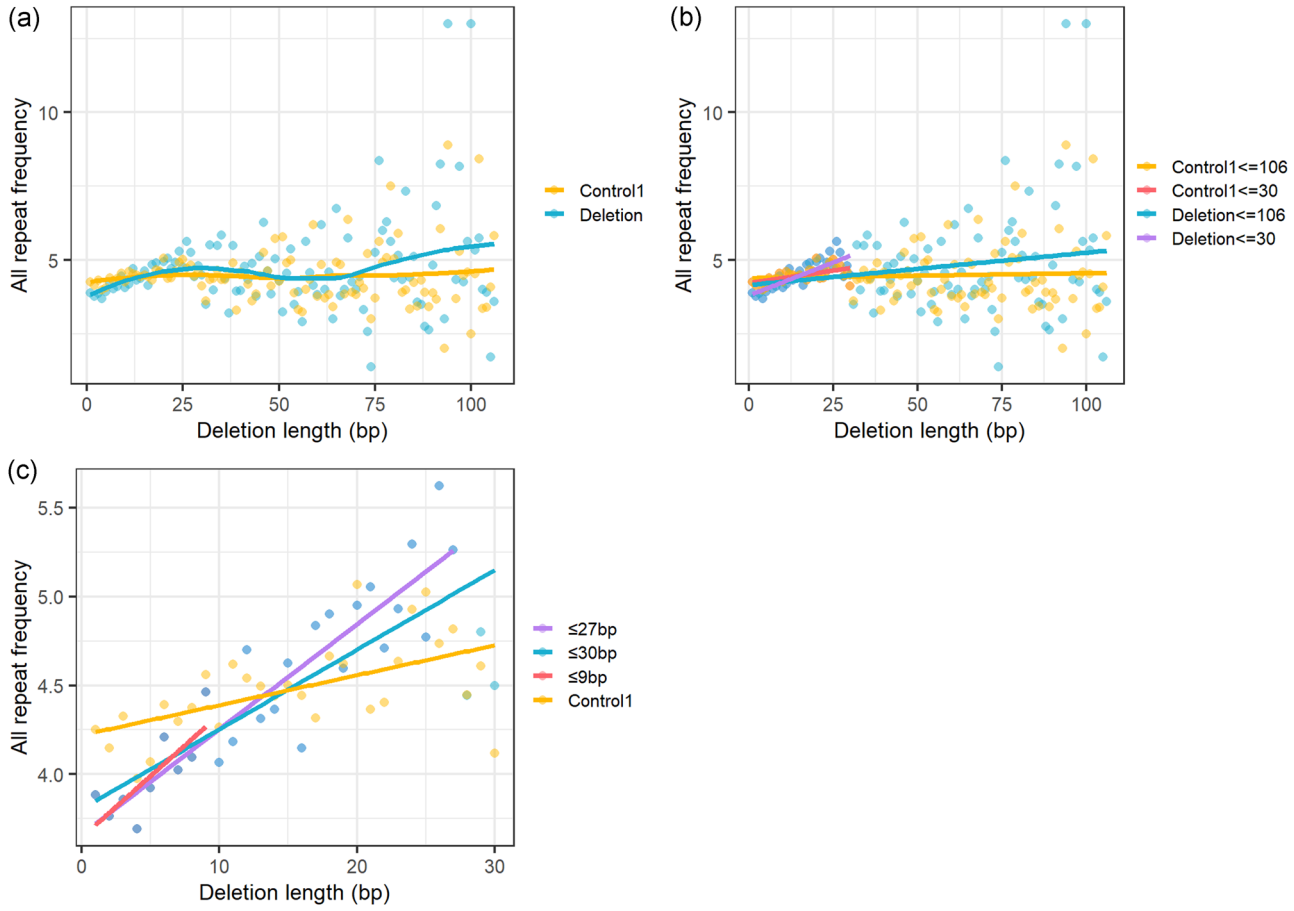

FIGURE 3 Relationship between deletion length and average non-B DNA-forming repeat frequency. (a) The relationship between deletion length and average repeat frequency within a 1 kb bin of breakpoints. (b) Correlations were observed between deletion length and the average repeat frequency for deletion lengths <31 bp, whereas no significant correlations were observed for control1 or deletion lengths >30 bp. (c) Significant correlations were observed between deletion length and repeat frequency in 1 kb sequence centered at breakpoints by different cut-offs for deletions of length ≤9, ≤27, and ≤30 bp, respectively

repeat frequency in control1 (Figure 3b). These findings indicate that the non-B DNA-forming repeat frequency in the vicinity of the breakpoints of deletions ≤27 bp in length was significantly and positively correlated with deletion length. When the deletion length was >30 bp, no significant correlation was observed between deletion length and the average non-B DNA-forming repeat frequency (Figure 3b). Thus, we speculate that 30 bp could represent a natural cut-off that serves to separate the pathogenic deletions into two relatively distinct (albeit overlapping) groups, with the larger deletions (with length >30 bp) having more complicated mechanisms of formation than the shorter deletions.

The relationship between the frequencies of the different types of non-B DNA-forming repeat and the deletion length is shown in Figure S6. For G-quadruplex-forming (GQ) sequences, a strong correlation (PCC = 0.87, p = 3.48E−10) was observed between deletion length and repeat frequency when the deletion length was ≤30 bp. For IR, DR, and STR, strong correlations (PCC = 0.72 and p = 1.3E−2, PCC = 0.76 and p = 5E−6, and PCC = 0.73 and p = 1.57E−5, respectively) were observed when the deletion length was ≤11, ≤27, and ≤27 bp, respectively. However, no strong correlation was observed between deletion length and the average frequencies of MR and Z-DNA-forming repeats. Taken together, for DR, GQ, and STR, the frequencies of these repeats were significantly correlated with deletion length when the deletions were ≤30 bp; for IR, the repeat frequencies were significantly correlated with deletion length when the deletions were ≤10 bp. These results suggest that a more precise cut-off to separate deletions mechanistically into microdeletions and gross deletions might lie between 10 and 30 bp.

To further investigate the non-B DNA-forming repeat frequency and distribution in the vicinity of breakpoints of deletions of different lengths, we used 30 bp as a cut-off to divide the pathogenic deletions in the HGMD-deletion data set into gross deletions and microdeletions and then analyzed the frequency of DR, GQ, and STR repeats in the vicinity of the breakpoints. We observed two frequency peaks of DR and STR repeats for deletions >30 bp and two frequency valleys for deletions ≤30 bp (Figure 4a,c).

However, no obvious frequency peak or valley was observed for GQ repeats flanking deletions >30 bp whereas a valley was found around the breakpoint location of deletions ≤30 bp (Figure 4b). When we divided the GQ repeats into G-rich and C-rich, we found that the

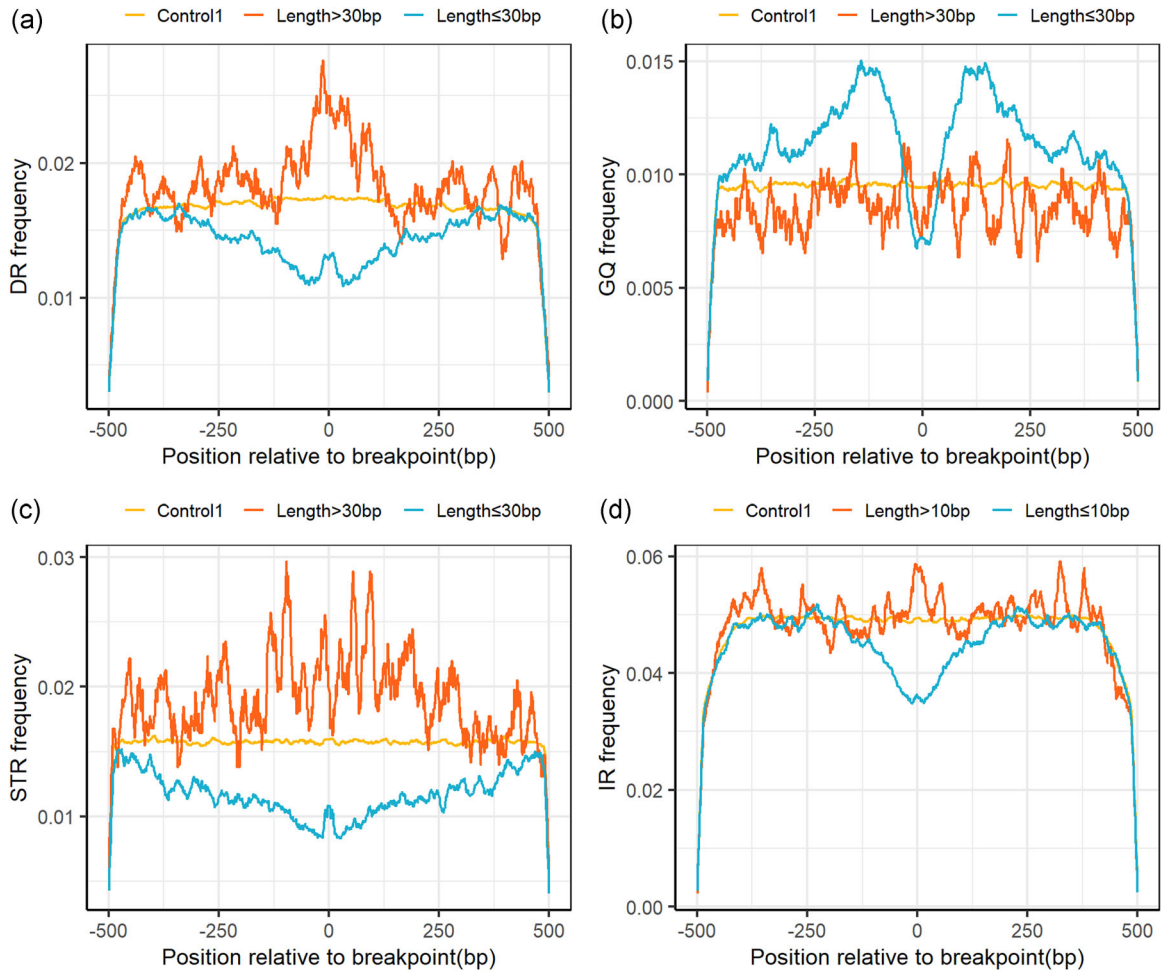

FIGURE 4 Repeat frequencies occurring near the breakpoints of deletions of different length. (a–d) are the average frequencies of direct repeats (DR), G-quadruplex-forming (QG), short tandem repeats (STR), and inverted repeats (IR), respectively

frequencies of G-rich GQ repeats and C-rich GQ repeats around breakpoints of short and long pathogenic deletions were close, and exhibited valleys around the breakpoints of deletions with length ≤30 bp (Figure S4C). The underlying reason for the absence of any obvious frequency peak of GQ repeats for deletions with length >30 bp appears to be due to the fact that G4 structures arising from GQ repeats may cause DNA polymerase pausing when associated with certain short motifs, which in turn promotes short deletions. Indeed, when we analyzed the probability of co-occurrence of GQ around deletions with short motifs found at DNA polymerase pause sites (Table S6), 91.1% of the GQs co-occurred together with such short motifs.

We also used 10 bp as a cut-off to divide the deletions into microdeletions and gross deletions and to analyze the frequency of IR in the vicinity of breakpoints. The frequencies of IR repeats showed a peak around the breakpoint of deletions with length >10 bp, and a valley at the breakpoint of deletions with length ≤10 bp (Figure 4d). These results suggest that the deletions separated by a cut-off into two groups had different properties in terms of the frequencies of non-B DNA-forming repeats in the vicinity of breakpoints. The patterns observed for the frequencies of non-B DNA-forming repeats in the vicinity of deletion breakpoints contrasted with the flat lines seen in controls (Figure 4), supporting the conclusion that either a 30 or a 10 bp cut-off can functionally distinguish microdeletions from gross deletions.

In summary, the frequency and distribution of non-B DNA forming repeats in the vicinity of pathogenic deletion breakpoints were clearly different when comparing deletions ≤30 and >30 bp (Figure 4). These differences may reflect heterogeneity in the underlying causative mechanisms responsible for both groups of deletion. For the breakpoints of deletions ≤30 bp, the number of non-B DNA-forming repeats increased in the breakpoint flanking regions in a "mirror image" fashion, suggesting that these breakpoints are either rarely located within non-B DNA forming sequences or that limited resection occurs before repair. Nevertheless, the increase in the frequency of these repeats at breakpoint flanking regions supports the view that non-B DNA structures induced nearby DNA breakage or polymerase stalling. Indeed, a comparable pattern of non-B DNA-forming sequences was not observed in either the control data set or in pathogenic deletions >30 bp. Rather, the most striking difference

between the ≤30 bp and >30 bp deletions was observed with respect to the distribution of direct repeats, which exhibited the highest frequency directly at breakpoints, suggesting replication slippage to be the initiating event for the genetic alteration.

3.2 | Non-B DNA-forming repeat motifs associated with deletions

Next, we wished to ascertain whether the short deletions and long deletions were associated with different types of repeat motif. Six types of non-B DNA-forming repeat, DR, GQ, IR, MR, Z-DNA, and STR, were investigated in this study. For each type of repeat, we obtained the top 10 most frequent sequences occurring in the vicinity of breakpoints of deletions with length >30 or ≤30 bp (Figure S7). Interestingly, most repeat motifs occurring in the vicinity of short deletions were different from the repeat motifs occurring in the vicinity of the long deletions (Figure S7). For DR (Figure S7A and B), all of the top 10 repeat motifs in deletions >30 bp were single nucleotide repeats whereas in deletions ≤30 bp, only one of the top 10 repeats in DR was a single base repeat. Meanwhile, for MR, we observed six single nucleotide repeat motifs (all motifs were nucleotide poly-A repeats) among the deletions >30 bp whereas only three single nucleotide repeats were found in the deletions ≤30 bp (Figure S7E and F). Thus, there may be a preference for single nucleotide repeats (poly A, poly T, poly C, or poly G) around deletion breakpoints ≥30 bp. From Figures S7I and J, we can see that seven of the top 10 repeat motifs occurring in STR are shared between the long deletions and the short deletions. We also noted that the sequence preference of Z-DNA repeats in long deletions is similar to the sequence preference associated with short deletions (Figure S7K and L). The underlying reason may be that for the STR and Z-DNA repeats, the cut-off in terms of partitioning the deletions into short and long groups does not lie around 30 bp (Figure S6). Frequencies of Z-DNA repeats were not found to correlate with the deletion length. When Z-DNA was divided into two groups according to deletion length, a frequency peak was observed at the breakpoints (Figure S8F) of long deletions (length >20 bp) but not at the breakpoints of short deletions (≤20 bp). Thus, if we use the frequency of Z-DNA to define gross deletions, 20 bp may be the appropriate cut-off.

3.3 | Relationship between GC content and deletion length

We next determined the GC content within the 1 kb bins centered at the breakpoints in the HGMD-deletion data set and the control1 data set. As shown in Figure 5a, the GC content was at its maximum at precisely 1 bp from the breakpoint and was invariably higher for pathogenic deletions than for the control1 data set (Student's t test $p < 2.2E{-}16$). The average GC content was then determined for deletions of different lengths. When the deletion length was ≤29 bp, the correlation between deletion length and GC content attained its highest value, with PCC = 0.87 ($p = 6.0E{-}10$) (Figure 5b). Indeed, GC content correlated positively (PCC = 0.71 and $p = 7.3E{-}7$) with deletion length up to a length of ≤38 bp. These results suggest that, in relation to GC content, 29–38 bp represents a potential cut-off that can serve to divide pathogenic deletions into gross deletions and microdeletions. When we used either 29 or 38 bp as a cut-off to partition the deletions into two groups, the GC content of the short deletions was higher than that of the longer deletions at the breakpoint (Figure S9). Thus, short and long deletions partitioned by the cut-off exhibit differences in GC content at the breakpoints.

The correlation between deletion length and GC content was intriguing because high GC content is expected to generate local folded-back hairpin-loop DNA structures with high thermodynamic stability. Therefore, we assessed the relationships between GC content and hairpin-loop structures by first dividing the human reference genome into ~6 million 500-base sequences and determining for each sequence the highest free-energy (ΔG) value with the potential to fold into complex hairpin-loop structures. Most genomic sequences displayed weakly stable hairpin-loops, peaking at a ΔG value of −37.5 kcal/mol (Figure 5c). Next, we determined the GC content for each sequence genome-wide and assessed the relationships between ΔG and GC content; these showed that, as predicted, GC content correlates with the stability of hairpin-loop structures, and that the bulk of the genome exhibits a GC content of ~0.35 (Figure 5d). Of note, the GC content of microdeletions was comparatively higher (0.45–0.55), implying that these mutations occurred within genomic regions prone to fold into metastable hairpin-look structures.

Analyses of copy-number variants in healthy individuals (Abyzov et al., 2015; Dhokarh & Abyzov, 2016; Hinds et al., 2006), of pathologic complex rearrangements (Carvalho et al., 2013; Wang et al., 2015) and of common fragile sites (Twayana et al., 2021), have shown that these tend to occur within genomic domains that are intrinsically unstable, such that they tend to co-occur with increased densities of SNPs and micro-mutations, pointing to error-prone DNA replication and repair. Therefore, given that our gross deletions occurred within a wide range of GC content (Figure 5b), we compared the number of SNPs within ±500 bp of their junctions with that of a control data set comprising 1000 random sites genome-wide (control2). Somewhat surprisingly, the number of SNPs flanking the gross deletions was lower than expected. SNP densities have been shown to vary within gene regions with the extent of sequence conservation (Castle, 2011); therefore, it is possible that, since our gross deletions are invariably pathogenic, they occurred within functional genomic regions that have been under selective constraint, and hence have incurred reduced variability in the human population. In summary, our data support the view that most deletions have occurred in regions of high GC content, and that the length of microdeletions correlates directly with GC content.

3.4 | Motif frequency and deletion length

The motif analysis was performed to determine the frequencies of a series of specific DNA sequence motifs around the breakpoints of the

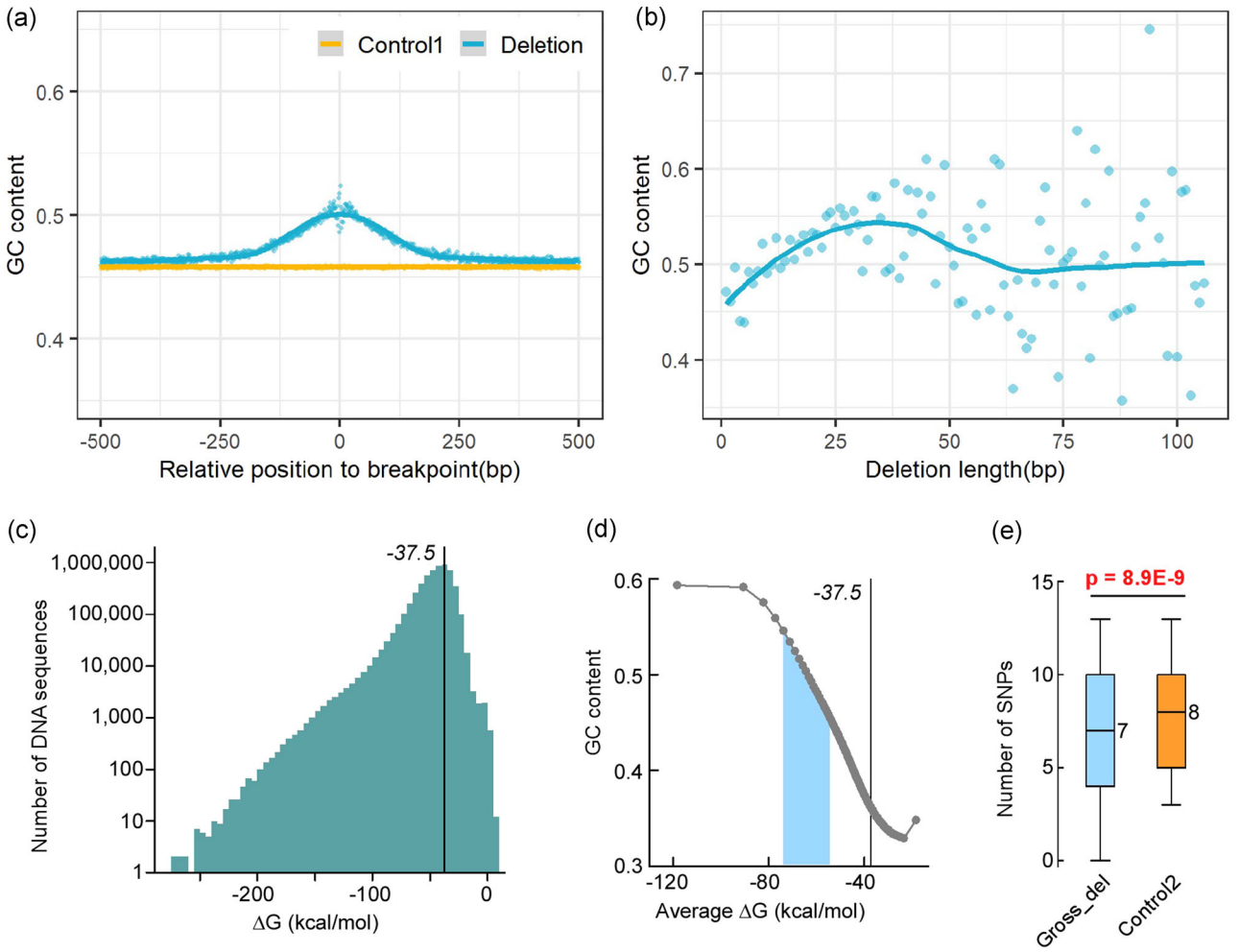

FIGURE 5 GC content in the vicinity of deletion breakpoints and the relationship between deletions and SNPs. (a) GC content in the vicinity of all the pathogenic deletion breakpoints and the simulated data. (b) Relationship between deletion length and GC content. When deletion length was less than 38 bp, it was significantly correlated with GC content (PCC = 0.71 and $p = 7.3E{-}7$). (c) Histogram of genome-wide distribution of ΔG values from mfold for folded-back harpin-loop structures. *Reference line*, ΔG value at the distribution peak. (d) Dot-plot of the relationship between ΔG and GC content genome-wide. *Shaded area*, ΔG values for GC content 0.45–0.55, corresponding approximately to deletions <30 bp. *Reference line*, as in panel (c). (e) Box plot of number of SNPs for gross deletions and control2. Medians and *p* values from Mann–Whitney rank-sum test are shown. Outliers were removed for the sake of clarity. SNP, single-nucleotide polymorphisms

pathogenic deletions. In total, 78 motifs (Table S4) were surveyed from previous publications (Abeysinghe et al., 2003; Ball et al., 2005; Chuzhanova et al., 2009). For each deletion from the HGMD data set, we calculated the motif frequency at each location in 1 kb bins centered at the breakpoints. Each deletion in the HGMD data set had 100 simulated deletions in the control0 data set, for which we also calculated the frequency of motifs. Considering all motifs together, we compared the motif frequencies in the vicinity of the breakpoints of the pathogenic deletions (HGMD-deletion data set) to the motif frequencies in the vicinity of breakpoints in deletions from the control0 data set. We found that the motif frequencies flanking the pathogenic breakpoints decreased gradually with distance from 150 bp to the breakpoint, and then attained their highest values precisely one base pair from the breakpoint itself (Figure S10), reflecting the likely contributions of these motifs to the formation of the deletions. By contrast, the motif frequencies in the vicinity of the deletion breakpoints from the control0 data set were remarkably similar irrespective of their distances from the breakpoints.

When we considered the frequencies of individual motifs in the vicinity of breakpoints, the distributions could be classified into four subtypes (Table S7), "Valleys," "Peaks," "M shapes," and "Others" (Figure S11-S16). In total, 22 motifs were grouped as "Valleys" (Figure S11 and S12) whose frequencies decreased with decreasing distance to the breakpoints and reached their lowest values at the breakpoints themselves. Among these motifs, motif25, motif26, motif44, motif45, motif49, and motif66 are recombination hotspots which are rarely seen around deletion breakpoints. Motif77 (RRRRRRRRRR) and motif78 (YYYYYYYYYY) are long polypurine/polypyrimidine tracts which have been noted to be over-represented at translocation breakpoints (Abeysinghe et al., 2003). A total of 28 motifs were grouped as "Peaks" (Figure S13 and S14), and their frequencies increased with decreasing distance to the breakpoints and reached their highest values

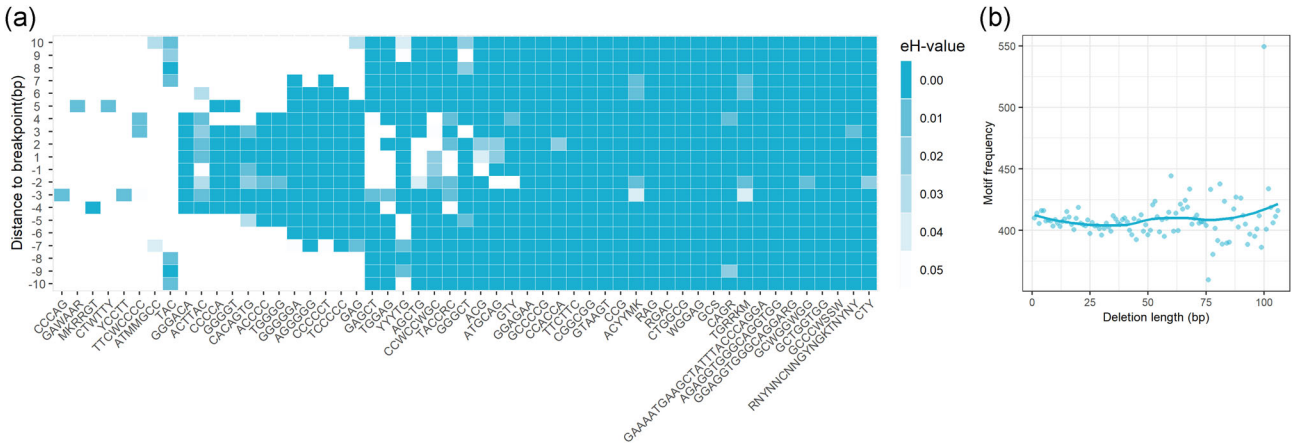

FIGURE 6 Sequence motifs around the breakpoints of deletions. (a) eH-values for the difference between frequencies of motif occurrence in 10 bp bins centered at the breakpoints of the deletion data set and the simulated data set; 16 motifs were found to occur more frequently (eH-value <0.01) in 10 bp bins centered at the breakpoints of the pathogenic deletion breakpoints than in 10 bp bins centered at the breakpoints of the control data set including simulated breakpoints. (b) Relationship between deletion length and average motif frequency; each point represents the average motif frequency occurring in the vicinity of deletions of a certain length

precisely at the breakpoints. Among the motifs in "peak" style, the motif2 (CTY), motif3 (GTY), motif48 (TAC), and motif51 (RAG) are vertebrate/plant topoisomerase I consensus cleavage sites, and motif6 (RNYNNCNNGYNGKTNYNY) is the vertebrate topoisomerase II consensus cleavage site, which have been reported to occur around deletion breakpoints (Abeysinghe et al., 2003; Ball et al., 2005). The motif29 (ACG) and motif30 (GCS) are DNA polymerase α pause site core sequences, whereas motif31 (WGGAG) is a DNA polymerase arrest site which may slow or stop the replication fork, increasing the time during which the single-stranded DNA is exposed in the replication fork, thereby increasing the chance of DNA breakage (Abeysinghe et al., 2003; Chuzhanova et al., 2009). The other motifs in the "peak" have been reported to occur in deletion hotspots (Abeysinghe et al., 2003; Ball et al., 2005; Chuzhanova et al., 2009).

In total, 14 motifs were grouped in an "M shape" (Figure S15) being characterized by frequencies that were distributed as an "M" shaped curve. Six out of seven immunoglobulins heavy chain class switch repeats exhibit an "M shape" in motif frequency distribution. Among them, motif11 (GGGCT), motif12 (GGGGT), motif13 (TGGGG), motif41 (ACCCC), and motif53 (CCCCA) exhibit high GC content, while motif14 (TGAGC) is characterized by low GC content. The motifs with high GC content occur at a relatively high frequency at a distance of 100–250 bp from the breakpoint, which may be related to the Okazaki fragments. As far as we are aware, this is a novel finding.

Finally, 11 motifs were grouped as "Others" (Figure S16) and were characterized by frequencies that were unrelated to distance from the breakpoints. Many of these "patterns" are exclusive to the pathogenic deletion data set and hence may indicate specific sequence differences between both datasets that are functionally relevant and predispose these regions to instability.

We counted the frequency of each motif in 10 bp bins centered at each breakpoint of the HGMD-deletion data set and the 100 simulated breakpoints. Then, we calculated the "experience hit" eH-values to assess the significance of each motif in the vicinity of the control breakpoints and the average eH-value of this motif over all the deletion breakpoints in the HGMD-deletion data set. The eH-value indicates the number of times the number of the motifs in the vicinity of the simulated breakpoints of the control data set was larger than the number of motifs in the vicinity of the pathogenic deletion breakpoints, divided by 100. We found that 23 motifs occurred more frequently (eH-value < 0.05) in 10 bp bins centered at the breakpoints of the pathogenic deletion data set than at the breakpoints from the simulated data set (Figure 6a). These motifs were "CTY," "RNYNNCNNGYNGKTNYNY," "GCCCWSSW," "GCTGGTGG," "GCWGGWGG," "GGAGGTGGGCAGGARG," "AGAGGTGGGCAGGTGG," "GAAAATGAAGCTATTTACCCAGGA," "TGRRKM," "CAGR," "GCS," "WGGAG," "CTGGCG," "RGAC," "RAG," "ACYYMK," "CCG," "GTAAGT," "CGGCGG," "TTCTTC," "CCACCA," "GCCCCG," "GGAGAA" (Table 2), which included four motifs identified by Ball et al. (2005). The one-sided Fisher's Exact test was used to examine if the motifs identified by Ball et al. overrepresented as motifs occurred more frequently in 10 bp bins centered at the breakpoints of the pathogenic deletion data set than at the breakpoints from the simulated data set. No significant result was identified with OR = 0.35 and $p = 0.055$. We calculated the average frequencies of all 78 motifs in 1 kb bins centered at the deletion breakpoints to explore the relationship between motif frequency and deletion length (Figure 6b) and identified six motifs for which the frequencies significantly correlated with deletion length (PCC > 0.7 and $p < 1E{-}6$) (Figure S17).

3.5 | Microhomology analysis for deletions and control1 data

To ascertain microhomologies, we used MHcut, which searches for homologous sequences within the flanking sequences of deletion variants. Of the 15,453 deletions with a minimum size of 3 bp, 40%

TABLE 2 Sequence motifs present more frequently (eH-value <0.05) in 10 bp bins centered at the breakpoints of the pathogenic deletion data set (HGMD-deletion) than in the breakpoints from the simulated data set

Motif sequence	Motif description	Average eH-value
GCCCWSSW	Translin target sites	0
GCTGGTGG	χ element	0
GGAGGTGGGCAGGARG	Human hypervariable minisatellite core sequence	0
AGAGGTGGGCAGGTGG	Human hypervariable minisatellite recombination sequence	0
GAAAATGAAGCTATTTACCCAGGA	Mariner transposon-like element (30end)	0
GCS	DNA polymerase α pause site core sequence	0
WGGAG	DNA polymerase arrest site	0
CTGGCG	DNA polymerase α frameshift hotspots	0
RGAC	Murine MHC deletion hotspot	0
RAG	Vertebrate/plant topoisomerase I consensus cleavage site	0
CCG	Fragile X breakpoint cluster repeat	0
GTAAGT	Indel hotspot	0
CGGCGG	Human Fra(X) breakpoint cluster	0
TTCTTC	Hamster and human APRT deletion hotspot	0
GCCCCG	"Super-hotspot" motifs	0
GGAGAA	"Super-hotspot" motifs	0
RNYNNCNNGYNGKTNYNY	Vertebrate topoisomerase II consensus cleavage site	5.00E−04
GCWGGWGG	Human minisatellite conserved sequence/χ-like element	5.00E−04
CTY	Vertebrate/plant topoisomerase I consensus cleavage sites	0.001
CCACCA	"Super-hotspot" motifs	0.001
CAGR	Murine MHC deletion hotspot	0.0015
TGRRKM	Deletion hotspot consensus sequence	0.0035
ACYYMK	Deletion hotspot consensus sequence	0.0035

(6195) were flanked by microhomologies of at least 3 bp, which is significantly higher than the corresponding probability (7.3% ± 0.2%) from control1 (*t* test $p < 2.2E{-}6$). For the remaining deletions, 59.4% of 1 bp deletions were found with at least 1 bp flanking microhomologies (control1: 28.2% ± 0.2%), and 71.3% of 2 bp deletions were detected with at least 2 bp flanking microhomologies (control1: 8.7% ± 0.1%), thereby implicating microhomologies as a common enriched characteristic feature of pathogenic deletion breakpoints. When we divided the pathogenic deletions in the HGMD data set into two groups using 30 bp as a cutoff, we found that 42% of sequences flanking deletions of length <30 bp have microhomologies while 29% of sequences flanking longer deletions have microhomologies. The χ^2 test indicated that the short deletions (length <30 bp) were enriched ($p < 2.2E{-}16$) with microhomologies as compared to the longer deletions. However, there was no significant correlation between the frequency of microhomologies and deletion length.

3.6 | Gross deletions and microdeletions are naturally partitioned

Our analysis indicates that the frequencies of non-B DNA-forming repeat, GC content, and specific sequence motifs all correlated with

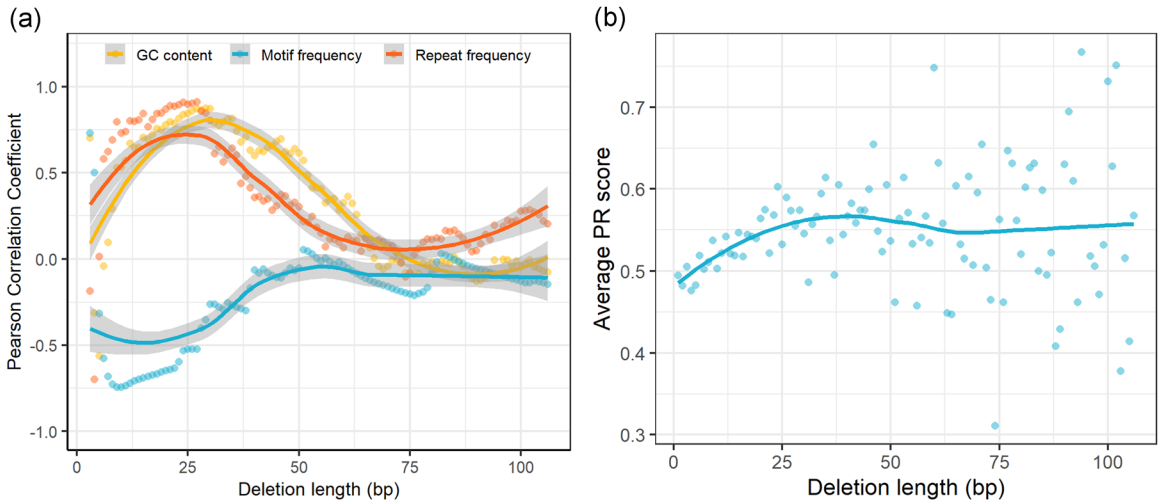

FIGURE 7 The Pearson Correlation Coefficient (PCC) and percentile ranking (PR) scores for motif frequency, GC content, or repeat frequency against deletion length. (a) Distribution of PCC against deletion length. The PCC values represent the correlations between deletion length and motif frequency, GC content, or repeat frequency. (b) Relationship between deletion length and PR score

the length of the deletions when deletion length was shorter than a given threshold (Figures 3a, 5b, and 6b). The PCC values plotted against deletion lengths are shown in Figure 7a. Here, PCC represents the extent of the correlation between deletion length and the frequencies of non-B DNA-forming repeats, GC content, and the frequencies of the sequence motifs being explored. As indicated in Figure 7a, when the deletion length was <25 bp, the PCC values pertaining to motif frequency and deletion length were negatively correlated. The PCC of the correlation between non-B DNA-forming repeat frequencies and deletion length attained its maximum value when the deletion length was 25 bp. The highest PCC value for the correlation between the deletion length and GC content was observed when the deletion length was 29 bp. Thus, we conclude that 25–30 bp may be a natural threshold to functionally distinguish gross deletions from microdeletions in terms of the underlying generative mechanisms.

3.7 | Can we score the deletions so as to separate the gross deletions and microdeletions naturally?

For each deletion, we calculated the non-B DNA-forming repeat frequency, GC content, and motif frequency in the region around it. Subsequently, we obtained the percentile ranking (PR) of the deletions in the HGMD-repeat database according to the cumulative non-B DNA-forming repeat frequency, GC content, and motif frequency. Then, each deletion was scored by summing the PR of the deletion in terms of the frequency of non-B DNA-forming repeats, GC content, and motif frequencies in the HGMD-deletion database. This score was termed the PR score. We then investigated the correlation between the PR scores of deletions and the deletion lengths. As shown in Figure 7b, when the deletion length was less than 46 bp, the average PR score for deletions of each length was significantly (PCC = 0.71 and $p = 4.1E{-}8$) correlated with deletion length. When the deletion length was >46 bp, no significant correlation was observed between the average PR score for deletions of each length and the deletion length. When we investigated the relationship between PR scores and deletion length with respect to repeat frequencies, GC content, and motif frequencies, respectively, we found that the deletion length (<31 bp) was significantly ($p = 8.8E{-}9$) correlated with the PR scores of non-B DNA-forming repeat frequency, and the deletion length (<47 bp) was significantly ($p = 5.0E{-}8$) correlated with the PR scores of GC content (Figure S18). These findings suggest that a deletion length of around 30–47 bp could serve as a possible natural cutoff to partition microdeletions and gross deletions on the basis of their PR scores calculated from the non-B DNA-forming repeat frequency, GC content, and motif frequency.

4 | DISCUSSION

Irrespective of whether we consider microdeletions or gross deletions, the mechanisms underlying pathogenic deletions appear to be strongly influenced by the local DNA sequence environment (Kondrashov & Rogozin, 2004; Krawczak & Cooper, 1991). The role of non-B DNA structures in the formation of cancer-associated deletions, as well as deletions in the germline and in mitochondrial sequences, has been appreciated for some time (Bacolla et al., 2016, 2019; Damas et al., 2014; Dong et al., 2014; Fontana & Gahlon, 2020; Pabis, 2021; Svetec Miklenić & Svetec, 2021; Zhao et al., 2010). Such non-B DNA structures often have key regulatory functions in DNA replication and transcription but may also cause genomic instability (Lemmens et al., 2015; Zhao et al., 2010). Furthermore, many deletions in the human genome are mediated by retrotransposon repeat-dependent mechanisms (Fujimoto et al., 2021; Mendez-Dorantes et al., 2020; Morales et al., 2021;

Vocke et al., 2021). Similarly, many studies have indicated a role for GC content and DNA motif sequences in the formation of microdeletions and gross deletions (Cooper et al., 2010; Visser, Shimokawa et al., 2005). However, the role of these sequence features in the formation of deletions of different lengths has not yet been methodically examined by robust statistical analyses. Meanwhile, the somewhat arbitrary definitions traditionally employed to distinguish between microdeletions and gross deletions have become blurred. We, therefore, collected 42,098 pathogenic deletions that display a length continuum stretching from 1 to 28,394,429 bp, from which we used 40,037 deletions with length <107 bp to perform a comprehensive analysis of the relationship between deletion length and non-B DNA-forming sequences, GC content, specific sequence motifs, the thermodynamic stability of fold-back hairpin-loops and microhomologies.

To our knowledge, this is the first study to demonstrate that very short deletions (≤8 bp) have a low probability of co-occurrence with non-B DNA-forming repeats. However, when the deletion length is >8 bp but ≤30 bp, the non-B DNA-forming repeat frequency neighboring deletion breakpoints is significantly and positively correlated with deletion length (Figure 3). By contrast, we observed no significant correlation between deletion length and repeat frequencies for deletions >30 bp, a finding that illustrates the complexity of the mechanisms of formation associated with long deletions versus short deletions.

In this study, we observed two frequency peaks of DR and STR repeats for deletions >30 bp and two frequency valleys for deletions <31 bp (Figure 4a,c). Replication slippage has long been recognized as a common cause of deletions (MacLean et al., 2006). When DNA is being processed during replication, direct repeats and short tandem repeats may produce slipped structures. During this process, the primer and template strands can transiently dissociate and then reassociate in misaligned configurations. If the primer strand containing the newly synthesized first direct repeat dissociates from the template strand and misaligns (slipping forward) at the second direct repeat, continued DNA synthesis will lead to the deletion of one of the direct repeats as well as the intervening sequence (Ball et al., 2005). When the slipped structure is formed, the DNA loops that do not form double strands are relatively easy to break and hence tend to be lost. The longer repeats will give rise to longer single stranded DNA loops with a greater chance of deletion. Moreover, the longer the repeat, the greater the chance of forming a misaligned configuration leading to deletion.

This study confirmed and extended previous observations that deletions of all sizes tend to be concentrated in GC-rich regions of the genome. Indeed, high GC content has been associated with a high level of mutation in general, not just deletions (Abeysinghe et al., 2003; Albano et al., 2010; Kiktev et al., 2018; Twayana et al., 2021; Zheng et al., 2013). Furthermore, we found that when deletion length was less than 38 bp, deletion length and GC content were positively correlated, with the correlation attaining its highest value (PCC = 0.87, p = 6.0E−10) when the deletion length was ≤29 bp. A previous study found that increased GC content contributes to the stabilization of non-B DNA structures, thereby enhancing the propensity of deletions to occur (Tanay & Siggia, 2008). Along the same line, we observed a direct relationship between GC content and the stability of fold-back hairpin-loop structures. Thus, given the direct correlation of GC content, stability of fold-back hairpin-loop structures, and co-occurrence of specific sequence motifs with short (<29 bp) deletions, we propose that the underlying mutational mechanism may involve hairpin-loop bypass through template switching during DNA replication (Northam et al., 2014). The increased frequency of GQ-forming repeats in regions flanking the breakpoints of deletions ≤30 bp may also have contributed, perhaps by transiently pausing DNA synthesis.

Variable GC content has been associated with both increased and decreased mutation rates depending upon local sequence context (Carlson et al., 2018) and evolutionary conservation (Castle, 2011). Our finding that pathological gross deletions occurred within regions of reduced SNP density is consistent with both the functional impact of, and the evolutionary constraints acting upon, the loci involved.

Previous studies have reported the involvement of a number of different sequence motifs in the DNA breakage events leading to microdeletions and microinsertions (Ball et al., 2005). Several studies have been performed on the sequence motifs in the vicinity of large genomic rearrangement breakpoints including large deletions (Abeysinghe et al., 2003; Dittwald et al., 2013; Férec et al., 2006; Hillmer et al., 2017; Jahic et al., 2017; Visser, Shimokawa, et al., 2005; J. Vogt et al., 2014). Here we collected a large number of inherited pathogenic deletions, representing a continuum of lengths from 1 bp to 28,394,429 bp, and determined the frequency of occurrence of 78 sequence motifs known to be overrepresented or underrepresented in the vicinity of breakpoints or sites of gene conversion in the human genome (Abeysinghe et al., 2003; Ball et al., 2005; Chuzhanova et al., 2009). We found that the sequence motif frequency was significantly and negatively (PCC = −0.62, p = 3.2E−2) correlated with deletion length when deletions were ≤12 bp. However, the relationship between motif frequency and deletion length may well be dependent upon the type of motif in question. As shown in Figures S11–S16, the motif frequencies are distributed quite differently in the vicinity of the deletion breakpoints; thus, further studies are required to identify the underlying reasons responsible for the relationship between deletions and the frequencies of specific motifs.

Here we observed that non-B DNA-forming sequences such as DR, IR, and STR were more abundant at the breakpoints and in breakpoint-flanking regions of deletions >30 bp than of deletions ≤30 bp (Figure 4). These repeats may form non-B DNA structures that cause replication stalling followed by replication fork repriming downstream, thereby leading to the deletions, a mechanism described as Fork Stalling and Template Switching (FoSTeS) (Lee et al., 2007). Replication errors mediated by these repeats may cause deletions >30 bp more frequently than deletions ≤30 bp in length. In particular, direct repeats and simple tandem repeats were overrepresented immediately juxtaposed to the breakpoints of deletions >30 bp (Figure 4a,c), indicative of a specific role for these repeats in

deletion formation. Both types of motifs may form slipped structures if they are base-paired with the complementary strand in a misaligned fashion, causing hairpins or looped-out bases which may then stall DNA replication (Zhao et al., 2010). Large deletions were also enriched in short A-tract-containing motifs (Figure S17), known to contain flexible and genetically unstable hinges (Bacolla et al., 2015), thereby supporting the view that, as with small deletions, noncanonical DNA conformations have contributed to the process of gross deletion mutagenesis. We suggest that the repair of stalled replication forks at these noncanonical DNA conformations may have involved more complex mechanisms than the replication bypass we propose for small deletions.

Microhomology-mediated end joining (MMEJ) plays an important role in double-strand repair and causes pathogenic deletion and translocation variants in the human genome (McVey & Lee, 2008; Verdin et al., 2013). MMEJ repairs DNA breaks via the use of substantial microhomology and creates precise deletions without introducing insertions or other mutations at the breakpoints. We identified microhomologies within the breakpoint flanking regions of 60% of the HGMD deletions indicating that MMEJ is an important mechanism underlying pathogenic deletions in humans. This is in accord with the findings of Grajcarek et al. (2019) who identified microhomologies at the breakpoints of 57% of the deletions included in ClinVar. Additionally, we found that more than 42% of the breakpoint flanking regions of short deletions (<30 bp) have microhomologies, somewhat higher than for the breakpoint flanking regions within long deletions (29%). To our knowledge, this is the first investigation to compare the occurrence of microhomologies in short and long deletions.

It is well known that replication-based mechanisms are often involved in the formation of deletions and duplications of various sizes (Ankala et al., 2012; Geng et al., 2021; Hambarde et al., 2021; Tsutakawa et al., 2021; Seo et al., 2020; Vissers et al., 2009; Zhao et al., 2010). Our findings suggest that these mechanisms contribute to both the formation of pathogenic microdeletions <30 bp and gross deletions ≥30 bp. However, the different frequencies and distribution profiles of non-B DNA-forming sequence motifs at the breakpoints and within breakpoint-flanking regions of both groups of deletions suggest that the replication errors underlying these deletions are induced by different types, and perhaps different sizes, of non-B DNA structure.

Overall, this study suggests 25–30 bp as a threshold that can be used to distinguish gross deletions and microdeletions in terms of their likely underlying mechanisms of mutagenesis. This notional threshold is based on the observation of the correlations between deletion length, non-B DNA-forming repeats frequencies, GC content, and sequence motif frequencies (Figure 7a). For deletion lengths greater than 30 bp, correlations start to weaken, and they tend to disappear at lengths greater than 50 bp. Although establishing a threshold to distinguish gross deletions from microdeletions is to some extent dependent on the intended research purpose, there is value in being able to draw distinctions based upon objective analyses. The approach and results reported here provide a path forward that should allow us to move away from arbitrary dividing lines and arrive at information-based knowledge concerning the rather different generative mechanisms underlying microdeletions and gross deletions.

ACKNOWLEDGMENTS

This study was supported in part through computational resources provided by Bioinformatics and Omics Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, and through HPC resources from the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (URL: http://www.tacc.utexas.edu). This study was funded by grants from the National Key R&D Program of China (2020YFB0204803), the Natural Science Foundation of China (81801132, 81971190, 61772566), Guangdong Key Field R&D Plan (2019B020228001 and 2018B010109006), Introducing Innovative and Entrepreneurial Teams (2016ZT06D211), Guangzhou S&T Research Plan (202007030010) to H.Z. and by National Institutes of Health (NIH) grants (P01 CA092584 and R35 CA220430) to J.A.T. P.D.S., E.V.B and D.N.C. acknowledge the financial support of Qiagen Inc through a License Agreement with Cardiff University.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

A pipeline is available at https://github.com/Qimengling/deletion_score_pipe for anyone with novel deletions. This pipeline enables users to calculate the frequency of non-B-forming DNA repeats, GC content, and specific motif frequency, and to obtain a deletion score according to the percentile ranking in the HGMD-deletion database.

ORCID

Mengling Qi https://orcid.org/0000-0003-4956-2031
Albino Bacolla https://orcid.org/0000-0003-0206-8423
Huiying Zhao https://orcid.org/0000-0001-9134-536X

REFERENCES

Abelleyro, M. M., Radic, C. P., Marchione, V. D., Waisman, K., Tetzlaff, T., Neme, D., & De Brasi, C. D. (2020). Molecular insights into the mechanism of nonrecurrent F8 structural variants: Full breakpoint characterization and bioinformatics of DNA elements implicated in the upmost severe phenotype in hemophilia A. *Human Mutation*, *41*(4), 825–836. https://doi.org/10.1002/humu.23977

Abeysinghe, S. S., Chuzhanova, N., Krawczak, M., Ball, E. V., & Cooper, D. N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Human Mutation*, *22*(3), 229–244. https://doi.org/10.1002/humu.10254

Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stutz, A. M., Parrish, N. F., & Gerstein, M. B. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications*, *6*, 7256. https://doi.org/10.1038/ncomms8256

Albano, F., Anelli, L., Zagaria, A., Coccaro, N., Casieri, P., Rossi, A. R., & Specchia, G. (2010). Non-random distribution of genomic features in breakpoint regions involved in chronic myeloid leukemia cases with variant t(9;22) or additional chromosomal rearrangements. *Molecular Cancer*, *9*, 120. https://doi.org/10.1186/1476-4598-9-120

Ankala, A., Kohn, J. N., Hegde, A., Meka, A., Ephrem, C. L., Askree, S. H., & Hegde, M. R. (2012). Aberrant firing of replication origins potentially explains intragenic nonrecurrent rearrangements within genes, including the human DMD gene. *Genome Research*, *22*(1), 25–34. https://doi.org/10.1101/gr.123463.111

Arlt, M. F., Mulle, J. G., Schaibley, V. M., Ragland, R. L., Durkin, S. G., Warren, S. T., & Glover, T. W. (2009). Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *American Journal of Human Genetics*, *84*(3), 339–350. https://doi.org/10.1016/j.ajhg.2009.01.024

Bacolla, A., Jaworski, A., Larson, J. E., Jakupciak, J. P., Chuzhanova, N., Abeysinghe, S. S., & Wells, R. D. (2004). Breakpoints of gross deletions coincide with non-B DNA conformations. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(39), 14162–14167. https://doi.org/10.1073/pnas.0405974101

Bacolla, A., Tainer, J. A., Vasquez, K. M., & Cooper, D. N. (2016). Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Research*, *44*(12), 5673–5688. https://doi.org/10.1093/nar/gkw261

Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J. E., & Wells, R. D. (2006). The involvement of non-B DNA structures in gross chromosomal rearrangements. *DNA Repair (Amst)*, *5*(9–10), 1161–1170. https://doi.org/10.1016/j.dnarep.2006.05.032

Bacolla, A., Ye, Z., Ahmed, Z., & Tainer, J. A. (2019). Cancer mutational burden is shaped by G4 DNA, replication stress and mitochondrial dysfunction. *Progress in Biophysics and Molecular Biology*, *147*, 47–61. https://doi.org/10.1016/j.pbiomolbio.2019.03.004

Bacolla, A., Zhu, X., Chen, H., Howells, K., Cooper, D. N., & Vasquez, K. M. (2015). Local DNA dynamics shape mutational patterns of mononucleotide repeats in human genomes. *Nucleic Acids Research*, *43*(10), 5065–5080. https://doi.org/10.1093/nar/gkv364

Ball, E. V., Stenson, P. D., Abeysinghe, S. S., Krawczak, M., Cooper, D. N., & Chuzhanova, N. A. (2005). Microdeletions and microinsertions causing human genetic disease: Common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human Mutation*, *26*(3), 205–213.

Bauters, M., Van Esch, H., Friez, M. J., Boespflug-Tanguy, O., Zenker, M., Vianna-Morgante, A. M., & Froyen, G. (2008). Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Research*, *18*(6), 847–858. https://doi.org/10.1101/gr.075903.107

Béna, F., Gimelli, S., Migliavacca, E., Brun-Druc, N., Buiting, K., Antonarakis, S. E., & Sharp, A. J. (2010). A recurrent 14q32.2 microdeletion mediated by expanded TGG repeats. *Human Molecular Genetics*, *19*(10), 1967–1973. https://doi.org/10.1093/hmg/ddq075

Brown, R. E., & Freudenreich, C. H. (2021). Structure-forming repeats and their impact on genome stability. *Current Opinion in Genetics & Development*, *67*, 41–51. https://doi.org/10.1016/j.gde.2020.10.006

Carlson, J., Locke, A. E., Flickinger, M., Zawistowski, M., Levy, S., Myers, R. M., & Zöllner, S. (2018). Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature Communications*, *9*(1), 3753. https://doi.org/10.1038/s41467-018-05936-5

Carvalho, C. M., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Carvalho, C. M., Pehlivan, D., Ramocki, M. B., Fang, P., Alleva, B., Franco, L. M., & Lupski, J. R. (2013). Replicative mechanisms for CNV formation are error prone. *Nature Genetics*, *45*(11), 1319–1326. https://doi.org/10.1038/ng.2768

Carvalho, C. M., Zhang, F., Liu, P., Patel, A., Sahoo, T., Bacino, C. A., & Lupski, J. R. (2009). Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Human Molecular Genetics*, *18*(12), 2188–2203. https://doi.org/10.1093/hmg/ddp151

Castle, J. C. (2011). SNPs occur in regions with less genomic sequence conservation. *PLOS One*, *6*(6), e20660. https://doi.org/10.1371/journal.pone.0020660

Cer, R. Z., Bruce, K. H., Mudunuri, U. S., Yi, M., Volfovsky, N., Luke, B. T., & Stephens, R. M. (2011). Non-B DB: A database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Research*, *39*, D383–D391. https://doi.org/10.1093/nar/gkq1170

Cer, R. Z., Donohue, D. E., Mudunuri, U. S., Temiz, N. A., Loss, M. A., Starner, N. J., & Stephens, R. M. (2013). Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Research*, *41*(Database issue), D94–D100. https://doi.org/10.1093/nar/gks955

Chuzhanova, N., Chen, J. M., Bacolla, A., Patrinos, G. P., Ferec, C., Wells, R. D., & Cooper, D. N. (2009). Gene conversion causing human inherited disease: Evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. *Human Mutation*, *30*(8), 1189–1198. https://doi.org/10.1002/humu.21020

Cooper, D. N., Bacolla, A., Férec, C., Vasquez, K. M., Kehrer-Sawatzki, H., & Chen, J. M. (2011). On the sequence-directed nature of human gene mutation: The role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. *Human Mutation*, *32*(10), 1075–1099. https://doi.org/10.1002/humu.21557

Cooper, D. N., Ball, E. V., & Mort, M. (2010). Chromosomal distribution of disease genes in the human genome. *Genetic Testing and Molecular Biomarkers*, *14*(4), 441–446. https://doi.org/10.1089/gtmb.2010.0081

Cukier, H. N., Kunkle, B. W., Vardarajan, B. N., Rolati, S., Hamilton-Nelson, K. L., Kohli, M. A., & Pericak-Vance, M. A. (2016). ABCA7 frameshift deletion associated with Alzheimer disease in African Americans. *Neurology: Genetics*, *2*(3), e79. https://doi.org/10.1212/nxg.0000000000000079

Damas, J., Carneiro, J., Amorim, A., & Pereira, F. (2014). MitoBreak: The mitochondrial DNA breakpoints database. *Nucleic Acids Research*, *42*, D1261–D1268. https://doi.org/10.1093/nar/gkt982

Del Mundo, I. M. A., Zewail-Foote, M., Kerwin, S. M., & Vasquez, K. M. (2017). Alternative DNA structure formation in the mutagenic human c-MYC promoter. *Nucleic Acids Research*, *45*(8), 4929–4943. https://doi.org/10.1093/nar/gkx100

Demaerel, W., Mostovoy, Y., Yilmaz, F., Vervoort, L., Pastor, S., Hestand, M. S., & Vermeesch, J. R. (2019). The 22q11 low copy repeats are characterized by unprecedented size and structural variability. *Genome Research*, *29*(9), 1389–1401. https://doi.org/10.1101/gr.248682.119

Dhokarh, D., & Abyzov, A. (2016). Elevated variant density around SV breakpoints in germline lineage lends support to error-prone replication hypothesis. *Genome Research*, *26*(7), 874–881. https://doi.org/10.1101/gr.205484.116

Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M. Y., & Stankiewicz, P. (2013). NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Research*, *23*(9), 1395–1409. https://doi.org/10.1101/gr.152454.112

Dong, D. W., Pereira, F., Barrett, S. P., Kolesar, J. E., Cao, K., Damas, J., & Kaufman, B. A. (2014). Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. *BMC Genomics*, *15*(1), 677. https://doi.org/10.1186/1471-2164-15-677

Dutta, A., Eckelmann, B., Adhikari, S., Ahmed, K. M., Sengupta, S., Pandey, A., Hegde, P. M., Tsai, M.-S., Tainer, J. A., Weinfeld, M., Hegde, M. L., & Mitra, S. (2017). Microhomology-mediated end joining is activated in irradiated human cells due to phosphorylation-dependent formation of

the XRCC1 repair complex. *Nucleic Acids Research*, *45*(5), 2565–2599. https://doi.org/10.1093/nar/gkw1262

Eckelmann, B. J., Bacolla, A., Wang, H., Ye, Z., Guerrero, E. N., Jiang, W., El-Zein, R., Hegde, M. L., Tomkinson, A. E., Tainer, J. A., & Mitra, S. (2020). XRCC1 promotes replication restart, nascent fork degradation and mutagenic DNA repair in BRCA2-deficient cells. *NAR Cancer*, *2*(3). https://academic.oup.com/narcancer

Férec, C., Casals, T., Chuzhanova, N., Macek, M., Jr., Bienvenu, T., Holubova, A., & Chen, J. M. (2006). Gross genomic rearrangements involving deletions in the CFTR gene: Characterization of six new events from a large cohort of hitherto unidentified cystic fibrosis chromosomes and meta-analysis of the underlying mechanisms. *European Journal of Human Genetics*, *14*(5), 567–576. https://doi.org/10.1038/sj.ejhg.5201590

Fontana, G. A., & Gahlon, H. L. (2020). Mechanisms of replication and repair in mitochondrial DNA deletion formation. *Nucleic Acids Research*, *48*(20), 11244–11258. https://doi.org/10.1093/nar/gkaa804

Fujimoto, A., Wong, J. H., Yoshii, Y., Akiyama, S., Tanaka, A., Yagi, H., & Shimada, M. (2021). Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Medicine*, *13*(1), 65. https://doi.org/10.1186/s13073-021-00883-1

Gadgil, R. Y., Romer, E. J., Goodman, C. C., Rider, S. D., Jr., Damewood, F. J., Barthelemy, J. R., & Leffak, M. (2020). Replication stress at microsatellites causes DNA double-strand breaks and break-induced replication. *Journal of Biological Chemistry*, *295*(45), 15378–15397. https://doi.org/10.1074/jbc.RA120.013495

Geng, C., Tong, Y., Zhang, S., Ling, C., Wu, X., Wang, D., & Dai, Y. (2021). Sequence and structure characteristics of 22 deletion breakpoints in intron 44 of the DMD gene based on long-read sequencing. *Frontiers in Genetics*, *12*, 638220. https://doi.org/10.3389/fgene.2021.638220

Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M., & Nik-Zainal, S. (2018). Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Research*, *28*(9), 1264–1271. https://doi.org/10.1101/gr.231688.117

Ghosh, A., & Bansal, M. (2003). A glossary of DNA structures from A to Z. *Acta Crystallographica. Section D, Biological Crystallography*, *59*(Pt 4), 620–626. https://doi.org/10.1107/s0907444903003251

Grajcarek, J., Monlong, J., Nishinaka-Arai, Y., Nakamura, M., Nagai, M., Matsuo, S., & Woltjen, K. (2019). Genome-wide microhomologies enable precise template-free editing of biologically relevant deletion mutations. *Nature Communications*, *10*(1), 4856. https://doi.org/10.1038/s41467-019-12829-8

Guiblet, W. M., Cremona, M. A., Harris, R. S., Chen, D., Eckert, K. A., Chiaromonte, F., & Makova, K. D. (2021). Non-B DNA: A major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Research*, *49*(3), 1497–1516. https://doi.org/10.1093/nar/gkaa1269

Guo, X., Shi, J., Cai, Q., Shu, X. O., He, J., Wen, W., & Long, J. (2018). Use of deep whole-genome sequencing data to identify structure risk variants in breast cancer susceptibility genes. *Human Molecular Genetics*, *27*(5), 853–859. https://doi.org/10.1093/hmg/ddy005

Hambarde, S., Tsai, C.-L., Pandita, R. K., Bacolla, A., Maitra, A., Charaka, V., Hunt, C. R., Kumar, R., Limbo, O., Le Meur, R., Chazin, W. J., Tsutakawa, S. E., Russell, P., Schlacher, K., Pandita, T. K., & Tainer, J. A. (2021). EXO5-DNA structure and BLM interactions direct DNA resection critical for ATR-dependent replication restart. *Molecular Cell*, *81*(14), 2989–3006. https://doi.org/10.1016/j.molcel.2021.05.027

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., Weber, R., & Haussler, D. (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research*, *13*(1), 13–26. https://doi.org/10.1101/gr.844103

Harel, T., & Lupski, J. R. (2018). Genomic disorders 20 years on-mechanisms for clinical manifestations. *Clinical Genetics*, *93*(3), 439–449. https://doi.org/10.1111/cge.13146

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLOS Genetics*, *5*(1), e1000327. https://doi.org/10.1371/journal.pgen.1000327

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, *10*(8), 551–564. https://doi.org/10.1038/nrg2593

Hillmer, M., Summerer, A., Mautner, V. F., Högel, J., Cooper, D. N., & Kehrer-Sawatzki, H. (2017). Consideration of the haplotype diversity at nonallelic homologous recombination hotspots improves the precision of rearrangement breakpoint identification. *Human Mutation*, *38*(12), 1711–1722. https://doi.org/10.1002/humu.23319

Hillmer, M., Wagner, D., Summerer, A., Daiber, M., Mautner, V. F., Messiaen, L., & Kehrer-Sawatzki, H. (2016). Fine mapping of meiotic NAHR-associated crossovers causing large NF1 deletions. *Human Molecular Genetics*, *25*(3), 484–496. https://doi.org/10.1093/hmg/ddv487

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X., & Frazer, K. A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genetics*, *38*(1), 82–85. https://doi.org/10.1038/ng1695

Hu, Q., Lu, H., Wang, H., Li, S., Truong, L., Li, J., & Wu, X. (2019). Break-induced replication plays a prominent role in long-range repeat-mediated deletion. *EMBO Journal*, *38*(24), e101751. https://doi.org/10.15252/embj.2019101751

Inoue, K., & Lupski, J. R. (2002). Molecular mechanisms for genomic disorders. *Annual Review of Genomics and Human Genetics*, *3*, 199–242. https://doi.org/10.1146/annurev.genom.3.032802.120023

Jahic, A., Hinreiner, S., Emberger, W., Hehr, U., Zuchner, S., & Beetz, C. (2017). Doublet-mediated DNA rearrangement—A novel and potentially underestimated mechanism for the formation of recurrent pathogenic deletions. *Human Mutation*, *38*(3), 275–278. https://doi.org/10.1002/humu.23162

Kamat, M. A., Bacolla, A., Cooper, D. N., & Chuzhanova, N. (2016). A role for non-B DNA forming sequences in mediating microlesions causing human inherited disease. *Human Mutation*, *37*(1), 65–73. https://doi.org/10.1002/humu.22917

Kato, T., Inagaki, H., Kogo, H., Ohye, T., Yamada, K., Emanuel, B. S., & Kurahashi, H. (2008). Two different forms of palindrome resolution in the human genome: Deletion or translocation. *Human Molecular Genetics*, *17*(8), 1184–1191. https://doi.org/10.1093/hmg/ddn008

Keegan, N. P., Wilton, S. D., & Fletcher, S. (2019). Breakpoint junction features of seven DMD deletion mutations. *Human Genome Variation*, *6*, 39. https://doi.org/10.1038/s41439-019-0070-x

Keute, M., Miller, M. T., Krishnan, M. L., Sadhwani, A., Chamberlain, S., Thibert, R. L., & Hipp, J. F. (2020). Angelman syndrome genotypes manifest varying degrees of clinical severity and developmental impairment. *Molecular Psychiatry*, *26*, 3625–3633. https://doi.org/10.1038/s41380-020-0858-6

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., & Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, *143*(5), 837–847. https://doi.org/10.1016/j.cell.2010.10.027

Kiktev, D. A., Sheng, Z., Lobachev, K. S., & Petes, T. D. (2018). GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(30), E7109–E7118. https://doi.org/10.1073/pnas.1807334115

Kondrashov, A. S., & Rogozin, I. B. (2004). Context of deletions and insertions in human coding sequences. *Human Mutation*, *23*(2), 177–185. https://doi.org/10.1002/humu.10312

Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., & Levens, D. (2017). Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst*, *4*(3), 344–356. https://doi.org/10.1016/j.cels.2017.01.013

Krawczak, M., & Cooper, D. N. (1991). Gene deletions causing human genetic disease: Mechanisms of mutagenesis and the role of the local DNA sequence environment. *Human Genetics*, *86*(5), 425–441. https://doi.org/10.1007/bf00194629

Lee, J. A., Carvalho, C. M., & Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, *131*(7), 1235–1247. https://doi.org/10.1016/j.cell.2007.11.037

Lemmens, B., van Schendel, R., & Tijsterman, M. (2015). Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers. *Nature Communications*, *6*, 8909. https://doi.org/10.1038/ncomms9909

Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T., & Hurles, M. E. (2019). Similarities and differences in patterns of germline mutation between mice and humans. *Nature Communications*, *10*(1), 4053. https://doi.org/10.1038/s41467-019-12023-w

Liu, P., Carvalho, C. M., Hastings, P. J., & Lupski, J. R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. *Current Opinion in Genetics & Development*, *22*(3), 211–220. https://doi.org/10.1016/j.gde.2012.02.012

MacLean, H. E., Favaloro, J. M., Warne, G. L., & Zajac, J. D. (2006). Double-strand DNA break repair with replication slippage on two strands: a novel mechanism of deletion formation. *Human Mutation*, *27*(5), 483–489. https://doi.org/10.1002/humu.20327

Maranchie, J. K., Afonso, A., Albert, P. S., Kalyandrug, S., Phillips, J. L., Zhou, S., & Linehan, W. M. (2004). Solid renal tumor severity in von Hippel Lindau disease is related to germline deletion length and location. *Human Mutation*, *23*(1), 40–46. https://doi.org/10.1002/humu.10302

Marey, I., Ben Yaou, R., Deburgrave, N., Vasson, A., Nectoux, J., Leturcq, F., & Cossee, M. (2016). Non random distribution of DMD deletion breakpoints and implication of double-strand breaks repair and replication error repair mechanisms. *The Journal of Neuromuscular Diseases*, *3*(2), 227–245. https://doi.org/10.3233/jnd-150134

McVey, M., & Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): Deleted sequences and alternative endings. *Trends in Genetics*, *24*(11), 529–538. https://doi.org/10.1016/j.tig.2008.08.007

Mendez-Dorantes, C., Tsai, L. J., Jahanshir, E., Lopezcolorado, F. W., & Stark, J. M. (2020). BLM has contrary effects on repeat-mediated deletions, based on the distance of DNA DSBs to a repeat and repeat divergence. *Cell Reports*, *30*(5), 1342–1357. https://doi.org/10.1016/j.celrep.2020.01.001

Morales, M. E., Kaul, T., Walker, J., Everett, C., White, T., & Deininger, P. (2021). Altered DNA repair creates novel Alu/Alu repeat-mediated deletions. *Human Mutation*, *42*(5), 600–613. https://doi.org/10.1002/humu.24193

Nambot, S., Thevenon, J., Kuentz, P., Duffourd, Y., Tisserant, E., Bruel, A. L., & Thauvin-Robinet, C. (2018). Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: Substantial interest of prospective annual reanalysis. *Genetics in Medicine*, *20*(6), 645–654. https://doi.org/10.1038/gim.2017.162

Northam, M. R., Moore, E. A., Mertz, T. M., Binz, S. K., Stith, C. M., Stepchenkova, E. I., & Shcherbakova, P. V. (2014). DNA polymerases zeta and Rev1 mediate error-prone bypass of non-B DNA structures. *Nucleic Acids Research*, *42*(1), 290–306. https://doi.org/10.1093/nar/gkt830

Pabis, K. (2021). Triplex and other DNA motifs show motif-specific associations with mitochondrial DNA deletions and species lifespan. *Mechanisms of Ageing and Development*, *194*, 111429. https://doi.org/10.1016/j.mad.2021.111429

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2020). Biostrings: Efficient manipulation of biological strings.

Prihar, G., Verkkoniem, A., Perez-Tur, J., Crook, R., Lincoln, S., Houlden, H., & Haltia, M. (1999). Alzheimer disease PS-1 exon 9 deletion defined. *Nature Medicine (New York, NY, United States)*, *5*(10), 1090. https://doi.org/10.1038/13383

Romiguier, J., Ranwez, V., Douzery, E. J., & Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research*, *20*(8), 1001–1009. https://doi.org/10.1101/gr.104372.109

Sahoo, T., Peters, S. U., Madduri, N. S., Glaze, D. G., German, J. R., Bird, L. M., & Bacino, C. A. (2006). Microarray-based comparative genomic hybridization testing in deletion bearing patients with Angelman syndrome: Genotype-phenotype correlations. *Journal of Medical Genetics*, *43*(6), 512–516. https://doi.org/10.1136/jmg.2005.036913

Sato, D., Lionel, A. C., Leblond, C. S., Prasad, A., Pinto, D., Walker, S., & Scherer, S. W. (2012). SHANK1 deletions in males with autism spectrum disorder. *American Journal of Human Genetics*, *90*(5), 879–887. https://doi.org/10.1016/j.ajhg.2012.03.017

Seo, S. H., Bacolla, A., Yoo, D., Koo, Y. J., Cho, S. I., Kim, M. J., & Jeon, B. (2020). Replication-based rearrangements are a common mechanism for SNCA duplication in Parkinson's disease. *Movement Disorders*, *35*(5), 868–876. https://doi.org/10.1002/mds.27998

Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., & Eichler, E. E. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics*, *38*(9), 1038–1042. https://doi.org/10.1038/ng1862

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., & Cooper, D. N. (2020). The Human Gene Mutation Database (HGMD (®)): Optimizing its use in a clinical diagnostic or research setting. *Human Genetics*, *139*, 1197–1207. https://doi.org/10.1007/s00439-020-02199-3

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., & Cooper, D. N. (2014). The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, *133*(1), 1–9. https://doi.org/10.1007/s00439-013-1358-4

Summerer, A., Mautner, V. F., Upadhyaya, M., Claes, K. B. M., Högel, J., Cooper, D. N., & Kehrer-Sawatzki, H. (2018). Extreme clustering of type-1 NF1 deletion breakpoints co-locating with G-quadruplex forming sequences. *Human Genetics*, *137*(6-7), 511–520. https://doi.org/10.1007/s00439-018-1904-1

Svetec Miklenić, M., & Svetec, I. K. (2021). Palindromes in DNA-A risk for genome stability and implications in cancer. *International Journal of Molecular Sciences*, *22*(6), 2840. https://doi.org/10.3390/ijms22062840

Tan, E. K. (2016). Chromosomal deletion at 22q11.2 and Parkinson's disease. *Lancet Neurology*, *15*(6), 538–540. https://doi.org/10.1016/s1474-4422(16)00115-0

Tanay, A., & Siggia, E. D. (2008). Sequence context affects the rate of short insertions and deletions in flies and primates. *Genome Biology*, *9*(2), R37. https://doi.org/10.1186/gb-2008-9-2-r37

Tsutakawa, S. E., Bacolla, A., Katsonis, P., Bralić, A., Hamdan, S. M., Lichtarge, O., Tainer, J. A., & Tsai, C.-L. (2021). Decoding Cancer Variants of Unknown Significance for Helicase–Nuclease–RPA Complexes Orchestrating DNA Repair During Transcription and Replication. *Frontiers in Molecular Biosciences*, 8. https://doi.org/10.3389/fmolb.2021.791792

Tsutakawa, S. E., Thompson, M. J., Arvai, A. S., Neil Alexander, J., Shaw Steven, J., Algasaier, S. I., Kim, J. C., Finger, L. D., Jardine, E., Gotham, V. J. B., Sarker, A. H., Her, M. Z., Rashid, F., Hamdan, S. M.,

Mirkin, S. M., Grasby, J. A., & Tainer, J. A. (2017). Phosphate steering by Flap Endonuclease 1 promotes 5′-flap specificity and incision to prevent genome instability. *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms15855

Twayana, S., Bacolla, A., Barreto-Galvez, A., De-Paula, R. B., Drosopoulos, W. C., Kosiyatrakul, S. T., & Schildkraut, C. L. (2021). Translesion polymerase eta both facilitates DNA replication and promotes 3 increased human genetic variation at common fragile sites 4. *Proceedings of the National Academy of Sciences of the United States of America*, *118*, e2106477118.

Vaags, A. K., Lionel, A. C., Sato, D., Goodenberger, M., Stein, Q. P., Curran, S., & Scherer, S. W. (2012). Rare deletions at the neurexin 3 locus in autism spectrum disorder. *American Journal of Human Genetics*, *90*(1), 133–141. https://doi.org/10.1016/j.ajhg.2011.11.025

Verdin, H., D'Haene, B., Beysen, D., Novikova, Y., Menten, B., Sante, T., & De Baere, E. (2013). Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. *PLOS Genetics*, *9*(3), e1003358. https://doi.org/10.1371/journal.pgen.1003358

Visser, R., Shimokawa, O., Harada, N., Kinoshita, A., Ohta, T., Niikawa, N., & Matsumoto, N. (2005). Identification of a 3.0-kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-Mb microdeletion. *American Journal of Human Genetics*, *76*(1), 52–67. https://doi.org/10.1086/426950

Visser, R., Shimokawa, O., Harada, N., Niikawa, N., & Matsumoto, N. (2005). Non-hotspot-related breakpoints of common deletions in Sotos syndrome are located within destabilised DNA regions. *Journal of Medical Genetics*, *42*(11), e66. https://doi.org/10.1136/jmg.2005.034355

Vissers, L. E., Bhatt, S. S., Janssen, I. M., Xia, Z., Lalani, S. R., Pfundt, R., & Stankiewicz, P. (2009). Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. *Human Molecular Genetics*, *18*(19), 3579–3593. https://doi.org/10.1093/hmg/ddp306

Vocke, C. D., Ricketts, C. J., Schmidt, L. S., Ball, M. W., Middelton, L. A., Zbar, B., & Linehan, W. M. (2021). Comprehensive characterization of Alu-mediated breakpoints in germline VHL gene deletions and rearrangements in patients from 71 VHL families. *Human Mutation*, *42*(5), 520–529. https://doi.org/10.1002/humu.24194

Vogt, J., Bengesser, K., Claes, K. B., Wimmer, K., Mautner, V. F., van Minkelen, R., & Kehrer-Sawatzki, H. (2014). SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. *Genome Biology*, *15*(6), R80. https://doi.org/10.1186/gb-2014-15-6-r80

Vogt, P. H., Bender, U., Deibel, B., Kiesewetter, F., Zimmer, J., & Strowitzki, T. (2021). Human AZFb deletions cause distinct testicular pathologies depending on their extensions in Yq11 and the Y haplogroup: New cases and review of literature. *Cell & Bioscience*, *11*(1), 60. https://doi.org/10.1186/s13578-021-00551-2

Wang, Y., Su, P., Hu, B., Zhu, W., Li, Q., Yuan, P., & Wang, Y. (2015). Characterization of 26 deletion CNVs reveals the frequent occurrence of micro-mutations within the breakpoint-flanking regions and frequent repair of double-strand breaks by templated insertions derived from remote genomic regions. *Human Genetics*, *134*(6), 589–603. https://doi.org/10.1007/s00439-015-1539-4

Wells, R. D. (2007). Non-B DNA conformations, mutagenesis and disease. *Trends in Biochemical Sciences*, *32*(6), 271–278. https://doi.org/10.1016/j.tibs.2007.04.003

Wu, X., Lu, Y., Ding, Q., You, G., Dai, J., Xi, X., & Wang, X. (2014). Characterisation of large F9 deletions in seven unrelated patients with severe haemophilia B. *Thrombosis and Haemostasis*, *112*(3), 459–465. https://doi.org/10.1160/th13-12-1060

Xu, J., Mo, Z., Ye, D., Wang, M., Liu, F., Jin, G., & Sun, Y. (2012). Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4. *Nature Genetics*, *44*(11), 1231–1235. https://doi.org/10.1038/ng.2424

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, *41*(7), 849–853. https://doi.org/10.1038/ng.399

Zhang, F., Seeman, P., Liu, P., Weterman, M. A., Gonzaga-Jauregui, C., Towne, C. F., & Lupski, J. R. (2010). Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. *American Journal of Human Genetics*, *86*(6), 892–903. https://doi.org/10.1016/j.ajhg.2010.05.001

Zhao, J., Bacolla, A., Wang, G., & Vasquez, K. M. (2010). Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Science*, *67*(1), 43–62. https://doi.org/10.1007/s00018-009-0131-2

Zheng, S., Fu, J., Vegesna, R., Mao, Y., Heathcock, L. E., Torres-Garcia, W., & Verhaak, R. G. (2013). A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival. *Genes and Development*, *27*(13), 1462–1472. https://doi.org/10.1101/gad.213686.113

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Qi, M., Stenson, P. D., Ball, E. V., Tainer, J. A., Albino, B., Kehrer-Sawatzki, H., Cooper, D. N., & Zhao, H. (2022). Distinct sequence features underlie microdeletions and gross deletions in the human genome. *Human Mutation*, 43, 328–346. https://doi.org/10.1002/humu.24314