HUMAN MUTATION 23:177–185 (2004)

RESEARCH ARTICLE

Context of Deletions and Insertions in Human Coding Sequences

Alexey S. Kondrashov* and Igor B. Rogozin

National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland

Communicated by Mark H. Paalman

We studied the dependence of the rate of short deletions and insertions on their contexts using the data on mutations within coding exons at 19 human loci that cause mendelian diseases. We confirm that periodic sequences consisting of three to five or more nucleotides are mutagenic. Mutability of sequences with strongly biased nucleotide composition is also elevated, even when mutations within homonucleotide runs longer than three nucleotides are ignored. In contrast, no elevated mutation rates have been detected for imperfect direct or inverted repeats. Among known candidate contexts, the indel context GTAAGT and regions with purine-pyrimidine imbalance between the two DNA strands are mutagenic in our sample, and many others are not mutagenic. Data on mutation hot spots suggest two novel contexts that increase the deletion rate. Comprehensive analysis of mutability of all possible contexts of lengths four, six, and eight indicates a substantially elevated deletion rate within YYYTG and similar sequences, which is one of the two contexts revealed by the hot spots. Possible contexts that increase the insertion rate (AT(A/C)(A/C)GCC and TACCRC) and decrease deletion (TATCGC) or insertion (GCGG) rates have also been identified. Two-thirds of deletions remove a repeat, and over 80% of insertions create a repeat, i.e., they are duplications. Hum Mutat 23:177–185, 2004. Published 2003 Wiley-Liss, Inc.†

KEY WORDS: mutation; hot spot; deletion; insertion; mutable motif; nucleotide context; repeat; microsatellite

DATABASES:
ftp://ftp.ncbi.nih.gov/pub/kondrashov/context (compilation of data on the 19 loci analyzed in this study)

INTRODUCTION

Mutability varies substantially along nucleotide sequences. At some extremely mutable sites, mutation rates exceed the average per site rate by an order of magnitude or more. However, such mutation hot spots [Benzer, 1961; Coulondre et al., 1978] are rare in human coding sequences [Kondrashov, 2003], with the only exception being substitution hot spots at methylated CpG sites [Cooper and Youssoufian, 1988]. Thus, at the majority of sites, local mutation rates deviate from the average by no more than a factor of two to five.

Often, elevated or reduced mutability at a site can be correlated to its local sequence context. Mutagenic contexts (or at-risk sequence motifs [ARMs] [Gordenin and Resnick, 1998]) can be defined either by a particular sequence (textual contexts), or by some relationship within the sequence (relational contexts). Several types of relationships, including small-scale periodicity (microsatellites) and direct and inverted repeats are known to be mutagenic [Streisinger et al., 1966; Drake and Baltz, 1976; Miller, 1983; Jeffreys et al., 1985; Ripley, 1990; Cooper and Krawczak, 1993; Gordenin and Resnick, 1998; Strauss, 1999; Bebenek and Kunkel, 2000]. A known textual mutagenic context of deletions in humans is TGRR(G/T)R [Krawczak et al., 1998]. A context causing complex mutations (indels), GTAAGT, has recently been identified [Chuzhanova et al., 2003a].

Properties of sites with unusually high (or low) mutation rates can shed light on the mechanisms of spontaneous mutation [Miller, 1983; Horsfall et al., 1990; Boulikas, 1992; Dogliotti et al., 1998; Rogozin et al., 2001a; Rogozin and Pavlov, 2003; Maki, 2002]. For example, comparison of mutational hot spots at the human APC locus with the error spectrum of DNA polymerase β suggests that at least some mutations at this locus are caused by errors of this polymerase [Muniappan and Thilly, 2002]. A similar comparison suggests that DNA polymerase η is involved in somatic hypermutation of mammalian immunoglobulin genes [Rogozin et al., 2001b; Pavlov et al., 2002].

Received 17 April 2003; accepted revised manuscript 15 October 2003.

*Correspondence to: Alexey S. Kondrashov, National Center for Biotechnology Information, Building 38A, National Institutes of Health, Bethesda, MD 20894. E-mail: kondrashov@ncbi.nlm.nih.gov

DOI 10.1002/humu.10312
Published online in Wiley InterScience (www.interscience.wiley.com).

Here, we analyze local contexts of deletions and insertions in coding regions of 19 human loci that cause mendelian diseases. We consider only deletions and insertions, because such mutations, at least when causing a frameshift, always lead to loss-of-function phenotypes. In contrast, phenotypic ascertainment of nucleotide substitutions is incomplete and involves unavoidable biases, which obscures patterns in mutation.

MATERIALS AND METHODS

Sets of Deletions and Insertions

We used the same set of 20 loci as in Kondrashov [2003], except for F8 (F8C), in which deletions and insertions are traditionally not described properly (only the codon where a mutation happened is usually reported, which is often insufficient to determine exactly how the DNA sequence has been changed). Sets of deletions and insertions were updated, in December 2002, for the following loci: PAX6 (pax6.hgu.mrc.ac.uk), PKD1 [Rossetti et al., 2002], RB1 (www.d-lohmann.de/Rb/mutations.html), ABCD1 (www.x-ald.nl), AR (ww2.mcgill.ca/androgendb), DMD [Mendell et al., 2001], F9 (www.kcl.ac.uk/ip/petergreen/haemB-database.html), IDS [Rathmann et al., 1996], IL2RG (www.nhgri.-nih.gov/DIR/GMBB/SCID), and OTC [Tuchman et al., 2002]. The 19 sets of deletions and insertions used for our study are available at (ftp://ftp.ncbi.nih.gov/pub/kondrashov/context). The total numbers of deletions and insertions in these sets are 1829 and 385, respectively.

Hot Spots

We regarded as a hot spot each site at which a particular deletion or insertion has been found at least three times. This threshold was obtained using the CLUSTERM program [Glazko et al., 1998; Rogozin et al., 2001a]. With our samples, the local mutation rate at a hot spot so defined is at least 10^{-8}, which exceeds the average per nucleotide rate of deletions or insertions by factors of 20 or 50, respectively [Kondrashov, 2003].

Analysis of the Impact of a Context

A context may contain sites of two types: 1) those where mutations are taken into account (denoted by uppercase letters (a context must contain at least one such site); and 2) those which only determine whether the context is present at a particular location (denoted by lowercase letters). For example, a context aTGc is present (exactly) if and only if the sequence contains the segment ". . . atgc . . ."; however only mutations that affect the two central nucleotides of such segments will be taken into account.

For a context, we calculated n+ and n–, the numbers of nucleotides in the coding exons of a locus that belong and do not belong to it (only to uppercase sites), and d+ and d– (i+ and i–), the numbers of deletions (insertions) that occurred within and outside of the uppercase sites of the context. A nucleotide was considered as belonging to the context when it was covered by the context on either DNA strand. When the exact position of a mutation was uncertain (for example, a mutation that transforms ". . . atgta . . ." into ". . . ata . . ." can be a deletion of either tg or gt), each possible position was included with the weight 1/q, where q is the total number of possible positions for the mutation. For a deletion, every deleted nucleotide was considered as a site where the deletion occurred. For an insertion, both nucleotides that flank the inserted sequence were considered as sites of the insertion.

The impact of the context on the per nucleotide deletion rate at the mth locus was described by the ratio of the densities of deletions within and outside the context, Rm = (d+/n+)/(d–/n–) (loci at which n+ = 0 were treated as missing data; for reasonable contexts, d– and n– are always nonzero). After this, the average impact, I, and its standard error, E, were calculated for the set of Rm values corresponding to the 19 loci. Insertions were treated analogously.

Nucleotides that belong to mutagenic periodic sequences (i.e., homonucleotide runs longer than three nucleotides, sequences in which a segment of length two is presented more than two times, or sequences in which a segment of length three, four, or five is presented at least twice; see below) were ignored, together with mutations at these sequences, when other contexts were investigated. In some cases, only subsets of mutations (e.g., only deletions of length one) were analyzed. An *ad hoc* C program performing the analyses is available at ftp://ftp.ncbi.nih.gov/pub/kondrashov/context.

Choice of Potentially Important Contexts

The analysis described above tests the impact of a particular context on the mutation rate. We identified contexts to be tested in four ways. The first two ways rely on the existing data on spontaneous mutation, and the other two ways do not use any preexisting information.

First, we considered known ARMs [Gordenin and Resnick, 1998], all of which are relational contexts. Second, we tested textual contexts known or suspected to affect mutation in other datasets. This information was collected from the literature (cited below) and from the compilation of recombination signals and mutational hot spots (ftp.bionet.nsc.ru/pub/biology/dbms/RE-COMB.ZIP).

Third, we looked for common contexts in mutation hot spots using the MEME [Grundy et al., 1996] and REGRT [Berikov and Rogozin, 1999] programs. Fourth, we identified potential contexts automatically. This was done as follows. First, we tested the impact on mutability of all possible 4^L contexts of length L (L = 4, 6, or 8). For each such context, all sequences that deviate from it by no more than k nucleotides were treated as belonging to this context. After this, we selected a small fraction of the most (or the least) mutable contexts, and performed their classification using single-link clustering [Kondrashov and Shabalina, 2002]. For this purpose, two contexts were considered similar if and only if they differed from each other by a single substitution. For several of the most populous classes, we derived their consensus sequences and studied their impacts on mutability.

Analysis of the Impact of Imperfect Direct or Inverted Repeats

It has been suggested that deletions and insertions may result from repair of short heteroduplexes formed by complementary regions within imperfect direct or inverted repeats [Ripley and Glickman, 1983; Golding and Glickman, 1985]. We attempted to detect such heteroduplex-repair mutagenesis using a modification of a Monte Carlo procedure implemented in the CONSEN program [Rogozin and Kondrashav, 1992; Rogozin and Pavlov, 2003]. A weight Wj of site j is N*M/L, where N is the number of deletions/insertions at this site that are compatible with the heteroduplex-repair mechanism, M is the number of complementary nucleotides in a potential heteroduplex ($M > 4$), and L is the distance between two regions of direct or inverted repeats ($5 < L < 100$). The average of Wj, W, was calculated for all sites in the mutation target sequence. The distribution of average statistical weights W_{random} was calculated for 10,000 groups of random sites. Each group contained a number of mutations equal to the observed one with the same distribution of mutations throughout the sites. Based on the distribution W_{random}, a probability that W is below W_{random}, $P(W \leq W_{random})$ was calculated.

RESULTS

Hot Spots

A total of 50 deletion hot spots and 10 insertion hot spots were detected at the 19 loci. Only 21 deletion hot spots occurred within periodic contexts; eight deletion hot spots occurred within yyYTG contexts, two occurred at the acACTTacaa motif, and the rest involved diverse sequences without obvious common features (Table 1). Only seven hot spots involved deletions of one nucleotide, and deletions of length four were responsible for 16 hot spots. In contrast, deletions of one nucleotide were five times more common than deletions of four nucleotides among all deletions in human coding sequences (see Fig. 5 of Kondrashov [2003]).

Most of the insertion hot spots were located within periodic sequences, and most of the corresponding insertions were only 1 nucleotide long (Table 1). The difference between the prevalences of periodic sequences in deletion vs. insertion hot spots is statistically significant (by the Fisher exact test, $P = 0.03$).

Mutation at Periodic Sequences

Figures 1, 2, and 3 present data on the mutation rates in periodic sequences. Sequences with period equal to one (homonucleotide runs) are mutagenic when they are four or more nucleotides long (Fig. 1). Sequences with period equal to two are mutagenic when the number of identical segments of two nucleotides is three or more (Fig. 2; there was not enough data for insertions into such sequences). In both cases, the mutation rate grows rapidly with the number of identical segments. When the period is three nucleotides or longer, even two identical segments in tandem are mutagenic, at least for deletions, and the mutation rate increases with the length of the period (Fig. 3; in our data there were too few sequences with three or more such segments to study the dependence of the mutation rate on the number of identical segments).

For all mutagenic periodic sequences (i.e., for those of length ≥ 4 with a repeated segment of length one, or of length ≥ 6 with a repeated segment of length two, or with at least two repeated segments of length ≥ 3), their average impacts on the rates of deletion and insertion were 2.27 ± 0.16 and 2.01 ± 0.25, respectively. Over one-third of all deletions (628), and over 60% (236) of all insertions occur within such periodic sequences.

Mutation of Sequences With Biased Nucleotide Composition

Even when we ignore mutations within homonucleotide runs longer than three nucleotides, which are mutagenic per se, short sequences that mostly consist of just one nucleotide have elevated mutation rates. For example, the impacts of sequences of length six with five identical nucleotides on the rates of deletion and insertion are 2.48 ± 0.41 and 2.84 ± 1.44, respectively. For sequences of length eight with six or seven identical nucleotides, the corresponding impacts on the rates of deletion and insertion are 2.99 ± 0.55 and 2.30 ± 0.86, respectively.

Mutation Within Imperfect Direct or Inverted Repeats

We did not observe an increased mutation rate at imperfect direct or inverted repeats. For mutations that can be interpreted as products of heteroduplex repair events, $P(W < W_{random})$ varied between 0.12 and 0.96. Thus, the observed cooccurrence of deletions/insertions and repeats was not statistically significant.

Mutation at Textual Contexts

Table 2 lists two known textual contexts that were found to increase the deletion rate, as well as some other previously studied contexts which were not significantly mutagenic in our dataset.

Screening of all contexts of length eight (under $k = 2$) reveals 59 contexts with high deletion rates, each of which had $I > 2.5$ and $I–2^*E > 1.0$ (these conditions ensure that the context increases the deletion rate substantially, and that this increase is statistically significant, $P < 0.05$). Classification of these contexts produces 31 classes, three of which each contain more than five members (Table 3). We can see that all these classes contain, in three different phases, essentially the same context, which also appears in eight deletion hot spots (Table 1). If, as suggested by the hot spots, we define this context as yyYTG (or CARrr in the opposite strand) and allow one deviation from the exact context ($k = 1$), its impacts on deletion and insertions rates are 3.19 ± 0.72 and 1.18 ± 0.33, respectively. If we define this context as cyCTGt ($k = 1$), its impacts on deletion and insertions rates are 2.24 ± 0.42 and 1.36 ± 0.37, respectively. Screening of all contexts of lengths four (with $k = 0$) and six (with $k = 1$) revealed the same mutable context (data not reported). Essentially the same context has also been found by the MEME and REGRT programs. However, all other predictions made by these programs on the basis of hot spots were not confirmed when the complete gene sequences were taken into account (data not reported).

Screening of all contexts of length six (with $k = 1$) revealed 28 contexts with low deletion rates, each of which had $I < 0.5$ and $I–2^*E < 1.0$. Their classification produced 24 classes, 23 with one context each, and one with five contexts. The impacts of the consensus sequence of this largest class, TATCGC ($k = 1$) on deletion and insertion rates are 0.24 ± 0.087 and 2.62 ± 0.97, respectively. Screening of all contexts of lengths eight and four did not reveal additional clear-cut contexts with low deletion rates.

Screening of all contexts of length eight (with $k=2$) revealed 82 contexts with high insertion rates, each of which had $I>2.5$ and $I–2^*E>1.0$. Their classification produced only two classes with more than three members. The impacts of the consensus sequence of the first class, AT(A/C)(A/C)GCC ($k=1$) on deletion and insertion rates are 1.15 ± 0.30 and 2.66 ± 0.64, respectively. The corresponding impacts of the consensus sequence of the

TABLE 1. Deletion and Insertion Hotspots*

Locus	GenBank Accession	Hotspot position[a]	Number of cases	Sequence[b]
Deletions				
Periodic				
APC	M74088.1	523	6	cactaaaaga**ATAGatagt**cttcct
APC	M74088.1	3939	109	tagcagaaat**AAAAGaaaag**attgga
APC	M74088.1	4403	8	actgctgaaa**AGagagagag**tgg
APC	M74088.1	4403	3	actgctgaaa**AGAGagagag**tggac
JAG1	U73936.1	2531	12	ctccaggtga**CAGTcagt**gtgatga
NF1	M82814.1	706	6	aggaattaac**TGTTtgtt**cagaaga
RB1	L11910.1	162204	3	tttgctctag**Cccccta**ccttg
TSC2	X75621.1	4436	3	cgcaggggca**AGagag**tagagag
VHL	NM_000551.1	296	15	tcccaggtca**TCTtctg**caatcgc
ABCD1	XM_010174.3	870	5	tggccaactc**GGAgga**gatcgcct
AVPR2	L22206.1	1428	7	agtgattgtg**GTCg**tctatgtgct
CYBB	NM_000397.2	655	3	ctctttgtga**TCTtctt**cattggc
CYBB	NM_000397.2	713	3	ggcagaccgc**AGagag**tttggct
EMD	X86810.1	1103	5	gcggctctcg**Ccccccc**agctcg
F9	K02402.1	33801	4	acataatatt**GAGga**gacagaaca
F9	K02402.1	34075	5	agagttccac**TTGttga**ccgagcc
F9	K02402.1	34122	3	caccatctat**AACaa**catgttctg
IL2RG	NM_000206.1	826	5	ggctccatgg**GATTgatt**atcagcc
Periodic-like				
APC	M74088.1	3198	69	aagatgaaat**AAAA**Caaagtgagcaa
APC	M74088.1	3199	3	agatgaaata**AAA**Caaagtgagcaa
APC	M74088.1	3940	4	agcagaaata**AAA**Gaaaagattgga
yyYTG motif				
APC	M74088.1	4629	5	ggaatgaaa*cAGa*atcagagcag
JAG1	U73936.1	1898	4	ccagcaa*cccCTgttt*gaatggg
NF1	M82814.1	1752	3	ccaagaaaa*cAGggg*cccgaaac
ABCD1	XM_010174.3	1415	41	gatgtggaa*cAGggg*atcatctg
AR	M20132.1	2660	5	acttcgc*ccc*T*g*atctggtttt
AVPR2	L22206.1	927	3	gccagatg*ccCTg*tgtcgggccg
CYBB	NM_000397.2	1052	3	catccgc*ccc*T*g*aggaagactt
F9	K02402.1	23351	3	tggaagagtt*TCtg*tttcacaaa
acACTTacaa motif				
APC	M74088.1	2818	5	acattcaaa*cACTTacaa*tttcact
NF1	M82814.1	6998	6	aggacctg*acACTTacaa*cagtcaa
Others				
APC	M74088.1.1	1933	3	aggtgggataTtacggaatgtg
APC	M74088.1.1	2909	4	aatgatagttTAAAtagtgtcagta
APC	M74088.1.1	3181	6	acccaaacacATAATagaagatgaaa
APC	M74088.1.1	3216	16	gtgagcaaagACAAtcaaggaatca
APC	M74088.1.1	3217	3	tgagcaaagaCAATcaaggaatcaa
APC	M74088.1.1	3594	3	catcacagaaACagtcattttca
APC	M74088.1.1	3614	3	tcattctcaaAGagttcatctgg
APC	M74088.1.1	3988	3	gagcgaagttCcagcagtgtca
JAG1	U73936.1	3574	3	gccgttgcagAAGTaagagttcaga
NF1	M82814.1	1228	3	gggaagataaCTctgtcattttc
NF1	M82814.1	3181	3	gcattgaaacAATgatgttaaatc
NF1	M82814.1	3667	3	tgtcaaacttACTCaatgccaacgt
RB1	L11910.1	170400	4	atcattcgggGTGAgtattttcttt
RB1	L11910.1	153353	4	ttataaaaaaGgttagtagatg
ABCD1	XM_010174.3	1412	4	gtggatgtggAAcaggggatcat
EMD	X86810.1	1350	4	gaggacgcttTACTCtaccagagcaa
F9	K02402.1	23495	5	ggttgttggtGGAgaagatgccaa
IDS	NM_000202.2	924	3	accttgcctgACAAacagagcactg
IDS	NM_000202.2	1544	3	ggaacttgtgTCtttttccca
Insertions				
Periodic				
APC	M74088.1.1	1873	3	ttggttggca**CTctct**tactt
APC	M74088.1.1	4040	3	agggttctag**Ttttat**cttc**a**
JAG1	U73936.1	1612	5	agtgtgtgtg**Cccccca**ca**g**
NF1	M82814.1	2237	4	tgcagcggaa**Ccccccc**aat
AVPR2	L22206.1	1328	3	agagaggcct**Gggggg**cgcc
BTK	NM_000061	714	3	accggaagac**Aaaaaaa**gcct
CYBB	NM_000397.2	750	15	gtttgtgaac**Aaaaaaat**ctc
EMD	X86810.1	2420	4	cgtgctcctg**GGGCTgggctgggct**
Other				
TSC2	X75621.1	3115	3	atggctcgatAacgtcttctc
IL2RG	NM_000206.1	373	3	tgcaaaaaaaTggagatccac

*For each locus, the GenBank accession of the sequence used in the corresponding locus-specific database is presented. Deleted or inserted nucleotides are capitalized.

[a]The position of a hotspot is the leftmost possible position of the first deleted or inserted nucleotide, i. e., the number of the first capitalized nucleotide. Such nucleotides are also underlined. Nucleotides are numbered as in the sequence whose accession is provided.

[b]Periodic sequences affected by mutations are in Bold. Two deletion motifs are in italics. Y=T or C.

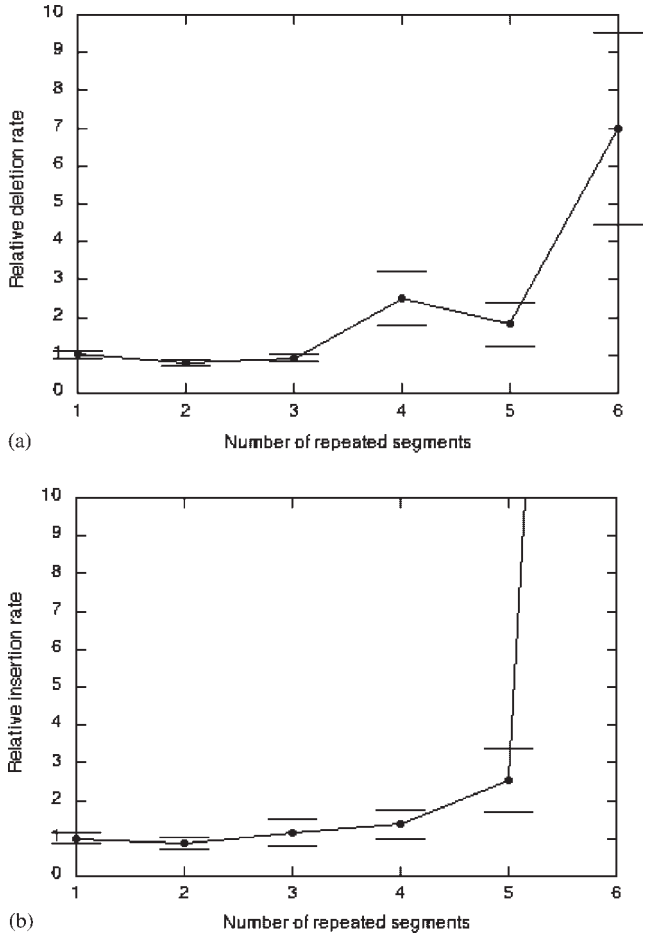

FIGURE 1. **The impact of periodic sequences with period of one on deletion (a) and insertion (b) rate. Bars show the standard errors.**

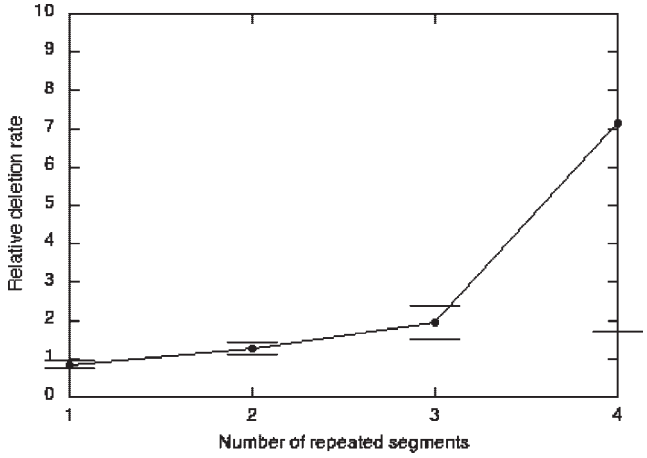

FIGURE 2. **The impact of periodic sequences with period of two on deletion rate.**

second class, TACCRC (k=1), on deletion and insertion rates are 0.74 ± 0.21 and 3.36 ± 1.54, respectively. Screening of all contexts of lengths six and four did not reveal additional clear-cut contexts with high insertion rate.

Screenings of all contexts of lengths four, six, and eight produced several classes of contexts with low insertion

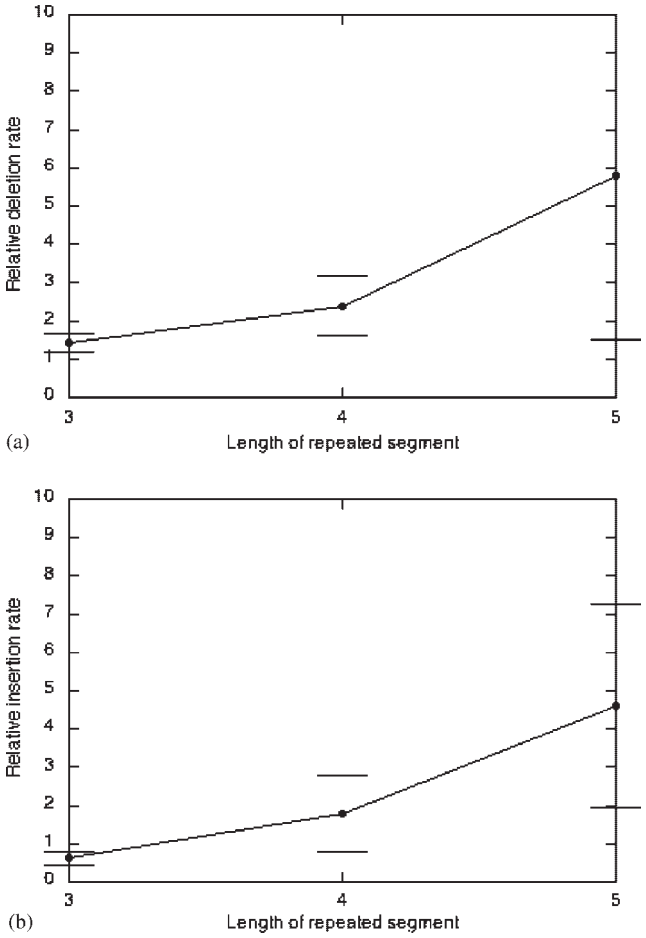

FIGURE 3. **The impact of exact, direct repeats of lengths three, four, and five on deletion (a) and insertion (b) rate.**

rates, whose consensus sequences shared one common motif, GCGG. The impacts of GCGG sequence (k=0) on rates of deletion and insertions are 0.55 ± 0.25 and 0.07 ± 0.07, respectively.

Repeat Removals and Duplications

Among all deletions, 66% (1212) lead to removal of a repeat ("deduplication"), in the sense that the deleted sequence is identical to a sequence bordering the site of deletion. Among all insertions, 81% (311) are duplications, i.e., the inserted sequence is identical to a sequence bordering the site of insertion.

DISCUSSION

Data on disease-causing deletions and insertions at autosomal dominant or X-linked loci are suitable for studying contexts of mutation. Indeed, drastic, frameshift alleles of such loci must persist in the population for only few generations, so that different patients must carry independent mutations. Even at loci-causing late-onset (e.g., APC; Bjork et al. [1999]) or relatively mild (e.g., JAG1; Crosnier et al. [1999]) diseases, 50% or more of patients carry de novo mutations (see Kondrashov [2003] for review), indicating short persistent times, at least for complete loss-of-function alleles.

TABLE 2. Impact on the Mutation Rate of Some Previously Described Textual Contexts

Context[a]	Deviations[b]	Impact on deletion rate[c]	Impact on insertion rate[c]	Comments	Reference
GTAAGT	(1)	1.87 (0.37)	0.44 (0.19)	Indel context	Chuzhanova et al. [2003]
RrRRRrr	(1)	1.72 (0.28)	1.36 (0.29)	R/Y imbalance	Boulikas [1992]
(A/T)(A/T)(A/T)(A/T)(A/T)	(1)	1.34 (0.16)	2.41 (0.86)	AT-rich	Boulikas [1992]
(G/C)(G/C)(G/C)(G/C)(G/C)	(1)	1.24 (0.40)	0.83 (0.11)	GC-rich	Boulikas [1992]
TGRR(G/T)R	(0)	0.83 (0.15)	2.10 (0.83)	Deletion hotspot motif	Krawczak et al. [1998]
GGGCAGGARG	(3)	0.90 (0.13)	1.16 (0.40)	Human minilatellite core	Jeffreys et al. [1985]
TGAAGA	(0)	1.01 (0.38)	0.86 (0.67)	Polymerase α arrest motif	Todorova and Danieli [1997]
GYTRGYRG	(1)	0.97 (0.25)	0.40 (0.19)	χ site consensus	Myers and Stahl [1994]
RYYYRR	(0)	0.75 (0.10)	0.96 (0.24)	Topoisomerase I, motif 1	Shen and Shen [1990]
YYRY	(0)	1.07 (0.15)	1.36 (0.17)	Topoisomerase I, motif 2	Shen and Shen [1990]
GTN(A/T)AYATTNATNNG	(3)	0.92 (0.18)	1.75 (0.58)	Topoisomerase II	Dobbs et al. [1994]
CG	(0)	1.01 (0.20)	0.73 (0.23)	Substitution context	Cooper and Youssoufian [1988]

[a]R=A or G, Y=T or C.
[b]The maximal allowed numbers of deviations from the exact context are shown.
[c]The average impact I of a context on mutability at the 19 loci and its standard error E (in parentheses) are presented.

TABLE 3. Three Most Populous Classes of All Sequences of Length 8 With Large Positive Impacts on Deletion Rate

	Class 1	Class 2	Class 3
	CTCTTGTT	CGCTTTTT	CCTGTTTT
	CCCTTGTT	CACTGTTT	CCTCTGTT
	CCACTGTT	CCCTGTTT	TCTGTGTT
	TCCCTGTT	CGCTGTTT	CCTGTGTT
	ACCCTGTT	CTGTGTTT	CCTGTTAT
	CCCCTGTT	CCGTGTTT	CCTGTTTC
	TCGCTGTT	CACTGTCT	
	CCCTTGGT	CCCTGTGT	
	CCCCTGTC	CTCTGTTA	
		CCCTGTTA	
		CTGTGTTA	
		CTGTGTCA	
		CCTTGTTC	
		CTCTGTTC	
		CCCTGTTC	
		CTGTGTTC	
		CTCTGTTG	
Consensus:	CCCCTGTT	CTCTGTTT C	CCTGTTTT G

Our analysis confirms that sequence periodicity is mutagenic [Streisinger et al., 1966; Miller, 1983; Ripley, 1990; Gordenin and Resnick, 1998; Bebenek and Kunkel, 2000]. The impact of periodicity rapidly increases with the number and length of repeated sequence segments (Figs. 1–3). Similar results were obtained by Greenblatt et al. [1996] and Halandoga et al. [2001] for somatic mutations. However, periodicity per se does not determine the mutation rate exactly. Some periodic sequences are mutation hot spots (Table 1), but many others with the same patterns of periodicity are not, and periodic hot spots of deletions and of insertions do not overlap. On average, deletions in human coding sequences are approximately three times more common than insertions [Kondrashov, 2003], however within some periodic sequences, insertions are much more common than deletions. Expanding disease-causing microsatellites (CTG)n, (CGG)n, and (GAA)n are well-known examples of such sequences [Mitas, 1997; Petruska et al., 1998]. In contrast, periodic sequences that are more prone to deletions than to insertions will disappear, unless maintained by purifying selection.

Contexts containing primarily one nucleotide (e.g., AAAGACAA) are also mutagenic, even when we disregard mutagenic homonucleotide runs. This suggests that a relaxed version of Streisinger's model [Streisinger et al., 1966], allowing some deviations from exact periodicity at periodic contexts, is still applicable to spontaneous mutation in human protein-coding genes.

We did not observe any significant increase in mutation at contexts that contain inverted or direct repeats separated by 5–100 nucleotides. Thus, our data offer no support for the short heteroduplex repair model of mutation [Ripley and Glickman, 1983; Golding and Glickman, 1985].

Among the contexts known or suspected to be mutagenic, our data support only two. Contexts with R/Y imbalance between strands [Boulikas, 1992], and a motif of complex mutations (indels) GTAAGT [Chuzhanova et al., 2003a] were found to increase the deletion (but not the insertion) rate. Also, our data showed that AT-rich sequences may be marginally mutagenic.

We found two new contexts that increase the deletion rate. The more common one is yyYTG (Table 3). This context is present in eight deletion hot spots (Table 1). A similar motif ytG (hot spot of deletions of one nucleotide G) was observed in the spectra of errors produced by *E. coli* DNA polymerases I in vitro [Papanicolaou and Ripley, 1989]. Also, (CTG)n is prone to duplication events in several human disease-causing genes (reviewed by Mitas [1997]). Three independent observations of error-prone synthesis of CTG-containing sequences in vivo and in vitro suggest a general property of different DNA polymerases. Another deletion motif, acACTTacaa (k = 0), has been encountered only in two hot spots (Table 1). Among all the deletions at hot spots, four-nucleotide-long deletions were overrepresented (Table 1). An excess of four-nucleotide-long deletions has also been found among spontaneous mutations in the *E. coli lacI* gene [Schaaper et al., 1986].

We have found two new contexts that increase the insertion rate. Although statistically significant, they probably should still be treated with caution, since the amount of data on insertions was four times below that on deletions. Eight out of 10 insertion hot spots produce single-nucleotide insertions and are located within periodic sequences (Table 1). These features differentiate them from deletion hot spots and suggest that mechanisms of deletions and insertions have different context properties. This is also supported by the absence of any overlaps between deletion and insertion hot spots (Table 1).

We have also identified one context each as a deletion and an insertion cold spot, TATCGC and GCGG. Mutation cold spots may be harder to identify than hot spots, since mutations from a sample may be absent within a particular context simply by chance. However, a large number of sites in our sample of exons of 19 loci belong to our insertion cold spot sequence (tetranucleotide with no deviations allowed) or deletion cold spot sequence (hexanucleotide with one deviation allowed). Thus, these cold spot contexts may well be real.

Some mutation-affecting contexts, such as the CpG motif, which facilitates substitution [Coorper and Youssoufian, 1988] can be defined unequivocally. Often, however, a large number of similar short sequences are known to have higher (or lower) mutabilities, and the context is hard to define. Sometimes, it may be desirable to consider several related contexts [Rogozin and Pavlov, 2003]. For example, mutation hot spots associated with somatic hypermutation in immunoglobulin genes have been reported as rGy(a/t), G being the mutable base, or gaRy(a/t) (see Rogozin and Pavlov [2003]). Rogozin et al. [2001b] proposed a statistical method to evaluate the relative merits of different consensus sequences. However, statistical analysis of 15 mutational spectra in immunoglobulin genes suggested not one sequence, but two sequences, rGy(a/t) and aGy(a/t), that had the same best score. Both motifs were used for further analysis of errors made by DNA polymerases in vitro [Rogozin et al., 2001b]. Here, we considered different variants of mutable motifs. We suggested the yyYTG consensus sequence, however several hot spots have one mismatch with this sequence (e.g., the deletion hot spot in the position 5012, Table 1), and were not included in the yyYTG set (Table 1). Thus, some other variants of suggested mutable contexts of deletions/insertions might exist, although a more accurate formal description of these motifs awaits larger datasets.

Perhaps the patterns in deletions and insertions observed within our sample of 19 human loci are representative of other coding genes. However, intergenic regions may have substantially different patterns in deletions and insertions, since local properties of noncoding and coding sequences are not the same (for example, noncoding sequences contain more repetitive fragments and fewer CpG sites). The ratio of deletion and insertion rates in noncoding regions is not yet known.

We considered only deletions and insertions that are no longer than 10 nucleotides. Longer deletions and insertions have been ignored, because they are rare in the datasets that we used and are often not exactly described [Kondrashov, 2003]. Overall, long deletions and insertions are infrequent compared to short deletions and insertions [Weber et al., 2002; Britten et al., 2003]. Long deletions often occur between direct or inverted repeats [Efstratiadis et al., 1980; Albertini et al., 1982; Ehrlich et al., 1993; Gordenin and Resnick, 1998; Smit, 1999; Sinden et al., 1999] or between repetitive elements [Prak and Kazazian, 2000; Makalowski, 2000; Rogozin et al., 2000; Deininger and Batzer, 2002; Kazazian and Goodier, 2002]. Other patterns in long deletions and translocations have also been recently reported [Abeysinghe et al., 2003; Chuzhanova et al., 2003b].

Mutations in periodic contexts cannot be used as fingerprints for identifying DNA polymerases and/or repair enzymes, since many of them are error-prone (at least DNA polymerases are [Bebenek and Kunkel, 2000]) in such contexts. Fortunately, many mutagenic contexts described here consist of nonperiodic sequences. In vitro studies [Pavlov et al., 2002; Muniappan and Thilly, 2002] may identify DNA polymerases that are error-prone within these contexts, and thus shed light on the mechanisms of spontaneous mutation.

ACKNOWLEDGMENTS

We thank Olga Sinitsina and Elena Vasunina for discussions and for providing information about mutable motifs from the compilation of recombination signals and mutational hot spots. Two anonymous reviewers made several very useful suggestions.

REFERENCES

Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer I: nucleotide composition and recombination-associated motifs. Hum Mutat 22:229–244.

Albertini AM, Hofer M, Calos MP, Miller JH. 1982. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell 29:319–328.

Bebenek K, Kunkel TA. 2000. Streisinger revisited: DNA synthesis errors mediated by substrate misalignments. Cold Spring Harb Symp Quant Biol 65:81–91.

Benzer S. 1961. On the topography of the genetic fine structure. Proc Natl Acad Sci USA 47:403–415.

Berikov VB, Rogozin IB. 1999. Regression trees for analysis of mutational spectra in nucleotide sequences. Bioinformatics 15:553–562.

Bjork J, Akerbrant H, Iselius L, Alm T, Hultcrantz R. 1999. Epidemiology of familial adenomatous polyposis in Sweden: changes over time and differences in phenotype between males and females. Scand J Gastroentero 34:1230–1235.

Boulikas T. 1992. Evolutionary consequences of nonrandom damage and repair of chromatin domains. J Mol Evol 35:156–180.

Britten RJ, Rowen L, Williams J, Cameron RA. 2003. Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci USA 100:4661–4665.

Chuzhanova NA, Anassis EJ, Ball EV, Krawczak M, Cooper DN. 2003a. Meta-analysis of indels causing human genetic disease:

mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat 21:28–44.

Chuzhanova N, Abeysinghe SS, Krawczak M, Ball EV, Cooper DN. 2003b. Translocation and gross deletion breakpoints in human inherited disease and cancer II: potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. Hum Mutat 22:245–251.

Cooper DN, Krawczak M. 1993. Human gene mutation. Oxford: Bios Scientific Publishers. 402p.

Cooper DN, Youssoufian H. 1988. The CpG dinucleotide and human genetic disease. Hum Genet 78:151–155.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274:775–780.

Crosnier C, Driancourt C, Raynaud N, Dhorne-Pollet S, Pollet N, Bernard O, Hadchouel M, Meunier-Rotival M. 1999. Mutations in JAGGED1 gene are predominantly sporadic in alagille syndrome. Gastroenterology 116:1141–1148.

Deininger PL, Batzer MA. 2002. Mammalian retroelements. Genome Res 12:1455–1465.

Dobbs CL, Shaiu W-L, Benbow RM. 1994. Modular sequence elements associated with origin regions in eukaryotic chromosomal DNA. Nucleic Acids Res 22:2479–2489.

Dogliotti E, Hainaut P, Hernandez T, D'errico M, Demarini DM. 1998. Mutation spectra resulting from carcinogenic exposure: from model systems to cancer-related genes. Recent Results Cancer Res 154:97–124.

Drake JW, Baltz RH. 1976. The biochemistry of mutagenesis. Annu Rev Biochem 45:11–37.

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ. 1980. The structure and evolution of the human β-globin gene family. Cell 21:653–668.

Ehrlich SD, Bierne H, d'Alencon E, Vilette D, Petranovic M, Noirot P, Michel B. 1993. Mechanisms of illegitimate recombination. Gene 135:161–166.

Glazko GV, Milanesi L, Rogozin IB. 1998. The subclass approach for mutational spectrum analysis: application of the SEM algorithm. J Theor Biol 192:475–487.

Golding GB, Glickman BW. 1985. Sequence-directed mutagenesis: evidence from a phylogenetic history of human alpha-interferon genes. Proc Natl Acad Sci USA 82:8577–8581.

Gordenin DA, Resnick MA. 1998. Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability. Mutat Res 400:45–58.

Greenblatt MS, Grollman AP, Harris CC. 1996. Deletions and insertions in the p53 tumor suppressor gene in human cancers: confirmation of the DNA polymerase slippage misalignment model. Cancer Res 56:2130–2136.

Grundy WN, Bailey TL, Elkan CP. 1996. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. Comput Appl Biosci 12:303–310.

Halandoga A, Still JG, Hill KA, Sommer SS. 2001. Spontaneous microdeletions and microinsertions in a transgenic mouse mutation detection system: analysis of age, tissue, and sequence specificity. Environ Mol Mutagen 37:311–323.

Horsfall MJ, Gordon AJ, Burns PA, Zielenska M, Van Der Vliet GM, Glickman BW. 1990. Mutational specificity of alkylating agents and the influence of DNA repair. Environ Mol Mutagen 15:107–122.

Jeffreys AJ, Wilson V, Lay ST. 1985. Hypervariable "minisatellite" regions in human DNA. Nature 314:67–73.

Kazazian HH Jr, Goodier JL. 2002. LINE drive. retrotransposition and genome instability. Cell 110:277–280.

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474–488.

Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21:12–27.

Kondrashov AS, Shabalina SV. 2002. Classification of common conserved sequences in mammalian intergenic regions. Hum Mol Genet 11:669–674.

Makalowski W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene 259:61–67.

Maki H. 2002. Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. Annu Rev Genet 36:279–303.

Mendell JR, Buzin CH, Feng J, Yan J, Serrano C, Sangani DS, Wall C, Prior TW, Sommer SS. 2001. Diagnosis of Duchenne dystrophy by enhanced detection of small mutations. Neurology 57:645–650.

Miller JH. 1983. Mutational specificity in bacteria. Annu Rev Genet 17:215–238.

Mitas M. 1997. Trinucleotide repeats associated with human disease. Nucleic Acids Res 25:2245–2254.

Muniappan BP, Thilly WG. 2002. The DNA polymerase β replication error spectrum in the adenomatous polyposis coli gene contains human colon tumor mutational hotspots. Cancer Res 62:3271–3275.

Myers RS, Stahl FW. 1994. Chi and the RecBCD enzyme of *Escherichia coli*. Annu Rev Genet 28:49–70.

Papanicolaou C, Ripley LS. 1989. Polymerase-specific differences in the DNA intermediates of frameshift mutagenesis. In vitro synthesis errors of *Escherichia coli* DNA polymerase I and its large fragment derivative. J Mol Biol 207:335–353.

Pavlov YI, Rogozin IB, Galkin AP, Aksenova AY, Hanaoka F, Rada C, Kunkel TA. 2002. Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase η during copying of a mouse immunoglobulin κ light chain transgene. Proc Natl Acad Sci USA 99:9954–9959.

Petruska J, Hartenstine MJ, Goodman MF. 1998. Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. J Biol Chem 273:5204–5210.

Prak ET, Kazazian HH Jr. 2000. Mobile elements and the human genome. Nat Rev Genet 1:134–144.

Rathmann M, Bunge S, Beck M, Kresse H, Tylki-Szymanska A, Gal A. 1996. Mucopolysaccharidosis type II (Hunter syndrome): mutation "hot spots" in the iduronate-2-sufatase gene. Am J Hum Genet 59:1202–1209.

Ripley LS. 1990. Frameshift mutation: determinants of specificity. Annu Rev Genet 24:189–213.

Ripley LS, Glickman BW. 1983. Unique self-complementarity of palindromic sequences provides DNA structural intermediates for mutation. Cold Spring Harb Symp Quant Biol 47(Pt 2):851–861.

Rogozin IB, Kolchanov NA. 1992. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochim Biophys Acta 1171:11–18.

Rogozin IB, Mayorov VI, Lavrentieva MV, Milanesi L, Adkison LR. 2000. Prediction and phylogenetic analysis of mammalian short interspersed elements (SINEs). Brief Bioinform 1:260–274.

Rogozin IB, Kondrashov FA, Glazko GV. 2001a. Use of mutation spectra analysis software. Hum Mutat 17:83–102.

Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. 2001b. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. Nat Immunol 2:530–536.

Rogozin IB, Pavlov YI. 2003. Theoretical analysis of mutational hotspots and their DNA sequence context specificity. Mutat Res 544:65–85.

Rossetti S, Chauveau D, Walker D, Saggar-Malik A, Winearls CG, Torres VE, Harris PC. 2002. A complete mutation screen of the ADPKD genes by DHPLC. Kidney Int 61:1588–1599.

Schaaper RM, Danforth BN, Glickman BW. 1986. Mechanisms of spontaneous mutagenesis: an analysis of the spectrum of spontaneous mutation in the *Escherichia coli* lacI gene. J Mol Biol 189:273–284.

Shen CC, Shen C-KJ. 1990. Specificity and flexibility of the recognition of DNA helical structure by eukaryotic topoisomerase I. J Mol Biol 212:67–78.

Sinden RR, Hashem VI, Rosche WA. 1999. DNA-directed mutations. Leading and lagging strand specificity. Ann NY Acad Sci 870:173–189.

Smit AFA. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657–663.

Strauss BS. 1999. Frameshift mutation, microsatellites and mismatch repair. Mutat Res 437:195–203.

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M. 1966. Frameshift mutations and the genetic code. Cold Spring Harb Symp Quant Biol 31:77–84.

Todorova A, Danieli GA. 1997. Large majority of single-nucleotide mutations along the dystrophin gene can be explained by more than one mechanism of mutagenesis. Hum Mutat 9:537–547.

Tuchman M, Jaleel N, Morizono H, Sheehy L, Lynch MG. 2002. Mutations and polymorphisms in the human ornithine transcarbamylase gene. Hum Mutat 19:93–107.

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. 2002. Human diallelic insertion/deletion polymorphisms. Am J Hum Genet 71:854–862.