Supplemental Information
Somatic microindels in human cancer: The insertions are highly error-prone and derive from nearby but not adjacent sense and antisense templates

The subset of 1-M and N-1 microindels do not obviously fit the alternative model of a base substitution with an adjacent deletion or insertion.

About half the microindels have one base inserted (1-M microindels). These mutations might be described by an alternative model hypothesizing a one base substitution with an adjacent deletion, but are not consistent with the signatures of either single-base substitutions or microdeletions within the *TP53* gene. The 31 *TP53* somatic microindels with one bp inserted (1-M microindels, SI Table 6A) or the 6 *TP53* somatic microindels with one bp deleted (N-1 microindels, SI Table 6B) might be hypothesized to result from a pure microdeletion or pure microinsertion, respectively, and a single base substitution at either end. However, the observed substitutions by this alternative model do not reflect the hotspots, and in general, the spectrum of substitutions in the IARC database, providing evidence against this hypothesis. SI Table 6 shows the hypothesized substitutions in this alternative model of 1-M and N-1 microindels. When the hypothesized substitutions are categorized by their observed frequency in the IARC database, there are dramatically fewer "hot" substitutions than expected ($p<0.000000005$ and $p=0.01$ for 1-M and N-1 microindels, respectively; SI Table 7). For example, 35% of the substitutions in the IARC database are substitutions that are reported more than 100 times, so if 1-M microindels were actually due to a pure microdeletion and single base substitution, then the hypothesized substitutions for about 11 of the 31 1-M microindels would be expected to be among those that appear more than 100 times in the IARC database, yet none are. In addition, the deletions postulated by this model do not have the signature of pure microdeletions (1-4) and the distribution of deletion sizes in 1-M microindels is not different from that in the other microindel types that cannot derive from this alternative mechanism ($p=0.3$, data not shown).

On the other hand, when the hypothesized substitutions in 1-M microindels are categorized by their observed frequency in the IARC database, the distribution does not quite fit the expected distribution for randomized *TP53* substitutions (p=0.02 compared to the substitutions in the IARC database uniformly distributed over the set of unique substitutions). This difference may reflect non-randomness in the location of microindels (see discussion of recurroids) or the involvement of base substitutions having a mutation signature dramatically different from the general signature of base substitutions within *TP53*. We conclude that the alternative model for 1-M and N-1 microindels is possible for some of the events, but the signatures for the hypothesized base substitutions, deletions and insertions differ from the expected signatures of these types of mutations.

Nonhomologous end joining is a potential mechanism for microindels.

One mechanism considered for the origin of microindels is nonhomologous end joining (NHEJ), a common pathway for the repair of DNA double strand breaks (DSB) in multicellular eukaryotes (5). NHEJ involves modification of the broken ends to create a compatibility that facilitates rejoining. Before modification, the DNA ends created by DSBs are generally incompatible, but if microhomology of 1 to 4 bp occurs at the ends, end joining is facilitated (6,7). While not essential, any microhomology biases the joining for a preferred alignment. NHEJ is characterized by a diversity of DSB sequence ends, alignments and end-processing. An *in vitro* system used to study NHEJ showed evidence of at least four principal mechanisms of NHEJ, likely accomplished by different protein complexes and leading to a diversity of outcomes (7). NHEJ produces primarily small deletions (6,7) and to a lesser extent small insertions (8). NHEJ appears mechanistically to be a potential source of microindels, however a review of the literature did not find microindels to be a common signature outcome of NHEJ.

Microindels were recently observed as a minority of events in a plasmid based NHEJ assay in mycobacteria (9).

HPRT **microindels are similar to** ***TP53*** **microindels.**

The frequency of microindels requires a large database of mutations to obtain multiple microindels. To obtain a sample of microindels not in cancers for comparison, microindels were analyzed in *HPRT*. *HPRT* microindels were extracted from the Human *HPRT* Mutation Database (http://www.ibiblio.org/dnam/mainpage.html). Eleven *HPRT* microindels were identified: six somatic mutations with known mutagen exposure, four somatic mutations with no known mutagen exposure, and one germline mutation (SI Table 9). Among these *HPRT* microindels, 1-2 microindels are the most common type (45%, 5/11). This small sample of *HPRT* microindels has features that are generally similar to those of the sample of *TP53* microindels.

Supplemental Figure Legends

SI Figure 4. Spectrum of *TP53* somatic microindels. The microindels are shown on the genomic sequence (GenBank accession X54156.1) containing the *TP53* gene. The nucleotide range displayed includes exons 3 through 9. Jagged vertical lines indicate breaks in the spectrum where intronic regions have been removed. Solid vertical lines indicate microindel counts.

SI Figure 5. Distribution of the net sequence length change resulting from *TP53* somatic microindels. The inset triangle shows the distribution of microindels classified by the number of nucleotides inserted and deleted (e.g., 9: 1-2 indicates that there are 9 microindels with 1 nt inserted and 2 nt deleted).

SI Figure 6. Microindels graphed by type. The counts of microindels of each type, defined by the number of nucleotides inserted and deleted, are graphed in a 2D grid with the length of the inserted sequence along the vertical axis and the length of deleted sequence along the horizontal axis. A) All *TP53* somatic microindels. B) Microindels in the Human Germline Mutation Database (HGMD).

SI Figure 7. Comparison of size distributions of inserted and deleted sequences. Comparison of inserted (A) or deleted (B) sequence sizes in *TP53* pure microinsertions (A) or pure microdeletions (B), *TP53* somatic microindels, and germline (HGMD) microindels.

SI Figure 8. Spectrum of *TP53* pure microinsertions. The pure microinsertions are shown on the genomic sequence (GenBank accession X54156.1) containing the *TP53* gene. The nucleotide range displayed includes exons 3 through 9. Dashed vertical lines indicate the

beginning (short dashes) and ends (long dashes) of exons. Solid vertical lines indicate microinsertion counts.

SI Figure 9. Alternative serial replication slippage models. Serial replication slippage models (10,11) in cis (SRScis) and trans (SRStrans) were applied in attempts to explain, respectively, the first and last microindels shown in Fig. 2. A. Leading strand SRScis model for microindel 13337_13345 del CATCTTATC ins GCCCCT. B. Lagging strand SRScis model for microindel 13337_13345 del CATCTTATC ins GCCCCT. C. Leading-lagging-leading strand SRStrans model for microindel 123599_123602 del CTCA ins TGAGTACTATGAG. D. Lagging-leading-lagging strand SRStrans model for microindel 123599_123602 del CTCA ins TGAGTACTATGAG.

SI Table 2. Comparison of Somatic and Germline Microindels, Microinsertions and Microdeletions

	Somatic	Germline	
Mutation Types	*TP53*	*F8* and *F9**	HGMD†
Base substitutions	18,629 (88.4%)	980 (82.1%)	n/a
Pure microinsertions (≤50 bp)	596 (2.8%)	50 (4.2%)	n/a
Pure microdeletions (≤50 bp)	1,782 (8.5%)	158 (13.2%)	n/a
Microindels (net of ≤50 bp)	66 (0.3%)	5 (0.4%)	155 (0.4%)
Total mutation events	21,073‡	1,193	~36,000
Pure microinsertions (1-20 bp)	578 (97.0%)	48 (96.0%)	n/a
Pure microinsertions (21-50 bp)	18 (3.0%)	2 (4.0%)	n/a
Pure microdeletions (1-20 bp)	1,649 (92.5%)	157 (99.4%)	n/a
Pure microdeletions (21-50 bp)	133 (7.5%)	1 (0.6%)	n/a
Microindels, 1-20 bp net change	62 (93.9%)	5 (100%)	155
Microindels, 21-50 bp net change	4 (6.1%)	0	n/a§
Microindels, Net gain	11 (16.7%)	4 (80.0%)	58 (37.4%)
Microindels, Net loss	55 (83.3%)	1 (20.0%)	97 (62.6%)
Microindels, 1-2 type	9 (13.6%)¶	0	31 (20.0%)¶
Microindels, 2-1 type	4 (6.1%)¶	0	13 (8.4%)¶
Microindels, 1-M type	32 (48.5%)¶	0	55 (35.5%)¶
Microindels, N-1 type	6 (9.1%)¶	1 (20.0%)¶	21 (13.5%)¶

*Current values for mutations collected in our unpublished data; part of the *F9* data was summarized in Sommer et al. (2001). For *F8*, hotspot mutations were excluded, including the common inversion hotspot mutation and single base insertions/deletions at three 7-9 bp oligo A tracts within the coding sequence.

†The version of HGMD (Human Gene Mutation Database) used in the Chuzhanova et al. (2003) meta-analysis of indels, excluding those resulting in zero net length change (i.e., tandem-base mutations). "n/a" indicates that the count is not available.

‡Less than 21,587 since complex mutations, tandem-base mutations, and insertions and deletions larger than 50 bp are excluded.

§HGMD includes microindels in which the deleted and inserted sequences are ~22 bp or less.

¶Percentage of microindels.

SI Table 3. Duplication errors for microindels in which the inserted sequence is at least six nucleotides.

Type	Gene	Nucleotide*	Insertion Size	Errors
Sense duplication of nearby but not adjacent sequence	*TP53*	13337	6	0
	PTEN	30639	8	1
Sense duplication overlapping the deleted sequence	*TP53*	13145	8	2
	CFTR	68869	6	1
		110441	12	1
	EGFR	162287	11	1
		162295	7	1
Antisense duplication of nearby but not adjacent sequence	*TP53*	13959	7	0
		13387	11	2
	PTEN	94480	9	3
	HPRT	33307	6	0
		13163	8	1
Antisense duplication overlapping the deleted sequence	*TP53*	13320	8	0
	CFTR	123599	13	3
		Total	120	16 (13%)

*Nucleotide numbering according to genomic reference sequences listed in Fig. 2 legend.

SI Table 4. Putative mechanisms for microindels in which the inserted sequence is at least six nucleotides.

	Errors*	
	Zero	One or More
Sense duplication of nearby but not adjacent sequence	H	H
Antisense duplication of nearby but not adjacent sequence	HHM	HHH
Sense duplication overlapping the deleted sequence	M	HHHHH
Antisense duplication overlapping the deleted sequence	H	H
Sense duplication of adjacent sequence		M
Antisense duplication of adjacent sequence		M
Putative Adduct Block	M	
Other	M	

*Each "H" represents one microindel in the human *TP53*, *CFTR, EGFR, PTEN* and *HPRT* data herein. Each "M" represents one mouse microindel in the *lacI* data in Gonzalez et al (13).

SI Table 5A. Somatic *TP53* Microindels

Nucleotide Range*	Ins-Del	5' Flanking Sequence‡ (20 bp)	Sequence Inserted†	Sequence Deleted‡	3' Flanking Sequence‡ (20 bp)	Cancer Topography	F/M	Age	PubMed ID
11912_11918	2-7	CTCTTGTCTTTCAGACTTCC	AG	TGAAAAC	AACGTTCTGgtaaggacaag	Liver & Intrahepatic Bile Ducts			12759240
12139_12140	1-2	AATGCCAGAGGCTGCTCCCC	T	GC	GTGGCCCCTGCACCAGCAGC	Lung & Bronchus			1324794
12234_12245	2-12	TCCCTTCCCAGAAAACCTAC	AG	CAGGGCAGCTAC	GGTTTCCGTCTGGGCTTCTT	Breast	F		7862445
12240_12245	4-6	CCCAGAAAACCTACCAGGGC	TTAA	AGCTAC	GGTTTCCGTCTGGGCTTCTT	Breast	F	84	7862445
12268_12287	1-20	TTTCCGTCTGGGCTTCTTGC	C	ATTCTGGGACAGCCAAGTCT	GTGACTTGCACGgtcagttg	Mouth			12434417
12286	2-1	GCATTCTGGGACAGCCAAGT	GA	C	TGTGACTTGCACGgtcagtt	Rectum	M	59	9735666
12288_12302	1-15	ATTCTGGGACAGCCAAGTCT	T	GTGACTTGCACGgtc	agttgccctgaggggctggc	Breast	F	38	7862445
12294_12300	1-7	GGACAGCCAAGTCTGTGACT	A	TGCACGg	tcagttgccctgaggggctg	Breast	F		8611423
13057_13061	1-5	tccttcctcttcctacagTA	A	CTCCC	CTGCCCTCAACAAGATGTTT	Renal Pelvis	F	71	10761705
13074	2-1	gTACTCCCCTGCCCTCAACA	TG	A	GATGTTTTGCCAACTGGCCA	Mouth			10225439
13082_13083	1-2	CTGCCCTCAACAAGATGTTT	C	TG	CCAACTGGCCAAGACCTGCC	Breast	F		12203794
13093_13100	1-8	AAGATGTTTTGCCAACTGGC	G	CAAGACCT	GCCCTGTGCAGCTGTGGGTT	Ovary	F	41	9550561
13116_13117	1-2	GACCTGCCCTGTGCAGCTGT	A	GG	GTTGATTCCACACCCCCGCC	Breast	F	71	14697642
13122_13123	1-2	CCCTGTGCAGCTGTGGGTTG	G	AT	TCCACACCCCCGCCCGGCAC	Lung & Bronchus	M	69	10353731
13127_13133	1-7	TGCAGCTGTGGGTTGATTCC	N†	ACACCCC	CGCCCGGCACCCGCGTCCGC	Ovary	F		8934544
13132_13133	1-2	CTGTGGGTTGATTCCACACC	T	CC	CGCCCGGCACCCGCGTCCGC	Skin§			11668523
13134_13135	1-2	GTGGGTTGATTCCACACCCC	T	CG	CCCGGCACCCGCGTCCGCGC	Larynx	M	49	7882279
13145_13155	8-11¶	CCACACCCCCGCCCGGCACC	TCGCGTCG	CGCGTCCGCGC	CATGGCCATCTACAAGCAGT	Colorectum, NOS	M	69	1319835
13157_13160	3-4	CCGGCACCCGCGTCCGCGCC	TCT	ATGG	CCATCTACAAGCAGTCACAG	Oropharynx	M		21225878
13184_13198	4-15	TCTACAAGCAGTCACAGCAC	NNNN†	ATGACGGAGGTTGTG	AGGCGCTGCCCCCACCATGA	Ovary	F		9723023
13201_13209	3-9	CACATGACGGAGGTTGTGAG	CCT	GCGCTGCCC	CCACCATGAGCGCTGCTCAG	Breast	F	37	8950983
13209_13217	1-9	GGAGGTTGTGAGGCGCTGCC	A	CCCACCATG	AGCGCTGCTCAGATAGCGAT	Ovary	F	49	21199244
13320_13334	8-15	tcctcactgattgctcttag	CAGACCTA	GTCTGGCCCCTCCTC	AGCATCTTATCCGAGTGGAA	Stomach			8180781
13333_13368	1-36	ctcttagGTCTGGCCCCTCC	C	TCAGCATCTTATCCGAGTGGAAGGA AATTTGCGTGT	GGAGTATTTGGATGACAGAA	Renal Pelvis	F	66	9761125
13337_13345	6-9	TAGGTCTGGCCCCTCCTCAG	GCCCCT	CATCTTATC	CGAGTGGAAGGAAATTTGCG	Ovary	F		9891239
13381_13382	3-2	TTGCGTGTGGAGTATTTGGA	ATT	TG	ACAGAAACACTTTTCGACAT	Skin§		28	8319200
13385_13387	1-3	GTGTGGAGTATTTGGATGAC	T	AGA	AACACTTTTCGACATAGTGTG	Brain	F	26	9224526
13387_13388	3-2	GTGGAGTATTTGGATGACAG	CCC	AA	ACACTTTTCGACATAGTGTG	Hematopoietic & Reticuloendothelial Systems			7727782
13387_13392	11-6	GTGGAGTATTTGGATGACAG	CCCACACGCAT	AAACAC	TTTTCGACATAGTGTGGTGG	Hematopoietic & Reticuloendothelial Systems			8289498
13397_13404	2-8	TGGATGACAGAAACACTTTT	AT	CGACATAG	TGTGGTGGTGCCCTATGAGC	Ovary	F		8481915
13397_13404	1-8	TGGATGACAGAAACACTTTT	A	CGACATAG	TGTGGTGGTGCCCTATGAGC	Colorectum, NOS			9516972
13413_13417	4-5	TTTTCGACATAGTGTGGTGG	NNNN†	TGCCC	TATGAGCCGCCTGAGgtctg	Breast	F	54	9569050
13425_13427	2-3	TGTGGTGGTGCCCTATGAGC	TG	CGC	CTGAGgtctggtttgcaact	Skin§			7997263
13431	3-1	GGTGCCCTATGAGCCGCCTG	TCT	A	Ggtctggtttgcaactgggg	Lung & Bronchus			15161705
14011_14026	1-16	atctcctagGTTGGCTCTGA	T	CTGTACCACCATCCAC	TACAACTACATGTGTAACAG	Esophagus	M	62	7768632
14013_14014	1-2	ctcctagGTTGGCTCTGACT	A	GT	ACCACCATCCACTACAACTA	Mouth			10225439
14029_14032	1-4	GACTGTACCACCATCCACTA	G	CAAC	TACATGTGTAACAGTTCCTG	Liver & Intrahepatic Bile Ducts	M	40	10212000

14037_14050	2-14	CACCATCCACTACAACTACA	CG	TGTGTAACAGTTCC	TGCATGGGCGGCATGAACCG	Ovary	F	50	21199244
14040_14042	2-3	CATCCACTACAACTACATGT	TT	GTA	ACAGTTCCTGCATGGGCGGC	Breast	F	63	8611423
14042_14045	1-4	TCCACTACAACTACATGTGT	T	AACA	GTTCCTGCATGGGCGGCATG	Hematopoietic & Reticuloendothelial Systems	M	84	1959992
14042_14045	1-4	TCCACTACAACTACATGTGT	T	AACA	GTTCCTGCATGGGCGGCATG	Liver & Intrahepatic Bile Ducts	F	63	15288479
14045_14054	1-10	ACTACAACTACATGTGTAAC	G	AGTTCCTGCA	TGGGCGGCATGAACCGGAGG	Corpus Uteri	F	76	11733960
14055_14068	1-14	CATGTGTAACAGTTCCTGCA	A	TGGGCGGCATGAAC	CGGAGGCCCATCCTCACCAT	Breast	F		11051239
14066_14072	1-7	GTTCCTGCATGGGCGGCATG	G	AACCGGA	GGCCCATCCTCACCATCATC	Bladder			21278281
14070_14073	3-4	CTGCATGGGCGGCATGAACC	AGA	GGAG	GCCCATCCTCACCATCATCA	Colorectum, NOS			15541358
14072_14073	1-2	GCATGGGCGGCATGAACCGG	T	AG	GCCCATCCTCACCATCATCA	Colorectum, NOS			12921629
14074_14078	4-5	ATGGGCGGCATGAACCGGAG	ACCC	GCCCA	TCCTCACCATCATCACACTG	Breast	F		9288052
14088_14089	4-2	CCGGAGGCCCATCCTCACCA	ATCA	TC	ATCACACTGGAAGACTCCAG	Esophagus	M	72	8722219
14105	2-1	CCATCATCACACTGGAAGAC	GC	T	CCAGgtcaggagccacttgc	Kidney			7633433
14453_14475	1-23	cttttcctatcctgagtagT	A	GGTAATCTACTGGGACGGAACAG	CTTTGAGGTGCGTGTTTGTG	Lung & Bronchus			8934544
14462_14476	1-15	tcctgagtagTGGTAATCTA	A	CTGGGACGGAACAGC	TTTGAGGTGCGTGTTTGTGC	Lung & Bronchus	F	73	7767998
14486_14488	4-3	GACGGAACAGCTTTGAGGTG	TGTC	CGT	GTTTGTGCCTGTCCTGGGAG	Esophagus	F	54	8722219
14487	2-1	ACGGAACAGCTTTGAGGTGC	CC	G	TGTTTGTGCCTGTCCTGGGA	Hematopoietic & Reticuloendothelial Systems	F	65	1423304
14511_14515	1-5	TTGTGCCTGTCCTGGGAGAG	C	ACCGG	CGCACAGAGGAAGAGAATCT	Larynx	M		21225878
14514	3-1	TGCCTGTCCTGGGAGAGACC	CCC	G	GCGCACAGAGGAAGAGAATC	Hematopoietic & Reticuloendothelial Systems			8289498
14515_14518	3-4	GCCTGTCCTGGGAGAGACCG	ACG	GCGC	ACAGAGGAAGAGAATCTCCG	Bones, Joints & Articular Cartilage of Other & Unspecified Sites			8336944
14515_14518	2-4	GCCTGTCCTGGGAGAGACCG	AG	GCGC	ACAGAGGAAGAGAATCTCCG	Pancreas	M	75	8569192
14537_14538	1-2	GCACAGAGGAAGAGAATCTC	T	CG	CAAGAAAGGGGAGCCTCACC	Skin§			10498902
14538_14553	1-16	CACAGAGGAAGAGAATCTCC	T	GCAAGAAAGGGGAGCC	TCACCACGAGCTGCCCCCAG	Corpus Uteri	F	77	14976538
14551_14554	3-4	AATCTCCGCAAGAAAGGGGA	ACC	GCCT	CACCACGAGCTGCCCCCAGG	Breast	F		21217967
14555_14578	1-24	TCCGCAAGAAAGGGGAGCCT	T	CACCACGAGCTGCCCCCAGGGAGC	ACTAAGCGAGgtaagcaagc	Ovary	F		10682669
14555_14578	2-24	TCCGCAAGAAAGGGGAGCCT	GT	CACCACGAGCTGCCCCCAGGGAGC	ACTAAGCGAGgtaagcaagc	Ovary	F		10682669
14557_14560	3-4	CGCAAGAAAGGGGAGCCTCA	TTA	CCAC	GAGCTGCCCCCAGGGAGCAC	Skin§	F	28	21090798
14571_14575	3-5	GCCTCACCACGAGCTGCCCC	AGG	CAGGG	AGCACTAAGCGAGgtaagca	Adrenal Gland			11454518
14726_14729	3-4	TCCCCAGCCAAAGAAGAAAC	GAC	CACT	GGATGGAGAATATTTCACCC	Mouth			10225439
14755_14757	1-3	GAGAATATTTCACCCTTCAG	c	gta	ctaagtcttgggacctctta	Pancreas	F	64	8569192
Total	**66**								

*Range of deleted nucleotides. Numbering according to GenBank genomic sequence accession X54156.1.
†Inserted sequence was extracted from the primary publication, but was not available for three of the microindels (inserted sequence shown as N's).
‡Upper case letters indicate exon sequence; lower case letters indicate intron sequence.
§Excludes Skin of Vulva, Skin of Penis, Skin of Scrotum.
¶We note that an alternate interpretation of the 8-11 microindel at nucleotide 13145 is as a rare doublet mutation composed of two rare types of mutation: an insertion that does not repeat the adjacent base separated by six nucleotides from a 1-5 microindel.

SI Table 5B. Germline *TP53* Microindels

Nucleotide Range*	Ins-Del	5' Flanking Sequence† (20 bp)	Sequence Inserted	Sequence Deleted	3' Flanking Sequence† (20 bp)	Phenotype	F/M	Age	PubMed ID
12246_12256	5-11	AAACCTACCAGGGCAGCTAC	ATTCA	GGTTTCCGTCT	GGGCTTCTTGCATTCTGGGA	Li-Fraumeni Syndrome	M		8118819
13477_13479	5-3	gggaggaggggttaagggtg	agtta	gtt	gtcagtggccctccgggtga	Putative Neutral Variant‡	M		9180930
13959_13964	7-6	tgcttgccacaggtctcccc	cagagcc	aaggcg	cactggcctcatcttgggcc	Putative Neutral Variant§		39	1686725
Total	**3**								

*Range of deleted nucleotides. Numbering according to GenBank genomic sequence accession X54156.1.
†Upper case letters indicate exon sequence; lower case letters indicate intron sequence.
‡Putative neutral variant at IVS6+45 reported in two individuals.
§Putative neutral variant at IVS6-41.
‡,§Normal tissue was not analyzed. Does not obviously affect splicing.

SI Table 6A. Reported *TP53* Substitutions at 5' and 3' Deleted Nucleotides of Somatic *TP53* 1-M Microindels

Nucleotide Range*	Ins-Del	Reported *TP53* Substitutions at 5' deleted nucleotide (instances, percentile)†,‡	Sequence Inserted	Sequence Deleted‡	Reported *TP53* Substitutions at 3' deleted nucleotide (instances, percentile)†,‡	PubMed ID
12139_12140	1-2	12139G>T(1, 32%)	T	GC		1324794
12268_12287	1-20		C	ATTCTGGGACAGCCAAGTCT		12434417
12288_12302	1-15	12288G>T(1, 32%)	T	GTGACTTGCACGgtc		7862445
12294_12300	1-7	12294T>A(1, 32%)	A	TGCACGg	12300g>A(4, 65%)	8611423
13057_13061	1-5	13057C>A(6, 74%)	A	CTCCC		10761705
13082_13083	1-2	13082T>C(8, 79%)	C	TG	13083G>C(5, 71%)	12203794
13093_13100	1-8	13093C>G(2, 48%)	G	CAAGACCT	13100T>G(1, 32%)	9550561
13116_13117	1-2	13116G>A(43, 96%)*	A	GG	13117G>A(44, 97%)*	14697642
13122_13123	1-2		G	AT	13123T>G(2, 48%)	10353731
13132_13133	1-2	13132C>T(10, 83%)	T	CC	13133C>T(27, 94%)	11668523
13134_13135	1-2	13134C>T(67, 98%)*	T	CG	13135G>T(1, 32%)	7882279
13209_13217	1-9	13209C>A(2, 48%)	A	CCCACCATG	13217G>A(9, 82%)	21199244
13333_13368	1-36	13333T>C(1, 32%)	C	TCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGT	13368T>C(3, 58%)	9761125
13385_13387	1-3	13385A>T(11, 85%)	T	AGA	13387A>T(1, 32%)	9224526
13397_13404	1-8	13397C>A(1, 32%)	A	CGACATAG	13404G>A(15, 90%)	9516972
14011_14026	1-16	14011C>T(3, 58%)	T	CTGTACCACCATCCAC	14026C>T(6, 74%)	7768632
14013_14014	1-2	14013G>A(3, 58%)	A	GT	14014T>A(5, 71%)	10225439
14029_14032	1-4	14029C>G(1, 32%)	G	CAAC		10212000
14042_14045	1-4	14042A>T(8, 79%)	T	AACA	14045A>T(5, 71%)	1959992
14042_14045	1-4	14042A>T(8, 79%)	T	AACA	14045A>T(5, 71%)	15288479
14045_14054	1-10	14045A>G(19, 92%)	G	AGTTCCTGCA	14054A>G(2, 48%)	11733960
14055_14068	1-14	14055T>A(4, 65%)	A	TGGGCGGCATGAAC	14068C>A(2, 48%)	11051239
14066_14072	1-7	14066A>G(7, 77%)	G	AACCGGA	14072A>G(39, 96%)*	21278281
14072_14073	1-2	14072A>T(32, 95%)*	T	AG	14073G>T(57, 97%)*	12921629
14453_14475	1-23	14453G>A(6, 74%)	A	GGTAATCTACTGGGACGGAACAG	14475G>A(8, 79%)	8934544
14462_14476	1-15	14462C>A(5, 71%)	A	CTGGGACGGAACAGC	14476C>A(1, 32%)	7767998
14511_14515	1-5	14511A>C(3, 58%)	C	ACCGG	14515G>C(1, 32%)	21225878
14537_14538	1-2	14537C>T(8, 79%)	T	CG	14538G>T(8, 79%)	10498902
14538_14553	1-16	14538G>T(8, 79%)	T	GCAAGAAAGGGGAGCC	14553C>T(7, 77%)	14976538
14555_14578	1-24	14555C>T(9, 82%)	T	CACCACGAGCTGCCCCCAGGGAGC		10682669
14755_14757	1-3	14755g>c(1, 32%)	c	gta		8569192
Total	**31**					

*Range of deleted nucleotides. Numbering according to GenBank genomic sequence accession X54156.1.

†Reported instances in the IARC *TP53* Mutation Database version 10 [Olivier et al., 2002] of substitutions from the deleted nucleotide to the inserted nucleotide. Percentages are the substitution percentiles, the percentage of substitutions with the same or fewer instances recorded in the database. Substitutions at or above the 95th percentile are marked with an asterisk.

‡Upper case letters indicate exon sequence; lower case letters indicate intron sequence.

SI Table 6B. Reported *TP53* Substitutions at 5' and 3' Inserted Nucleotides of Somatic *TP53* N-1 Microindels

Nucleotide Range*	Ins-Del	Reported *TP53* Substitutions at 5' deleted nucleotide (instances, percentile)†,‡	Sequence Inserted	Sequence Deleted‡	Reported *TP53* Substitutions at 3' deleted nucleotide (instances, percentile)†,‡	PubMed ID
12286	2-1		GA	C		9735666
13074	2-1	13074A>T(9, 82%)	TG	A	13074A>G(46, 97%)	10225439
13431	3-1	13431A>T(1, 32%)	TCT	A	13431A>T(1, 32%)	15161705
14105	2-1	14105T>G(3, 58%)	GC	T	14105T>C(2, 48%)	7633433
14487	2-1	14487G>C(30, 95%)	CC	G	14487G>C(30, 95%)	1423304
14514	3-1	14514G>C(15, 90%)	CCC	G	14514G>C(15, 90%)	8289498
Total	**6**					

*Range of deleted nucleotides. Numbering according to GenBank genomic sequence accession X54156.1.
†Reported instances in the IARC *TP53* Mutation Database version 10 [Olivier et al., 2002] of substitutions from the deleted nucleotide to the inserted nucleotide. Percentages are the substitution percentiles, the percentage of substitutions with the same or fewer instances recorded in the database.
‡Upper case letters indicate exon sequence; lower case letters indicate intron sequence.

SI Table 7. *TP53* Substitutions Categorized by Observed Frequency in the IARC *TP53* Mutation Database*

#Instances in IARC *TP53* Mutation Database	Hypothesized Substitutions at 5' and 3' Deleted Nucleotides in 1-M Microindels†	Hypothesized Substitutions at 5' and 3' Inserted Nucleotides in N-1 Microindels‡	Substitutions in IARC *TP53* Mutation Database	Uniform Distribution over Unique Substitutions in IARC *TP53* Mutation Database
1	12 (22.6%)	2 (20.0%)	597 (3.2%)	5977 (32.1%)
2-5	16 (30.2%)	2 (20.0%)	2244 (12.0%)	7249 (38.9%)
6-10	15 (28.3%)	1 (10.0%)	1790 (9.6%)	2333 (12.5%)
11-100	10 (18.9%)	5 (50.0%)	7565 (40.6%)	2874 (15.4%)
>100	0 (0.0%)	0 (0.0%)	6447 (34.6%)	210 (1.1%)
Total	**53**	**10**	**18643**	**18643**

*IARC *TP53* Mutation Database version 10 [Olivier et al., 2002].
†Substitions based on an alternative model of a deletion of one less nucleotide (either on the 5' or 3' end) with substitution of that nucleotide (instead of the insertion).
‡Substitions based on an alternative model of a insertion of one less nucleotide (either on the 5' or 3' end) with substitution of that nucleotide (instead of the deletion).

SI Table 8. EGFR Recurroids*

Nucleotide Range†	Ins-Del	5' Flanking Sequence (20 bp)	Sequence Inserted	Sequence Deleted	3' Flanking Sequence (20 bp)	Database ID*	Recurroid Site Type	PubMed ID
2235_2252	3-18	AAAATTCCCGTCGCTATCAA	AAT	GGAATTAAGAGAAGCAAC	ATCTCCGAAAGCCAACAAGG	2248	Insertion Identical	16052218
2235_2255	3-21	AAAATTCCCGTCGCTATCAA	AAT	GGAATTAAGAGAAGCAACATC	TCCGAAAGCCAACAAGGAAA	1856		15741570
2236_2248	4-13	AAATTCCCGTCGCTATCAAG	AGAC	GAATTAAGAGAAG	CAACATCTCCGAAAGCCAAC	85	Deletion Identical	15604253
2236_2248	4-13	AAATTCCCGTCGCTATCAAG	CAAC	GAATTAAGAGAAG	CAACATCTCCGAAAGCCAAC	2341		15899142
2237_2252	1-16	AATTCCCGTCGCTATCAAGG	T	AATTAAGAGAAGCAAC	ATCTCCGAAAGCCAACAAGG	1857	Hybrid Site: 2 variants of Insertion Identical	15741570
2237_2255	1-19	AATTCCCGTCGCTATCAAGG	T	AATTAAGAGAAGCAACATC	TCCGAAAGCCAACAAGGAAA	25		15118125
2245_2252	1-8	TCGCTATCAAGGAATTAAGA	T	GAAGCAAC	ATCTCCGAAAGCCAACAAGG	33		15329413
2237_2253	2-17	AATTCCCGTCGCTATCAAGG	TC	AATTAAGAGAAGCAACA	TCTCCGAAAGCCAACAAGGA	2801		16152581
2237_2256	2-20	AATTCCCGTCGCTATCAAGG	TC	AATTAAGAGAAGCAACATCT	CCGAAAGCCAACAAGGAAAT	3165		16533793
2239_2248	1-10	TTCCCGTCGCTATCAAGGAA	C	TTAAGAGAAG	CAACATCTCCGAAAGCCAAC	22	Hybrid Site: Deletion Identical plus two variants of Insertion Identical	15118125
2239_2251	1-13	TTCCCGTCGCTATCAAGGAA	C	TTAAGAGAAGCAA	CATCTCCGAAAGCCAACAAG	83		15604253
2241_2248	1-8	CCCGTCGCTATCAAGGAATT	C	AAGAGAAG	CAACATCTCCGAAAGCCAAC	2246		16052218
2241_2251	1-11	CCCGTCGCTATCAAGGAATT	C	AAGAGAAGCAA	TCTCCGAAAGCCAACAAGGA	2249		16052218
2239_2252	2-14	TTCCCGTCGCTATCAAGGAA	CA	TTAAGAGAAGCAAC	ATCTCCGAAAGCCAACAAGG	89		15604253
2239_2258	2-20	TTCCCGTCGCTATCAAGGAA	CA	TTAAGAGAAGCAACATCTCC	GAAAGCCAACAAGGAAATCC	84		15604253
2239_2252	5-14	TTCCCGTCGCTATCAAGGAA	CCAAT	TTAAGAGAAGCAAC	ATCTCCGAAAGCCAACAAGG	2306		15958609
2252_2276	1-25	CAAGGAATTAAGAGAAGCAA	A	CATCTCCGAAAGCCAACAAGGAAAT	CCTCGATGAAGCCTACGTGA	2004	Deletion Identical	15815931
2252_2276	1-25	CAAGGAATTAAGAGAAGCAA	G	CATCTCCGAAAGCCAACAAGGAAAT	CCTCGATGAAGCCTACGTGA	104		15604253

*Data from the Epidermal Growth Factor Receptor (EGFR) Mutation Database (http://www.cityofhope.org/cmdl/egfr_db).

†Range of deleted nucleotides. Numbering according to GenBank cDNA sequence accession NM_005228.3 (coding sequence starts at base 247).

SI Table 9A. Somatic HPRT Microindels*

Nucleotide Range†	Ins-Del	5' Flanking Sequence (20 bp)	Sequence Inserted	Sequence Deleted	3' Flanking Sequence (20 bp)	Database ID*	Mutagen	PubMed ID
13163_13176	8-14	ATGAACCAGGTTATGACCTT	AGGAAGAA	GATTTATTTTGCAT	ACCTAATCATTATGCTGAGG	978	Smoker	1394847
13222_13224	1-3	GAAAGGGTGTTTATTCCTCA	A	TGG	ACTAATTATGGACAGgtaag	1140	Spontaneous	8513767
15063_15064	1-2	TTTGCTGACCTGCTGGATTA	T	CA	TCAAAGCACTGAATAGAAAT	327	UV (G1)	2005888
33294_33306	2-13	tttgaaagGATATAATTGAC	TC	ACTGGCAAAACAA	TGCAGACTTTGCTTTCCTTG	1351	Spontaneous	8404873
33310_33317	3-8	TGACACTGGCAAAACAATGC	CGA	AGACTTTG	CTTTCCTTGGTCAGGCAGTA	1352	Spontaneous	8404873
33331_33332	1-2	GACTTTGCTTTCCTTGGTCA	T	GG	CAGTATAATCCAAAGATGGT	879	BPDE (S)	1902394
38166_38167	1-2	acattttgtaattaacagCT	C	TG	CTGGTGAAAAGGACCCCACG	1260	MNNG (G1/S)	8504428
38444	2-1	CCCTTGACTATAATGAATAC	CC	T	TCAGGGATTTGAATgtaagt	954	Spontaneous	1394847
38448_38449	1-2	TGACTATAATGAATACTTCA	A	GG	GATTTGAATgtaagtaattg	284	UV (S)	2005888
38452_38453	1-2	TATAATGAATACTTCAGGGA	C	TT	TGAATgtaagtaattgcttc	184	UV (G1)	2005888
Total	**10**							

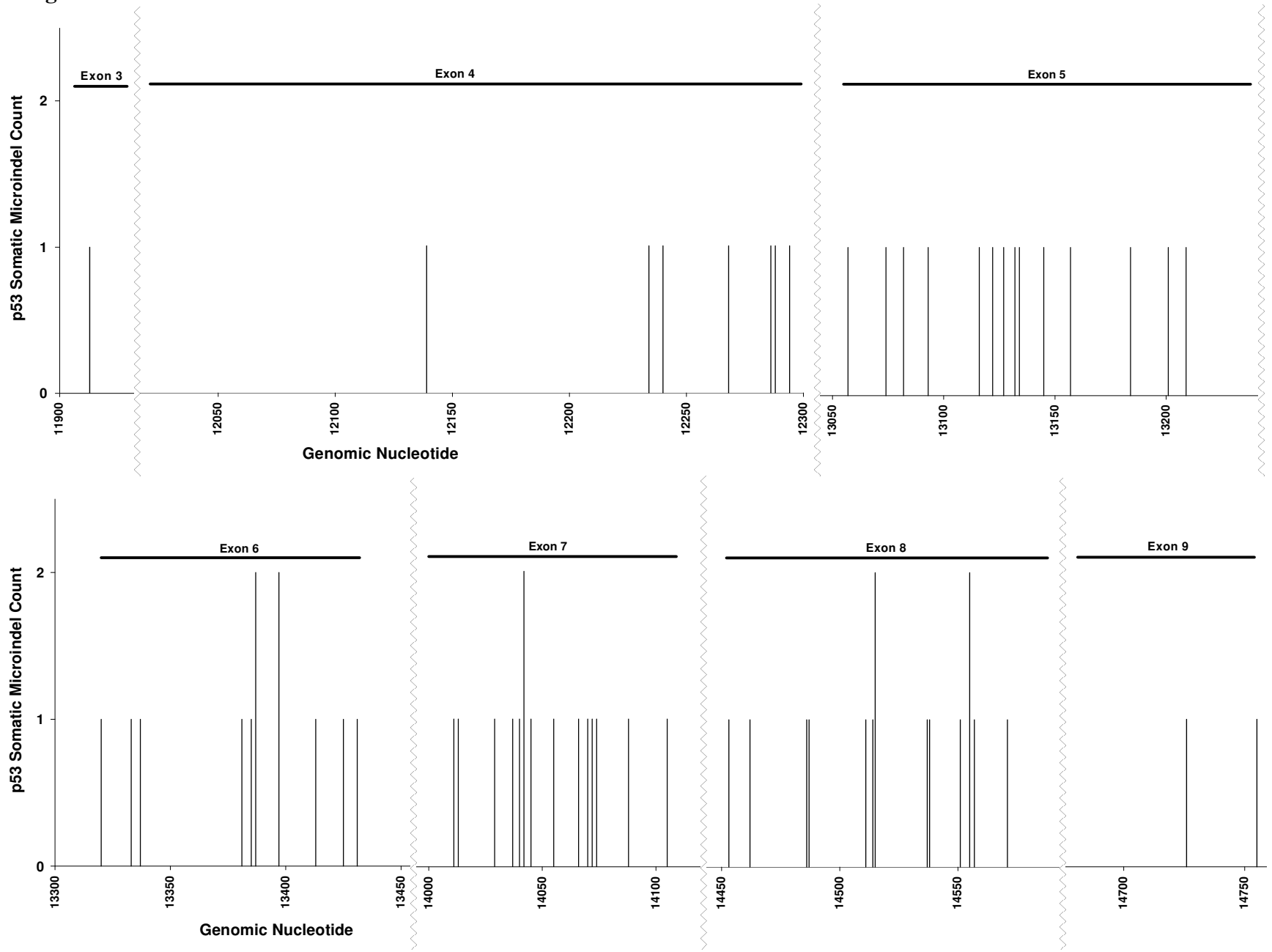
*Data from the Human HPRT Mutation Database (http://www.ibiblio.org/dnam/mainpage.html).
†Range of deleted nucleotides. Numbering according to GenBank genomic sequence accession NC_000023.9 region 133421923 to 133462362.

SI Table 9B. Germline HPRT Microindels*

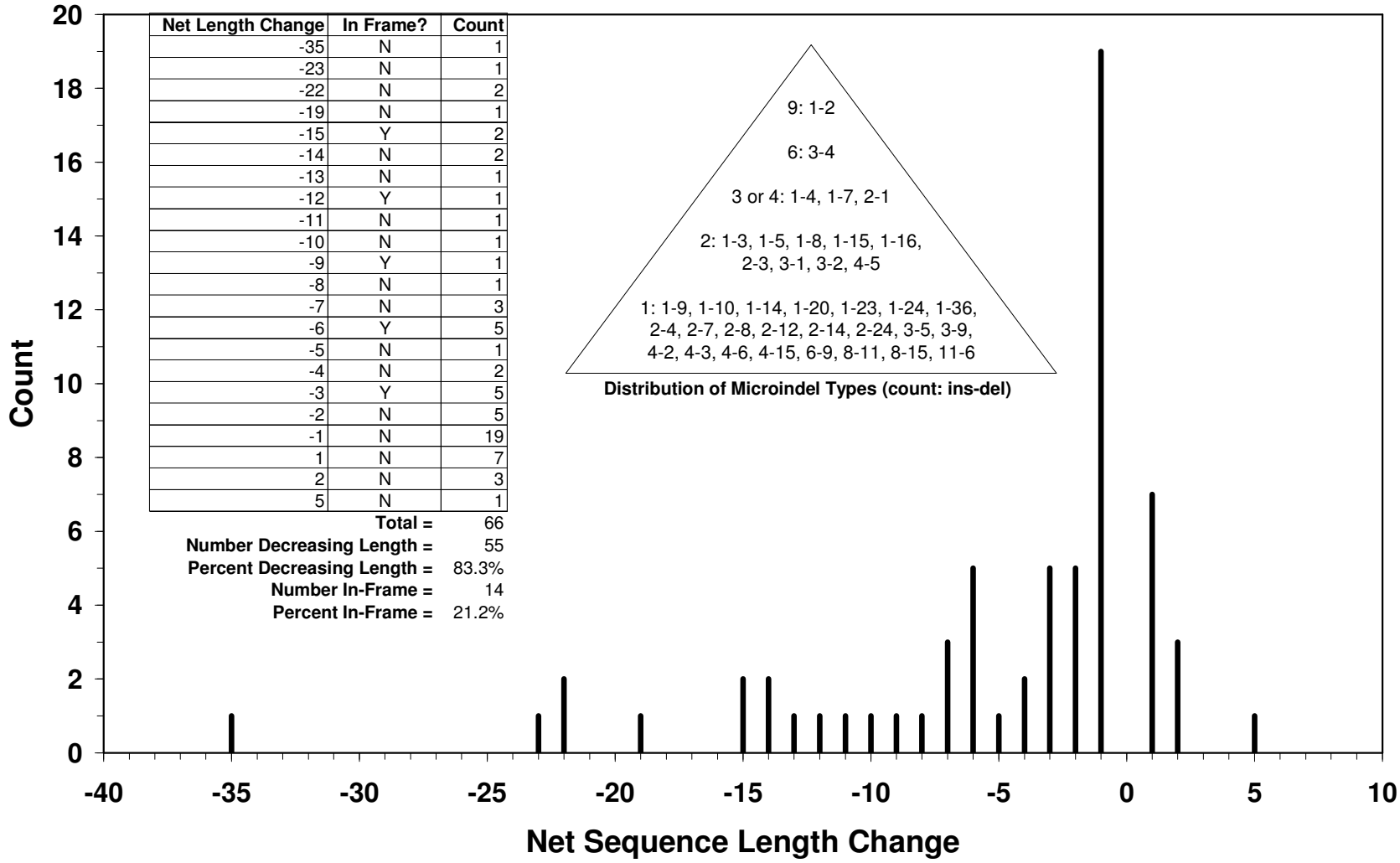
Nucleotide Range†	Ins-Del	5' Flanking Sequence (20 bp)	Sequence Inserted	Sequence Deleted	3' Flanking Sequence (20 bp)	Database ID*	Phenotype	PubMed ID
33307_33311	6-5	AATTGACACTGGCAAAACAA	AGCAAA	TGCAG	ACTTTGCTTTCCTTGGTCAG	140	Lesch-Nyhan	2928313
Total	**1**							

*Data from the Human HPRT Mutation Database (http://www.ibiblio.org/dnam/mainpage.html).
†Range of deleted nucleotides. Numbering according to GenBank genomic sequence accession NC_000023.9 region 133421923 to 133462362.

SI Figure 4


Exon 3
Exon 4
Exon 5
p53 Somatic Microindel Count
0
1
2
11900
12050
12100
12150
12200
12250
12300
13050
13100
13150
13200
Genomic Nucleotide
Exon 6
Exon 7
Exon 8
Exon 9
p53 Somatic Microindel Count
0
1
2
13300
13350
13400
13450
14000
14050
14100
14450
14500
14550
14700
14750
Genomic Nucleotide

SI Figure 5



Net Length Change	In Frame?	Count
-35	N	1
-23	N	1
-22	N	2
-19	N	1
-15	Y	2
-14	N	2
-13	N	1
-12	Y	1
-11	N	1
-10	N	1
-9	Y	1
-8	N	1
-7	N	3
-6	Y	5
-5	N	1
-4	N	2
-3	Y	5
-2	N	5
-1	N	19
1	N	7
2	N	3
5	N	1
Total =		66
Number Decreasing Length =		55
Percent Decreasing Length =		83.3%
Number In-Frame =		14
Percent In-Frame =		21.2%

9: 1-2

6: 3-4

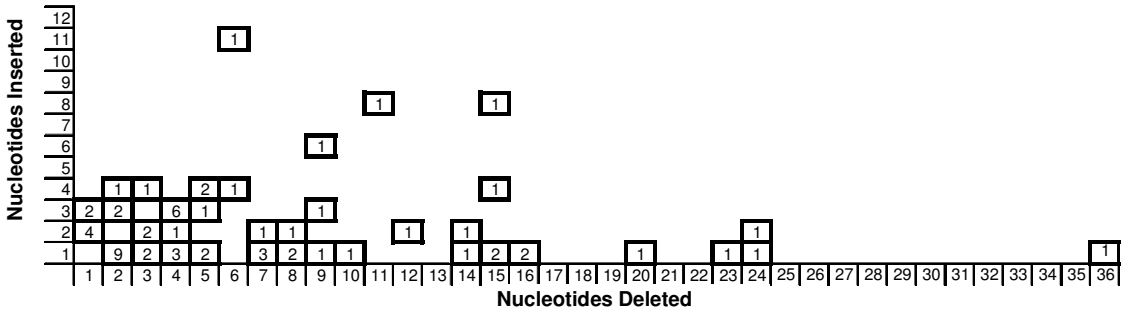
3 or 4: 1-4, 1-7, 2-1

2: 1-3, 1-5, 1-8, 1-15, 1-16, 2-3, 3-1, 3-2, 4-5

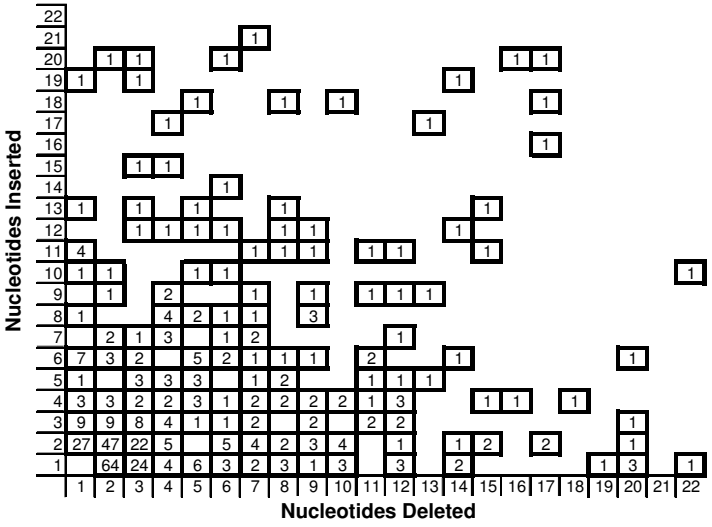
1: 1-9, 1-10, 1-14, 1-20, 1-23, 1-24, 1-36, 2-4, 2-7, 2-8, 2-12, 2-14, 2-24, 3-5, 3-9, 4-2, 4-3, 4-6, 4-15, 6-9, 8-11, 8-15, 11-6

Distribution of Microindel Types (count: ins-del)

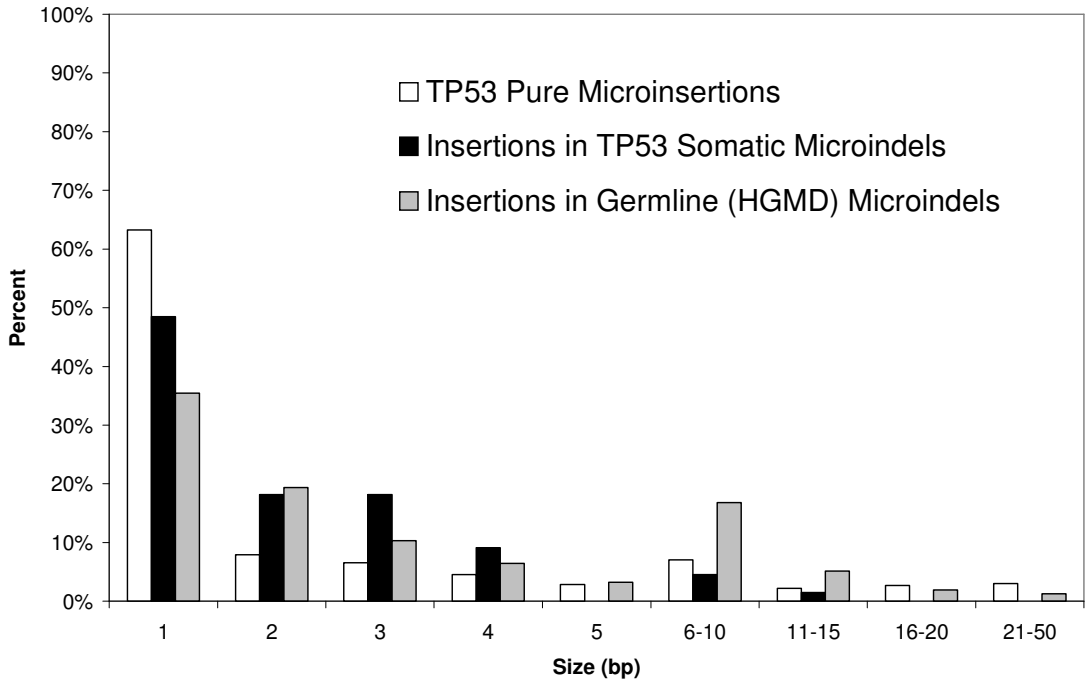
SI Figure 6A



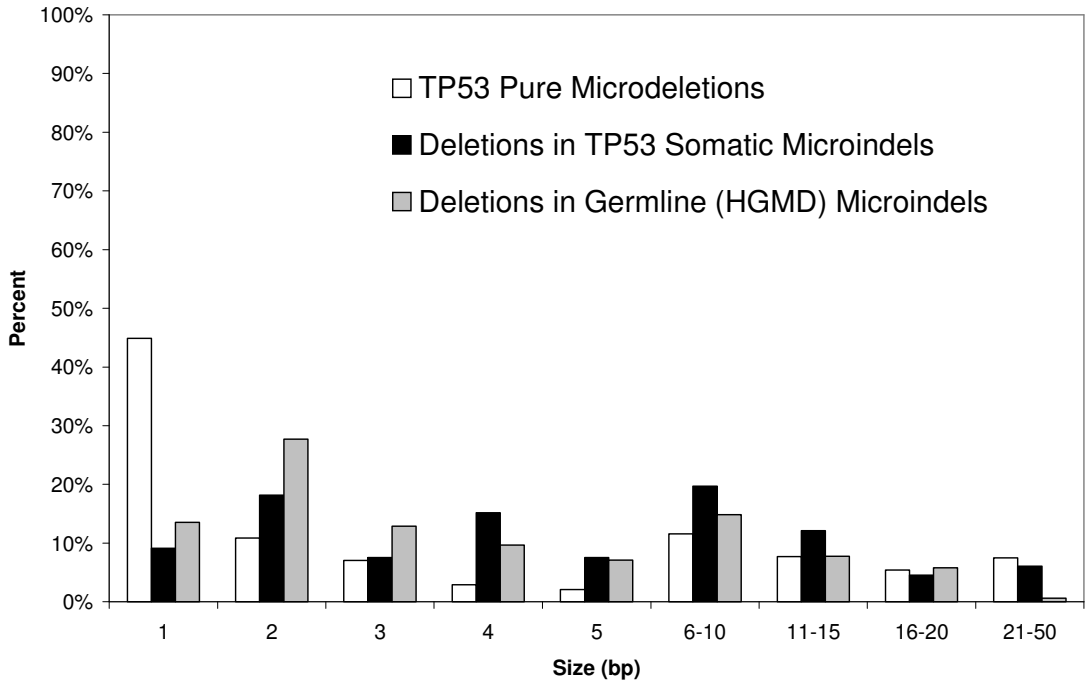
SI Figure 6B



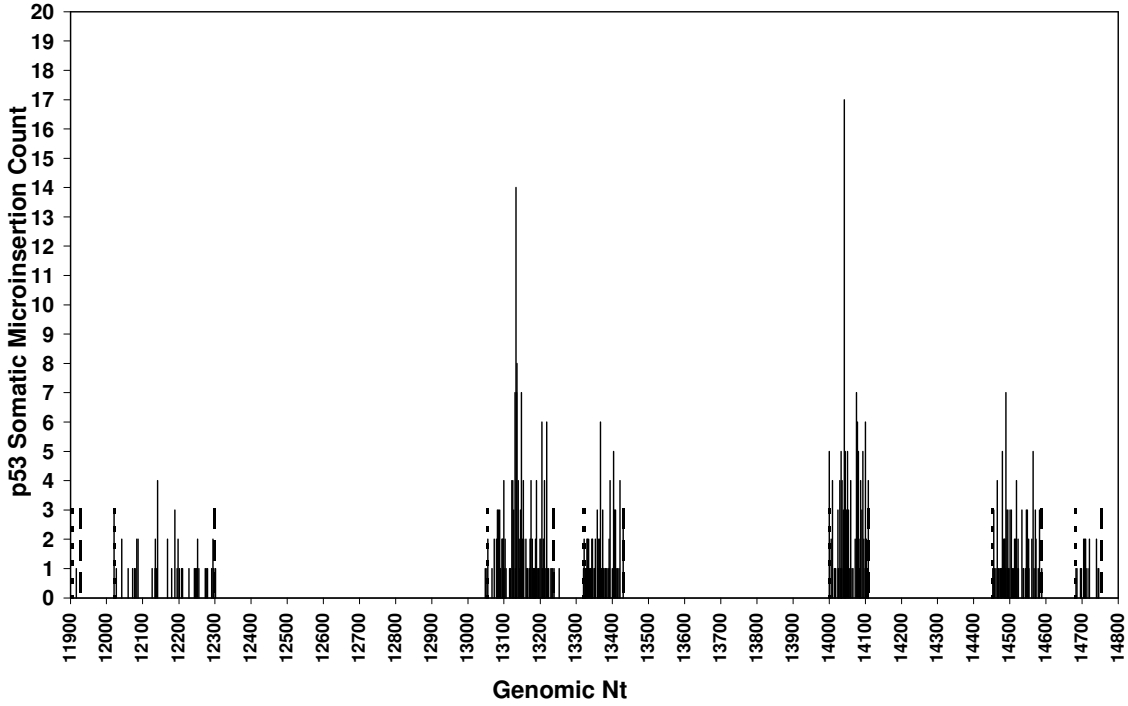
SI Figure 7A



SI Figure 7B

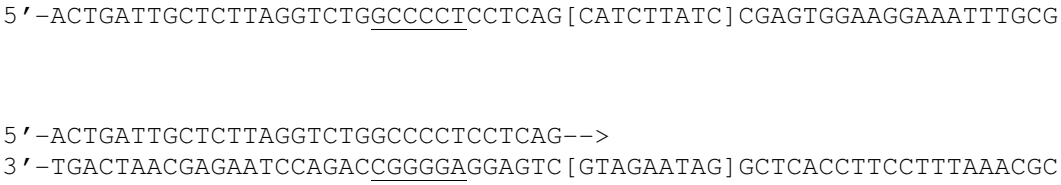


SI Figure 8

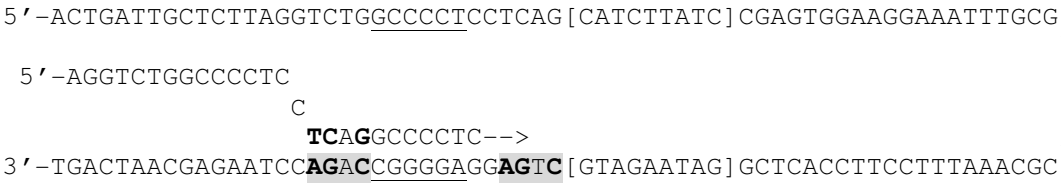

p53 Somatic Microinsertion Count
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
11900
12000
12100
12200
12300
12400
12500
12600
12700
12800
12900
13000
13100
13200
13300
13400
13500
13600
13700
13800
13900
14000
14100
14200
14300
14400
14500
14600
14700
14800
Genomic Nt

SI Figure 9 panel A

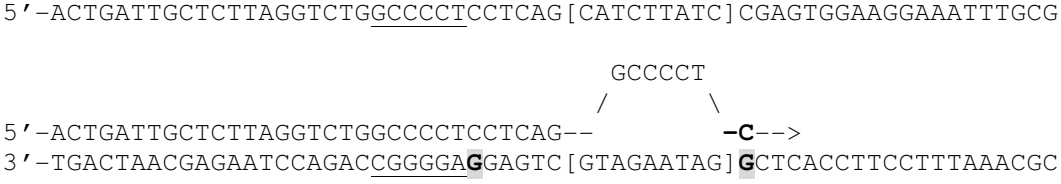
Step 1: Leading strand synthesis up to the deleted region (in brackets).



Step 2: Dissociation from leading strand, backward slippage, re-association with leading strand before insertion template at highlighted repeat region (matching bases in bold), and leading strand synthesis through insertion template (underlined) plus one additional base.

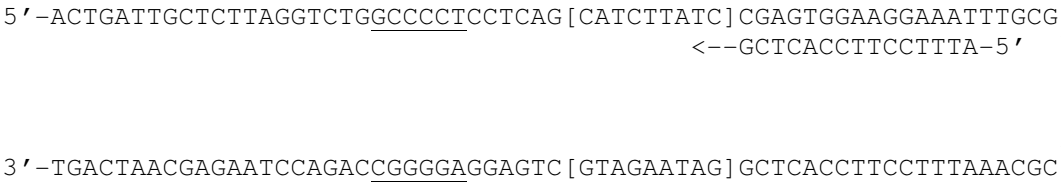


Step 3: Dissociation from leading strand, forward slippage, re-association with leading strand beyond deleted region at highlighted "repeat" region. At this point, normal synthesis resumes.

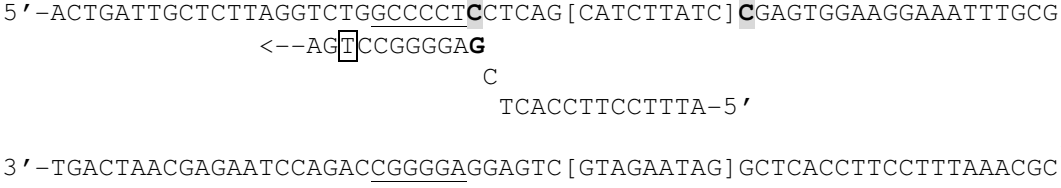


SI Figure 9 panel B

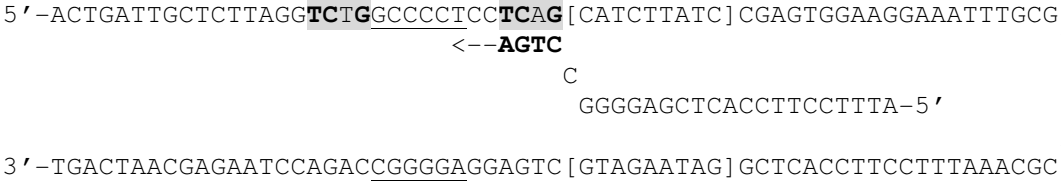
Step 1: Lagging strand synthesis up to the deleted region (in brackets).



Step 2: Dissociation from lagging strand, forward slippage, re-association with lagging strand before insertion template at highlighted "repeat" region, and lagging strand synthesis through insertion template (underlined) plus four additional bases. One of these additional bases would have had to be synthesized erroneously as a T (boxed).

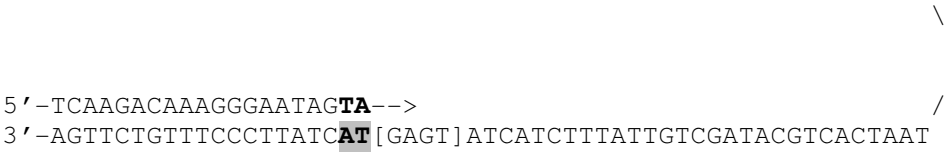


Step 3: Dissociation from lagging strand, backward slippage, re-association with lagging strand after deleted region at highlighted repeat region. At this point, normal synthesis resumes.

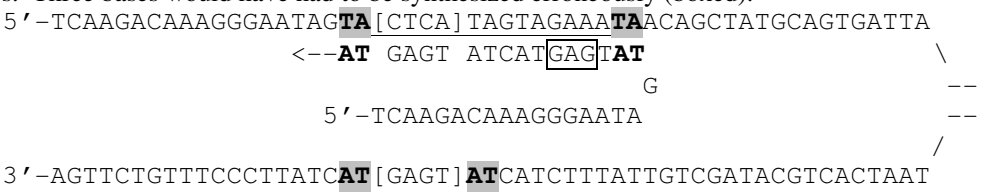


SI Figure 9 panel C

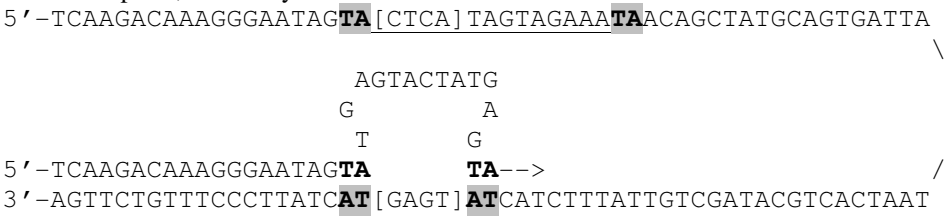
Step 1: Leading strand synthesis up through highlighted repeat region (matching bases in bold)..



Step 2: Dissociation from leading strand, association with lagging strand before insertion template at highlighted repeat region (matching bases in bold), and lagging strand synthesis through insertion template (underlined) plus two additional bases. Three bases would have had to be synthesized erroneously (boxed).

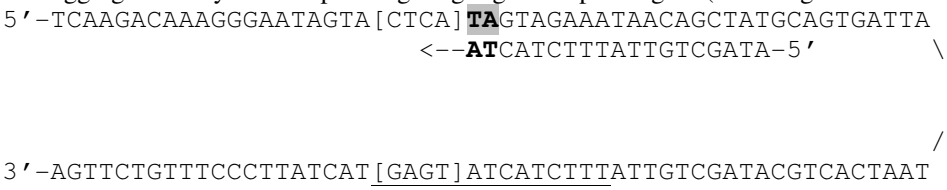


Step 3: Dissociation from lagging strand, re-association with leading strand beyond deleted region at highlighted "repeat" region. At this point, normal synthesis resumes.

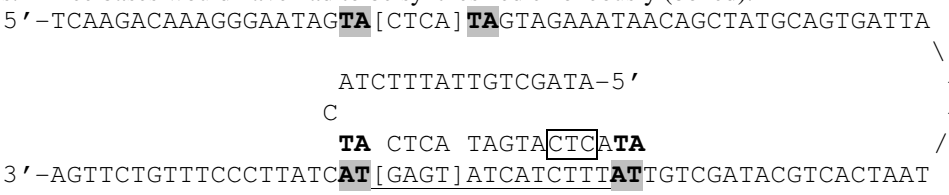


SI Figure 9 panel D

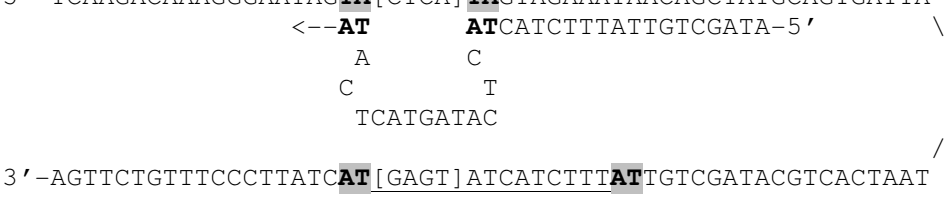
Step 1: Lagging strand synthesis up through highlighted repeat region (matching bases in bold).



Step 2: Dissociation from lagging strand, association with leading strand before insertion template at highlighted repeat region (matching bases in bold), and leading strand synthesis through insertion template (underlined) plus two additional bases. Three bases would have had to be synthesized erroneously (boxed).



Step 3: Dissociation from leading strand, re-association with lagging strand beyond deleted region at highlighted "repeat" region. At this point, normal synthesis resumes.



References for Supplemental Information

1. Halangoda,A., Still,J.G., Hill,K.A., and Sommer,S.S. (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.

2. Ball,E.V., Stenson,P.D., Abeysinghe,S.S., Krawczak,M., Cooper,D.N., and Chuzhanova,N.A. (2005) Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat.*, **26**, 205-213.

3. Ketterling,R.P., Sommer,S.S. (1994) Microdeletions in the factor IX gene: three recurrences associated with a quasipalindromic sequence. *Hum.Molec.Genet.*, **3**, 191-192.

4. Cooper,D.N., Krawczak,M. (1993) *Human Gene Mutation*. Bios Scientific Publishers, Oxford, England.

5. Lieber,M.R., Ma,Y., Pannicke,U., and Schwarz,K. (2003) Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol*, **4**, 712-720.

6. Honma,M., Sakuraba,M., Koizumi,T., Takashima,Y., Sakamoto,H., and Hayashi,M. (2007) Non-homologous end-joining for repairing I-SceI-induced DNA double strand breaks in human cells. *DNA Repair (Amst)*, **6**, 781-788.

7. Willers,H., Husson,J., Lee,L.W., Hubbe,P., Gazemeier,F., Powell,S.N., and Dahm-Daphi,J. (2006) Distinct mechanisms of nonhomologous end joining in the repair of site-directed chromosomal breaks with noncomplementary and complementary ends. *Radiat.Res*, **166**, 567-574.

8. Moore,J.K., Haber,J.E. (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol Cell Biol*, **16**, 2164-2173.

9. Aniukwu,J., Glickman,M.S., and Shuman,S. (2008) The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends. *Genes Dev*, **22**, 512-527.

10. Chen,J.M., Chuzhanova,N., Stenson,P.D., Ferec,C., and Cooper,D.N. (2005) Complex gene rearrangements caused by serial replication slippage. *Hum Mutat.*, **26**, 125-134.

11. Chen,J.M., Chuzhanova,N., Stenson,P.D., Ferec,C., and Cooper,D.N. (2005) Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. *Hum Mutat.*, **26**, 362-373.