Nucleic Acids Research, 2001, Vol. 29, No. 21 **4405–4413**

Hypermutability at a poly(A/T) tract in the human germline

Andrea L. Bacon, Malcolm G. Dunlop and Susan M. Farrington*

University of Edinburgh Department of Oncology and MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Received July 3, 2001; Revised and Accepted September 6, 2001

ABSTRACT

Poly(A/T) tracts are abundant simple sequence repeats (SSRs) within the human genome. They constitute part of the coding sequence of a variety of genes, encoding polylysine stretches that are important for protein function. Assessment of poly(A/T) tract stability is also used to identify microsatellite unstable colorectal cancers, which are characteristic of tumours defective in DNA mismatch repair. Despite their importance, little is known about the stability of poly(A/T) SSRs in the human germline. We have determined the stability of a paradigm poly(A/T) tract, BAT-40, by study of population allele frequencies, mutation frequency in families and mutation frequency in sperm DNA. We show that the locus is polymorphic, with a level of heterozygosity of 59.7%. Germline mutation was observed in 13 of 187 germline transmissions (7.0%) in 10 families suggesting BAT-40 is unstable in the germline. Further evidence for germline instability at BAT-40 was provided by small pool PCR analysis of matched blood and sperm DNA templates, revealing a significantly elevated frequency of mutation in the germline ($P < 0.001$). These findings provide insight into poly(A/T) tract stability in the germline. They also have relevance to the study of gene expression and to determination of microsatellite instability in tumours.

INTRODUCTION

Simple sequence repeats (SSRs) occur ubiquitously throughout the genome. Many are highly polymorphic, making them of particular importance to the study of evolution and the mapping of disease genes (1). SSRs such as poly(A/T) and $(CA)_n$ are routinely used in determining microsatellite instability (MSI) status in mismatch repair (MMR) deficient colorectal cancers (CRCs) (2–5). The proportion of markers that display mutational shifts in the tumour directs subsequent analysis for germline MMR mutations that lead to this MSI phenotype (6). Factors affecting the mutation frequency of microsatellites have been studied in MMR deficient tumours since analysis is rapid and material readily available (7,8). We have previously demonstrated that constitutional genotype at a given microsatellite locus influences the propensity for instability in the presence of defective MMR (7). Such studies highlight the need for a well-characterised panel of markers to be used for such assessments, in order for them to be employed with confidence.

It has been shown that mechanisms generating mutations in microsatellite unstable (MSI^+) tumours have relevance to understanding the evolution of such sequences in the germline (8). Investigation of determinants of germline mutation at SSR loci is laborious and frequently requires analysis of many hundreds of gametes in family studies (9,10). However, the development of small pool PCR (SP-PCR) techniques in studying germline stability at minisatellites has facilitated investigations of such mutations at other SSRs (11,12).

To date, many studies of germline mutation at SSR loci have focused on understanding sequence instability of the trinucleotide repeat disorders (13–16). In addition, investigations have been carried out on germline stability of dinucleotide repeat markers, including long CA stretches (10). However, there has been little investigation of the stability of mononucleotide tracts in the germline. This is surprising since poly(A/T) repeats are the most abundant simple repetitive sequence motif in the human genome (17) largely due to the poly(A/T) tails of scattered retrotransposed sequences such as long (LINEs) and short interspersed elements (SINEs) (17,18). Coding poly(A/T) sequence tracts have been identified with repeat lengths of up to 27 bp and within introns they may occur up to 70 bp in repeat length (17). Any process influencing the fidelity of replication at coding sequence mononucleotide tracts will clearly have important functional effects. The transforming growth factor beta receptor type 2 (*TGFBR2*) gene contains a $poly(A/T)_{10}$ tract in exon 3 that has been shown to be mutated in up to 90% of MSI^+ tumours, resulting in inactivation of the gene (19–21).

In view of the prevalence of poly(A/T) stretches and their functional relevance we were interested to gain insight into the inherent stability of such sequences. The microsatellite BAT-40 is a paradigm mononucleotide marker consisting of 40 adenine repeats located in intron 2 of the 3-beta-hydroxysteroid dehydrogenase (*3-beta-HSD*) gene on chromosome 11 (22). BAT-40 is highly sensitive to the effects of defective MMR, since it is susceptible to mutation in >95% MSI^+ tumours and thus is used routinely in the analysis of MSI (5,21,23,24). Previous studies have also demonstrated that BAT-40 exhibits significant

*To whom correspondence should be addressed at: Colon Cancer Genetics Group, MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. Tel: +44 131 467 8422; Fax: +44 131 343 2620; Email: susan.farrington@hgu.mrc.ac.uk

polymorphism within populations (25,26). Hence we hypothesised that BAT-40 and other long poly(A/T) repeats might be unstable in the germline.

We analysed germline stability at the BAT-40 locus as a paradigm poly(A/T) tract, in order to gain insight into the generation of new mutations at such sequences. Assessment of the degree of mutability at such a locus might have considerable relevance to the generation of mutations at that locus in MMR deficient tumours.

MATERIALS AND METHODS

DNA sample groups

Genotyping was carried out on the constitutional DNA of 102 unrelated Scottish individuals and 35 unrelated CEPH family members from the nine families listed below.

A Scottish Family, K-435, was used for pedigree analysis. Family relationships were confirmed by genotyping, using a panel of microsatellite markers (data not shown). This family was identified previously as being an HNPCC kindred with affected individuals displaying tumour instability as determined using a panel of microsatellite markers. Proband MD-473 was previously determined to be heterozygous at BAT-40 with two alleles differing by 12 bp in length thus making individual allele identification obvious. MD numbers represent our laboratory sample identification system. DNA samples were available from the peripheral blood leukocytes of 20 individuals from this family.

DNA from nine CEPH families (66, 1331, 1341, 1346, 1362, 1377, 1423, 13293 and 13294) was used for further pedigree analysis to provide a total of 176 germline transmissions for study.

SP-PCR was carried out on matched constitutional and germline DNA samples from two further unrelated individuals. MD-949 carries a germline mutation in the MMR gene, human *MLH1*, resulting in a deletion of exon 12 (codons 347–470). MD-c1 is a healthy control individual.

Preparation of constitutional DNA and BAT-40 genotyping

Constitutional DNA was extracted from blood using Nucleon II DNA extraction protocol (Scotlab Bioscience, Strathclyde).

BAT-40 alleles were amplified from DNA templates in triplicate using primers described previously (24). These primers amplify a 126 bp product containing the standard 40 A residues according to the genomic sequence of the *3-beta-HSD* gene (GenBank accession no. M38180) (22). We were not able to assess the number of adenine repeats in a given sized allele directly, since repeated attempts at sequencing across the BAT-40 locus were unsuccessful. Therefore repeat length is based on the theoretical predicted amplified sequence both in this and in the majority of other studies that report BAT-40 repeat length.

PCR reactions were performed in a final volume of 25 µl using the Expand High Fidelity PCR system (Boehringer Mannheim, Germany). Final reaction concentrations were 1× PCR buffer II, 0.2 mM dNTPs, 100 ng oligonucleotide primer, 100 ng DNA and 0.87 U of Expand high fidelity PCR mix. Amplifications were performed using an Omnigene PCR thermal Cycler (Hybaid) at 94°C for 3 min for 1 cycle; 94°C for 1 min, 55°C for 1 min, 72°C for 1 min for 35 cycles; 72°C for 5 min for 1 cycle. PCR products were size analysed using an ABI310 Automated Genetic Analyser, using Genescan software.

Sperm DNA preparation and SP-PCR

DNA was extracted from sperm as described by Jeffreys *et al.* (27). Pelleted semen was rinsed three times with 20 ml 1× SSC followed by six washes with 20 ml 1× SSC and 1% SDS to lyse any seminal leukocytes and epithelial cells. The residual sperm pellet was incubated in 1× SSC and 1 M 2-mercaptoethanol at room temperature for 5 min and the reduced sperm lysed by addition of SDS to 1%. Sperm DNA was collected after phenol extraction by ethanol precipitation and resuspended in Tris–EDTA pH 7.7. DNA concentration was calculated using a spectrophotometer to measure the optical density of DNA samples in triplicate and also by running samples against standards of known DNA quantity using gel electrophoresis.

SP-PCR and analysis of PCR products was performed as described (7). DNA from matched blood and sperm samples were serially diluted to a final concentration of 15–20 pg per PCR reaction. This results in an estimated five to six template BAT-40 alleles in total, per reaction (assuming 6 pg of DNA per diploid genome). Limiting dilutions were carried out and final DNA concentrations resulted in the detection of a product in ~30% of analyses corresponding to products that represent single DNA templates (11,28). PCR reactions were then performed as described above in 96 well plates using the Expand High Fidelity PCR System (Boehringer Mannheim, Germany). Avoidance of contamination was paramount when amplifying dilute DNA templates. Therefore all reactions were carried out in a Class-2 containment hood. All pipettes and plastics used in the preparation of the SP-PCR reactions were UV irradiated for 20 min in a Template Tamer (Oncor). Buffer solution and sterile water were opened under sterile conditions and also subjected to UV irradiation. On each plate, 16 wells were DNA free to provide negative controls and positive controls containing 100 ng of undiluted sample DNA was prepared in two wells on every plate to ensure reproducibility of ABI310 profiles between plates. Matched sperm and blood DNA samples and SP-PCRs were prepared simultaneously, using the same reagents to allow direct comparison. Amplifications were performed as above and SP-PCR products and positive controls were size analysed using ABI310 Automated Genetic Analyser and Genescan software. All 16 negative controls from every SP-PCR plate were also analysed and if a product was observed in any negative sample, the entire plate was discarded.

Determination of the origin of new alleles

In instances where the origin of 'new/mutant' alleles was inferred, this was done in the same manner as in published studies (e.g. 10). The origin of 'new/mutant' alleles is such that if there were two possibilities, the shortest mutational step was considered to be the actual one. For example, if one progenitor allele differed by one repeat and the other by two repeats, when compared to the mutant, a one step mutation was inferred. If two progenitor alleles exhibited the same difference when compared to the new/mutant one, the origin was declared ambiguous.

Statistical analysis

Statistical comparison of population allele size frequency at the BAT-40 locus was carried out using a Mann–Whitney U test on the Minitab (V.13) statistical package. Significance was taken at the 5% level.

For comparison of mutation levels between matched sperm and blood samples, the frequency of mutant alleles in each sample was expressed as the number of alleles that were mutant in length divided by the total number of alleles detected (normal and mutant). Accordingly the frequency of mutants was not the exact number of cells with alterations but represents the relative proportions of alleles. Statistical analyses were then performed using a chi-squared analysis on Minitab (V.13) statistical package, and significance taken at the 5% level.

RESULTS

The BAT-40 poly(A/T) locus is polymorphic

BAT-40 genotypes were defined in 104 unrelated Scottish individuals and 35 unrelated CEPH family members (Table 1). Representative ABI310 profiles are shown in Figure 1. PCR products displaying a single complex of peaks with a near normal distribution were counted as homozygous (Fig. 1A). Those with extra peaks were regarded as BAT-40 heterozygous (Fig. 1B–D). The allele traces of the BAT-40 mononucleotide marker are complex with 'stutter' peaks evident due to DNA polymerase slippage. However, 'bona fide' allele sizes are taken to be the predominant peak in each separate peak complex in accordance with previous studies (Fig. 1) (26,29). The predominant peak is that with the greatest peak area as indicated by the ABI310 genetic analyser software. Our own previous analysis of this locus by SP-PCR analysis of multiple single alleles in three individuals has also validated this method of allele sizing at the BAT-40 locus. The most predominant peaks as genotyped from constitutional DNA are detectable as individual alleles by SP-PCR (Fig. 2). PCR error is evident when amplifying BAT-40, by nature of the stutter bands that are observed (Figs 1 and 2). However, reproducibility of the prominent peaks in a given individual assures confidence in the sizes given (see Materials and Methods). Where the genotype of an individual could not be confirmed by reproducibly detecting the same peaks in the allele trace, these individuals were discarded from further analysis (three cases).

The distribution of BAT-40 heterozygous genotypes also indicates that amplification and detection of two BAT-40 alleles of different sizes is due to the difference in genotype and not technical artefact (Table 1).

Allele frequency and distribution for both the Scottish and the CEPH cohorts are shown in Figure 3. Of the 139 samples analysed, a total of 83 demonstrated heterozygosity at the BAT-40 locus (59.7%). Levels of heterozygosity were similar between the cohorts, 58.7% (61/104) in the Scottish population and 62.9% (22/35) in the CEPH cohort. As expected for a highly polymorphic marker the overall distribution of allele sizes was not significantly different between these two populations ($P = 0.056$) (Fig. 3). Allelic size variation was from –15 to +11 as compared to the most frequent allele. Taking into account variation in the cohorts studied, these data are in line with a previous study reporting polymorphism at this

Table 1. Levels of heterozygosity at the BAT-40 locus

	Allele set (bp)	Frequency detected in Scottish cohort (% $n = 104$)	Frequency detected in CEPH cohort (% $n = 35$)
Heterozygotes	108/119	0	2.9
	108/123	1.9	0
	109/124	0	2.9
	108/125	1	0
	109/122	1	0
	111/123	0	2.9
	111/124	1	0
	112/124	1	0
	118/119	1	0
	118/121	2.9	0
	118/123	1.9	0
	118/124	1	0
	119/120	1	0
	119/121	1.9	0
	119/122	6.7	5.7
	119/123	1	2.9
	119/124	1.9	0
	119/126	1	0
	119/127	1	2.9
	120/121	1	0
	120/122	5.7	11.4
	120/123	9.6	5.7
	120/124	1.9	5.7
	120/125	1.0	0
	121/122	1.0	0
	121/123	3.8	8.6
	121/124	0	2.9
	122/123	0	2.9
	122/124	1.9	0
	122/125	1.0	0
	123/124	1.0	0
	123/125	1.9	0
	123/128	1	0
	124/125	0	2.9
	124/127	0	2.9
	124/134	1	0
Homozygotes	117/117	1.9	0
	118/118	2.9	0
	119/119	4.8	0
	120/120	3.8	8.6
	121/121	3.8	0
	122/122	4.8	0
	123/123	16.3	22.9
	124/124	3.8	5.7

Frequency of individual BAT-40 genotypes in the Scottish and CEPH cohorts analysed are indicated. Allele sets are given in base pairs and are grouped according to whether or not the genotype is heterozygous.

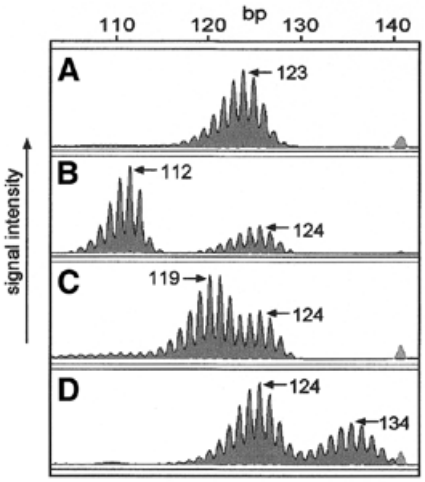

Figure 1. ABI310 traces of BAT-40 poly(A/T) PCR products show the polymorphic nature of this microsatellite marker in samples from a Scottish cohort. Blood DNA shows a single complex of peaks, the highest being 123 bp (**A**). (**B**–**D**) Heterozygosity at BAT-40 as illustrated by separate peak complexes.

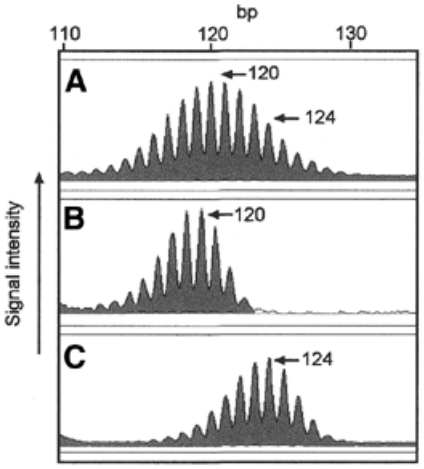

Figure 2. Genotyping of the BAT-40 locus in constitutional DNA is validated by SP-PCR analysis. (**A**) Constitutional BAT-40 allele sizes in cell line lbl-1261 are revealed as 120 and 124 bp from analysis of undiluted DNA. (**B** and **C**) In SP-PCR analysis of DNA from the same cell line, individual alleles of 120 and 124 bp are easily determined and confirm the genotype revealed from the undiluted DNA. (Data taken from ref. 33).

mononucleotide marker for a large number of different alleles, and are suggestive of the frequent generation of new alleles at this locus (25). The most frequent allele in both cohorts in this study (123 bp) corresponds to a 37-adenine tract as calculated from the genomic sequence (GenBank accession no. M38180). However, the BAT-40 heterozygosity reported here differs in both frequency and size distribution to that observed in a Japanese study despite the use of similar methodology (29).

Germline hypermutability at BAT-40 in pedigree analysis

Since BAT-40 displays high levels of polymorphism in populations and has been previously demonstrated to be extremely

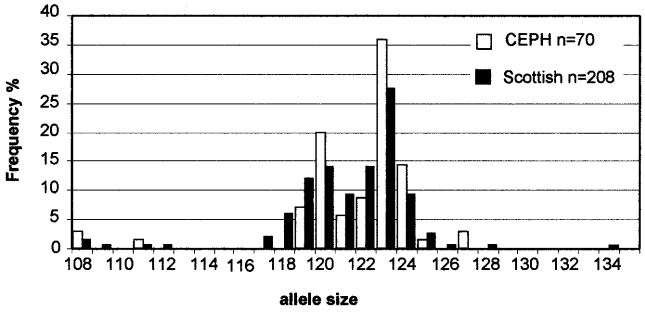

Figure 3. Comparisons of BAT-40 allele frequencies between Scottish and CEPH populations. The sizes of each allele are given in base pairs. The estimated size of the standard BAT-40 allele with 40 adenine repeats is 126 bp as calculated from the predicted PCR product size (GenBank accession no. M38180). There is no statistical difference in the distribution of alleles within the two cohorts ($P = 0.073$).

susceptible to instability in MMR deficient tumours (21), we reasoned that BAT-40 may be inherently unstable and that germline mutations might be detectable in family studies. A Scottish family, K-435, was chosen to determine whether a high level of germline mutation occurs within this population. Proband MD-473 had previously been identified as heterozygous at BAT-40, with two distinct sized alleles at this locus (112/124). Analysing the meiotic stability of BAT-40 alleles that are easily distinguished by size allows for the most accurate assessment of individual allele stability at a complex locus such as BAT-40. DNA from 20 available individuals from K-435 was genotyped at the BAT-40 locus (Fig. 4A). There were 11 germline transmissions available for study. Within the family there was striking evidence of a germline mutation in the allele transmission from MD-1303 to MD-449. MD-1303 is heterozygous for BAT-40 with an allele set of 120/124 but her daughter (MD-449) is homozygous for two 112 BAT-40 alleles (Fig. 4). DNA was unavailable from the father of MD-449 who is very likely to have carried at least one 112 allele, inferred from sibling and progeny genotypes. Therefore the mutation is implicated as being maternal in origin showing loss of repeats at the BAT-40 locus. The 112/112 homozygous allele is highly unlikely to have arisen by dropout of the larger 120 or 124 bp allele during PCR because four family members including MD-439, the sister of MD-449, had 112/124 genotypes that were easily detected under the PCR conditions used (Fig. 4). This indicates that the technique reliably detects the larger alleles. Furthermore, the 120 bp allele in MD-1303 was faithfully amplified (Fig. 4B). Genotyping for all members of this family was confirmed in triplicate and previous genotyping at microsatellite markers confirmed that MD-1301 was indeed the mother of the twins MD-449 and MD-439 (data not shown).

The observation of a germline mutation in only 11 allele transmissions is striking, since only one mutation event in 3000–5000 transmissions has been reported for CA repeats (10). This led us to further investigate the possibility that BAT-40 is highly unstable in the germline and that mutant alleles might be transmitted in subsequent generations, to manifest as polymorphism within the population.

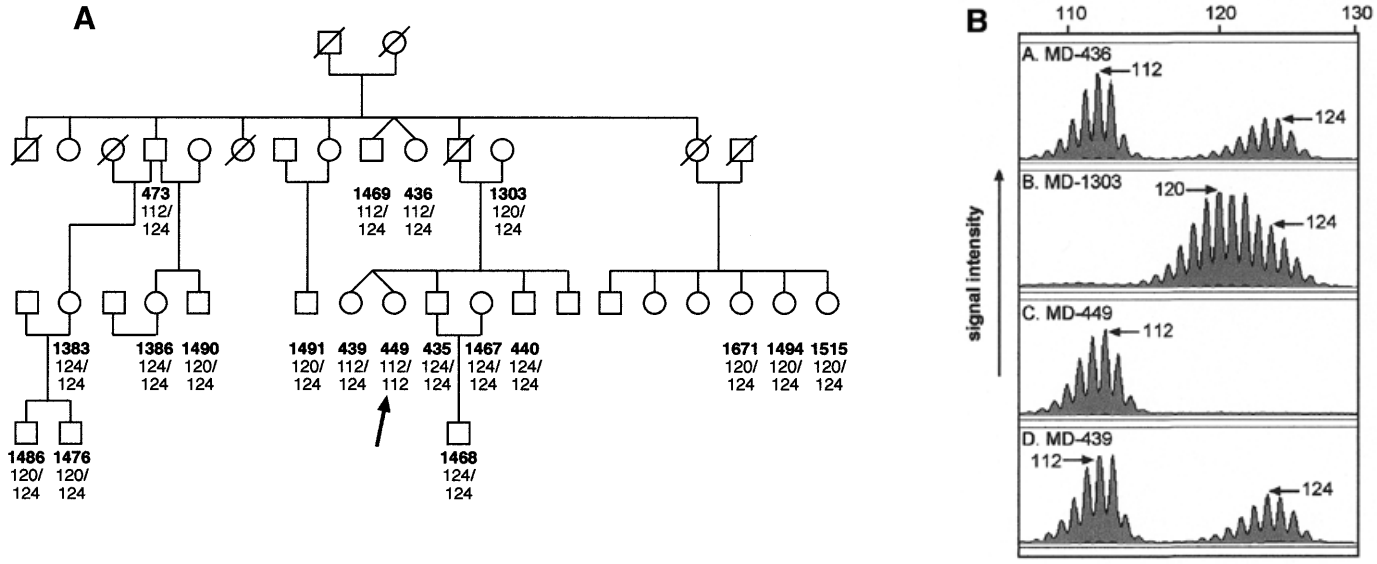

Figure 4. (**A**) BAT-40 genotypes of blood DNA from available samples of pedigree K-435. A single incidence of transmissible germline hypermutability is highlighted. Mother MD-1303 has allele set 120/124 whereas daughter MD-449 (arrow) is homozygous for 112. (**B**) The ABI310 profiles of the BAT-40 alleles of MD-1303 and MD-449 are shown. There is no indication of the presence of either of MD-1303's alleles in MD-449. Although DNA was not available from the father it is likely that he carried at least one 112 allele as inferred from siblings such as his sister MD-436. The profile of MD-439 also demonstrates that there is not a problem in detecting the 124 allele in the presence of the 112 allele.

Germline hypermutability in CEPH family analysis

To further explore germline instability at BAT-40, we analysed BAT-40 alleles in a CEPH family panel. Nine CEPH families were genotyped at the BAT-40 locus, totalling 176 germline transmissions analysed, and 12 putative mutations (6.8%) were identified (Table 1). However, in all cases, heterozygous parental alleles differed by only a few base pairs and the mutations indicated involved small (1 bp) changes (Table 1 and Fig. 5). Of the 88 maternal transmissions analysed, three were mutant at BAT-40 (3.4%) and of the 88 paternal transmissions analysed, nine were mutant (10.2%). This difference was not statistically significant ($\chi^2 = 3.22$, $P = 0.073$). Insertions and deletions appeared to occur equally.

The mutations identified in CEPH families provided further support for our initial observation that BAT-40 is hypermutable in the germline. The CEPH data add further weight to the identification of the germline mutation in family K-435 and also to the evidence from the high levels of heterozygosity at BAT-40 in the caucasian population study demonstrating high levels of heterozygosity at BAT-40. However, although the CEPH mutations were reproducible, the small changes observed in the complex BAT-40 profile led us to devise a further, rigorous method to analyse susceptibility of this locus to mutation in the germline.

Inherent hypermutablity of BAT-40 in the germline demonstrated by small-pool PCR analysis of sperm DNA

SP-PCR analysis of germline DNA has important advantages over family studies for analysing germline stability at complex loci (11,15). The method overcomes the practical constraints encountered during pedigree analyses, which suffer limitations from the small number of mutants that can be identified. In contrast, many hundreds of gametes can be analysed from a single semen sample and consequently, a greater variation in allele size changes may be available for identification. In addition, the dilution of the DNA sample aids unambiguous allele identification at a hypermutable locus. Mutation frequency as detected in sperm DNA has been shown to reflect estimations from studies in pedigrees (11). Comparisons of sperm DNA and constitutional DNA templates have shown that SP-PCR reliably discriminates alleles in both and that there are no demonstrable differences in technical artefact between the two sample templates (11). In addition SP-PCR has been demonstrated to reliably detect differences in intra-allelic mutation frequency in sperm DNA (11,14–16). For SP-PCR of sperm DNA, study subjects were selected as being constitutionally heterozygous at the BAT-40 locus with individual wild-type alleles easily distinguished by size. MD-c1 had allele sizes 120/124 and MD-949 had alleles of size 121/124. Correct identification of constitutional allele sizes was further supported in the SP-PCR analysis where individual alleles of the same predominant allele size were detected (Fig. 6). Approximately 100 SP-PCR products were analysed per sample. BAT-40 allele sizes typed from constitutional and sperm DNA templates by SP-PCR are shown in Figure 7. Mutant alleles were detected in sperm DNA by comparison with constitutional genotype (Fig. 6). The frequency of mutant alleles detected in each sample is summarised in Table 3. In both MD-c1 and MD-949 matched samples, there was a significantly greater number of mutant alleles detected in sperm DNA compared to that of matched blood leukocyte DNA samples ($\chi^2 = 19.32$, $P < 0.001$ and $\chi^2 = 13.82$, $P < 0.001$ for MD-c1 and MD-949, respectively). A total of 9/198 (4.5%) alleles in the leukocyte DNA were mutant compared to a total of 64/255 (25.1%) mutant alleles in the sperm templates, indicating an almost 6-fold increase in mutation accumulation in the germline. The proportion of mutant alleles in blood and sperm was

Table 2. Putative BAT-40 mutations detected in germline transmissions of CEPH families

Family	Genotype (father)	Genotype (mother)	Genotype (child)
66	122/123 (f, -01)	124/127 (m, -02)*	123/126 (c/m, -03)
	123/123 (fm, -12)*	124/127, (mm, -14)	124/127 (m, -02)
1331	123/123 (f, -01)*	119/123 (m, -02)	119/124 (c/m, -17)
1341	120/123 (f, -01)*	123/125 (m, -02)[a]	119/125 (c/f, -05)
	120/123 (f, -01)*	123/125 (m, -02)[a]	119/125 (c/f, -08)
1346	123/123 (f, -01)*	122/127 (m, -02)	124/127 (c/f, -08)
	123/123 (f, -01)*	122/127 (m, -02)	124/127 (c/f, -09)
1362	120/120 (fm, -15)	121/123 (mm, -16)*	120/120 (m, -02)
	121/123 (f, -01)*	120/120 (m, -02)	120/120 (c/f, -04)
1377	120/124 (ff, -10)	120/122 (mf, -11)*	120/121 (f, -01)
	120/121 (f, -01)*	119/122 (m, -02)	122/123 (c/m, -08)
13293	120/123 (f, -01)*	108/109 (M, -02)	109/124 (c/m, -09)

Transmissions in which a germline mutation was indicated are described. The genotype of the child along with both parents is presented. The inferred origin is highlighted by an asterisk. The CEPH pedigree number, individual identification number and family relationship is given as standard CEPH nomenclature.
[a]In one case the sample failed to amplify and genotype was inferred from the relatives.

equivalent in MD-c1 and MD-949 samples suggesting that constitutional heterozygous DNA MMR gene mutation does not influence mutation rate in the male gamete in this case.

Although both addition and deletion mutations were identified, we observed a bias towards repeat losses over gains in the sperm ($\chi^2 = 11.0$, $P < 0.001$ and $\chi^2 = 10.3$, $P < 0.001$ for MD-c1 and MD-949 sperm samples, respectively).

DISCUSSION

Long poly(A/T) tracts are abundant in the human genome, occurring at the 3′ untranslated region (UTR) of genes (30) and within intronic sequences as well as coding regions (17). Length variation at such repetitive sequences has importance with respect to gene expression and function (21,30–32). In addition, poly(A/T) markers play an important role in the classification of microsatellite unstable CRCs (6). Here we report a series of studies aimed at defining germline stability of a paradigm poly(A/T) repeat locus. By analysis of two cohorts we show that BAT-40 is a highly polymorphic locus with an observed level of heterozygosity of 59.7%. This is similar to a previous analysis of a CEPH cohort in which a level of heterozygosity of 72% was reported. However, the level of BAT-40 heterozygosity detected here and in a CEPH cohort by Zhou *et al.* (25) is considerably higher that that observed in a Japanese cohort (14.6%) (29). Although this might be explained in part by variation in allele heterozygosity between populations, the Japanese study cohort were from hospital based samples and may not be representative of the true Japanese population frequencies. Our confirmation that BAT-40 is a highly polymorphic locus suggests that generation of new alleles by slippage and mutation at this locus might be quite common.

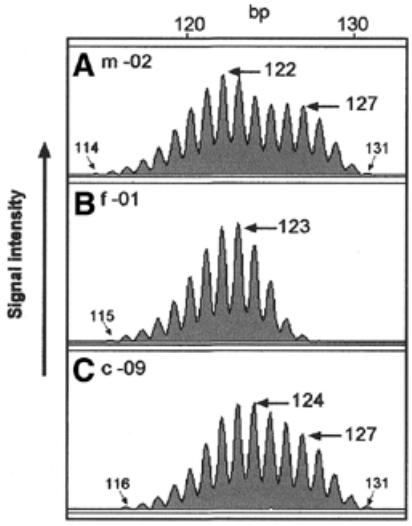

Figure 5. Representative example of a putative BAT-40 germline mutation in CEPH family 1346. While the mother -02 (**A**) has a BAT-40 genotype of 122/127, the father -01 (**B**) appears homozygous for a 123 bp allele. The mother's 127 bp allele is detected in child -09 (**C**) but the most predominant peak in the first complex is at 124 bp. This would indicate a 1 bp mutation at BAT-40 had occurred in the germline of the father. Sizing of the extreme stutter bands also indicates a 1 bp mutation of the father's 123 bp allele in c-09 and confirms the presence of the 127 allele derived from m-02.

Instability at BAT-40 is well documented in MMR deficient tumours (6,21) and we have also observed instability at this locus in cells derived from normal tissue that are deficient in MMR (33). These data also indicate that this repeat locus might be particularly unstable. Identification of a germline mutation at BAT-40 in a Scottish pedigree suggests that this locus is also

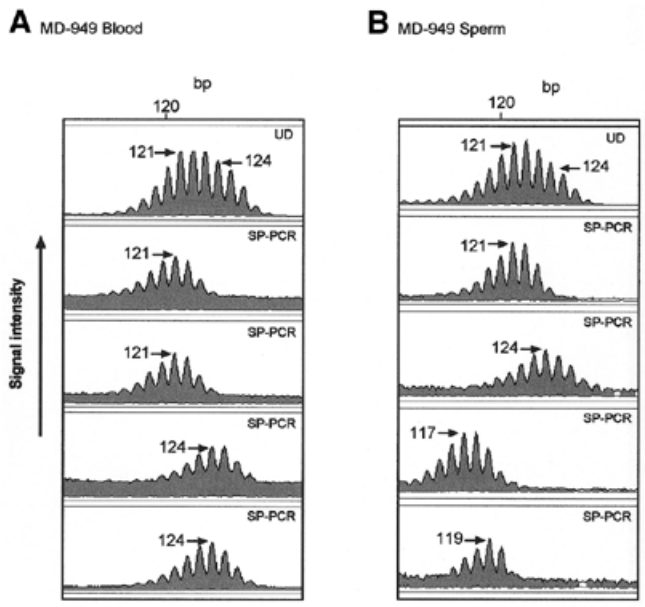

Figure 6. Representative ABI310 traces of BAT-40 alleles detected by SP-PCR in matched blood (**A**) and sperm (**B**) DNA. Almost all BAT-40 SP-PCR products amplified from blood DNA revealed individual alleles with predominant peaks of the same size as those in undiluted (UD) DNA. For MD-949 these were 121 and 124 bp. The majority of BAT-40 SP-PCR products amplified from sperm DNA were also of wild-type allele size as shown. However, a significant number of mutant alleles were detected. Mutants of 117 and 119 bp are illustrated here.

highly unstable in the germline. Explanations such as failure to amplify larger alleles as an explanation for apparent germline mutations, such as that in MD-449, are highly unlikely since larger alleles were detected reliably in the presence of a 112 allele in other family members. Our initial observation in a single family is supported by pedigree analysis of a further nine CEPH families, albeit with less dramatic examples, biased in part by the nature of the parental genotypes. Hence, we chose individuals with easily distinguishable allele sizes for analysis of sperm DNA. Analysis of matched sperm and blood DNA at BAT-40 by SP-PCR demonstrated a statistically significant increase in the proportion of mutant alleles in sperm compared to somatic DNA. This argues strongly that the mutations detected by SP-PCR of sperm DNA are indeed authentic. We employed rigorous controls to ensure against contamination (see Materials and Methods). The SP-PCR approach has been previously shown to detect mutant alleles with equal fidelity in sperm and constitutional DNA templates as demonstrated by direct comparisons between mutation rates detected by SP-PCR of sperm compared to family studies (11). These studies have consistently validated the SP-PCR approach (11,14).

The presence of an inactivating *MLH1* mutation in the germline of one individual did not further influence the level of instability at BAT-40 in the sperm DNA. Intriguingly, there were shorter mutant alleles predominating in the sperm DNA despite the fact that the SP-PCR technique reliably detected both large and short constitutional alleles.

The results presented here provide compelling evidence that BAT-40 is inherently highly unstable in the germline. It will be of interest to determine whether this phenomenon is common and what length of poly(A/T) tract represents a threshold at which instability becomes likely.

Hypermutability at the BAT-40 locus provides an explanation for the wide spectrum of allelic variants present in the

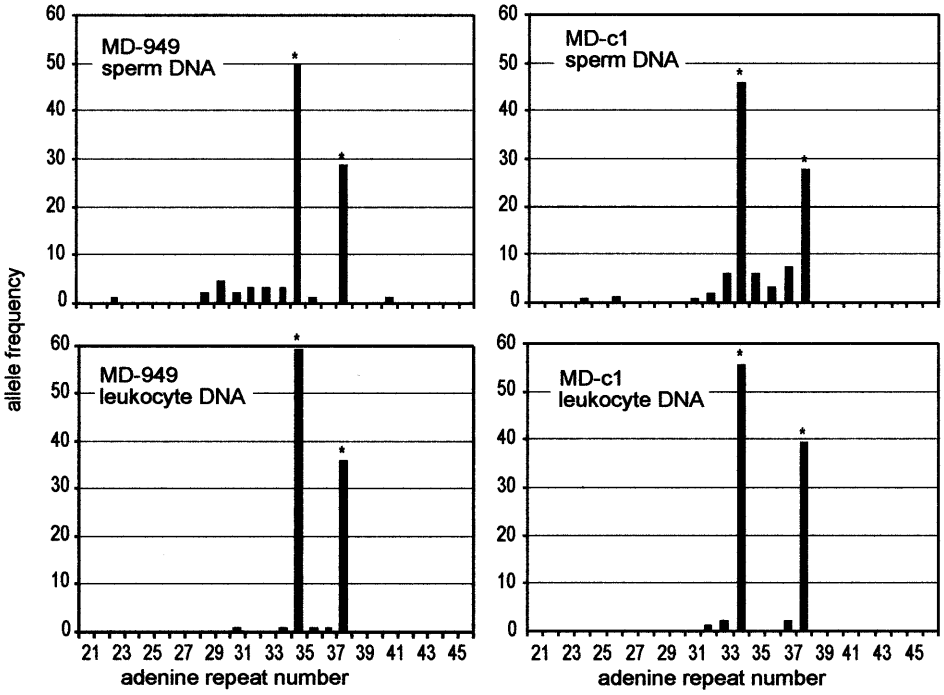

Figure 7. BAT-40 allele sizes in matched constitutional and sperm DNA detected by SP-PCR. The predominant allele sizes for each individual are indicated by asterisks, as detected from analysis of undiluted DNA. MD-949 is a CRC patient with a germline mutation in the human *MLH1* gene. MD-c1 is a normal healthy control individual.

Table 3. Summary of mutant alleles detected by SP-PCR in matched sperm and blood DNA from samples MD-949 and MD-c1

Sample	Total no. alleles	Mutants (frequency)
MD-949 Sperm	91	20 (0.22)
MD-949 Blood	99	4 (0.04) ($\chi^2 = 13.82$, $P < 0.001$)
MD-c1 Sperm	164	44 (0.27)
MD-c1 Blood	99	5 (0.05) ($\chi^2 = 19.32$, $P < 0.001$)

Scottish, CEPH and other populations studied, since transmission of new germline variants can become established within the population.

The evidence that BAT-40 represents a poly(A/T) tract within the genomic structure of a gene and exhibits instability in the germline might be of importance in understanding the mechanisms generating mutations at other such polymorphic repeat loci. Indeed this phenomenon may be common to many poly(A/T) tracts and further study of such sequences is merited to elucidate whether this is a widespread phenomenon. Poly(A/T) tracts are ubiquitous at the 3′UTR of all coding genes where the stability in length of the poly(A) tail is of known functional importance to the stability of the mRNA species (30). Of further relevance, these repetitive tracts are common in intronic sequence (17), and shortening of intronic mononucleotides has also been shown to have functional consequences. For instance, aberrant splice variants of the ATM gene that result in ataxia-telangiectasia can arise as a consequence of shortened intronic mononucleotide tracts (31). In addition the shortened poly$(T)_5$ variant in intron 8 of the cystic fibrosis transmembrane conductance regulator gene causes congenital bilateral absence of the vas deferens when associated with a cystic fibrosis mutation on the other allele (34). Mutation of poly(A/T) tracts within exonic sequences have also been shown to contribute to carcinogenesis and this is exemplified by mutation of the TGFBR2 gene in MSI$^+$ CRCs (20,21). Hence, it seems reasonable to speculate that the mechanism of inherent instability elucidated here might also have relevance to a number of genes containing such repeats.

BAT-40 is used routinely as a marker in determining tumour genomic stability in relation to defective DNA MMR, due to its extreme sensitivity to mutation in the absence of MMR activity (5,21,23). Microsatellite markers that display such germline hypermutability should be used with caution in view of the likelihood of mitotic instability. Very unstable markers may be too sensitive to provide the specificity to MMR defects that is clearly required in such screening strategies. The evidence reported here supports a growing number of studies that highlight the importance of understanding inherent characteristics influencing marker stability when they are used in clinical analyses (7,23,26,35).

ACKNOWLEDGEMENTS

This work was supported by the following grants: Cancer Research Campaign (SP2326/0201), Scottish Health Department (K/MRS/50/C2723) and Urquhart Charitable Trust. A.L.B. was funded by an MRC PhD Student Fellowship and S.M.F. by an RSE Personal Research Fellowship.

REFERENCES

1. Weissenbach,J., Gyapay,G., Dib,C., Vignal,A., Morissette,J., Millasseau,P., Vaysseix,G. and Lathrop,M. (1992) A second-generation linkage map of the human. *Nature*, **359**, 794–801.
2. Aaltonen,L.A., Peltomaki,P., Leach,F.S., Sistonen,P., Pylkkanen,L., Mecklin,J.P., Jarvinen,H., Powell,S.M., Jen,J. and Hamilton,S.R. (1993) Clues to the pathogenesis of familial colorectal cancer. *Science*, **260**, 812–816.
3. Thibodeau,S.N., Bren,G. and Schaid,D. (1993) Microsatellite instability in cancer of the proximal colon. *Science*, **260**, 816–819.
4. Ionov,Y., Peinado,M.A., Malkhosyan,S., Shibata,D. and Perucho,M. (1993) Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, **363**, 558–561.
5. Rodriguez-Bigas,M.A., Boland,C.R., Hamilton,S.R., Henson,D.E., Jass,J.R., Khan,P.M., Lynch,H., Perucho,M., Smyrk,T., Sobin,L. and Srivastava,S. (1997) A National Cancer Institute workshop on hereditary nonpolyposis colorectal cancer syndrome: meeting highlights and Bethesda guidelines. *J. Natl Cancer Inst.*, **89**, 1758–1762.
6. Dietmaier,W., Wallinger,S., Bocker,T., Kullmann,F., Fishel,R. and Ruschoff,J. (1997) Diagnostic microsatellite instability: definition and correlation with mismatch repair protein expression. *Cancer Res.*, **57**, 4749–4756.
7. Bacon,A.L., Farrington,S.M. and Dunlop,M.G. (2000) Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. *Hum. Mol. Genet.*, **9**, 2707–2713.
8. Sturzeneker,R., Bevilacqua,R.A., Haddad,L.A., Simpson,A.J. and Pena,S.D. (2000) Microsatellite instability in tumors as a model to study the process of microsatellite. *Hum. Mol. Genet.*, **9**, 347–352.
9. Brinkmann,B., Klintschar,M., Neuhuber,F., Huhne,J. and Rolf,B. (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.*, **62**, 1408–1415.
10. Weber,J.L. and Wong,C. (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.*, **2**, 1123–1128.
11. Jeffreys,A.J., Tamaki,K., MacLeod,A., Monckton,D.G., Neil,D.L. and Armour,J.A. (1994) Complex gene conversion events in germline mutation at human minisatellites. *Nature Genet.*, **6**, 136–145.
12. May,C.A., Jeffreys,A.J. and Armour,J.A. (1996) Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). *Hum. Mol. Genet.*, **5**, 1823–1833.
13. Monckton,D.G., Cayuela,M.L., Gould,F.K., Brock,G.J., Silva,R. and Ashizawa,T. (1999) Very large (CAG)(n) DNA repeat expansions in the sperm of two spinocerebellar ataxia type 7 males. *Hum. Mol. Genet.*, **8**, 2473–2478.
14. Crawford,D.C., Wilson,B. and Sherman,S.L. (2000) Factors involved in the initial mutation of the fragile X CGG repeat as determined by sperm small pool PCR. *Hum. Mol. Genet.*, **9**, 2909–2918.
15. Kunst,C.B., Leeflang,E.P., Iber,J.C., Arnheim,N. and Warren,S.T. (1997) The effect of FMR1 CGG repeat interruptions on mutation frequency as measured by sperm typing. *J. Med. Genet.*, **34**, 627–631.
16. Mornet,E., Chateau,C., Hirst,M.C., Thepot,F., Taillandier,A., Cibois,O. and Serre,J.L. (1996) Analysis of germline variation at the FMR1 CGG repeat shows variation in the normal-premutated borderline range. *Hum. Mol. Genet.*, **5**, 821–825.
17. Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.
18. International human genome sequencing consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
19. Markowitz,S. (2000) TGF-beta receptors and DNA repair genes, coupled targets in a pathway of human colon carcinogenesis. *Biochim. Biophys. Acta*, **1470**, M13–M20.
20. Markowitz,S., Wang,J., Myeroff,L., Parsons,R., Sun,L., Lutterbaugh,J., Fan,R.S., Zborowska,E., Kinzler,K.W. and Vogelstein,B. (1995) Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science*, **268**, 1336–1338.
21. Parsons,R., Myeroff,L.L., Liu,B., Willson,J.K., Markowitz,S.D., Kinzler,K.W. and Vogelstein,B. (1995) Microsatellite instability and mutations of the transforming growth factor beta type II receptor gene in colorectal cancer. *Cancer Res.*, **55**, 5548–5550.

22. Lachance,Y., Luu-The,V., Labrie,C., Simard,J., Dumont,M., de Launoit,Y., Guerin,S., Leblanc,G. and Labrie,F. (1990) Characterization of human 3 beta-hydroxysteroid dehydrogenase/delta 5-delta 4-isomerase gene and its expression in mammalian cells [published erratum appears in *J. Biol. Chem.* (1992), **267**, 3551]. *J. Biol. Chem.*, **265**, 20469–20475.
23. Boland,C.R., Thibodeau,S.N., Hamilton,S.R., Sidransky,D., Eshleman,J.R., Burt,R.W., Meltzer,S.J., Rodriguez-Bigas,M.A., Fodde,R., Ranzani,G.N. and Srivastava,S. (1998) A National Cancer Institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res.*, **58**, 5248–5257.
24. Liu,B., Farrington,S.M., Petersen,G.M., Hamilton,S.R., Parsons,R., Papadopoulos,N., Fujiwara,T., Jen,J., Kinzler,K.W., Wyllie,A.H., Vogelstein,B. and Dunlop,M. (1995) Genetic instability occurs in the majority of young patients with colorectal cancer. *Nature Med.*, **1**, 348–352.
25. Zhou,X.P., Hoang,J.M., Cottu,P., Thomas,G. and Hamelin,R. (1997) Allelic profiles of mononucleotide repeat microsatellites in control individuals and in colorectal tumours with and without replication errors. *Oncogene*, **15**, 1713–1718.
26. Samowitz,W.S., Slattery,M.L., Potter,J.D. and Leppert,M.F. (1999) BAT-26 and BAT-40 instability in colorectal adenomas and carcinomas and germline polymorphisms. *Am. J. Pathol.*, **154**, 1637–1641.
27. Jeffreys,A.J., Neumann,R. and Wilson,V. (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell*, **60**, 473–485.
28. Vilkki,S., Tsao,J.L., Loukola,A., Poyhonen,M., Vierimaa,O., Herva,R., Aaltonen,L.A. and Shibata,D. (2001) Extensive somatic microsatellite mutations in normal human tissue. *Cancer Res.*, **61**, 4541–4544.
29. Yokozaki,H. (2000) Distribution of germline BAT-40 poly-adenine tract microsatellite variants in the Japanese. *Int. J. Mol. Med.*, **6**, 445–448.
30. Bernstein,P. and Ross,J. (1989) Poly(A), poly(A) binding protein and the regulation of mRNA stability. *Trends Biochem. Sci.*, **14**, 373–377.
31. Ejima,Y., Yang,L. and Sasaki,M.S. (2000) Aberrant splicing of the ATM gene associated with shortening of the intronic mononucleotide tract in human colon tumour cell lines: a novel mutation target of microsatellite instability. *Int. J. Cancer*, **86**, 262–268.
32. Chu,C.S., Trapnell,B.C., Curristin,S., Cutting,G.R. and Crystal,R.G. (1993) Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nature Genet.*, **3**, 151–156.
33. Bacon,A.L., Farrington,S.M. and Dunlop,M.G. (2001) Mutation frequency in coding and non-coding repeat sequences in mismatch repair deficient cells derived from normal human tissue. *Oncogene,* **20**, in press.
34. Chillon,M., Casals,T., Mercier,B., Bassas,L., Lissens,W., Silber,S., Romey,M.C., Ruiz-Romero,J., Verlingue,C. and Claustres,M. (1995) Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. *N. Engl. J. Med.*, **332**, 1475–1480.
35. Frazier,M.L., Sinicrope,F.A., Amos,C.I., Cleary,K.R., Lynch,P.M., Levin,B. and Luthra,R. (1999) Loci for efficient detection of microsatellite instability in hereditary non-polyposis colorectal cancer. *Oncol. Rep.*, **6**, 497–505.