Human Molecular Genetics, 1999, Vol. 8, No. 2 **173–183**

ARTICLE

Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism

Esther P. Leeflang[1], Simon Tavaré[1,2], Paul Marjoram[2], Carolyn O. S. Neal[1], Jayalakshmi Srinidhi[3], Heather MacFarlane[3], Marcy E. MacDonald[3], James F. Gusella[3], Margot de Young[4], Nancy S. Wexler[5,6] and Norman Arnheim[1,*]

[1]Molecular Biology Program and [2]Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1340, USA, [3]Molecular Neurogenetics Unit, Massachusetts General Hospital, Boston, MA 02129, USA, [4]Association of Friends of Families with Huntington Disease, Maracaibo, Zulia, Venezuela, [5]Department of Neurology, Columbia University, New York, NY 10032, USA and [6]Hereditary Disease Foundation, 1427 7th Street, Suite 2, Santa Monica, CA 90401, USA

Received September 11, 1998; Revised and Accepted November 20, 1998

Trinucleotide repeat disease alleles can undergo 'dynamic' mutations in which repeat number may change when a gene is transmitted from parent to offspring. By typing >3500 sperm, we determined the size distribution of Huntington's disease (HD) germline mutations produced by 26 individuals from the Venezuelan cohort with CAG/CTG repeat numbers ranging from 37 to 62. Both the mutation frequency and mean change in allele size increased with increasing somatic repeat number. The mutation frequencies averaged 82% and, for individuals with at least 50 repeats, 98%. The extraordinarily high mutation frequency levels are most consistent with a mutation process that occurs throughout germline mitotic divisions, rather than resulting from a single meiotic event. In several cases, the mean change in repeat number differed significantly among individuals with similar somatic allele sizes. This individual variation could not be attributed to age in a simple way or to '*cis*' sequences, suggesting the influence of genetic background or other factors. A familial effect is suggested in one family where both the father and son gave highly unusual spectra compared with other individuals matched for age and repeat number. A statistical model based on incomplete processing of Okazaki fragments during DNA replication was found to provide an excellent fit to the data but variation in parameter values among individuals suggests that the molecular mechanism might be more complex.

INTRODUCTION

At least 11 neurodegenerative diseases result from expansion in the number of trinucleotide repeats in or adjacent to a protein coding gene (1–4). For most of these diseases the unstable sequence consists of CAG/CTG triplets. A remarkable characteristic of trinucleotide repeat disease alleles is that they can undergo dynamic mutations in which repeat number may change when the disease gene is transmitted from an affected parent to the offspring. The molecular basis of this dynamic mutation process is of great fundamental interest and stands in contrast to the stable transmission of other disease mutations.

Trinucleotide repeat instability is influenced by the sex of the transmitting parent, the number of repeats and the purity of the repeat tract (1–4). The extent to which other factors such as age, genetic background or interallelic effects may contribute to dynamic mutation is less certain and more difficult to analyze (5–9). Since only a small number of offspring is conceived by any one affected parent, it is difficult to study the relationship between mutation characteristics and other variables in a single individual. Pooling transmission data from many different individuals may confound the analysis of important variables. For example, the pooled size distribution of mutant alleles found among offspring of affected parents with the same mutant allele size could reflect a random sampling from the mutation size distribution characteristic of that somatic allele size in any individual. On the other hand, the pooled distribution could reflect sampling from mutation size distributions that differed significantly among the parents. Fortunately, the large sample sizes afforded by single genome

*To whom correspondence should be addressed. Tel: +1 213 740 7675; Fax: +1 213 740 8631; Email: arnheim@molbio.usc.edu

analyses such as single sperm typing allow accurate estimates of the germline mutation frequency and the size distribution of mutant sperm (the mutation spectrum) in individual males. Each sperm represents a potential paternal transmission.

A detailed description of the human dynamic mutation process in individual males is useful in three regards. First, it can define a precise quantitative standard against which the results from experimental model systems can be judged for their applicability to human dynamic mutations. Second, it can provide clues as to what molecular mechanisms may be contributing to the mutation process. Finally, with the proper experimental design the data can be used to examine the role of genetic and possibly environmental factors in genetic instability.

Huntington's disease (HD) is associated with progressive disordered movements, decline in cognitive function and emotional disturbance. Disease-causing alleles exhibit significant instability especially when transmitted paternally (9–17). In order to begin to unravel the contributions of different variables to dynamic mutation at the HD locus, we studied 26 men from the large Venezuelan HD cohort (18). We generated mutation spectra from individuals with somatic allele sizes ranging from 37 to 62 repeats. We used sperm from affected individuals or individuals at risk, including siblings, cousins and a father and his son. Our data were compared with the mutation spectra expected under a simple Okazaki fragment processing model of trinucleotide repeat instability.

RESULTS

A number of studies have shown that somatic variation in HD allele size is extremely limited compared with germline variation (17,19,20). Consequently, trinucleotide repeat mutations can be detected if the HD allele repeat number in each sperm is compared with the allele size in somatic DNA from the same individual (19). Data from the analysis of 27 sperm samples, including three published previously (19), are shown in Table 1. Included is the age of each donor, the somatic HD allele repeat number, the observed expansion and contraction mutation frequencies and the mean and standard deviation of the change in repeat number. In Figure 1, we show the 27 mutation spectra. The histograms of the allele sizes only include sperm with the original HD somatic allele size or a size derived from it by mutation. The data can be obtained electronically at http://hto-e.usc.edu/datasets/leeflang98

Table 1. Donor identification, somatic allele size $(CAG)_n$, heterozygosity (E) or homozygosity (O) at the adjacent CCG tract, paternal (P) or maternal (M) inheritance of the allele tested

Donor ID	$(CAG)_n$	Age (years)	Sample size	Mean	Standard deviation	% expansion	% contraction	E/O	P/M	$p_S \times 10^{-4}$	p_D	$p_L \times 10^{-3}$
A	62	18	168	32.51	11.15	99	0	E	P	120.0	0.51	400.0
B	62	17	118	9.74	6.90	94	4	O	M	210.0	0.00	310.0
C	53	21	100	12.98	9.82	89	8	E	M	200.0	0.45	140.0
D	51	29	95	20.35	12.25	97	3	O	P	210.0	0.15	170.0
E	50	29	61	15.39	13.16	87	10	O	P	260.0	0.23	120.0
F	49	23	94	10.41	8.24	95	3	O	M	8.0	0.78	41.0
G	48	31	113	3.36	5.12	65	19	O	M	6.4	0.78	8.0
H	48	26	109	2.67	5.51	59	29	O	M	13.0	0.88	5.6
I	47	34	114	4.88	7.65	79	16	O	P	13.0	0.86	6.8
J	46	27	113	0.86	3.22	45	34	O	M	11.0	1.00	0.5
K	45	52	136	3.40	6.51	63	26	O	M	5.1	0.86	2.8
L	45	41	367	4.19	6.40	72	15	O	M	10.0	0.80	5.8
M	45	31	63	6.41	6.07	98	0	O	M	0.0	0.66	21.0
N	45	29	106	2.98	5.44	75	13	O	M	6.1	0.73	8.6
O	44	54	128	1.56	4.10	45	27	E	P	2.3	0.83	1.5
P	44	52	107	23.37	30.31	85	8	E	M	27.0	1.00	9.6
Q	44	44	152	0.67	3.54	42	41	O	M	5.6	0.90	0.6
R	44	30	158	1.20	2.18	61	22	O	P	7.4	0.00	11.0
S	44	21	116	1.13	2.26	53	22	O	M	8.6	0.48	11.0
T	43	52	121	2.56	3.61	77	12	O	M	4.3	0.51	5.5
U	43	17	121	0.42	1.26	41	19	O	M	6.4	0.27	9.8
V	42	35	112	1.75	2.28	70	12	O		3.6	0.29	9.3
W	41	38	191	0.65	1.91	49	23	O	M	3.5	0.35	2.8
X	40	18	183	–0.01	1.42	31	33	E	M	11.0	0.55	0.8
Y	39	65	180	2.73	6.82	69	15	E	P	2.4	0.76	2.4
Z	39	63	210	2.26	3.26	73	18	E	P	3.3	0.47	4.2
AA	37	37	136	0.04	1.10	26	24	E	M	2.0	0.16	0.3

p_S, type I mutation rate; p_D and p_L, type II mutation rates.

Variation in germline mutation spectra within an individual

We examined whether differences in mutation spectra existed between sperm samples taken at age 63 (sample Z) and age 65 (sample Y) from the same donor. The mutation spectra show indistinguishable patterns except for two sperm in the older sample with allele sizes considerably larger than those seen in the age 63 sample. There is no statistically significant difference between the two samples in the mutation frequency ($P = 0.08$) or the mean change in repeat number ($P = 0.40$).

Variation in germline mutation spectra between individuals

Qualitatively, the mutation spectra of different individuals can sometimes vary dramatically, even among individuals with similar ages and somatic allele lengths. Compare, for example, the spectra of individuals A and B shown in Figure 1. Quantitatively, this individual variation is reflected in the differences in mean allele size in the sperm (Table 1). Figure 2 illustrates the general feature that the larger the somatic allele length the greater the average allele length of the mutant sperm.

The mutation frequency for each individual's HD allele was calculated from the mutation spectra by dividing the number of sperm which differed in size from the originally inherited HD allele by the total number of HD sperm. The mutation frequency always exceeds 50% and in most cases is >80%. For individuals with at least 50 repeats the average mutation frequency was 98%. The expansion mutation frequency in an individual almost always exceeds that for contractions. As shown in Figure 3, there is also a clear trend for the mutation frequency to increase with allele size.

Effects of parental origin on HD allele instability

It is possible that the parental origin of the HD allele shows an effect on instability. Therefore we divided the sample of 25 distinct individuals with known origin into two groups depending on whether the HD allele was paternally (seven donors) or maternally (18 donors) inherited (Table 1). Using permutation tests (see Statistical analysis), we concluded that parental origin has no discernable effect on the average somatic repeat number. This being the case we further compared the average of the mean change in repeat number and the average mutation frequency in sperm. No effect of parental origin was detected.

Interallelic effects on instability

There is a $(CCG)_n$ polymorphism almost immediately adjacent to the HD CAG repeat tract (21,22). Our donors could be divided into those that are homozygous for the $(CCG)_7$ allele (19 donors) and those heterozygous for $(CCG)_7$ and $(CCG)_{10}$ (seven donors) (Table 1). Using the permutation approach described above, we also found that $(CCG)_n$ genotype has no significant effect on the average somatic repeat number. Similarly, we found no effect of the $(CCG)_n$ polymorphism on the average of the mean change in repeat number and the average mutation frequency in sperm.

Models of trinucleotide repeat mutation

Our focus is on using our data to help us to understand the molecular mechanism of dynamic mutation at the HD locus. In the next sections, we describe a model (23,24) based on Okazaki fragment processing and show how it can be fitted to the observed mutation spectra.

Molecular events occurring during recombination or DNA replication have been hypothesized to explain instability at microsatellite repeat loci. Based on model organisms, there is little experimental support for a recombination mechanism. Studies on *Escherichia coli* and yeast strains carrying mutants that dramatically reduce recombination result in little increase in microsatellite (including trinucleotide repeat) instability (25–29).

Triplet repeat expansions may occur during DNA replication as a result of slippage or deficient Okazaki fragment processing. Slippage is thought to be the basis of microsatellite repeat mutations observed in yeast and in certain human colon cancers (26,30–34). Loops created by replication slippage can lead to either expansion or contraction mutations depending on whether loop formation occurs on the nascent or template strand, respectively.

Recent experiments in yeast suggest that triplet repeat expansions occur during DNA replication of the lagging strand (24,28,35–37). DNA flaps generated during lagging strand synthesis are normally processed by the Okazaki fragment flap endonuclease Fen-1 (38,39). Flaps containing certain trinucleotide repeat sequences are hypothesized to resist Fen-1 cleavage (23). This could lead to loop formation on the nascent strand and result primarily in expansion mutations.

Using the mutation spectra to estimate the allele-specific HD mutation rate

Assuming that germline mutations in the HD gene are generated during mitotic DNA replication, the mutation rate per cell division might be estimated if the number of cell divisions that preceded sperm formation is known. However, dynamic mutations are unlike classical mutations: not all mutant sperm may have experienced the same number of mutation events. Thus, a sperm that has gained 20 repeats compared with somatic DNA could have undergone the expansion due to a single mutation event at one division, could have experienced multiple, but smaller, expansion events occurring over many divisions or could have had an extensive history of both expansions and contractions. Due to this uncertainty, new methods of mutation rate analysis need to be devised. Any estimate of the dynamic mutation rate per cell division using mutation frequency data must consider both the nature of the mutation event (how many repeats are added or subtracted per event) as well as the total number of mutation events that have occurred over the sperm's DNA replication history.

Our approach to estimating the HD mutation rate is based on comparing the observed mutation spectra with spectra obtained by modeling the mutation process and the history of spermatogonial stem cell divisions. We note that for most of the CAG/CTG diseases, greater instability is observed in male than in female transmissions (1–4); this may be influenced by the continual mitotic divisions experienced by spermatogonia. One candidate model, proposed earlier in the development of replication slippage models for microsatellite mutations, is the stepwise mutation model (40,41). In our setting, this model allows for the addition or deletion of a single repeat when copying a given triplet on either the lagging or leading strand. This mechanism can be asymmetrical, in the sense that a triplet need not be deleted or added with equal probability. When fitted to the data presented

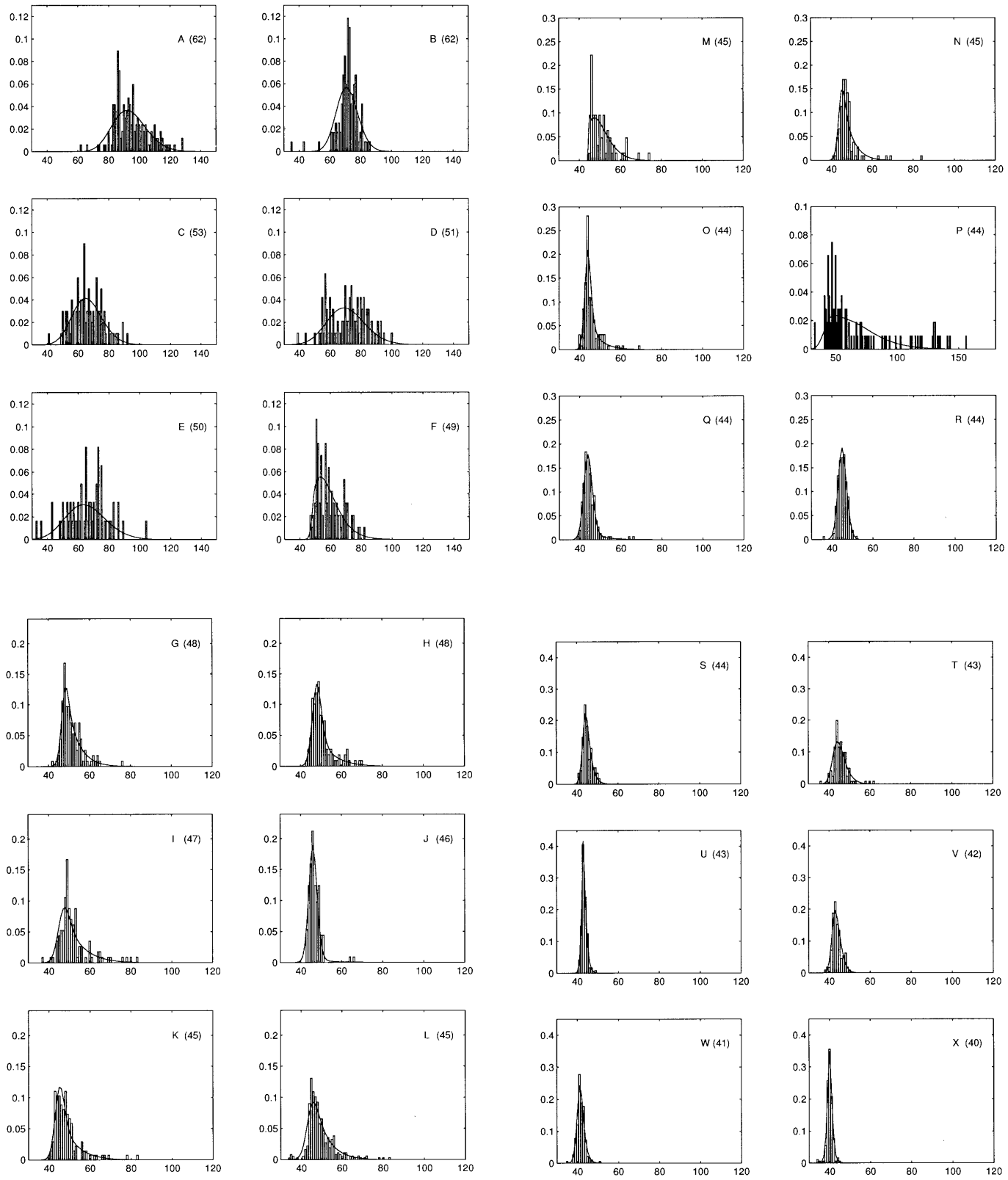

Figure 1. Mutation spectra of sperm samples. The vertical bars are actual data, shown as fractions of the sample size. Data on donors D, E and a portion of Z were previously published (19). The curves trace the distribution expected according to the model described in the text. Note the three distinct *x*-axis scales that are used.

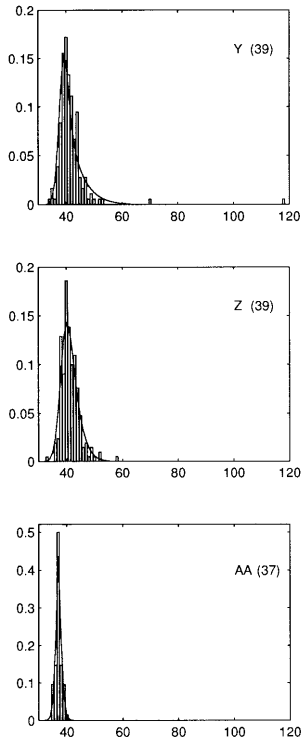

Figure 1. *Continued*

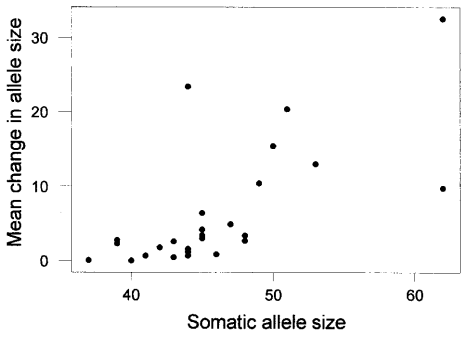

Figure 2. Relationship between the mean change in allele size resulting from mutation and number of repeats in the somatic HD allele of each donor.

here, this model provided an adequate fit to the five donors with somatic allele sizes of at least 50 repeats and to five of those eight with sizes ≤43 repeats. In contrast, the model fitted only one of the 14 donors with somatic allele sizes of 44–49 repeats (further details on the model are provided in P. Marjoram *et al.*, in preparation). This result casts serious doubt on the adequacy of so simple a molecular model for the expansion mechanism. As noted in Leeflang *et al.* (19), any model has to allow for the addition of much larger numbers of repeats during any one replication event. A mechanism for this is described in the next sections.

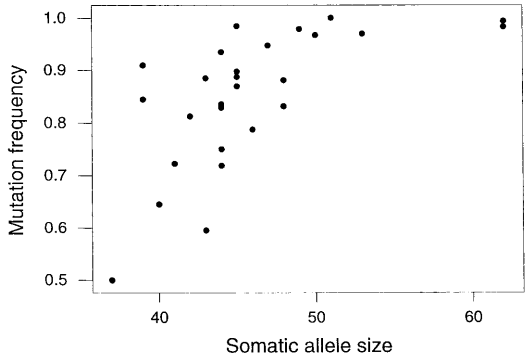

Figure 3. Relationship between mutation frequency and somatic HD allele size of each donor.

The Okazaki fragment processing model of trinucleotide repeat mutations

Recent experiments have shown that yeast carrying a large CAG/CTG repeat tract and an interruption in the *RAD27* gene (the *Saccharomyces cerevisiae* homolog of Fen-1) exhibit a marked increase in the frequency of expansion mutations compared with wild-type yeast cells (24,28). RAD27-deficient strains also exhibit duplication mutations between short direct repeats as well as expansion mutations in dinucleotide repeat tracts (35,37). In addition to these new data, it was previously noted that significant trinucleotide repeat instability in humans is first manifested when the length of the repeated region approaches the size of an average mammalian Okazaki fragment (42–44). Data on the strand preference of certain mutations in *E.coli* (45,46) also support the idea that trinucleotide repeat expansion mutations may be preferentially initiated during synthesis of the lagging strand.

During DNA replication of the lagging strand, Okazaki fragments are initiated sequentially in the direction opposite to that of the advancing replication fork. As a consequence, an Okazaki fragment may have its 5′-end displaced by polymerase extension of the immediately upstream new Okazaki fragment. Normally, the flap would be expected to be eliminated by Fen-1. Gordenin *et al.* (23) argued that a displaced flap containing CAG or CTG sequences might form a thermodynamically favorable secondary structure, a well-known characteristic of CAG/CTG sequences *in vitro* (47–53) and *in vivo* (54,55). Based on the biochemical properties of Fen-1 (38,39), Gordenin *et al.* (23) further postulated that formation of any hairpin-like structure at the 5′-end of the flap would be expected to inhibit the ability of Fen-1 to remove the flap through its endonuclease activity. Ligation of the 5′-end of a flap (that had not been removed by Fen-1) to the 3′-end of the upstream Okazaki fragment will result in a nascent DNA strand with an increased number of repeats equal to the length of the flap (23,24,36) and result in an expansion mutation.

Modeling HD mutations

We modeled the trinucleotide repeat mutation process using a simplified Okazaki fragment model of mutation. We posit two different steps in the mutation process. The first is the requirement

that the Okazaki fragment is initiated within the CAG/CTG repeat tract to allow for secondary structure formation at its 5′-end if displaced by the upstream Okazaki fragment. To model this, we use experimental data on the size distribution of mammalian Okazaki fragment length (56). In the second step we ask how many repeats are contained in a flap displaced by the upstream Okazaki fragment. This determines the size of the expansion mutation after integration of the unprocessed flap into the nascent DNA strand.

In addition to this mechanism we include another one, based on DNA slippage, which allows for the loss (as well as addition) of one repeat when replicating each repeat on the leading strand. This second process is important since mutation spectra with smaller disease alleles contain a sizable number of short contractions. While we chose to model the slippage mechanism occurring on the leading strand only, the model with slippage on both strands can be analyzed in exactly the same way (see Probabilistic model).

Finally, the cell division history of each individual sperm is estimated by considering the age of the donor at the time the sample was collected. Before spermatogenesis begins at puberty (assumed to occur at age 13), the spermatogonial stem cells have undergone an estimated 34 divisions since formation of the zygote. After puberty the stem cells divide ~23 times/year (57). As described in more detail in Materials and Methods, these features can be combined into a probability model that allows us to calculate the chance that a repeat tract in a sperm from a donor of given somatic allele size and age has any particular number of repeats. Previous models of the HD expansion process have not included cell division history (19,58).

Fitting the model to the data

There are three parameters to be estimated in our model: p_S gives the probability that a slippage mutation occurs while replicating a triplet on the leading strand (this mutation results in the addition or deletion of a triplet with equal probability); p_D is the parameter of the displacement process during Okazaki fragment synthesis; and p_L is the ligation probability, the chance that a flap is incorporated into the nascent strand. Estimates of these parameters were determined for each individual data set (Table 1) by the method of maximum likelihood, using the approach outlined in Estimation of parameters.

Most striking is the agreement between observed and expected mutation spectra (Fig. 1). Other formal statistical tests of the adequacy of the fits are described in Estimation of parameters. The general adequacy of the fits stands in marked contrast to the simpler stepwise mutation model alluded to above. This agreement between model and data suggests that errors in Okazaki fragment processing provide a plausible mechanism for the dynamic mutation process at the HD locus.

One feature of the fits that suggests a more complicated mutation mechanism is the variability in the pattern of the parameter estimates themselves. For example, a simple model might have predicted that the p_S values be the same for each individual. However, the standard errors of the parameter estimates (found by simulation) show that the slippage parameter p_S varies significantly among the donors. In particular, the values for the longer somatic alleles are considerably bigger than those for the shorter somatic alleles. Similarly, the ligation probabilities p_L are considerably bigger for the larger somatic alleles. The biological basis of this variation is not yet understood.

DISCUSSION

Dynamic mutations are likely to occur during germline mitotic divisions

Trinucleotide repeat instability could arise in the germline or post-zygotically. The latter was proposed to be the basis for expansions in males inheriting a fragile X pre-mutation (59,60), but recent studies have concluded that expansion to a full mutation is more likely to occur in the germline rather than following fertilization but prior to segregation of the germline (61). If the germline is the site of trinucleotide repeat mutations, then the question of whether instability is a pre-meiotic or meiotic phenomenon still remains unknown.

Consideration of our HD data presents difficulties for a single event meiotic model. In the case of the largest HD alleles, the meiotic mutation rate would have to be close to unity since, in individuals with at least 50 repeats, an average of 98% of the sperm carrying the HD allele have undergone a mutation. If HD mutation events occurred once during replication, at the last premeiotic S phase for example, two events would be required to explain the almost 100% observed mutation frequency.

First, a mutation must arise on both the leading and lagging strands of the replicating chromosome carrying the HD allele. This would yield two heteroduplex chromatids each with one DNA strand with the original HD allele size and the other DNA strand with the newly mutated HD allele size.

Second, each heteroduplex chromatid must undergo repair so that the mutated loop-containing strand is retained and the unmutated original HD strand is lost. A repair bias of almost 50:1 would be required to account for the data on donors with a 98% average mutation frequency. Trinucleotide repeat sequences form palindrome-like structures and the repair of palindromic loops in mammalian cells has been reported to be biased by only as much as 2:1 in favor of retaining the loop structure (62). Additional experimental data on large palindromic loop repair are clearly needed.

It would also be difficult to explain the highest HD mutation frequencies (in individuals with at least 50 repeats) based on a meiotic recombination event. If some form of recombination occurred between the normal allele and the HD allele that produced two new mutant HD alleles we would have observed drastically reduced numbers of sperm with the normal allele, which was not the case. Recombination between the HD-containing sister chromatids is a possibility, but would have to occur in almost every meiosis and result only in increases in repeat number on both sister chromatids to explain our data. No mechanism that could accomplish this in a single step has been identified.

The above considerations also apply to previously published transmission data on fragile X syndrome and myotonic dystrophy. Mothers with 90–129 repeat premutation alleles passed on the full fragile X mutation (>200 repeats) to their offspring in 98% of the transmissions (6). Similarly, parents carrying DM alleles in the size range 50–300 repeats transmit expansion mutations 87% of the time (63,64).

Additional support exists for a germline mitotic model. First, in *E.coli* (25,27) and yeast (24,28,29,65–67) long inserted trinucleotide repeat sequences exhibit instability in the absence of

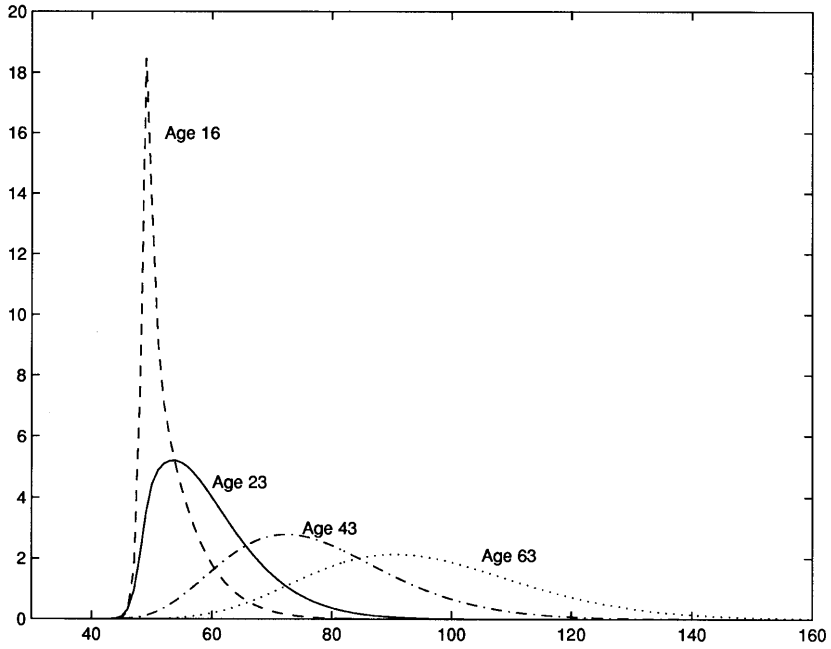

Figure 4. The expected distribution of allele sizes in sperm from donor F at ages 16, 23, 43 and 63 as predicted according to the model.

meiosis. Second, in many of the trinucleotide repeat diseases, some somatic instability is detected (1–4). In HD and a number of other diseases, somatic variation is often less than in the germline. The basis for this difference among the diseases is unknown but may be due to properties unique to the different loci, perhaps relating to the position of replication origins used in germline versus somatic tissues. Finally, analysis of human somatic cells from patients with large myotonic dystrophy disease alleles demonstrated a gradual increase in allele size as a function of an increasing number of generations in culture and supports the mitotic division model (68).

HD dynamic mutations may result from a defect in Okazaki fragment processing

Although it can be argued that HD mutations arise during mitotic DNA replication, the fact that our data fit a simple model based on a defect in Okazaki fragment processing does not prove that this is the molecular mechanism. In addition to the integration of the unprocessed flap into the nascent strand that is incorporated in our model, double-strand break formation at the site of the flap might also lead to mutation through recombination or end joining (23,35). Mitotic recombination over many generations of germline cell divisions would also be compatible with the high mutation frequency and large changes in repeat number seen in sperm. However, studies on trinucleotide repeats in *E.coli* (25,27) and yeast (28,29) as well as on microsatellite repeats in yeast (26,69) do not support an important role for recombination, while the recent data on trinucleotide repeats in yeast strains carrying *RAD27* mutants are consistent with some form of the Okazaki fragment processing model. As more is understood about the molecular mechanisms of instability we can refine the Okazaki fragment model to look for better fits to the data and a simpler pattern in the parameter values.

The role of paternal age in determining the mutation spectra

Our analysis suggests that there may be a paternal age effect on the HD mutation spectrum. Using the model which incorporates spermatogonial stem cell division, we can predict the effect of age on the mutation spectra of individual sperm donors. For example we determined the expected mutation spectrum of donor P (44 repeats) at the age his son (donor A, 62 repeats) was conceived. The frequency of alleles with at least 62 repeats in the father's sperm would be expected to have been >27% (data not shown).

Figure 4 shows the predicted mutation spectra for the 23-year-old donor F at ages 16, 23, 43 and 63, calculated from the model using his estimated parameter values obtained from Table 1. These spectra can be compared with somatic mutation spectra at the myotonic dystrophy locus derived from two tissue samples from a single affected individual taken 5 years apart (70,71). From these data it was suggested that the range and mean allele size would increase but the modal frequency would decrease with age (71). The predicted age effect on HD germline mutation in donor F is virtually the same (Fig. 4). It is striking that this conclusion is based strictly on using the parameter values determined from our model of the HD mutation process.

To prove an effect of age independent of all other genetic variables experimentally, sperm samples donated throughout the lifetime of an individual would be ideal (72). Two samples were taken from the same donor at ages 63 and 65 (Z and Y). No significant effect of age was detected in any of the variables (Table 1 and Fig. 1). This is not unexpected since 2 years accounts only for a small fraction of this donor's sperm cell division history (46/1230 = 3.7%). To maximize the direct detection of an age effect, samples from a donor with as large an HD allele and as close in age to puberty as possible should be taken for comparison with later donations. This strategy of course is complicated by the early onset and severity of HD in individuals carrying the largest

alleles. The predictions made for donor F at ages 43 and 63, for example, are unlikely to be tested. Since comparisons made among individuals of different ages with similar repeat numbers do not invariably show a correlation of age with mutation frequency or mean change in repeat number (Table 1), it is likely that additional factors which vary among individuals play a role.

Effect of genetic background on instability

Contributions to individual variation in mutation spectra may come from individual differences in genetic background. At the HD locus this variation has been proposed to result from '*cis*' factors (72). We also detect individual variation but can disregard the role of sequences tightly linked to the HD disease allele in our study group. All the HD chromosomes we examined derive from the same founder chromosome about eight generations (~200 years) ago (18).

Recent analysis of instability in MJD patients reported that a single nucleotide polymorphism on the normal MJD allele influences the instability of the disease-causing chromosome (5). As noted earlier we did not detect any influence on HD instability depending on whether the donors were homozygous for the $(CCG)_7$ allele or were $(CCG)_7/(CCG)_{10}$ heterozygotes. We conclude that there is no detectable interallelic effect on HD instability in our data due to this polymorphism.

We note that HD P and his son HD A each have exceptionally high mean changes in allele size when compared with individuals matched for age and repeat number (Table 1). This is suggestive of a familial influence. Familial influences on repeat instability at the FMR1 locus have been reported, but whether the effect is due to a linked or unlinked gene(s) is not known (6,8). Candidates for genes with alleles that might influence repeat instability include those involved in replication or DNA repair, including, but not limited to, classical mismatch repair. Indeed, human polymorphisms have recently been detected in a significant number of DNA repair genes (73).

Factors involved in instability may be identified using genetic approaches adapted for the detection of quantitative trait loci. Ideally, a simple assay that does not depend on single genome assays yet provides quantitative information on instability is desirable. If, in a small number of cases, the discrete data generated by single genome analysis can be compared with the data obtained from the simpler total sperm DNA PCR assay (13), the relationship between the two different read-outs of instability could be determined. This would provide the experimental basis for a large-scale genetic analysis of instability.

MATERIALS AND METHODS

Single sperm isolation and preparation and PCR

Single sperm isolation and preparation, PCR and product analysis were performed as described (19), unless otherwise noted. First round PCR was designed so as to amplify five independent loci simultaneously. In addition to the HD and D4S127 loci, primers for the DM, SCA-1 and SBMA loci were included in the first round PCR reaction. All primers were at a final concentration of 0.5 μM, with the exception of the D4S127 locus, which were at 0.05 μM each. Primer IT2-B (5′-TCACGGTCGGTGCAGCGG-CTCCT) was used in place of IT2 at the HD locus. In the cases where external genomic contamination was suspected, second round PCRs were performed at the non-HD linked loci. For second round PCR, cycle number varied at the HD locus from 21 to 30 and 35, depending on the donor. Twenty three cycles were carried out at the D4S127 locus.

Reliability of the sperm typing data

We have shown previously that the well-known PCR stutter that occurs during the amplification of microsatellite repeats does not significantly affect the estimates of individual allele sizes (19). Mutations in HD allele size are not due to CGG expansions in the CGG repeat region adjacent to the CAG repeat tract (19). An excellent agreement between sperm typing data (on single individuals with HD and SBMA disease alleles) and the available data on paternal transmissions studied in families has been shown previously (19,74,75).

Statistical analysis

The hypothesis that parental origin of the HD allele has no effect on stability can be examined using permutation tests. In the data, seven individuals inherited the HD allele from their father. The average somatic allele size was 48.1 repeats, their average mutation frequency was 90% and the average mutation expansion size in the sperm was 11.23 repeats (Table 1). We compared these averages with those from 10 000 random selections of seven individuals from the 25 individuals we sampled. The random selections result in an empirical distribution of the statistics under the null hypothesis that every selection is equally likely. To examine the effect of parental origin on stability we compared the observed values of the statistics for the allocation actually seen in the data with the values seen in the random selections. For example, 18% of simulated average somatic allele sizes exceeded the observed value of 48.1 repeats, suggesting that somatic allele size is not influenced by paternal inheritance. We compared the mutation frequency and mean change in repeat number in the same way. No significant departures were detected (although only 5.4% of the simulated selections had an average mean change in repeat number as large as that observed in the data). We conclude that we cannot detect a major effect of parental imprinting on instability.

A similar approach can be applied to assess interallelic effects on stability. Once more, somatic allele size does not appear to be influenced by the CCG polymorphism. A similar conclusion applies to the average of the mean change in repeat length in the sperm (for example, 7.3% of the 10 000 simulated samples had a higher average mean change in repeat length than observed in the data).

Probabilistic model

The sperm that were sampled in our experiments are the end products of a number of spermatogonial mitotic stem cell divisions (together with a small number of divisions during the spermatogenesis cycle). It is estimated that there are 34 mitotic divisions prior to the onset of puberty, during which time the spermatogonial stem cell population has undergone very rapid expansion, resulting in some 10^9 stem cells at puberty. Further, it is estimated that there are 23 stem cell divisions/year after puberty (here assumed to be at age 13) (57). We assume that at each division one daughter cell initiates the spermatogenesis cycle, the other remaining a stem cell. Because of the rapid proliferation of stem cells during the growth phase and the subsequent nature of

replication of these cells, a random sample of sperm is likely to have effectively independent mutational histories (because their most recent common ancestor cell is likely to be close to the time of the first differentiated stem cells). It is therefore reasonable to treat sperm from a given donor as independent of each other. The number n of stem cell divisions through which a sperm from a donor of age a has gone may be calculated from the formula

$$n = 34 + 23(a - 13) \quad \mathbf{1}$$

We assume the same value of n for each sperm from a given donor.

We assume that expansions and contractions in repeat numbers arise primarily during the n mitotic divisions before meiosis. We postulate two mutation mechanisms at work, which are asymmetric with respect to the two strands of DNA. One mechanism (Type I) adds or deletes a single triplet during copying of the leading strand. The second (Type II) is responsible for larger additions in repeats and occurs only during lagging strand synthesis, as described in the earlier section on Okazaki fragment formation. To model the effects of mutation through the cell line leading to a given sperm, we must first follow a randomly chosen cell from its initiation as a stem cell though to puberty and from there along the stem cell 'backbone' to the mature sperm. We then model the length of the repeat tract on a randomly chosen strand through this cell lineage.

Because we model an asymmetric mutation process we must keep track of which strands being synthesized are leading or lagging strands. We label strands (i,x); $i = 0$ if they will be the template for lagging strand synthesis and $i = 1$ if they will be the template for leading strand synthesis. x gives the number of repeats on the strand. The probability that a mutation will result in y repeats when a $(0,x)$ strand is copied is $r_0(x,y)$. For a $(1,x)$ strand, the corresponding probability is $r_1(x,y)$.

The succession of states (I_0,X_0), …, (I_n,X_n) forms a Markov chain, starting from $(I_0,X_0) = (0,L)$ with probability 1/2 or $(1,L)$ with probability 1/2. Let $P_n[(i,x)\,(j,y)]$ denote the n step transition probabilities of this chain; these give the probability that a molecule starting with type (i,x) produces a molecule of type (j,y) after n replications. The probability $q(v)$ that the strand at the nth generation is of length v is

$$q(v) = 1/2\{P_n[(0,L)\,(0,v)] + P_n[(0,L)\,(1,v)]\} + 1/2\{P_n[(1,L)\,(0,v)] + P_n[(1,L)\,(1,v)]\} \quad \mathbf{2}$$

for $v = 1, 2, 3, \ldots$.

A model for expansions and contractions

It remains to determine the transition probabilities $r_0(x,y)$ and $r_1(x,y)$. We begin with the simpler case of leading strand synthesis. We suppose that triplets are copied independently of one another. Replication of a triplet on a leading strand results in a slippage mutation with probability p_S. This mutation results in gain or loss of a single triplet (a Type I mutation), each with probability 1/2. The triplet is copied without error with probability $1 - p_S$. The probability distribution $\{r_0(x,y), y = 0, 1, 2, \ldots\}$ can be found by convoluting x times the distribution of the number of repeats added or subtracted when copying a single triplet.

Next we describe how $r_1(x,y)$ may be computed. Okazaki fragments have a length distribution that has been determined empirically (56). Using this distribution and the assumption that the origin of replication is a long way from the triplet repeat region, we first approximate the distribution of the length z of the overlap between an Okazaki fragment and the start of the triplet repeat region. If $z > x$, the Okazaki fragment does not initiate in the repeat region and the length of the repeat region remains x. On the other hand, if $x \geq z$ an Okazaki fragment ends in the repeat region and there is the possibility of flap formation by the next fragment. This fragment can displace an amount D, which must be between 1 and z triplets in length, of the Okazaki fragment that ends in the repeat region. We model the distribution of the displacement length D as a geometric random variable with probability p_D: the chance that $D = k$ is proportional to p_D^k, the constant of proportionality being determined by the fact that D must be between 1 and z. The two limiting cases are $p_D = 0$ (corresponding to a potential displacement of exactly one repeat) and $p_D = 1$ (corresponding to a potential displacement being uniformly distributed between 1 and z repeats).

This choice of model is predicated on the assumption that displacements are more likely to be short than long. Such a displacement is incorporated into the nascent strand with probability p_L, resulting in a repeat region of length $y = x + k$. With probability $1 - p_L$, the displacement is not incorporated and the repeat region remains of length x. For repeat tracts of the size seen in these sperm, it is unlikely that more than one Okazaki fragment will end in a repeat region, so we assume that at most one will. The probabilities $r_1(x,y)$ may now be computed numerically. Finally, from the values of $r_0(x,y)$ and $r_1(x,y)$ we can construct numerically the one step transition matrix P and from that the n step matrix $P_n = P^n$, the nth power of P.

Estimation of parameters

For each data set we know the age a of the donor and his somatic allele length L. We can therefore determine the number of stem cell divisions using equation **1**. The probability distribution of allele sizes can be determined from equation **2** and the model in the previous section. We assume that allele sizes from a given donor are independent and identically distributed copies of the distribution $q(v)$, $v = 0, 1, \ldots$ in equation **2**. In order to estimate the parameters p_S, p_D and p_L, we use the method of maximum likelihood (76). For a given donor whose sample has n_i copies of sperm with i repeats, the log likelihood $l(p_S, p_D, p_L)$ is (up to a constant independent of the parameters)

$$l(p_S, p_D, p_L) = \Sigma_i\, n_i \log q(i)$$

and this expression is maximized numerically.

As noted in the text, there is in general good agreement between the data and the fitted model. For somatic alleles of at least 50 repeats, the fitted models show a symmetrical, bell-shaped estimated mutation spectrum, in stark contrast to the smaller allele sizes. The fits are adequate for all but donor P, who produced a very broad range of sperm allele sizes that is not captured well in our fitted model. To assess these fits further, we used conventional goodness-of-fit tests and we also simulated sperm samples for the larger somatic allele sizes using the estimated parameter values and compared the simulated mutation spectra with those obtained in the data. There was no systematic lack-of-fit revealed by the simulations.

There are clearly some sperm that appear to have anomalous size compared with the mass of the data. For example, donor Y has a sperm of length 118. In order to assess the sensitivity of the parameter estimates on such an outlying data point, we refitted the

model with that measurement removed. This resulted in new estimates of $p_S = 2.5 \times 10^{-4}$, $p_D = 0.50$ and $p_L = 3.7 \times 10^{-3}$. The parameter p_D changed most in absolute terms, from an initial estimate of $p_D = 0.76$. The other parameter estimates were less influenced by the potential outlier.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grants R37 GM37645 (N.A.), NS16367 (J.F.G.), NS22031 (N.W.), NS32765 (M.M.) and NS16367 (M.M.) and by grant BIR 95-04393 (S.T. and P.M.) from the National Science Foundation. E.P.L. was partially supported by a grant from the HDF.

REFERENCES

1. Wells, R.D., Warren, S.T. and Sarmiento, M. (1998) *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, San Diego, CA.
2. Gusella, J.F. and MacDonald, M.E. (1996) Trinucleotide instability: a repeating theme in human inherited disorders. *Annu. Rev. Med*., **47**, 201–209.
3. Ashley, C.T. and Warren, S.T. (1995) Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet*., **29**, 703–728.
4. Sutherland, G.R. and Richards, R.I. (1995) Simple tandem DNA repeats and human genetic disease. *Proc. Natl Acad. Sci. USA*, **92**, 3636–3641.
5. Igarashi, S., Takiyama, Y., Cancel, G., Rogaeva, E.A., Sasaki, H., Wakisaka, A., Zhou, Y.X., Takano, H., Endo, K. *et al*. (1996) Intergenerational instability of the CAG repeat of the gene for Machado–Joseph Disease (MJD1) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat. *Hum. Mol. Genet*., **5**, 923–932.
6. Nolin, S.L., Lewis, F.A., Ye, L.L., Houck, G.E., Glicksman, A.E., Limprasert, P., Li, S.Y., Zhong, N., Ashley, A.E. *et al*. (1996) Familial transmission of the FMR1 CGG repeat. *Am. J. Hum. Genet*., **59**, 1252–1261.
7. Goldberg, Y.P., Kremer, B., Andrew, S.E., Theilmann, J., Graham, R.K., Squitieri, F., Telenius, H., Adam, S., Sajoo, A. *et al*. (1993) Molecular analysis of new mutations for Huntington's disease—intermediate alleles and sex of origin effects. *Nature Genet*., **5**, 174–179.
8. Murray, A., Macpherson, J.N., Pound, M.C., Sharrock, A., Youings, S.A., Dennis, N.R., McKechnie, N., Linehan, P., Morton, N.E. *et al*. (1997) The role of size, sequence and haplotype in the stability of FraxA and FraxE alleles during transmission. *Hum. Mol. Genet*., **6**, 173–184.
9. Norremolle, A., Sorensen, S.A., Fenger, K. and Hasholt, L. (1995) Correlation between magnitude of CAG repeat length alterations and length of the paternal repeat in paternally inherited Huntington's disease. *Clin. Genet*., **47**, 113–117.
10. Lucotte, G., Gerard, N., Aouizerate, A., Loirat, F. and Hazout, S. (1997) Patterns of meiotic variability of the $(CAG)_n$ repeat in the Huntington disease gene. *Genet. Couns*., **8**, 77–81.
11. Zuhlke, C., Riess, O., Bockel, B., Lange, H. and Thies, U. (1993) Mitotic stability and meiotic variability of the $(CAG)_n$ repeat in the Huntington disease gene. *Hum. Mol. Genet*., **2**, 2063–2067.
12. Kremer, B., Almqvist, E., Theilmann, J., Spence, N., Telenius, H., Goldberg, Y.P. and Hayden, M.R. (1995) Sex-dependent mechanisms for expansions and contractions of the CAG repeat on affected Huntington disease chromosomes. *Am. J. Hum. Genet*., **57**, 343–350.
13. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., Gray, J., Conneally, P., Young, A., Penney, J., Hollingsworth, Z., Shoulson, I., Lazzarini, A., Falek, A., Koroshetz, W., Sax, D., Bird, E., Vonsattel, J., Bonilla, E., Alvir, J., Bickham Conde, J. *et al*. (1993) Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature Genet*., **4**, 387–392.
14. De Rooij, K.E., De Koning Gans, P.A.M., Skraastad, M.I., Belfroid, R.D.M., Vegter-Van Der Vlis, M., Roos, R.A.C., Bakker, E., Van Ommen, G.J.B., Den Dunnen, J.T. *et al*. (1993) Dynamic mutation in Dutch Huntington's disease patients: increased paternal repeat instability extending to within the normal size range. *J. Med. Genet*., **30**, 996–1002.
15. Novelletto, A., Persichetti, F., Sabbadini, G., Mandich, P., Bellone, E., Ajmar, F., Pergola, M., Del, S.-L., MacDonald, M.E., Gusella, J.F. *et al*. (1994) Analysis of the trinucleotide repeat expansion in Italian families affected with Huntington disease. *Hum. Mol. Genet*., **3**, 93–98.
16. Trottier, Y., Biancalana, V. and Mandel, J.L. (1994) Instability of CAG repeats in Huntington's disease: relation to parental transmission and age of onset. *J. Med. Genet*., **31**, 377–382.
17. MacDonald, M.E., Barnes, G., Srinidhi, J., Duyao, M.P., Ambrose, C.M., Myers, R.H., Gray, J., Conneally, P.M., Young, A., Penney, J., Shoulson, I., Hollingsworth, Z., Koroshetz, W., Bird, E., Vonsattel, J.P., Bonilla, E., Moscowitz, C., Penchaszadeh, G., Brzustowicz, L., Alvir, J., Bickham Conde, J., Cha, J.-H., Dure, L., Gomez, F. *et al*. (1993) Gametic but not somatic instability of CAG repeat length in Huntington's disease. *J. Med. Genet*., **30**, 982–986.
18. Wexler, N.S., Rose, E.A. and Housman, D.E. (1991) Molecular approaches to hereditary diseases of the nervous system: Huntington's disease as a paradigm. *Annu. Rev. Neurol*., **14**, 503–529.
19. Leeflang, E.P., Zhang, L., Tavare, S., Hubert, R., Srinidhi, J., Macdonald, M.E., Myers, R.H., Deyoung, M., Wexler, N.S. *et al*. (1995) Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum. *Hum. Mol. Genet*., **4**, 1519–1526.
20. Telenius, H., Kremer, B., Goldberg, Y.P., Theilmann, J., Andrew, S.E., Zeisler, J., Adam, S., Greenberg, C., Ives, E.J., Clarke, L.A. *et al*. (1994) Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nature Genet*., **6**, 409–414. [Erratum, *Nature Genet*., 1994, **7**, 113.]
21. Andrew, S.E., Goldberg, Y.P., Theilmann, J., Zeisler, J. and Hayden, M.R. (1994) A CCG repeat polymorphism adjacent to the CAG repeat in the Huntington disease gene, implications for diagnostic accuracy and predictive testing. *Hum. Mol. Genet*., **3**, 65–67.
22. Rubinsztein, D.C., Leggo, J., Barton, D.E. and Ferguson-Smith, M.A. (1993) Site of (CCG) polymorphism in the HD gene. *Nature Genet*., **5**, 214–215.
23. Gordenin, D.A., Kunkel, T.A. and Resnick, M.A. (1997) Repeat expansion—all in a flap? *Nature Genet*., **16**, 116–118.
24. Schweitzer, J.K. and Livingston, D.M. (1998) Expansions of CAG repeat tracts are frequent in a yeast mutant defective in Okazaki fragment maturation. *Hum Mol. Genet*., **7**, 69–74.
25. Levinson, G. and Gutman, G.A. (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res*., **15**, 5323–5338.
26. Henderson, S.T. and Petes, T.D. (1992) Instability of simple sequence DNA in *Saccharomyces cerevisiae*. *Mol. Cell. Biol*., **12**, 2749–2757.
27. Bacolla, A., Bowater, R.P. and Wells, R.D. (1998) In Wells, R.D., Warren, S.T. and Sarmiento, M. (eds), *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, San Diego, CA, p. 829.
28. Freudenreich, C.H., Kantrow, S.M. and Zakian, V.A. (1998) Expansion and length-dependent fragility of CTG repeats in yeast. *Science*, **279**, 853–856.
29. Miret, J.J., Pessoa-Brandao, L. and Lahue, R.S. (1997) Instability of CAG and CTG trinucleotide repeats in *Saccharomyces cerevisiae*. *Mol. Cell. Biol*., **17**, 3382–3387.
30. Kolodner, R.D. and Alani, E. (1994) Mismatch repair and cancer susceptibility. *Curr. Opin. Biotechnol*., **5**, 585–594.
31. Modrich, P. and Lahue, R. (1996) Mismatch repair in replication fidelity, genetic recombination and cancer biology. *Annu. Rev. Biochem*., **65**, 101–133.
32. Aaltonen, L.A., Peltomaki, P., Leach, F.S., Sistonen, P., Pylkkanen, L., Mecklin, J.P., Jarvinen, H., Powell, S.M., Jen, J., Hamilton, S.R. *et al*. (1993) Clues to the pathogenesis of familial colorectal cancer. *Science*, **260**, 812–816.
33. 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.
34. Peltomaki, P., Aaltonen, L.A., Sistonen, P., Pylkkanen, L., Mecklin, J.P., Jarvinen, H., Green, J.S., Jass, J.R., Weber, J.L., Leach, F.S. *et al*. (1993) Genetic mapping of a locus predisposing to human colorectal cancer. *Science*, **260**, 810–812.
35. Tishkoff, D.X., Filosi, N., Gaida, G.M. and Kolodner, R.D. (1997) A novel mutation avoidance mechanism dependent on *S. cerevisiae* RAD27 is distinct from DNA mismatch repair [see comments]. *Cell*, **88**, 253–263.
36. Kokoska, R.J., Stefanovic, L., Tran, H.T., Resnick, M.A., Gordenin, D.A. and Petes, T.D. (1998) Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (rad27) and DNA polymerase delta (pol3-t). *Mol. Cell. Biol*., **18**, 2779–2788.
37. Johnson, R.E., Kovvali, G.K., Prakash, L. and Prakash, S. (1998) Role of yeast Rth1 nuclease and its homologs in mutation avoidance, DNA repair and DNA replication. *Curr. Genet*., **34**, 21–29.

38. Murante, R.S., Rust, L. and Bambara, R.A. (1995) Calf 5′ to 3′ exo/endonuclease must slide from a 5′ end of the substrate to perform structure-specific cleavage. *J. Biol. Chem.*, **270**, 30377–30383.
39. Lieber, M.R. (1997) The FEN-1 family of structure-specific nucleases in eukaryotic DNA replication, recombination and repair. *Bioessays*, **19**, 233–240.
40. Shriver, M.D., Jin, L., Chakraborty, R. and Boerwinkle, E. (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics*, **134**, 983–993.
41. Valdes, A.M., Slatkin, M. and Freimer, N.B. (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics*, **133**, 737–749.
42. Richards, R.I. and Sutherland, G.R. (1994) Simple repeat DNA is not replicated simply. *Nature Genet.*, **6**, 114–116.
43. Eichler, E.E., Holden, J.J.A., Popovich, B.W., Reiss, A.L., Snow, K., Thibodeau, S.N., Richards, C.S., Ward, P.A. and Nelson, D.L. (1994) Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nature Genet.*, **8**, 88–94.
44. Kunst, C.B. and Warren, S.T. (1994) Cryptic and polar variation of the fragile-X repeat could result in predisposing normal alleles. *Cell*, **77**, 853–861.
45. Trinh, T.Q. and Sinden, R.R. (1991) Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*. *Nature*, **352**, 544–547.
46. Veaute, X. and Fuchs, R.P. (1993) Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand. *Science*, **261**, 598–600.
47. Gacy, A.M., Goellner, G., Juranic, N., Macura, S. and McMurray, C.T. (1995) Trinucleotide repeats that expand in human disease form hairpin structures in-vitro. *Cell*, **81**, 533–540.
48. Mitas, M., Yu, A., Dill, J., Kamp, T.J. and Chambers, E.J. (1995) Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat $(CTG)_{15}$. *Nucleic Acids Res.*, **23**, 1050–1059.
49. Mariappan, S.V.S., Garcia, A.E. and Gupta, G. (1996) Structure and dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy. *Nucleic Acids Res.*, **24**, 775–783.
50. Mitchell, J.E., Newbury, S.F. and McClellan, J.A. (1995) Compact structures of $d(CNG)_n$ oligonucleotides in solution and their possible relevance to fragile-X and related human genetic diseases. *Nucleic Acids Res.*, **23**, 1876–1881.
51. Petruska, J., Arnheim, N. and Goodman, M.F. (1996) Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucleic Acids Res.*, **24**, 1992–1998.
52. Pearson, C.E. and Sinden, R.R. (1996) Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile-X loci. *Biochemistry*, **35**, 5041–5053.
53. Smith, G.K., Jie, J., Fox, G.E. and Gao, X.L. (1995) DNA CTG triplet repeats involved in dynamic mutations of neurologically related gene sequences form stable duplexes. *Nucleic Acids Res.*, **23**, 4303–4311.
54. Darlow, J.M. and Leach, D.R. (1995) The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in *Escherichia coli* suggest hairpin folding preferences *in vivo*. *Genetics*, **141**, 825–832.
55. Moore, H., Greenwell, P.W., Liu, C.-P., Arnheim, N. and Petes, T.D. (1998) Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc. Natl Acad. Sci. USA*, in press.
56. Burhans, W.C., Vassilev, L.T., Caddle, M.S., Heintz, N.H. and DePamphilis, M.L. (1990) Identification of an origin of bidirectional DNA replication in mammalian chromosomes. *Cell*, **62**, 955–965.
57. Drost, J.B. and Lee, W.R. (1995) Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among *Drosophila*, mouse and human. *Environ. Mol. Mutagen.*, **25**, 48–64.
58. Bat, O., Kimmel, M. and Axelrod, D.E. (1997) Computer simulation of expansions of DNA triplet repeats in the fragile X syndrome and Huntington's disease. *J. Theor. Biol.*, **188**, 53–67.
59. Devys, D., Biancalana, V., Rousseau, F., Boue, J., Mandel, J.L. and Oberle, I. (1992) Analysis of full fragile X mutations in fetal tissues and monozygotic twins indicate that abnormal methylation and somatic heterogeneity are established early in development. *Am. J. Med. Genet.*, **43**, 208–216.
60. Wohrle, D., Hennig, I., Vogel, W. and Steinbach, P. (1993) Mitotic stability of fragile X mutations in differentiated cells indicates early post-conceptional trinucleotide repeat expansion [see comments]. *Nature Genet.*, **4**, 140–142.
61. Malter, H.E., Iber, J.C., Willemsen, R., de Graaff, E., Tarleton, J.C., Leisti, J., Warren, S.T. and Oostra, B.A. (1997) Characterization of the full fragile X syndrome mutation in fetal gametes. *Nature Genet.*, **15**, 165–169.
62. Taghian, D.G., Hough, H. and Nickoloff, J.A. (1998) Biased short tract repair of palindromic loop mismatches in mammalian cells. *Genetics*, **148**, 1257–1268.
63. Harley, H.G., Rundle, S.A., MacMillan, J.C., Myring, J., Brook, J.D., Crow, S., Reardon, W., Fenton, I., Shaw, D.J. and Harper, P.S. (1993) Size of the unstable CTG repeat sequence in relation to phenotype and parental transmission in myotonic dystrophy. *Am. J. Hum. Genet.*, **52**, 1164–1174.
64. Barcelo, J.M., Mahadevan, M.S., Tsilfidis, C., MacKenzie, A.E. and Korneluk, R.G. (1993) Intergenerational stability of the myotonic dystrophy protomutation. *Hum. Mol. Genet.*, **2**, 705–709.
65. Freudenreich, C.H., Stavenhagen, J.B. and Zakian, V.A. (1997) Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. *Mol. Cell. Biol.*, **17**, 2090–2098.
66. Maurer, D.J., Ocallaghan, B.L. and Livingston, D.M. (1996) Orientation dependence of trinucleotide CAG repeat instability in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **16**, 6617–6622.
67. Schweitzer, J.K. and Livingston, D.M. (1997) Destabilization of CAG trinucleotide repeat tracts by mismatch repair mutations in yeast. *Hum. Mol. Genet.*, **6**, 349–355.
68. Steinbach, P., Wohrle, D., Glaser, D. and Vogel, W. (1998) In Wells, R.D., Warren, S.T. and Sarmiento, M. (eds), *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, San Diego, CA, p. 829.
69. Wierdl, M., Dominska, M. and Petes, T.D. (1998) Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics*, in press.
70. Wong, L.J., Ashizawa, T., Monckton, D.G., Caskey, C.T. and Richards, C.S. (1995) Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. *Am. J. Hum. Genet.*, **56**, 114–122.
71. Monckton, D.G., Wong, L.-J.C.A.T. and Casky, C.T. (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analysis. *Hum. Mol. Genet.*, **4**, 1–8.
72. Telenius, H., Almqvist, E., Kremer, B., Spence, N., Squitieri, F., Nichol, K., Grandell, U., Starr, E., Benjamin, C., Castaldo, I. *et al.* (1995) Somatic mosaicism in sperm is associated with intergenerational $(CAG)_n$ changes in Huntington disease. *Hum. Mol. Genet.*, **4**, 189–195. [Erratum, *Hum. Mol. Genet.*, 1995, **4**, 974.]
73. Shen, M.R., Jones, I.M. and Mohrenweiser, H. (1998) Nonconservative amino acid substitution variants exist at polymorphic frequency in DNA repair genes in healthy humans. *Cancer Res.*, **58**, 604–608.
74. Zhang, L., Leeflang, E.P., Yu, J. and Arnheim, N. (1994) Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. *Nature Genet.*, **7**, 531–535.
75. Zhang, L., Fischbeck, K.H. and Arnheim, N. (1995) CAG repeat length variation in sperm from a patient with Kennedy's disease. *Hum. Mol. Genet.*, **4**, 303–305.
76. Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco, CA.