Segregation Distortion of the CTG Repeats at the Myotonic Dystrophy Locus

Ranajit Chakraborty,[1] David N. Stivers,[1] Ranjan Deka,[2] Ling M. Yu,[2] Mark D. Shriver,[2] and Robert E. Ferrell[2]

[1]Human Genetics Center, University of Texas Houston Health Science Center, Houston; and [2]Department of Human Genetics, The University of Pittsburgh Graduate School of Public Health, Pittsburgh

Summary

Myotonic dystrophy (DM), an autosomal dominant neuromuscular disease, is caused by a CTG-repeat expansion, with affected individuals having $\geqslant$50 repeats of this trinucleotide, at the DMPK locus of human chromosome 19q13.3. Severely affected individuals die early in life; the milder form of this disease reduces reproductive ability. Alleles in the normal range of CTG repeats are not as unstable as the $(CTG)_{\geqslant 50}$ alleles. In the DM families, anticipation and parental bias of allelic expansions have been noted. However, data on mechanism of maintenance of DM in populations are conflicting. We present a maximum-likelihood model for examining segregation distortion of CTG-repeat alleles in normal families. Analyzing 726 meiotic events in 95 nuclear families from the CEPH panel pedigrees, we find evidence of preferential transmission of larger alleles (of size $\leqslant$29 repeats) from females (the probability of transmission of larger alleles is .565 ± 0.03, different from .5 at $P \approx .028$). There is no evidence of segregation distortion during male meiosis. We propose a hypothesis that preferential transmission of larger CTG-repeat alleles during female meiosis can compensate for mutational contraction of repeats within the normal allelic size range, and reduced viability and fertility of affected individuals. Thus, the pool of premutant alleles at the DM locus can be maintained in populations, which can subsequently mutate to the full mutation status to give rise to DM.

Introduction

Myotonic dystrophy (DM), the most common form of adult muscular dystrophy, is an autosomal dominant disease characterized by a high degree of clinical heterogeneity (Harper 1989). At the molecular level, however, DM is a homogeneous disease that maps to the protein kinase (DMPK) gene locus on chromosome 19q13.3 (Whitehead et al. 1982; Stallings et al. 1988; Brook et al. 1992; Fu et al. 1992), where in the 3′ UTR of the DMPK gene, CTG-repeat expansions larger than 50 repeats occur exclusively in affected individuals (Harley et al. 1992; Mahadevan et al. 1992; Yamagata et al. 1992; Mulley et al. 1993). Molecular analysis of the DMPK gene has revealed that the DM-associated expanded CTG-repeats are of Eurasian origin (Harley et al. 1992; Yamagata et al. 1992; Mahadevan et al. 1993*a, b*), although a different pathway of the origin of DM in families of African descent has also been noted (Krahe et al. 1995).

Epidemiological data on DM occurrences in populations of different origins show that the current disease incidence is globally quite variable, with the highest frequency ($\approx$1/8,000) in Western Europe and North American whites (Harper 1989), the lowest (rare) in Africans (Dada 1973; Krahe et al. 1995), intermediate (1/18,000) in Japan, and slightly lower in other Southeast Asians (Ashizawa and Epstein 1991; Davies et al. 1992). While DM shares the characteristics of anticipation (i.e., increased severity and decreased age at onset in successive generations in disease-prone families) with other trinucleotide expansion–causing neurological disorders (e.g., Huntington disease [HD] and fragile X syndrome), the reduced reproductive fitness within a few generations is a usual pattern unique to DM- and fragile X syndrome–prone families (Carey et al. 1994). Thus, in terms of evolutionary origin of DM and its maintenance in populations, it is interesting to ask how DM, in spite of its reduced fitness in affecteds, can be maintained for hundreds of generations in human populations. This question is even more relevant because, in the normal allele size ranges (i.e., CTG repeats of length $\leqslant$50 repeats) at this locus, mutations leading to contraction of allele sizes have been noted (Zhang et al. 1994), and, furthermore, most of the presently extant DM chromosomes apparently evolved from a small pool of founder haplotypes (Imbert et al. 1993; Neville et al. 1994; Krahe et al. 1995). A recent observation that alleles of CTG-repeat sizes $\geqslant$19 are preferentially transmitted to the children by healthy fathers has been suggested as a possible mechanism of replenishing the premutant DM-

Received October 13, 1995; accepted for publication April 1, 1996.

Address for correspondence and reprints: Dr. Ranajit Chakraborty, Human Genetics Center, The University of Texas Houston Health Science Center, P. O. Box 20334, Houston, TX 77225.

0002-9297/96/5901-0016$02.00

allele pools in a population (Carey et al. 1994). While the observation that the DM alleles are preferentially transmitted to sons in disease-prone families (Gennarelli et al. 1994) lends further support to the meiotic-drive hypothesis of Carey et al. (1994), these observations together do not totally explain how all CTG repeat sizes below the full mutation range (i.e., of sizes $\geqslant 50$ repeats) can be maintained in a population. More recently, Hurst et al. (1995) contested the statistical validity of the meiotic-drive hypothesis by claiming that the observed segregation distortion may be an artifact of testing multiple hypotheses from the same data.

The purpose of this research is to present new data on this subject and to suggest a simpler procedure for testing the hypothesis of segregation distortion at the DM CTG-repeat locus. To address this question, we investigated CTG-repeat size transmissions to children in 40 CEPH panel pedigrees by molecular analysis, which allows us to analyze segregation distortions in a more detailed manner. In addition to testing whether the $(CTG)_{\geqslant 19}$ alleles are preferentially transmitted during male and female meiosis, we asked a more general question with regard to segregation distortion, namely, whether the heterozygous parents (of each sex) transmit their alleles with equal probability. Thus, for all mothers and fathers, the alleles were scored as long and short, irrespective of the absolute size of the CTG repeat, so that the hypothesis of segregation distortion translates into p_f and/or p_m (the probability of transmitting the longer allele from fathers and mothers, respectively) significantly different from .5. A maximum-likelihood method of parameter estimation is suggested from the entire data, from which the most parsimonious model was selected by using the Akaike criterion, AIC (Akaike 1974), starting from the most general model ($0 \leqslant p_m \neq p_f \leqslant 1$) to the simplest one ($p_m = .5 = p_f$). Comparative evaluations of the likelihood ratios of these models indicate that in the CEPH-panel pedigrees, within the normal size ranges of the CTG alleles at the DMPK locus, mothers transmit the longer alleles preferentially, while in paternal transmissions no segregation distortion is observed. In contrast, these families do not exhibit any preferential transmission of parental alleles when the CTG-repeat sizes are classified into classes such as 5, 11–15, <19, and $\geqslant$19, as done by Carey et al. (1994). Finally, we discuss the relevance of our findings in relation to maintenance of DM in populations, in the light of worldwide distributions of normal CTG-repeat sizes reported elsewhere (Deka et al. 1996).

Material and Methods

Data and Laboratory Methods

Transformed DNA cell lines from parents and children of 40 CEPH-panel pedigrees (Dausset et al. 1990) from the repository were utilized in this study. CTG repeats were determined by amplification of 100 ng of DNA in a total volume of 25 μl reaction mixture containing standard PCR buffer, 200 μM each dNTP, 1 unit of *Taq* polymerase. The primer sequences are as described in Fu et al. (1992). The forward primer was end-labeled using [γ^{33} P]ATP and polynucleotide kinase T4. The amplified products were separated on 6% denaturing polyacrylamide gels. Following electrophoresis, the gels were dried, and allelic fragments were visualized by autoradiography. Repeat sizes were determined by comparison to an M13 sequence ladder and control samples. This genotyping procedure yielded data on segregation of CTG-repeat alleles in 367 children from 95 nuclear families from the 40 CEPH-panel pedigrees. Table 1 lists the sample sizes for each category (explained in the next section) of nuclear families, along with the number of pedigrees contributing from which these families are derived.

Statistical Analysis

For the purpose of estimating preferential transmission of alleles, we considered a model where a heterozygous parent is labeled having a short and long allele, when the alleles in the parental genotype showed two CTG-repeat sizes. Let p_f and p_m denote the transmission probabilities for the (comparatively) longer allele from the father and mother, respectively. With these type of groupings of genotype data, there are four different types of nuclear families where (i) only mother is heterozygous, (ii) only father is heterozygous, (iii) both father

Table 1

Distribution of Nuclear Families, by Family Type and Sample Size for Data Analysis

Family Types	No. of Nuclear Families	No. of Children	No. of Pedigrees Found
Both parents genotyped:			
Family type i	14	73	13
Family type ii	18	53	16
Family type iii	8	39	8
Family type iv	46	188	28
Both parents homozygous	1	6	1
Genotype of one parent available:			
Heterozygous mother	4	4	4
Homozygous mother	0	0	0
Heterozygous father	2	2	2
Homozygous father	2	2	2
Total	95	367	75[a]

[a] Of the total 40 pedigrees, several contributed nuclear families to more than one type; this explains the column total >40. The 40 pedigrees have the following structure: 10 with no grandparents typed, 1 with a single grandparent (mother's mother) typed, 4 with 2 grandparents typed (both paternal), 6 with 3 grandparents typed, and 19 with all 4 grandparents typed.

and mother are heterozygous for the same two CTG-repeat alleles (i.e., mother and father share both alleles), and (iv) both father and mother are heterozygous with at least one unshared allele between them. The distinction between the families of types iii and iv is subtle but important for determining which children got the longer allele from the mother as opposed to from the father. For example, when both parents are heterozygous for the $(CTG)_5$ and $(CTG)_{15}$ alleles, we know that both parents transmitted the longer alleles to an offspring of genotype $(CTG)_{15}$-$(CTG)_{15}$. However, in such families we do not know whether the father or the mother transmitted the longer allele to an offspring of genotype $(CTG)_5$-$(CTG)_{15}$. In other words, for nuclear families of type iii, described above, the transmission of short or long alleles from each specific parent to heterozygous children is not unequivocally known. However, barring the unlikely event of two mutations, such children receive the shorter allele from one parent and the longer from the other. For type iv families, all parental transmission of alleles to each offspring are unequivocally known.

The distribution of nuclear families, shown in table 1, indicates that 86 nuclear families can be classified into the above four categories, in which we have information on transmission of shorter versus longer alleles in a total of 353 children. In one nuclear family, both parents were homozygous (for different CTG-repeat alleles), and hence, the six children from this family did not provide any information about the segregation ratio parameters p_f and/or p_m. In addition, we also have data on eight other nuclear families, in each of which only one parent's genotype was known. Of these, the parent genotyped was heterozygous in six families (table 1). While, in principle, corresponding likelihood functions can be written for such incomplete data, for our main analysis, we excluded all such incomplete data (for reasons explained in the Discussion section). However, we also evaluated the effect of inclusion of such data on the parameter estimates (see Results).

In table 2 we present the data and notations of frequencies of transmission of shorter and longer alleles from each parent. The entire sample can then be analyzed by a single likelihood function in relation to the two parameters p_m and p_f, given by

$$\begin{aligned} L = {} & \text{Const} \times p_m^{m_1}(1 - p_m)^{n_1 - m_1} \\ & \times p_f^{m_2}(1 - p_f)^{n_2 - m_2}(p_m p_f)^{m_{31}} \\ & \times [p_m(1 - p_f) + p_f(1 - p_m)]^{m_{32}} \\ & \times [(1 - p_m)(1 - p_f)]^{n_3 - m_{31} - m_{32}}(p_m p_f)^{m_{41}} \\ & \times [p_m(1 - p_f)]^{m_{42}}[p_f(1 - p_m)]^{m_{43}} \\ & \times [(1 - p_m)(1 - p_f)]^{n_4 - m_{41} - m_{42} - m_{43}}, \end{aligned} \qquad (1)$$

from which the traditional maximum-likelihood estimates of p_m and p_f were obtained. For p_m and p_f within the interval 0–1, explicit closed form estimators of these parameters do not exist. However, the likelihood equations $[(\partial \ln L)/\partial p_m] = 0$ and $[(\partial \ln L)/\partial p_f] = 0$ are biquadratic equations, and hence they were solved iteratively to obtain the maximum-likelihood estimates of p_m and p_f, without any restrictions imposed on these segregation ratios. We obtained the standard errors of the estimates by inverting the information matrix, using procedures as described in Rao (1973). Under the subhypotheses (*a*) $0 \leq p_m = p_f \leq 1$; (*b*) $p_m = .5$, $0 \leq p_f \leq 1$; and (*c*) $p_f = .5$, $0 \leq p_m \leq 1$ the above likelihood function takes simpler forms for which explicit solutions of the maximum-likelihood estimates exist. In the appendix, we present these estimators along with their standard errors. By using these estimates, the maximum values of log likelihood were evaluated for each model for selecting the most parsimonious model using Akaike's criterion (Akaike 1974). The significance of deviations of p_m and p_f from the traditional Mendelian segregation ratio were tested by the likelihood-ratio test criterion (Rao 1973).

To examine whether the allelic transmissions of CTG repeats in children of these CEPH-panel pedigrees support the observation of Carey et al. (1994), we also pooled the repeat alleles in classes $(CTG)_5$, $(CTG)_{11-13,15}$, $(CTG)_{<19}$, and $(CTG)_{\geq 19}$. For the specific types (as defined by Carey et al. 1994) of heterozygous parents of each sex, we estimated the probability of transmission of the larger allele to examine whether it deviated significantly from the expected 50% ratio. Following the method of Hurst et al. (1995), the goodness-of-fit χ^2 statistics were computed for testing the departure from the expected 1:1 ratio of segregation of smaller and larger alleles.

Results

Reexamination of the Hypothesis of Carey et al.

In 182 parents and 367 children scored for CTG-repeat sizes, no allele of size >29 CTG repeats was observed in these healthy subjects. Of the 726 meioses, only one mutation, from $(CTG)_{24}$ to $(CTG)_{25}$, from a maternal transmission, was found. This mutation is the same one observed earlier, and confirmed as a true in vivo mutation (Weber and Wong 1993). There were 70 fathers and 68 mothers heterozygous for CTG repeats at this locus. When allelic transmissions from heterozygous parents were examined by grouping CTG-repeat sizes (such as heterozygous for $(CTG)_{<19}$ versus $(CTG)_{\geq 19}$, $(CTG)_5$ versus $(CTG)_{11-13,15}$, and $(CTG)_5$ versus $(CTG)_{\geq 19}$) and by sex of origin of meiosis, we find no significant segregation distortion for any subdivisions of the parental allelic types (panels *A–C*, table 3). In other words, the present data do not confirm the findings of Carey et al. (1994).

Table 2

Frequencies (and Notations) of Transmissions of Longer (L) and Shorter (S) CTG-Repeat Alleles from Parents in 86 Nuclear Families from 40 CEPH Pedigrees

	Type of Family			
	Only Mother Heterozygous (L_mS_m)	Only Father Heterozygous (L_fS_f)	Mother and Father Heterozygous for Same Two Alleles	Mother and Father Heterozygous with at Least One Unshared Allele
No. of families	14 (N_1)	18 (N_2)	8 (N_3)	46 (N_4)
No. of children:				
L_m	37 (m_1)	. . .	. . .	. . .
S_m	36 ($n_1 - m_1$)	. . .	. . .	. . .
L_f	. . .	26 (m_2)	. . .	. . .
S_f	. . .	27 ($n_2 - m_2$)	. . .	. . .
LL	. . .	. . .	21 (m_{31})	. . .
LS	. . .	. . .	13 (m_{32})	. . .
SS	. . .	. . .	5 ($n_3 - m_{31} - m_{32}$)	. . .
L_mL_f	. . .	. . .	. . .	41 (m_{41})
L_mS_f	. . .	. . .	. . .	63 (m_{42})
S_mL_f	. . .	. . .	. . .	45 (m_{43})
S_mS_f	. . .	. . .	. . .	39 ($n_4 - m_{41} - m_{42} - m_{43}$)
Total	73 (n_1)	53 (n_2)	39 (n_3)	188 (n_4)

NOTE.—In addition, there are four nuclear families where the mothers are heterozygous but fathers are untyped and two families with heterozygous fathers and untyped mothers. Pooling them in categories i and ii gives $N_1 = 18$; $m_1 = 39$; $n_1 = 77$, and $N_2 = 20$; $m_2 = 28$; $n_2 = 55$, respectively.

Preferential Transmission of Larger CTG Repeats during Female Meiosis

In contrast, when all heterozygous parents are considered, and the allelic transmissions during each specific male and female meiosis were scored as transmission of shorter (S) and longer (L) alleles (complete data as shown in table 2), the maximum-likelihood estimators of p_m and p_f, their standard errors, and test results based on the likelihood ratios are shown in table 4. The most parsimonious model (based on the likelihood ratio as well as the Akaike's criterion) is when p_f is assumed to be .5 and $\hat{p}_m = .565 \pm .030$ (see table 4), which is significantly different from the null model, which is $p_m = p_f = .5$ ($P \approx .028$). This suggests that in the CEPH pedigrees the larger CTG-repeat sizes are preferentially transmitted during female meiosis at the DMPK locus, and the male meiosis does not indicate any significant segregation distortion.

As mentioned before, the results shown in table 4 do not include data on nuclear families where the genotype of one parent was unknown. When the parent genotyped was heterozygous, we may assume that the unknown parent's genotype does not provide any information regarding the parameter p_m or p_f, so that such families may be pooled with type i or type ii families. While this strategy is not totally justifiable (explained below), with inclusion of such families, there is virtually no change in the conclusions that can be derived from the parameters obtained in table 4. For example, the most parsimonious model still gives parameter estimates $\hat{p}_m = .564 \pm .029$ when p_f is assumed to be .5. Merging of such incompletely typed families with the more complete data did not pose any problem in our sample, since there was no nuclear family where the typed parent and the child were both heterozygous for the same two CTG-repeat alleles.

A further comment regarding the preferential transmission of larger alleles from the mothers (but not from fathers) is also relevant in this context. It is apparent from the data in table 2 that the segregation patterns appear somewhat different for families of types i and ii from those of types iii and iv, since the p_m and p_f estimates are indistinguishably different from .5, when data from only type i and type ii families are used. In contrast, when only type iii and iv families are used, the most general model yields parameter estimates $\hat{p}_m = .584 \pm .034$, and $\hat{p}_f = .495 \pm .035$ with $-\ln L = 302.62$, not significantly different from the parsimonious model $\hat{p}_m = .584 \pm .034$, $p_f = .5$ with $-\ln L = 302.64$. We may also note that of the 40 pedigrees, the nuclear families contributing to data on type iii and type iv families come from 33 different pedigrees. Some of these two-generation pedigrees also contribute to several nuclear families of types i and ii. When data from the remaining seven pedigrees (contributing to family types i and ii only) are excluded, the segregation ratio estimates from type i and type ii families become $\hat{p}_m = 30/54 = .556$, and $\hat{p}_f = 15/31 = .484$, respectively, which are statistically indistinguishable from the respective estimates from the type iii and type iv families derived from the same 33 pedigrees. From these, we may conclude

Table 3

Transmission of CTG Alleles at the DM Locus, Classified by Allele Size of Parents

Sex of Origin of Meiosis	LAS[a]	SAS[b]	Total	% LAS	χ^2_1	P-Value
A. Parent Heterozygous for $(CTG)_{<19}$ and $(CTG)_{\geqslant 19}$ Alleles:						
Female meiosis	30	28	58	51.7	.07	.791
Male meiosis	16	20	36	44.4	.44	.509
Total	46	48	94	48.9	.04	.841
B. Parent Heterozygous for $(CTG)_5$ and $(CTG)_{11-13,15}$ Alleles:						
Female meiosis	58	51	109	53.2	.45	.503
Male meiosis	56	48	104	53.8	.62	.431
Unknown sex	15	15	30	50.0	.0	≈1.0
Total	129	114	243	53.1	.93	.335
C. Parent Heterozygous for $(CTG)_5$ and $(CTG)_{\geqslant 19}$ Alleles:						
Female meiosis	18	19	37	48.6	.03	.862
Male meiosis	2	3	5	40.0	.20	.655
Total	20	22	42	47.6	.10	.752

[a] SAS = number of smalier-size alleles segregating.
[b] LAS = number of larger-size alleles segregating.

that the preferential transmission of larger alleles during female meiosis, as observed in the present sample, is observed in a great majority of the pedigrees (33 of 40 analyzed), and in these pedigrees there is no heterogeneity of parameter estimates in families of types i and ii versus iii and iv.

Discussion and Conclusion

The above observations are different from the ones by Carey et al. (1994). The smaller sample size of our data may have contributed to this difference of conclusions. Nevertheless, our method of testing for sex-of-origin–specific segregation distortion is not compromised by the statistical artifact of multiple testing, noted by Hurst et al. (1995). Several implications of the present findings are noteworthy in relation to the question of maintenance of the expanded CTG repeats in populations.

First, limited data that currently exist suggest the possibility of mutations of CTG repeats at the DM locus being allele size–dependent within the normal size ranges of CTG alleles (Zhang et al. 1994). If the rate of mutation increases with repeat size and contractions outnumber the expansions in mutations within the normal allele size range, a preferential transmission of larger alleles can maintain disease frequency of DM. The equilibrium-frequency distribution of allele sizes within the normal size range, will of course, depend on the magnitude of segregation distortion as well as the extent of contraction bias of mutations. The dynamics of effects of these compensatory factors is somewhat complex and will be reported

Table 4

Maximum-Likelihood Estimates of Segregation Ratio (p_m and p_f) under Different Models

	Parameter Estimates ± SE				
Model	p_m	p_f	$-\ln L$	χ^2	AIC
H_1: $0 \leqslant p_m, p_f \leqslant 1$	.565 ± .030	.495 ± .030	390.61	. . .	785.22
H_{0_1}: $0 \leqslant p_m \leqslant 1$; $p_f = .5$	.565 ± .030	.5	390.62	.02	783.24
H_{0_2}: $0 \leqslant p_f \leqslant 1$; $p_m = .5$	.5	.498 ± .031	393.01	4.80*	788.02
H_{0_3}: $0 \leqslant p_f = p_m \leqslant 1$	.531 ± .021	.531 ± .021	391.10	.98	784.20
H_{0_4}: $p_m = p_f = .5$	.5	.5	393.01	4.80*	786.02

Note.—$\chi^2 = -2(\ln L_{0_i} - \ln L_1)$ where L_{0_i} is the likelihood under the ith null, and L_1 is the likelihood under unrestricted model; AIC $= -2[\ln L_H -$ number of parameters estimated].

* $P < .05$.

elsewhere. However, it has implicit implications for analyzing the incomplete family data for detecting segregation distortion. For example, consider a family where the father is heterozygous for CTG repeats 12 and 14. Suppose that a child of this family has genotype $(CTG)_{12}$-$(CTG)_{16}$. Without data on any further children from this family, we only know that the mother with unknown genotype has at least one copy of the CTG_{16} allele. Conditioned on the three possible maternal genotypes ($[CTG]_{16}$-$[CTG]_{16}$, $[CTG]_{16}$-$[CTG]_{\leqslant 16}$, and $[CTG]_{16}$-$[CTG]_{\geqslant 16}$), under the present formulation, the likelihood functions for this incomplete family data become $(1 - p_f)$, $p_m(1 - p_f)$, and $(1 - p_m)(1 - p_f)$, respectively. To incorporate this family in our estimation procedure, we must multiply these conditional probabilities by maternal genotype frequencies, which, in turn, would depend on the frequencies of $(CTG)_{16}$, $(CTG)_{\leqslant 16}$, and $(CTG)_{\geqslant 16}$ alleles in the population. We cannot substitute arbitrary allele frequencies, since the allele frequencies in a population, under the presence of segregation distortions, are actually dependent on the parameters p_m and p_f as well. An empirical solution to get around this would be to substitute genotype frequencies for the unknown parents by their Hardy-Weinberg expectations (HWE) based on population-allele frequencies. Of course, data on adherence with HWE, based on samples from the same population, are needed before evaluating such likelihood functions empirically. From our genotyping efforts in the present sample, we had 127 unrelated individuals typed, in which the observed genotype frequencies of CTG repeats are shown in table 5. The gene count estimates of CTG-repeat alleles from this data is shown in figure 1. The likelihood ratio for testing concordance with HWE from this data becomes 87.03, for which the permutation-based (Chakraborty et al. 1994) empirical probability is .229, indicating that the assumption of HWE for CTG-repeat genotypes is valid for the present sample. We, therefore, examined whether our initial parameter estimates were substantially altered, when the incomplete family data are incorporated with such empirical evaluation of the likelihood function for the entire data. The results are virtually indistinguishable from the ones shown in table 4, which suggests that our observation on preferential transmission of larger alleles during female meiosis is consistent with the implication that, together with such segregation distortion, there must be one or more compensatory factor (such as contraction bias of mutation and/or selective disadvantage of genotypes with full mutation) acting at the DMPK locus, keeping the normal size allele frequencies at equilibrium in the population. Second, recent observations by others (Krahe et al. 1995; Zerylnick et al. 1995), as well as our own study in worldwide populations (Deka et al. 1996), suggest that there could have been more than a single pathway of generating higher repeat sizes within the normal size range of alleles at this locus. In particular, our analysis (Deka et al. 1996) indicates that a gradual increase of CTG-repeat sizes within the normal size range, as opposed to few sudden jumps from $(CTG)_{\leqslant 5}$ to $(CTG)_{\geqslant 19}$ under the *Alu*-insertion background (as proposed by Imbert et al. 1993) is supported better

Table 5

Observed CTG Genotype Frequencies in 127 Unrelated Individuals from 40 CEPH-Panel Pedigrees

	CTG Repeat								
CTG-Repeat	5	8	10	11	12	13	14	16	22
5	18								
8	...	...							
10	1	...	1						
11	17	...	...	1					
12	11	1	1	3	2				
13	15	...	1	6	3	3			
14	2	2	1	...	2	1	...		
16	7	...	...	2	1	...	...	...	
17	...	...	...	1	...	...	...	...	
18	...	...	...	...	...	...	...	1	
20	2	...	...	1	...	...	2	...	
21	2	...	...	1	1	2	...	...	
22	1	...	1	...	...	...	...	...	1
23	1	...	...	1	...	...	...	...	...
24	2	...	...	...	...	1	...	...	...
25	...	...	...	1	...	1	...	...	...
28	...	...	...	1	...	...	...	...	...
29	1	...	...	...	...	...	...	...	...

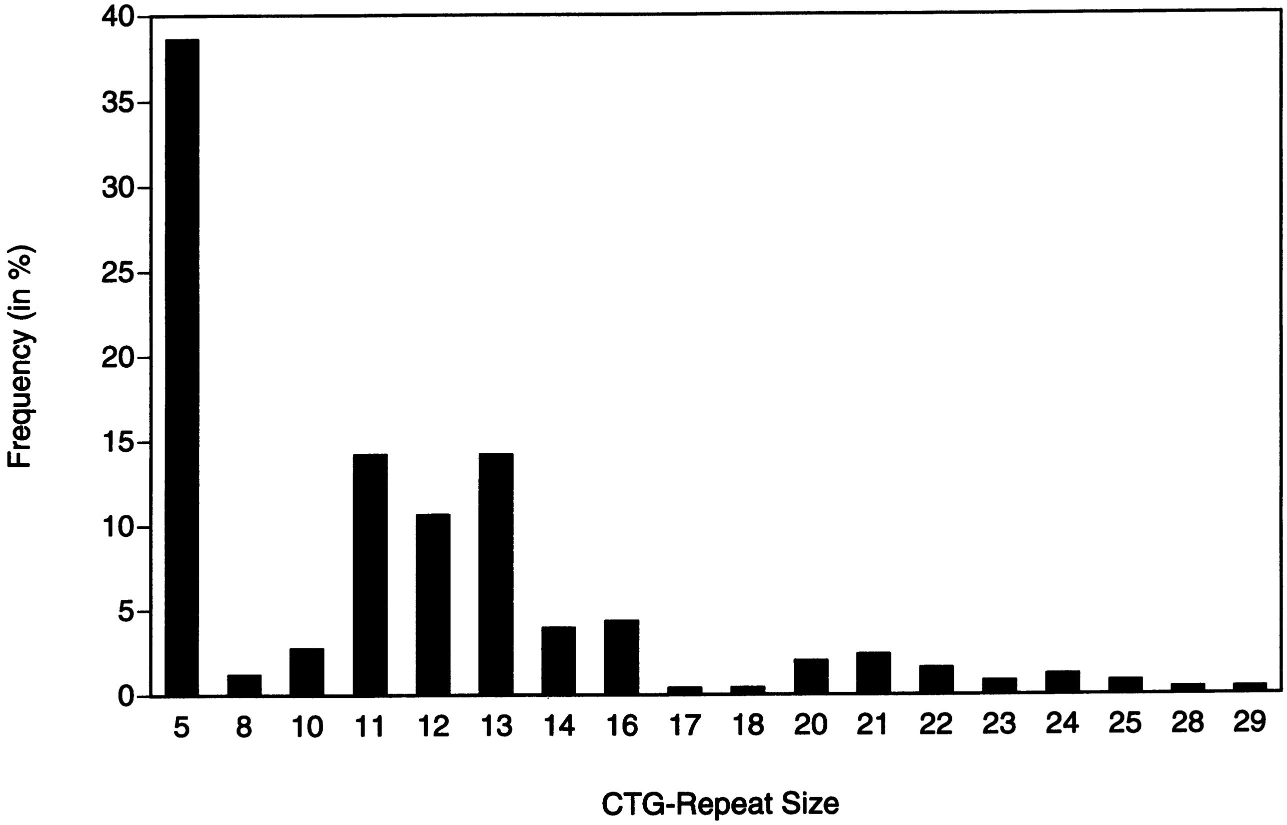

Figure 1 Allele frequency distribution of CTG-repeat alleles at the DMPK locus in 127 unrelated individuals from the CEPH-panel pedigrees.

with worldwide molecular data on the haplotype distributions at the DMPK gene region. Should this be the case, a general meiotic drive toward preferential transmission and/or differential survival of larger-size alleles may be required to counter the opposing force of contracting mutations in the normal size range for generating premutant CTG-repeat sizes. From the above findings we conclude that segregation distortion for larger CTG-repeat alleles over the entire normal range of allele sizes can explain the present data, and that this preferential transmission can maintain DM disease frequency in populations, particularly when the smaller-size alleles are constantly being replenished by contraction mutations occurring at higher rates for large-size alleles. A technical comment on the analysis of Carey et al. (1994) is also worthwhile to note here. They reported 24 transmitted alleles from parents of unknown phase (see table 1 of Carey et al. 1994). Since this refers to families where both parents are heterozygous for the same two alleles and the children are heterozygous, unequal transmissions of large and small alleles cannot be detected where the sex of origin of transmitted alleles is not known. They report 11 small alleles and 13 large alleles from these subgroup of parents, which contradicts the above logic.

Finally, we must recall that analyses such as the ones conducted here cannot determine whether the segregation distortion for larger alleles is due to pre- or postzygotic selection. Our data is also at variance with increased male-to-male transmission of larger-size DM alleles (Gennarelli et al. 1994; Wieringa 1994), found in DM patients. Of course, it is quite possible that the allele-transmission pattern and its mechanism may be different in the normal, premutation, and full-mutation sizes of repeat alleles (Ashizawa et al. 1994; Monckton et al. 1995). Nevertheless, data from other families, particularly in which the haplotypic associations of CTG-repeat alleles are different, are needed for fully characterizing the detailed nature of meiotic drive at this locus, if it truly exists.

Acknowledgments

This research was supported by U.S. Public Health Service research grants GM 45861 and GM 41399 from the National Institutes of Health and by a W. M. Keck Center grant for advanced training in computational biology to the University of Pittsburgh and Carnegie Mellon University. Genotyping of the CTG-repeat polymorphism was possible through the use of transformed cell lines from the CEPH collaborative research program, made available to one of us (R.E.F.).

Appendix

Maximum-Likelihood Estimation for Segregation Distortion Models

The General Model

Define alleles as short (S) and long (L) for any individual heterozygous for the CTG-repeat alleles. Note that the alleles labeled in this manner are only comparative, and, thus, S and L do not refer to the absolute size of CTG-repeat alleles. Let p_m and p_f denote the probability of transmitting L from a mother and a father, respectively. For nuclear family data from four types of families as represented in table 2, the logarithm of the likelihood function for parameters (p_m and p_f) of the general model ($0 \leqslant p_m \neq p_f \leqslant 1$) may be simplified as

$$\begin{aligned}\ln L = C_0 &+ C_1 \ln(p_m) + C_2 \ln(1 - p_m) \\ &+ C_3 \ln(p_f) + C_4 \ln(1 - p_f) \\ &+ C_5 \ln(p_m + p_f - 2p_m p_f)\,,\end{aligned} \tag{A1}$$

where C_0 is a constant, and

$$C_1 = m_1 + m_{31} + m_{41} + m_{42}\,, \tag{A2}$$

$$\begin{aligned}C_2 = (n_1 - m_1) &+ (n_3 - m_{31} - m_{32}) \\ &+ (n_4 - m_{41} - m_{42})\,,\end{aligned} \tag{A3}$$

$$C_3 = C_2 + m_{31} + m_{41} + m_{43}\,, \tag{A4}$$

$$\begin{aligned}C_4 = (n_2 - m_2) &+ (n_3 - m_{31} - m_{32}) \\ &+ (n_4 - m_{41} - m_{43})\,,\end{aligned} \tag{A5}$$

and

$$C_5 = m_{32}\,. \tag{A6}$$

Thus, the maximum-likelihood estimates of p_m and p_f are solutions to the two equations

$$\frac{C_1}{p_m} - \frac{C_2}{1 - p_m} + \frac{(1 - 2p_f)C_5}{p_m + p_f - 2p_m p_f} = 0\,, \tag{A7}$$

and

$$\frac{C_3}{p_f} - \frac{C_4}{1 - p_f} + \frac{(1 - 2p_m)C_5}{p_m + p_f - 2p_m p_f} = 0\,. \tag{A8}$$

Equations (A7) and (A8) are biquadratic equations with two unknown parameters (p_m and p_f). These always yield one set of solutions satisfying $0 \leqslant \hat{p}_m, \hat{p}_f \leqslant 1$, which can be found by solving the equations iteratively. The coefficients $C_1, C_2, \ldots, C_5$ are linear combinations of multinomial variates; therefore, by following the method of Rao (1973), the sampling variances of these maximum likelihood estimators may be written as

$$V(\hat{p}_m) = \frac{I_{ff}}{I_{mm} I_{ff} - I_{mf}^2} \tag{A9}$$

and

$$V(\hat{p}_f) = \frac{I_{mm}}{I_{mm} I_{ff} - I_{mf}^2}\,, \tag{A10}$$

where

$$\begin{aligned}I_{mm} &= -E\left[\frac{\partial^2 \ln L}{\partial p_m^2}\right] \\ &= \frac{n_1 + n_4 + p_f n_3}{p_m} \\ &\quad + \frac{n_1 + n_4 + (1 - p_f)n_3}{1 - p_m} \\ &\quad + \frac{(1 - 2p_f)^2 n_3}{p_m + p_f - 2p_m p_f}\,,\end{aligned} \tag{A11}$$

$$\begin{aligned}I_{ff} &= -E\left[\frac{\partial^2 \ln L}{\partial p_f^2}\right] \\ &= \frac{n_1 + n_4 + p_m n_3}{p_f} \\ &\quad + \frac{n_1 + n_4 + (1 - p_m)n_3}{1 - p_f} \\ &\quad + \frac{(1 - 2p_m)^2 n_3}{p_m + p_f - 2p_m p_f}\,,\end{aligned} \tag{A12}$$

and

$$I_{mf} = -E\left[\frac{\partial^2 \ln L}{\partial p_f \partial p_m}\right] = \frac{n_3}{p_m + p_f - 2p_m p_f}\,. \tag{A13}$$

Thus, the sampling variances of $\hat{p}_m$ and $\hat{p}_f$ can be estimated by substituting the parameters with their respective estimates in equations (A11) and (A12).

Model $0 \leqslant p_m = p_f \leqslant 1$

Under this model, when the common values of the two parameters is denoted by p, the likelihood equation

$$\frac{\partial \ln L}{\partial p} = 0 \tag{A14}$$

yields the closed-form solution

$$\hat{p} = \frac{m_1 + m_2 + 2m_{31} + m_{32} + 2m_{41} + m_{42} + m_{43}}{n_1 + n_2 + 2n_3 + 2n_4}, \tag{A15}$$

whose sampling variance may be estimated from

$$V(\hat{p}) = \frac{p(1 - p)}{n_1 + n_2 + 2n_3 + 2n_4}. \tag{A16}$$

by substituting $\hat{p}$ (eq. A15) for p.

Models with p_m or p_f = .5, and the Other Unknown

Under these models, as in the previous models, the maximum-likelihood estimators have closed form solutions. When p_m or p_f is .5, observation of the number of children with one short and one long allele (m_{32}) in type iii families does not contribute any information regarding the unknown parameter (for either p_m or p_f). Therefore, when $p_f = .5$, the maximum-likelihood estimator of p_m becomes

$$\hat{p}_m = \frac{m_1 + m_{31} + m_{41} + m_{42}}{n_1 + (n_3 - m_{32}) + n_4}, \tag{A17}$$

whose sampling variance equals

$$V(\hat{p}_m) = \frac{2p_m(1 - p_m)}{2n_1 + n_3 + 2n_4}. \tag{A18}$$

Likewise, when $p_m = .5$, the estimator for p_f becomes

$$\hat{p}_f = \frac{m_2 + m_{31} + m_{41} + m_{43}}{n_2 + (n_3 - m_{32}) + n_4}, \tag{A19}$$

with a sampling variance

$$V(\hat{p}_f) = \frac{2p_f(1 - p_f)}{2n_2 + n_3 + 2n_4}. \tag{A20}$$

The likelihood-ratio test criteria for testing the departure from any of these hypotheses can be evaluated by computing $\chi^2 = -2[\ln L - \ln L_0]$, where $\ln L$ is the evaluated likelihood function for the general model and $\ln L_0$ is computed by substituting the estimated parameters under the specific subhypotheses. The degrees of freedom for the respective statistics are the differences of the number of parameters estimated under each subhypothesis from the general model. The model selection is based on the Akaike's criterion (Akaike 1974)

$$\text{AIC} = -2[\ln L - \text{number of parameters}], \tag{A21}$$

which attains the minimum value for the most parsimonious model. The results of numerical evaluations based on these methods for the data shown in table 2 are given in table 4.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Control AC 19:716–723

Ashizawa T, Anvret M, Baiget M, Barceló JM, Brunner H, Cobo AM, Dallapiccola B, et al (1994) Characteristics of intergenerational contraction of the CTG repeat in myotonic dystrophy. Am J Hum Genet 54:414–423

Ashizawa T, Epstein HF (1991) Ethnic distribution of the myotonic dystrophy gene. Lancet 338:642–643

Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, et al (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide CTG repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell 68:799–808

Carey N, Johnson K, Nokelainen P, Peltonen L, Savontaus M-L, Juvonen V, Anvret M, et al (1994) Meiotic drive at the myotonic dystrophy locus? Nat Genet 6:117–118

Chakraborty R, Zhong Y, Jin L, Budowle B (1994) Nondetectability of restriction fragments and independence of DNA fragment sizes within and between loci in RFLP typing of DNA. Am J Hum Genet 55:391–401

Dada TO (1973) Dystrophia myotonica in Nigerian family. East Afr Med J 50:213–228

Dausset J, Conn H, Cohen D, Lathrop M, Lalouel J-M, White R (1990) Center d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575–577

Davies J, Yamagata H, Shelbourne P, Buxton J, Ogihara T, Nokelainen P, Nakagawa M, et al (1992) Comparison of the myotonic dystrophy associated CTG repeat in European and Japanese populations. J Med Genet 29:766–769

Deka R, Majumder PP, Shriver MD, Stivers DN, Zhong Y, Yu LM, Barrantes R, et al (1996) Distribution and evolution of CTG repeats at the myotonin protein kinase gene in human populations. Genome Res 6:142–154

Fu Y-H, Pizzuti A, Fenwick RG, King J, Rajnarayan S, Dunne PW, Dubel J, et al (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science 255: 1256–1547

Gennarelli M, Dallapiccola B, Baiget M, Martoreli I, Novelli G (1994) Meiotic drive at the myotonic dystrophy locus. J Med Genet 31:980

Harley HG, Brook JD, Rundle SA, Crow S, Reardon W, Buckler AJ, Harper PS, et al (1992) Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. Nature 355:545–546

Harper P (1989) Myotonic dystrophy, 2d ed. WB Saunders, London

Hurst GDD, Hurst LD, Barrett JA (1995) Meiotic drive and myotonic dystrophy. Nat Genet 10:132–133

Imbert G, Kretz C, Johnson K, Mandel J-L (1993) Origin of the expansion mutation in myotonic dystrophy. Nat Genet 4:72–76

Krahe R, Eckhart M, Ogunniyi AO, Osuntokun BO, Siciliano

MJ, Ashizawa T (1995) De novo myotonic dystrophy mutation in a Nigerian kindred. Am J Hum Genet 56:1067–1074

Mahadevan MS, Amemiya C, Jansen G, Sabourin L, Baird S, Neville CE, Wormkamp N, et al (1993*a*) Structure and genomic sequence of myotonic dystrophy (DM kinase) gene. Hum Mol Genet 2:299–304

Mahadevan MS, Foitzik MA, Surh LC, Korneluk RG (1993*b*) Characterization and polymerase chain reaction (PCR) detection of an Alu deletion polymorphism in total linkage disequilibrium with myotonic dystrophy. Genomics 15: 446–448

Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Neville C, et al (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3′ untranslated region of the gene. Science 255:1253–1255

Monckton DG, Wong LC, Ashizawa T, Caskey CT (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analysis. Hum Mol Genet 4:1–8

Mulley JC, Staples A, Donnelly A, Gedeon AK, Hecht BK, Nicholson GA, Haan EA, et al (1993) Explanation for exclusive maternal origin for congenital form of myotonic dystrophy. Lancet 341:236–237

Neville CE, Mahadevan MS, Barcelo JM, Korneluk RG (1994) High resolution genetic analysis suggests one ancestral predisposing haplotype for the origin of the myotonic dystrophy mutation. Hum Mol Genet 3:45–57

Rao CR (1973) Linear statistical inference and its applications. Wiley, New York

Stallings RL, Olson E, Strauss AW, Thompson LH, Bachinski LL, Siciliano MJ (1988) Human creatine kinase genes on chromosomes 15 and 19, and proximity of the gene for the muscle form to the genes for apolipoprotein C2 and excision repair. Am J Hum Genet 43:144–151

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Whitehead AS, Solomon E, Chambers S, Bodmer WF, Pover S, Fey G (1982) Assignment of the structural gene for the third component of human complement to chromosome 19. Proc Natl Acad Sci USA 79:5021–5025

Wieringa B (1994) Myotonic dystrophy reviewed: back to the future. Hum Mol Genet 3:1–7

Yamagata H, Miki T, Ogihara T, Nakagawa M, Higuchi I, Osame M, Shelbourne P, et al (1992) Expansion of unstable DNA region in Japanese myotonic dystrophy patients. Lancet 339:692

Zerylnick C, Torroni A, Sherman SL, Warren ST (1995) Normal variation at the myotonic dystrophy locus in global human populations. Am J Hum Genet 56:123–130

Zhang L, Leeflang EP, Yu J, Arnheim N (1994) Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat Genet 7:531–535