The cryptic CGG repeat responsible for the fragile X syndrome, located in the 5′-UTR of FMR1, is unique compared with the many other triplet repeat-causing diseases, making it ideal for identifying factors involved in repeat expansion that may be common to other triplet repeat diseases. To date, a number of factors have been identified which may influence repeat instability, including the number and position of interspersed AGGs, length of the 3′ pure CGG repeat and haplotype background. However, nearly all such data were derived from studies of Caucasians. Using a large African-American population, we present the only comprehensive examination of factors associated with CGG repeat instability in a non-Caucasian population. Among Caucasians, susceptible alleles were thought to come from those in the intermediate repeat range (41–60 repeats); however, we find that susceptible alleles may come from a larger repeat pool (35–60 repeats) and are better defined by their pure CGG repeat and/or presence of only one AGG interruption. These results demonstrate the existence of different susceptible alleles among world populations and may account for the similar prevalence of the fragile X syndrome in African-Americans compared with Caucasians despite the lower frequency of intermediate-sized alleles in the African-American population. Finally, we show that repeat structures among unaffected African-Americans with the most frequent fragile X haplotype background are either pure or contain a single distal interruption. We propose that the lack of a proximal most interruption is a novel factor involved in CGG repeat instability.

INTRODUCTION

Over a dozen neurological and neuromuscular diseases have been attributed to the amplification of a CAG, CTG, GAA or CGG repeat. However, despite the seemingly growing list of diseases triggered by triplet repeat expansion, relatively little is known about the mechanisms or pathways that drive the variability and ultimately the instability witnessed at these loci (1–4). Presumably, these diseases share some common characteristics and common mutational pathways in repeat instability; therefore, detailed study of one locus may provide insight into other triplet repeat loci. The fragile X CGG repeat locus serves as an ideal model to identify factors involved in triplet repeat expansion, due to its cryptic repeat and relative frequency among the general population.

The fragile X syndrome is an X-linked disorder that affects ~1 in 4000 males in the general Caucasian population and is considered one of the most common forms of inherited mental retardation (5–9). It was among the first triplet repeat disorders cloned and described (10,11) with >95% of fragile X syndrome cases caused by the hyperexpansion of a CGG repeat located in the 5′-UTR of the FMR1 gene (12,13). Hyperexpansion of the repeat causes hypermethylation (14,15) and histone deacetylation (16) of a nearby CpG island which subsequently silences the transcription of FMR1 (17). Thus, it is thought that the lack of the encoded protein FMRP, an RNA-binding protein, is responsible for the fragile X phenotype (18).

Because of the lack of an accurate animal model for CGG repeat instability (19–21), investigators have relied heavily on family studies and large population studies to examine repeat behavior and identify factors involved in trinucleotide repeat instability at the fragile X locus. It was appreciated early in fragile X studies that affected individuals had >200 CGG repeats, termed the full mutation (13). Examination of families with affected individuals demonstrated that asymptomatic carriers had large repeat sizes ranging from ~55 to <200 repeats, termed premutations (13). Premutations have been shown to be unstable when passed from parent to offspring with the risk of expanding to the full mutation being correlated to the size of the maternal premutation allele (12,13,22–26). In contrast to the premutation allele, the normal CGG repeat ranged from 6 to ~55 repeats and was usually transmitted from parent to offspring in a stable manner (13).

Sequence analysis of the CGG repeat array revealed that the repeat is not pure but is periodically interspersed with an AGG usually every 9–10 CGGs (24,27–30). Although repeat sizes within the normal range often contained two repeats, premuta-
tion alleles often had only one repeat at the 5′ end of the repeat or none at all (27,30,31), implicating purity of the 3′ end of the repeat as an important factor for CGG repeat instability. More specifically, family studies and population studies have suggested that alleles with >34 (31) or maybe even >24 pure repeats (29) at the 3′ end of the structure may be predisposed to expansion to the premutation state.

A multi-allelic model has been proposed to model the progress of a normal, stable allele through several defined allelic states before it is essentially eliminated from the general population as a full mutation allele (32). However, the biological details of this process are unknown. Using population studies in Caucasians, Eicher et al. (33) expanded on the multi-allelic model by describing two mutational pathways that may lead to the full mutation at different mutational rates. In essence, Eichler et al. (33) showed that two different fragile X haplotypes constructed from flanking short tandem repeat (STR) markers (see Fig. 1 for location and nomenclature) are associated with two different CGG repeat structures. The repeat alleles on the 2-1-3 haplotype, that are associated with large, usually highly interspersed structures, were postulated to progress slowly to the full mutation by the addition of a few repeats at the 3′ end of the repeat, possibly by replication slippage. In contrast, alleles on the 6-4-4/5 haplotype, that are associated with ‘asymmetrical’ interspersion patterns such as 9+12+9, were postulated to progress more rapidly to the full mutation. Eichler et al. (33) further postulated that cis-acting factors in linkage disequilibrium with these two haplotype backgrounds may influence the loss or retention of AGG interruptions.

In a re-examination of factors associated with instability, Gunter et al. (34) tested a large Caucasian population for associations between the single nucleotide polymorphism (SNP) ATL1 and the CGG repeat structures. The rarer allele (ATL1-G) was associated with both intermediate alleles (41–60 repeats) and mutated alleles in a Caucasian population. Also, ATL1-G was found to be associated with 9+ structures, the repeat structures that account for the majority of Caucasian intermediate and premutation alleles. In contrast, ATL1-A was associated with 10+ structures, which were conspicuously less frequent among the Caucasian intermediate and premutation alleles. The position of the 5′-most repeat was then proposed as another factor that may be responsible for eventual repeat instability in addition to purity of the 3′ end of the repeat and possible cis-acting factors.

The vast majority of population and family data collected on fragile X repeat structures are based on western European Caucasian X chromosomes. Few reports on the prevalence (35–39), repeat structure (40–45) or haplotype associations (46–50) are available for other racial/ethnic groups.

To fully understand the CGG repeat dynamics and to potentially apply the findings to other repeat sequences in the genome, the structure and behavior within the different racial/ethnic populations need to be examined. Previously, we have reported the prevalence of FMR1 allelic forms and performed haplotype studies based on overall repeat size in a large African-American population (9,51). We extend these studies by examining the CGG repeat interspersion pattern data on a large, unaffected African-American population. Using these
Table 1. Diversity of the CGG repeat array within the African-American population compared with other world populations

<table>
<thead>
<tr>
<th>Population</th>
<th>FRAXA No. of CGG repeat alleles</th>
<th>No. of CGG repeat patterns</th>
<th>Expected heterozygosity</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-Americans (n = 213)</td>
<td>20</td>
<td>55</td>
<td>0.816 ± 0.027</td>
<td>Present study</td>
</tr>
<tr>
<td>African-Americans (n = 72)</td>
<td>13</td>
<td>31</td>
<td>0.807 ± 0.046</td>
<td>Eichler et al. (40)</td>
</tr>
<tr>
<td>Africans (n = 37)</td>
<td>12</td>
<td>19</td>
<td>0.839 ± 0.060</td>
<td>Mandenka and Wolof from Kunst et al. (44)</td>
</tr>
<tr>
<td>Africans (n = 19)</td>
<td>6</td>
<td>8</td>
<td>0.789 ± 0.094</td>
<td>Baka pygmy and Mbuti pygmy from Eichler and Nelson (43)</td>
</tr>
<tr>
<td>Caucasians (n = 200)</td>
<td>27</td>
<td>52</td>
<td>0.836 ± 0.026</td>
<td>Eichler et al. (33)</td>
</tr>
<tr>
<td>Hispanics (n = 73)</td>
<td>9</td>
<td>17</td>
<td>0.656 ± 0.056</td>
<td>Eichler et al. (40)</td>
</tr>
<tr>
<td>Asians (n = 74)</td>
<td>10</td>
<td>17</td>
<td>0.633 ± 0.056</td>
<td>Eichler et al. (40)</td>
</tr>
<tr>
<td>Native Americans (n = 28)</td>
<td>3</td>
<td>5</td>
<td>0.559 ± 0.094</td>
<td>Navajo and Mataco from Kunst et al. (44)</td>
</tr>
</tbody>
</table>

Table 2. Frequency of CGG repeat interspersion patterns among an unaffected African-American population (n = 213)

<table>
<thead>
<tr>
<th>Interspersion pattern</th>
<th>No. (frequency)</th>
<th>Interspersion pattern</th>
<th>No. (frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10+9+9</td>
<td>45 (0.211)</td>
<td>10+30</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+9+9</td>
<td>45 (0.211)</td>
<td>10+7+19</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+9+10</td>
<td>16 (0.075)</td>
<td>10+8+10</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+10+9</td>
<td>13 (0.061)</td>
<td>10+9+11</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+9+10</td>
<td>6 (0.028)</td>
<td>10+9+21</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+10+9</td>
<td>6 (0.028)</td>
<td>10+9+6+10</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+9+9+9</td>
<td>6 (0.028)</td>
<td>11</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+11</td>
<td>4 (0.019)</td>
<td>11+30</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>14+9</td>
<td>4 (0.019)</td>
<td>11+9+9</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+20</td>
<td>3 (0.014)</td>
<td>12+10</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+22</td>
<td>3 (0.014)</td>
<td>12+20</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+9</td>
<td>3 (0.014)</td>
<td>12+29</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>19+9</td>
<td>3 (0.014)</td>
<td>13+10</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>20+9</td>
<td>3 (0.014)</td>
<td>30</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>23</td>
<td>3 (0.014)</td>
<td>33</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+12</td>
<td>3 (0.014)</td>
<td>36</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+12+9</td>
<td>3 (0.014)</td>
<td>7+9</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+14</td>
<td>2 (0.009)</td>
<td>9+10+11</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+19</td>
<td>2 (0.009)</td>
<td>9+18</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>12+9</td>
<td>2 (0.009)</td>
<td>9+21</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>22+9</td>
<td>2 (0.009)</td>
<td>9+22</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+10+10</td>
<td>2 (0.009)</td>
<td>9+25</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>9+20</td>
<td>2 (0.009)</td>
<td>9+28</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+13</td>
<td>1 (0.005)</td>
<td>9+29</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+21</td>
<td>1 (0.005)</td>
<td>9+9+12</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+24</td>
<td>1 (0.005)</td>
<td>9+9+13</td>
<td>1 (0.005)</td>
</tr>
<tr>
<td>10+27</td>
<td>1 (0.005)</td>
<td>9+9+19</td>
<td>1 (0.005)</td>
</tr>
</tbody>
</table>
In Human Molecular Genetics, 2000, Vol. 9, No. 12, data in conjunction with previous findings from this population we demonstrate that: (i) diversity within the African-American population is better defined by CGG repeat structure compared with overall repeat size; (ii) frequency and structure of ‘susceptible’ alleles in African-Americans are different from those in Caucasians; (iii) associations between repeat structure and STR- and SNP-based haplotypes are dependent on genetic age and/or history; and (iv) frequency of alleles lacking the 5’-most AGG interruption in African-Americans suggests a novel factor involved in CGG repeat instability.

RESULTS

Repeat structure in the unaffected African-American population

To describe the general distribution of the CGG repeat structures within an unaffected African-American population, the CGG repeat array of 213 chromosomes were randomly selected from a previously described unaffected African-American population (9,51,52). Of 213 chromosomes with 20 different CGG repeat size alleles, 55 different CGG repeat structures were identified (Table 1). As expected for a genetically older population, the unaffected African-American population displayed a greater number of CGG repeat structures, as well as a higher expected heterozygosity compared with other populations such as Asians, Native Americans, Hispanics and Caucasians (33,40,44) (Table 1). Also, the expected heterozygosity for the African-American population reported here is similar to those reported for other African-American and African populations (40,43,44) (Table 1).

Among unaffected African-Americans, the most common CGG repeat structures observed were 10+9+9 (0.211), 9+9+9 (0.211), 10+9+10 (0.075) and 10+10+9 (0.061) (Table 2). Most African-American CGG repeat interspersion patterns contained either one (0.249) or two AGG interruptions (0.681), whereas the remaining patterns contained either three AGG interruptions (0.038) or none at all (0.033) (Table 2). In general, when grouped by the position of the 5’-most interruption, alleles with the 10+n repeat pattern (0.474) were the most common, followed by those with the 9+n pattern (0.35). Other more rare alleles showed variability regarding the position of the 5’-most interruption (Table 2).

Repeat structures among intermediate alleles

In Caucasian populations, intermediate alleles (41–60 repeats) with >24 pure repeats at the 3’ end of the repeat array are hypothesized to be prone to repeat instability (29,31). Of the 213 unaffected African-American chromosomes, only four intermediate alleles were identified and three had >24 pure CGG repeats at the 3’ end of the repeat (Table 2). Of the 213 unaffected African-American chromosomes, only four intermediate alleles were identified and three had >24 pure CGG repeats at the 3’ end of the repeat (Table 2). Because the frequency of intermediate alleles is low in the African-American population (9), we returned to the original large, population-based African-American sample and sequenced all alleles with >34 CGG repeats to estimate more accurately the frequency of the different repeat structures. By returning to the original study, we were able to determine the
repeat structure for an additional 32 African-American chromosomes ranging from 35 to 49 repeats. These 32 additional chromosomes and the 20 chromosomes identified in the set of 213 chromosomes represent a random set of repeat structures among alleles with >34 CGG repeats (Fig. 2). Among these chromosomes, 22 of 52 (0.423) have >24 pure CGG repeats at the 3′ end of the repeat. All 22 of these alleles also have one proximal-most interruption or none at all (Fig. 2). In fact, alleles with >24 pure 3′ repeats account for ~92% (22/24) of the 35–49 repeat class alleles with a single or no AGG interruption (Table 3; Fig. 2). Adjusting for the frequency of this class of alleles among unaffected African-Americans [5.5% (9,52)], 2.3% of unaffected African-American alleles are expected to be between 35 and 49 repeats with one or no interruptions and have >24 pure repeats at the 3′ end.

**Association of repeat structure and FMR1 flanking markers**

The SNP ATL1, the closest of the surrounding markers to the CGG repeat, was shown to be tightly associated with the position of the 5′-most AGG interruption in a large, unaffected Caucasian population (34). More specifically, the most common allele in this population, ATL1-A, was strongly associated with 10+n structures, whereas the ATL1-G allele was strongly associated with 9+n structures (34). We previously reported that, in contrast to the Caucasian population, the most common ATL1 allele in African-Americans was G, possibly reflecting the higher frequency of the suspected 9+n structures in the African-American population compared with Caucasians (51). However, examination of the African-American CGG repeat structures here revealed that 33% (49/148) of the alleles on the ATL1-G background have the 10+n pattern (Fig. 3). Since the G allele was proposed to be the ancestral allele (34), this result was not surprising based on the expected diversity of the older population. The greater number of CGG repeat structures (47 versus 18) and STR-based haplotypes (79 versus 16) on the G background compared with the A background further supports that the G allele is the ancestral allele (data not shown).

In addition to observing the G allele being tightly linked to 9+n structures in Caucasians, Gunter et al. (34) also observed that the frequency of the G allele was significantly higher among both intermediate and full mutation alleles compared with the unaffected population. This observation led Gunter et al. (34) to propose that, in addition to purity of the 3′ end of the repeat, position of the 5′-most AGG might be another factor for instability. If 9+n structures were inherently more unstable than 10+n structures, one might expect to see an overabundance of 9+n structures among the African-American intermediate alleles as observed in Caucasians. In fact, 10+n structures and 9+n structures among the random set of alleles with 35–49 repeats were found in almost equal frequency (Fig. 2). If only alleles with 41–60 repeats were considered (i.e. the strict definition of intermediate alleles), the 10+n structures (12/18) were found to be more frequent than the 9+n structures (4/18) in the African-American population (Fig. 2). Superimposing ATL1 information on CGG repeat interspersion data also revealed the lack of association with the 5′-most AGG among African-American intermediate alleles (Fig. 2). Thus, the lack of association with the 5′-most AGG as well as the frequency of 10+n structures among African-American alleles suggests that the 9+n structure is not inherently susceptible to expansion.
We have previously reported that the lack of STR-based haplotype associations with CGG repeat size observed among African-American compared with Caucasians was primarily due to the age and/or history of the population rather than differences in mutational pathways (51). However, these data were based only on repeat size. Here we examine repeat structure as well. Of the 89 haplotypes found among the 213 African-American chromosomes, five were significantly associated at the \( P < 0.0003 \) level with a specific CGG repeat structure, and two approached significance \( (P < 0.0008) \) (see Statistical methods; Table 4). Like the Caucasian population reported by Eichler et al. (33), the most common haplotype \((7-3-4+)\) was associated with the common structure \((10+9+9)\) in African-Americans (Table 4). Also, like the Caucasian population, the haplotype \((6-4-4)\) was associated with the ‘asymmetrical’ repeat pattern \((9+12+9)\), although this association only approached significance possibly due to the small sample size.

The remaining associations are unique to the African-American population with three haplotype backgrounds \((7-3-4+, 7-3-6\) and \(7-1-3)\) found on African-American fragile X chromosomes and four on related backgrounds.

To understand better possible susceptibility factors, we examined the CGG repeat structures on STR-based haplotypes known to be found among African-American fragile X chromosomes. Again, to enrich our sample for such chromosomes, we returned to our original population-based sample (9) and identified an additional 39 unaffected African-American alleles for sequencing. Approximately one-third of African-American fragile X chromosomes are found on the \((+4-4-5)\)
haplotype background, whereas this haplotype is only found among 1.7% of unaffected African-Americans and has not yet been reported among unaffected Caucasians (51). Repeat structures on this haplotype background in unaffected African-Americans were striking: either no AGG interruptions or only one at the 3′-most end of the repeat were observed (51) (Fig. 4). Structures lacking a 5′-most interruption were not confined to the (+)4-4-5 haplotype and its related backgrounds (+)4-4-4, 4-4-5 and (+)5-4-5 as they were also found among the 6-4-5 and (+)7-4-6+ fragile X haplotypes (Fig. 4).

In contrast to the (+)4-4-5 haplotype, the second most common African-American fragile X and rare unaffected haplotype background [0.5% (51)], (-)3-4-5, demonstrated CGG repeat structures with one or two AGG interruptions with the first interruption in the 10th position. Interestingly, the 6-4-5 haplotype demonstrated the most CGG repeat structure heterogeneity: alleles had one, two or five interruptions and the placement of the 5′-most AGG occurred at the 5th, 10th or 20th position of the repeat (Fig. 4).

To date, the CGG repeat structures of African-American premutation alleles have not been described. Therefore, in order to gain insight into the mutational pathway of the structures lacking the 5′-most AGG, we ascertained an African-American premutation female with the haplotype background (+)4-4-5 (Fig. 5, individual II:2). After separating the X chromosomes using somatic cell hybrids (see Materials and Methods), the CGG repeat array of the premutation allele was sequenced and found to contain no AGG interruptions (Fig. 5 and data not shown). The normal CGG repeat was also sequenced and found to be the common 10+9+9 structure on the 7-3-4+ haplotype background (Fig. 5). Although only in one case, these results suggest that the structures lacking the 5′-most AGG may progress rapidly to the full mutation after the loss of the 3′-most AGG leading to pure CGG repeats. Alternatively, the pure premutation allele may represent the group of unaffected pure alleles found within the (+)4-4-5 haplotype group in African-Americans. Further ascertainment and characterization of African-American premutation alleles will be necessary to distinguish between the two possibilities.

### DISCUSSION

Reported here is the most comprehensive examination of factors implicated in CGG repeat instability in a non-Caucasian population. Within the African-American population, we observed a diversity of CGG repeat structures as well as diminished linkage disequilibrium between these structures and surrounding markers. While not unexpected for a genetically older population, these results taken within the context of other world population data on this locus provide a glimpse of the evolutionary history of the CGG repeat. From this, we are able to propose both the mechanisms of CGG repeat instability and the numerous factors that govern them.

### Population dynamics of the CGG repeat

Previously we showed a significantly different distribution of CGG repeat sized alleles in the African-American population compared with the Caucasian population. (9,52). The difference between the two populations was due mainly to a lower frequency of smaller alleles (20–23 repeats) and larger alleles (41–60 repeats) in the African-American population compared with Caucasians. This observation has also been made between other African-American and Caucasian populations (40). The difference in distributions led to a decreased diversity of CGG repeat sized alleles in African-Americans compared with Caucasians which was in marked contrast to the diversity
displayed by the flanking markers in the African-American population (51). The addition of CGG repeat structure information in the African-American population reported here, however, increased the diversity compared with other world populations at this locus despite the lack of diversity of CGG repeat size alleles (55 repeat structures versus 20 repeat sized alleles; Table 1). These results suggest that, over time, CGG repeat alleles at the tails of the size distribution are either eliminated from the population and/or assimilated into the common allele distribution by mutation leading to expansion or contraction of the repeat resulting in more diverse and sometimes unique CGG repeat structures (e.g. 9+10+11, 10+9+11, 10+8+10, 11+9+9, 12+20) (Table 1).

With respect the distribution of the CGG repeat structures in African-Americans, the 9+9+9 pattern was as frequent as the most common pattern found in Caucasians, the 10+9+9 pattern. Comparatively, the 9+9+9 pattern in the African-American population was found on many more STR-based haplotype backgrounds than the 10+9+9 pattern [30 versus 13 (data not shown)]. The diversity of backgrounds coupled with the high frequency of the 9+9+9 pattern in both Africans and African-Americans supports the hypothesis that this is the ancestral CGG repeat structure in humans (43,44).

When structures were classified by the number of AGG interruptions, the African-American distribution of interrupted structures did not differ from those reported for US or western European Caucasian populations (27,28,30,33), nor did it differ from a reported Mediterranean population (Table 3; Fisher’s exact test, \( P > 0.1000 \)) (38). The number of interruptions in African-American alleles, however, did differ from those reported in a Danish and Greenlandic population (42) and Tunisian Jewish population (35) (Table 3; Fisher’s exact test, \( P < 0.004 \)). The Greenlandic population had a higher frequency of highly interrupted alleles and, consequently, a lower frequency of alleles with one interruption. As the Greenlandic population is believed to be partly Eskimo and, therefore, of Asian origin (42), this population is more representative of the highly interspersed Native American CGG repeat structures previously reported (43,44). The Tunisian Jewish population, in contrast, had a higher frequency of pure repeat alleles compared with other world populations, of which almost half are between 17 and 25 pure repeats (52).

Defining susceptible repeat structures

Studies of Caucasian chromosomes have suggested that both the purity of the 3’ end of the repeat (>24 repeats) and the overall reduction in AGG interruptions are risk factors for instability. We examined the repeat structure of alleles with >34 CGG repeats in the African-American population to assess the frequency of such risk factors. We found that ~2% of the alleles in the African-American population have >24 pure 3’ repeats, which is similar to that estimated for a Caucasian population (29). As shown in Table 3, 46% of African-American alleles 35–49 repeats are either pure or have a single proximal interruption. Comparatively, only 21% (12/58) of Caucasian alleles with 35–49 repeats (27–29,33) have either one AGG or none at all (Table 3). This 2-fold difference between African-Americans and Caucasians is statistically significant (Fisher’s exact test, \( P = 0.0007 \)). Adjusting for the frequency of the 35–49 repeat class of alleles in Caucasians [0.080 (29)] and African-Americans (0.055), it is estimated that at least 1.7 and 2.3% of alleles in their respective populations have >24 pure 3’ repeats and either one or no interruptions. This result suggests that, compared with Caucasians, the African-American population may have a slight increased frequency of susceptible alleles as defined by 3’ purity and the number of interruptions.

As in our previous study of STR- and SNP-based haplotypes and repeat size, we did not find evidence for cis-acting factors related to mutational susceptibility (51). Nor did we find evidence for increased susceptibility of 9+n structures compared with 10+n structures. However, the haplotype association studies did reveal a potentially new susceptibility factor. That is, we did find evidence that the presence or absence of the proximal-most interruption may have an impact on repeat stability. As noted in Crawford et al. (51), the (+)4-4-5 and (-)3-4-5 haplotype backgrounds account for almost half of the African-American fragile X chromosomes, yet both are relatively infrequent in the unaffected African-American population. Also, the repeat sizes with these two haplotype backgrounds in the unaffected African-American population are not large (i.e. intermediate alleles), suggesting that the repeat structure is somehow susceptible to rapid expansion. Sequencing of the CGG repeat array among unaffected African-Americans with the (+)4-4-5 haplotype background revealed that these alleles were either pure or contained one AGG at the distal most position of the repeat array (Fig. 4). Thus, the lack of a 5’ interruption may be a novel factor involved in CGG repeat instability and may be an alternative pathway to those proposed from Caucasian association studies (i.e. the loss of the distal-most interruption or the slow increase of 3’ pure repeats).

Examination of the chromosomal backgrounds of the African-American CGG repeat structures missing the proximal-most interruption revealed that they were not confined to the (+)4-4-5 haplotype as they were also identified among the haplotypes (+)4-4-4, 4-4-5, (+)5-4-5, 6-4-5, (+)7-4-6+ and 5-4-3+ (Fig. 4). Thus, the mutation may be recurrent or may have occurred on a single haplotype background with FRAXAC1 allele 4, and the identification of the several haplotypes reported here may be the result of mutations at FRAXAC2 or recombination at DXS548, which have been shown to occur at these loci (53,54).

As suggested previously, the (+)4-4-5 haplotype and its variants may be related to the 6-4-4/5 haplotype (51). This particular haplotype has been shown to be associated with ‘asymmetrical’ CGG repeat structures in Caucasians (i.e. 9+10+9 and 9+12+9) (33). Indeed, loss of the proximal-most interruption of these asymmetrical structures could give rise to some of the structures observed in this African-American population. Also, the lack of African-American intermediate alleles missing the 5’-most interruption suggests that these repeat structures rapidly progress to the full mutation as was proposed for the ‘asymmetrical’ repeat structures (33). However, among other world populations, repeat structures missing the proximal-most interruption have been reported on haplotypes 7-3-4+ (33), 7-3-3+ (52) and 2-1-3 (55). For those reports typing for flanking markers, these structures have been found with FRAXAC1 alleles 3 and 4 (29,44) and DXS548 allele 6 (27). As recombination at FRAXAC1 is less likely due to its proximity to the CGG repeat, the number of haplotype
backgrounds represented among these particular structures in the African-American population and other world populations suggests that the loss of the 5′-most AGG is a recurrent mutation and may not necessarily originate from ‘asymmetrical’ structures.

Susceptible structures and their impact on prevalence

Discussing the frequency of susceptible structures and its impact on the prevalence of the fragile X syndrome in the context of world populations is difficult due to the lack of population-based studies for non-Caucasian populations. A non-population-based report on the fragile X syndrome in Israel suggested that the Tunisian Jewish population is at a nearly 10-fold increased risk compared with the Caucasian population (35). Analysis of the CGG repeat structure among the unaffected Tunisian Jewish population revealed a high frequency of intermediate alleles devoid of AGG interruptions on an unusual haplotype background found among Tunisian Jewish fragile X chromosomes (35). In contrast, although the Mediterranean population had a similar frequency of intermediate alleles without AGG interruptions (Table 3), Patsalis et al. (38) reported a frequency of the fragile X syndrome that is very similar to Caucasians.

If the elevated frequency of intermediate alleles with pure repeats were solely responsible for the increased prevalence of the fragile X syndrome for the Tunisian Jewish population, we would then expect to find a similar elevated prevalence for the Mediterranean population. Results to the contrary suggest that the Tunisian Jewish population has a higher frequency of yet another susceptible structure compared with the Mediterranean population. Based on our analysis of the African-American population, we suggest that the higher frequency of the fragile X syndrome for the Tunisian Jewish population is driven, in part, by the higher frequency of CGG repeat structures missing the proximal-most interruption. Indeed, 6/149 (0.04) unaffected Tunisian Jewish chromosomes have CGG repeat structures lacking the 5′-most interruption (35). This figure is comparable to African-Americans where 8/213 (0.038) structures among unaffected chromosomes lack the proximal-most interruption. However, if structures with only a single interruption are considered, 33% (6/18) of these structures in the unaffected Tunisian Jewish population lack the proximal-most interruption. Comparatively, only 5% (1/19) (38) and 15% (8/53) of single interruption structures lack the proximal-most repeat in the Hellenic and African-American populations, respectively.

Although the Tunisian Jewish population has clearly demonstrated an increased frequency of structures lacking the proximal-most repeat and an increased frequency of pure repeat alleles, it is unclear what impact these different structures have on the prevalence of the fragile X syndrome within the population. Even if the frequency for the two different susceptible structures is similar within a population, it may not be appropriate to deem them equally likely to confer risk to expansion. A recent study in the yeast Saccharomyces cerevisiae showed that a CTG repeat with the 3′-most interruption conferred less stability than a CTG repeat with the 5′-most interruption (56). However, the CTG repeat with the 3′-most interruption conferred greater stability compared with a pure CTG repeat. Thus, further research, both population based and model system based, is necessary to determine the frequency as well as the contribution of each type of susceptible structure on CGG repeat instability and its ultimate impact on fragile X prevalence.

Conclusions

We present here a systematic exploration of the factors thought to be associated with CGG repeat instability in a non-Caucasian population: purity of the 3′ repeat, haplotype background and position of the 5′-most AGG. As in a previous study, we demonstrate that associations identified in Caucasians with haplotype background are dependent on the population in which they are examined. We also show that, while purity of the 3′ end of the repeat plays a major role in CGG repeat instability, the number of interruptions among these alleles must also be considered. Finally, structures lacking the proximal-most interruption are identified as novel structures leading to CGG repeat instability. The precise impact that structures lacking the 5′ interruption have on CGG repeat instability and the prevalence of the fragile X syndrome has yet to be determined and is the subject of future study.

MATERIALS AND METHODS

Sequencing the CGG repeat array

A subset of unaffected African-Americans whose overall CGG repeat size had been previously determined (9,52) was subjected to sequencing for CGG repeat structure determination. Males were ascertained from a larger study of the prevalence of the fragile X syndrome in children with special needs detailed by Meadows et al. (52) and Crawford et al. (9). The primary sequencing protocol used here is detailed by Crawford et al. (51). The primers used in the sequencing reactions were: 5′-GACGGAGGCCGCGCTGCCAGG-3′, 5′-TCTCTCATCTTCTCTCAGCGCT-3′, 5′-GTGGGTGTGCGGCGTCGAGG-3′, 5′-CTCTGCTAGCGCCGGGACG-3′ and 5′-GGCCGGTGACCGAGCGGCC-3′ (28,57).

In cases of sequencing templates of limited DNA, the initial purified PCR product was subjected to a second PCR based on a previously published protocol (58). Amplification was carried out in a 25 µl final volume of 1× Pfu Plus Buffer (Stratagene, La Jolla, CA), 0.25 mM dNTPs, 12.4% DMSO, 0.7 M betaine, 1.25 U of Pfu polymerase (Stratagene) and 0.4 pmol each of primers C and F (9). The reactions were heated to 95°C for 5 min and then subjected to 35 cycles of denaturation at 95°C for 1 min, annealing at 65°C for 1 min, and elongation at 72°C for 1 min and one cycle of 72°C for 7 min. The PCR products were then run on a 1.5% agarose gel, purified and sequenced as described by Crawford et al. (51).

Genotyping

The FRAXA CGG repeat size for normal alleles was determined by a fluorescent sequencer method previously described (52). CGG repeat size for the premutation allele described in Figure 5 was also determined by the fluorescent method described by Meadows et al. (52), but with a modified PCR protocol. In short, the premutation CGG repeat was amplified with primer C fluorescently tagged with a phosphoramidite dye (Applied Biosystems, Foster City, CA) and primer F (13)
using the Expand Long PCR protocol (9,59). The PCR products were loaded onto a 6% polyacrylamide gel ( Gibco BRL, Grand Island, NY), and the data were collected and analyzed from an ABI 373 Stretch with Genescan software as described by Meadows et al. (52).

A description of the markers reported here including their position in relation to the CGG repeat (FRAXA) and nomenclature is described in Figure 1. The STR-based haplotype consists of microsatellites DXS548, FRAXAC1 and FRAXAC2, and the SNP is AFLP1. Both the previously published and the newly characterized STR- and SNP-based haplotypes were determined by protocols in the literature (10,34).

Somatic cell hybrid

Blood was obtained from an African-American premutation female under informed consent, and a lymphoblastoid cell line was established using EVB as previously described (60). Fusion was carried out as described using PEG1500 (61). Following fusion, the cells were trypsinized, collected and split into 12 dishes in selection media (F12 + 6% fetal bovine serum + 100 µM hypoxanthine + 1 µM azaguanine). After ~3 weeks, single clones were collected and DNA was extracted from the clones for genotyping and sequencing using the methods described above.

Statistical methods

We first tested the independence of the 55 different CGG repeat structures (rows) against the 89 different STR-based haplotype backgrounds (columns) using the Statxact software (Cytel Software, Cambridge, MA) to compare the observed table with 10,000 randomly generated tables. The Monte-Carlo estimate of the P-value (P < 0.0001) indicated a departure from independence. Each cell then was tested for a departure from independence using Fisher’s exact test. Because we tested a total of 147 cells, we adjusted the level of significance to a P-value of <0.0003.

Expected heterozygosities were calculated using either overall CGG repeat length information and repeat structure information as previously described (62).

ACKNOWLEDGEMENTS

We would like to thank Lisa F. Taft, Elizabeth F. Scott, Mary L. Leslie and Lisa Shubek for ascertaining samples. We would also like to thank Lorri Griffin in the GCRC for the establishment of lymphoblastoid cell lines (M01-RR-00039) and Dr Mark Hirst, Dr Sarah L. Nolin, Jane Iber, Aileen Kenneson, Kellen L. Meadows, James L. Newman and Dr Sayed Hosseini for technical assistance. We also thank Charles E. Schwartz and the Greenwood Genetic Center (Greenwood, SC) for helpful discussions concerning the fragile X syndrome and African-Americans. Finally, we would like to thank the study subjects whose participation made this work possible. S.T.W. is an investigator of the Howard Hughes Medical Institute. This work was supported by NIH grants HD35576, HD29909 and HD08443.

REFERENCES


