Original Article

DOI: 10.5582/irdr.2014.01028

Distribution of AGG interruption patterns within nine world populations

Carolyn M. Yrigollen[1], Stefan Sweha[1], Blythe Durbin-Johnson[2], Lili Zhou[3], Elizabeth Berry-Kravis[3], Isabel Fernandez-Carvajal[4], Sultana MH Faradz[5], Khaled Amiri[6], Huda Shaheen[6], Roberta Polli[7], Luis Murillo-Bonilla[8], Gabriel de Jesus Silva Arevalo[9], Patricia Cogram[10], Alessandra Murgia[7], Flora Tassone[1,11,*]

[1] *Department of Biochemistry and Molecular Medicine, University of California Davis, School of Medicine, Davis, CA, USA;*
[2] *Department of Public Health Sciences, University of California Davis, School of Medicine, Davis, CA, USA;*
[3] *Department of Pediatrics, Neurological Sciences, Biochemistry, Rush University Medical Center, Chicago, IL, USA;*
[4] *Laboratorio de Enfermedades genéticas y cribado neonatal, Departamento de Genetica Molecular de la Enfermedad, Instituto de Biologìa y Genética Molecular Universidad de Valladolid-CSIC, Valladolid, Spain;*
[5] *Center for Biomedical Research, Diponegoro University, Semarang, Central Java, Indonesia;*
[6] *Department of Biology, College of Science, United Arab University, United Arab Emirates;*
[7] *Laboratory of Molecular Genetics of Neurodevelopment, Department of Women's and Children's Health, University of Padova, Italy;*
[8] *Autonomous University of Guadalajara, Faculty of Medicine, Guadalajara, Mexico;*
[9] *Genetic and Neurometabolic Clinic, Obras Sociales Santo Hermano Pedro, Antigua Guatemala. Center by Biomedical Research, Medicine school San Carlos University, Guatemala Central America;*
[10] *Biomedicine Division, Fraunhofer Chile Research Foundation, Santiago, Chile;*
[11] *M.I.N.D. Institute, University of California Davis Medical Center, Davis, CA, USA.*

Summary **The CGG trinucleotide repeat within the *FMR1* gene is associated with multiple clinical disorders, including fragile X-associated tremor/ataxia syndrome, fragile X-associated primary ovarian insufficiency, and fragile X syndrome. Differences in the distribution and prevalence of CGG repeat length and of AGG interruption patterns have been reported among different populations and ethnicities. In this study we characterized the AGG interruption patterns within 3,065 normal CGG repeat alleles from nine world populations including Australia, Chile, United Arab Emirates, Guatemala, Indonesia, Italy, Mexico, Spain, and United States. Additionally, we compared these populations with those previously reported, and summarized the similarities and differences. We observed significant differences in AGG interruption patterns. Frequencies of longer alleles, longer uninterrupted CGG repeat segments and alleles with greater than 2 AGG interruptions varied between cohorts. The prevalence of fragile X syndrome and *FMR1* associated disorders in various populations is thought to be affected by the total length of the CGG repeat and may also be influenced by the AGG distribution pattern. Thus, the results of this study may be important in considering the risk of fragile X-related conditions in various populations.**

Keywords: AGG interruptions, *FMR1* allele, CGG repeat, expansion, ethnicity

**Address correspondence to:*
Dr. Flora Tassone, Department of Biochemistry and Molecular Medicine, University of California Davis, School of Medicine, 2700 Stockton Blvd, Suite 2102, Sacramento, CA 95817, USA; M.I.N.D. Institute, University of California Davis Medical Center, 2805 50th Street Sacramento, CA 95817, USA.
E-mail: ftassone@ucdavis.edu

1. Introduction

Fragile X syndrome (FXS) and *FMR1* associated disorders are predominantly the result of an expansion of a trinucleotide repeat element located within the 5' UTR of the Fragile X Mental Retardation 1 gene (*FMR1*). In normal individuals the triplet repeat number varies in

length from 5 to 44 CGG repeats. Intermediate alleles are between 45 and 54 repeats, premutation alleles are between 55 and 200 CGG repeats and above 200 CGG repeats are full mutation alleles (*1*). *FMR1* full mutations cause FXS, while premutation alleles lead to fragile X-associated tremor/ataxia syndrome (FXTAS) in an estimated 40% of males and 8-16% of females with the mutation, and fragile X-associated primary ovarian insufficiency (FXPOI) in approximately 20% of female premutation carriers (*2*).

The CGG repeat element, like other trinucleotide repeats, is prone to expansion during transmission from parent to child (*3*). While the mechanism that gives rise to CGG repeat expansion in *FMR1* is not understood, evidence suggests repair of single-strand breaks in the meiotically arrested oocytes form loops, which may be incorporated into the DNA through mismatch repair resulting in an expansion (*4*).

Normal alleles most frequently have 2 AGG interruptions, less frequently they have 1 AGG interruption or 0 AGG interruptions, and rarely greater than 2 AGG interruptions. Within normal alleles the patterns most commonly seen are 9 or 10 CGG repeat segments between interruptions (*5,6*). The 9-A-9-A-9 and 10-A-9-A-9 AGG interruption patterns predominate in all populations that have been studied, evidence that these two patterns were present 200,000 years ago during early divergence of human races or that a strong selection pressure exists at this locus (*7*).

In intermediate and premutation alleles the AGG interruptions tend to occur at the 5' end of the locus and the pure CGG stretch, defined as the longest stretch of uninterrupted CGG repeats, is located at the 3' end (*8,9*). The loss of AGG interruptions appear to have occurred multiple times during human evolution (*10*) but can be a late event in the mutation pathway that leads to expansion (*11*). It is rare for AGG interruptions to be lost during transmission, but observation of its occurrence has been reported (*12-14*).

A normal allele without an AGG interruption has been shown to have an increase mutational rate compared to an allele of similar size containing an AGG interruption (*15-17*). Differences in the distribution of AGG interruption patterns between ethnicities, has been reported, including differences in the frequency of alleles that exceed 35 CGG repeats in length and lack AGG interruptions. These higher frequencies are associated with increased prevalence of FXS (*18*). Conversely, highly interspersed CGG repeat alleles have been observed in the Basque, Native American, and Asian populations, which also have lower estimated FXS prevalence rates (*19,20*).

The presence of AGG interruptions does not seem to affect the transcriptional or translational expression of the *FMR1* gene (*21-24*). However, the presence of AGG interruptions in both intermediate and premutation alleles has been shown to decrease the rate of instability (any change in CGG repeat size) and magnitude of size change in both paternal and maternal transmissions (*12,25,26*).

While the distribution of CGG repeat total length has been reported in a number of populations (*27*), fewer studies have reported the distribution of AGG interruptions within populations. This study reports on the AGG interruption patterns in a total of 3,065 normal alleles (9-40 CGG repeats) from 1,989 participants (males: n = 794; females: n = 1,195) from 9 countries: Australia, Chile, Emirates, Guatemala, Indonesia, Italy, Mexico, Spain, and USA. We compare these results with previous studies that reported AGG interruption patterns in global populations (Figure 1).

Our findings indicate that variations in CGG repeat allele sizes and AGG interruption pattern distributions exist between populations. Two populations (Australia and Indonesia), from the nine newly described, had a higher frequency of long pure CGG repeat stretches (greater than 20 pure CGG repeats), and the USA population had a lower frequency of these long pure stretches. These differences may be important when considering the burden of *FMR1* associated disorders in different populations.

2. Materials and Methods

2.1. *Participants*

Genomic DNA from unrelated individuals with at least one normal *FMR1* allele was included in this study (n = 3,065 alleles). These samples were previously screened to determine the prevalence rates of expanded alleles. Cohorts from Australia (n = 201) (*28*), Chile (n = 77), the United Arab Emirates (n = 263), Guatemala (n = 151) (*29*), Indonesia (n = 312) (*30*), Italy (n = 67), Mexico (n = 277), Spain (n = 358) (*31*), and the United States (n = 1,359) (*32*) were included. Individuals were recruited from the general population for the Italy, Spain, and United States samples. From the USA cohort, participants were from two different geographical areas: Sacramento (California) and Chicago (Illinois). The remaining samples were recruited from high-risk populations including intellectual disabilities, individuals with a family history of FXS and individuals with Parkinsonism. DNA isolation and AGG interruption genotyping were performed at the UC Davis MIND Institute Molecular Laboratory as previously described (*25,32*), except 67 alleles extracted and genotyped in Italy, following IRB approved protocols at the correspondent institutions. Only AGG interruption patterns of unrelated normal alleles less than or equal to 40 CGG repeats in length, therefore within the normal size range (*33-35*), were included in the study.

2.2. *Statistical analysis*

Distributions of categorical variables were compared

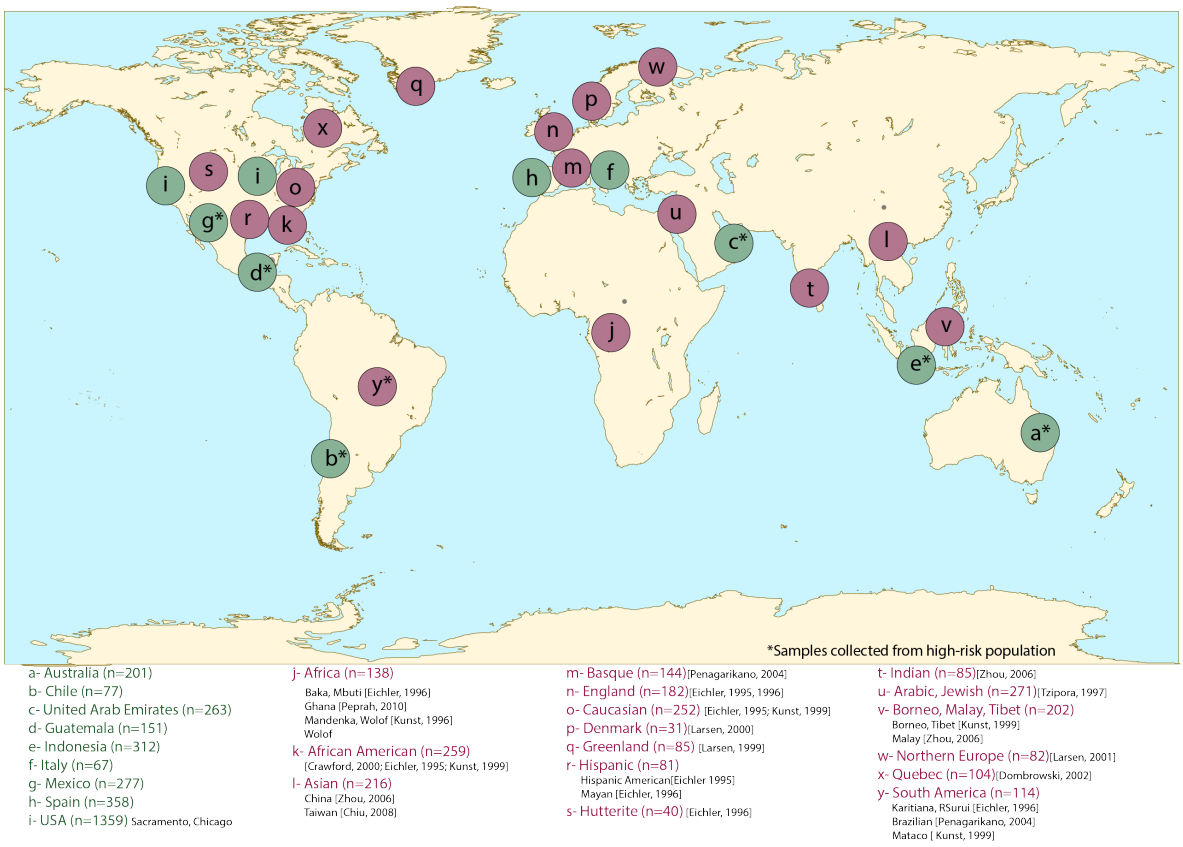

Figure 1. The distribution of 25 global populations with AGG interruption patterns described. AGG interruption patterns were compared between the 9 newly characterized cohorts (a-i, in green) and with previously published studies (j-y, in red). Populations from previous studies were combined if geographical proximity was present to increase sample sizes. Cohorts with samples collected from high-risk populations are denoted with an asterisk, total sample size for each cohort and the studies reporting their AGG patterns are provided next to the cohort's name.

among countries using chi-square tests. Chi-square test *p*-values were obtained by Monte Carlo simulation when the sample size assumptions for use of the chi-square distribution were not met.

In order to identify specific AGG interspersion patterns, total CGG lengths, pure CGG stretches, or AGG interruptions whose frequency in a given population was significantly higher or lower than would be expected under homogeneity, the adjusted residuals from the chi-square table (*36*) were compared to a standard normal distribution and the resulting *p*-values were adjusted for multiple testing using the Bonferroni correction. All analyses were conducted using R, version 2.13.0 (*37*).

3. Results

In the nine populations we determined the number and position of the AGG interruption within each CGG alleles and thus determined the AGG interruption pattern in 3,065 alleles. We observed 30 different CGG repeat lengths ranging from 9 to 40 CGG repeats and 231 different AGG interruption patterns, each allele contained no AGG interruptions up to 3 AGG interruptions. Consistent with previous population based studies of the CGG repeat locus, 29 and 30 CGG repeats were the most common allele sizes in all 9 populations. Indonesia was the only population with a greater proportion of alleles with 29 (39%) than 30 (28%) CGG repeats. Two AGG interruptions were present in at least 56% of the alleles genotyped for each population; 1 AGG interruption occurred in at least 11% of the alleles genotyped (Figure 2, Table 1).

3.1. *The distribution of total CGG length, pure CGG stretch, and number of AGG interruptions differs between populations*

The mode of total CGG length was 30 in subjects from all countries examined except in Indonesia where it was 29 (Figure 3). The relative proportions of subjects with a total of 29 CGG repeats, 30 CGG repeats, or a value other than 29 or 30 differed significantly by country ($p < 0.001$) (Table 1). Likewise, the relative proportions of subjects with a pure stretch of 9 CGG repeats, 10 CGG repeats, or a value other than 9 or 10 differed significantly by country ($p < 0.001$). Examination of adjusted residuals suggests that significantly more alleles from Australia ($p < 0.001$), Emirates ($p = 0.007$) and Spain ($p < 0.001$) had total CGG lengths other than 29 or 30 and Australia ($p < 0.001$) and Spain ($p < 0.001$) had pure stretch lengths other than 9 or

10 repeats. Further, significantly more alleles from Indonesia ($p < 0.001$) had a total length of 29 CGG repeats, and significantly fewer USA ($p < 0.001$) alleles had total CGG lengths other than 29 and 30. Indonesia had significantly fewer alleles with a pure stretch of 10 CGG repeats ($p < 0.001$), Spain had fewer alleles with a pure stretch of 9 CGG repeats ($p < 0.001$) and USA had more alleles with 10 pure CGG repeats ($p < 0.001$) than was expected under homogeneity, where homogeneity would assume the same allele frequencies are present between populations.

The proportion of alleles with 3 AGG interruptions differs significantly by country ($p < 0.001$) (Table 1); examination of adjusted residuals suggests that significantly more alleles from Indonesia ($p < 0.001$) and Australia ($p < 0.001$), and significantly fewer from USA ($p = 0.005$) had AGG interruptions than was expected under homogeneity.

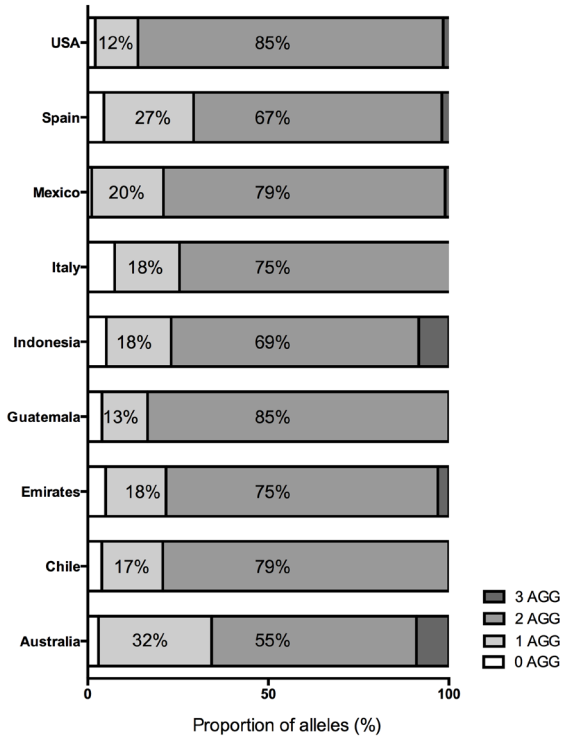

Figure 2. Distribution of number of AGG interruptions. For the nine newly characterized populations the proportion of alleles with 0 to 3 AGG interruptions is graphically represented. Alleles with 2 AGG interruptions were the most common in each cohort, followed by 1 AGG interruption. Four AGG interruptions were observed in Australia, United Arab Emirates, Indonesia, and Spain only. Within the nine populations no alleles were identified with more than 3 AGG interruptions.

3.2. *AGG interspersion patterns by country*

The most common AGG interspersion pattern was 10-A-9-A-9 in all countries except Indonesia. In Indonesia the most common AGG interspersion pattern was 9-A-9-A-9, 10-A-9-A-9 was the second most common pattern. The distribution of AGG interspersion patterns differed significantly by country ($p < 0.001$) and the most common patterns are shown in Supplementary Table 1 (*http://www.irdrjournal.com/docindex.php?year=2014&kanno=4*). Examination of adjusted residuals suggested that significantly more alleles from Australia had the patterns 9-A-9-A-9-A-9 ($p < 0.001$, 8%), 9-A-9-A-19 ($p = 0.012$, 2%), and 10-A-9 ($p < 0.001$, 11%); significantly more alleles from the Emirates had the pattern 10-A-10-A-9 ($p < 0.001$, 6%), 11-A-9-A-9 ($p = 0.022$, 2%), and 9-A-10-A-9 ($p < 0.001$, 7%). Indonesia had significantly more alleles with 30 CGG repeats and no AGG interruptions ($p = 0.010$, 2%), 9-A-13 ($p < 0.001$, 3%), 9-A-22 ($p < 0.001$, 3%), 9-A-9-A-9 ($p = 0.004$, 35%), and 9-A-9-A-6-A-9 ($p < 0.001$, 6%) patterns; significantly more Spanish alleles had the patterns 10-A-9 ($p < 0.001$, 9%), 13-A-9 ($p = 0.013$, 4%) and 9-A-12-A-9 ($p < 0.001$, 4%), and significantly more USA alleles had the pattern 10-A-9-A-9 ($p < 0.001$, 44%) than was expected under homogeneity. Likewise, fewer alleles from Australia and Spain had the pattern 9-A-9-A-9 (both $p < 0.001$, 9% and 14%, respectively), fewer alleles from Indonesia have the pattern 10-A-9-A-9 ($p < 0.001$, 21%), fewer

Table 1. Summary of allele structure in nine populations

Items	Australia	Chile	Emirates	Guatemala	Indonesia	Italy	Mexico	Spain	USA
Total length									
29	23 (11%)	19 (25%)	58 (22%)	47 (31%)	122 (39%)	11 (16%)	96 (35%)	56 (16%)	413 (30%)
30	72 (36%)	41 (53%)	104 (40%)	71 (47%)	86 (28%)	30 (45%)	106 (38%)	143 (40%)	677 (50%)
Other	106 (53%)	17 (22%)	101 (38%)	33 (22%)	104 (33%)	26 (39%)	75 (27%)	159 (44%)	269 (20%)
Pure Stretch									
9	34 (17%)	17 (22%)	57 (22%)	43 (28%)	139 (45%)	11 (16%)	96 (35%)	56 (16%)	422 (31%)
10	100 (50%)	48 (62%)	136 (52%)	75 (50%)	78 (25%)	35 (52%)	127 (46%)	190 (53%)	756 (56%)
Other	67 (33%)	12 (16%)	70 (27%)	33 (22%)	95 (30%)	21 (31%)	54 (19%)	112 (31%)	181 (13%)
Number of AGG									
0	6 (3%)	3 (4%)	13 (5%)	6 (4%)	16 (5%)	5 (7%)	3 (1%)	16 (4%)	28 (2%)
1	63 (31%)	13 (17%)	44 (17%)	19 (13%)	56 (18%)	12 (18%)	55 (20%)	89 (25%)	161 (12%)
2	114 (57%)	61 (79%)	198 (75%)	126 (83%)	214 (69%)	50 (75%)	216 (78%)	246 (69%)	1148 (84%)
3	18 (9%)	0 (0%)	8 (3%)	0 (0%)	26 (8%)	0 (0%)	3 (1%)	7 (2%)	22 (2%)

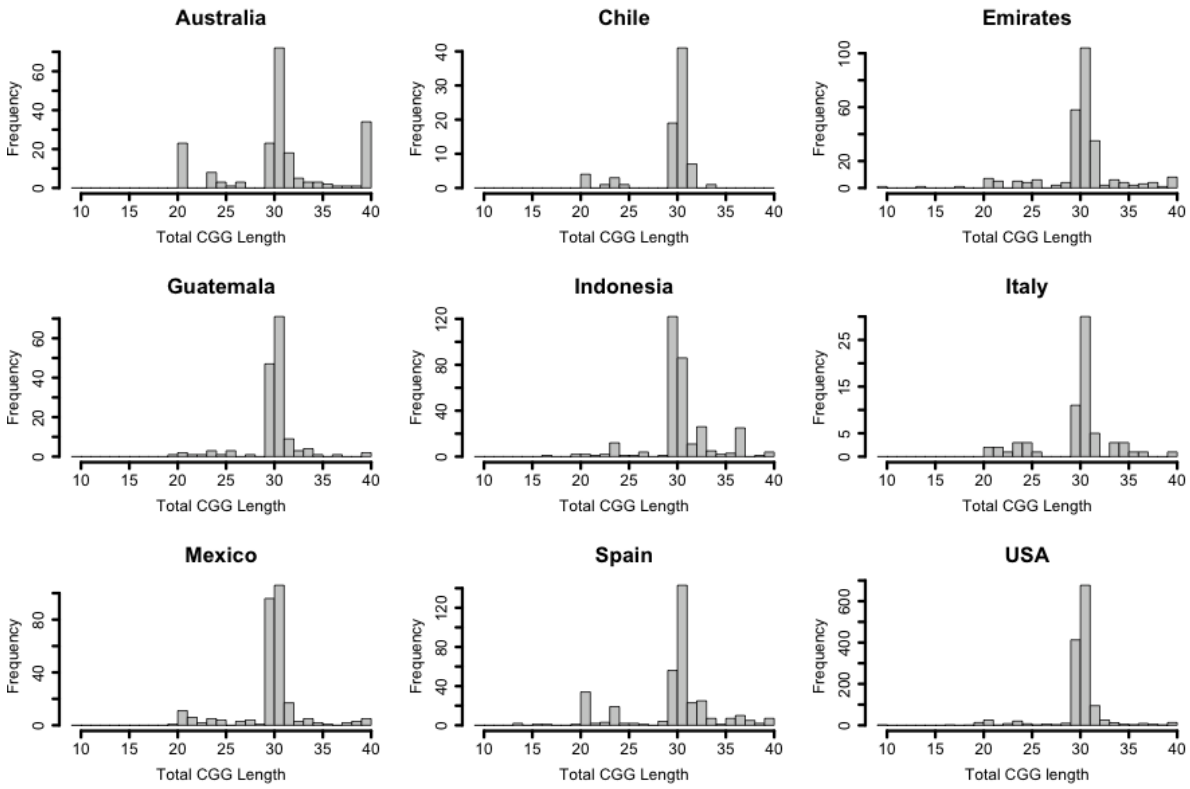

Figure 3. Histogram of total CGG length for 9 populations. The most common total length of CGG repeat sizes for the 9 populations are 30 and 29 CGG repeats, 30 is the most common for every population except Indonesia. Populations show difference in less prominent modes including some previously identified (20 CGG repeats, 23 CGG repeats, and 36 CGG repeats).

alleles from USA had the pattern 10-A-9 ($p < 0.001$, 2%) than was expected under homogeneity.

In 201 normal alleles genotyped from Australia, we observed 74 AGG interruption patterns (Supplementary Table 1; *http://www.irdrjournal.com/docindex.php?year=2014&kanno=4*). A larger proportion of alleles were in the high normal range than observed in the other populations, between 32 and 40 CGG repeats in length, and approximately 9% of the genotyped alleles contained three AGG interruptions. In 77 normal alleles genotyped from Chile, 17 AGG interruption patterns were present. In 151 normal alleles genotyped from Guatemala 39 AGG interruption patterns were observed however no remarkable patterns were observed. In 263 normal alleles genotyped from the United Arab Emirates, 77 different AGG interruption patterns were observed out of which twenty were only observed in the Emirates population. Approximately 1% of the alleles had 3 AGG interruptions. In the 312 normal alleles genotyped from Indonesia, 60 AGG interruption patterns were observed. A large portion of normal alleles with 3 AGG interruptions (8%), with the majority of these alleles having the 9-A-9-A-6-A-9 pattern (6.4% of patterns) was observed in Indonesia. The 9-A-9-A-6-A-9 pattern and CGG length of 36 repeats has been observed in previous studies to occur within Indonesian and Asian cohorts (*30,38,39*). Fifty AGG interruption patterns were observed in 277 *FMR1* alleles genotyped from Mexico. Six distinct AGG interruption patterns were observed only in the Mexico cohort, although these were each observed only once. Eighty-four AGG interruption patterns were observed in 358 normal alleles from Spain. Eleven AGG interruption patterns were only observed in the Spain cohort. Twenty-four AGG interruption patterns were observed in 67 normal CGG repeat alleles from Italy. Three patterns were observed in the Italy cohort only.

3.3. *Regional differences in frequencies of AGG interruption patterns were observed within the USA samples (Chicago and Sacramento area)*

The largest cohort of this study was from the United States, and consisted of samples from a larger collection of newborn blood spots that were collected in both the Sacramento and Chicago area (*32*). The Chicago cohort was comprised of individuals identified as Caucasian (n = 153 alleles), African American (n = 225 alleles), Hispanic (n = 223 alleles), Asian (n = 42 alleles), Southeast Asian (n = 14 alleles), Native American (n = 14 alleles), and other (n = 5 alleles). The Sacramento cohort was comprised of individuals identified as Caucasian (n = 156 alleles), African American (n = 24 alleles), Hispanic (n = 123 alleles), Asian (n = 42 alleles), Pacific Islander (n = 6 alleles), Native American (n = 4 alleles), and other (n = 328 alleles).

There were 105 AGG interruption patterns observed in 1,359 normal alleles. Twenty-five AGG interruption patterns were observed in the USA cohort and were not observed in the other 9 populations.

The most common total CGG length in both Sacramento and Chicago was 30, pure CGG stretch was 10, and number of AGG interruptions was 2 (Supplementary Figure 1; *http://www.irdrjournal.com/docindex.php?year=2014&kanno=4*). Compared to Chicago, Sacramento had a higher frequency of alleles with total CGG lengths of 30 and pure CGG stretches of 10 than would be expected under homogeneity (both, $p < 0.001$). The two cities were similar in the proportion of alleles with a pure stretch that was greater than 20 CGG repeats ($p = 0.2322$), and a total length that was greater than 35 CGG repeats ($p = 0.7471$).

The proportion of alleles with 3 interruptions did not differ significantly between Sacramento and Chicago ($p = 0.8271$). The most common pattern in both cities was 10-A-9-A-9. However, the overall distribution of AGG interspersion patterns differed significantly between Sacramento and Chicago ($p < 0.001$). Examination of adjusted residuals reveals that more alleles in Sacramento had the pattern 10-A-9-A-9 ($p < 0.001$) and more alleles in Chicago had the pattern 9-A-9-A-9 ($p = 0.043$) than would be expected under homogeneity.

3.4. *Previously studied populations*

The data from the 9 populations were compared with data from previously published studies including samples collected and sequenced from Quebec (*11*), Taiwan (*40*), Norway, Saami, Nenets (*41*), Greenland (*42*), African American (*43*), Denmark (*16*), Basque (*44*), Caucasian, Mataco, Tibet, Navajo, Borneo, Mandenka, Wolof, African American (*19*), Brazil (*45*), China, Malay, India (*17*), and sub-saharah West Africa (*46*). AGG interruption patterns were determined by *mnl I* digestion for samples collected from England, Hispanic American, African American, and Asian American (*5*), Tunisian Jews, Sephardic Jews, Ashkenazic Jews, and Arabs (*18*), Suriu, Mayan, Karitiana, Baka, Mbuti, and Hutterite (*7*).

Collections were combined as indicated in Figure 1 in order to increase sample sizes; alleles from the Navajo population were excluded because they did not reach a sufficient sample size. The distribution of total CGG repeats length, pure CGG stretch, and number of AGG interruptions was significantly different in the 25 global populations. Seven populations had higher proportions of alleles with more than 35 CGG repeats (Asian, $p < 0.001$; Australia, $p < 0.001$; Caucasian, $p = 0.013$; Denmark, $p = 0.001$; Greenland, $p < 0.001$; India, $p < 0.001$; and Indonesia, $p = 0.007$). Four populations had a larger proportion of alleles with less than 35 CGG repeats (Chile, $p = 0.016$; Guatemala, $p = 0.016$; Hispanic American, $p = 0.044$; and USA, $p < 0.001$). When pure CGG repeat stretch was compared in the 25 populations, 6 populations (Australia, $p < 0.001$; Africa, $p = 0.001$; African American, $p = 0.043$; Basque, $p = 0.043$; Indonesia, $p = 0.001$; and Jewish & Arabic, $p < 0.001$) had higher frequencies of alleles with greater than 20 pure CGG repeats. Seven populations (Asia, $p < 0.036$; Greenland, $p = 0.014$; Hispanic American, $p = 0.006$; Malay, $p = 0.004$; and USA, $p < 0.001$) had higher frequencies of alleles with less than 20 pure CGG repeats. The populations with highly interspersed alleles included Asia ($p < 0.001$), Australia ($p = 0.002$), Greenland ($p < 0.001$), India ($p < 0.001$), Indonesia ($p < 0.001$), and Malay ($p = 0.034$); the USA had significantly less AGG interruptions then expected under homogeneity ($p < 0.001$).

4. Discussion

Differences in the frequency of AGG interruption patterns within the CGG repeat locus of *FMR1* have been previously reported to vary between ethnicities, and suggested that such differences can affect the mutation rate of this locus. We have genotyped 3,065 alleles from 9 global cohorts to investigate how AGG interruption patterns vary between geographic and ethnic populations. The distribution of CGG repeat total length, and AGG interruption patterns were found to be significantly different between populations. Consistent with previous studies two AGG interruption patterns, 10-A-9-A-9 and 9-A-9-A-9, were the most common in all nine populations reported in this study, and in the 14 previously published population studies (Supplementary Table 1; *http://www.irdrjournal.com/docindex.php?year=2014&kanno=4*). 10-A-9-A-9 was the most common allele for all populations except in the African American, Asia, Indonesia, and Malay, Borneo, and Tibet cohort where 9-A-9-A-9 was the most common pattern. The frequency of the 9-A-9-A-9 pattern in Asian ethnic groups was consistent with what has previously been shown (*17,40*), and in the African American group the 9-A-9-A-9 pattern was only 1% higher in frequency than the 10-A-9-A-9 pattern. It is unknown whether these two patterns have a biological advantage, however, CGG repeat length in the normal allele has been shown to alter translational efficiency (*47*) with the highest translational efficiency occurring at 30 CGG repeats. Thus, the common lengths may provide alleles within the optimum size range with the lowest mutation rate.

We combined the AGG interruption pattern results of the 9 population cohorts genotyped for this study to the 16 cohorts from previous published studies. The results showed that six populations had a higher frequency of alleles with a total length greater than 35 repeats, and five populations had a higher frequency of alleles with an uninterrupted stretch greater than 20 repeats. Australia, Denmark, and Quebec had both, suggesting that an increased frequency of expanded

alleles, intermediate, premutation, and full mutation alleles may be present in these populations. It should be noted that as the Australian cohort was part of a high risk screening study, a sample bias affecting these results could be present given that intermediate prevalence rates were found to be increased compared to the general population (*28*). However, only alleles not greater than 40 CGG repeats were included in this study and importantly the distribution of CGG repeat length was not statistically different from the one observed in a group of 3,091 alleles (1,091 male and 2,000 female alleles) derived from Australian newborns from the state of Victoria ($p = 0.3052$). In these two population-based samples the frequencies of GZ alleles were 1.3% (> 40 CGG repeats) and 0.4% (> 44 CGG repeats), in male newborns; and 5.5% (> 40 CGG repeats) and 1.4% (> 44 CGG repeats), in female newborns (unpublished data). The frequency of premutation alleles was 0.3% in both male and female samples. In Canada prevalence estimates for intermediate alleles is 1:86 in females, and for premutation alleles is 1:813 in males and 1:241 in females (Dombrowski *et al.*, 2002). No prevalence estimates are available for the Denmark population. These prevalence rates are not higher than those estimated in other populations and also we do not have any information regarding whether these prevalence rates are increasing or decreasing with generation.

Guatemala, Hispanic American, Mexico, and USA cohorts had a smaller proportion of alleles in the high normal range, and a smaller proportion of alleles with greater than 20 uninterrupted CGG repeats, suggesting increased stability of the normal allele in these populations. However, both Guatemala and Mexico cohorts were collected for high risk screening studies, and sample collection bias may also be present in these two populations. In the USA population prevalence rates for intermediate alleles were estimated to be 1:112 for males and 1:66 for females, and for premutation alleles were estimated to be 1:430 for males and 1:209 for females (*32*). The Hispanic American, Guatamala, and Mexico populations do not have estimated prevalence rates. The prevalence estimates for the USA population are neither in agreement or disagreement with the population having an increased stability compared to the other studied populations.

A comparison of Sacramento and Chicago showed similarities in the distribution of AGG interruption patterns, and the proportion of alleles in the high normal and intermediate range, and with more than 20 uninterrupted CGG repeats (Supplementary Figure 1; *http://www.irdrjournal.com/docindex.php?year=2014&kanno=4*). Interestingly, Sacramento has an increased prevalence of premutation alleles (males, 1:305; females, 1:172) when compared to Chicago (males, 1:308; females, 1:894) (*32*). Both cohorts were collected and genotyped in the same study and were collected as part of a pilot study newborn screening for FXS.

One limitation of this study is the possible sampling bias within the six newly described population cohorts that were collected from high-risk screening studies. A sample bias may also likely be present in the Sacramento and Chicago cohorts that make up the USA population because AGG interruption data was available mainly for samples that were genotyped by the CGG linker PCR assay (*32*) when initial genotyping of females resulted in only one allele. Another limitation of this study, and of the other published studies, is represented by the small sample sizes. The expected variability introduced by sampling error inhibits strong comparisons between prevalence rates of intermediate, premutation, and full mutation alleles and AGG interruption pattern distribution; limitations that could be reduced with increasing cohort sizes.

The AGG interruption patterns within the CGG repeat locus of *FMR1* further characterize the alleles beyond repeat length. The results of the study agree with what is known about the CGG repeat distribution in the nine countries, including increased frequency of the 9-A-9-A-6-A-9 pattern in Asian ethnicities where the 36 CGG repeat length is more frequent. Population structure is important to consider when studying the CGG repeat locus, sub-populations have consistently shown significant differences in the literature, including differences between ethnic and geographic groups. Our results suggest that AGG interruption pattern distributions in populations could be associated with differing prevalence of categorically non-normal alleles, however larger cohort sizes and more prevalence rates will be needed for many ethnicities to confirm these observations.

Acknowledgements

The project described was supported by the NICHD grant HD02274, and by the National Center for Advancing Translational Science, National Institutes of Health, through grant # UL1 TR000002. This work is dedicated to the memory of Matteo.

References

1. Maddalena A, Richards CS, McGinniss MJ, Brothman A, Desnick RJ, Grier RE, Hirsch B, Jacky P, McDowell GA, Popovich B, Watson M, Wolff DJ. Technical standards and guidelines for fragile X: The first of a series of disease-specific supplements to the Standards and Guidelines for Clinical Genetics Laboratories of the American College of Medical Genetics. Quality Assurance Subcommittee of the Laboratory Practice Committee. Genet Med. 2001; 3:200-205.
2. Hagerman R, Hagerman P. Advances in clinical and molecular understanding of the *FMR1* premutation and fragile X-associated tremor/ataxia syndrome. Lancet Neurol. 2013; 12:786-798.
3. Fu YH, Kuhl DP, Pizzuti A, *et al*. Variation of the CGG

repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox. Cell. 1991; 67:1047-1058.
4. McMurray CT. Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet. 2010; 11:786-799.
5. Eichler EE, Hammond HA, Macpherson JN, Ward PA, Nelson DL. Population survey of the human *FMR1* CGG repeat substructure suggests biased polarity for the loss of AGG interruptions. Hum Mol Genet. 1995; 4:2199-2208.
6. Kunst CB, Warren ST. Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. Cell. 1994; 77:853-861.
7. Eichler EE, Nelson DL. Genetic variation and evolutionary stability of the *FMR1* CGG repeat in six closed human populations. Am J Med Genet. 1996; 64:220-225.
8. Eichler EE, Holden JJ, Popovich BW, Reiss AL, Snow K, Thibodeau SN, Richards CS, Ward PA, Nelson DL. Length of uninterrupted CGG repeats determines instability in the *FMR1* gene. Nat Genet. 1994; 8:88-94.
9. Snow K, Tester DJ, Kruckeberg KE, Schaid DJ, Thibodeau SN. Sequence analysis of the fragile X trinucleotide repeat: Implications for the origin of the fragile X mutation. Hum Mole Genet. 1994; 3:1543-1551.
10. Eichler EE, Macpherson JN, Murray A, Jacobs PA, Chakravarti A, Nelson DL. Haplotype and interspersion analysis of the *FMR1* CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome. Hum Mol Genet. 1996; 5:319-330.
11. Dombrowski C, Levesque S, Morel ML, Rouillard P, Morgan K, Rousseau F. Premutation and intermediate-size *FMR1* alleles in 10572 males from the general population: Loss of an AGG interruption is a late event in the generation of fragile X syndrome alleles. Hum Mol Genet. 2002; 11:371-378.
12. Nolin SL, Sah S, Glicksman A, *et al*. Fragile X AGG analysis provides new risk predictions for 45-69 repeat alleles. Am J Med Genet A. 2013; 161A:771-778.
13. Fernandez-Carvajal I, Lopez Posadas B, Pan R, Raske C, Hagerman PJ, Tassone F. Expansion of an *FMR1* grey-zone allele to a full mutation in two generations. J Mol Diapn. 2009; 11:306-310.
14. Terracciano A, Pomponi MG, Marino GM, Chiurazzi P, Rinaldi MM, Dobosz M, Neri G. Expansion to full mutation of a *FMR1* intermediate allele over two generations. Eur Hum Genet. 2004; 12:333-336.
15. Kunst CB, Leeflang EP, Iber JC, Arnheim N, Warren ST. The effect of *FMR1* CGG repeat interruptions on mutation frequency as measured by sperm typing. J Med Genet. 1997; 34:627-631.
16. Larsen LA, Armstrong JS, Grønskov K, Hjalgrim H, Macpherson JN, Brøndum-Nielsen K, Hasholt L, Nørgaard-Pedersen B, Vuust J. Haplotype and AGG-interspersion analysis of *FMR1* $(CGG)_n$ alleles in the Danish population: Implications for multiple mutational pathways towards fragile X alleles. Am J Med Genet. 2000; 93:99-106.
17. Zhou Y, Tang K, Law HY, Ng IS, Lee CG, Chong SS. *FMR1* CGG repeat patterns and flanking haplotypes in three Asian populations and their relationship with repeat instability. AnnHum Genet. 2006; 70:784-796.
18. Falik-Zaccai TC, Shachak E, Yalon M, Lis Z, Borochowitz Z, Macpherson JN, Nelson DL, Eichler EE. Predisposition to the fragile X syndrome in Jews of Tunisian descent is due to the absence of AGG interruptions on a rare Mediterranean haplotype. Am J Hum Genet. 1997; 60:103-112.
19. Kunst CB, Zerylnick C, Karickhoff L, Eichler E, Bullard J, Chalifoux M, Holden JJ, Torroni A, Nelson DL, Warren ST. FMR1 in global populations. Am J Hum Genet. 1996; 58:513-522.
20. Arrieta MI, Ramírez JM, Télez M, Flores P, Criado B, Barasoain M, Huerta I, González AJ. Analysis of the Fragile X trinucleotide repeat in Basques: Association of premutation and intermediate sizes, anchoring AGGs and linked microsatellites with Unstable Alleles. Curr Genomics. 2008; 9:191-199.
21. Ludwig AL, Raske C, Tassone F, Garcia-Arocena D, Hershey JW, Hagerman PJ. Translation of the *FMR1* mRNA is not influenced by AGG interruptions. Nucleic Acids Res. 2009; 37:6896-6904.
22. Peprah E, He W, Allen E, Oliver T, Boyne A, Sherman SL. Examination of FMR1 transcript and protein levels among 74 premutation carriers. J Human Genet. 2010; 55:66-68.
23. Yrigollen CM, Tassone F, Durbin-Johnson B, Tassone F. The role of AGG interruptions in the transcription of *FMR1* premutation alleles. PloS One. 2011; 6:e21728.
24. Tassone F, Beilina A, Carosi C, Albertosi S, Bagni C, Li L, Glover K, Bentley D, Hagerman PJ. Elevated *FMR1* mRNA in premutation carriers is due to increased transcription. RNA. 2007; 13:555-562.
25. Yrigollen CM, Durbin-Johnson B, Gane L, Nelson DL, Hagerman R, Hagerman PJ, Tassone F. AGG interruptions within the maternal *FMR1* gene reduce the risk of offspring with fragile X syndrome. Genet Med. 2012; 14:729-736.
26. Yrigollen CM, Martorell L, Durbin-Johnson B, Naudo M, Genoves J, Murgia A, Polli R, Zhou L, Barbouth D, Rupchock A, Finucane B, Latham GJ, Hadd A, Berry-Kravis E, Tassone F. AGG interruptions and maternal age affect *FMR1* CGG repeat allele stability during transmission. J Neurodev Disord. 2014;6:24.
27. Peprah E. Fragile X syndrome: The *FMR1* CGG repeat distribution among world populations. Ann Hum Genet. 2012; 76:178-191.
28. Loesch DZ, Tassone F, Lo J, Slater HR, Hills LV, Bui MQ, Silburn PA, Mellick GD. New evidence for, and challenges in, linking small CGG repeat expansion *FMR1* alleles with Parkinson's disease. Cli Genet. 2013; 84:382-385.
29. Yuhas J, Walichiewicz P, Pan R, Zhang W, Casillas EM, Hagerman RJ, Tassone F. High-risk fragile x screening in Guatemala: Use of a new blood spot polymerase chain reaction technique. Genet Test Mol Biomarkers. 2009; 13:855-859.
30. Winarni TI, Utari A, Mundhofir FE, Tong T, Durbin-Johnson B, Faradz SM, Tassone F. Identification of expanded alleles of the *FMR1* gene among high-risk population in Indonesia by using blood spot screening. Genet Test Mol Biomarkers. 2012; 16:162-166.
31. Fernandez-Carvajal I, Walichiewicz P, Xiaosen X, Pan R, Hagerman PJ, Tassone F. Screening for expanded alleles of the *FMR1* gene in blood spots from newborn males in a Spanish population. J Mol Diagn. 2009; 11:324-329.
32. Tassone F, Iong KP, Tong TH, Lo J, Gane LW, Berry-Kravis E, Nguyen D, Mu LY, Laffin J, Bailey DB, Hagerman RJ. *FMR1* CGG allele size and prevalence

ascertained through newborn screening in the United States. Genome Med. 2012; 4:100.
33. Loesch DZ, Bui QM, Huggins RM, Mitchell RJ, Hagerman RJ, Tassone F. Transcript levels of the intermediate size or grey zone fragile X mental retardation 1 alleles are raised, and correlate with the number of CGG repeats. J Med Genet. 2007; 44:200-204.
34. Hall D, Tassone F, Klepitskaya O, Leehey M. Fragile X-associated tremor ataxia syndrome in *FMR1* gray zone allele carriers. Mov Disord. 2012; 27:296-300.
35. Seltzer MM, Baker MW, Hong J, Maenner M, Greenberg J, Mandel D. Prevalence of CGG expansions of the *FMR1* gene in a US population-based sample. Am J Med Genet B Neuropsychiatr Genet. 2012; 159B:589-597.
36. Agresti A. An introduction to categorical data analysis Hoboken, NJ: Wiley-Interscience; 2007; 2nd:[xvii, 372 p. ill. 25 cm.].
37. Team RC. R: A language and environment for statistical computing. 3.0.3 ed. Vienna, Austria: R Foundation for Statistical Computing; 2014.
38. Faradz SM, Pattiiha MZ, Leigh DA, Jenkins M, Leggo J, Buckley MF, Holden JJ. Genetic diversity at the *FMR1* locus in the Indonesian population. Ann Hum Genet. 2000; 64:329-339.
39. Faradz SM, Leggo J, Murray A, Lam-Po-Tang PR, Buckley MF, Holden JJ. Distribution of *FMR*1 and *FMR2* alleles in Javanese individuals with developmental disability and confirmation of a specific AGG-interruption pattern in Asian populations. Ann Hum Genet. 2001; 65:127-135.
40. Chiu HH, Tseng YT, Hsiao HP, Hsiao HH. The AGG interruption pattern within the CGG repeat of the *FMR1* gene among Taiwanese population. J Genet. 2008; 87:275-277.
41. Larsen LA, Vuust J, Nystad M, Evseeva I, Van Ghelue M, Tranebjaerg L. Analysis of *FMR1* $(CGG)_n$ alleles and DXS548-FRAXAC1 haplotypes in three European circumpolar populations: Traces of genetic relationship with Asia. Eur J Hum Genet. 2001; 9:724-727.
42. Larsen LA, Armstrong JS, Grønskov K, Hjalgrim H, Brøndum-Nielsen K, Hasholt L, Nørgaard-Pedersen B, Vuust J. Analysis of *FMR1* $(CGG)_n$ alleles and FRAXA microsatellite haplotypes in the population of Greenland: Implications for the population of the New World from Asia. Eur J Hum Genet. 1999; 7:771-777.
43. Crawford DC, Meadows KL, Newman JL, Taft LF, Scott E, Leslie M, Shubek L, Holmgreen P, Yeargin-Allsopp M, Boyle C, Sherman SL. Prevalence of the fragile X syndrome in African-Americans. Am J Med Genet. 2002; 110:226-233.
44. Penagarikano O, Gil A, Télez M, Ortega B, Flores P, Veiga I, Peixoto A, Criado B, Arrieta I. A new insight into fragile X syndrome among Basque population. Am J Med Genet A. 2004; 128A:250-255.
45. Angeli CB, Capelli LP, Auricchio MT, Leal-Mesquita ER, Ribeiro-dos-Santos AK, Ferrari I, Oliveira SF, Klautau-Guimarães Mde N, Vianna-Morgante AM,Mingroni-Netto RC. AGG interspersion patterns in the CGG repeat of the *FMR1* gene and linked DXS548/FRAXAC1 haplotypes in Brazilian populations. Am J Med Genet A. 2005; 132A:210-214.
46. Peprah EK, Allen EG, Williams SM, Woodard LM, Sherman SL. Genetic diversity of the fragile X syndrome gene (*FMR1*) in a large Sub-Saharan West African population. Ann Hum Genet. 2010; 74:316-325.
47. Chen LS, Tassone F, Sahota P, Hagerman PJ. The $(CGG)_n$ repeat element within the 5' untranslated region of the *FMR1* message provides both positive and negative cis effects on *in vivo* translation of a downstream reporter. Hum Mol Genet. 2003; 12:3067-3074.

(Received October 31, 2014; Accepted November 28, 2014)