Hum Genet (1997) 101 : 214–218

ORIGINAL INVESTIGATION

**Mark C. Hirst · Tadao Arinami · Charles D. Laird**

# Sequence analysis of long FMR1 arrays in the Japanese population: insights into the generation of long $(CGG)_n$ tracts

Received: 18 March 1997 / Accepted: 18 July 1997

**Abstract** The human fragile-X syndrome is associated with expansions of a $(CGG)_n$ triplet repeat within the FMR1 gene. Whilst normal FMR1 arrays consist of variable numbers of $(CGG)_{7–13}$ blocks punctuated with single AGG triplets, unstable arrays contain longer blocks of uninterrupted $(CGG)_n$. The degree of instability, and subsequent risk of expansion to the fragile-X mutation, is dependent upon the length of this uninterrupted repeat. Detailed analyses of normal FMR1 array structures suggest that longer uninterrupted blocks of repeat could arise either through a process of gradual slippage or a more dramatic loss of an intervening AGG triplet. Up to 15% of Japanese and Chinese individuals have FMR1 triplet arrays centred on 36 repeats in length, a modal group not found in Caucasians. As longer FMR1 arrays have been associated with high-risk fragile-X haplotypes in some populations, we investigated the nature of these larger arrays. Sequence analysis revealed that the unusual length is due to the presence of a novel $(CGG)_6$ block within the array. Several haplotypically related arrays contain blocks of $(CGG)_{16}$ or $(CGG)_{15}$, consistent with the fusion of adjacent $(CGG)_9$ and $(CGG)_6$ blocks after loss of the intervening AGG triplet. This is compatible with inferences from the Caucasian population that AGG loss is a mechanism by which long blocks of identical repeats are generated.

M. C. Hirst (✉)
Institute of Molecular Medicine, The John Radcliffe,
Headley Way, Headington, Oxford, OX3 9DS, UK

M. C. Hirst · C. D. Laird
Program in Molecular Medicine,
Fred Hutchinson Cancer Research Center, 1124 Columbia Street,
Seattle, WA 98104, USA

T. Arinami
Department of Medical Genetics, Institute of Basic Sciences,
University of Tsukuba, Tsukuba, Ibaraki, 305 Japan

## Introduction

Fragile-X syndrome is the most common form of inherited mental handicap after Down syndrome. It is caused through a dramatic expansion of an unstable triplet repeat (Fu et al. 1991; Oberlé et al. 1991; Verkerk et al. 1991; Yu et al. 1991) which results in loss of gene transcription (Pierreti et al. 1991). The triplet array lies within the 5′-untranslated portion of the FMR1 gene (Verkerk et al. 1991). Its length varies within the normal population from 6 to 52 copies and the distribution of array lengths is multi-modal (Fu et al. 1991), a feature caused by the underlying compound nature of the array. In most arrays, two or three smaller $(CGG)_{7–13}$ blocks are interspersed with single AGG triplets, giving a symmetrical and highly ordered modular structure with major modal group lengths exhibiting a ten-repeat periodicity (Hirst et al. 1994; Kunst and Warren 1994; Snow et al. 1994; Zhong et al. 1995). Allelic diversity results from the variable number and length of these $(CGG)_{7–13}$ blocks.

In contrast to normal, stable arrays, fragile-X premutation chromosomes carry arrays longer than 54 repeats that are either entirely uninterrupted or have long portions of $(CGG)_n$ at their 3′ end (Eichler et al. 1994; Hirst et al. 1994; Snow et al. 1994; Zhong et al. 1995). Expansion to fragile-X mutation length (> 200 repeats) appears to occur exclusively within the uninterrupted repeat, and the degree of array instability is related to its length (Eichler et al. 1994; Snow et al. 1994). Several unstable arrays, not known to be associated with fragile-X syndrome, have 34 and 31 perfect CGG repeats, demonstrating that a low level of instability exists below the length normally considered as a premutation (Eichler et al. 1994; Snow et al. 1994). Many other arrays within the normal size range carry long portions of $(CGG)_n$ at their 3′ end, with over 10% having $(CGG)_{>17}$ (Hirst 1995). The similarity in their structure with unstable and premutation arrays and their association with certain high-risk fragile-X haplotypes has led to the suggestion that some of these may be precursors for recurrent expansion into the premutation range

(Hirst et al. 1994; Kunst and Warren 1994; Snow et al. 1994). Little is actually known about how uninterrupted arrays arise but, based upon the observation that arrays with long 3′ $(CGG)_n$ blocks are often of modal repeat length, it is assumed that the loss of an interrupting AGG triplet is involved (Hirst et al. 1994). This is supported by larger population studies of interspersion pattern and haplotype analysis which confirm a strong 3′ bias for the loss of the AGG (Eichler et al. 1995). This is also supported by studies of the Bornean and Mandenka populations, where loss of the most 3′ AGGs from a common progenitor allele has occurred (Kunst et al. 1996). Further studies have suggested that fragile-X alleles have arisen through two independent pathways; one through loss of AGG from arrays with an asymmetric AGG interspersion pattern and a second in alleles carrying two AGG interspersions through a more gradual slippage within the most 3′ $(CGG)_n$ block (Eichler et al. 1996).

Population and ethnic variations in length of FMR1 arrays might provide insights into the origins of unstable arrays and the mechanisms by which fragile-X mutations arise. Whilst the distributions of FMR1 array lengths are similar for most human ethnic groups studied (Fu et al. 1991; Kunst et al. 1996), the Japanese and Chinese populations are an exception (Arinami et al. 1993; Richards et al. 1994; Zhong et al. 1994). Two major modal groups of array lengths of 21 and 29–30 repeats exist in the Caucasian population. In the Japanese and Chinese populations, the shorter length group is reduced in frequency and the 29–30 group accounts for almost 70% of individuals. In addition, a third modal group of 36 repeats (the 36-modal group) is found in 7–15% of individuals. These longer arrays are notably shifted from the ten-repeat periodicity observed in the other major modal groups. Haplotype analysis of the Japanese fragile-X population has shown evidence for founder chromosomes, but the association is with a different haplotype from that found in Caucasians (Richards et al. 1994). Whilst some association of array lengths and haplotypes was found, alleles longer than 31 repeats appear not to be uniquely associated with the high-risk fragile-X haplotypes (Richards et al. 1994). The Japanese and Chinese populations have a very low allele diversity for the markers analysed, however, and an association between certain FMR1 array lengths and high-risk haplotypes might not have been detected. In order to characterise these novel length FMR1 alleles, we have sequenced 21 arrays from unrelated individuals in the Japanese 36-modal group and find that most contain a novel $(CGG)_6$ block of repeats. These data explain both the shift from the common population modal lengths and provide evidence of AGG loss in the generation of longer, uninterrupted arrays.

## Materials and methods

### PCR and sequence analysis

Genomic DNAs from 21 normal unrelated individuals in the 36-modal group were selected from the 24 identified in a previous study of 370 Japanese males (Arinami et al. 1993). PCR amplification and sequencing were carried out essentially as described previously, except that cycling conditions were modified for an MJ Research thermal cycler (Hirst et al. 1994). Amplification conditions were an initial denaturation at 104° C for 5 min, followed by 35 cycles of 104° C for 30 s and 70° C for 10 min. Each 20-µl reaction contained 20 mM TRIS-HCl (pH 8.8), 10 mM KCl, 1.5 mM $MgCl_2$, 10 µM $(NH_4)_2SO_4$, 0.1% Triton X-100, 100 µg/ml BSA, 0.5 µM each oligonucleotide (721 and 723), 200 µM each dNTP, 5% DMSO and 1 U wild-type *Pfu* polymerase (Stratagene). Products were purified through 2% low melting point agarose and isolated after Gelase solublisation (Epicenter Technologies) with Wizard PCR Prep reagents (Promega). One-tenth of each product was then sequenced with $^{32}P$ end-labelled primers 172 and 170 using the exo$^-$ *Pfu* cyclist kit (Stratagene) under conditions suitable for an MJ Research thermal cycler (30 cycles of 104° C, 30 s denaturation; 70° C, 1 min annealing/elongation). The products were resolved in 6% denaturing polyacrylamide gels (National Diagnostics) and visualised by autoradiography. All oligonucleotide sequences and position numbers are taken from HSFXDNA (Genbank):

721: 5′-AGCCCCGCACTTCCACCACCAGCTCCTCCA (complementary to 2617–2647);

723: 5′-TTCACTTCCGGTGGAGGGCCGCCTCTGAGC (2876–2838);

170: 5′-GGCGGTGACGGAGGCGCC (2678–2695);

171; 5′-CCTGCTAGCGCCGGGAGC (complementary to 2807–2824).

### Haplotype analysis

Alleles for the dinucleotide repeat FRAXAC1 were determined by PCR amplification and denaturating gel electrophoresis as described previously (Hirst et al. 1993).

## Results

### Sequence of the 36-modal group FMR1 arrays

The sequences of 21 FMR1 trinucleotide arrays from the 36-modal group of Japanese alleles were obtained by direct sequence analysis of PCR amplification products (Fig. 1). A summary of the array lengths, internal structures and flanking haplotypes is shown in Table 1. The array lengths in this analysis were found to be one triplet longer than the original estimates of Arinami et al. (1993), which were based upon studies by Fu et al. (1991). These have since been adjusted by one additional triplet repeat (Kunst and Warren 1994). All the arrays analysed in the 36-modal group have a 5′ $(CGG)_9$ block and 15 of the 21 have an identical structure consisting of $(CGG)_9AGG(CGG)_9AGG(CGG)_6AGG(CGG)_9$ (abbreviated to 9A9A6A9). An internal $(CGG)_6$ block, plus an additional interspersed AGG triplet, results in the 36-modal group being seven repeats longer than the common 29-repeat allele. Six arrays contained a block of repeats 15 or more in length, with 16 being the most common length. Four arrays have this longer stretch of uninterrupted repeat in the middle of the array, whilst two are at the 3′ end including one $(CGG)_{26}$ (Table 1).

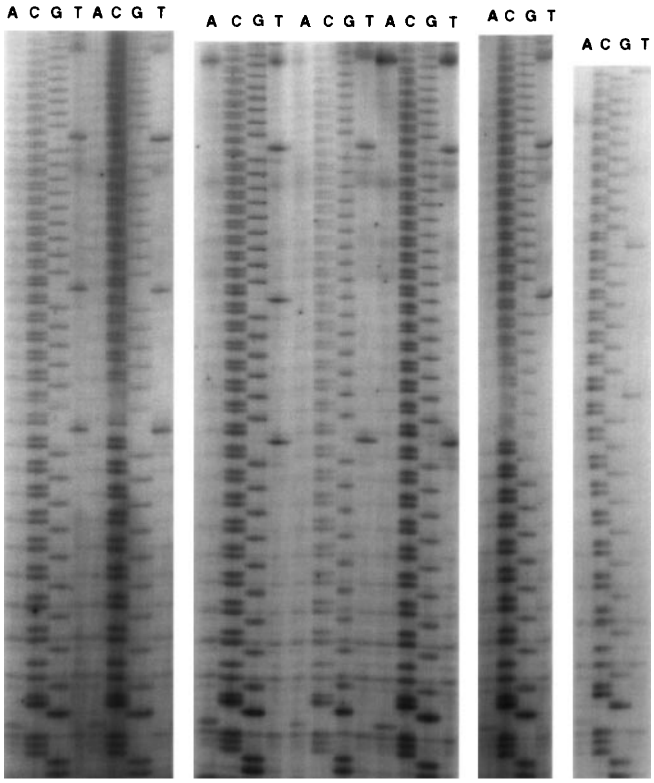

**Fig. 1** Sequence analysis of seven FMR1 arrays. Arrays are shown sequenced with the 172 primer from the 3′ end of the array; thus, the triplet repeat array is read as CCG and the interspersed triplet as CCT. From *left to right*, arrays from the 36-modal group are samples 6(9A9A6A9), 10 (9A9A6A9), 13 (9A9A6A9), 24 (9A16A9), 9 (9A16A9), 12 (9A9A16). One repeat array of 27 repeats with a structure of 10A6A9 (sample 22) is shown on the *far right*

### Haplotype analysis

The flanking FRAXAC1 dinucleotide repeat was analysed to determine the haplotypic background for this group of arrays. With one exception, all the arrays analysed are associated with the D allele, confirming that this group of arrays are haplotypically related. This is in agreement with previously published data (Arinami et al. 1993; Richards et al. 1994).

### The presence of $(CGG)_6$ in other arrays

To investigate the presence of the unusual $(CGG)_6$ block in the Japanese population, we examined the possibility that it might be present within other arrays. A ten-repeat periodicity from the 36-modal group length would give arrays with lengths of 26 or 46 repeats. The Japanese and Chinese populations, however, show no modal groups of these lengths. Only one array in either of these size ranges (sample 22; 27 repeats) was available for investigation and this was found to contain a $(CGG)_6$ block with a structure of 10A6A9 (Table 1). FRAXAC1 typing of this chromosome demonstrated that it was associated with the C allele. Interestingly, this allele is strongly associated with the 30-repeat arrray in the Japanese population which commonly has the structure 10A9A9 (M. C. Hirst, unpublished data). It is possible that this C-10A6A9 (FRAXAC1/FMR1) array configuration was generated by recombination between a 36-repeat (D-9A9A6A9) and a 30-repeat (C-10A9A9) array resulting in an array with the 5′FRAXAC1 allele and a hybrid FRM1 triplet array structure.

**Table 1** Summary of array lengths and structures and flanking haplotypes of 21 FMR1 trinucleotide arrays. FMR1 array structures are abbreviated to represent the number of CGG triplets and the position of the interspersed AGG. For example, 9A9A6A9 represents an array with an internal structure 5′$(CGG)_9$-(**A**GG)-$(CGG)_9$-(**A**GG)-$(CGG)_6$-(**A**GG)-$(CGG)_9$

| Sample | Array length | Array structure | FRAXAC1 |
|---|---|---|---|
| 6 | 36 | 9A9A6A9 | C |
| 5 | 35 | 9A15A9 | D |
| 7 | 36 | 9A9A6A9 | D |
| 8 | 36 | 9A16A9 | D |
| 9 | 36 | 9A16A9 | D |
| 10 | 36 | 9A9A6A9 | D |
| 11 | 36 | 9A26 | D |
| 12 | 36 | 9A9A16 | D |
| 13 | 36 | 9A9A6A9 | D |
| 15 | 36 | 9A9A6A9 | D |
| 16 | 36 | 9A9A6A9 | D |
| 17 | 36 | 9A9A6A9 | D |
| 18 | 36 | 9A9A6A9 | D |
| 19 | 36 | 9A9A6A9 | D |
| 20 | 36 | 9A9A6A9 | D |
| 21 | 36 | 9A9A6A9 | D |
| 23 | 36 | 9A9A6A9 | D |
| 24 | 36 | 9A16A9 | D |
| 25 | 36 | 9A9A6A9 | D |
| 26 | 36 | 9A9A6A9 | D |
| 27 | 36 | 9A9A6A9 | D |
| 22 | 27 | 10A6A9 | C |

## Discussion

We have investigated the molecular structure of FMR1 arrays belonging to the 36-repeat modal group of the Japanese population. They contain a $(CGG)_6$ block, which accounts for their unusual length and the shift from the common modal length of 29 repeats. The occurrence of longer blocks of 15 or 16 repeats within a group of haplotypically related arrays normally carrying adjacent $(CGG)_6$ and $(CGG)_9$ is strong evidence that they have arisen by mutation or deletion of the intervening AGG triplet (see Fig. 2). Furthermore, blocks of $(CGG)_{16}$ are consistent with the fusion of the adjacent $(CGG)_6$ and $(CGG)_9$ blocks through mutation of the intervening *A*GG to *C*GG, resulting in no overall change in array length. The array with $(CGG)_{26}$ could have been generated through further mutation of a second AGG triplet (Fig. 2). The array carrying a $(CGG)_{15}$ could have arisen from a deletion of the intervening AGG triplet

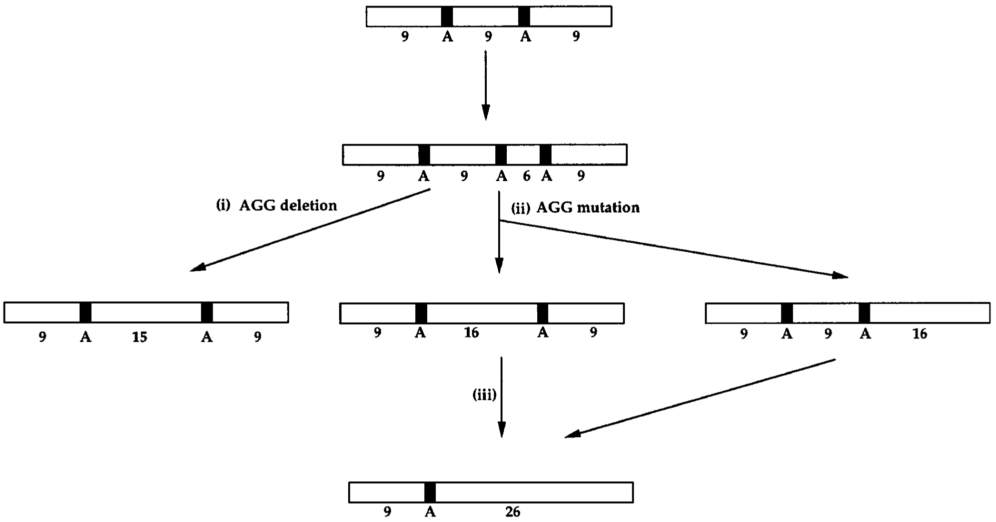

**Fig. 2** The suggested relationship between FMR1 triplet arrays identified in the study of the Japanese 36-modal group. A progenitor 36-repeat array could have been generated by insertion of a AGG(CGG)$_6$ block of repeats on a 29-repeat D-9A9A9 array through replication slippage or an unequal cross-over event. The loss of the interspersed AGG triplet by *(i)* deletion, or *(ii)* mutation at the proximal or distal site would result in a longer block of (CGG)$_{15}$ or (CGG)$_{16}$, respectively. Further loss of AGG would lead to the generation of longer 3′(CGG)$_{26}$, a possible precursor for further expansion into the fragile-X premutation range

in a 9A9A6A9 array. If deletion of the AGG were the predominant mechanism of loss, we would expect to observe more arrays carrying (CGG)$_{15}$. The numbers of alleles in this study are too small to draw any firm conclusions regarding the relative frequency of these mutation events.

Interestingly, in this group of alleles, most longer (CGG)$_{16}$ are internal to the array, contrasting strongly with the 3′ polarity with which they occur in the Caucasian population (Hirst et al. 1994; Eichler et al. 1995). This difference might be specifically influenced by the presence of the (CGG)$_6$ block itself. Indeed, Eichler et al. (1996) suggested that the asymmetry of AGG interspersion within an array might predispose to its loss. Our data on the 36-modal group support this conclusion and suggest that loss of the proximal AGG might occur more frequently than the distal AGG (Fig. 2). This might be influenced by secondary structure formed within replication intermediates or a failure to detect mispaired bases.

Small modal groups centred on 36 repeats have also recently been found in the Bornean and Tibetan populations and sequence analysis identified a (CGG)$_6$ in some FMR1 alleles (Kunst et al. 1996). Thus the (CGG)$_6$ appears restricted to Asian lineages, suggesting that it most likely arose after divergence of these from the Caucasian lineage 150 000–200 000 years ago (Bowcock et al. 1994; Horai et al. 1995). The association with the FRAXAC1-D allele and the presence of a 5′(CGG)$_9$ strongly suggests that it arose from the common 29-repeat D-9A9A9 allele found in all lineages (Fig. 2). This might have occurred through an unequal exchange or recombination event or through replication slippage.

The significantly different haplotype frequencies between fragile-X and normal chromosomes suggest that many fragile-X chromosomes have descended from a small number of founder mutations, although this could also reflect recurrent expansion from a pool of precursor arrays (Richards et al. 1992; Hirst et al. 1993; Oudet et al. 1993). An association of longer-than-average FMR1 arrays with high-risk haplotypes indicated that these founder chromosomes might originate from normal arrays with a higher repeat number (Richards et al. 1992; Jacobs et al. 1993; Oudet et al. 1993). Sequence data indicate that, whilst overall length is an important factor, it is the length of uninterrupted (CGG)$_n$ that is a crucial factor in determining instability. Arrays with long 3′ blocks of (CGG)$_n$ occur more frequently on chromosomes carrying high-risk haplotypes (Hirst et al. 1994; Snow et al. 1994). Their similarity with premutation arrays suggests that these might represent precursor arrays that could progress to longer stretches of uninterrupted repeat through slippage or further loss of an AGG triplet. Thus, the generation of these alleles carrying longer than average (CGG)$_n$ may in itself represent a founder event. Almost 75% of arrays in the 36-modal group have an internal AGG repeat and are unlikely to carry any significant risk of expansion. This confirms observations from haplotype analysis, which showed that when taken as a single grouping, arrays longer than 31 repeats are not associated with the Japanese high-risk fragile-X haplotypes (Richards et al. 1994). However, one group of alleles with the FRAXAC1-D haplotype was strongly associated with fragile-X syndrome in the Japanese population (Richards et al. 1994). It is possible that this small subgroup could have arisen from the 36-modal group. Further studies are currently being performed with additional markers, although the level of allelic diversity in the Japanese and Chinese populations at flanking dinucleotide markers could mask such associations (Richards et al. 1994; Zhong et al. 1994).

In summary, we have shown that arrays within the Asian-specific 36-modal group have a novel structure. The presence of longer (CGG)$_n$ blocks strongly suggests that loss of the interspersed AGG triplet occurs relatively frequently on these asymmetric alleles. This variation in normal length arrays might reflect the mechanism whereby longer founder fragile-X precursor chromosomes arose by additional AGG loss or slippage. A more extensive investigation of haplotypes and FMR1 array structures in the Japanese population, and of fragile-X arrays will be necessary further to address this question.

**Acknowledgements** M. C. H. is funded by the Wellcome Trust and was supported as a Visiting Investigator by the Fred Hutchinson Cancer Research Center (FHCRC), Seattle. Support was also provided by the FHCRC, NIH (C. L.) and the Ministry of Education, Science and Culture of Japan (T. A.). We thank colleagues at the FHCRC and Dr. Rosalind Harding (IMM) for many useful discussions and comments.

## References

Arinami T, Asano M, Kobayashi K, Yanagi H, Hamaguchi H (1993) Data on the CGG repeat at the fragile-X site in the non-retarded Japanese population and family suggest the presence of a subgroup of normal alleles predisposing to mutate. Hum Genet 92 : 431–436

Bowcock A, Ruiz-Linares A, Minch E, Kidd K, Cavalli-Sforza L (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368 : 455–457

Eichler E, Holden J, Popovich B, Reiss A, Snow K, Thibodeau S, Richards C, Ward P, Nelson D (1994) Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Nat Genet 8 : 88–94

Eichler E, Hammond H, MacPherson J, Ward P, Nelson D (1995) Population survey of the human FRM1 CGG repeat substructure suggests biased polarity for the loss of AGG interruptions. Hum Mol Genet 4 : 2199–2208

Eichler E, MacPherson J, Murray A, Jacobs P, Chakravarti A, Nelson D (1996) 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 5 : 319–330

Fu Y, Kuhl D, Pizzuti A, Piereti M, Sutcliffe J, Richards S, Verkerk A, Holden J, Fenwick R, Warren S, Oostra B, Nelson D, Caskey C (1991) Variation of the CGG repeat at the fragile-X site results in genetic instability: resolution of the Sherman paradox. Cell 61 : 1–20

Hirst M (1995) FMR1 triplet arrays: paying the price of perfection. J Med Genet 32 : 761–763

Hirst M, Knight S, Christodoulou Z, Grewal P, Fryns J, Davies K (1993) Origins of the fragile-X syndrome mutation. J Med Genet 30 : 647–650

Hirst M, Grewal P, Davies K (1994) Precursor arrays for triplet repeat expansion at the fragile-X locus. Hum Mol Genet 3 : 1553–1560

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequence analysis of mitochondrial DNAs. Proc Natl Acad Sci USA 92 : 532–536

Jacobs P, Bullman H, MacPherson J, Youings S, Rooney V, Watson A, Dennis N (1993) Population studies of the fragile X: a molecular approach. J Med Genet 30 : 454–459

Kunst C, Warren S (1994) Cryptic and polar variation of the fragile-X repeat could result in predisposing normal alleles. Cell 77 : 853–861

Kunst C, Zerylnick C, Karickhoff L, Eichler E, Bullard J, Chalifoux M, Holden J, Torroni A, Nelson D, Warren S (1996) FMR1 in global populations. Am J Hum Genet 58 : 513–522

Oberlé I, Rousseau F, Heitz D, Kretz C, Devys D, Hanauer A, Boue J, Bertheas M, Mandel J (1991) Instability of a 550-bp DNA fragment and abnormal methylation in fragile-X syndrome. Science 252 : 1097–1102

Oudet C, Mornet E, Serre J, Thomas F, Lentes-Zengerling S, Kretz C, Deluchat C, Tejada I, Boue A, Mandel J (1993) Linkage disequilibrium between the fragile-X mutation and two closely linked CA repeats suggests that fragile-X chromosomes are derived from a small number of founder chromosomes. Am J Hum Genet 52 : 297–304

Pierreti M, Zhang F, Fu Y, Warren S, Oostra B, Caskey C, Nelson D (1991) Absence of expression of the FMR1 gene in fragile-X patients. Cell 66 : 817–822

Richards R, Holman K, Friend K, Kremer E, Hillen D, Staples A, Brown W, Goonewardena P, Tarleton J, Schwartz C, Sutherland G (1992) Evidence of founder chromosomes in fragile-X syndrome. Nat Genet 1 : 257–260

Richards R, Kondo I, Holman K, Yamauchi M, Seki N, Kishi K, Staples A, Sutherland G, Hori T (1994) Haplotype analysis at the FRAXA locus in the Japanese population. Am J Med Genet 51 : 412–416

Snow K, Tester D, Kruckeberg K, Schaid D, Thibodeau S (1994) Sequence analysis of the fragile-X trinucleotide repeat: implications for the origin of the fragile-X mutation. Hum Mol Genet 3 : 1543–1551

Verkerk A, Pieretti M, Sutcliffe J, Fu Y, Kuhl D, Pizzuti A, Reiner O, Richards S, Victoria M, Zhang F, Eussen B, Ommen G van, Blonden L, Riggins G, Chastain J, Kunst C, Galjaad H, Caskey C, Nelson D, Oostra B, Warren S (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile-X syndrome. Cell 65 : 905–914

Yu S, Pritchard M, Kremer E, Lynch M, Nancarrow J, Baker E, Holman K, Mulley J, Warren S, Schlessinger D, Sutherland G, Richards R (1991) Fragile-X genotype characterised by an unstable region of DNA. Science 252 : 1179–1181

Zhong N, Liu X, Gou S, Houck G, Li S, Dobkin C, Brown W (1994) Distribution of FMR-1 and associated microsatellite alleles in a normal Chinese population. Am J Med Genet 51 : 417–422

Zhong N, Yang W, Dobkin C, Brown W (1995) Fragile-X gene instability: anchoring AGGs and linked microsatellites. Am J Hum Genet 57 : 351–361