 Human Molecular Genetics, 1996, Vol. 5, No. 3 **319–330**

## ARTICLE

# Haplotype and interspersion analysis of the FMR1 CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome

**Evan E. Eichler, James N. Macpherson[1], Anna Murray[1], Patricia A. Jacobs[1], Aravinda Chakravarti[2] and David L. Nelson***

Department of Molecular and Human Genetics, Human Genome Center, Baylor College of Medicine, Houston, TX 77030, USA, [1]Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, Wiltshire SP2 8BJ, UK and [2]Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-495, USA

Received October 24, 1995; Revised and Accepted December 20, 1995

**To understand the origins of the fragile X syndrome and factors predisposing alleles to instability and hyperexpansion, we have compared the haplotype (using markers FRAXAC1, FRAXAC2, and DXS548) and AGG interspersion patterns of the FMR1 CGG repeat for 214 normal and 16 premutation chromosomes. Association testing between interspersion pattern and haplotype reveals a highly significant ($P$ <0.002) non-random distribution, indicating that all three markers are useful in phylogenetic reconstruction of mutational change. Parsimony analysis of the FMR1 CGG repeat substructure predicts that loss of AGG interruptions has occurred independently on many haplotypes associated with the fragile X syndrome, partially explaining the haplotype diversity of this disease. Among haplotypes found in linkage disequilibrium with the fragile X mutation, two different modes of mutation and predisposition to instability have been identified. One pathway has involved the frequent and recurrent loss of AGG interruptions from rare asymmetrical ancestral array structures. Intergenerational transmission studies suggest that these predisposed chromosomes progress relatively rapidly to the disease state. In contrast, the second mutational pathway involves a single haplotype which has maintained two AGG interruptions. Parsimony analysis of CGG repeat substructure within this haplotype suggests that larger alleles have been generated by gradual increments of CGG repeats distal to the most 3′ interruption. Pedigree analysis of the intergenerational stability of alleles of this haplotype confirms a gradual progression toward instability thresholds. As a result, a large reservoir of chromosomes carrying large repeats on this haplotype exists. These chromosomes are predisposed to disease. The present data support a model in which there are at least two different mutational pathways predisposing alleles to instability and hyperexpansion associated with the fragile X syndrome.**

## INTRODUCTION

Investigations into the origin of the fragile X syndrome have traditionally involved the use of three polymorphic loci (DXS548, FRAXAC1, FRAXAC2) found within 150 kb of the FMR1 CGG repeat (1,2) (Fig. 1). Haplotype analysis of normal and mutant chromosomes among diverse population groups has indicated that a substantial proportion of all fragile X chromosomes shows linkage disequilibrium with a small subset of DXS548-FRAXAC2-FRAXAC1 haplotypes (3–15). In more genetically isolated populations (such as Finland), linkage disequilibrium has been shown to be even more pronounced with 75% of all fragile X chromosomes occurring on single rare DXS548-FRAXAC2 haplotype (9,12). Genealogical studies of fragile X kindreds previously considered to be unrelated further confirmed the existence of founder effects for the fragile X syndrome, by demonstrating common descent for the fragile X chromosome from shared ancestors in the 17th and 18th centuries (11,16,17). These findings suggested that FMR1 CGG repeat alleles predisposed to the development of the fragile X syndrome could be carried silently through human pedigrees from five to 100 generations (4,12,15) prior to hyperexpansion and

*To whom correspondence should be addressed

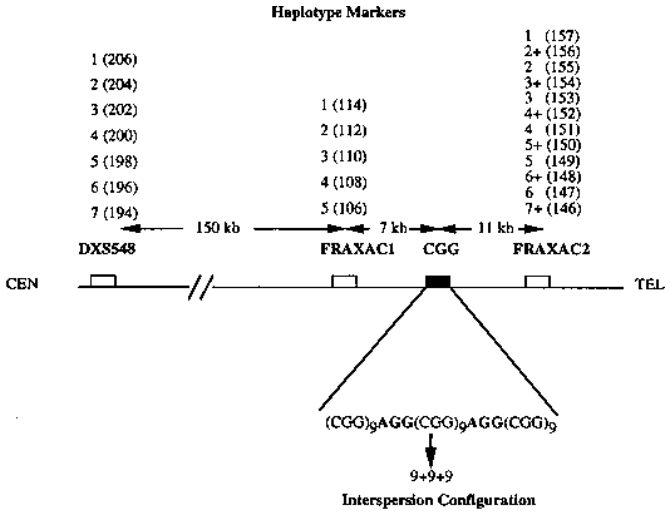


**Figure 1.** Interspersion and haplotype nomenclature. Summarized are the nomenclature and position of the haplotype markers used in this study, relative to the AGG-interspersed FMR1 CGG repeat. A previously described convention is employed to designate the genotypes of the various haplotype markers (see Materials and Methods) (10). Haplotype alleles are designated using a number in descending order of the length of the PCR product. Composite haplotypes are used in this study and are always indicated in the order, DXS548-FRAXAC1-FRAXAC2 (i.e. 7–3–4+ represents a haplotype which consists of DXS548 allele 7, FRAXAC1 allele 3, and FRAXAC2 allele 4+). A canonical FMR1 CGG repeat sequence (29 repeats) is shown with an AGG interruption once every nine CGG repeats. Such an allele is symbolized as 9+9+9.

hypermethylation associated with disease (18–22). The conclusion that new mutations are rare was difficult to reconcile with an X-linked dominant disorder in which the affected males rarely reproduce (3,15,23).

Based on initial linkage disequilibrium data coupled with the observation that no transition from a normal FMR1 CGG repeat allele (5–50 repeats) (18,24) to an unstable premutation allele (59–200 repeats) had ever been observed in fragile X pedigrees, Morton and Macpherson proposed a multi-allelic model to account for the latency of the mutation and the origin of the disease (25). The transition from a normal stable allele to a fully expanded FMR1 CGG repeat allele (>200 repeats) was initially postulated to occur through progression between four defined allelic states. A normal stable FMR1 CGG repeat allele (N) would infrequently become predisposed to modest instability as a small, yet relatively stable, insert (S) beyond 40 repeats in length. Such predisposed alleles could be maintained in the population for as many as 90 generations (15) before progressing to a larger unstable (Z) or premutation allele. The conversion of an unstable (Z) allele to a full mutation (L) occurs very rapidly requiring a female germline transmission. Implicit in this and other multi-step progression models (25–27) was a gradual increase in the size of the FMR1 CGG repeat as it progressed from N to S to Z.

Due to the longevity of predisposed (S) alleles in the human population, one prediction of the multi-step progression models is that haplotypes in linkage disequilibrium with the fragile X mutation should show an enrichment for alleles with longer lengths of repeat. Although one specific set of haplotypes appears to be enriched for high-end normal alleles (3,5,6), the majority of fragile X associated haplotypes do not demonstrate significant increases in total repeat length among pools of normal alleles (3,4,8,10,13). These results suggested that other haplotype-specific influences, either flanking the repeat or intrinsic to the repeat itself, were important in determining an allele's predisposition to instability and disease (3,10,14,28).

Several groups have investigated the possibility that the AGG substructure of the FMR1 CGG repeat may play a critical role in determining stability and susceptibility to disease (29–32). The results of these analyses revealed that the longest tract of pure (uninterrupted) CGG repeats determines instability at the FRAXA locus, with unstable intergenerational transmissions being observed at a threshold length of 34–37 pure repeats (30). Comparisons of interspersion pattern and chromosomal haplotypes indicated that disease haplotypes may be enriched for longer tracts of pure repeats (29). Other groups, however, have failed to find a significant correlation between the length of the longest tract of pure repeat lengths and haplotypes at-risk for the fragile X syndrome (31,32). From previous studies comparing haplotype and AGG interspersion patterns it has been difficult to reconstruct the phylogeny of fragile X lineages due to the limited number of premutation alleles studied (29,32). In addition, cross-study comparisons with other linkage disequilibrium data have been hampered by the consideration of only one or two of the microsatellite markers in the region (31,32).

In a comprehensive study considering all three haplotype markers (DSX548, FRAXAC1 and FRAXAC2), two groups of haplotypes which showed linkage disequilibrium with the fragile X mutation were distinguished among Caucasians (10). One group represented solely by DSX548-FRAXAC1-FRAXAC2 haplotype (2–1–3) accounted for 14% of all fragile X chromosomes (Table 1) and was found to be significantly enriched for high-end normal alleles (10). The second group, which accounted for approximately 30% of all fragile X cases (haplotypes 6–4–4 and 6–4–5; Table 1), demonstrated no such enrichment for lengthy normal alleles. It was suggested that these different properties among at-risk haplotypes might indicate different mutational pathways for the origin of the disease. It was proposed that the 2–1–3 haplotype represented an 'ancient' fragile X lineage whose intermediate (S) alleles had reached appreciable frequencies in the population (10,25). In contrast, the 6–4–5 and 6–4–4 haplotypes, as well as many of the other haplotypes found in equilibrium in the general and fragile X populations (Table 1), might be subject to a recurrent 'leap-frog' mutational event, which allowed these alleles to progress relatively rapidly toward instability and hyperexpansion thresholds associated with the disease. Although such mutational events would explain the disparity in distribution of high-end normal alleles among fragile X haplotypes as well as the considerable haplotype diversity associated with the disease, the molecular basis of these different mutational pathways was unknown.

Due to our observation that the longest tract of pure repeats was a more suitable indicator of FMR1 CGG repeat allelic stability, we speculated that the number of AGG interruptions contained within a repeat might determine the longevity of predisposed alleles within a population (30). To test this model, we have compared the AGG interspersion pattern and DXS548-FRAXAC1-FRAXAC2 haplotype of 200 normal chromosomes (10) and 30 premutation and high-end normal chromosomes in which intergenerational stability or instability had been ascertained (30,33). Association testing between the haplotype markers and interspersion patterns revealed a significant non-random distribution, which could be

**Table 1.** Fragile X haplotypes

| | Fragile X chromosomes | | | Normal chromosomes | | |
|---|---|---|---|---|---|---|
| *Haplotype* | *Number* | *freq* | *Expected Number* | | *freq* | *Association* |
| 644 | 7 | 0.159 | 2.4 | 11 | 0.055 | positive* |
| 645 | 6 | 0.136 | 1.5 | 7 | 0.035 | positive** |
| 213 | 6 | 0.136 | 1.3 | 6 | 0.03 | positive** |
| 834+ | 3 | 0.068 | 0.9 | 4 | 0.02 | |
| 734+ | 5 | 0.114 | 19.6 | 89 | 0.445 | negative** |
| 734 | 2 | 0.045 | 5.3 | 24 | 0.12 | |
| 113 | 1 | 0.023 | 1.1 | 5 | 0.025 | |
| 646+ | 1 | 0.023 | 0.4 | 2 | 0.01 | |
| 234 | 1 | 0.023 | 0.4 | 2 | 0.01 | |
| 634+ | 1 | 0.023 | 0.9 | 10 | 0.05 | |
| 212 | 1 | 0.023 | 0 | 0 | 0 | positive* |
| 613 | 1 | 0.023 | 0 | 0 | 0 | positive* |
| 713 | 3 | 0.068 | 0 | 0 | 0 | positive** |
| 724 | 3 | 0.068 | 0 | 0 | 0 | positive** |
| 724+ | 1 | 0.023 | 0 | 0 | 0 | positive* |
| 735 | 1 | 0.023 | 0 | 0 | 0 | positive* |
| 845 | 1 | 0.023 | 0 | 0 | 0 | positive* |
| Total | 44 | 1 | 44 | 154 | 0.77 | |

The frequency of DXS548-FRAXAC1-FRAXAC2 haplotypes among normal ($n$ = 200) and fragile X ($n$ = 44) males of Wessex, England is compared. This table is derived almost exclusively from a previously reported study (10) with the exception that an additional 12 normal chromosomes are included. Associations between haplotype and the fragile X mutation were determined by $\chi^2$ analysis of $2 \times 2$ contingency tables. Haplotypes found in linkage disequilibrium with the fragile X mutation are indciated by an asterisk. Significance of association is $P < 0.05$ (*) or $P < 0.01$ (**). Eighteen haplotypes representing 46 unaffected chromosomes were not observed among the fragile X population in this study.

used in the phylogenetic reconstruction of mutational events of FMR1 CGG repeat alleles. Parsimony analysis predicts that the loss of AGG interruptions has occurred independently and frequently among many of the haplotypes associated with the fragile X chromosome. In particular, two of the most prominent fragile X haplotypes (6–4–5 and 6–4–4; Table 1) show a correlation between the recurrent loss of AGG interruptions and asymmetrical array substructures. In contrast, one haplotype (2–1–3) is predicted to have been refractory to the loss of AGG interruptions. This haplotype's unique constraint to maintain two AGG interruptions has allowed FMR1 CGG repeat alleles to increase in length likely by replication slippage-like events distal to the last AGG interruption, resulting in the accumulation of high-end normal alleles in the population. Our data provide a molecular basis for the existence of two distinct mutational pathways for the origin of the fragile X mutation and suggest haplotype-specific influences in both the maintenance and loss of AGG interruptions.

## RESULTS

### Distribution of interspersion patterns among FRAXA haplotypes

Table 2 summarizes the distribution of 52 FMR1 CGG repeat AGG interspersion patterns among 28 DXS548-FRAXAC1-FRAXAC2 haplotypes for 200 unrelated male subjects from the Salisbury district of Wessex, England (Materials and Methods). The nomenclature and position of haplotype markers and the FMR1 CGG repeat interspersion patterns is described in Figure 1. In order to assess the randomness of distribution (Table 2), 5000 computer simulations were performed using the program ASSOC (15). The $\chi^2$ statistic was employed to determine the significance of association between rows (interspersion pattern) and columns (haplotype) in Table 2. Association testing revealed a highly significant non-random distribution between haplotypes and interspersion patterns ($\chi^2$ = 1381; $P < 0.002$; $n$ = 5000). In addition, each individual cell entry (association between row and column) was tested for significance. Significant associations ($P < 0.002$) between interspersion pattern and haplotype are indicated in Table 2 by an asterisk (*). Due to the complex polymorphic nature of FRAXAC2 and its reported potential for microsatellite hypermutability (34), various compressions of Table 2 were analysed to assess the utility of each marker independently against the interspersion configurations. All three markers demonstrated significant independent associations with FMR1 CGG repeat interspersion patterns: FRAXAC1, $\chi^2$ = 102.5, $P < 0.0001$; FRAXAC2, $\chi^2$ = 409.7, $P < 0.006$; DXS548, $\chi^2$ = 256.3, $P < 0.004$). In a similar fashion a highly significant ($P < 0.0001$) non-random association was determined between FRAXAC1 and FRAXAC2.

The three most common interspersion patterns (10+9+9, 9+9+9 and 10+9) account for 51.5% (103/200) of all X chromosomes and are distributed non-randomly among the various haplotypes (Table 2). The most abundant interspersion configuration (10+9+9) is concentrated primarily on FRAXAC1 haplotype 3 with the most significant associations ($P < 0.002$) occurring on DXS548-FRAXAC1-FRAXAC2 haplotype, 7–3–4+ (40/55 or 72.7% of all 10+9+9 chromosomes). Similarly, the 10+9 interspersion pattern shows a similar restriction to FRAXAC1 haplotype 3 with the most significant associations occurring on haplotypes 7–3–4+ and 6–3–4+ (Table 1). The 9+9+9 interspersion pattern demonstrates the greatest haplotype diversity being distributed among 9/28 (32.1%) of all observed haplotypes. Of these, the association of 9+9+9 with haplotypes 7–4–6+ and 1–1–3 are the most significant ($P < 0.002$).

Based on association testing between interspersion patterns and haplotype as well as visual inspection of the data in Table 2, it is apparent that haplotypes fall into one of two categories based on the position of the first AGG interruption. Both FRAXAC1 alleles 1 and 4 predominantly (51/53 or 96%) show interspersion patterns in which the first AGG interruption occurs at position 10 (9+n configurations). In contrast, haplotypes with FRAXAC1 marker 3 associate strongly with FMR1 CGG repeat interspersion patterns which possess the first AGG interruption at position 11 (10+n configurations) (Table 2). One notable exception is haplotype 7–3–4 which shows a substantial number of chromosomes with the 9+9+9 configuration and a significant association ($P < 0.002$) with the 13+9 interspersion pattern.

Although the position of the first AGG interruption generally divides haplotypes into one of two groups (relative to haplotype FRAXAC1), other significant associations are observed when one considers the other microsatellite haplotype markers and the position of the second AGG interruption. A total of 14 significant associations were identified using the DXS548-FRAXAC1-FRAXAC2 haplotype and the FMR1 CGG repeat interspersion pattern. (Only half as many significant associations were identified, using exclusively the FRAXAC1 marker). Using all three markers, 10 out of 14 of significant associations occur on FMR1 CGG repeat alleles with two AGG interruptions, and four out of 14 occur among alleles with a single interruption. For example, haplotypes 6–4–4 and 6–4–5 are virtually unique in possessing asymmetrical array patterns of the types, 9+10+9,

**Table 2.** FMR1 CGG repeat interspersion pattern distribution among DXS548-FRAXAC1-FRAXAC2 haplotypes

| | Haplotype: DXS548-FRAXAC1-FRAXAC2 | | | | | | | | | | | | | | | | | | | | | | | | | | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AGG Pattern | 113 | 213 | 313 | 714+ | 732 | 833 | 833+ | 733+ | 134 | 234 | 634 | 734 | 734+ | 634+ | 734+ | 834+ | 644 | 744 | 145 | 645 | 745 | 246 | 846 | 146+ | 246+ | 646+ | 746+ | 747+ | |
| 8–9 | | | | | | | | | | | | 1 | | | 1 | | | | | | | | | | | | | | 2 |
| 8–9+12 | | | 1* | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9–5+9+9 | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | 1 |
| 9+7+9+9 | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 9+9+9 | 3* | 1 | | 1 | | | | | | | | 5 | | 1 | 3– | | | | | | | | | | | 1 | 6* | 1 | 22 |
| 9+9+9+9+9 | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9+9+10 | | | | | | | | | | | | 1 | | | | | | | | | 1 | | | | | | | | 2 |
| 9+9+11 | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | 1 |
| 9+9+13 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9+9+20 | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9+9+23 | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9+9+29 | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9+9+24 | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9+10+9 | | | | | | | | | | | | 1 | | | | | | | | 4* | | | | | | | | | 5 |
| 9+11+9 | | | | | | | | | | | | | | | | | 4* | 1 | | | 1 | | | | | | | | 6 |
| 9+12+9 | | | | | | | | | | | | | | | | | 4* | | | 2 | | | | | 1* | | | | 6 |
| 9+15+9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| 9+16 | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | 1 |
| 9+20 | | | | | | | | | | | | | | | | | | | | | | 1 | 1 | | | | | | 2 |
| 9+21 | | | | | | | | | | | | | | | | | | | | | 2* | | | | | | | | 2 |
| 9+22 | | | | | | | | | | | | | | | | | 1 | | | 1 | | | | 1 | | | | | 3 |
| 9+26 | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | 1* | | 2 |
| 9+28 | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 10+6+9 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 10+9 | | | | | | 1 | | | | | | 1 | | 5* | 17* | 2 | | | | | | | | | | | | | 26 |
| 10+9+4+ | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | | 1 |
| 10+9+9 | | | | | 1 | | 1 | 2 | | | | 5 | 1 | 3 | 40* | 1 | | | 1 | | | | | | | | | | 55 |
| 10+9+10 | | | | | | | | | | | | | | | 10* | | | | | | | | | | | | | | 10 |
| 10+9+21 | | | | | | | | | | | | 1 | | | 1 | | | | | | | | | | | | | | 2 |
| 10+10+9 | | | | | | | | | | | | 1 | 1 | | 5 | | | | | | | | | | | | | | 7 |
| 10+10+10 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 10+11 | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | | 1 |
| 10+12+9 | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | 1 |
| 10+15 | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | 1 |
| 10+18 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 10+19 | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | | 1 |
| 10+20 | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 10+21 | | | | | | | | | | | | | 1 | | 1 | | | | | | | | | | | | | | 2 |
| 11+10 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 11+47 | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | 1 |
| 12+9 | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 12+10 | | | | | | | | 1 | | | | | 1 | | 1 | | | | | | | | | | | | | | 3 |
| 13+9 | | | | | | | | | 1 | 1 | 1 | 4* | | | | | | | | | | | | | | | | | 7 |
| 13+10 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 14 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 15 | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | | 1 |
| 21 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 23 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 26 | | | | | | | | | | 1 | | 1 | | 1 | | | | | | | | | | | | | | | 3 |
| 26–9 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 34 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| 54 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| total | 5 | 6 | 1 | 1 | 1 | 1 | 1 | 7 | 2 | 2 | 2 | 24 | 4 | 10 | 89 | 4 | 11 | 1 | 1 | 7 | 3 | 2 | 1 | 1 | 1 | 2 | 9 | 1 | 200 |

Fifty-two AGG interspersion configurations for the FMR1 CGG repeat are compared against 28 chromosomal haplotypes. Haplotypes are organized in columns and are grouped using the FRAXAC1 marker (since it shows the strongest association with interspersion pattern, $P$ <0.00001). Interspersion patterns are arranged in rows based on increasing length of the 5′ tract of CGG repeats proximal to the first interruption. Significant positive associations ($P$ <0.02) between interspersion pattern and haplotype were determined from ASSOC computer simulations and are indicated by bold and asterisk in the table. A single significant deficiency was observed and is indicated by a minus sign (haplotype 7–3–4+; interspersion pattern 9+9+9).

9+11+9 and 9+12+9 (Table 2). Haplotype 7–4–5, a relatively rare haplotype in this population survey, demonstrates a significant ($P$ <0.002) association with an interspersion pattern devoid of a second AGG interruption (9+21). Similarly, another rare haplotype, 2–4–6+, is associated with an equally rare interspersion pattern (9+15+9).

Due to the prevalence of 7–3–4+ haplotype in this survey (89/200 = 44.5% of all chromosomes), several significant interspersion configurations were identified, namely; 10+9, 10+9+9 and 10+9+10. In addition, a single significant deficiency was found with interspersion pattern 9+9+9, confirming once again a partition of haplotypes based on the position of the first AGG interruption. It is interesting that the majority of FMR1 CGG repeats completely devoid of interruptions occur also on haplotype 7–3–4+ and other closely allied haplotypes (7–3–4 and 6–3–4+).

It should be noted that many of the DXS548-FRAXAC1-FRAXAC2 haplotypes (17/28) do not demonstrate non-random associations with particular FMR1 CGG repeat interspersion patterns due to the under-representation of these haplotypes in this population survey. One haplotype, 2–1–3, is interesting in this regard. Although no significant association was found, five out of the six interspersion patterns occurring on this haplotype are unique, due to the presence of a long distal tract of CGG repeats (Table 2). Furthermore, the position of the second AGG is remarkably conserved among all FMR1 CGG repeat alleles on this chromosomal background, generating the general configuration pattern 9+9+n.

## Intra-haplotype comparisons of 1-AGG and 2/3-AGG interspersion configurations

FMR1 CGG repeat alleles may be categorized into different classes based on the total number of AGG interruptions which they possess (29–32). In this population survey of 200 random chromosomes, alleles were distributed as follows: nine (4.5%)

**Table 3.** Intra-haplotype comparisons between 2-AGG and 1-AGG allele configurations (>25 total repeat length)

| *Haplotype* | *2 and/or 3 AGG Interruption Configurations* | *1 AGG Interruption Configurations* |
|---|---|---|
| 733+ | 10+9+9* (30) | 10+19 (30) |
| 734 | 9+9+9 (29); 10+9+9*(30); 9+11+9 (31) | 10+20 (31) |
| 234+ | 10+10+10 (32); 10+9+10 (31) | 10+21 (32) |
| 734+ | 10+9+9* (30); 10+9+10 (31); 10+6+9 (27); 10+9+21 (43) | 10+21 (32); 10+18 (29); 26+9 (36) |
| 644 | 9+11+9* (31); 9+12+9* (32) | 9+26 (36); 9+22 (32) |
| 645 | 9+10+9* (30); 9+12+9 (32) | 9+22 (32) |
| 745 | 9+9+10 (30) | 9+21 (31)* |
| 246 | 9+11+9 (31) | 9+20 (30) |
| 646+ | 9+9+9 (29) * | 9+16 (26) |
| 746+ | 9+9+9 (29); 9+5+9+9 (35) | 9+26 (36) |
| 834+ | 10+9+9 (30) | 11+47 (58) |

FMR1 CGG repeat interspersion configurations of alleles with a single AGG interruption (1-AGG class) and alleles with two or more interruptions are compared. Only alleles whose total repeat length was greater than 25 triplet repeat units are considered. Total repeat length is indicated in brackets. In most cases, only the most significant (indicated by an asterisk) and/or the most frequent 2/3-AGG alleles are considered.

with no interruption; 59 (29.5%) with a single interruption; 129 (64.5%) with two AGG interruptions; and three (1.5%) with more than two AGG interspersions. Table 3 compares the most frequent and most significant 2-AGG configurations with 1-AGG interspersion patterns (>25 total repeats) within each haplotype. In most cases, the total overall repeat length of the 1-AGG class does not differ dramatically from the total repeat length of alleles with two AGG interruptions. Interestingly, the position of the first AGG interruption is also generally conserved between 2-AGG and 1-AGG classes within the same haplotype (Table 2). With one exception (Table 3; haplotype 7–3–4+; allele 26+9), there appears to be a considerable bias among normal alleles to lose the most distal AGG interruption.

## Haplotype analysis among premutations and 'proto-premutation' alleles

In an attempt to reconstruct the phylogenetic progression of normal alleles to the fully expanded state, we examined an additional 30 unrelated premutation and high-end repeat length 'normal' alleles in which intergenerational stability/instability had been ascertained (30). The FMR1 CGG repeat substructure, the haplotype and the stability for each premutation and evolving premutation are summarized in Table 4. The majority of haplotypes at high-risk for the fragile X syndrome (Table 1) possess FMR1 CGG repeat alleles which have a single AGG interruption or are completely devoid of AGGs (Table 4). Among those alleles in which a single interruption predominates (haplotypes 6–4–5, 6–4–4, 8–3–4+ and 6–4–6+), the position of the most 5′ AGG has been conserved between normal and premutation alleles (Table 2). Similarly, unstable high-end normal alleles, which likely represent intermediate states between normal stable alleles and premutation alleles, show a similar maintenance of the position of the first AGG interruption (Table 4).

Although the reduction or complete absence of AGG interruptions is common to most of the fragile X haplotypes, a specific subset of DXS548-FRAXAC1-FRAXAC2 haplotypes (2–1–3, 7–1–3 and 1–1–3) have interspersion patterns in which the position of the second AGG interruption has been maintained (Table 4). Intergenerational transmission studies of these FMR1 CGG repeat alleles indicate that these 2-AGG configurations can be remarkably stable (Table 4). One of the largest stable premutation-size alleles identified to date (66 total repeats;

**Table 4.** AGG interspersion patterns among premutation and 'protopremutation' haplotypes

| *Sample* | *DXS548* | *AC1* | *AC2* | *CGG* | *Repeat Substructure* | *AGG #* | *Stability* | *Reference* |
|---|---|---|---|---|---|---|---|---|
| 8125 | 4 | 0 | 3 | 50 | 9+9+30 | 2 | 2s | 30, 31 |
| 16,415 | 1 | 1 | 3 | 92 | 9+9+72 | 2 | fm | 30 |
| IX-1 | 2 | 1 | 3 | 49 | 9+9+29 | 2 | 2s | 30, 33 |
| 10882 | 2 | 1 | 3 | 51 | 9+9+31 | 2 | 5s | 30, 31 |
| VII-1 | 2 | 1 | 3 | 53 | 9+9+33 | 2 | 1s | 30, 33 |
| 1408-08 | 2 | 1 | 3 | 55 | 55 | 0 | 5u | 18, 30 |
| 15,135 | 2 | 1 | 3 | 66 | 9+9+46 | 2 | 3s | unpublished |
| 68-650 | 2 | 1 | 3 | 66 | 9+56 | 1 | fm | 30 |
| XIII-3 | 6 | 1 | 3 | 50 | 9+40 | 1 | 1u | 30, 33 |
| 80-813 | 7 | 1 | 3 | 82 | 9+9+62 | 2 | 1u | 30 |
| 91/1450 | 8 | 3 | 4+ | 97 | 97 | 0 | fm | 10 |
| C-1 | 7 | 3 | 4+ | 34 | 34 | 0 | 1s 1u | 30 |
| D-3 | 7 | 3 | 4+ | 54 | 12+41 | 1 | 1u | 30 |
| 92002 | 7 | 3 | 4+ | 61 | 12+48 | 1 | 1u | 30 |
| 21,578 | 7 | 3 | 4+ | 74 | 10+63 | 1 | fm | 30 |
| 91/213 | 7 | 3 | 4+ | 74 | 74 | 0 | fm | 10 |
| 90/2097 | 7 | 3 | 4 | 79 | 79 | 0 | 1m | 10 |
| B-1 | 8 | 3 | 4+ | 51 | 11+39 | 1 | 2u | 30 |
| 91/391 | 8 | 3 | 4+ | 83 | 11+71 | 1 | 1m | 10 |
| A-17 | 6 | 4 | 5 | 44 | 9+34 | 1 | 2s 1u | 30 |
| 10508 | 6 | 4 | 4 | 47 | 9+37 | 1 | 2s | 30, 31 |
| XVII-1 | 6 | 4 | 5 | 47 | 9+37 | 1 | 1u | 30, 33 |
| IV-1 | 6 | 4 | 5 | 48 | 9+38 | 1 | 1s | 30, 33 |
| 7544 | 6 | 4 | 5 | 59 | 59 | 0 | 1m | 30, 31 |
| 75-830 | 6 | 4 | 5 | 66 | 9+46 | 1 | 1m | 30 |
| 75-787 | 6 | 4 | 5 | 72 | 9+62 | 1 | 1m | 30 |
| 75-828 | 6 | 4 | 5 | 76 | 9+66 | 1 | 1m | 30 |
| 89/205 | 6 | 4 | 6+ | 82 | 9+72 | 1 | fm | 10 |
| A-4 | 6 | 4 | 5 | 83 | 9+73 | 1 | fm | 30 |
| 92/1600 | 6 | 4 | 4 | 84 | 9+74 | 1 | fm | 10 |
| 15-522 | 6 | 4 | 6+ | 104 | 9+94 | 1 | fm | 30 |
| 66-691 | 7 | 4 | 6+ | 110 | 9+100 | 1 | fm | 30 |

The DXS548-FRAXAC1-FRAXAC2 haplotype is compared with the FMR1 CGG repeat interspersion pattern of 15 premutation (documented progression to full mutation) and 15 protopremutation (evolving premutation) alleles, based on large total repeat length or observation of unstable intergenerational transmissions (30). Samples are derived from 30 unrelated pedigrees. The study of origin for each allele is indicated in the reference column (10,18,30,31,33). Samples are arranged according to haplotype and are grouped relative to the FRAXAC1 marker. The number of documented stable (s) and unstable (u) transmissions in each pedigree is summarized in the column labelled stability. Alleles which have been observed to progress to full mutation are denoted in this column as fm.

interspersion pattern 9+9+46; sample 15,135; Table 4), belongs to this conspicuous fragile X haplotype. Although this allele has a pure repeat tract which is much greater than the postulated instability threshold (34–37 CGG repeats) (30), only stable

transmissions have been observed to date within this pedigree over two generations (data not shown). Despite this unusual high degree of stability, alleles carrying 2-AGG interspersions have been observed to progress to full-mutation (samples 80–813 and 16,415; Table 4).

### Parsimony analysis of fragile X haplotypes

Parsimony analysis was performed using both heuristic, branch-and-bound and exhaustive tree searches on aligned normal and premutation CGG repeat sequences (Table 2, Materials and Methods). Majority rule (>50%) and strict consensus trees were generated for haplotypes which had at least four different sequences and multiple, equally parsimonious topologies. Based on character-state reconstructions for trees of the shortest length, parsimony predicts that the loss of AGG interruptions has occurred frequently and independently within eight out of the nine fragile X haplotypes which could be examined (data not shown). The phylogenetic relationship of interspersion patterns for four haplotypes is depicted in Figure 2.

The topology of the tree generated for haplotype 2–1–3 reveals a unique phylogeny of interspersion pattern among fragile X haplotypes (Fig. 2a). Character-state reconstructions, without defining an ancestral state, show that the position of the second AGG interruption has been a highly conserved and ancient characteristic of this haplotype. The predicted ancestral state is 9+9+9, with exclusion of ambiguous characters. Unlike other fragile X haplotypes, the 2–1–3 haplotype possesses an unusual sub-lineage which is enriched for long uninterrupted repeats and 2-AGG interruptions (compare Figure 2a with 2b,c and d), indicating that this particular haplotype may be refractory to the loss of AGG interruptions. Within this same sub-lineage, however, the loss of one or both AGG interruptions has been shown to occur in association with instability and the fragile X syndrome (Fig. 2a).

Since strong non-random associations between haplotype and interspersion pattern likely reveal historical information on founder interspersion configurations, association testing was used to identify the ancestral state (Fig. 2b,c) of two other haplotypes (6–4–5 and 6–4–4) which are found in positive association with the fragile X mutation (Table 1). Interestingly, the ancestral states defined by association testing were both asymmetrical with respect to the middle tract of CGG repeats

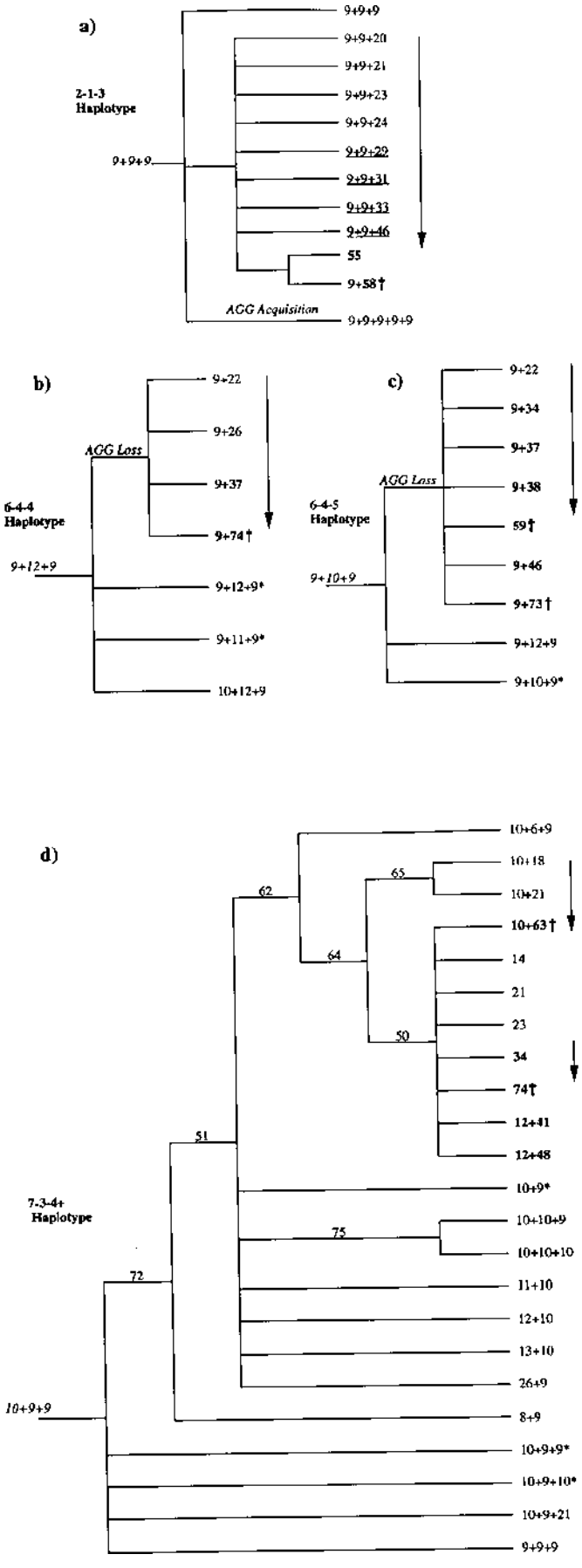


**Figure 2.** Interspersion phylogeny of fragile X haplotypes. Parsimony analysis (PAUP v.3.1.1) was used to reconstruct the phylogenetic relationships among various interspersion patterns within each haplotype (see Materials and methods). Phylogenetic trees were generated by performing heuristic, branch-and-bound and exhaustive tree searches from aligned FMR1 CGG repeat sequences and are depicted for four haplotypes: (**a**) 2–1–3, (**b**) 6–4–4, (**c**) 6–4–5 and (**d**) 7–3–4+. Ancestral states are shown in italics (see Materials and Methods). The most significant interspersion associations are once again indicated by an asterisk. Alleles in which unstable intergenerational transmissions have been observed are shown in bold. Large alleles in which only stable transmissions have been observed (see Table 4) are underlined. Unstable premutation alleles which have progressed to full mutation are further denoted by the symbol (†). The likely progression of predisposed allele to disease state is indicated by a vertical arrow. Although the single most parsimonious trees for 2–1–3, 6–4–4 and 6–4–5 (a, b, c) are presented (homoplasy index = 1.0), a total of 3000 equally parsimonious trees were generated for haplotype 7–3–4+. As a result, the majority-rule consensus (>50%) tree is depicted for haplotype 7–3–4+. The numbers at each node in the consensus tree represent the percentage of trees which exhibited similar topology for that particular branch position.

(9+10+9 and 9+12+9). A single phylogenetic tree was constructed for both the 6–4–5 and 6–4–4 haplotype (tree length = 6 and 7, respectively) after an exhaustive search. In contrast to haplotype 2–1–3, parsimony predicts the frequent or 'ancient' loss of AGG interruptions on these two chromosomal backgrounds. This has resulted in the formation of a large sub-tree of FMR1 CGG repeat lineages which are prone to instability and hyperexpansion (Fig. 2b,c). The loss of AGG interruptions in these two haplotypes is predicted to have occurred among normal-sized alleles (30–35 total repeats) without a dramatic change in the overall length of the repeat.

Due to the prevalence of the 7–3–4+ haplotype within this population, we also investigated the phylogenetic origins of unstable and premutation alleles within this genetic background. A majority rule (50%) consensus tree is presented for 3000 (MAXTREES limit) equally parsimonious trees (length = 7) generated by a heuristic tree search ($14.4 \times 10^6$ rearrangements), using the 10+9+9 configuration as the likely ancestral state (Fig. 2d). Character-state reconstructions predict that the loss of the most 3′ interruption has occurred by two different routes within this haplotype. Similar to 6–4–4 and 6–4–5, one subtree within 7–3–4+ indicates that the loss has occurred in a manner which preserves the overall length of the repeat (10+18 and 10+21). A closely related branch (10+63) is associated with the fragile X syndrome (Fig. 2d). Another pathway for the loss of the distal AGG interruption, which accounts for several different interspersion patterns within this haplotype, is the loss of an entire +9 array generating the 10+9 and related lineages (Fig. 2d). Character state reconstructions, furthermore, predict that the most proximal AGG interruption has also been subject to loss, with configurations completely devoid of interruptions occurring among normal and premutation chromosomes (Fig. 2d).

## DISCUSSION

### Founder effects?

Several groups have reported a significant founder effect phenomenon for the fragile X syndrome (3–15) with the mutation being enriched from two- to five-fold on specific DXS548, FRAXAC1 and FRAXAC2 haplotypes. Dependent upon the genetic homogeneity of the population under study (9,11,12), 50% of all fragile X chromosomes on average, share a few common genetic backgrounds (9,11,12). The remaining fragile X chromosomes either exist at equilibrium with common haplotypes in the population (10) or occur on rare haplotypic variants (10,28). These observations have prompted speculation that recurrent mutational events may occasionally occur on genetic backgrounds not at risk for the development of the disease, resulting in the formation of *de novo* fragile X lineages (10). Furthermore, based on multistep progression models for the progression of predisposed alleles to the disease state (25–27), it has been predicted that among haplotypes in which clear linkage disequilibrium was demonstrated, one might expect an enrichment of alleles with large repeats among normal chromosomes of that haplotype. Surprisingly, only one of the three major fragile X haplotypes (Table 1; 2–1–3) demonstrates such an enrichment (Table 2). This suggests haplotype-specific influences other than total repeat length are playing a role in predisposing chromosomes to the development of the fragile X syndrome (10,13,29).

### Loss of AGG interruptions: the molecular basis for diversity of fragile X haplotypes

In order to test the hypothesis that the loss of AGG interruptions might resolve these two founder effect discrepancies, we compared the AGG interspersion patterns of 200 normal and 30 premutation and high-end repeat alleles within various haplotypes (Table 2). Unlike previous investigations (32), we have found a significant ($P < 0.002$) non-random distribution between haplotype and FMR1 CGG repeat interspersion patterns. Furthermore, the use of all three markers was found to be more informative than any single haplotype marker. Phylogenetic reconstruction of CGG repeat mutational events within haplotype lineages, using association testing and parsimony analysis, predicts that the loss of the most 3′ AGG interruption has occurred frequently and independently on many normal and at-risk fragile X haplotypes (Fig. 2b,c,d, as examples). Parsimony analysis indicates that alleles in the normal population which have been subjected to the loss of AGG interruption belong to the same clades as alleles which are predisposed to instability and hyperexpansion (Fig. 2). Our data suggest that the loss of AGG interruptions likely accounts for the occurrence of the fragile X mutation among haplotypes found in equilibrium within the normal population (Table 1 and Fig. 2c). The frequent loss of AGG interruptions, thus, may account for the considerable haplotype diversity of the fragile X mutation.

Parsimony analysis and intra-haplotype comparisons of 1-AGG and 2-AGG class of alleles (Table 3 and Fig. 2b,c,d) confirms that there has been a preferential bias to lose the most 3′ AGG interruption (Eichler, in press). This appears to have occurred by a mechanism which tends to preserve the overall length of the repeat (Table 3). Previously it was suggested that the loss of AGG interruptions may be mediated by one of several different mutational processes such as unequal-sister chromatid exchange, AGG deletion or AGG to CGG transversion (29–32). Since the loss of AGG interruptions by non-reciprocal recombination events would be expected to alter the total number of CGG repeats (35), it is unlikely that this is the predominant mechanism. Our data would best support the deletion or conversion of an AGG to a CGG, resulting, perhaps, from deficiencies in mis-match repair processes (36,37). It should be noted, however, that not all 1-AGG and 2-AGG intra-haplotype comparisons show a conservation of total repeat length in the human population (6–4–6+ and 8–3–4+; Table 3). Indeed, parsimony analysis reveals that the loss of AGG interruptions may occur by the occasional deletion of an entire $AGG(CGG)_9$ array (Fig. 2c). Such an event could likely be mediated by non-reciprocal recombination processes.

### Founder haplotypes 6–4–5 and 6–4–4: the recurrent loss of AGG interruptions

Although the loss of AGG interruptions occurs frequently from a phylogenetic perspective, such losses could not occur with equal frequency among all haplotypes without completely obscuring founder effect phenomena. Differential propensities for the loss of AGG interruptions, then, must exist in order to account for observed haplotype linkage disequilibrium with the mutation (3–14). Two haplotypes in this study (6–4–4 and 6–4–5) appear particularly prone to the loss of AGG interruptions. Parsimony analysis indicates that a substantial proportion of unstable and premutation alleles have likely progressed from a

single AGG interruption ancestral state (Fig. 2b,c), with the position of the 5′ AGG interruption being remarkably conserved (Table 2). Not surprisingly, DXS548, FRAXAC1 and FRAXAC2 haplotypes equivalent to 6–4–5 and 6–4–4 have been shown to be enriched for fragile X chromosomes in virtually all human populations studied to date. These haplotypes generally account for 15–40% of all fragile X chromosomes among Japanese, American, Australian, Finnish, English and French populations (3,4,6,7,10,12,13). The strong association with the loss of AGG interruptions within these two haplotypes may reflect either an 'ancient' or recurrent event; two possibilities which parsimony analysis can not distinguish. The conspicuous absence of high-end normal FMR1 CGG repeat alleles among these haplotypes (10,13,24,29), however, strongly argues against the former hypothesis, suggesting that alleles progress relatively rapidly through instability and hyperexpansion thresholds without the accumulation of large intermediate (S) alleles (25). Another argument favoring the recurrent loss of the 3′ AGG interruption is the finding of global linkage disequilibrium among these haplotypes with the fragile X mutation. Conservative estimates based on multistep progression models of total repeat length (25) calculate that the longevity of predisposed (S) alleles in the population is approximately 90 generations or 2000 years (15). The Japanese and caucasian populations likely diverged from a common *Homo sapiens* stock, 150 000 to 200 000 years ago (38–41). The most prosaic explanation, then, for linkage disequilibrium for the same haplotypes within these two populations would be a recurrent mutational event (such as the loss of an AGG interruption), rather than a common ancestral mutation.

Due to the absence of normal alleles with large total repeat lengths (>40 repeats) among the 6–4–5 and 6–4–4 haplotypes, it was originally speculated that a recurrent mutational event might allow alleles on these haplotypes to 'leap-frog' toward instability and hyperexpansion thresholds associated with disease (10). Our analysis indicates that the loss of AGG interruptions could clearly provide the molecular basis for this 'leap-frog' mutational event. Occurring by a mechanism which tends to maintain the overall length of the repeat, the loss of AGG interruptions would generate alleles which 'leap-frog' *in cognito* by 10 and 12 repeat units toward their instability (35 pure repeats) and hyperexpansion thresholds (70 pure repeats) in a single generation (30). The longevity of such alleles due to increased pure repeat length would be dramatically reduced, obviating the accumulation of high-end normal repeat length intermediate alleles. The propensity for these chromosomes to frequently incur such mutations would confirm previous suggestions of haplotype-specific influences for these fragile X genetic backgrounds (29).

Although it is difficult at this point to identify the nature of these haplotype-specific influences, our analysis suggests that symmetry of repeat configurations may play a role in predisposing alleles to the loss of AGG interruptions. In both 6–4–5 and 6–4–4 haplotypes, significant associations were observed for configurations in which symmetry has been disrupted for the middle tract of the FMR1 CGG repeat (9+10+9 and 9+12+9; Table 2). Common symmetrical arrays such as 9+9+9 are conspicuously absent among these haplotypes. Parsimony analysis and association testing indicate that 9+12+9 and 9+10+9 have likely been the original precursors to unstable and premutation alleles within haplotypes 6–4–5 and 6–4–4; respectively (Table 3; Fig. 2b,c). Conversely among haplotypes,1–1–3 and 7–4–6+, in which there is a strong association for symmetrical arrays (9+9+9), fragile X chromosomes have rarely been observed or are found in equilibrium within the normal population (Table 1). Interestingly, one allele with a large tract of pure repeats (9+26; Table 2) has been observed in haplotype 7–4–6+. Although far too few alleles exist in this haplotype to reconstruct an unambiguous phylogeny by parsimony, due to the similar conservation of repeat size, the 9+26 may have originated from the asymmetrical allele, 9+9+5+9, by the consecutive loss of 2-AGG interruptions (9+9+5+9 to 9+15+9 to 9+26 see Table 2; 7–4–6+ and 6–4–6+ haplotypes). Recent repeat length (13,28) and interspersion analysis surveys of Asian populations show a preponderance of 9+9+5+9 (35 total repeat) interspersion configurations on DXS548-FRAXAC1 haplotype (7–4) and FRAXAC1-FRAXAC2 haplotype (4–6+) (Composite 7–4–6+) (13,28). Among the Japanese, at least, 15% of all fragile X cases occur on the 4–6+ haplotype, suggesting that once again a correlation exists between asymmetrical array patterns, loss of AGG interruptions and progression to disease. An alternative explanation for the correlation between asymmetric interspersion configurations and the loss of AGG interruptions may however be that asymmetry is simply a consequence rather than a cause of increased instability on these particular haplotypes. If this were true, other *cis*-acting factors closely linked to the CGG repeat would have to be invoked to explain the propensity for haplotype 6–4–5 and 6–4–4 to lose AGG interruptions.

### Founder haplotype 2–1–3: maintenance of two AGG interruptions

In many respects, the fragile X founder haplotype 2–1–3 (Table 1) represents the antithesis of 6–4–5 and 6–4–4. Computer simulation testing for non-random distributions of alleles within this haplotype reveals no significant association between haplotype and interspersion patterns, despite the fact that the majority of AGG configurations are unique. This lack of association is likely a reflection of the extreme FMR1 CGG repeat sequence heterogeneity of alleles within this haplotype (Table 2). As has been suggested earlier for the FRAXAC1 haplotype 1 (29), the positions of both AGG interruptions appear to have been highly conserved with most changes in repeat length occurring distal to the second 3′ AGG interruption. Character state reconstructions of the 2–1–3 phylogenetic tree confirms that the 9+9+n interspersion pattern represents the ancestral configuration of this haplotype (Fig. 2a). Parsimony analysis further suggests that 2–1–3 is the only haplotype in which FMR1 CGG repeat alleles have progressed toward instability and disease with the 2-AGG configuration. Although the loss of one or both AGG interruptions has occurred in association with instability in this haplotype, it should be noted that unique 2-AGG containing premutations have been observed among haplotypes closely allied to 2–1–3 (Table 4; haplotype 7–1–3 and 6–1–3). *In toto*, these observations suggest that the 2–1–3 haplotype has been particularly refractory to the loss of the most 3′ AGG interruption.

The constraint of the 2–1–3 haplotype to maintain two AGG interruptions defines a second mutational pathway for the origin of the fragile X syndrome (Fig. 3). Unlike other founder haplotypes, the 2–1–3 haplotype initially has not jumped in increments of 10 CGG repeat increments toward instability and hyperexpansion thresholds by the loss of an AGG interruption. Most changes in repeat length have occurred distal to the last AGG interruption, likely in small increments of one or two repeat

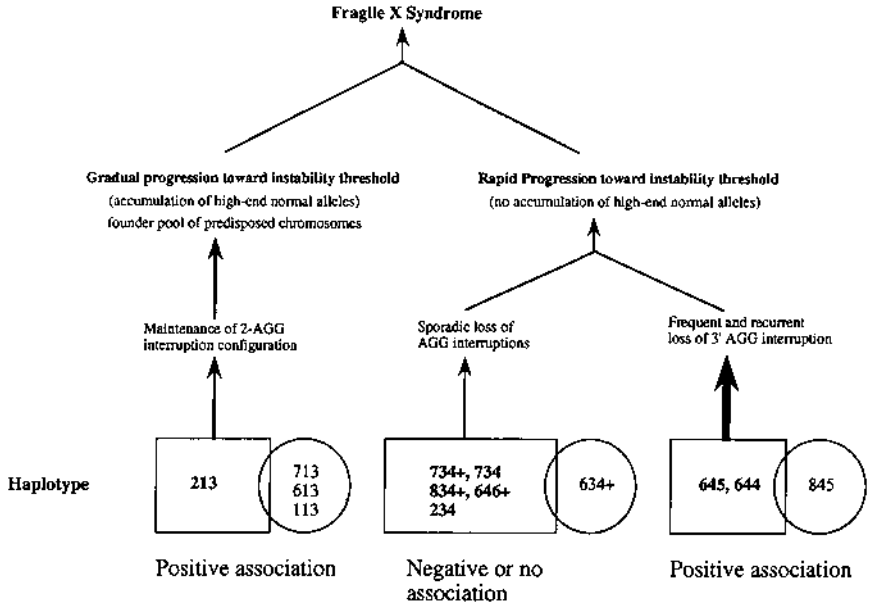


**Figure 3.** Origins of the fragile X mutation. Based on intrahaplotype comparisons of interspersion pattern, association testing and parsimony analysis, a model is presented depicting the possible origin of the fragile X mutation across a variety of chromosomal haplotypes. Haplotypes shown within squares are distinguished as being positively, negatively or showing no association with the fragile X mutation (10) (see Table 1). Encircled are fragile X haplotypes that have not yet been observed among unaffected chromosomes. Due to the considerable phylogenetic probability of recombination between the most proximal DXS548 marker (Fig. 1), these haplotypes are depicted overlapping closely-allied FRAXAC1-FRAXAC2 haplotypes. The vertical arrows indicate the likely mutational origin and progression to the diseased state as predicted by parsimony analysis. Due to the absence of normal chromosomes with FRAXAC1 haplotype marker 2, the origin of fragile X haplotypes (7–2–4 and 7–2–4+) could not be deduced.

units, by a mechanism similar to replication slippage (42). As a result, a continuum of repeat lengths has been generated (Table 2). Due to the persistence of 2-AGG interruptions, longer total repeat lengths are attained before uninterrupted pure CGG repeat instability and hyperexpansion thresholds can be reached (30). This pathway accommodates the postulated multistep progression of alleles toward disease (25,26) and explains why the 2–1–3 founder chromosome is the only haplotype which shows an enrichment for high-end normal FMR1 CGG repeat alleles (3–5).

Previous surveys of interspersion pattern and DXS548 and FRAXAC1 haplotypes among normal control populations found an unusually high proportion (60%) of long uninterrupted pure CGG repeat tracts (>21 CGG repeats) associated with haplotypes equivalent to 2–1–3 (29,31). Virtually every possible 9+9+n configuration ranging from 9+9+15 to 9+9+35 has been observed among normal 2–1–3 chromosomes (Table 2) (29,31). It is surprising that such a continuum has not been observed for haplotypes 6–4–4 and 6–4–5. Based on intergenerational transmission studies, we originally proposed an instability threshold of 34–37 pure CGG repeats (30). Interestingly, all alleles demonstrating unstable transmission in 'normal' human pedigrees were either devoid of AGG interruptions or constrained to a single AGG interruption, belonging primarily to haplotypes 6–4–4, 6–4–5 and 7–3–4+ (Table 4). No unstable transmissions among 2–1–3 haplotypes with two AGG interruptions have been observed. These findings suggest the number of AGG interruptions within the repeat may influence the instability threshold, such that 9+21 and 9+9+21 progress at different tempos toward their respective instability thresholds. Although the vast majority of alleles exhibit instability in human pedigrees when the length of the longest tract of repeats exceeds 37 pure repeats, 2-AGG allele configurations among haplotype 2–1–3 may be exceptional in this regard.

We have recently documented the stable transmission of one of the largest premutation-sized alleles, 9+9+46 (66 total triplet repeats; Sample 15,135, Table 4). Although the longest tract of pure repeats (46 CGG triplets) was well beyond the instability threshold, only stable transmissions were observed over two meioses within this pedigree. It will be interesting to determine whether other large and stable 'premutation-sized' alleles also belong to the 2–1–3 haplotype, possessing 2-AGG allele configurations (43). It has recently been demonstrated that contiguous CGG repeats may be capable of forming several unusual DNA structures including tetraplex, triad-DNA, unimolecular foldback and hairpin conformations (44–47). Such conformations may be important in mediating replication slippage. Furthermore, it has been postulated that AGG interruptions may confer a stabilizing influence by disrupting such conformations which may be prerequisite for CGG triplet repeat instability (45). One possible explanation, then, for the observation of increased stability on the 2–1–3 haplotype normal chromosomes, may be that 2-AGG FMR1 CGG repeat conformations have a greater stabilizing influence than haplotypes with a single AGG interruption, resulting in different thresholds for instability. Alternatively, the remarkable stability of large alleles of the 2–1–3 haplotype, may be the result of other *cis*-acting influences.

## Origins of the fragile X mutation

Phylogenetic reconstruction of fragile X lineages based on parsimony analysis, association testing and intrahaplotype com-

parisons of interspersion patterns among normal and premutation alleles, reveals that the origins of the fragile X syndrome are complex. Our analysis predicts that fragile X mutations have arisen by both founder effects and a recurrent mutational event involving the loss of the most 3′ AGG interruption. Among haplotypes which demonstrate linkage disequilibrium with the fragile X mutation (Table 1), two distinct mutational pathways have been delineated (Figs 2a,b,c, 3). The DXS548-FRAXAC1-FRAXAC2 haplotype 2–1–3 (Fig. 1) appears refractory to the loss of AGG interruptions, progressing relatively slowly toward instability and hyperexpansion thresholds associated with the disease. As a result, the 2–1–3 haplotype is enriched for high-end repeat length alleles, generating a large founder pool of chromosomes which are predisposed to the development of disease (Fig. 3). In contrast, haplotypes 6–4–5 and 6–4–4, based on parsimony analysis, appear particularly prone to the loss of the most 3′ AGG interruption from asymmetrical 2-AGG interspersion configurations (Fig. 2b,c). This occurs by a mechanism which tends to preserve the total overall length of the repeat (Fig. 2 and Table 3). Due to the observation of unstable transmission in 'normal' human pedigrees on this genetic background and the dearth of large intermediate alleles in random population surveys of these haplotypes, we propose that chromosomes which have lost an AGG interruption at the FRAXA locus progress rapidly toward the disease state (Fig. 3). The frequent and recurrent loss of AGG interruptions, is thus likely responsible for the observation of linkage disequilibrium with the 6–4–5 and 6–4–4 founder haplotypes (Fig. 3). Parsimony analysis, in addition, predicts that the loss of AGG interruptions has occurred independently on most DXS548-FRAXAC1-FRAXAC2 haplotypes. The sporadic loss of AGG interruptions and the genesis of *de novo* fragile X lineages (Figs 2c, 3) can account for the observation of disease among haplotypes found in linkage equilibrium with the mutation and partially explain the considerable haplotype diversity observed with the fragile X syndrome. Although our model of founder effects and recurrent loss of AGG interruptions can adequately explain the origin of 70% of all fragile X chromosomes, further haplotype analysis, interspersion determination, and repeat-length transmission studies will be necessary to confirm our observations and to explain the frequent occurrence of fragile X syndrome on haplotypes rare in the normal population.

## MATERIALS AND METHODS

### DNA samples

DNA samples from the normal control population, consisting of 200 unrelated non-fragile X male subjects, were collected from blood donors living in Wessex, Southern England as previously described (10). The remaining 30 premutation and large repeat-length normal alleles were derived from unrelated fragile X pedigrees or from families in which the intergenerational stability of the FMR1 CGG repeat had been documented (10,18,30,31,33). Six of the 15 premutation alleles were ascertained from the same population as the control samples (Wessex, England). The remaining premutation and unstable 'normal' FMR1 CGG repeat alleles were collected from North American populations of caucasian ethnic origin (Table 3).

### AGG interspersion analysis

PCR amplification of the FMR1 CGG repeat was performed using a previously developed protocol which replaces exo(–) *Pfu* polymerase (Stratagene) with *Taq* DNA polymerase (48). In order to reconstruct the AGG substructure of the FMR1 CGG repeat, PCR products were digested with *Mnl* I restriction enzyme (New England Biolabs), electrophoresed, transferred to nylon membrane (GeneScreen plus), and probed with a $\gamma^{32}$P end-labelled $(CGG)_{10}$ oligonucleotide as previously described (30). Based on the size of the FMR1 PCR product, the pattern of *Mnl* I digestion and the sizes of *Mnl* I fragments which hybridized to the CGG repeat oligonucleotide, the AGG interspersion configuration for each FMR1 CGG repeat allele was deduced. FMR1 CGG interspersion nomenclature is summarized in Figure 1. In our convention, a '+' sign designates the position of an AGG interruption and the number refers to the length of uninterrupted repeats.

### Haplotype determination

Three polymorphic microsatellite markers, DXS548, FRAXAC1 and FRAXAC2, which span 150 kb of the FMR1 CGG repeat locus were used to reconstruct the chromosomal haplotype of each normal, premutation and fragile X allele (Fig. 1 and Table 1). The genotypes for DXS548, FRAXAC1 and FRAXAC2 were determined using previously described PCR conditions and primers (4,6,10). For each polymorphic marker, alleles were designated using numbers in descending order of repeat length (Fig. 1). Due to the complex nature of the FRAXAC2 marker, which consists of two dinucleotide polymorphic repeat tracts and a polymorphic poly T tract (34) care was taken to distinguish between FRAXAC2 alleles which differ by a single bp. These are indicated in this study by a '+' sign. For example, a FRAXAC2 genotype '4+' is intermediate in size between genotype 3 and 4 differing by a single bp (Fig. 1). Composite haplotypes using all three markers are considered in this analysis and are configured based on the order of each marker (Fig. 1).

### Association testing

A total of 28 composite haplotypes (DXS548-FRAXAC1-FRAXAC2) were compared against 52 different AGG interspersion patterns (Table 2). The randomness of distribution between rows (AGG configuration) and columns (haplotypes) was tested using the association program ASSOC (AC, unpublished). Briefly, 5000 tables (replicates) were generated by computer simulation using row and column totals and assuming a random distribution between haplotype and interspersion pattern for Table 2. A $\chi^2$ value was then calculated comparing these randomly-generated tables to the observed table (Table 2). Significant departures of randomness ($P$ <0.002) for each cell entry (intersection between haplotype and interspersion pattern) were determined by analysing the individual ranges of distribution for each cell entry. The most significant association between haplotype and interspersion pattern ($P$ <0.02) are summarized in Table 1. In order to test the usefulness of the composite haplotype against each haplotype marker independently, three compressions of Table 2 were also generated (one for each marker) and non-random distributions were tested as described above (see Results).

### Parsimony analysis

PAUP (phylogenetic analysis using parsimony) version 3.1.1 (Illinois Natural History Survey) was employed to derive phylogenetic relationships among various AGG interspersion FMR1 CGG repeat patterns within each haplotype. Only those haplotypes which exhibited four or more different interspersion patterns were analysed due to software requirements for at least four distinct taxa. Deduced FMR1 CGG repeat sequences were encoded (Let C = CGG and A = AGG) to eliminate the possibility of gap formation within a trinucleotide repeat unit. Alignment of encoded data was performed using the ClustalW software package (default setting: gap penalty = 15.0 and gap extension penalty = 1.0). Reduction of gap penalty assignments below default setting parameters resulted in the generation of sequence alignments which were inconsistent with characteristic intra-haplotype AGG interspersion patterns. Phylogenetic trees were generated from aligned sequences using heuristic and exhaustive tree searches (Fig. 2). In cases where multiple, equally parsimonious trees were generated, a consensus tree was constructed using majority-rule (>50%) methods (Fig. 2d). Ancestral states were defined using the ANCSTATES command under the assumptions block of PAUP (49) and were chosen using the most significantly associated interspersion configuration ($P$ <0.002) within each haplotype. When no significant association could be found (haplotype 2–1–3), the likely ancestral state was deduced using the character-state reconstruction option (Fig. 2a). When ambiguous character states were obtained (i.e. either the presence or absence of a CGG trinucleotide were equally parsiminious), ancestral states were reconstructed excluding these character assignments. In such situations, the shortest, most parsimonious ancestral state was considered.

## ACKNOWLEDGEMENTS

We would like to thank Drs P. Ward, B. Popowich, J. Holden, A. Reiss, S. Richards and S. Thibodeau for providing some of the DNA samples used in this study. This work was supported in part by NIH grants (HD29256 and GM52982) to DLN.

## REFERENCES

1. Richards, R.I., Holman, K., Kozman, H., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Mulley, J. and Sutherland, G.R. (1991) Fragile X syndrome: genetic localisation by linkage mapping of two microsatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. *J. Med. Genet.* **28**, 818–23.
2. Riggins, G.J., Sherman, S.L., Oostra, B.A., Sutcliffe, J.S., Feitell, D., Nelson, D.L., van Oost, B.A., Smits, A.P.T., Ramos, F.J., Pfendner, E., Kuhl, D.P.A., Caskey, C.T. and Warren, S.T. (1992) Characterization of a highly polymorphic dinucleotide repeat 150 kb proximal to the fragile X site. *Am. J. Med. Genet.* **43**, 237–43.
3. Richards, R.I., Holman, K., Friend, K., Kremer, E., Hillen, D., Staples, A., Brown, W.T., Goonewardena, P., Tarleton, J., Schwartz, C. and Sutherland, G.R. (1992) Evidence of founder chromosomes in fragile X syndrome. *Nature Genet.* **1**, 257–60.
4. Oudet, C., Mornet, E., Serre, J.L., Thomas, F., Lentes-Zengerling, S., Kretz, C., Deluchat, C., Tejada, I., Boue, J., Boue, A. and Mandel, J.L. (1993) Linkage diequilibrium 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.
5. Buyle, S., Reyniers, E., Vits, L., De Boulle, K., Handig, I., Wuyts, F.L.E., Deelen, W., Halley, D.J.J., Oostra, B.A. and Willems, P.J. (1993) Founder effect in a Belgian-Dutch fragile X population. *Hum. Genet.* **92**, 269–72.
6. Jacobs, P.A., Bullman, H., Macpherson, J., Youings, S., Rooney, V., Watson, A. and Dennis, N.R. (1993) Population studies of the fragile (X): a molecular approach. *J. Med. Genet.* **30**, 454–59.
7. Hirst, M.C., Knight, S.J.L., Christodoulou, Z., Grewal, P.K., Fryns, J.P. and Davies, K.E. (1993) Origins of the fragile X syndrome mutation. *J. Med. Genet.* **30**, 647–50.
8. Arinami, T., Asano, M., Kobayashi, K., Yanagi, H. and 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**, 431–6.
9. Haataja, R., Vaisanen, M.L., Li, M., Ryynanen, M. and Leisti, J. (1994) The fragile X syndrome in Finland: demonstration of a founder effect by analysis of microsatellite haplotypes. *Hum. Genet.* **94**, 479–83.
10. Macpherson, J.N., Bullman, H., Youings, S.A. and Jacobs, P.A. (1994) Insert size and flanking haplotype in fragile X and normal populations: possible multiple orgins for the fragile X mutation. *Hum. Mol. Genet.* **3**, 399–405.
11. Malmgren, H., Gustavson, K.H., Oudet, C., Holmgren, G., Petterson, U. and Dahl, N. (1994) Strong founder effect for the fragile X syndrome in Sweden. *Eur. J. Hum. Genet.* **2**, 103–9.
12. Oudet, C., von Koskull, H., Nordstrom, A.M., Peippo, M. and Mandel, J.L. (1993) Striking founder effect for the fragile X syndrome in Finland. *Eur. J. Hum. Genet.* **1**, 181–189.
13. Richards, R.I., Kondo, I., Homan, K., Yamauchi, M., Seki, N., Kunikazu, K., Sutherland, G.R. and Hori, T. (1994) Haplotype analysis at the FRAXA locus in the Japanese population. *Am. J. Med. Genet.* **51**, 412–16.
14. Zhong, N., Ye, L., Dobkin, C. and Brown, W.T. (1994) Fragile X founder chromosome effects: Linkage disequilibrium or microsatellite heterogeneity? *Am. J. Med. Genet.* **5**, 405–11.
15. Chakravarti, A. (1992) Fragile X founder effect? *Nature Genet.* **1**, 237–8.
16. Smits, A.P.T., Dreesen, J.C.F.M., Smeets, D.F.C.M., de Die-Smulders, C., Spaans-van der Bijl, T., Govaerts, L.C.P., Warren, S.T., Oostra, B.A. and van Oost, B.A. (1992) The fragile X syndrome: no evidence for any recent mutations. *J. Med. Genet.* **30**, 94–6.
17. Drugge, U., Holmgren, G., Blomquist, H.K., Dahl, N., Gustavson, K.H. and Malmgren, H. (1992) A study of individuals possibly affected with the fragile X syndrome in a large Swedish family in the 18th and 20th centuries. *Am. J. Med. Genet.* **43**, 353–4.
18. Fu, Y.H., Kuhl, D.P.A., Pizzuti, A., Pieretti, M., Sutcliffe, J.S., Richards, S., Verkerk, A.J.M.H., Holden, J.J.A., Fenwick, R.G., Jr., Warren, S.T., Oostra, B.A., Nelson, D.L. and Caskey, C.T. (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67**, 1047–58.
19. Oberle, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boue, J., Bertheas, M.F. and Mandel, J.L. (1991) Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252**, 1097–102.
20. Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., Holman, K., Mulley, J.C., Warren, S.T., Schlessinger, D. *et al.* (1991) Fragile X genotype characterized by an unstable region of DNA. *Science* **252**, 1179–81.
21. Heitz, D., Rousseau, F., Devys, D., Saccone, S., Abderrahim, H., le Paslier, D., Cohen, D., Vincent, A., Toniolo, D., Della Valle, G., Johnson, S., Schlessinger, D., Oberlé, I. and Mandel, J.L. (1991) Isolation of sequences that span the fragile X and identification of a fragile X-related CpG island. *Science* **251**, 1236–9.
22. Kremer, E.J., Yu, S., Pritchard, M., Nagaraja, R., Heitz, D., Lynch, M., Baker, E., Hyland, V.J., Little, R.D., Wada, M. *et al.* (1991) Isolation of a human DNA sequence which spans the fragile X. *Am. J. Hum. Genet.* **49**, 656–61.
23. Haldane, J.B.S. (1935) *J. Genet.* **31**, 317–26.
24. Snow, K., Doud, L.K., Hagerman, R., Pergolizzi, R.G., Erster, S.H. and Thibodeau, S.N. (1993) Analysis of a CGG sequence at the FMR-1 locus in the fragile X families and in the general population. *Am. J. Hum. Genet.* **53**, 1217–28.
25. Morton, N.E. and Macpherson, J.N. (1992) Population genetics of the fragile x syndrome: a multiallelic model for the FMR1 locus. *Proc. Natl Acad. Sci. USA* **89**, 4215–17.
26. Morris, A., Morton, N.E., Collins, A., Macpherson, J., Nelson, D.L. and Sherman, S. (1995) An n-allele model for progressive amplifcation in the FMR1 locus. *Proc. Natl Acad. Sci. USA* **92**, 4833–7.
27. Kolehmainen, K. (1994) Population genetics of fragile X: a multiple allele model with variable risk of CGG repeat expansion. *Am. J. Med. Genet.* **51**, 428–35.
28. Zhong, N., Liu, X., Gou, S., Houck, G.E., Li, S., Dobkin, C. and Brown, W.T. (1994) Distribution of FMR1 and associated microsatellite alleles in a normal Chinese population. *Am. J. Med. Genet.* **51**, 417–22.

29. Kunst, C.B. and Warren, S.T. (1994) Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* **77**, 853–61.
30. Eichler, E.E., Holden, J.J.A., Popovich, B.W., Reiss, A.L., Snow, K., Thibodeau, S.N., Richards, C.S., Ward, P.A. and Nelson, D.L. (1994) Length of uninterrupted CGG repeats determines stability in the FMR1 gene. *Nature Genet.* **8**, 88–94.
31. Snow, K., Tester, D.J., Kruckenberg, K.E., Schaid, D.J. and Thibodeau, S.N. (1994) Sequence analysis of the fragile X trinucleotide repeat: implications for the origin of the fragile X mutation. *Hum. Mol. Genet.* **3**, 1543–51.
32. Hirst, M.C., Grewal, P.K. and Davies, K.E. (1994) Precursor arrays for triplet repeat expansion at the fragile X locus. *Hum. Mol. Genet.* **3**, 1553–60.
33. Reiss, A.L., Kazazian, H.H., Jr., Krebs, C.M., McAughan, A., Boehm, C.D., Abrams, M.T. and Nelson, D.L. (1994) Frequency and stability of the fragile X premutation. *Hum. Mol. Genet.* **3**, 393–8.
34. Zhong, N., Dobkin, C. and Brown, W.T. (1993) A complex mutable polymorphism located within the fragile X gene. *Nature Genet.* **5**, 248–53.
35. Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–17.
36. Fishel, R., Lescoe, M.K., Rao, M.R.S., Copeland, N.G., Jenkins, N.A., Garber, J., Kane, M. and Kolodner, R. (1993) The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–38.
37. Liu, B., Nicolaides, N.C., Markowitz, S., Willson, J.K.V., Parsons, R.E., Jen, J., Papadopolous, N., Peltomaki, P., de la Chapelle, A., Hamilton, S.R., Kinzler, K.W. and Vogelstein, B. (1995) Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. *Nature Genet.* **9**, 48–55.
38. Dorit, R.L., Akashi, H. and Gilbert, W. (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science* **268**, 1183–5.
39. Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. and Takahata, N. (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl Acad. Sci. USA* **92**, 532–6.
40. Paabo, S. (1995) The Y chromosome and the origin of all of us (men). *Science* **268**, 1141–2.
41. Bowcock, A.M., Ruiz-Linares, A., Minch, E., Kidd, J.R. and Cavalli-Sforza, L.L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–7.
42. Levinson, G. and Gutman, G.A. (1986) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–21.
43. Rousseau, F., Morgan, K., Rouillard, P. and Khandjian, E.W. (1995) Prevalence of carriers of premutation-size alleles of the FMR1 gene and implications for the populaton genetics of the fragile X syndrome. *Am. J. Hum. Genet.* in press.
44. Fry, M. and Loeb, L.A. (1994) The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. *Proc. Natl Acad. Sci. USA* **91**, 4950–4.
45. Gacy, A., Goellner, G., Juranic, N., Macura, S. and McMurray, C. (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* **81**, 533–40.
46. Kuryavyi, V.V. and Jovin, T.M. (1995) Triad-DNA: a model for trinucleotide repeats. *Nature Genet.* **9**, 339–41.
47. Smith, S.S., Laayoun, A., Lingeman, R.G., Baker, D.J. and Riley, J. (1994) Hypermethylation of telomere-like foldbacks at codon 12 of the human c-Ha-ras gene and the trinucleotide repeat of the FMR1 gene of fragile X. *J. Mol. Biol.* **243**, 143–51.
48. Chong, S.S., Eichler, E.E., Hughes, M.R. and Nelson, D.L. (1994) Robust amplification of the fragile X syndrome CGG repeat using *Pfu* polymerase: ethidium bromide detection of normal and premutation alleles. *Am. J. Med. Genet.* **51**, 522–6.
49. Swofford, D.L. (1993) *PAUP: Phylogenetic Analysis using Parsimony.* Illinois Natural History Survey, Champaign, Illinois.