Volume 17 Number 16 1989 Nucleic Acids Research

Hypervariability of simple sequences as a general source for polymorphic DNA markers

Diethard Tautz*

Max-Planck Institut für Entwicklungsbiologie, Spemannstrasse 35/II, 7400 Tübingen, FRG

Received June 27, 1989; Accepted July 21, 1989

ABSTRACT

Short simple sequence stretches occur as highly repetitive elements in all eukaryotic genomes and partially also in prokaryotes and eubacteria (1−6). They are thought to arise by slippage like events working on randomly occuring internally repetitive sequence stretches (3, 7, 8). This predicts that they should be generally hypervariable in length. I have used the polymerase chain reaction (PCR) process to show that several randomly chosen simple sequence loci with different nucleotide composition and from different species show extensive length polymorphisms. These simple sequence length polymorphisms (SSLP) may be usefully exploited for identity testing, population studies, linkage analysis and genome mapping.

INTRODUCTION

The reason for the ubiquitous occurence of simple sequences in organismal genomes has so far not been satisfactorily resolved. Simple sequences consist of stretches of monotonously repeated short nucleotide motifs. They are usually less than 100 bp long and are normally embedded within unique DNA stretches. They occur as interspersed repetitive elements in all eukaryotic and to a lesser extent also in prokaryotic and eubacterial genomes. Almost all permutations of mono-, di- and tri-nucleotide motifs, but also several longer ones can be found as building blocks of simple sequences (1−10). In eukaryotes one can expect to encounter at least one simple sequence stretch every 10 kb of DNA sequence. Cryptically simple sequences, which consist of less regular and internally scrambled motifs are equally, or even more frequent (7). They have been correlated with regions of high divergence on an evolutionary time scale (7,11−13). On the basis of these facts we have previously argued that simple sequences and cryptically simple sequences do not generally provide a defined function for the genome, but merely reflect internal genomic mechanisms, which have the tendency to dynamically produce and delete these sequences (3,7,14). This assumption predicts that simple sequence stretches should be generally hypervariable in length. I provide here direct evidence that this is indeed the case. Furthermore, I show that the degree of hypervariability is such that the length polymorphism within these regions may be exploited for relationship studies of individuals within and between populations and that they may serve as a general source for polymorphic DNA markers for genome mapping and linkage studies.

MATERIALS and METHODS

DNA-samples

Drosophila DNA was extracted from a pool of animals of several wild type strains, which were originally established as isofemale lines and kept in the laboratory for about 50−200

a)

```
                                         T                                                                                                   A
TAAGCTTGGGAATCATCTCGCCGACGGGCAGCGATATGGGCATCATGCTCGCCCCGCCCCAATCCTCGAAGAATAGTGCAATAATGCAAACGATATCACCCCAGCAACAGCAGCAGCAGCAGCAGCAG
  SerLeuGlyIleIleSerProThrGlySerAspMetGlyIleMetLeuAlaProProGlnSerSerLysAsnSerAlaIleMetGlnThrIleSerProGlnGlnGlnGlnGlnGlnGlnGlnGln

                                                                                   HaeIII               C
CAACAGCAGCAACATCAGCAGCAGCAA(CAG)3-6CAACAGCAGCAGCAGCAA(CAG)3-4CAACTCGGAGGCCTGGAGTTCGGTTCAGAGGGCTTGGACCTGAATGGATTTTGTGGATCTCCGGGTAA
GlnGlnGlnGlnHisGlnGlnGlnGln(Gln)3-6GlnGlnGlnGlnGlnGln(Gln)3-4GlnLeuGlyGlyLeuGluPheGlySerGluGlyLeuAspLeuAsnGlyPheCysGlySerProA
                                                                                                        Ala
                         --.---A--------A----------------ACTAA---
GTGGTCACTCATGATGGACTCTATGG.ACTCGCTAACTAGCTAACTAATCATTCTA.....CCATCCCAACTTGCAGACTCATTTCACTCGGGTCAAATGAATCCGCCCTCGATACAAAGTTCAAT
                         --G.........--------------------.....---             spSerPheHisSerGlyGlnMetAsnProProSerIleGlnSerSer
```

b)

```
AGAAAGATTAATAGACTGGAGTAAAGGAAATGGACCTCTGTCTCTCTCGCTGTCTCTTTTTTGAGGACTT(GT)16-17TGTGGTCAGTGGGGCTGGAATAAAAGTAGAATAGACCTGCAC
CTGCTGTGGCATCCATTCACAGAG
```

```
CTCCCCCACACAAAGAAGTTCTGTTCTCTTCCCTCTACCTTGACCTGAATGCACTGTGA(TG)15-25ACTCGTTCCCAGGTATGGAA
```

c)

```
TGAAATTCTTCATCAGTTGCCATGGTGGGGATTTTTCCCTCTCTTCTCTCTCTCTCTCTCTCTCTCTCTCCCCCCTTAACTGGTGAAACTGACAGTCTGCCTACATTAAC
```

generations. Human DNA was extracted from buccal cells of the saliva. These were obtained by rinsing the mouth for 10 seconds with 10−15ml of an isotonic salt solution and spinning down the cells for 5 min at 2000g (15). Whale DNA was isolated from skin samples of stranded whales and provided by W. Amos (Cambridge). The DNA extraction was done in all cases using a standard SDS-ProteinaseK-Phenol procedure (16).

Primers

Primers were synthesized on a Cyclone DNA synthesizer and purified on a 15% denaturing acrylamidgel, ensuring that only the desired synthesis products were obtained. The primers were 5'-endlabeled using γ^{32}P-ATP and Polynucleotidekinase according to a standard protocol (16).

Polymerase chain reaction

PCR reactions were performed in 10μl using the thermostable Taq-Polymerase under the recommended conditions of the supplier (Perkin Elmer Cetus)(17). 25 cycles were performed with 1 min at 95°C, 2 min at 45°C and 1.5 min at 72°C for each cycle. The final extension was for 5 min at 72°C. About 0.2μg of DNA was used for each PCR reaction, which included in addition to the unlabeled primers 1 pmol of one labeled primer. The reaction products were mixed with an equal volume of Formamide-dye solution and were resolved on a 6% denaturing acrylamide gel ('sequencing gel'(18)).

Isolation of simple sequence loci

10μg of DNA from the long finned pilot whale *Globicephala malaena* was digested to completion with AluI and HaeIII. The digestion products were resolved on an 1.2% agarose gel and a size fraction of 250−350nt was cut out and isolated. These fragments were blunt end cloned into a M13 vector (see ref. 16 for the details of cloning methods). Transformed clones were then hybridized in a plaque hybridization assay with a nicktranslated GA/CT probe (3). Hybridization was at 65°C in 5×SSPE (for the formulation of SSPE see ref. 16) with a final wash in 2× SSPE at 65°C. Positive clones were plaque purified and then sequenced. The clone for the locus in Fig. 1c had a total length of 246nt, only part of which is shown.

RESULTS

The polymerase chain reaction (PCR) technique allows to directly amplify defined DNA segments from genomic DNA (17). Simple sequence stretches are normally flanked by unique DNA stretches. It is therefore possible to choose primers, which specifically amplify a single simple sequence locus. The length of the amplified piece will normally be within the range of 50 to 300 nt, which allows to resolve the amplification products on a standard

Figure 1: *Simple sequence loci which were employed in this study.* The primers which were used to amplify each of the regions are underlined. (a) Region of the *Notch* gene of *D.melanogaster* (positions 1−379 in ref. 10). The sequence is displayed with its translation of the coding region and includes also the various sequenced variants. Three single nucleotide polymorphisms were found (indicated above the sequence), which were detected in independent clones and are thus unlikely to be PCR artefacts. The two length polymorphic CAG-stretch regions are indicated in the appropriate positions. They could be put at defined places, because the whole CAG-stretch region is interspersed in a defined pattern with CAA triplets. The two new length variants of the intron region are indicated below and above the main sequence. Dots represent deleted regions, dashes represent identities, the sequence comparison is limited to the relevant region. The HaeIII site which was used for separating the CAG-stretch region from the intron region is indicated in italics. (b) Sequence from the intergenic region of the human δ and β-globins (positions 43−211 in ref. 9) (upper sequence) and from the intron IV of the human cardiac muscle actin gene (20)(lower sequence). The different length classes found for the sequenced variants of these stretches are indicated. Other polymorphisms were not found. (c) Sequence of a randomly isolated whale locus, containing a CT-dinucleotide simple sequence stretch.

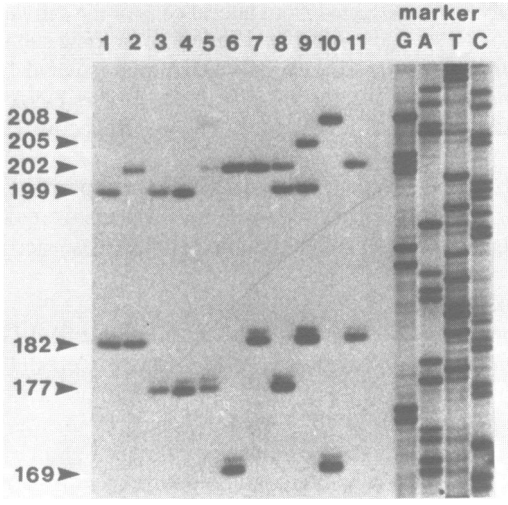

Figure 2: *Hypervariability of simple and cryptically simple sequences within the Notch gene of Drosophila.* The sizes of the various reaction products are given in nucleotides (left). The primer pair indicated in Fig. 1a was endlabeled with ^{32}P and used in a standard PCR reaction to amplify the respective DNA fragment from the DNA of 11 isofemale lines of *Drosophila melanogaster* wild-type strains. The DNA was then digested with HaeIII and resolved on a 6% sequencing gel. A sequencing reaction was taken as a marker(right). The 11 strains came from Canada (1), USA (2,3), Peru (4), Great Britain (5), Spain (6), Cyprus (7), Turkey (8), USSR (9), Japan (10) and Australia (11).

sequencing gel and to reproducibly record even single nucleotide length differences. I have chosen several different simple sequence loci in three different species with different simple sequence motifs.

CAG repeats in Drosophila

A first test for the variability of simple sequence loci was done for a region in the *Notch* gene of *Drosophila*. This region contains within the coding region a long stretch of CAG repeats, which are interspersed with CAA triplets (10,19). Next to this lies a small intron, which shows some indications of cryptic simplicity (Fig. 1a). The CAG stretch and the intron region were amplified using primers which flank both of them and which should amplify a piece of 379 nt according to the published sequence. Both primers were endlabeled with ^{32}P and the amplification products were digested with HaeIII before loading them on a sequencing gel. This allowed to analyse each region separately, since they should be split into fragments with sizes of 202 and 107 nt. Fig. 2 shows the amplification products for 11 independent isofemale lines of *D. melanogaster* from throughout the world. Both, the '202' nt fragment containing the CAG stretch and the '177' nt fragment containing the intron region show length variability. The CAG stretch shows variants, which differ by multiples of three, indicative of deletions or insertions of CAG triplets. The intron region shows variants which differ by five and seven nucleotides, which seem not to be related to any particular internal repeat structure. The novel variants were cloned and sequenced. The CAG stretch showed indeed variations in the number of CAG triplets which lengthen or shorten the stretch of Glutamin residues within the coding region (Fig. 1a). The intron region contains multiple mutation events, including the observed insertion of five and the deletion of eight nucleotides. These mutations occured in a region which shows several

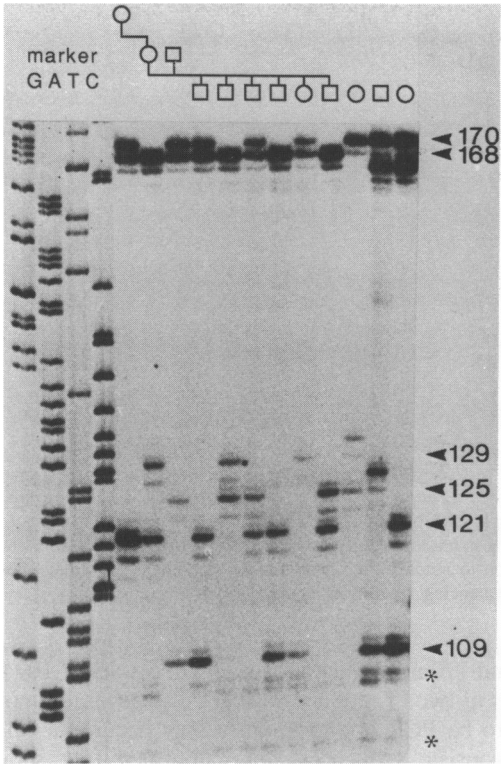

Figure 3: *Inheritance study for two simple sequence loci from humans.* The sizes of the various reaction products which are of relevance for the family study are given in nucleotides (right). Circles represent females, squares males. The two primer pairs indicated in Fig. 1b were endlabeled with ^{32}P and used in a standard PCR reaction. The reaction products were directly resolved on a 6% sequencing gel using a sequencing reaction as a marker (left). The asterisks a the bottom indicate artefactual PCR bands, which arise due to chance homologies with the primers (see text).

directly repeated CTAA motifs, though the deletions/insertions do not occur in a defined frame. The underlying mechanism for these mutation events seems not obvious, but these types of mutations were predicted in view of the universal occurence of cryptically simple regions (7).

GT repeats in humans

As a second test system, I have chosen two simple sequence loci from the intergenic region between the δ and β globin genes and from the intron region of the human cardiac muscle actin gene. Both are unique loci in the genome and are located on different chromosomes. The published sequences contain stretches of 17 and 25 GT dinucleotide pairs respectively (9,20) (Fig. 1b). Two primer pairs which flank these stretches were synthesized. The predicted length of the reaction product for the globin pair of primers is 168 nt, the one for the actin pair is 129 nt. The choice of these two length classes allowed to simultaneously amplify these two regions and to resolve the products on single gel lanes. Individuals from three generations of a family as well as three unrelated individuals were tested. A clear length polymorphism was observed for both loci (Fig. 3). The globin locus showed length

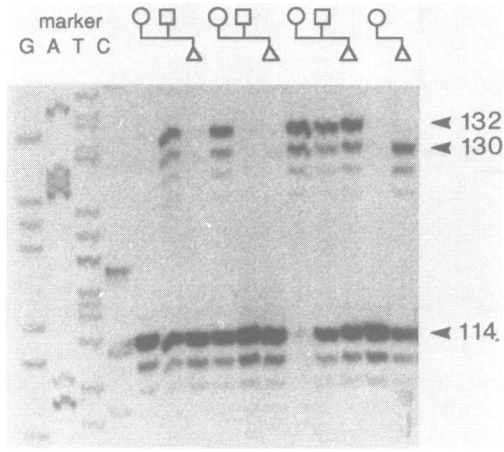

Figure 4: *Hypervariability of a randomly chosen simple sequence locus from whales.* The sizes of the various reaction products are given in nucleotides (right). Circles represent females, squares males. The triangles indicate that the sex of the offspring is not known. The primers indicated in Fig. 1c were endlabeled with ^{32}P and used in a standard PCR reaction to amplify the corresponding genomic fragments. These were resolved on a 6% sequencing gel using a sequencing reaction as a marker (left).

classes of 166, 168 and 170 nt, the actin locus length classes of 109 and 121−131nt, each differing by multiples of two. There are however also several minor bands visible which must be considered to be PCR artefacts, since they occured similarly, if single cloned variants of the appropriate sequences were amplified (not shown). They represent presumeably slippage events, which occur during the PCR replication process. Nonetheless, if the strongest bands are taken as the basis for the analysis, it is easily possible to interpret the patterns. Each individual contained at most two length variants at each locus as it is appropriate for diploid organisms. The family study shows the expected inheritance pattern with the descendants having inherited always one variant from the mother and one from the father. The patterns are interpreted as follows: the grandmother is heterozygous for the 168nt/170nt length classes at the globin locus and homozygous for the 121nt class at the actin locus. Her daughter has inherited the 168nt and the 121nt length classes. In addition, she shows a 129nt length class at the actin locus, but is homozygous for the globin locus. Her husband is heterozygous for the 170nt/168nt length classes at the globin locus and shows two new variants, namely a 125nt and a 109 nt length class at the actin locus. The six children show always a combination of each of the mentioned length classes. The three other individuals can not have been offspring of these parents, since they differ in one or more length classes, or since they show novel combinations of length classes. Several variants of the length polymorphic sequences were cloned and sequenced and showed variation only with respect to the number of GT dinucleotides (Fig. 1b), strongly suggesting that slippage like events have led to the observed length variability.

TC repeats in whales

To prove that a randomly cloned simple sequence locus can also be shown to be hypervariable, I have chosen pilot whales as a third test system. A DNA sample from one individual was digested with AluI and HaeIII and a size fraction of around 300 bp

was directly cloned into M13 phages. The phages were screened with an *in vitro* synthesized GA/TC dinucleotide probe. About one percent of the clones showed a positive signal and four of them were plaque purified and sequenced. All four contained the expected GA/TC stretch, though in different combinations with other simple sequence variants, or within a stretch of high cryptic simplicity. One of these sequences was chosen (Fig. 1c) and primers flanking it were used to amplify the appropriate piece from different individuals of a single population. Length classes of 114, 130 and 132 were observed which differ by multiples of two. Again there are also minor reaction products, which are apparently due to the same type of PCR artefacts as seen for the GT stretches of the human loci. Nonetheless, the pattern can be clearly analysed. Three mating pairs and their offspring show the expected inheritance pattern. A fourth individual, whose descent was in question, can be shown to be not related to any of the three tested males, since it shows a novel length class, which is not present in the other individuals (Fig. 4). The relationship of the individuals was independently confirmed using hypervariable minsatellite probes (W. Amos, pers. communication).

DISCUSSION

The above results prove that simple sequence stretches can be generally considered to be hypervariable in length. This variability does not seem to be limited to particular types of simple sequences and must therefore be a reflection of a general mechanism. This mechanism is most likely slippage, which would occur during replication or DNA repair (8,21,22). This would make simple sequence hypervariability different from the variability at mini satellite loci, which was suggested to be mainly due to recombination mechanisms (23,24). Little is known about the frequency at which slippage mutations occur. In prokaryotes they are certainly more frequent than spontaneous pointmutations (reviewed in ref. 8) and indirect evidence suggests that this is also true for eukaryotes (7,11−13, 22). From the above results one can infer that slippage mutations are sufficiently frequent to maintain a high degree of polymorphism within populations, but not frequent enough to occur in successive generations. Considering that simple sequence loci may comprise up to 5% of a typical eukaryotic genome (3), one has to conclude that a significant portion of the genome is subject to very dynamical turnover processes. The *Drosophila* results show furthermore that this is not limited to noncoding sequences, but that even the coding sequence of genes which have a direct influence on developmental pathways are subject to these processes. This effect has been previously predicted from evolutionary comparisons and may play a significant role in the evolution of new developmental traits (7,11−13,25).

Potential applications

Hypervariability within simple sequence loci opens a new way for determining individual identity. Since several independent loci may be amplified simultaneously and since all of these can be potentially analysed in a single gel lane, it seems possible that a set of about 10 oligonucleotide pairs is sufficient to uniquely identify individuals. Moreover, since sequencing gels can provide an absolute length standard, it seems possible that the identity information for an individual can be digitized and reproduced independently. Further advantages are the possibility to use minute amounts of DNA and to largely automate the process, using the existing technology which has been developed for the automation of DNA sequencing (26). The only potential problem is presented by the occurence of PCR artefacts. One type of artefact, namely slippage during the PCR amplification process is shown in Fig. 3 and 4. It is as yet unclear how this can be avoided. Preliminary tests

with a change of reaction conditions did not lead to significant improvements. However, if these side reactions should be a problem for certain types of analysis, it seems alternatively possible to use simple sequence loci which are less prone to this artefact (Fig. 2). Another problem is the frequent occurence of nonrelated amplification products (compare Fig. 2), which is due to chance homologies of the primers at other sites. This can however be easily solved by using a third primer (as the only labeled one) for each locus which lies within the target sequence and which is added only at a late stage during the amplification process (not shown), similarly as it has been suggested for the direct sequencing of PCR amplified bands (27). The only potential problem is presented by the occurence of PCR artefacts. One type of artefact, namely slippage during the PCR amplification process is evident in Fig. 3 and 4. It is as yet unclear how this can be avoided. Preliminary tests with a change of reaction conditions did not lead to significant improvements. However if these side reactions should be a problem for certain types of analysis, it seems alternatively possible to use simple sequence loci which are less prone to this artefact (Fig. 2). Another problem is the frequent occurence of nonrelated amplification products (compare Fig. 2), which is due to chance homologies of the primers at other sites. This can however be easily solved by either changing the reaction conditions, or by using a third primer (as the only labeled one) for each locus which lies within the target sequence and which is added only at a late stage during the amplification process (not shown).

Simple sequence length polymorphisms (SSLP) are effectively 'single locus' probes, which greatly facilitates the analysis of the obtained patterns in population studies (28). For the same reason, SSLPs may also be exploited for genome mapping and linkage studies. The various simple sequences are highly repetitive components of eukaryotic genomes and appropriate loci can be expected to occur every few thousand base pairs. They provide therefore an almost unlimited and easily accessible supply for polymorphic loci, which would make the sometimes tedious search for restriction fragment length polymorphisms (RFLP) superfluous. The full value of SSLPs will as yet have to be established. However in view of the range and the number of potential loci, it seems likely that they can provide a solution to any problem related to the exploitation of DNA polymorphisms.

While this paper was in preparation, two reports were published, which arrive at similar conclusions for GT/CA simple sequences in the human system (29,30).

ACKNOWLEDGEMENTS

I thank M. Kidwell, W. Amos and the members of my family for providing materials and DNA samples, C. Pfeifle for technical assistance, D. Weigel for comments on the manuscript and H. Jäckle for his constant support. This work is subject of patent applications. Enquiries should be addressed to Garching Instruments (Munich).

*Present address: Universität München, Institut für Genetik und Mikrobiologie, Maria-Ward-Strasse 1a, 8000 München 19, FRG

REFERENCES

1. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982) Proc. Natl. Acad. Sci. USA, **79**, 6465–6469.
2. Gebhard, W. and Zachau, H.G. (1983) J. Mol. Biol., **170**, 567–573.
3. Tautz, D. and Renz, M. (1984) Nucl. Acids Res., **12**, 4127–4138.
4. Gross, D.S. and Garrard, W.T. (1986) Mol. Cell. Biol., **6**, 3010–3013.
5. Greaves, D.R. and Patient, R.K. (1985) EMBO J., **4**, 2617–2626.
6. Schèfer, R., Ali, S. and Epplen, J.T. (1986) Chromosoma, **93**, 502–510.
7. Tautz, D., Trick, M. and Dover, G.A. (1986) Nature, **322**, 652–656.

8. Levinson, G. and Gutman, G.A. (1987) Mol. Biol. Evol., **4**, 203–221.
9. Miesfeld, R., Krystal, M. and Arnheim, N. (1981) Nucl. Acids Res., **9**, 5931–5947.
10. Wharton, K.A., Yedbovnick, B., Finnerty, V.G. and Artavanis-Tsakonas, S. (1985) Cell, **40**, 55–62 .
11. Tautz, D., Tautz, C., Webb, D. and Dover, G.A. (1987) J. Mol. Biol., **195**, 525–542.
12. Hancock, J.M. and Dover, G. A. (1988) Mol. Biol. Evol., **5**, 377–391.
13. Treier, M., Pfeifle, C. and Tautz, D. (1989) EMBO J., **8**, 1517–1525.
14. Tautz, D. (1983) Thesis, University of Tübingen.
15. Lench, N., Stanier, P. and Williamson, R. (1988/I) Lancet, **8599**, 1356–1358.
16. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning, a Laboratory Manual, Cold Spring Harbor Laboratory, New York.
17. Sakai, R.K. et al. (1988) Science, **239**, 487–491.
18. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc Natl Acad Sci USA, **74**, 5463–5467.
19. Wharton, K.A., Johansen, K.M., Xu, T. and Artavanis-Tsakonas, S. (1985) Cell, **43**, 567–581.
20. Hamada, H., Petrino, M., and Kakunaga, T. (1982) Proc. Natl. Acad. Sci. USA, **79**, 5901–5905.
21. Streisinger, G. et al. (1966) Cold Spring Harb. Symp. quant. Biol., **31**, 77–84.
22. Efstratiadis, A. et al. (1980) Cell, **21**, 653–668.
23. Levinson, G. and Gutman, G.A. (1987) Nucl. Acids Res., **15**, 5323–5338.
24. Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985) Nature, **314**, 67–73.
25. Jeffreys, A.J., Royle, N.J., Wilson, V. and Wong, Z. (1988) Nature, **322**, 278–281.
26. Yu, Q., Colot, H.V., Kyriacou, C.P., Hall, J.C. and Rosbash, M. (1987) Nature, **326**, 765–769.
27. Smith, L.M. et al. (1986) Nature, **321**, 674–679.
28. Lewin, R. (1989) Science, **243**, 1549–1550.
29. Weber, J.L. and May, P.E. (1989) Am. J. Hum. Genet., **44**, 388–396.
30. Litt, M. and Luty, J.A. (1989) Am. J. Hum. Genet., **44**, 397–401.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.