Recombinant DNA Techniques in the Diagnosis of Inherited Disorders

James F. Gusella
Neurogenetics Laboratory, Massachusetts General Hospital, Boston, Massachusetts 02114; and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115

Introduction

Recombinant DNA technology has revolutionized the study of human inherited disease. Initially, the impact was felt in the elucidation of the molecular bases for defects in known genes such as mutations in the globin genes causing thalassemia. More recently, the realization that molecular techniques could be used to directly detect DNA sequence variation in human populations has fostered a general approach to the investigation of the great majority of genetic diseases in which the biochemical deficit is not known (1, 2).

Restriction fragment–length polymorphism (RFLP)[1]

Until recently, detection of DNA sequence differences at specific human loci was necessarily indirect, and limited to those causing altered expression of a protein. Only a relatively small number of plasma and red cell proteins displayed sufficient electrophoretic or antigenic variation to be generally useful to genetic investigators. Now, with the availability of molecular techniques, sequence differences in the DNA of two chromosomes can be directly monitored in several ways which do not require that they affect expressed sequences. By far the easiest and most commonly employed approach is to cut genomic DNA with a restriction endonuclease, fractionate the digest by gel electrophoresis, transfer the DNA to a solid filter support, and hybridize with a cloned DNA probe for the region in question to determine the sizes of fragments produced from the corresponding locus. A given restriction endonuclease will only cleave DNA at particular recognition sites where a specific sequence of bases occurs. Any alteration in the primary sequence of the DNA at one of these sites will eliminate cutting by the enzyme and will consequently change the size of the restriction fragment produced (see Fig. 1). Inherited differences in the pattern of enzyme digestion have been termed restriction fragment–length polymorphisms (2). These can also result, of course, from more complex differences than single base changes, such as insertion or deletion of DNA within a restriction fragment. They can be found either at known gene loci, or using anonymous DNA sequences of no known function.

Address correspondence to Dr. Gusella, Massachusetts General Hospital, Jackson 11, Boston, MA 02114.
Received for publication 25 November 1985.

1. *Abbreviations used in this paper:* CF, cystic fibrosis; HD, Huntington's disease; PIC, polymorphism information content; RFLP, restriction fragment–length polymorphism(s).

J. Clin. Invest.

0021-9738/86/06/1723/04 $1.00
Volume 77, June 1986, 1723–1726

The value of an RFLP, like any other genetic marker, is that it permits the investigator to distinguish the two homologous chromosomes in an individual heterozygous at the polymorphic site. The geneticist is therefore able to infer from the alleles present in children which parental chromosome was transmitted in each case. If the RFLP is at a defective locus known to be present in an individual, it enables inferences to be made concerning the inheritance of the gene defect in the family. The practical application of RFLPs for known disease loci, particularly in prenatal diagnosis, is straightforward and has been used with many cloned genes responsible for specific inherited defects such as thalassemia, sickle cell anemia, growth hormone deficiency, and phenylketonuria (3–6).

In inherited diseases where the functional deficit is unknown, the availability of RFLPs as genetic markers has opened the door to genetic linkage investigations aimed at determining the chromosomal positions of the disease loci (1, 2). The basis for such studies is the fact that two genes located close to one another on the same chromosome will be transmitted together with a frequency far greater than chance alone would predict. In fact, alleles at such sites will only be transmitted separately when a recombination event occurs between the two loci. The chances for such a crossover increase with the distance between the genes. The degree of co-inheritance of two genes can therefore be used as a measure of their proximity on the chromosome. Similarly, if a disease and a genetic marker show a correlated pattern of inheritance, it implies that the disease gene is in the same chromosome region as the marker locus.

The search for RFLPs

In 1980, when human geneticists were first considering the use of RFLPs as genetic markers, there were three fundamental questions, the answers to which would determine the future of this approach. Could DNA polymorphisms be found frequently in the human genome? Would the level of polymorphism and individual allele frequencies be high enough to make RFLPs useful genetic markers? Could RFLPs be found in all regions of the genome, especially the many regions not containing a standard expressed marker?

Five years later, the answer to each of these questions is a resounding yes! At the most recent Human Gene Mapping Workshop (HGM8) held in Helsinki, Finland in August 1985, a reasonably up-to-date list of RFLPs was compiled (see Fig. 2) (7). Compared with approximately 30 expressed markers, there are now 333 RFLP markers. many of which have multiple alleles or multiple polymorphic sites. Although 88 of the RFLP markers represent known gene loci, 245 have been found using anonymous DNA probes. As genetic markers, of course, each are equally useful for monitoring the transmission of the chromosomal region in which they reside.

The rate of discovery of new RFLPs, outlined in Fig. 2 suggests that many more DNA markers will be found in the next few years. Underlying this rapid accumulation is the high degree of sequence heterozygosity in the human genome. Several investigations have indicated that on average 1 base in 250 to 500 differs between any two chromosomes chosen at random (8–11). Thus, for any given probe, a screen using multiple restriction enzymes has a very high probability of identifying at least one RFLP. Predictably, larger probes are more likely to detect polymorphisms than shorter segments (12).

Most RFLPs have been found by simply comparing the pattern of fragments detected by a cloned probe in restriction enzyme–digested DNA from a relatively small number of unrelated individuals (13). Differences in pattern with a given enzyme can often be interpreted directly as the gain or loss of a restriction enzyme site. If similar differences are seen with many enzymes, an insertion or deletion is usually the basis for the polymorphism (14). To report a new RFLP, it is incumbent upon the investigator to demonstrate a Mendelian pattern of inheritance. Barker, Schafer, and White (15) have proposed a useful parameter, the Log Relative Mendelian Likelihood, to estimate whether there is sufficient support for Mendelian segregation. The reporting investigator should type a number of unrelated individuals to estimate allele frequencies in the population and to test whether the distribution of genotypes deviates from that expected for Hardy-Weinberg equilibrium.

The vast majority of individual RFLPs described involve the presence or absence of a restriction enzyme site and therefore have two alleles (7). The use of a small set of unrelated individuals in the original screen for RFLPs favors the detection of polymorphisms with a high level of heterozygosity (13). The frequency of the less common allele for most RFLP markers falls in the range of 0.15–0.5, which corresponds to levels of heterozygosity of 25–50%. This compares very favorably with most of the standard expressed markers. Two allele systems are naturally limited in informativeness, however, since at least 50% of the individuals will be homozygous at the locus, and it will be impossible to distinguish transmission of the two homologs.

The polymorphism information content or PIC value is a parameter that can be used to express the usefulness of a particular marker system (2). It is the likelihood that a child of an arbitrary mating will yield information concerning transmission of the locus in question. PIC is a function of the number of alleles at a locus and of their individual frequencies in the general population. A fully informative marker would have a PIC of one. The maximum PIC for a two-allele system is 0.37, which indicates that on average only one mating in three will be informative. Few of the standard expressed markers exceed a PIC of 0.37. Fortunately, a major advantage of DNA markers is the ability to combine into haplotypes the individual alleles at RFLPs closely spaced in the genomic DNA, and thereby generate multiallele systems with greater informativeness (Fig. 1). For most markers of this kind, it is necessary to digest DNA using multiple different restriction enzymes to monitor each of several different two-allele systems.

The most useful DNA markers are those that detect multiallele systems due to length variation of the DNA within a particular fragment. Unfortunately, these markers are less frequent than single site variations. It has recently been determined, however, that the basis for this length variation is likely to be a high frequency of unequal crossovers within a set of short direct repeats (16). RFLPs of the insertion/deletion type have been described near the insulin, Harvey ras, alpha globin, and myoglobin genes as well as with several anonymous probes (7). At least one of these repeat units shows partial homology with the chi recombination signal of *Escherichia coli* (16). It is likely that a systematic search for similar loci using these repeats as probes will yield many more multiallele markers.

As indicated in Fig. 2, RFLPs have been found on every human chromosome including the Y chromosome. Their occurrence is not restricted to a small region of the genome. In fact, for some chromosomes, there are enough markers to have permitted the construction of preliminary genetic linkage maps by tracing the inheritance of all loci in large "reference" pedigrees (17, 18). Within a few years, it is probable that there will be a complete linkage map for the entire genome, permitting linkage analysis in all regions. It will then be possible to localize any disease gene showing a clearcut pattern of inheritance using a standard battery of highly informative linkage markers. Already the availability of large numbers of testable loci has begun to alter the way in which linkage data is analyzed. Formerly, individual markers were tested sequentially for linkage to a disease locus. Recently, computer programs have been produced that are capable of simultaneous analysis of multiple markers with known linkage relationships (19). It is possible that the improvement and extension of multipoint methods of analysis combined with a detailed human linkage map will eventually permit a linkage approach to diseases involving multifactorial inheritance, including complex behavioral disorders.

Diseases mapped by linkage to RFLPs

The use of RFLPs as genetic linkage markers has already resulted in the successful localization of a number of disease loci. The approach has been most intensively pursued on the X-chromosome. The first X-linked disease for which a linked RFLP was identified was Duchenne muscular dystrophy in 1982 (20). A concentrated effort on the Duchenne muscular dystrophy region has now produced a large number of linked genetic markers and has brought investigators to the very brink of identifying the gene defect itself (21–23). The ease of analysis of sex-linked disorders, due to the presence of a single allele at each locus in males, has allowed the detection of linked markers for Alport syndrome, Becker muscular dystrophy, X-linked Charcot-Marie-Tooth disease, choroideremia, Fragile X mental retardation syndrome, Menkes disease, ocular albinism, X-linked retinitis pigmentosa, and retinoschisis (24). In fact, given the 68 polymorphic DNA markers that have been generated for this chromosome, it is virtually guaranteed that a linkage can be found for any X-linked disorder with an adequate number of reasonable-sized pedigrees.

The first major success in the application of the RFLP approach to autosomes came with the discovery in 1983 of a marker tightly linked to Huntington's disease (HD) (25, 26). As a result, the defect has now been assigned to the terminal region of the chromosome 4 short arm (27). More recently, it has been reported that the alpha globin locus is genetically linked to the defect causing polycystic kidney disease (28). A large number of laboratories have undertaken the search for linked RFLPs for many other diseases including familial Alzheimer's disease, Von Recklinghausen's neurofibromatosis, central neurofibromatosis, Von Hippel-Lindau disease, Dystonia musculorum deformans, Tourette's syndrome, manic-depressive disorder, multiple endocrine neoplasia, etc. Every genetic disorder is a potential candidate for this approach if enough families are available for in-

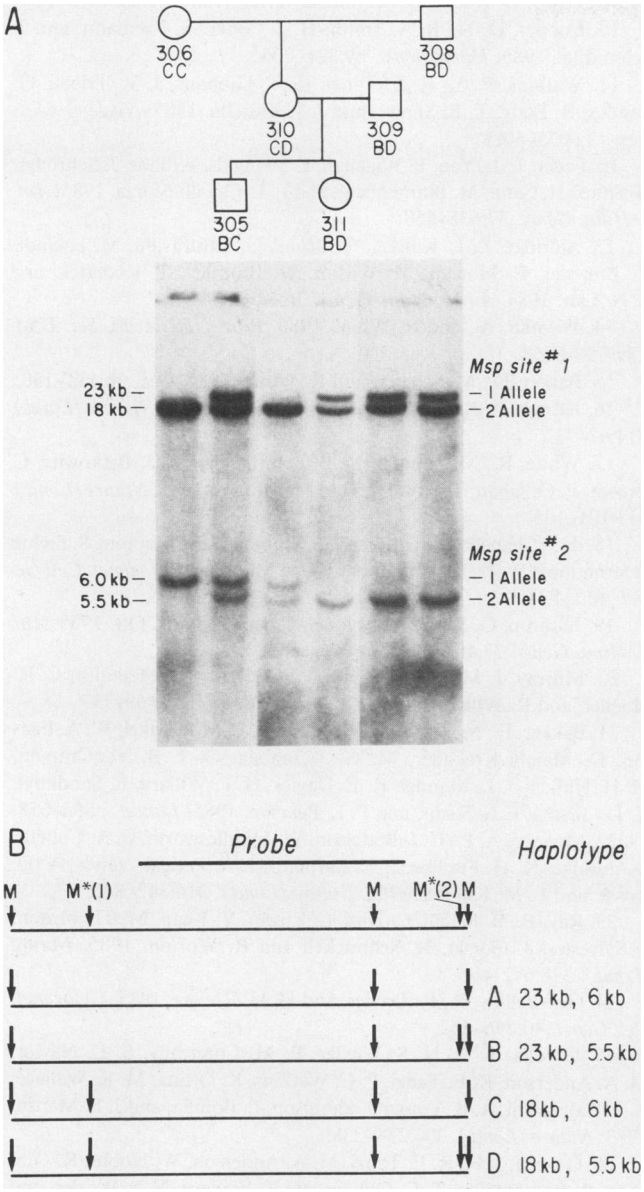

Figure 1. A DNA marker displaying restriction fragment length polymorphism. (*A*) The inheritance pattern of two polymorphic MspI sites is followed in a small family. Each gel lane contains DNA from a different family member. The fragments are detected by the probe G9, an anonymous segment of DNA from chromosome 4. The locus detected by this probe has been designated *D4S35* according to the nomenclature conventions set out in reference 7. The pattern of fragments observed can be interpreted with reference to the haplotypes shown in *B* to derive the genotype at this locus in each individual. No example of an A haplotype is seen in this particular family. (*B*) The relative positions of MspI digestion sites at the *D4S35* locus are shown. Polymorphic sites are indicated by an asterisk. The four possible haplotypes derived from the various combinations of presence or absence of the variable sites have been designated A, B, C, and D.

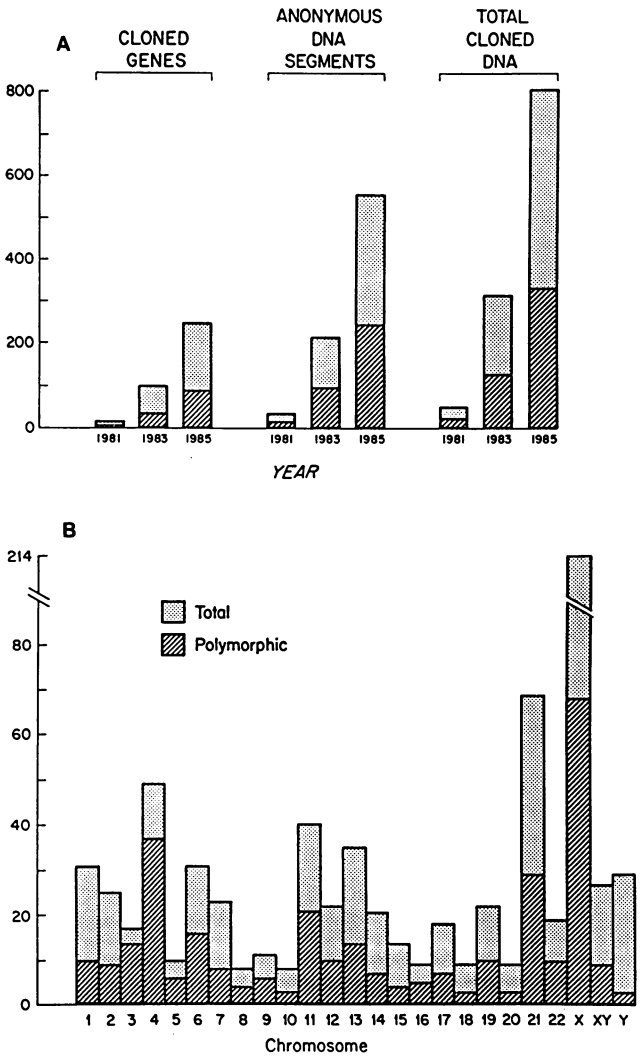

Figure 2. RFLP markers in the human genome. (*A*) The number of cloned genes and anonymous DNA sequences reported and compiled at Human Gene Mapping Workshops 6 (1981), 7 (1983), and 8 (1985) are shown along with the proportion of each that detects an RFLP. The majority of those sequences for which no RFLP is known have simply not yet been tested (7). (*B*) The number of RFLP markers on each human chromosome is shown. Sequences detecting loci on both the X and Y chromosomes are entered in the bar labeled XY (7).

vestigation. Even in cases where the chromosomal location of the defect is already known, such as myotonic dystrophy, RFLP analysis is resulting in identification of more tightly linked markers. The method is not limited to dominant disorders, as demonstrated very recently by the discovery of linked RFLP markers for cystic fibrosis (CF), the most common lethal recessive disease in caucasians. Within a very short time, several groups have identified DNA markers tightly linked to the CF locus on chromosome 7 (29–32).

Genetic linkage analysis is also one of the fastest methods of determining whether a particular cloned gene is a candidate for representing the site of the primary defect. The observation of a single recombination between a gene encoding a known protein and a disease locus can obviate months of biochemical investigations aimed at implicating the protein as cause of the disorder (4, 33).

Why map disease genes?

The mapping of human disease genes to particular chromosomal locations is not simply an exercise in information gathering

without any practical impact. The identification of a genetic marker linked to a disease gene has both immediate and long-term ramifications. A linked locus allows investigators to infer the inheritance of a disease locus in informative families even when no phenotype attributable to the defect can be determined. Thus, the first clinical impact of a linkage marker would, for many diseases, be to provide improved diagnosis. In many disorders, prenatal diagnosis would become available for the first time while for late onset disorders such as HD, presymptomatic diagnosis is a possibility.

The linked marker can also be used to provide more information about the genetic basis for the disease phenotype. For example, a fundamental issue in any disorder is whether the symptoms in all affected individuals are due to mutation at the same locus, or whether more than one defect (nonallelic heterogeneity) can result in the same phenotype. This is particularly important in a case such as HD where preclinical diagnosis is under consideration. Similar considerations apply to the potential use of RFLPs for prenatal diagnosis in CF. Nonallelic heterogeneity is a possibility in even apparently homogeneous disorders and the issue can only be resolved using genetic linkage analysis.

In addition to aiding in disease categorization, the linked marker can be used to explore the relationship of any of a large number of biochemical, physical, and behavioral parameters to the presence of the primary defect and to assess their involvement in the disease process. Perhaps the greatest impact of the chromosomal localization of a disease gene, however, is the avenue it opens toward cloning and characterizing the defective gene sequence without resort to a knowledge of the encoded protein. A number of novel techniques are currently being developed to aid in implementing this "reverse genetics" approach which may ultimately result in the isolation of many of the more important human disease genes (26). The identification of defective proteins through cloning of the coresponding gene sequences will undoubtedly improve our understanding of the biochemical bases for many disease phenotypes and may, in some cases, suggest effective forms of therapy.

Acknowledgments

This work was supported by National Institute of Neurological and Communicative Disorders and Stroke grants NS16367 (Huntington's Disease Center Without Walls) and NS20012, and by grants from the McKnight Foundation and Hereditary Disease Foundation. The author is supported by the Searle Scholar Program/Chicago Community Trust.

References

1. Housman, D., and J. F. Gusella. 1982. *In* Molecular Genetic Neuroscience: A New Hybrid. F. O. Schmitt, S. Bird, and F. E. Bloom, editors. Raven Press, New York. 415–424.

2. Botstein, D., R. L. White, M. Skolnick, and R. W. Davis. 1980. *Am. J. Hum. Genet.* 32:314–331.

3. Orkin, S. H., P. F. R. Little, H. H. Kazazian, and C. Boehm. 1982. *N. Eng. J. Med.* 307:32–36.

4. Phillips, J. A., J. S. Parks, B. L. Hjelle, J. E. Herd, and P. Plotnick. 1982. *J. Clin. Invest.* 70:489–495.

5. Woo, S. C., A. S. Lidsky, F. Guttler, C. Thirumalachary, and K. J. H. Robson. 1984. *J. Am. Med. Assoc. (JAMA).* 251:1998–2001.

6. Orkin, S. 1984. *Blood.* 63:249–253.

7. Willard, H. F., M. H. Skolnick, P. L. Pearson, and J. L. Mandel. 1985. *Cytogenet. Cell Genet.* 40:360–489.

8. Murray, J. C., K. A. Mills, C. M. Demopolous, S. Hornung, and A. G. Motulsky. 1984. *Proc. Natl. Acad. Sci. USA.* 81:3486–3490.

9. Cooper, D. N., and J. Schmidtke. 1984. *Hum. Genet.* 66:1–16.

10. Cooper, D. N., B. A. Smith, H. J. Cooke, S. Niemann, and J. Schmidtke. 1985. *Hum. Genet.* 69:201–205.

11. Watkins, P. C., R. E. Tanzi, K. T. Gibbons, J. V. Tricoli, G. Landes, R. Eddy, T. B. Shows, and J. F. Gusella. 1985. *Nucleic. Acids Res.* 13:6075–6088.

12. Feder, J., L. Yen, E. Wijsman, L. Wang, L. Wilkins, J. Schroder, N. Spurr, H. Cann, M. Blumenberg, and L. L. Cavalli-Sforza. 1985. *Am. J. Hum. Genet.* 37:635–650.

13. Aldridge, J., L. Kunkel, G. Bruns, U. Tantravahi, M. Lalande, T. Brewster, E. Moreau, M. Wilson, W. Bromley, T. Roderick, and S. A. Latt. 1984. *Am. J. Hum. Genet.* 36:546–564.

14. Wyman, A., and R. White. 1980. *Proc. Natl. Acad. Sci. USA.* 77:6754–6758.

15. Barker, D., M. Schafer, and R. White. 1984. *Cell.* 36:131–138.

16. Jeffreys, A. J., V. Wilson, and S. L. Thein. 1985. *Nature (Lond.).* 314:67–73.

17. White, R., M. Leppert, D. Bishop, D. Barker, J. Berkowitz, C. Brown, P. Callahan, T. Holm, and L. Jerominski. 1985. *Nature (Lond.).* 313:101–105.

18. de la Chapelle, A., editor. 1985. Human Gene Mapping 8: Eighth International Workshop in Human Gene Mapping. *Cytogenet. Cell Genet.* 40:1–823.

19. Lathrop, G. M., J. M. Lalouel, C. Julier, and J. Ott. 1985. *Am. J. Hum. Genet.* 37:482–499.

20. Murray, J. M., K. E. Davies, P. S. Harper, L. Meredith, C. R. Mueller, and R. Williamson. 1982. *Nature (Lond.).* 300:69–71.

21. Bakker, E., N. Goor, K. Wrogemann, L. M. Kunkel, W. A. Fenton, D. Majoor-Krakauer, M. G. J. Jahoda, G. J. B. VanOmmen, M. H. Hofker, J. L. Mandel, K. E. Davies, H. F. Willard, L. Sandkuyl, A. J. v. Essen, E. S. Sachs, and P. L. Pearson. 1985. *Lancet.* i:655–658.

22. Monaco, A. P., C. J. Bertelson, W. Middlesworth, C. A. Colletti, J. Aldridge, K. H. Fischbeck, R. Bartlett, M. A. Pericak-Vance, A. D. Roses, and L. M. Kunkel. 1985. *Nature (Lond.).* 316:842–845.

23. Ray, P., B. Belfall, C. Duff, C. Logan, V. Kean, M. Thompson, J. Sylvester, J. Gorski, R. Schmickel, and R. Worton. 1985. *Nature (Lond.).* 318:672–675.

24. Goodfellow, P., K. Davies, and H. H. Ropers. 1985. *Cytogenet. Cell Genet.* 40:296–352.

25. Gusella, J. F., N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi, A. B. Young, I. Shoulson, E. Bonilla, and J. B. Martin. 1983. *Nature (Lond.).* 306:234–238.

26. Gusella, J. F., R. E. Tanzi, M. A. Anderson, W. Hobbs, K. Gibbons, R. Raschtchian, T. C. Gilliam, M. R. Wallace, N. S. Wexler, and P. M. Conneally. 1984. *Science (Wash. DC).* 225:1320–1326.

27. Gusella, J. F., R. E. Tanzi, P. I. Bader, M. C. Phelan, R. Stevenson, M. R. Hayden, K. J. Hofman, A. G. Faryniarz, and K. Gibbons. 1985. *Nature (Lond.).* 318:75–78.

28. Reeders, S. T., M. H. Breuning, K. E. Davies, R. D. Nicholls, A. P. Jarman, D. R. Higgs, P. L. Pearson, and D. J. Weatherall. 1985. *Nature (Lond.).* 317:542–544.

29. Tsui, L., M. Buchwald, D. Barker, J. Braman, R. Knowlton, J. Schumm, H. Eiberg, J. Mohr, D. Kennedy, N. Plavsic, M. Zsiga, D. Markiewicz, G. Akots, C. Helms, T. Gravius, C. Parker, K. Rediker, and H. Donis-Keller. 1985. *Science (Wash. DC).* 230:1054–1057.

30. Knowlton, R. G., O. Cohen-Hagenauer, N. Van Cong, J. Frezal, V. A. Brown, D. Barker, J. C. Braman, J. W. Schumm, L. C. Tsui, M. Buchwald, and H. Donis-Keller. 1985. *Nature (Lond.).* 318:380–382.

31. White, R. L., S. Woodward, M. Leppert, P. O'Connell, M. Hoff, J. Herbst, J. M. Lalouel, M. Dean, and G. Van de Woude. 1985. *Nature (Lond.).* 318:382–384.

32. Wainwright, B., P. J. Scambler, J. Schmidtke, E. A. Watson, H. Y. Law, M. Farrall, H. J. Cooke, H. Eiberg, and R. Williamson. 1985. *Nature (Lond.).* 318:384–386.

33. Seizinger, B. R., R. E. Tanzi, T. C. Gilliam, J. Bader, D. Perry, A. Spence, M. Marazita, K. Gibbons, W. Hobbs, and J. F. Gusella. 1985. *Ann. NY Acad. Sci.* In press.