**Title**
Cloning and characterization of the Huntington disease region on human chromosome 4



**Author**
Zuo, Jian,

**Publication Date**
1993

Peer reviewed|Thesis/dissertation

# CLONING AND CHARACTERIZATION OF THE HUNTINGTON DISEASE REGION ON HUMAN CHROMOSOME 4

by

JIAN ZUO

**DISSERTATION**

**Submitted in partial satisfaction of the requirements for the degree of**

**DOCTOR OF PHILOSOPHY**

**in**

PHYSIOLOGY

**in the**

**GRADUATE DIVISION**

**of the**

**UNIVERSITY OF CALIFORNIA**

**San Francisco**

This thesis is dedicated to my family,

and to those who have suffered from Huntington disease.

## INVOLVEMENT OF COAUTHORS-ADVISOR'S STATEMENT

Jian Zuo's thesis work resulted in the publication of three research articles. All of these articles, including two that comprise Chapters One and Two of the thesis, involve coauthors. For Chapter One and Two, the authorship reflects the contributions of two other individuals in each chapter. For both chapters, the vast majority of the work was designed and carried out by Jian himself independently. The involvement of collaborators in this work is the inevitable consequence of research in a rapidly progressing and competitive field. In addition, Jian carried out the initial studies and contributed significantly on the work described in Appendix One. I consider Jian's thesis work very successful.

Richard M. Myers

Richard M. Myers

## ACKNOWLEDGEMENTS

Yes, we will question everything, everything once again. And what we find today we shall strike out from the record tomorrow and only write it in again when we have once more discovered it.

-----Galileo

It has been such a long trip. It is impossible to describe here what the graduate school means to me.

I have witnessed the development of two foremost scientific frontiers: human genetics and human genome project. I thank Rick Myers, my thesis advisor, for his encouragement, advice and support for getting me interested and making me independent in these research areas. After all these wonder years, I begin to appreciate what Rick told me when I first started, "Every student should cry at least three times in my lab." I also acknowledge Rick as a friend for encouraging me to pursue whatever I am interested in but not what the society thinks.

I would like to thank the members of my thesis committee, David Cox, Jane Gitschier, and Andrew Murray, for their criticisms on my proposals and their constant encouragement during the course of this research. Particularly, David has been a model human geneticist to me; his criticism and amusements in our lab meetings and parties have always inspired me.

I thank Jim Hudspeth for his advice and support at UCSF and for getting me interested in neuroscience. I also thank Ron Vale and Lou Reichardt for teaching me how to be a good scientist both during my rotations in their labs and

afterwards. Zach Hall, as the chairman of our department, has always helped me to get through the difficult times.

At UCSF, I have made many good friends. They worked with and taught me tremendously about life, science and America. In particular, Andy Peterson, Catrin Pritchard, Grant Hartzog have got me started in molecular biology and I am grateful for their constant intellectual and philosophical stimulation during my daily work. Carolyn Robbins has been extremely entertaining and helpful in some of the experiments in this competitive field. I regret that I cannot mention all of the current and former members of the Myers' lab and Hudspeth's lab, who have been working together as a harmonious family. It is this environment that has made me eager to come to work long hours almost every day and every weekend.

I thank Nori Kasahara, Karen Scribner for helping me during the first few years of graduate school at UCSF. Yi Rao has been a good friend for helping me throughout my graduate career and for working together on many occasions.

I thank Tal Teitz, who has made my life more enjoyable, and Ye Hu, my cousin, who is always there when I call her.

Finally I owe to my parents for their love, consistent encouragement and support, and for sharing any joy with me.

**Cloning and characterization of the Huntington disease region on human chromosome 4**

**Jian Zuo**

## Abstract

Huntington disease (HD) is an autosomal dominant neurodegenerative disease, characterized by late onset of symptoms, including chorea, psychiatric problems, and dementia. The gene responsible for the disease has been localized to a 2.2 million base pair (Mb) region on the short arm of chromosome 4. As part of a strategy to identify the HD gene on the basis of its chromosomal location, I have primarily worked on cloning of the 2.2 Mb HD region.

I first utilized the pulsed-field gel electrophoresis technique to determine the proximal positions of a large number of DNA probes isolated from the HD region. A long range restriction map with four rare cutting restriction sites, NotI, MluI, NruI and CspI, of the HD region was constructed.

I further developed a large number of PCR based sequence-tagged-sites (STS) from these DNA markers. These STSs were then used to screen the total human YAC libraries at Washington University. A total number of 28 YACs from these screens were obtained and characterized by determining their sizes of inserts, their probe-content, and by isolating and characterizing their ends. This maneuver resulted in a 2.3 redundant coverage of more than 2 Mb of the HD region.

To completely characterize the HD region, we developed a method to isolate cosmids from the HD region by screening a chromosome specific cosmid library with YAC probes. Based on the overlapping pattern of the characterized

YACs , we were able to put the cosmids into bins and further determine their restriction maps. An EcoRI restriction map of 90% of the HD region was constructed and all the available DNA markers from the HD region were mapped to the corresponding EcoRI fragments. Furthermore, the sites of NotI and MluI were determined by double restriction digestion in combination with EcoRI.

The contigs of cosmids and YACs, as well as the high resolution map, of the HD region provide necessary material for identification of polymorphic markers, cDNA clones, for characterization of the candidate genes for HD and for sequencing the entire HD region.

Richard M. Myers

Richard M. Myers

Thesis advisor

## TABLE OF CONTENTS

## LIST OF TABLES

### INTRODUCTION



### CHAPTER ONE



### CHAPTER TWO



### APPENDIX ONE

## LIST OF FIGURES

### INTRODUCTION



### CHAPTER ONE



### CHAPTER TWO

**APPENDIX ONE**

## INTRODUCTION

In 1872, at the age of 22, George Huntington made his most important contribution to science. He was the first person to describe the hereditary nature of a human neurological disorder. "The hereditary chorea, as I call it, is confined to certain and fortunately few families and has been transmitted to them, an heirloom from generations way back in the dim past." He described that "when either or both parents have shown manifestations of the disease, one or more of the offspring invariably suffer from the disease. It never skips a generation to again manifest itself in another." This disorder has characteristics of "all symptoms of common chorea, only in an aggravated degree, hardly ever manifesting itself until adult or middle life, and then coming on gradually but surely, increasing by degrees and often occupying years in its development, until the hapless sufferer is but a quivering wreck of his former self." (Huntington, 1872)

Huntington's description of the disorder was remarkably accurate and the disease was later called Huntington disease (HD). Since his description of the hereditary nature of the chorea was consistent with and only six years later than Mendel's classic observations of crosses in plants, HD has been a classical human genetic disease that has contributed to the fundamental concept of the genetic inheritance.

Despite the additions to the original description by George Huntington, HD is still recognized as a neurodegenerative disorder, with an autosomal dominant inheritance, beginning in mid-life, characterized by chorea and dementia, with death 15 to 20 years later. The cause of HD has baffled scientists for over a

century. In March 1993, a gene that potentially causes the disease was finally identified(THDCRG, 1993).

Here I briefly summarize the current knowledge and the historical developments in the studies of HD.

## Clinical features of Huntington disease

### The Motor Disorder

Involuntary movements and abnormal voluntary movements are the two components of chorea in HD patients (Folstein, 1989). The most common symptoms are jerkiness, clumsiness and mild uncoordination, which usually herald the onset of chorea. During the early stage of the disease, a typical sign is the spontaneous choreiform movements elicited in patients on squeezing the examiner's two fingers for five seconds with constant pressure. Another characteristic symptom is the inability of the patient to protrude his/her tongue for more than five seconds. The failure to walk in straight line and to perform complex facial movements are also early symptoms of chorea. Rapid, flickering involuntary movements, failure of saccadic eye movements and disruption of coordination are the usual signs of the early symptoms. As the disease progresses, chorea becomes more frequent, pervasive and of higher amplitude. This is characterized by ceaseless writhing, jerking and twisting of different parts of the body, particularly hands and feet. About half of the HD patients also develop rigidity and spasticity. At the end of the 15 to 20 year disease progression, patients usually die of aspiration pneumonia, choking, heart attack, hematomas or suicide.

Cognitive and emotional disorders

In over half of the patients, cognitive and emotional disturbances precede the development of abnormal movements, sometimes by many years (Folstein, 1989; Wexler et al., 1991). Disruption of memory, organizational ability and judgment cause premature loss of employment. Despite severe cognitive impairments, patients retain the knowledge of their own identities and their friends and family's identities. They are often aware of their failures in intellectual ability. Depression is the most frequent psychiatric problem. More than 25% of HD patients attempt suicide. Mania, hallucinations, and delusions are less frequent, while irritability, apathy, or explosive outbursts are common.

Juvenile onset patients have distinctive symptoms (Wexler et al., 1991). Before the age of twenty, they develop a unique conjunction of parkinsonian symptoms, such as rigidity and tremor, with choreic and dystonic features and, frequently seizures. Their cognitive and emotional disturbances are similar to the adult onset patients but the course of disease is much accelerated, as it culminates in death in eight to ten years.

Since there are variations in the clinical features of chorea in the HD patients, it is easily misdiagnosed with other diseases such as Parkinson's disease Alzheimer's disease, schizophrenia and Tradeoff dyskinesia (Folstein, 1989; Harper, 1991; Hayden et al., 1988). To distinguish HD from other diseases, family history is an important criteria. In addition, CT scanning can also be used to trace the degeneration of neurons in the basal ganglia. However, definitive proof for the diagnosis of HD is not available before a postmortem brain examination.

## Pathophysiology of Huntington disease

HD is a neurodegenerative disease. The most obvious histological degeneration in HD patients is in the basal ganglia, particularly caudate nucleus and putamen (Wexler et al., 1991). The basal ganglia, embryologically derived from the telecephalon, consists of striatum, globus pallidus, substantia nigra and subthalamic nucleus. It serves as a highly complex processing station to integrate cortical and subcortical inputs involved in visual, labyrinthine, and proprioceptive information. The defects in this region of the brain are expected to manifest the disruption of coordination of movement, cognition and affective status as in the HD patients.

Which brain region and cell types are affected earliest and most severely are the crucial questions in unraveling the physiological basis of the defect. The genetically programmed cell death starts in the tail of the caudate and follows a medial-to-lateral gradient, whereas degeneration in putamen progresses dorsolaterally (Martin & Gusella, 1986). The ventral putamen and nucleus accumbens are spared. As the disease progresses, there is also cell loss in the cortex, external segment of globus pallidus, and later in the hypothalamus and cerebellum. The spiny neurons and their axons of the striatum are most severely depleted in the HD brain. These spiny neurons project to targets outside the caudate and they all contain the neurotransmitter gamma-aminobutyric acid (GABA). Some spiny neurons also contain substance P and dynorphin or enkephalin in addition to GABA. The other structural neurons, aspiny interneurons, are mostly unaffected. These neurons contain mainly acetylcholine, somatostatin and neuropeptide Y. The striatum maintains its architecture in patches and matrix despite the extensive neuronal loss even at very late stages of the disease.

To date, the earliest reported abnormality of neurons in the caudate nucleus of HD patients is the curling, branching, and arborizations of the dendrites of medium size spiny neurons. These changes indicate that the neurons undergo simultaneous degeneration and regeneration as the consequence of the primary defect (Graveland et al., 1985). As neuronal loss progresses, astroglial cells in the striatum become more predominant. Postmortem studies of brain tissues from patients have revealed that levels of as many as 30 neurotransmitters, biosynthetic enzymes, and receptor binding sites in the striatum are abnormal (Martin and Gusella, 1986). While such increases or decreases may contribute to the symptoms of the disease, they are likely to be the secondary effects.

These earliest morphological changes in the dendrites of the spiny neurons in HD patients may provide some clues for understanding the nature of the defect. Considering the mechanism of programmed cell death, two morphological types of cell death have been distinguished: apoptosis and necrosis (Clarke, 1990; Raff, 1992). Apoptosis is considered to be part of normal development, while necrosis is the cellular response to the abnormal environmental stimuli. In the process of apoptosis, chromatin in the nucleus is first condensed, nuclear DNA is cleaved by a $Ca^{++}/Zn^{++}$-dependent endonuclease at internucleosome sites, the nuclear envelope shrinks down to small spheres, the plasma membrane convolutes and forms blebs, and the cytoplasm is reduced in volume. Eventually the cell is engulfed by phagocytes so that inflammation is avoided. Necrosis, on the other hand, involves swelling of mitochondria, dilation of endoplasmic reticulum and formation of vacuoles in the cytoplasm. Eventually the nucleus swells and cell membranes rupture, leading to inflammation. These morphological phenotypes indicate that there are at least two distinguishable mechanisms involved in cell death processes.

Numerous neurodegenerative diseases offer good examples for studying programmed cell death in CNS development. An example of necrotic cell death in the

mammalian CNS is given by a mouse mutant, retinal degenerate slow (rds). The rds gene encodes a membrane structural protein, peripherin, in rod and cone outer segments. The expression of the mutated protein causes an abnormal disc morphology and vesicle formation before the rod outer segments degenerate (Travis et al., 1989). In humans, vacuolated neurons have been reported in some instances of Alzheimer's disease, in which the amyloid precursor protein gene, another membrane glycoprotein, is mutated (Goate et al., 1991). Another example of necrosis in the CNS is the Menkes disease, which manifests as progressive neurodegeneration with accumulation of copper in mitochondria. The mutation in the Menkes disease gene, which encodes an intracellular membrane associated copper transporter, likely causes the defective copper homeostasis in neurons (Vulpe et al., 1993). These genes may be involved in a cell death program characteristic of necrosis. By contrast, no genes have been identified to be responsible for developmental neurodegenerative phenotypes in mammals with characteristics of apoptosis.

It is unknown whether cell death in the HD brain is characteristic of apoptosis or necrosis, since it is very difficult to study the morphological changes at subcellular level that are associated with the earliest symptoms of the disease. Having the HD gene may provide a way to determine the types of cell death in the HD brains.

## Genetics of Huntington disease

HD is a classical autosomal dominant disease. The complete penetrance of HD has distinguished it from other autosomal dominant diseases. Its low mutation rate, the mode of action of the mutation, and the late onset of the disease are the most interesting and intriguing features of this genetic disease.

## Low mutation rate

In spite of the diversity of the clinical symptoms of the HD, the number of mutations responsible for the disease is thought to be low. First of all, many HD patients in different parts of the world can apparently be traced back to a single ancestral founder in 17th century in the north-west of Europe (Hayden, 1981). The best example is the large Venezuelan HD pedigree, which consists of more than two hundred and fifty affected individuals, and more than one thousand individuals at risk. All these individuals are believed to be the descendants of a Spanish sailor who resided around Lake Maracaibo in the early 1860s. Most strikingly, the single ancestral HD mutation in this pedigree has manifested a complete spectrum of symptoms of the disease, which strongly argues that the diversity of clinical features of the disease is not due to the heterogeneity of the mutations or allelic mutations (Bates & Lehrach, 1993). In addition, the notion of low mutation rate is supported by prevalence studies, which show that in a population of 100, 000, the number of HD patients is about 1 in Northern America and 4-7 in Northern Europe (Bates and Lehrach, 1993; Hayden, 1981). The prevalence in Asians and African Blacks is much lower. In addition, there are a few reports of putative sporadic HD patients but they fail to transmit the disease to succeeding generations (Harper, 1991; Hayden, 1981), even though they have been later proven to contain the mutation in the HD gene (THDCRG, 1993). Since the HD mutation has been identified now, it's thus possible to directly determine the mutation rate.

## Mode of action of the mutation

A dominant phenotype can result from either a gain-of-function mutation, a dominant negative mutation, or a haplo-insufficiency mutation. A gain-of-function mutation enhances the normal function of the gene, while a dominant negative mutation acts by inhibiting the normal function of the gene (Herskowitz, 1987). Underproduction of the normal protein below a certain threshold, in cases of haplo-insufficiency, can also result in a dominant phenotype.

Several lines of evidence might give some clues for the mode of action of the HD mutation. In the Venezuela HD pedigree, several sibs from a mating between two affected individuals have been shown to have a high probability of being homozygous for the HD mutation (Wexler et al., 1987). In a separate study, one of the four offspring in a New England HD family showed a 95% chance of being homozygous (Myers et al., 1989). These likely homozygotes, which were confirmed when the HD gene was recently identified, showed no obvious clinical differences from the heterozygous individuals in the same families. The fact that two copies of the mutated gene have the same effect as one copy does not distinguish whether the HD mutation is a gain-of-function, a dominant negative or a haplo-insufficiency. In addition, patients with Wolf-Hirschhorn syndrome (WHS), a distinctive chromosomal disorder, have deletions of the distal parts of their chromosome 4 short arms, where the HD mutation is located (see below) (Gusella et al., 1985). These WHS patients have no symptoms of HD. Thus, two normal copies of the HD gene seem to have the same effect as one normal copy, indicating that the null mutation of the HD gene has no effect on its function resulted from another normal copy of the HD gene. However the interpretations of these studies on WHS are inconclusive due to the early death of the patients in infancy and due to the lack of neuropathological

examination of their brain structure. The nature of the HD mutation therefore remains to be elucidated by further studying the function of the HD gene.

Onset of disease

Although the symptoms of HD generally occur in the 4th decade with an average onset of 38 years of age, the age of onset varies from early in childhood to late in the 7th decade (Hayden, 1981). About 5-6% of HD cases develop symptoms at an early stage of life, usually before 20 years of age. More than three quarters of these juvenile HD patients have inherited the disease from their fathers rather than from their mothers, while those who develop symptoms earlier than 10 years of age are exclusively the offspring of paternal transmission of the disease (Harper, 1991). The tendency to develop the symptoms through paternal transmission earlier than through maternal transmission is well-documented in a study of large numbers of HD families (Conneally, 1984). It has been proposed that this phenomenon is likely due to protection by a maternal factor that interacts with the HD gene (Harper, 1991). Alternatively, a modifying gene in the sex chromosome may influence or interact with the HD gene. Or the HD gene itself is subject to imprinting. However there is no direct evidence to distinguish these three possibilities.

Anticipation is a genetic term to describe the earlier age of onset or more severe symptoms of a disease in succeeding generations. Anticipation in the paternal transmission of HD has been studied primarily in HD patients in the U.S. (Ridley et al., 1991) and is particularly interesting since three other genetic diseases, Fragile X syndrome, Myotonic dystrophy and Kennedy's disease, have also been shown to manifest similar anticipation in both age of onset and severity of symptoms (Caskey et al., 1992). The recent identification of the HD mutation

showed that a similar mechanism, trinucleotide repeat expansion, underlies all these defects (THDCRG, 1993). A comparison of these four diseases in the context of the triplet repeat expansion will be given in the last section of this introduction.

## Positional Cloning of the HD Gene

A positional cloning strategy for the identification of a genetic defect underlying a disease has unique features that distinguish it from other approaches. This strategy utilizes chromosomal rearrangements or meiotic crossover events to identify the location of a gene on one of the chromosomes. It does not require any prior knowledge about the biochemical properties or anatomical location of the gene product. This approach, first applied in its modern molecular form in the 1980's, has opened a new frontier for studying human genetic diseases.

Historically the first clue for identifying a mutant locus has come from cytogenetic evidence. Karyotype analysis of HD patients has revealed no abnormality in their chromosomal structure that is directly associated with HD (Bates and Lehrach, 1993; Pritchard et al., 1992). One report described a family in which a balanced reciprocal translocation between the long arm of chromosome 4 and the short arm of chromosome 5 segregates with HD in 2 generations, but there is no further study on this family (Froster-Iskenius et al., 1986). This translocation event may be a coincidence that happened only in these HD chromosomes long after the HD mutation.

Without chromosomal rearrangements, such as translocations, inversions or deletions, it was difficult to determine the location of disease loci. Thus, meiotic crossovers were needed to localize the HD gene.

Genetic linkage studies

An alternative approach to cytogenetic methods for determine the location of the HD gene is to trace the segregation of disease phenotypes with polymorphic genetic markers. Although protein polymorphisms have traditionally been used as a genetic marker, the ability to detect DNA polymorphisms by using restriction enzymes offers a powerful approach to determine chromosomal locations of genetic disease loci and even to construct a genetic map of the entire human genome (Bostein et al., 1980). More specifically, if a polymorphic DNA probe detects a particular restriction fragment that mostly cosegregates with the disease in a particular pedigree, the likelihood that this marker is genetically linked to the disease locus is significantly higher than that it is not linked to or is randomly associated with the disease locus.

In 1983, Gusella et al tested the possible linkage of genomic DNA probes to the HD locus (Gusella et al., 1983; Harper, 1991). Two important factors facilitated their initial success. First, the pedigree structure of a few large HD families were extensively studied. Blood samples and lymphoblastoid cell lines from a large Venezualan pedigree, for example, were established. Second, among the first dozen random genomic DNA clones they tested, a marker, G8 at the D4S10 locus, showed extremely high linkage to the HD locus.

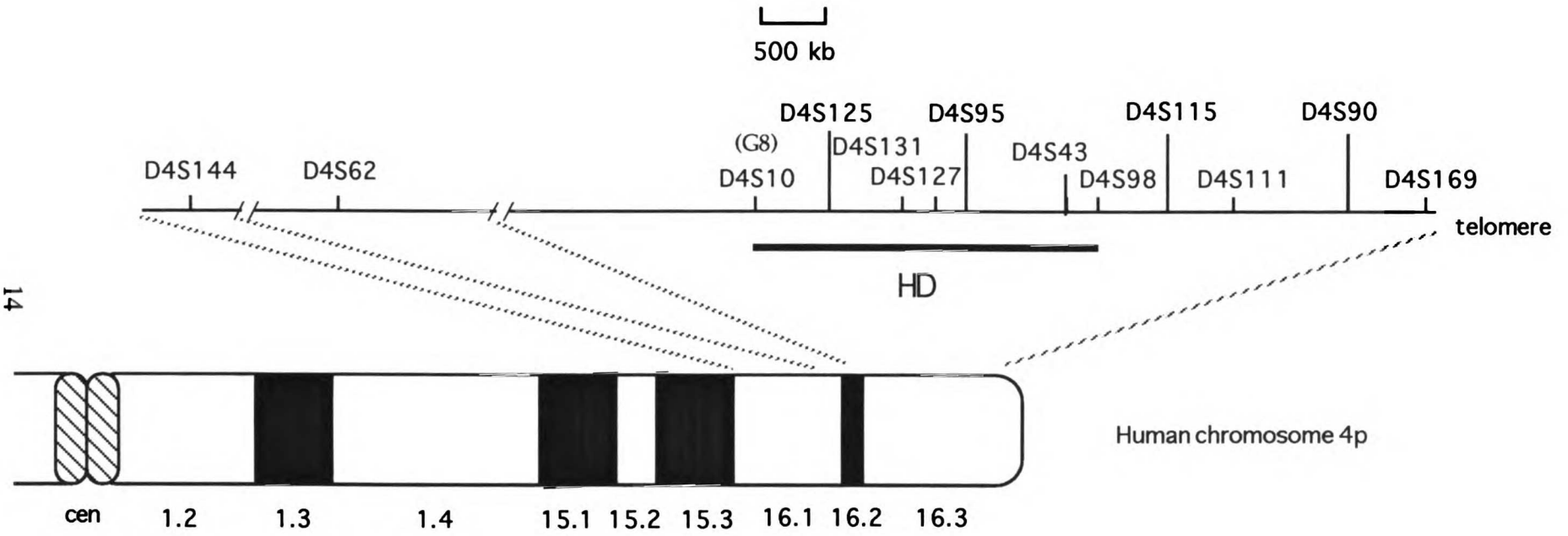
This result demonstrated as one of the first few examples that a disease gene can be identified by such a positional approach. It also illustrated the feasibility of constructing a linkage map of the entire human genome. The hunt for the HD gene has therefore pioneered the positional cloning approach. However it has also confronted almost every problem in this fast growing field.

The initial linkage study used two large HD pedigrees: the Venezualan pedigree and an American family. It was therefore reasonable to ask whether different HD pedigrees with different geographic origins all have the same genetic linkage to the G8 marker, as was expected from the previous studies on the low mutation rate of HD. A total of 63 HD families with wide range of origins were surveyed for the linkage to the G8 marker (Conneally et al., 1989). Six small families gave zero or mildly negative lod scores, while 57 families gave positive lod scores. The combined maximum lod score was 87.69 at a recombination fraction of 4% from G8. These results suggested, but did not prove, that there is a single genetic locus responsible for HD. This study did not eliminate the possibility of multiple closely-linked genetic loci for HD mutations. In spite of these uncertainties, the apparent lack of heterogeneity is consistent with the low mutation rate for HD, and therefore HD is likely to be caused by mutations in a region 4 cM away from the G8 marker.

In the initial linkage report, Gusella et al also determined that the G8 marker is on chromosome 4 by Southern blot analysis of human-rodent somatic cell hybrids (Gusella et al., 1983). Subsequent analyses of the G8 marker indicated it is located near the telomere of the chromosome 4 short arm (4p16.3 region and see Figure 1). This was the consensus of three independent studies: in situ hybridization of G8 to the metaphase chromosome spreads (Landegent et al., 1986; Magenis et al., 1986), Southern blot analysis of somatic cell hybrids with breaks on chromosome 4 (somatic mapping panels and see below) (MacDonald et al., 1987; Smith et al., 1988), and analysis of Wolf-Hirschhorn patients with deletions of the terminal cytogenetic band on the short arm of chromosome 4 (Gusella et al., 1985).

Figure 1: Composite map of 4p16.3. Cytogenetic bands are indicated below the human chromosome 4p. A composite expanded map of 4p16 is drawn above the map of 4p. The proximal positions of eleven markers (see text) from 4p16.3 region are labeled by vertical lines with their D4S numbers on top, while the marker G8 is specifically indicated on top of the D4S10 locus. Two other markers D4S144 and D4S62 from 4p16.1 and 4p16.2 respectively are also included for the reference to other 4p16.3 markers. The horizontal bar below the composite map of 4p16.3 represents the 2.2 Mb HD region between D4S10 and D4S98.

**Figure 1: Composite map of 4p16.3**



A more accurate location of the HD locus was further defined by multi-point linkage analysis (Gilliam et al., 1987b). The only two other available markers on the short arm of chromosome 4 at the time were D4S62, which is located in 4p16.2, and D4S144, which maps to 4p16.1 (see Figure 1). Each marker was analyzed for linkage to the HD locus and then was compared with each other marker and with G8. This analysis established the order of these loci as: cen-D4S144-D4S62-D4S10-HD-tel. Therefore the exact location of the HD locus was about 4 cM distal to G8, towards the telomere.

Refining the location of the HD locus in 4p16.3

The positional cloning strategy involves isolating two closest markers flanking the disease locus after the initial linkage studies, cloning of the entire region between these two flanking markers, and then searching for mutational events that are associated with the disease. This strategy is straight-forward, but requires a combination of expertise in several new technologies, a large amount of labor, and a highly organized effort. Cloning of the cystic fibrosis gene and neurofibromatosis I gene illustrated this notion (Cawthon et al., 1990; Kerem et al., 1989; Riordan et al., 1989; Viskochil et al., 1990).

a) Isolation of DNA probes from the HD region

To identify a flanking marker distal to the HD locus, it was essential to isolate DNA markers from the 4p16.3 region, particularly distal to D4S10 locus (G8). These markers needed to be tested for their polymorphisms in HD pedigrees, so that informative crossover events in HD families can be identified.

Several different approaches have been taken by different groups to achieve this goal. Gilliam et al analyzed 194 random clones from a flow-sorted chromosome 4 specific lambda phage library by hybridizing them to a mapping panel of somatic cell hybrids containing different portions of chromosome 4 (see figure 2) (Gilliam et al., 1987a). One single probe, C4H at D4S43 locus, was identified to map distal to D4S10 locus. Similar studies by Youngman et al resulted another probe, D5 at D4S90 locus, which maps distal to D4S10 (Youngman et al., 1989). To enrich the DNA source for screening, two different somatic hybrid cell lines were used (see figure 2). HHW693 contained human 4p15.1 to 4ptel and a large portion of 5p as its only human sources. Using a phage genomic library constructed from this hybrid, Smith et al screened three hundred phages containing human inserts and found that five of them map to 4p16.3 (Smith et al., 1988). These phages were named D4S95-99. Similarly, a genomic NotI linking library was constructed by Pohl et al from this hybrid (HHW693) and five more clones were mapped to 4p16.3, which were named D4S113, D4S114, D4S111, P107 and 417 (Pohl et al., 1988). Independently, a radiation hybrid cell line, C25, which contained mainly 4p16 and little other human chromosome 4, was therefore highly enriched for the HD region (Cox et al., 1989). This hybrid was used to construct a genomic phage library and seven new probes were identified from the HD region, namely D4S131-136 and D4S169 (Pritchard et al., 1989). In addition, two other probes, JZ1 (D4S380) and JZ6, were later found to be from the 4p16.3 region (Zuo and Myers, unpublished data) (Zuo et al., 1992). The third approach was to construct chromosome jumping libraries and screen with known probes such as G8. This approach resulted in three new probes, D4S81, J102, and D4S112 (Richards et al., 1988). Lastly, a random genomic clone, D4S125, was identified by screening with a consensus probe for a

variable number of tandem repeat sequence and later was mapped to 4p16.3 (Nakamura et al., 1988).

Thus 22 new probes were identified within a few years after the localization of the HD locus to 4p16.3. In subsequent analyses, many of these probes were found to be near each other in clusters, and the distribution of these probes in 4p16.3 was not even (Bates et al., 1991; Zuo et al., 1992). This was perhaps due to the use of human repetitive sequences as probes for identification of genomic clones with human inserts over the hybrid host background. Thus the use of additional approaches, such as mapping random genomic clones by somatic cell hybrid methods, should be able to overcome this problem and more evenly distributed probes should be obtained by a combination of different approaches in future genomic analysis. The effort to identify DNA markers from the 4p16.3 region led to the rapid development of various new technologies that had an impact on studies of other regions of the genome.

b) Physical mapping of the 4p16.3 region

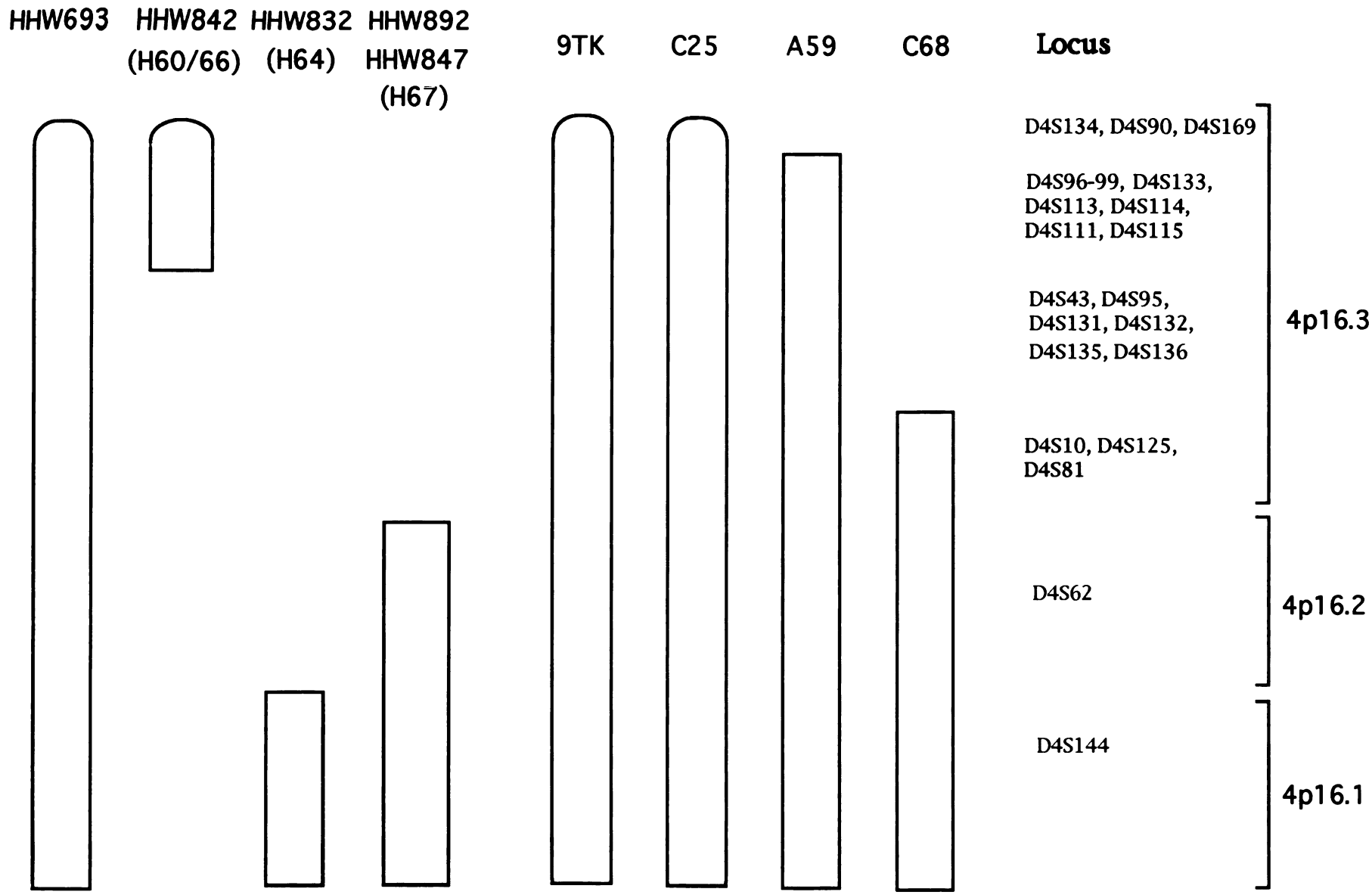
In order to identify DNA markers flanking the HD locus, a physical map of these new markers was needed. Several strategies were used to analyze the genomes of other species and smaller parts of the human genome. Because none of these strategies was definitive, a combination of different methods was the key to construct a consensus map.

(1). Somatic cell hybrid mapping. Different human cell lines that carry various cytogenetic deletions or translocations involving 4p were fused with rodent cells. To obtain hybrid cell lines retaining 4p, different selection strategies were designed. HHW693, for example, was derived from an individual with a balanced translocation between 4p and 5p. It was formed by fusing leukocytes

Figure 2: Mapping panel of somatic cell hybrids and radiation hybrids for 4p16. Human chromosomal portions of 4p16 that were retained in these hybrids are indicated with their names (see text) on top of each hybrid in the map. The 4p telomere is indicated by a half circle. Each subchromsomal interval is defined by the break-points of two different hybrids. DNA markers that are shown to be located within the same interval are listed as their D4S numbers on the side. The proximal intervals of 4p16 cytogenetic bands are indicated as well.

Figure 2

## Somatic Cell Hybrids and Radiation Hybrids



from this patient to a hamster cell line, and then by a positive selection of a temperature-sensitive gene near the 5q centromere and a negative selection of two genes at 5q (Wasmuth et al., 1986). The resulting hybrids were later found to contain different parts of human 4p and therefore were used as a source to map the newly isolated DNA probes from the HD region (see figure 2).

(2). Radiation hybrid mapping. The resolution of somatic cell hybrid mapping was at about 2 Mb. To increase the resolution for mapping, X-rays were used to break human chromosomes into small fragments (Cox et al., 1990). These small fragments were then recovered in rodent cells. The frequency of breakage between two loci was directly related to their physical distance. In contrast to meiotic recombination, there seemed to be no hot spots for breakage in human genome and this method therefore offered a fairly accurate estimate of physical distance between loci, with a resolution of 1 cR or 50 kb. In addition, the loci did not need to be polymorphic as in meiotic mapping. 9TK, for example, was a hamster-human hybrid with the entire chromosome 4 as its sole human DNA source (see figure 2) (Cox et al., 1989). It was exposed to a lethal dose of X-rays, and then fused to a hamster cell line and further selected for HPRT complementation. One line, C25, retained small fragments of human DNA non-selectively, and was found to contain D4S62, D4S10 and many other markers in 4p16.3 markers. Another line, A59, also retained D4S10 and the distal portion of 4p16.3 but missed the most telomeric probes. Hybrid C68 possessed most of the chromosome 4 but carried a terminal deletion distal to D4S10. The break point was between D4S125 and D4S131 loci (Zuo and Myers, unpublished results) (Cox et al., 1989; Pritchard et al., 1989; Zuo et al., 1992). This method was also applied to the distal part of 4p16.3 region by another group (Altherr et al., 1992). Eleven markers were analyzed in this study and the most likely order as well as their

distances were determined. The results from these two studies were fairly consistent (Cox and Myers, unpublished results).

The somatic cell hybrid and radiation hybrid mapping panels were helpful for determining the proximal locations of newly-isolated DNA probes, and these probes, in return, refined the exact maps of retention of these hybrids.

(3). Fluorescence in situ hybridization (FISH). Fluorescent molecules can be deposited in chromatin at the sites of specific DNA sequences by the use of FISH. DNA probes are labeled with reporter molecules and the target chromosomes or nuclei are denatured and then reannealed in the presence of the labeled probes. The reporter molecules are later visualized with fluorescently-labeled affinity reagents. Cosmids can be mapped by FISH with a resolution of >3Mb on metaphase chromosomes (Trask, 1991). However since chromatin is less condensed in interphase than in metaphase, the folding of chromatin in interphase cell nuclei makes it easier than in metaphase to study the distances between pairs of DNA probes by FISH. Pair-wise interphase measurements based on a random walk model were therefore used to determine both distances and relative orders of DNA probes at higher resolution (van den Engh et al., 1992). Thirteen cosmids from the distal portion of 4p16.3 region were analyzed by this method, and they were separated by an average distance of 300 kb with a potential resolution of 100 kb distance. This method provided an independent confirmation of the distance between probes.

(4). Pulsed-field gel electrophoresis (PFGE). To construct a long range restriction map of a large region of a genome, one usually digests genomic DNA with several different rare-cutting restriction enzymes, such as NotI, MluI, NruI, CspI and SplI. These enzymes have eight or six nucleotide recognition sites and cut about every 50-200 kb in genomic DNA. Some of them cut less frequently or less completely due to the methylation of some CpG dinucleotide sequences at

their recognition sites (Bird, 1986). Large fragments of genomic DNA digested with these enzymes can be separated by PFGE. When different probes are hybridized to genomic PFGE blots, the maximum distance between two probes is the size of the DNA fragment that is shared by the two probes.

However, there are several problems with this technique. Sometimes two unrelated fragments fortuitously migrate together in the gels. A combination of single and double digestion with different enzymes is usually necessary to obtain definitive results. Moreover, some regions of human chromosomes, such as subtelomeric regions, are GC rich and therefore rare-cutting restriction site rich. The PFGE mapping effort is thus difficult due to the large number of rare cutting sites.

More than twenty probes were analyzed by PFGE with NotI, MluI and NruI, and almost the entire 4p16.3 region was mapped with two gaps (Bucan et al., 1990). A cluster of these three sites at one end of a segment was later confirmed to represent the end of the chromosome by a human telomere specific probe, analysis of a telomeric yeast artificial chromosome (Bates et al., 1990) and a contig of phage clones at the telomere (Pritchard et al., 1990). In addition, one gap was later closed by digestion with another rare-cutting enzyme SplI (Bates et al., 1991) and by using different DNA probes (Zuo et al., 1992). Although there still remains one gap in the 4p16.3 region, its size is almost negligible by fluorescent in situ hybridization (van den Engh et al., 1992) and by radiation hybrid mapping (Altherr et al., 1992) described above.

c) Identification of recombinant chromosomes in the HD families

With a large number of probes and a long range restriction map of the region, one should first identify informative meiotic crossover events in the 4p16.3 region in the HD families to narrow down the HD region.

Assuming that about 200 HD families have been studied from different parts of the world and the number of meiosis per HD family is about 10, there shall be overall about 120 recombinant events in the 6 cM 4p16.3 region. However the number of published crossover events in HD pedigrees from this region is much less.

Due to the variable onset of age for the disease, many recombinant individuals in HD families are still at risk or likely recombinants, so that they can not be used as informative crossovers. Among the handful of informative recombinant events, three categories were found. Based on published data, more than fifteen recombinant chromosomes with marker-marker recombination placed the disease locus distal to D4S10 (see Figure 1) (Curtis et al., 1989; Holloway et al., 1989; Richards et al., 1988; Skraastad et al., 1989; Wasmuth et al., 1988; Youngman et al., 1989). At least four other crossover events were identified that placed the HD locus proximal to D4S98 (Barron et al., 1991; MacDonald et al., 1989; MacDonald et al., 1991; Snell et al., 1992; Whaley et al., 1991). These first two groups of recombinants were legitimate since they were well characterized HD patients and all had recombination events that were observed between markers in the region. Two of these recombination events define the physical limits or boundaries of the HD region: from D4S10 to D4S98, a segment of 2.2 Mbp in length (see Figure 1) (Bates et al., 1991; Snell et al., 1992). However, three other HD individuals from three different HD families inherited the normal version of the chromosomes from their affected parents, for all the available informative markers throughout the region from D4S10 to D4S169, which is about 60 kb from the 4p telomere (see Figure 1) (Bates et al., 1990; MacDonald et

al., 1989; Pritchard et al., 1990; Robbins et al., 1989; Whaley et al., 1988; Youngman et al., 1992). Under the assumption that there is only a single recombination in each individual, these putative recombinants put the disease locus between D4S169 locus and the telomere. However, this region of chromosome 4 is homologous to those of four other chromosomes, and contains many repetitive sequences (Pritchard et al., 1992; Youngman et al., 1992). It appeares more likely then that either there are double or multiple recombination events in the 4p16.3 region, or the disease locus was out side the 4p16.3 region in these three unusual HD families (Pritchard et al., 1992). Appendix One of this thesis describes our detailed analysis on one of these unusual HD families.

Interestingly, all recombination events in the first group described above, which put the disease locus distal to D4S10 towards 4p telomere, happened within a 500 kb region distal to D4S10 (Bates and Lehrach, 1993). This result indicates that there is a recombination hot spot within the segment between D4S10 and D4S125 (see Figure 1). A hot spot in the same region also existed in the genetic map constructed by studying a large collection of non-HD pedigrees from France (CEPH) (Buetow et al., 1991). Furthermore, there also seemed to be no statistically significant difference between the normal and HD population in the genetic maps in the 4p16.3 region (Buetow et al., 1991; MacDonald et al., 1989).

The search for new informative crossover events that can further narrow the HD region has been fruitless, particularly because of this recombination hot spot.

d) Haplotype and linkage disequilibrium analyses

With few informative recombinant events in the 2.2 Mb HD region, the localization of the disease gene remained a challenge. In general, recombinational mapping rarely provides resolution finer than 1 cM. Linkage disequilibrium mapping, however, is based on the observation that affected chromosomes descended from a common ancestral mutation should show a distinctive haplotype in the immediate vicinity of the gene, reflecting the haplotype of the ancestral chromosome. It therefore offers increased mapping resolution because it exploits recombination events occuring over the entire history of a population. In HD, if the presumed founder mutation occurred in a particular chromosome in early 17th century in Northern Europe, this founder chromosome could then have been transmitted through a large number of meioses to all its offspring around the world who had the disease. The expected number of meioses since the early 17th century up to now should be about ten times higher than that occurred only in the current available HD families. If the founder effect is true, one should find a common small segment within the 2.2 Mb region that is shared by all HD chromosomes. This region, called a common haplotype, would be likely to contain the HD mutation. The presumed founder effect, such as in the case of HD, is the key for this haplotype analysis.

To distinguish different haplotypes within the 2.2 Mbp region, a large number of highly polymorphic markers throughout the region were typed in both HD and normal chromosomes (MacDonald et al., 1992). The results of this study were striking. Among 78 HD chromosomes that were typed, 28 shared a single haplotype around D4S127 and D4S95, a small segment of 500 kb, and the rest of the HD chromosomes had completely different haplotypes around this small segment (see Figure 1). These results suggested that: 1) about one third of the HD chromosomes had the same ancestral origin; 2) many different origins also contributed to the disease; and most importantly, 3) the mutation

responsible for this one third of the HD chromosomes was located within this 500 kb segment.

This observation is also consistent with independent studies on linkage disequilibrium at several other loci (MacDonald et al., 1991; Snell et al., 1989; Theilmann et al., 1989). Since the number of generations after the founder mutation was not large enough, the alleles surrounding the HD mutant did not have enough time to be in complete equilibrium with the general population. Thus one would expect to find at a particular single locus a significant non-random association of a particular allele with the HD allele. This was the basis for single site linkage disequilibrium analysis of the HD mutation. As a result of studies on 97 independent HD families, significant disequilibria at D4S95 and D4S127 with the HD locus were observed, while the disequilibria at D4S43 and other loci were, in general, not as significant as at D4S95 and D4S127 (MacDonald et al., 1991; Snell et al., 1989; Theilmann et al., 1989). Other markers, which displayed random association with HD, were interspersed with markers that displayed non-random association. This lack of a simple peak of linkage disequilibrium in the HD region was initially frustrating, but it could be explained by either high mutation rates at some particular markers, or more than a single founder mutation (MacDonald et al., 1991).

A combination of the linkage disequilibrium data and haplotype analysis would indicate that the HD mutation is within the 500 kb around D4S95 and D4S127 (see Figure 1).

Molecular cloning and analysis of the HD region

a) Cloning the HD region

After the genetic analysis of recombinant HD individuals, which defined the physical limits of the 2.2 Mb HD region, and at about the same time as the haplotype and linkage disequilibrium analyses were conducted, cloning of the 2.2 Mb HD region between D4S10 and D4S98 was pursued by two groups (Bates et al., 1992; Zuo et al., 1992). The attainment of continuous overlapping clones in HD region would not only provide information on the structural complexity of the HD region but also facilitate the isolation of cDNAs, and polymorphic probes, which ultimately helped identify the HD gene.

There were several approaches to dissect and clone this 2.2 Mb region between D4S10 and D4S98. A phage library was constructed from a gel-purified NotI fragment near 4p telomere of a somatic cell hybrid, C25 and this library was used to construct a phage contig of a 300 kb genomic fragment (Pritchard et al., 1990). This approach in general can be very efficient and specific for enrichment of the region of interest. However, the HD region is large and rich in NotI sites, and these NotI sites were methylated differently in peripheral leukocyte DNA from in the somatic cell hybrids. Therefore, to enrich the HD region from the rodent hybrid background, a complete and independent NotI map of the hybrid line needed to be constructed, and a high density of probes were also required.

Alternatively, micro-dissection of a particular region of human chromosome could be both achieved by laser beam and by enzymatic cleavage through triple-helix formation (Hadano et al., 1991; Strobel et al., 1991). Although these techniques offered some promises in genome analysis, very few labs were able to perform these types of analysis at the time.

Thirdly, different genomic libraries had been constructed from both total human genome and specific chromosomes. Chromosomal walking in cosmids, for example, was widely used to clone 100 to 200 kb genomic DNA. But this approach also depended on whether the distribution of probes was high and even, and depended on

whether the genomic region was clonable in cosmid vectors. In the HD region, there were several large portions lacking probes (Bates et al., 1991; Bucan et al., 1990; Zuo et al., 1992). The speed of walking would therefore be extremely slow.

Finally, the recently developed yeast artificial chromosome (YAC) cloning system allows the cloning of large fragments of genomic DNA from 100 kb to at least 1 Mb (Burke et al., 1987). This YAC mapping approach was proven to be successful in cloning the Cystic Fibrosis and Duchenne Muscular Dystrophy regions, where probe densities were high (Green & Olson, 1990; Monaco et al., 1992). In Chapter One of this thesis, this YAC cloning approach will be described in greater detail. Our YAC contig of the HD region had covered almost the entire HD region (Zuo et al., 1992), and was consistent with the results reported by another group (Bates et al., 1992).

The YAC approach has been recently used successfully in cloning entire human chromosomes and is one of the main goals of the human genome project for the next few years (Foote et al., 1992; Schlessinger et al., 1991). However, several disadvantages were found in this cloning system. First, more than half of all the YAC clones that have been analyzed contain non-contiguous segments of genomic DNA or chimeric inserts. This appeared to be the result of homologous recombination in yeast cells during library construction or propagation (Green et al., 1991). Second, some YACs also contained additional small rearrangements, deletions and inversions, that are not easily detected (Bates et al., 1992; Zuo et al., 1992). Third, it is difficult to isolate YAC DNA from the background of yeast chromosomes in a yeast strain, because the copy number of YAC is about one per cell (Burke et al., 1987).

These problems with the YAC cloning system have hampered the complete characterization of the HD region. However, high resolution genome mapping could be facilitated by other cloning systems with small, more stable inserts and manageable DNA yields. Cosmids and P1 clones are such examples. To utilize the YAC mapping information that had been generated through the collaborative effort of the human

genome project, we described a strategy, in Chapter Two of this thesis, to construct a high resolution map of the HD region. The complete characterization of the HD region facilitated identification of the HD mutation and in helping to determine the genomic structure of the HD gene, and our map was also consistent with that reported by another group (Baxendale et al., 1993; THDCRG, 1993; Zuo et al., 1993).

b) Identification of genes and mutations in the HD region.

Identifying genesin cloned genomic DNA remains a formidable challenge to positional cloning efforts, mainly because the following three reasons. Only 2-4% of human genomic sequences represent exons. A large number of repetitive sequences, such as Alu (SINE) and Kpn (LINE), are present in most genomic clones. In addition, not every gene is expressed in any particular cDNA library due to temporal and spatial regulation.

Several approaches have been employed to identify genes in cloned genomic DNA. First, direct screening of cDNA libraries with cosmid and YAC DNA has been used to isolate cDNAs from specific genomic regions, by extensively pre-annealing the labeled DNA probes with an excess amount of unlabeled human placental DNA to minimize the interference of repeats in the probes (Elvin et al., 1990; Wallace et al., 1990). This approach suffers from poor signal intensity. Large numbers of both false positive and false negative signals are often observed.

Secondly, many CpG base pairs in the human genome are associated with genes at their 5' ends (Bird, 1986). These CpG islands are indicators for the locations of the genes. In genomic DNA, unmethylated CpG rich sites are often associated with expressed genes, while methylated ones are often not expressed. Therefore, the identification of clusters of rare-cutting, CpG rich restriction sites, such as NotI, BssHII, EagI and SacII, can be used to obtain potential exon probes to increase the efficiency for

screening cDNA libraries. However, not all genes are associated with CpG islands, and this approach requires efficient ways to identify the islands, such as subcloning cosmid or YAC clones.

A third approach to isolate genes from cloned genomic DNA is to use "zoo blots", which identify genomic cloned DNA probes that are conserved across different species, and these probes can be used as indicators for exons. This approach is labor intensive.

Screening heteronuclear RNA derived from human and rodent hybrid cell lines is another approach to isolate genes from genomic DNA. It utilizes the enriched human sources of hybrid lines, and can be used to obtain genes from entire chromosomes or subchromosome regions present in the hybrids (Corbo et al., 1990; Liu et al., 1989). However, in general only constitutively expressed genes could be isolated from these cell lines.

Exon trapping is another technique to identify genes. This technique utilizes consensus sequences for exon and intron junctions to identify exons in genomic clones (Buckler et al., 1991; Duyk et al., 1990). If a genomic clone contains a splice donor site, a splicing event will happen between this donor site and an acceptor site designed in the detecting vector. Several different detection systems were designed and they all seem to be successful. Splicing events can be rescued and screened for from genomic cosmid clones by either colony colors or decreases in sizes of PCR products. However, this approach fails to identify intronless genes and spliced exons that have different splicing junctions. In addition, this approach also suffers from cryptic splicing events.

Most recently, YACs and cosmids were immobilized on blots and cDNA was hybridized to, and cDNA that didn't hybridized to YAC and cosmid DNA was then washed away. Therefore cDNA from the specific region, where YAC and cosmid clones lie, can be enriched (Lovett et al., 1991). The enrichment can be further improved by hybridizing biotinylated genomic DNA to cDNA probes in solution and subsequently capturing the hybridized complexes using streptavidin-coated magnetic beads (Tagle et

al., 1993). This approach has advantages of screening several hundred thousand base pairs at once and obtaining low abundant messages.

It's also important to know how many genes are in the 2.2 Mb HD region. The number of genes in HD region can be estimated based on the gene densities of other known genomic regions. For example, a minimum of seven genes in MHC locus of 170 kb genomic DNA on mouse chromosome 17 were found, and 23 gene sequences from polycystic kidney disease locus of 300 kb region on human chromosome 16 were also reported (Abe et al., 1988). Considering the large number of CpG islands in the HD region, one would expect a large number of genes in this 2.2 Mb region. By analyzing six random cosmid clones from the 4p16.3 region, it was estimated that the number of genes from the 2.2 Mbp HD region was about 100 (Carlock et al., 1992). This estimate is also consistent with other estimates from a region near 4p telomere outside the 2.2 Mbp HD region (Weber et al., 1991).

From the 2.2 Mbp HD region, several genes have been identified by various methods. A fibroblast growth factor receptor was identified near D4S98 locus (see Figure 1) (Thompson et al., 1991); a-adducin and a G-protein coupled receptor kinase have been found near D4S95 and D4S127 loci (Ambrose et al., 1992; Goldberg et al., 1992; Taylor et al., 1992). In addition, two small transcripts from D4S43 were also identified (Gilliam et al., 1987a). And most recently, the putative HD gene was also identified proximal to D4S127 (THDCRG, 1993).

Mutation analysis on genes that have been identified is very difficult but is achievable. Mutations that involve small rearrangements, such as deletions, insertions and inversions, can be screened for by using cloned cosmid DNA on genomic Southern blots from different HD patients. For point mutations, PCR primers from cloned DNA sequences can be designed and used for screening genomic DNA from different patients. Denaturing gradient gel electrophoresis can be used for detection of mismatches in PCR products from each individuals (Myers et al., 1987). GC clamped

DGGE is a modified version of the detection system to increase the chance of finding mutations (Sheffield et al., 1992; Sheffield et al., 1989), and genomic DGGE was designed for using different probes on a single gel (Burmeister et al., 1991). Single-strand conformation polymorphism (SSCP) is another method to detect mutations (Orita et al., 1989). Although every method has its own disadvantages, it should be in theory possible to detect all mutations in a specific genomic clone.

A straight-forward way to identify mutations is to sequence all the genomic region of interest and compare the sequence differences among different patients. Sequencing a 2.2 Mb region is feasible, although it is costly and slow (Sulston, J. et al.,1992). New technologies are emerging for fast sequencing and for comparing sequence differences between two individuals (Lisitsyn et al., 1993; Nelson et al., 1993).

In summary, finding the HD mutation in the 2.2 Mb region was like looking for a needle in a haystack. Since the initial linkage study in 1983, the hunt for the HD gene was hampered by the lack of a cytogenetic abnormality, informative recombinants, and a simple peak of linkage disequilibrium to pinpoint the exact location of the mutation.

## Identification of the putative Huntington disease gene

While our work in Chapter Two of this thesis was under review for publication, two papers were published related to the identification of the putative HD gene (Goldberg et al., 1993; THDCRG, 1993).

### Identification of the putative HD mutation

A (CAG)n repeat was found in the proximal vicinity of the D4S127 locus (see Figure 1), where strong linkage disequalibrium was observed. This triplet

repeat is expanded over 42 copies in the 73 HD chromosomes that have been examined, while 173 normal chromosomes have less than 34 copies of the repeat. In addition, the copy number of the triplet repeat seems to be inversely correlated with the age of onset of the disease, especially in the juvenile HD cases. Finally, two "sporadic" HD patients without family history of HD show a moderate expansion of the repeat into the abnormal range while their normal siblings have normal range of copy numbers of the repeat. These three lines of evidence strongly support the notion that the expansion of the triplet repeat correlates with or even causes the disease.

One large transcript (IT15) from lymphoblast cell lines, about 11 kb in size, was found containing the CAG repeat. This transcript has an open reading frame with two potential translation start sites. The CAG repeat is located near the 5′ end (either translated or non-translated region) of the gene. The predicted size of the translated product is about 348 kD. It is ubiquitously expressed in all tissues examined. Interestingly its message was also detected in homozygous HD patients.

In another paper by Goldberg et al, an Alu retrotransposition was observed in an intron of the a-adducin gene at the distal vicinity of the D4S95 locus (see Figure 1), where a strong linkage disequilibrium was observed. This Alu insertion, which is located about 200 kb distal to the CAG repeat in the gene described by THDCRG, was found only in HD patients of two ancestrally related families, but not in any other HD patients nor in any of the 1,000 normal individuals tested. This observation indicated that the Alu retrotransposition event probably had nothing to do with the disease in these patients.

Several questions need to be further addressed. First, the distribution of the repeat in normal chromosomes has not yet been determined. Particularly, in the 173 normal chromosome pool, there are nine normal chromosomes that had the

same core haplotype as the one third of the 74 HD chromosomes. It would be interesting to know if these nine normal chromosomes all had expanded CAG repeats close to the abnormal range.

Secondly, the two affected individuals in the unusual family 217 that we studied in Appendix One of this thesis, had normal numbers of the CAG repeat, while their affected parent and another affected sib showed abnormal numbers of the repeat. These two inconsistent individuals might have been misdiagnosed; or their brain tissues might have the expanded versions of the repeat, while their lymphoblastoid cells had normal numbers of the repeat due to mitotic instability of the repeat.

Thirdly, until a point mutation is found in a sporadic HD patient without expansion of the triplet repeat, or a mouse model is created by homologous recombination or gene knock out, it still remains possible that the CAG repeat expansion is the consequence but not the cause of the disease.

Considering the possible effect of the unstable triplet repeat on the genomic elements in its vicinity, it is also possible that other genes in its vicinity could contribute to the disease.

Functions of the putative HD gene

These results raised several questions concerning the functions of the putative HD gene and the mechanisms of the HD mutation. Since another cDNA probe (GT149), isolated by Goldberg et al from the same region as IT15, detected two bands, 10 kb and 12 kb in size, in Northern blots containing RNA from different tissues. It is thus conceivable that the putative HD gene is alternatively spliced in different tissues. The putative HD gene has no significant homology to any published sequences in GenBank and other databases. But if the CAG repeat

is in the coding region of the gene, it shall encode a poly-glutamine stretch that resembles some transcription activators, which were initially detected as Opa repeats in Drosophila (Johnson & McKnight, 1989; Peterson et al., 1990; Wharton et al., 1985). In addition, there is one half of the Leucine-zipper motif in the middle of the gene, indicating that the putative HD gene may encode a DNA binding protein. Furthermore, in the view of the ubiquitous expression of the putative HD gene, it is intriguing why specific neurons in the brain die first.

The presence of the IT15 message in the homozygous HD individuals does not indicate the presence of the mature protein. Therefore it still remains to be elucidated what the mode of action of the HD mutation is. In addition, the presence of the message in the HD homozygotes is also consistent with the possibility that the putative HD gene is not the real but still elusive HD gene.

It is known that the juvenile onset HD patients mostly inherite their disease alleles from their affected fathers rather than from their affected mothers. If the CAG repeat expansion is the cause of the disease, there then may be some difference between male meiosis and female meiosis, resulting in the large expansion of the triplet repeat. Alternatively, since the putative HD gene is expressed ubiquitously, there may be some gametic selection against the large expansion of the repeat.

Considering the low mutation rate of HD, it is also interesting that the triplet expansion appears to occur only in a selected group of chromosomes with particular haplotypes. This is consistent with the presumed founder effect.

Common features of the unstable trinucleotide repeats in four genetic diseases

The expansion of the CAG repeat in the HD chromosomes has significant similarity to three other genetic diseases, namely Kennedy's disease (KD), Fragile

X syndrome (FAX), and Myotonic Dystrophy (MD). These three inherited diseases are known for the expansion of trinucleotide repeats in the affected individuals and have been extensively studied (Caskey et al., 1992; Richards & Sutherland, 1992). Comparison between HD and these three diseases will greatly facilitate the understanding of the mechanism of the HD mutation and their common features are summarized in Table 1.

First, the patterns of inheritance of these four different diseases are different. KD and FAX are both X-linked recessive, while MD and HD are both autosomal dominant.

The gene for KD is an androgen receptor (AR), where a CAG repeat is found to be expanded in patients at the 5′ end of the coding region (La et al., 1991). In FAX, a CGG repeat is found to be expanded in affected patients at the 5′ non-coding region of the FMR-1 gene (Kremer et al., 1991; Verkerk et al., 1991). In MD, a GCT repeat is found to be expanded in the patients at the 3′ untranslated region of the MD gene (Brook et al., 1992; Fu et al., 1992; Mahadevan et al., 1992). Whether the CAG repeat is translated in the putative HD gene is not certain. In addition, a GGN repeat is also found in the AR, while a CCG repeat is found in the putative HD gene just down stream of the CAG repeat. Although these repeats have not been studied extensively, their locations in the vicinity of the unstable CAG repeat are intriguing.

The normal ranges of copy numbers of these repeats in these genes are between 10 to 60, such that the total length of the repeat region is less than 200 bp in unaffected individuals. In both FAX and MD, there are constant pools of apparently normal chromosomes with copy numbers of repeats between 60 and 100 that have extremely high chances of expanding into the abnormal range of over 100 (Caskey et al., 1992; Yu et al., 1991). These pools of chromosomes are thought to be premutated. In HD case, there are about Therefore it is interesting

**Table 1: Comparison of four diseases with triplet expansion**

| | Huntington Disease | Myotonic Dystrophy | Kennedy's Disease | Fragile X Syndrome |
|---|---|---|---|---|
| Inheritance | autosomal dominant | autosomal dominant | X-linked recessive | X-linked recessive |
| Gene/ Function | DNA binding? | Protein kinase? | Androgen receptor | Unknown |
| Repeat/ Location | CAG/ 5'-Coding? | CAG/3'UTR | CAG/ 5'-coding | CGG/5'UTR |
| Size of repeat | 11-40 | 20-35 | ? | 38-50 |
| Sizes of disease alleles | 40-100 | 50-100 (premutated); 100-2,000 (affected). | 40-62 | 52-200 (premutated); 200-2,000 (affected). |
| Expression | Ubiquitous | Ubiquitous; High in muscle | Ubiquitous; High in motor neurons | Ubiquitous; High in brain and testes |
| Founder effect/Linkage disequilibrium | Yes | Strong | Unknown? | Yes |
| Parental sex bias | More paternal for early onset | More maternal for expansion | More paternal instability | More maternal for expansion |
| Mode of mutation | Abnormal protein? Gain-of-function? | Decreased message? /protein? | Abnormal protein. Loss-of-function? | Undetectable message. Methylation? |
| Anticipation | 6% (juvenile) severity/ onset | Strong | Variable severity | Strong |

to see if those normal chromosomes with the most common haplotype as one third of the HD chromosomes are also premutated to have higher copy number of the CAG repeats than the normal population. The apparent similarity of the distribution of the CAG repeats between KD and HD argued that the the CAG repeat might be in the coding region of the putative HD gene due to the selection on the number of glutamine residues in the mature protein.

Both expansion and reduction of the triplet repeats have been observed in all these diseases (O'Hoy et al., 1993). Except in KD, the triplet repeats in the other three diseases are able to expand dramatically in either premutated or affected chromosomes during meiosis. The mild expansion seen in the case of KD may indicate early prenatal lethality in the case of large expansion in its coding sequence. There may be some differences in the male and female meiosis on the large expansion of the triplet repeats, but the differences can be the result of gametic selection in those premutated or affected individuals in one particular gender. Therefore the preferential paternal inheritance in juvenile HD cases may or may not result from the differences between male and female meioses.

The mutations in these four diseases resulted in various changes in the expression of the corresponding genes. The AR protein activity is not detectable in affected homozygous individuals, probably resulting from inactivation of AR or the loss of its function. In FAX patients, the FMR-1 transcript is undetectable probably due to the CGG expansion and methylation of the CpG island at the 5′ end of the gene (Pieretti et al., 1991). The level of transcription of the MD gene in MD patients may be decreased (Caskey et al., 1992). IT15 was detected in both heterozygous and homozygous HD patients. From these analyses, it is not known how the triplet repeat expansion in four different diseases causes different patterns of gene expression. The locations of these repeats in the gene structure may provide some clues. For example, the CGG repeat in the FMR-1

gene is located at the 5′ untranslated region and therefore is expected to have some effect on the level of FMR-1 gene expression. Indeed, the message of FMR-1 is not detectable in the affected patients.

It is intriguing that these four diseases all develop at certain stages of development, although the corresponding genes are constantly expressed, and why specific cell types were affected while the genes are ubiquitously expressed. The level of expression of the normal gene products in different tissues may explain why some tissues are mostly affected due to the most abundant level of expression. In KD, for example, AR is expressed mostly in spinal and bulbar motor neurons and therefore the patients develop progressive motor weakness and atrophy. The highest expression of FMR-1 in brain and testis correlates well with head enlargement and macroorchidism phenotype in patients. It is also true in MD patients whose skeletal and cardiac muscles are functionally disrupted most severely where the expression of MD gene is high. It is thus of interest to determine the level of IT15 in different tissues and particularly different regions of the brain. Tissue specificity in these diseases may have nothing to do with the trinucleotide repeat instability but may be a general phenomenon.

The normal functions of these four genes may provide some clues to their mutant phenotypes when they are mutated. The role of MD gene is thought to be involved in a highly specialized signal transduction pathway, because it contains a ATP-binding site and a catalytic domain for protein kinase. Protein kinases often consist of non identical subunits; an alteration in the level of one subunit can affect the overall function, leading to a dominant mutant phenotype. AR has a DNA binding domain and a steroid binding domain. The mutation in AR may alter the DNA or hormone binding properties, resulting in a loss of its function and leading to a recessive phenotype. FMR-1 is largely a novel gene except a putative nuclear localizing signal. The putative HD gene shows no

homology to any genes in databases except some features of DNA binding proteins and it may normally function as a multimer. The alteration of the DNA binding activity can therefore result in a gain-of-function of its transcription activity.

Possible mechanisms of unstable trinucleotide repeats in genetic diseases

How these triplet repeats expand or reduce in size and how their expansions cause the disease phenotypes are largely unknown.

One common feature among these genetic diseases is that they are all thought to have founder effects. This feature indicates that there may be a small number of chromosomes with particular haplotypes that are prone to expansion of the triplet repeats. In other words, these founder chromosomes may have been premutated and predisposed to this trinucleotide repeat instability. The critical question is: what is the difference between these founder chromosomes and other normal chromosomes? The relatively high copy number of these repeats is certainly a difference, but may not be the only one or may be the consequence of other common features among these founder chromosomes. Several appealing mechanistic models can be drawn based on this founder effect (Caskey et al., 1992; Richards and Sutherland, 1992).

First, DNA polymerase slippage during replication could explain some unstable features of these repeats, but it is not easy to explain the founder effect by this model. In another word, the slippage model can not easily explain why only certain chromosomes are subject to slippage. If a stochastic process by polymerase somehow determines that only a certain pool of founder chromosomes are subject to slippage, then one will expect to have constant pools of predisposed chromosomes to be expanded at the repeat loci. This prediction

seems to be true in the cases of MD and FAX, but not obvious in the cases of HD and KD.

Alternatively, there can be a mutation in ancient chromosomes causing a change in nucleosome structure where the repeats are located. The premutation can simply be a point mutation, for example, so that it creates a larger perfect triplet repeat by combining two smaller perfect repeats. This perfect repeat can therefore be much longer and thus unstable due to the nucleosomal structure and subject to expansion or reduction. The sizes of the normal triplet repeats in these four diseases are between 150 to 200 bps, which are roughly coincident with the sizes of the nucleosomes, while the abnormal ranges are over 200 bps.

Finally, assuming there is an ancient mutation in the vicinity of a triplet repeat only on these founder chromosomes, and this mutation happens to create a perfect DNA protein binding site for some DNA binding proteins, these proteins can accidentally create some expansion or reduction of the triplet repeat in its vicinity. This model can explain the low mutation rates for these diseases and the resulting unstable triplet repeats in the disease chromosomes. It takes three rare events for the diseases to develop: first, the premutations happens to create the DNA binding sites; second, the trinucleotide repeats happens to be in the vicinity of the binding sites; third, the locations of these trinucleotide repeats happens to be in functional genes so that the expansion or reduction of the repeats will manifest in their functions, leading to the disease phenotypes. If this model is correct, one will expect that combination of screening for the DNA binding sites for these proteins, trinucleotide repeats and functional genes will identify more genetic diseases with similar founder effects.

The first and the third models proposed here may have some common features. A polymerase slippage may be the result of some point mutations in the vincinity of the repeats. Or the proposed DNA binding protein may well be

the polymerase. In the view of the fact that triplet repeats are not as unstable in mouse as in human, there may be some differences in how polymerases work in human and mouse, giving some clues for the mechanism of the triplet expansion.

In conclusion, the identification of the CAG repeat expansion in HD had marked the end of a 10 year search for the causative mutation, but opened a new era devoted to understanding the relation between expansion within a house-keeping gene of unknown function and programmed cell death of selected neurons. Luck had certainly played an important role in the initial search for the trinucleotide repeat expansion from the genomic and cDNA clones in the HD region. How many other diseases can be caused by the similar mutation is still under speculation.

## References

Abe, K., Wei, J.F., Wei, F.S., Hsu, Y.C., Uehara, H., Artzt, K. & Bennett, D. (1988) Searching for coding sequences in the mammalian genome: the H-2K region of the mouse MHC is replete with genes expressed in embryos. . Embo J 7: 3441-9.

Altherr, M.R., Plummer, S., Bates, G., MacDonald, M., Taylor, S., Lehrach, H., Frischauf, A.M., Gusella, J.F., Boehnke, M. & Wasmuth, J.J. (1992) Radiation hybrid map spanning the Huntington disease gene region of chromosome 4. . Genomics 13: 1040-6.

Ambrose, C., James, M. et al. (1992) A novel G protein-coupled receptor kinase gene cloned from 4p16.3. Hum.Mol.Genet. 1: 697-704.

Barron, L., Curtis, A., Shrimpton, A.E., Holloway, S., May, H., Snell, R.G. et al. (1991) Linkage disequilibrium and recombination make a telomeric site for the Huntington's disease gene unlikely. J. Med. Genet. 28: 520-522.

Bates, G. & Lehrach, H. (1993) The Huntington's disease gene-still a needle in a haystack? Hum.Mol.Genet. 2: 343-347.

Bates, G.P., MacDonald, M.E., Baxendale, S., Sedlacek, Z., Youngman, S., Romano, D., Whaley, W.L., Allitto, B.A., Poustka, A. & Gusella, J.F. (1990) A yeast artificial chromosome telomere clone spanning a possible location of the Huntington disease gene. . Am J Hum Genet 46: 762-75.

Bates, G.P., MacDonald, M.E., Baxendale, S., Youngman, S., Lin, C., Whaley, W.L., Wasmuth, J.J., Gusella, J.F. & Lehrach, H. (1991) Defined physical limits of the Huntington disease gene candidate region. . Am J Hum Genet 49: 7-16.

Bates, G.P., Valdes, J., Hummerich, H., Baxendale, S., Le Paslier, D.L., Monaco, A.P., Tagle, D., MacDonald, M.E., Altherr, M., Ross, M., Brownstein, B.H., Bentley, D., Wasmuth, J.J., Gusella, J.F., Cohen, D., Collins, F. & Lehrach, H. (1992) Characterization of a yeast artificial chromosome contig spanning the Huntington's disease gene candidate region. Nature genetics 1: 180-187.

Baxendale, S., MacDonald, M.E., Mott, R., Francis, F., Lin, C., Kirby, S.F., James, M., Zehetner, G., Hummerich, H., Valdes, J., Collins, F.C., Deaven, L.J., Gusella, J.F., Lehrach, H. & Bates, G.P. (1993) A cosmid contig and high resolution restriction map of the 2 megabase region containing the Huntington's disease gene. Nature Genet. 4: 181-186.

Bird, A.P. (1986) CpG-islands and the function of DNA methylation. Nature 321: 209-213.

Bostein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am.J.Hum.Genet. 32: 314-331.

Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, J.P. & Hudson, T. (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member [published erratum appears in Cell 1992 Apr 17;69(2):385] . Cell 68: 799-808.

Bucan, M., Zimmer, M., Whaley, W.L., Poustka, A., Youngman, S., Allitto, B.A., Ormondroyd, E., Smith, B., Pohl, T.M. & MacDonald, M. (1990) Physical maps of 4p16.3, the area expected to contain the Huntington disease mutation. . Genomics 6: 1-15.

Buckler, A.J., Chang, D.D., Graw, S.L., Brook, J.D., Haber, D.A., Sharp, P.A. & Housman, D.E. (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc.Natl.Acad.Sci.USA 88: 4005-4009.

Buetow, K.H., Shiang, R., Yang, P., Nakamura, Y., Lathrop, G.M., White, R., Wasmuth, J.J., Wood, S., Berdahl, L.D. & Leysens, N.J. (1991) A detailed multipoint map of human chromosome 4 provides evidence for linkage

heterogeneity and position-specific recombination rates. . Am J Hum Genet 48: 911-25.

Burke, D.T., Carle, G.F. & Olson, M.V. (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. . Science 236: 806-12.

Burmeister, M., diSibio, G., Cox, D.R. & Myers, R.M. (1991) Identification of polymorphisms by genomic denaturing gradient gel electrophoresis: application to the proximal region of human chromosome 21. . Nucleic Acids Res 19: 1475-81.

Carlock, L., Wisniewski, D., Lorincz, M., Pandrangi, A. & Vo, T. (1992) An estimate of the number of genes in the Huntington disease gene region and the identification of 13 transcripts in the 4p16.3 segment. Genomics 13: 1108-1118.

Caskey, C.T., Pizzuti, A., Fu, Y., Fenwick Jr., R.G. & Nelson, D.L. (1992) Triplet repeat mutations in human disease. Science 256: 784-789.

Cawthon, R.M., Weiss, R., Xu, G.F., Viskochil, D., Culver, M., Stevens, J., Robertson, M., Dunn, D., Gesteland, R. & O'Connell, P. (1990) A major segment of the neurofibromatosis type 1 gene: cDNA sequence, genomic structure, and point mutations [published erratum appears in Cell 1990 Aug 10;62(3):following 608] . Cell 62: 193-201.

Clarke, P.G.H. (1990) Developmental cell death: morphological diversity and multiple mechanisms. Anat. Embryol. 181: 195-213.

Conneally, P.M. (1984) Huntington's disease: genetics and epidemiology. Am.J.Hum.Genet. 36: 506-526.

Conneally, P.M., Haines, J.L., Tanzi, R.E., Wexler, N.S., Penchaszadeh, G.K., Harper, P.S., Folstein, S.E., Cassiman, J.J., Myers, R.H. & Young, A.B. (1989) Huntington disease: no evidence for locus heterogeneity. . Genomics 5: 304-8.

Corbo, L., Maley, J.A., Nelson, D.L. & Caskey, C.T. (1990) Direct cloning of human transcripts with HnRNA from hybrid cell lines. . Science 249: 652-5.

Cox, D.R., Burmeister, M., Price, E.R., Kim, S. & Myers, R.M. (1990) Radiation Hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science 250: 245-250.

Cox, D.R., Pritchard, C.A., Uglum, E., Casher, D., Kobori, J. & Myers, R.M. (1989) Segregation of the Huntington disease region of human chromosome 4 in a somatic cell hybrid. Genomics 4: 397-407.

Curtis, A., Millan, F., Holloway, S., Mennie, M., Crosbie, A., Raeburn, J.A. & Brock, D.J. (1989) Presymptomatic testing for Huntington's disease. A case complicated by recombination within the D4S10 locus. . Hum Genet 81: 188-90.

Duyk, G.M., Kim, S., Myers, R.M. & Cox, D.R. (1990) Exon trapping: a genetic screening to identify candidate transcribed sequences in cloned mammalian genomic DNA. Proc.Natl.Acad.Sci.USA 87: 8995-8999.

Elvin, P., Slynn, G., Black, D., Graham, A., Butler, R., Riley, J., Anand, R. & Markham, A.F. (1990) Isolation of cDNA clones using yeast artificial chromosome probes. Nucleic Acids Res 18: 3913-7.

Folstein, S.E. (1989) Huntington's disease: a disorder of families. Johns Hopkins University Press, Baltimore.

Foote, S., Vollrath, D., Hilton, A. & Page, D.C. (1992) The human Y chromosome: overlapping DNA clones spanning the euchromatic region. Science 258: 60-6.

Froster-Iskenius, U.G., Hayden, M.R., Wang, H.S., Kalousek, D.K., Horsman, D., Pfeiffer, R.A., Schottky, A. & Schwinger, E. (1986) A family with Huntington disease and reciprocal translocation 4;5. Am. J. Human Genet. 38: 759-767.

Fu, Y.H., Pizzuti, A., Fenwick, R.J., King, J., Rajnarayan, S., Dunne, P.W., Dubel, J., Nasser, G.A., Ashizawa, T. & de, J.P. (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. . Science 255: 1256-8.

Gilliam, T.C., Bucan, M., MacDonald, M.E., Zimmer, M., Haines, J.L., Cheng, S.V., Pohl, T.M., Meyers, R.H., Whaley, W.L. & Allitto, B.A. (1987a) A DNA segment encoding two genes very tightly linked to Huntington's disease. . Science 238: 950-2.

Gilliam, T.C., Tanzi, R.E., Haines, J.L., Bonner, T.I., Faryniarz, A.G., Hobbs, W.J., MacDonald, M.E., Cheng, S.V., Folstein, S.E. & Conneally, P.M. (1987b) Localization of the Huntington's disease gene to a small segment of chromosome 4 flanked by D4S10 and the telomere. . Cell 50: 565-71.

Goate, A., Chartier, H.M., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N. & James, L. (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease [see comments]. Nature 349: 704-6.

Goldberg, Y.P., Lin, B.Y., Andrew, S.E. et al. (1992) Cloning and mapping of the a-adducin gene close to D4S95 and assessment of its relationship to Huntington disease. Hum.Mol.Genet. 1: 669-676.

Goldberg, Y.P., Rommens, J.M., Andrew, S.E., Hutchinson, G.B., Lin, B., Theilmann, J., Graham, R., Glaves, M.L., Starr, E., McDonald, H. et al. (1993) Identification of an Alu retrotransposition event in close proximity to a strong candidate gene for Huntington's disease. . Nature 362: 370-3.

Graveland, G.A., Williams, R.S. & DiFiglia, M. (1985) Evidence for degenerative and regenerative changes in neostriatal spiny neurons in Huntington's disease. Science 227: 770-773.

Green, E.D. & Olson, M.V. (1990) Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping. . Science 250: 94-8.

Green, E.D., Riethman, H.C., Dutchik, J.E. & Olson, M.V. (1991) Detection and characterization of chimeric yeast artificial chromosome clones. Genomics 11: 658-669.

Gusella, J.F., Tanzi, R.E., Anderson, M.A. & et al. (1985) Deletion of Huntington's disease-linked G8 (D4S10) locus in the Wolf Hirschhorn syndrome. Nature 318: 75-76.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E. & et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234-8.

Hadano, S., Watanabe, M., Yokoi, H., Kogi, M., Kondo, I., Tsuchiya, H., Kanazawa, I., Wakasa, K. & Ikeda, J.E. (1991) Laser microdissection and single unique primer PCR allow generation of regional chromosome DNA clones from a single human chromosome. . Genomics 11: 364-73.

Harper, P.S. (1991) Huntington's disease. Saunders, London.

Hayden, M.R. (1981) Huntington's chorea. Springer-Verlag, Berlin.

Hayden, M.R., Hewitt, J., Wasmuth, J.J., Kastelein, J.J., Langlois, S., Conneally, M., Haines, J., Smith, B., Hilbert, C. & Allard, D. (1988) A polymorphic DNA marker that represents a conserved expressed sequence in the region of the Huntington disease gene. . Am J Hum Genet 42: 125-31.

Herskowitz, I. (1987) Functional inactivation of genes by dominant negative mutations. . Nature 329: 219-22.

Holloway, S., Millan, F.A., Curtis, A., Mennie, M. & Brock, D.J. (1989) Genetic linkage between Huntington's disease and D4S10 (G8) in Scottish families. . Clin Genet 35: 133-8.

Huntington, G. (1872) On chorea. Med.Surg.Rep. 26: 317-321.

Johnson, P.F. & McKnight, S.L. (1989) Eukaryotic transcriptional regulatory proteins. . Annu Rev Biochem 58: 799-839.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. & Tsui, L.C. (1989) Identification of the cystic fibrosis gene: genetic analysis. . Science 245: 1073-80.

Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S.T., Schlessinger, D., Sutherland, G.R. & Richards, R.I. (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. . Science 252: 1711-4.

La, S.A., Wilson, E.M., Lubahn, D.B., Harding, A.E. & Fischbeck, K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. . Nature 352: 77-9.

Landegent, J.E., Jansen, in, de, Wal, N., Fisser, G.Y., Bakker, E., van, der, Ploeg, M & Pearson, P.L. (1986) Fine mapping of the Huntington disease linked D4S10 locus by non-radioactive in situ hybridization. . Hum Genet 73: 354-7.

Lisitsyn, N., Lisitsyn, N. & Wigler, M. (1993) Cloning the differences between two complex genomes. Science 259: 946-51.

Liu, P., Legerski, R. & Siciliano, M.J. (1989) Isolation of human transcribed sequences from human-rodent somatic cell hybrids. . Science 246: 813-5.

Lovett, M., Kere, J. & Hinton, L.M. (1991) Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. Proc.Natl.Acad.Sci.USA 88: 9628-9632.

MacDonald, M.E., Anderson, M.A., Gilliam, T.C., Tranejaerg, L., Carpenter, N.J., Magenis, E., Hayden, M.R., Healey, S.T., Bonner, T.I. & Gusella, J.F. (1987) A somatic cell hybrid panel for localizing DNA segments near the Huntington's disease gene. . Genomics 1: 29-34.

MacDonald, M.E., Haines, J.L., Zimmer, M., Cheng, S.V., Youngman, S., Whaley, W.L. & et al. (1989) Recombination events suggest potential sites for the Huntington's disease gene. Neuron 3: 183-190.

MacDonald, M.E., Lin, C., Srinidhi, L., Bates, G., Altherr, M., Whaley, W.L., Lehrach, H., Wasmuth, J. & Gusella, J.F. (1991) Complex patterns of linkage disequilibrium in the Huntington disease region. . Am J Hum Genet 49: 723-34.

MacDonald, M.E., Novelletto, A., Lin, C. & et al. (1992) The Huntington's disease candidate region exhibits many different haplotypes. Nature Genet. 1: 99-103.

Magenis, R.E., Gusella, J., Weliky, K., Olson, S., Haight, G., Toth, F.S. & Sheehy, R. (1986) Huntington disease-linked restriction fragment length polymorphism localized within band p16.1 of chromosome 4 by in situ hybridization. . Am J Hum Genet 39: 383-91.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J. & O'Hoy, K. (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. . Science 255: 1253-5.

Martin, J.B. & Gusella, J.F. (1986) Huntington's disease Pathogenesis and management. New Eng. J. Med. 315: 1267-1276.

Monaco, A.P., Walker, A.P., Millwood, I., Larin, Z. & Lehrach, H. (1992) A yeast artificial chromosome contig containing the complete Duchenne muscular dystrophy gene. . Genomics 12: 465-73.

Myers, H.M., Leavitt, J., Lindsay, A.F. & et al (1989) Homozygote for Huntington's disease. Am.J.Hum.Genet. 45: 615-618.

Myers, R.M., Maniatis, T. & Lerman, L.S. (1987) Detection and localization of single base changes by denaturing gradient gel electrophoresis. . Methods Enzymol 155: 501-27.

Nakamura, Y., Culver, M., O'Connell, P.O., Leppert, M., Lathrop, G.M., Lalouel, J.M. & et al. (1988) Isolation and mapping of a polymorphic DNA sequence (pYNZ32) on chromosome 4p (D4S125). Nucl. Acids. Res. 16: 4186.

Nelson, S.F., McCusker, J.H., Sander, M.A., Kee, Y., Modrich, P. & Brown, P.O. (1993) Genomic mismatch scanning: a new approach to genetic linkage mapping. Nature Genet. 4:11-7.

O'Hoy, K.L., Tsilfidis, C., Mahadevan, M.S., Neville, C.E., Barcelo, J., Hunter, A.G. & Korneluk, R.G. (1993) Reduction in size of the myotonic dystrophy trinucleotide repeat mutation during transmission. . Science 259: 809-12.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. & Sekiya, T. (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. . Proc Natl Acad Sci U S A 86: 2766-70.

Peterson, M.G., Tanese, N., Pugh, B.F. & Tjian, R. (1990) Functional domains and upstream activation properties of cloned human TATA binding protein [published erratum appears in Science 1990 Aug 24;249(4971):844] . Science 248: 1625-30.

Pieretti, M., Zhang, F.P., Fu, Y.H., Warren, S.T., Oostra, B.A., Caskey, C.T. & Nelson, D.L. (1991) Absence of expression of the FMR-1 gene in fragile X syndrome. . Cell 66: 817-22.

Pohl, T.M., Zimmer, M., MacDonald, M.E., Smith, B., Bucan, M., Poustka, A., Volinia, S., Searle, S., Zehetner, G. & Wasmuth, J.J. (1988) Construction of a NotI linking library and isolation of new markers close to the Huntington's disease gene. . Nucleic Acids Res 16: 9185-98.

Pritchard, C., Casher, D., Bull, L., Cox, D.R. & Myers, R.M. (1990) A cloned DNA segment from the telomeric region of human chromosome 4p is not detectably rearranged in Huntington disease patients. Proc. Natl. Acad. Sci. USA 87: 7309-13.

Pritchard, C., Zhu, N., Zuo, J., Bull, L., Pericak-Vance, M.A., Vance, J. & et al. (1992) Recombination of 4p16 DNA markers in an unusual family with Huntington disease. Am. J. Hum. Genet.50:1218-30.

Pritchard, C.A., Casher, D., Uglum, E., Cox, D.R. & Myers, R.M. (1989) Isolation and field-inversion gel electrophoresis of DNA markers located close to the Huntington disease gene. Genomics 4: 408-18.

Raff, M.C. (1992) Social controls on cell survival and cell death. Nature 356: 397-400.

Richards, J.E., Gilliam, T.C., Cole, J.L., Drumm, M.L., Wasmuth, J.J., Gusella, J.F. & al, e. (1988) Chromosome jumping from D4S10 (G8) toward the Huntington disease gene. Proc. Natl. Acad. Sci. USA 85: 6437-41.

Richards, R.I. & Sutherland, G.R. (1992) Dynamic mutations: a new class of mutations causing human disease. . Cell 70: 709-12.

Ridley, R.M., Frith, C.D., Farrer, L.A. & Conneally, P.M. (1991) Patterns of inheritance of the symptoms of Huntington's disease suggestive of an effect of genomic imprinting. . J Med Genet 28: 224-31.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N. & Chou, J.L. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA [published erratum appears in Science 1989 Sep 29;245(4925):1437] . Science 245: 1066-73.

Robbins, C., Theilmann, J., Youngman, S., Haines, J., Altherr, M.J., Harper, P.S., Payne, C., Junker, A., Wasmuth, J. & Hayden, M.R. (1989) Evidence from family studies that the gene causing Huntington disease is telomeric to D4S95 and D4S90. . Am J Hum Genet 44: 422-5.

Schlessinger, D., Little, R.D., Freije, D., Abidi, F., Zucchi, I., Porta, G., Pilia, G., Nagaraja, R. & et al. (1991) Yeast artificial chromosome-based genome mapping: some lessons from Xq24-q28. genomics 11: 783-793.

Sheffield, V.C., Beck, J.S., Stone, E.M. & Myers, R.M. (1992) A simple and efficient method for attachment of a 40-base pair, GC-rich sequence to PCR-amplified DNA. . Biotechniques 12: 386-8.

Sheffield, V.C., Cox, D.R., Lerman, L.S. & Myers, R.M. (1989) Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes. . Proc Natl Acad Sci U S A 86: 232-6.

Skraastad, M.I., Bakker, E., de, L.L., Vegter, van, d.V.M., Klein, B.E., van, O.G. & Pearson, P.L. (1989) Mapping of recombinants near the Huntington disease locus by using G8 (D4S10) and newly isolated markers in the D4S10 region. . Am J Hum Genet 44: 560-6.

Smith, B., Skarecky, D., Bengtsson, U., Magenis, R.E., Carpenter, N. & Wasmuth, J.J. (1988) Isolation of DNA markers in the direction of the Huntington disease gene from the G8 locus. . Am J Hum Genet 42: 335-44.

Snell, R.G., Lazarou, L.P., Youngman, S., Quarrell, O.W., Wasmuth, J.J., Shaw, D.J. & Harper, P.S. (1989) Linkage disequilibrium in Huntington's disease: an improved localisation for the gene. . J Med Genet 26: 673-5.

Snell, R.G., Thompson, L.M., Tagle, D.A., Holloway, T.L., Barnes, G., Harley, H.G., Sandkuijl, L.A., MacDonald, M.E., Collins, F.S., Gusella, J.F., Harper, P.S. & Shaw, D.J. (1992) A recombination event that redefines the Huntington disease region. Am. J. Hum. Genet. 51: 357-362.

Strobel, S.A., Doucette, S.L., Riba, L., Housman, D.E. & Dervan, P.B. (1991) Site-specific cleavage of human chromosome 4 mediated by triple-helix formation. . Science 254: 1639-42.

Tagle, D.A., Swaroop, M., Lovett, M. & Collins, F.S. (1993) Magnetic bead capture of expressed sequences encoded within large genomic segments. . Nature 361: 751-3.

Taylor, S.A.M., Snell, R.G. & et al. (1992) Cloning of the a-adducin gene from the Huntington's disease candidate region of chromosome 4 by exon amplification. Nature Genet. 2: 223-227.

The Huntington's Disease Cooperative Research Group (THDCRG) (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72: 971-83.

Theilmann, J., Kanani, S., Shiang, R., Robbins, C., Huggins, M. & et al. (1989) Non-random association between alleles detected by D4S95 and D4S98 and the Huntington's disease gene. Med. Genet. 26: 676-681.

Thompson, L.M., Plummer, S., Schalling, M., Altherr, M.R., Gusella, J.F., Housman, D.E. & Wasmuth, J.J. (1991) A gene encoding a fibroblast growth factor receptor isolated from the Huntington disease gene region of human chromosome 4. . Genomics 11: 1133-42.

Trask, B.J. (1991) Fluorescence in situ hybridization: applications in cytogenetics and gene mapping. Trends Genet 7: 149-54.

Travis, G.H., Brennan, M.B., danielson, P.E., Kozak, C.A. & Sutcliffe, J.G. (1989) Identification of a photorecptor-specific mRNA encoded by the gene responsible for retinal degeneration slow (rds). Nature 338: 70-73.

van den Engh, G., Sachs, R. & Trask, B.J. (1992) Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. Science 257: 1410-1412.

Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F. & Zhang, F.P. (1991) Identification of a gene (FMR-

1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. . Cell 65: 905-14.

Viskochil, D., Buchberg, A.M., Xu, G., Cawthon, R.M., Stevens, J., Wolff, R.K., Culver, M., Carey, J.C., Copeland, N.G. & Jenkins, N.A. (1990) Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. . Cell 62: 187-92.

Vulpe, C., Levinson, B., Whitney, S., Packman, S. & Gitschier, J. (1993) Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper-transporting ATPase. Nature Genet. 3: 7-13.

Wallace, M.R., Marchuk, D.A., Andersen, L.B., Letcher, R., Odeh, H.M., Saulino, A.M., Fountain, J.W., Brereton, A., Nicholson, J. & Mitchell, A.L. (1990) Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients [published erratum appears in Science 1990 Dec 21;250(4988):1749] . Science 249: 181-6.

Wasmuth, J.J., Carlock, L.R., Smith, B. & Immken, L.L. (1986) A cell hybrid and recombinant DNA library that facilitate identification of polymorphic loci in the vicinity of the Huntington disease gene. . Am J Hum Genet 39: 397-403.

Wasmuth, J.J., Hewitt, J., Smith, B., Allard, D., Haines, J.L., Skarecky, D., Partlow, E. & Hayden, M.R. (1988) A highly polymorphic locus very tightly linked to the Huntington's disease gene. . Nature 332: 734-6.

Weber, B., Collins, C., Kowbel, D., Riess, O. & Hayden, M.R. (1991) Identification of multiple CpG islands and associated conserved sequences in a candidate region for the Huntington disease gene. . Genomics 11: 1113-24.

Wexler, N.S., Rose, E.A. & Houseman, D.E. (1991) Molecular approaches to hereditary diseases of the nervous system: Huntington's disease as a paradigm. Annu.Rev.Neurosci. 14: 503-529.

Wexler, N.S., Young, A.B., Tanzi, R.E. & et al. (1987) Homozygotes for Huntington's disease. Nature 326: 194-197.

Whaley, W.L., Bates, G.P., Novelletto, A., Sedlacek, Z., Cheng, S., Romano, D., Ormondroyd, E., Allitto, B., Lin, C. & Youngman, S. (1991) Mapping of cosmid clones in Huntington's disease region of chromosome 4. . Somat Cell Mol Genet 17: 83-91.

Whaley, W.L., Michiels, F., MacDonald, M.E., Romano, D., Zimmer, M., Smith, B., Leavitt, J., Bucan, M., Haines, J.L. & Gilliam, T.C. (1988) Mapping of D4S98/S114/S113 confines the Huntington's defect to a reduced physical region at the telomere of chromosome 4. . Nucleic Acids Res 16: 11769-80.

Wharton, K.A., Yedvobnick, B., Finnerty, V.G. & Artavanis, T.S. (1985) opa: a novel family of transcribed repeats shared by the Notch locus and other developmentally regulated loci in D. melanogaster. Cell 40: 55-62.

Youngman, S., Bates, G.P., Williams, S., McClatchey, A.I., Baxendale, S., Sedlacek, Z., Altherr, M., Wasmuth, J.J., MacDonald, M.E. & Gusella, J.F. (1992) The

telomeric 60 kb of chromosome arm 4p is homologous to telomeric regions on 13p, 15p, 21p, and 22p. . Genomics 14: 350-6.

Youngman, S., Sarfarazi, M., Bucan, M., MacDonald, M., Smith, B., Zimmer, M., Gilliam, C., Frischauf, A.M., Wasmuth, J.J. & Gusella, J.F. (1989) A new DNA marker (D4S90) is located terminally on the short arm of chromosome 4, close to the Huntington disease gene. . Genomics 5: 802-9.

Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., Holman, K., Mulley, J.C., Warren, S.T. & Schlessinger, D. (1991) Fragile X genotype characterized by an unstable region of DNA. . Science 252: 1179-81.

Zuo, J., Robbins, C., Baharloo, S., Cox, D.R. & Myers, R.M. (1993) Construction of cosmid contigs and high-resolution restriction mapping of the Huntington disease region of human chromosome 4. Hum.Mol.Genet. In press:

Zuo, J., Robbins, C., Taillon-Miller, P., Cox, D. & Myers, R.M. (1992) Cloning of the Huntington disease region in yeast artificial chromosomes. Human Molecular Genetics 1: 149-159.

# CHAPTER ONE:

Cloning of the Huntington disease region in yeast artificial chromosomes.

The text of this chapter is a reprint of the material as it appears in Human Molecular Genetics Vol.1 No.3. 149-159, 1992

The gene responsible for Huntington disease has been localized to a 2.5 million base pair (Mb) region between the loci *D4S10* and *D4S168* on the short arm of chromosome 4. As part of a strategy to clone the HD gene on the basis of its chromosomal location, we first mapped thirteen DNA probes from the HD region by pulsed-field gel electrophoresis, leading to a long range restriction map of the 2.5 Mb HD region with NotI, MluI, NruI and CspI restriction sites. This map facilitated the isolation of genomic DNA from the HD region as a set of overlapping yeast artificial chromosome (YAC) clones. Twenty-eight YAC clones were identified by screening human YAC libraries with twelve PCR-based sequence-tagged sites (STSs) from the region. We assembled the YAC clones into overlapping sets by hybridizing them to a large number of DNA probes from the HD region, including the STSs. In addition, we isolated the ends of the human DNA inserts of most of the YAC clones to assist in the construction of the contig. Although almost half of the YACs appear to contain chimeric inserts and several contain internal deletions or other rearrangements, we were able to obtain over 2.2 Mb of the HD region in YACs, including one continuous segment of 2.0 Mb covering the region that most likely contains the HD gene. Ten of the twenty eight YAC clones comprise a minimal set spanning the 2.2 Mb. These clones had provided reagents for isolation of candidate genes for HD, for the identification of polymorphic markers, such as di- and tri-nucleotide repeats, and for the construction of cosmid contigs and high-resolution restriction map of the HD region described in Chapter Two.

*Human Molecular Genetics, Vol. 1, No. 3* **149–159**

# Cloning of the Huntington disease region in yeast artificial chromosomes

Jian Zuo[1], Carolyn Robbins[1], Patricia Taillon-Miller[4], David R.Cox[2,3] and Richard M.Myers[1,2]*

Departments of [1]Physiology, [2]Biochemistry and Biophysics and [3]Psychiatry, University of California at San Francisco, 513 Parnassus Avenue, San Francisco, CA 94143-0444 and [4]Center for Genetics in Medicine, Washington University School of Medicine, Box 8232, 4566 Scott Avenue, St Louis, MO 63110, USA

Received May 6, 1992; Revised and Accepted May 15, 1992

## ABSTRACT

**The gene responsible for Huntington disease has been localized to a 2.5 million base pair (Mb) region between the loci *D4S10* and *D4S168* on the short arm of chromosome 4. As part of a strategy to clone the HD gene on the basis of its chromosomal location, we isolated genomic DNA from the HD region as a set of overlapping yeast artificial chromosome (YAC) clones. Twenty-eight YAC clones were identified by screening human YAC libraries with twelve PCR-based sequence-tagged sites (STSs) from the region. We assembled the YAC clones into overlapping sets by hybridizing them to a large number of DNA probes from the HD region, including the STSs. In addition, we isolated the ends of the human DNA inserts of most of the YAC clones to assist in the construction of the contig. Although almost half of the YACs appear to contain chimeric inserts and several contain internal deletions or other rearrangements, we were able to obtain over 2.2 Mb of the HD region in YACs, including one continuous segment of 2.0 Mb covering the region that most likely contains the HD gene. Ten of the twenty eight YAC clones comprise a minimal set spanning the 2.2 Mb. These clones provide reagents for the complete characterization of this region of the genome and for the eventual isolation of the HD gene.**

## INTRODUCTION

Huntington disease (HD) is a neurodegenerative disorder characterized by late age onset of symptoms, including chorea, psychiatric problems, and dementia. The disease is inherited in an autosomal dominant fashion and shows age-dependent penetrance [1]. Spiny neurons in the striatum have been found to be severely depleted in the brains of HD patients, but the primary biochemical defect responsible for the disease is not known [2]. Intensive efforts are underway that use meiotic linkage mapping and genomic cloning strategies to isolate the gene responsible for HD based on its location in the genome. In 1983, Gusella et al. [3] identified a probe, G8, that showed tight linkage to the HD gene in a large Venezuelan pedigree. This probe was found to identify a locus, *D4S10*, on the short arm of chromosome 4. Subsequent multi-point linkage analysis of the Venezuelan pedigree and two American families further localized the gene defect within a six million base pair (Mb) region between *D4S10* and the 4p telomere [4]. More recent analysis of recombination events has provided strong evidence that the HD gene lies distal to *D4S10* [5, 6] and proximal to *D4S168* [7, 8, 9, 10], a region that has been physically mapped and determined to contain about 2.5 Mb [5]. This location for the disease gene is also supported by linkage disequilibrium data from several groups [9, 11, 12, 13].

Positional cloning strategies have generally relied on defining the boundaries of the disease locus as tightly as possible by analysis of recombinants, followed by cloning the DNA between flanking DNA markers, isolating candidate genes from the cloned segment, and identifying a causative mutation in one of the candidate genes. The isolation of large segments of genomic DNA has been greatly facilitated by the yeast artificial chromosome (YAC) cloning system, which allows the cloning of human genomic DNA inserts up to at least 1 Mb in length [14, 15]. YAC clones carrying DNA sequences from the region of interest can be identified from libraries by hybridization of DNA probes and by screening pools of YAC clones in the libraries by the polymerase chain reaction (PCR) approaches [16, 17]. In this study, we describe the isolation of twenty eight YAC clones from the 2.5 Mb HD region. By analyzing the probe content, sizes and ends of the YACs, we determined that the YACs cover most of the 2.5 Mb region and that one contiguous set of overlapping clones covers a 2 Mb segment that most likely contains the gene.

## RESULTS

### Physical Mapping of the HD region

Our plan for obtaining the YAC clones was based on an already-existing high resolution physical map of the HD region [5, 18, 19]. Our strategy was to devise STSs from DNA markers on this map that are spaced evenly throughout the region. The distances between each STS were such that YAC clones isolated with one STS would have a high likelihood of overlapping with YAC clones isolated from adjacent STSs, based on the known average size of inserts in the YAC libraries we used. While the locations of seven of the DNA probes we used were previously known[5, 18, 19], a large number of the probes had not been mapped precisely. We mapped seven of these (four bacteriophage lambda probes, JZ1, 300, 336 and 251 [19, 20], two YAC end probes, A10LN3 and F11LN4, and probe 126E1.3 from cosmid 226F1) by pulsed-field gel electrophoresis (PFGE). Peripheral leukocyte DNA from a normal individual was digested with NotI, MluI, NruI, CspI and with all possible double digest combinations of these enzymes. After blotting the gels, we sequentially hybridized

* To whom correspondence should be addressed

**Table 1.** PFGE Fragment Sizes (kb)

| Probe | N | N/M | N/R | N/C | M | M/R | M/C | R | R/C | C |
|---|---|---|---|---|---|---|---|---|---|---|
| 126E1.3 | 370 | 370 | 370 | 280 | 350 | 350 | 220 | LM | 700 | LM |
| YNZ32RP3 | 370 | 370 | 370 | 300, (200)<br>(30) | 350 | 350 | 250 | LM | 700<br>(200) | LM,(250)<br>(30) |
| 221A11 1.7RI | 80 | 80 | 80 | (200)<br>80 | (450,350)<br>(280),220<br>(150) | (450)<br>(350)<br>(220) | (450)<br>(200)<br>100 | 360 | (360),200<br>70 | 200 |
| 300α | 480 | (620),480<br>380,240<br>(170) | 360<br>210 | 480 | 660<br>420<br>270 | (660), 460<br>350, 270<br>170 | 480, 400<br>290, (200) | LM<br>360 | 450<br>(360)<br>240 | 750<br>500 |
| 674D | 480 | 480, 400<br>(170) | 170 | 480 | 660, 420<br>(250), 200 | (320)<br>200 | 480, 400<br>230,180 | 680 | 150 | (750)<br>500 |
| 336b | 20 | 20 | 20 | 20 | 360 | 360 | 360 | 680 | (430)<br>340 | 340 |
| A10LN3 | 270 | 270 | 270 | 270 | 360 | 360 | 360 | 680 | (400)<br>360 | 360 |
| F11LN4 | 300 | 210 | (300), 180 | 210 | 220 | 220, 110 | 220 | (680) | (220), 110 | 220 |
| XP500 | 300 | 210 | (300), 100 | 210 | 220 | 220, 110 | 220 | 350 | 220, 110 | 220 |
| C4H | 180 | | | | 300 | | | | 260 | 250 |
| 251e | 100 | 100 | 100 | 100 | 300 | | | | 260 | 250 |
| 731C | 20 | | | | (330), 220 | | | | | |

Note: N-NotI; M-MluI; R-NruI; C-CspI; DNA probes are described in Materials and Methods; ( )-weak band; LM-limiting mobility.

them with each of the probes from HD region, and compared the fragment sizes obtained with each probe (Table 1).

From these experiments, we determined that the map position of probe 300α (from locus *D4S131*) is between *D4S95* and *D4S180* because it recognized the same 480 kb NotI fragment and 660 kb MluI fragment as probe 674D from locus *D4S95*, but it hybridized to a different 360 kb NruI fragment (Table 1). Probe JZ1-1, near locus *D4S180*, was used to isolate a contig of cosmids, one of which is cosmid clone 221A11, from a chromosome 4-specific cosmid library provided by Los Alamos National Laboratory (LANL). A probe from this cosmid, 221A11 1.7RI, hybridized to a 360 kb NruI fragment, which is the same as that recognized by probe 300α at *D4S131* and very likely by probe L19ps11 at *D4S180* [5]. On this basis, we deduce that the order of these probes is centromere-JZ1-1-300α-674-telomere. The orientation of this probe cluster was confirmed by a cosmid walk from probe YNZ32RP3 at *D4S125*, which led to cosmids that overlap JZ1-1 but not 300α. Probe 336b from lambda phage 336 at locus *D4S136* hybridized to the same 680 kb NruI fragment as did probe 674D; this NruI fragment is very likely the same as that recognized by a probe from *D4S181* in Bates et al [5]. However, probe 336b hybridized to a small NotI fragment that does not hybridize to probes from *D4S95* and *D4S181*, identifying a NotI fragment between these latter two markers not reported in reference 5 but later reported in reference 21.

Probe F11LN4, the YAC end from YAC D42F11, recognized the same 300 kb NotI and 220 kb MluI fragments as probe XP500 at *D4S43*, but recognized a different NruI fragment, indicating that F11LN4 is very close and proximal to *D4S43*. Probe 251e from phage 251 at locus *D4S135* recognized the same 250 kb CspI fragment as probe C4H, and hybridized to a small NotI fragment not recognized by any other probe in our study. These results, in combination with results from SalI digests (data not shown), led to the assignment of the location of this probe between *D4S43* and *D4S98*. Figure 1 shows the map positions of twelve DNA probes that we analyzed by PFGE, as well as two DNA probes (pHD3 and 337) that had been previously mapped [5,38].

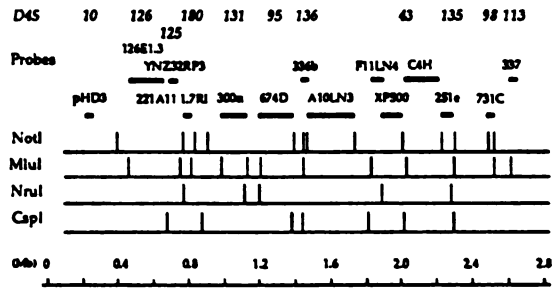



**Figure 1.** Long-range restriction map of the HD region, including the positions of fourteen of the DNA probes used in our study. The dark horizontal lines define the limits of the locations of each probe. The locations of cleavage sites of four rare-cutter restriction enzymes are shown as vertical lines. DNA fragment sizes on pulsed-field gels to which each probe hybridized are shown in Table 1. The locations of two DNA probes, pHD3 (a subclone of the G8 probe) and 337, are based on data in references 5 and 38.

## YAC Library Screening

Based on the physical mapping studies described above, we generated STSs from 10 DNA markers spaced an average of 250 kb apart, with two large intervals (a 500 kb interval between STSs at loci *D4S136* and *D4S43* and a 300 kb interval between STSs at loci *D4S180* and *D4S131*). As described in a later section, we developed two additional STSs from the ends of a YAC clone (A109E7) isolated in the initial screening to obtain markers in these two large intervals. We used these twelve STSs to screen four different YAC libraries (libraries A – D) constructed in the pYAC4 vector at the Center for Genetics in Medicine at Washington University by using a combination of PCR and hybridization approaches [14, 17, 22]. In some cases, we used PCR screening for all steps, including the identification of the microtiter well carrying a single YAC clone. We also used a few

**Table 2.** STSs Used in This Study

| STS | D4S | Primer Set | | Product Size (bp) | PCR Conditions* | Sequence Source# |
|---|---|---|---|---|---|---|
| HD196 | 10 | G8-PS-9: | 5'-ACC TGG ATC TCG GGC TTC-3' | 196 | 60°C, TNK | reference 44 |
| | | G8-PS-10: | 5'-GAA CAC AGA ATG GGC TGC-3' | | | |
| HD176 | 126 | 126-PS-3: | 5'-CTT CGC CCT TTC ACT GTG ATT GTA-3' | 176 | 60°C, TNK | Cos27G11/T3 |
| | | 126-PS-4: | 5'-GAC ATG TTC TGG ATT GTA TAG TGT G-3' | | | |
| HD227 | 125 | YNZ32-PS-1: | 5'-TTC TCT GTG TGC TGC AGA ATT TGG-3' | 230 | 55°C, TNK | pYNZ32/M13 |
| | | YNZ32-PS-2: | 5'-ATC CTG AAA GAG TTT CCT GGC AGG-3' | | | |
| HD200 | 125 | YNZ32-PS-5: | 5'-ACA TTT ACT AAA CTC TGG TCT GAG G-3' | 200 | 65°C, TNK | reference 45 |
| | | YNZ32-PS-6: | 5'-CCG GCA GCT GAG GAG GTG CCT CTG C-3' | | | |
| HD212 | 180 | 180-PS-3: | 5'-CAC ATC TTC CTG TTT CTT TGA ACA TC-3' | 212 | 60°C, TNK | Cos189F4/T3 |
| | | 180-PS-4: | 5'-GAG AGA CCC CAG AGT CCA GCA G-3' | | | |
| HD251 | | E7L-PS-1: | 5'-GGC AAT ACA GCA AGA TCT CAT CTA-3' | 251 | 60°C, TNK | pE7L12h |
| | | E7L-PS-2: | 5'-GAG AAT GTT AGT TAA CAT TCA TAG-3' | | | |
| HD183 | 95 | 674-PS-3: | 5'-TGG CTG AGT TCA GCT CAG TGC AGG C-3' | 183 | 60°C, TNK | Cos674B/T3 |
| | | 674-PS-4: | 5'-TCT CAC TCC AGT CTG CAG AAC TGG C-3' | | | |
| HD250 | 136 | 336-PS-1: | 5'-CTG ACT TGA TCC AAT CCA AAG GAA AG-3' | 250 | 60°C, A | pDan336b |
| | | 336-PS-2: | 5'-TTG AAC CTA GTA GGC GGA AGT TGC AC-3' | | | |
| HDE7R183 | | E7R-PS-1: | 5'-TAC CAA ACT TTC AGA CTT TGA AAG-3' | 183 | 60°C, TNK | pE7R14e |
| | | E7R-PS-2: | 5'-CCT GTG ATT TTT CAT TTT TTC TGG-3' | | | |
| HD500 | 43 | XP500-PS-1: | 5'-CAA GTA ACT TCC AAG GGT GAC-3' | 500 | 55°C, TNK | pXP500 |
| | | XP500-PS-3: | 5'-GTG GGC ACG GCT TGA TTC ACG GC-3' | | | |
| HD120 | 135 | 251-PS-1: | 5'-GCT TCC CTG TAA TGT TCT CAC AC-3' | 120 | 60°C, A | pDan251e |
| | | 251-PS-2: | 5'-CCC CTA TAT GCT CTG ATC TTG G-3' | | | |
| HD800 | 98 | 731-PS-1: | 5'-GCA GAG GCA CAG CCT TTG GC-3' | 800 | 65°C, TNK | pBS731 |
| | | 731-PS-2: | 5'-GAG GCA CCT GTC CTC CTC CCC-3' | | | |

Note: * Only annealing temperature and buffer (TNK or A) are indicated; The remaining PCR parameters are the same for all reactions (see Materials and Methods); #Cosmids and plasmids are described in Materials and Methods; T3 and M13 are primers used for sequencing.

of the DNA probes to screen a fifth YAC library, the E library, courteously provided by David Schlessinger at Washington University. The average insert size of the YACs in these libraries is about 300 kb, so the 89,000 total clones that were screened represent approximately eight-fold coverage of the human genome. In all, we obtained twenty-eight YAC clones (Table 3). All of the STSs led to the isolation of at least one positive clone, and on average, we obtained 2.3 YACs per STS. We determined the sizes of these YAC clones by PFGE, and found that the average insert size was 330 kb, with a range of 80 to 850 kb (Table 3 and Figure 2).

A few of the YAC-containing strains that we obtained in our screening carried multiple YACs or YACs that appeared to be mitotically unstable. PFGE analysis of single colonies of YAC strains D42A10 and B221E12 indicated the presence of more than one ethidium-bromide staining band in addition to the yeast chromosomes on the gel (Figure 2). In each case, both bands stained with ethidium equally and hybridized to YAC vector sequences, but only one band hybridized to the original DNA probe from which the STS was derived. Similar results have been previously reported in other YAC studies [23], and are thought to be the consequence of stable uptake of two different YACs during transformation. The YACs in several strains (A218C11, B134B4, B221E12, D42F11 and A1E4) appear to have undergone deletion or other rearrangement during propagation, as non-stoichiometric amounts of additional bands were present on pulsed-field gels (see A218C11 in Figure 2, for example). In other cases, different subclones from a single YAC strain contained YAC bands on PFGE that were different in size (these include A218C11, B134B4, D42F11, B206C11, A1E4, D110H3 and B50A12). In these cases, we obtained subclones of the strain with the largest size insert and used these subclones in subsequent analysis. In addition, YAC clone A1E4 showed small rearrangements around C4H (data not shown). These eight unstable clones span the HD region and five of them are from the region between *D4S43* and *D4S136* (Table 3).

### Probe Content Mapping of YACs

To determine the relationships of the twenty eight YACs to each other and to further assess their content, we utilized a large number of DNA markers from the HD region. Our approach was similar to that used by Monaco et al. [24] and Palmieri et al.[25], and involved determining the presence or absence of a large number of densely-spaced DNA probes in each of the YACs. Results from this 'probe content mapping' strategy indicate which YACs have overlapping inserts and allow a YAC contig to be established from the clones. The probes we used in this study include fourteen landmarks that we mapped (Figure 1) and twenty-one other probes, most of which are YAC end probes (Table 3). The order of the latter 21 probes in relation to the 14 landmarks was deduced based on the best fit of the sizes of YACs with minimal rearrangements. We sequentially hybridized each of 35 DNA probes (Table 3) to filters containing: 1) colonies from each YAC strain, 2) YAC DNAs separated on pulsed-field gels, and 3) YAC and human genomic DNAs digested with EcoRI and separated on standard agarose gels. These hybridizations confirmed the positive PCR signal of each YAC strain, gave the sizes of the EcoRI fragment containing the probe, and confirmed that the fragment sizes recognized by the probes were the same in genomic DNA and YAC DNA.

This probe content mapping approach, in combination with the PFGE mapping described above, allowed us to estimate the extent of each YAC in the 2.5 Mb region and to determine the continuity of the YAC contig (Table 3; Figure 1). For example, we found that YAC clone D42F11 contained all the probes from *D4S136* to probe F11LN4 (Table 3), and the size of the YAC (580 kb) is consistent with the distance between these two markers on the physical map (about 500 kb; Figure 1). This result indicates that

**Table 3.** Probe Content Mapping of YACs

| YAC | STS | Size (kb) | DNA Probes | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | pHD3 | C11LM1 | H2LH3 | 126E1.3 | G4RN4 | YNZ32RP3 | C10LN4 | 198H5.4R1 | H4RN4 | 221A11.7R1 | J/1.1 | G12LM1 | H4LM1.2 | E7L12b | B4LM1 | 300r | A10RN3 | 674D | G12RH3 | 336b | E7R14e |
| A218C11* (cWST50) | HD196 | 350 | + | + | + | - | - | - | | | - | | | | | | | | | | | | |
| A231H2 (cWST51) | HD196 | 140 | + | + | + | - | - | - | | | - | | | | | | | | | | | | |
| B111A3 (cWST52) | HD196 | 260 | + | + | + | - | - | - | | | - | | | | | | | | | | | | |
| B213G4 (cWST53) | HD196 | 160 | + | + | + | - | - | - | | | - | | | | | | | | | | | | |
| B165G4 (cWST47) | HD176 | 280 | - | - | - | + | + | - | - | - | - | - | - | - | - | - | - | | | | - | | |
| E155H5 (cWST55) | HD200 /HD227 | 140 | - | | - | | - | + | - | - | - | - | | | | | | | | | | | |
| D23C10 (cWST56) | HD212 | 580 | - | | - | | | - | + | + | + | + | + | + | + | - | | - | | | | | |
| D39H4 (cWST57) | HD212 | 290 | - | | | | | - | - | - | + | + | + | + | + | - | | - | | | | | |
| A187G12 (cWST58) | HD251 | 370 | | - | - | | | - | - | - | - | - | - | + | + | + | + | + | | + | + | - | |
| B134B4* (cWST59) | HD251 | 380 | | - | - | | | - | - | - | - | - | - | - | - | + | + | - | | - | - | | |
| D79E6 (cWST40) | HD183 | 850 | | - | - | | - | - | - | - | - | - | - | + | + | + | + | + | + | + | + | - | |
| A109E7 (cWST18) | HD250 /251/E7R 183 | 350 | | - | - | | - | - | - | - | - | - | - | - | - | + | + | + | - | - | - | + | + |

| YAC | STS | Size (kb) | A10R N3 | 674D | G12R H3 | 336b | E7R14e | A10L N3 | E4LH3 | H3RN3 | F11LN4 | XP500 | C4H | C5LH3 | B9LH3 | 251e | A12LN3 | F2LN4 | 73IC | 337 | 113E2RI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D102A10 (cWST61) | HDE7R 183 | 420/850 | + | + | + | + | + | + | - | - | - | - | | - | | | | | | | |
| D42F11* (cWST62) | HDE7R 183 | 580 | - | - | - | + | + | + | + | + | + | - | | - | - | | - | | | | |
| D41C7 (cWST63) | HDE7R 183 | 370 | - | - | - | + | + | - | | | | - | | | | | | | | | |
| D93B7 (cWST64) | HDE7R 183 | 80 | - | - | | - | + | - | | | | - | | | | | | | | | |
| B221E12* (cWST65) | HDE7R 183 | 200/250 | - | - | - | + | + | - | | - | - | - | | - | | | | | | | |
| B206C11* (cWST66) | HDE7R 183 | 800 | | | - | - | + | + | - | - | + | - | | - | | | | | | | |
| A1E4* (cWST36) | HD500 | 320 | | - | - | - | - | - | + | + | + | + | + | - | - | - | - | | | | |
| D110H3* (cWST35) | HD500 | 430 | - | - | - | - | - | - | + | + | - | + | + | + | + | + | - | - | - | | - |
| D89D11 (cWST33) | HD500 | 280 | | - | - | - | - | | - | - | - | + | - | - | - | - | - | | - | | |
| D131A2 (cWST34) | HD500 | 100 | | - | | | | | - | | - | + | - | - | - | - | - | | - | | |
| A38E2 (cWST8) | HD120 | 300 | | - | - | - | - | | - | - | - | - | - | - | - | + | + | + | - | | |
| A291C5 (cWST14) | HD120 | 280 | | | - | - | - | | | - | - | - | - | + | + | + | + | - | | | |
| A52B9 (cWST9) | HD120 | 280 | | | - | | - | | - | - | - | - | - | - | + | + | - | - | - | | |
| A131G7 (cWST10) | HD120 | 260 | | - | - | | - | | | - | - | - | - | - | + | + | + | + | - | | |
| B50A12* (cWST38) | HD800 | 160 | | | - | | - | | | - | - | - | | - | - | | + | + | + | + | |
| A141B4 (cWST39) | HD800 | 200 | | | - | | | | | | - | - | | - | | | + | - | + | | |

Note: YACs are named according to both their library positions (A218C11 for example) and standard nomenclature at the Center for Genetics in Medicine at Washington University (cWST50). * YACs showed some degree of instability (see text). DNA probes are described in Materials and Methods and Table 4; their orders are deduced from their sizes and data in Figure 1 assuming minimal number of rearrangements. - negative hybridization. + positive hybridization. Blank space: not determined.

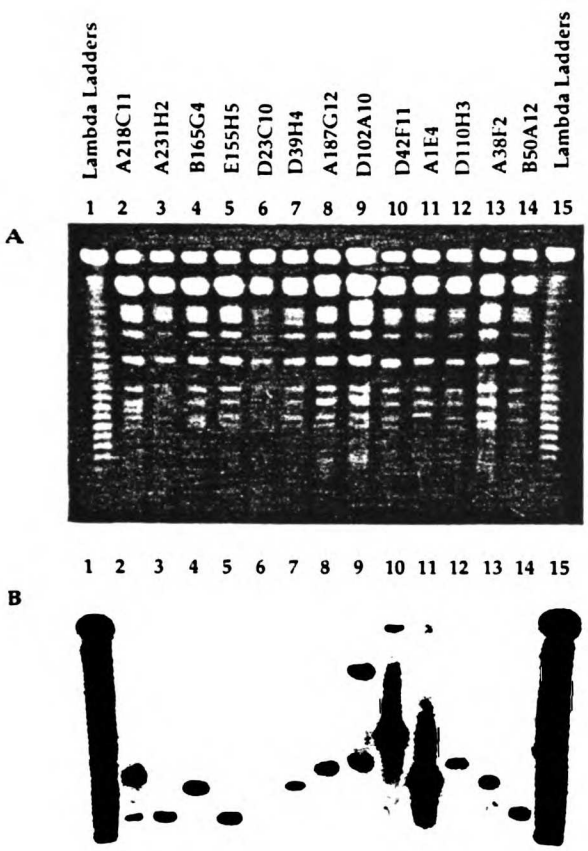



**Figure 2.** PFGE analysis of 13 YAC clones from the HD region. A. Ethidium bromide staining of a pulsed-field gel containing 13 of the YAC clones isolated in this study. B. Autoradiogram of a blot of the gel shown in (A) probed with a YAC end probe. The left end of YAC clone D42F11, designated F11LN4, was hybridized to the blot as described in Materials and Methods. This end probe contained 405 bp of the human genomic DNA sequence from the left end of the YAC insert and 75 bp from the left YAC vector arm. The size of the lambda ladder marker ranges from 48.5 kb and increases 48.5 kb with each successively larger band to approximately 1,000 kb. The hybridization signals seen in lanes 2–9 and 12–14 are due to cross-hybridization of the YAC end probe with the YAC vector sequences present in every lane. Similarly, the YAC vector sequences in the probe crosshybridize to the lambda ladder. The much stronger signals seen in lanes 10 and 11 are largely due to specific hybridization of the 405 bp human segment of the YAC end probe to these two YACs. The hybridization of this end, which was isolated from the YAC shown in lane 10, to YAC A1E4 in lane 11 indicates that these two YACs are overlapping. Sizes of all the YACs were deduced from the vector and specific hybridization signals on this and other autoradiograms.

most, if not all, of the genomic DNA insert in this YAC clone is derived from the HD region. On the other hand, we found that YAC clone D41C7 hybridized to only two closely-spaced DNA probes, 336b and E7R14e, suggesting that the majority of the 370 kb insert is not derived from the HD region. We found one YAC clone, A109E7, that appears to contain an internal deletion (Table 3), as it contains DNA markers 300α and 336b, but does not contain the DNA marker 674D, which is known to lie between the two former markers on the map (Figure 1). This result was confirmed by analysis of additional probes, A10RN3 and G12RH3, that were derived as described below from the ends of two YACs (Table 3). Based on the analysis of YAC ends E4LH3, H3RN3, F11LN4, A12LN3 and F2LN4, we deduced the presence of similar internal deletions in three other YACs (B206C11, D110H3 and A141B4; Table 3).

Each DNA probe that we analyzed in this manner was found to be present in at least two different YACs, with the exception of probes from *D4S126*, *D4S125* and *D4S113*. In each of these three cases, only a single YAC was identified.

### Isolation and Characterization of YAC Ends

While the probe content mapping experiments described above allowed us to establish relationships between many of the YAC clones, it did not allow us to assemble a complete unambiguous contig of the region. For this reason, and to further characterize the YAC clones, we isolated the ends of the human DNA inserts of most of the YAC clones by using inverse PCR [26], Alu-vector PCR [27], and a cosmid hybridization scheme (Table 4 and Materials and Methods). In all, we isolated 37 ends from 23 different YAC clones (Table 4). In our hands, the inverse-PCR method was the most successful, leading to the isolation of the majority of the ends (32/37). We obtained two YAC ends (pE7L12h and pE7R14e), both from YAC clone A109E7, by Alu-vector PCR. We isolated four YAC ends (E6RH2, C7R, B7(L or R) and A12R) by identifying cosmid clones that recognize the ends of the YACs (Materials and Methods). Finally, we used many of the YAC ends that we isolated by the two PCR approaches as a hybridization probe to isolate a cosmid DNA clone corresponding to the end.

We characterized the YAC ends isolated by the two PCR-based approaches to determine whether they are derived from the HD region. We hybridized each YAC end separately to Southern blots containing genomic DNA from human cells (HeLa), hamster cells (GM459), and two different somatic hybrid cells. One of these hybrid cells was 9TK, which is a hamster cell line carrying an intact human chromosome 4 as its only human material [28]. The other hybrid cell was C25, a radiation hybrid cell line that carries about 25 Mb of the short arm of human chromosome 4 from just distal to RAF1P1 to the 4p telomere [20]. We also hybridized each YAC end to Southern blots containing EcoRI digests of all the YAC clones from each subregion of the 2.5 Mb segment. The results of one such YAC end analysis are shown in Figure 3. In this case, the results confirmed that YAC end A10LN3 is derived from DNA in the 2.5 Mb region (Figure 3A). In addition, this YAC end hybridized not only to itself but also to another YAC clone, D42F11 (Figure 3B), indicating that the inserts of these two YAC clones overlap.

In addition to using PCR-amplified YAC ends as hybridization probes to characterize their location in the genome, we used many of the amplified ends to isolate cosmid clones from a flow-sorted chromosome 4-specific cosmid library (Table 4). Positive identification of chromosome 4-specific cosmids with a PCR-amplified YAC end provided further evidence that the end was derived from chromosome 4 and provided additional cloned DNA material from YAC ends that aided in establishing the YAC contig (see below).

Our characterization of the 37 YAC ends is summarized in Table 4. For each YAC end, the table indicates the sizes of the genomic EcoRI DNA fragments that the end recognizes, the hybridization results with the cell line DNAs, whether the end

**Table 4.** Isolation and Characterization of YAC End Fragments

| YAC End* | Enzyme for Circularization | Size of Inverse PCR Product | Size of Genomic Fragment | Hybridization to: Human | 9TK | C25 | Other YACs | Cloned Cosmid or Plasmid | From HD Region[f] |
|---|---|---|---|---|---|---|---|---|---|
| C11LMI | MboI | 450 | 5.2 | + | + | + | + | C | Y |
| C11RAI | AccI | 750 | 10 | + | | | | | |
| H2LH3 | HaeIII | 200 | 5.2 | | | | + | C | Y |
| H2RN3 | NlaIII | 210 | 11 | + | | | | C | |
| G4LMI | MboI | 410 | 4/2.2+11a | +/+ | +/− | −/− | − | | N |
| G4RN4 | NlaIV | 500 | 2.1 | + | + | + | − | C | Y |
| H5LMI | MboI | 450 | .45 | + | + | − | | | Y |
| H5RN4 | NlaIV | 200 | .45 | | | | − | C | Y |
| C10LN4 | NlaIV | 450 | 5.2+5.1 | + | + | | − | C | |
| H4RN4 | NlaIV | 220 | 3 | + | + | | + | C | Y |
| H4LMI-2 | MboI | 450 | 1.4 | + | | | + | C | Y |
| G12LMI | MboI | 300 | 5.8 | + | | | + | C | Y |
| G12RH3 | HaeIII | 360 | 15 | + | + | + | + | | Y |
| B4LMI | MboI | 450 | 1.2 | | | | + | | Y |
| B4RH3 | HaeIII | 500 | 0.8 | | | | − | | N |
| E6LN4 | NlaIV | 450 | 3.8 | + | − | − | − | | N |
| E6RH2 | HincII | 670 | 9.5 | | | | + | P/C | Y |
| E7L12hb | | | 2.8 | + | + | + | + | P | Y |
| E7R14eb | | | 6 | + | + | + | + | P/C | Y |
| A10LN3 | NlaIII | 900 | 5 | + | + | + | + | C | Y |
| A10RN3 | NlaIII | 350 | 12 | + | + | + | + | | Y |
| C7LMI-1 | MboI | 400 | 6.5 | + | − | − | − | | N |
| C7Rc | | | 3.4 or 1.4 | | | | + | C | Y |
| E12RAI-1 | AccI | 450 | | | | | | | |
| B7L orRd | | | 6.2 | | | | + | C | Y |
| F11LN4 | NlaIV | 480 | 0.8 | + | + | + | + | C | Y |
| F11RH3 | HaeIII | 250 | | | | | | | |
| E4LH3 | HaeIII | 300 | 7.5 | | | | + | C | Y |
| D11LN3 | NlaIII | 150 | 3.5 | + | + | + | + | C | Y |
| H3LN3 | NlaIII | 200 | | | | | | | |
| H3RN3 | NlaIII | 300 | 7.5 | | | | + | C | Y |
| F2LN4 | NlaIV | 350 | 2.3 | | | | + | | Y |
| C5LH3 | HaeIII | 220 | 2.8 | | | | + | C | Y |
| B9LH3 | HaeIII | 400 | 3.0 | | | | + | C | Y |
| B4LN4 | NlaIV | 160 | 10 | | | | − | C | Y |
| A12LN3 | NlaIII | 250 | 5.2 | | | | + | | Y |
| A12R | | | | | | | | C | Y |

Note: * YAC end C11LMI, for example, stands for YAC A218C11 left end fromMboI enzyme digest for inverse PCR.
[a]: Unlike the 4 kb band, two extra bands (2.2 and 11 kb) were found to hybridize only to 9TK but not to C25;
[b]: Two end clones by Alu-vector PCR were cloned into plasmids and E7R14e was also found in cosmids;
[c]: A cosmid contained the right end of the YAC but the order of the two EcoRI fragments at the YAC end was not determined;
[d]: A cosmid covered an end of the YAC but which of the left or right end it represented was not determined.
+: positive hybridization, −: negative hybridization, blank space: not determined or ambiguous.
Y/N: End is/not from the HD region, C: cosmid clone; P: plasmid clone; P/C: both cosmid and plasmid clones.

hybridizes to additional YAC clones, and whether the end was also used to isolate a cosmid clone. Thirteen of the 37 YAC ends gave definitive results, that is, unique hybridization signals to DNA fragments of the appropriate size were observed, with all three genomic DNA samples (HeLa, 9TK and C25) on the Southern blots. Several other YAC ends gave definitive hybridization results with one or two of these genomic DNA samples, but gave weak or ambiguous signals with the other samples. The remaining YAC ends did not give signals with any of the genomic samples, because the end probes were either repetitive or too small to work well as a hybridization probe against total genomic DNA. Nevertheless, many of these YAC ends were still useful because they gave good signals when used as hybridization probes against DNA from all the YAC clones (for example, see YAC ends G12LMI and B9LH3, Table 4) and against the gridded chromosome 4 cosmid library.

For ten of the YAC clones, we were able to obtain and successfully characterize both ends of the YAC insert. On the basis of this analysis, six of these YACs (E155H5, D39H4, A187G12, A109E7, D102A10 and B50A12) appear to be derived from DNA only within the HD region, whereas the other four YACs (B165G4, D79E6, B134B4 and D41C7) are chimeric. We found that one of the chimeric YAC clones, B165G4, contains two segments of DNA derived from different regions of chromosome 4 (Table 4 and data not shown).

We were able to make further deductions about the degree of chimerism in the remaining YAC clones by determining their probe content and comparing their sizes to the physical map. A total of four YAC clones (D42F11, D89D11, A1E4 and A38F2) were analyzed in this way, and the latter three were deduced to be chimeric. These results, combined with the analysis of the ten YAC clones for which both ends were isolated and analyzed, suggest that the chimerism frequency in the YAC clones is about 50% (7/14). This result is in good agreement with previous analyses of YAC clones from other regions of the genome that were isolated from the same libraries [23, 29].

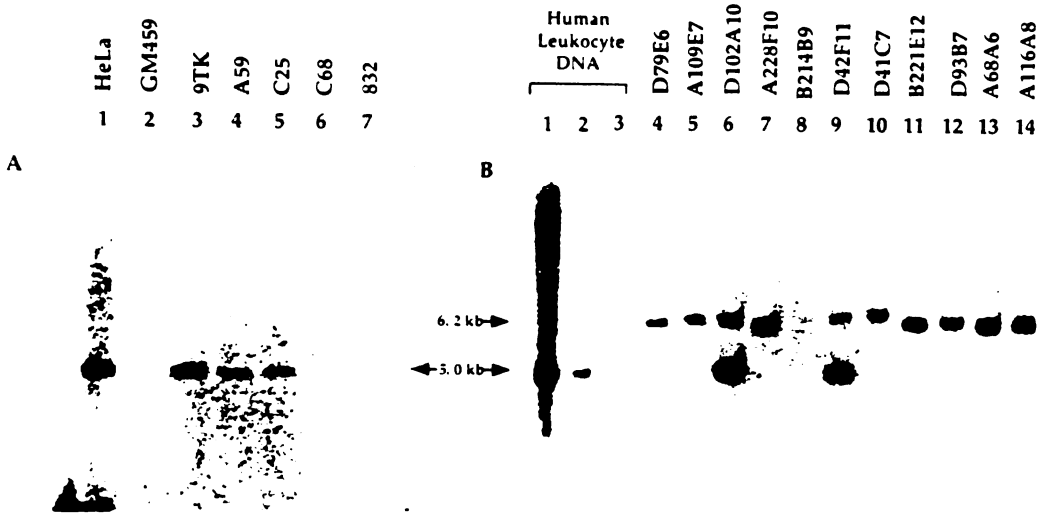



**Figure 3.** YAC end analysis. A. Localization of the YAC end to human chromosome 4p16. YAC end A10LN3, which was isolated from the left end of YAC clone D102A10, was labeled and hybridized to a blot containing EcoRI-digested genomic DNA from: human HeLa cells (lane 1), hamster GM459 cells (lane 2), hamster/human hybrid cell 9TK, which contains chromosome 4 as its only human material (lane 3) [28], radiation hybrid cell A59, a hamster cell that contains human 4p16.3 as most of its human material (lane 4), radiation hybrid cell C25, a hamster cell that contains human 4p16 as most of its human material (lane 5), radiation hybrid cell C68 [20], a hamster cell that contains much of human chromosome 4 but lacks the segment from just distal to*D4S180* to pter (lane 6), and hybrid cell 832 [17], which contains sequences from human chromosome 4p but lacks the HD region (lane 7). B. Hybridization of the YAC end to selected YAC clones. Genomic DNA from human leukocytes from three individuals (lanes 1–3) and DNA from eleven YAC clones was digested with EcoRI, electrophoresed on a standard agarose gel, blotted, and the blot was hybridized with YAC end A10LN3. Signals of the appropriate size, 5.0 kb, are present in all three leukocyte samples (lanes 1–3), in YAC D102A10, the YAC from which the end was isolated (lane 6) and in YAC D42F11 (lane 9), indicating that these two YACs overlap. This YAC end probe contains 147 bp of the left arm vector sequences, which results in a signal of approximately 6.2 kb in all YAC samples. It is likely that the variation of about 200 bp in the size of vector signal is due to telomere length variation in the pYAC4 vector [43]. The weak signal seen at 6.2 kb in lane 8 is due to loading several-fold less YAC DNA in this lane compared to other lanes. The YAC clone shown in lane 8 (B214B9) was isolated from the Washington University YAC libraries and provided by Drs. Francis Collins and John Wasmuth, but was not characterized further by us. Digests shown in lanes 7, 13, and 14 are YACs from other regions of the genome that were used as controls.

In addition to establishing contigs and analyzing the chimerism of these YAC clones by isolating their ends, we used DNA sequences from two YAC ends (E7L12H and E7R14e; Table 3) to generate new STSs that were then used to screen the YAC libraries. These two walking steps resulted in the isolation of eight additional YACs that had not been identified in the initial screens (Table 3). The large number of YAC end clones also provided important probes in two regions where there were no probes available before, one between *D4S180* and *D4S131*, and the other between *D4S136* and *D4S43* [5, 10, 30].

### Contig Establishment

We combined information from the physical map of the HD region (Figure 1), the probe content analysis and molecular weight determination of the YAC clones (Table 3), and the analysis of YAC ends (Table 4) to generate a map of the 28 YAC clones (Figure 4). In addition to allowing contigs to be established, this analysis allowed a refinement of the estimated locations of all the DNA probes used in this study. This analysis resulted in the identification of a continuous overlapping set of YAC clones from just distal to *D4S125* to just proximal to *D4S168*, a physical distance of about 2 Mb (Figure 4). While this contig contains 22 YAC clones, a set of seven of these clones (D23C10, A187G12, D102A10, D42F11, D110H3, A38F2 and B50A12) can be used to represent a minimal overlapping set.

Despite intensive efforts to extend the YAC contig towards the centromere, we were unable to do so because we obtained only a single small non-chimeric YAC (E155H5) with two STSs from *D4S125*, and a single chimeric YAC (B165G4) for DNA markers in the *D4S126* region by screening YAC libraries covering eight genome-equivalents (Table 3; Figure 4). YAC clone E155H5 covers 140 kb of continuous sequence from the *D4S125* region, and we have extended the contig, connecting YAC E155H5 with YAC D23C10, by isolating cosmid clones from the region (data not shown). We determined that YAC clone B165G4, which is 280 kb in length, is chimeric and contains only about 80 kb of DNA from the HD region based on hybridization to cosmid clones from the region. These hybridization results did not allow this YAC to be connected to other YACs either proximal or distal to it (Table 3).

Finally, we analyzed the four YACs that were isolated with DNA probes in the *D4S10* region, which is just proximal to the boundary for the HD gene. Probe content and YAC end analysis indicated that YAC clones A218C11 and A231H2 extend towards the centromere from the pHD3 marker (Table 3) and provide only about 30 kb of DNA extending towards the HD region (Figure 4). While the two other YAC clones, B111A3 and B213G4, from this locus may extend further into the HD region, we did not characterize them extensively so their map locations are uncertain. Based on analysis of YACs and cosmids from the

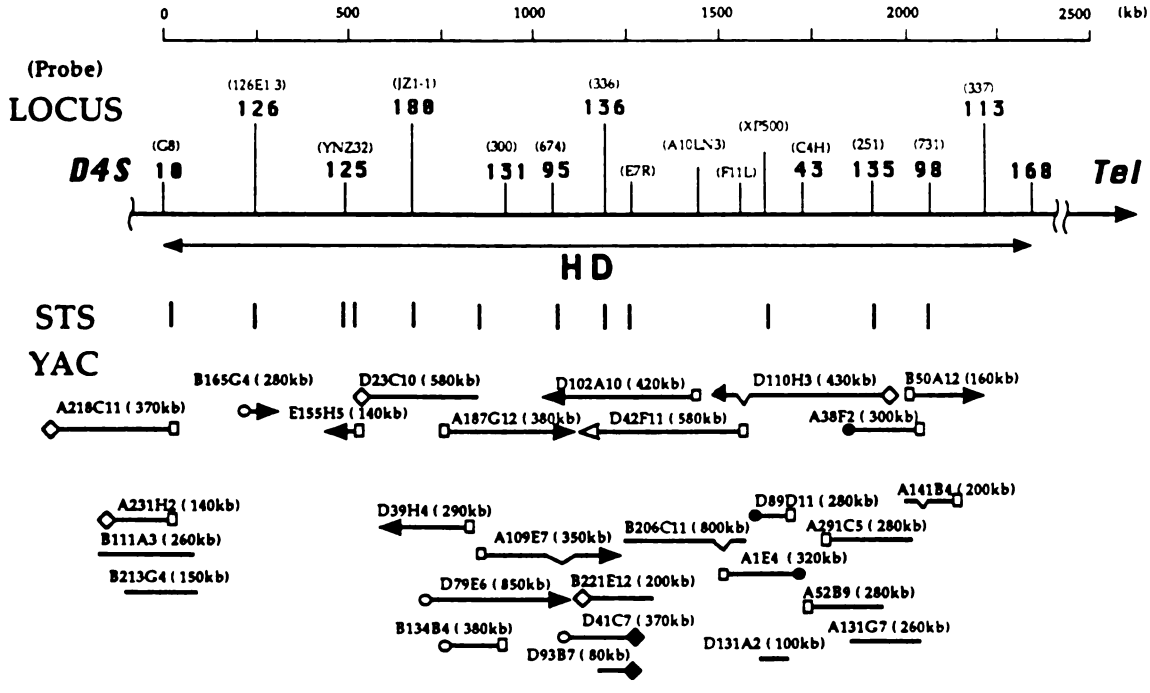



**Figure 4.** A map of the 28 YAC clones isolated in this study, based on probe content mapping, YAC size determination and YAC end analysis. A physical map of the DNA probes is shown at the top of the figure. Those probes for which STSs were generated are shown as vertical lines below the physical map. YAC clones are shown as horizontal lines below that, and the name of each YAC clone and its size in kb (in parentheses) are shown above each clone. Open squares indicate those YAC ends that were isolated from near the left arm of the YAC vector, and solid arrowheads indicate those YAC ends that were isolated from near the right arm of the YAC vector. The open arrowhead indicates a YAC end that was isolated but was not useful as a probe because it probably contains repetitive sequences. Open circles indicate those YAC ends that were isolated and shown to be a chimeric end, whereas closed circles indicate YAC ends that were not isolated but deduced to be a chimeric end as described in the text. Open diamonds indicate YAC ends that were isolated but not characterized, and closed diamonds indicate YAC ends that were isolated by cosmid clone hybridization. Discontinuities shown in YAC clones A109E7, B206C11, D110H3 and A141B4 indicate that these YACs are known to have internal deletions. The upper two horizontal lines of YACs are the ten YAC clones that make up the minimum set covering most of the HD region.

*D4S10*, *D4S126*, and *D4S125* region and the physical distance of about 500 kb between *D4S10* and *D4S125*, the two gaps not covered in this segment are less than a total of 200 kb. These results, combined with the results described above, indicate that we have covered at least 2.2 Mb of the 2.5 Mb HD region in YAC clones.

## DISCUSSION

In this paper, we describe the cloning of at least 2.2 Mb of the 2.5 Mb region of human chromosome 4p16.3 in which the Huntington disease locus resides. This cloning required the use of a large number of DNA probes from the HD region. Because many of the DNA probes we used in our study had not yet been placed on the physical map, we first performed pulsed-field gel electrophoresis to determine their locations; we also used PFGE to confirm the map positions of many of the other probes that had been previously reported. We adapted PCR-based assays for twelve of the most evenly-spaced probes, and these STSs were used to screen five YAC libraries. We analyzed YAC clones obtained in the screening by determining their sizes and probe content, and by isolating and characterizing the ends of most of the YACs. This analysis allowed us to establish the relationships between most of the YACs and build a contig spanning most of the HD region. In all, we isolated and characterized 28 YAC clones, and identified 10 YAC clones that make up a minimal set that covers 2.2 Mb. Seven YACs from this set of 10 clones cover a continuous stretch of about 2 Mb from around *D4S125* to a region distal to *D4S113*. The remaining three YACs from this set cover more than 0.2 Mb of the remaining 0.5 Mb of the HD region, between *D4S10* and *D4S125*, but do not overlap with each other to form a contig.

The HD region has a number of sites for rare cutting, methylation sensitive restriction enzymes, NotI, MluI, NruI and CspI (Figure 1). We found that YACs containing DNA from this region have many more detectable sites for these enzymes than human genomic DNA, presumably due to the lack of CpG methylation in yeast. These differences between yeast and mammalian DNA complicate the use of restriction mapping for determining correspondence between the genomic and YAC maps. While we occasionally used rare-cutter restriction analysis to help confirm the orientation of the YACs, we found that STSs, probe-content mapping and YAC end characterization were more useful for establishing the YAC contig in this region of the genome.

We screened six to eight genome-equivalents for each STS, yielding the set of 28 YAC clones. The number of YACs we obtained is lower than expected; based on random distribution

and the observed average insert size in the YAC libraries we screened, at least 72 YAC clones should have been obtained. These results suggest that clones from the region are underrepresented in the libraries. Nevertheless, for most of the 2.5 Mb region, we obtained multiple YAC clones for each DNA marker used in the screening, which allowed us to establish overlap relationships between the clones. However, one section within the HD region, which covers an approximately 500 kb stretch between *D4S10* and *D4S125*, was particularly difficult to obtain in YAC clones despite screening eight genome equivalents. In spite of this difficulty, we have been able to obtain large numbers of cosmid clones corresponding to this region by walking and direct screening experiments, such that at least half of it is covered by cosmid clones (unpublished results). This result suggests that regions of the genome that are difficult to clone in YACs will not be unattainable, but may be isolated in other vector/host systems.

Eight of the YAC clones from the set of 28 reported here demonstrated some degree of instability. Five of these clones are in the vicinity of the D4S43 locus, suggesting that this region of the genome contains sequences that are prone to deletions or other rearrangements in yeast. Similar instabilities of YACs have been reported by Neil et al [31].

According to published accounts, the boundaries of the HD region as defined by recombination events are between *D4S10* and *D4S168*, which spans 2.5 Mb. The proximal boundary is defined by a large number of crossover events that occur in an apparent hotspot of recombination a few hundred kb distal to *D4S10* [5, 32]. This boundary thus eliminates several hundred kb of DNA between *D4S10* and these recombination events as a location for the HD gene. The distal boundary is currently defined in the literature as being proximal to *D4S168* [5, 10], although an additional recombination event has been identified that places the gene about 300 kb proximal to *D4S168* (R. Snell, personal communication). Thus, the set of YACs reported here covers the distal boundary of the HD region, and lacks less than 200 kb around the proximal boundary. Furthermore, the 2 Mb continuous set of YACs from *D4S125* to near *D4S168* that we identified here covers all the DNA markers that show linkage disequilibrium with the HD mutation [9, 11, 12, 13]. These results together suggest that we very likely have cloned the entire segment that contains the gene despite the small gaps in the proximal region.

The YAC clones presented here provide an important resource to facilitate the search for the mutation responsible for HD. The YACs can be used directly as hybridization probes to screen cDNA libraries for candidate genes, or as reagents to select for coding sequences by exon selection [33, 34], exon amplification [35], or exon trapping [36]. In addition, the YACs can be used to screen cosmid or other smaller-insert libraries to obtain DNA for gene isolation and mutational analysis. Indeed, we have obtained about half of the 2.5 Mb HD region in overlapping cosmids, largely by labeling the YACs and screening a gridded chromosome 4-specific cosmid library and by conventional cosmid walking.

## MATERIALS AND METHODS

### DNA Probes

Probe pHD3 is a subclone containing a 2 kb HindIII EcoRI fragment from pK082, a plasmid clone from *D4S10* that was provided by Dr. Jim Gusella. Probe 126E1.3 is a 1.3 kb EcoRI fragment from cosmid 226E1 at *D4S126* that was isolated from the Los Alamos National Laboratory (LANL) gridded chromosome 4 cosmid library. 27G11 is another such cosmid overlapping with 226E1. Probe YNZ32RP3 is a 3 kb EcoRI PstI fragment from pYNZ32 (locus *D4S125*) and was obtained from the American Type Culture Collection. Probe 198H5.4RI is a 0.4 kb EcoRI fragment from cosmid 198H5 isolated from the LANL cosmid library by walking from YNZ32RP3. Probe JZ1-1 is a 1.2 kb EcoRI XhoI fragment of pJZ1-1, which is a plasmid subclone from lambda phage JZ1, which was isolated from radiation hybrid C25 [19, 20] and localized to a segment near *D4S180* by PFGE (Table 1; Figure 1). Probe 221A11.1.7RI was obtained from cosmid 221A11, which was obtained by walking from JZ1-1; cosmid 189F4, also obtained by walking from JZ1-1, overlaps cosmid 221A11. Probe 300α is a 0.8 kb HindIII fragment from pDan300α, which is a 3.5 kb EcoRI/BamHI plasmid subclone from lambda phage 300 (at *D4S131* [19]). Probe 674D is a 1.1 kb EcoRI fragment from plasmid pBS674 (at *D4S95* [37]). Probe 336b is a 1.3 kb SacI fragment of pDan336b, which is a plasmid subclone from lambda phage 336 (at *D4S136* [19]). Probe XP500 is a 0.5 kb PstI/XbaI insert from plasmid pXP500, which was provided by Dr. Marcy MacDonald. Probe C4H is a 3 kb HindIII insert from plasmid pBRC4H, also provided by Dr. Marcy MacDonald. Probe 251e is a 0.7 kb EcoRI/HindIII fragment from lambda phage 251 (at *D4S135* [19]). Probe 731C is a 1 kb BamHI insert from plasmid pBS731 (at *D4S98* [37]), provided by Dr. John Wasmuth. Probe 337 is a 260 bp PstI/Sau3A insert from plasmid p337 (at *D4S113* [38]), which was provided by Dr. Marcy MacDonald. Probe 113E.2RI is a 2 kb EcoRI fragment from cosmid 113E, which we isolated from the LANL library by screening with probe 337.

### Physical Mapping

Genomic DNA from blood leukocytes was prepared in agarose blocks and digested with several rare-cutter restriction enzymes as described [8]. All PFGE analysis was performed on a Bio-Rad CHEF Mapper™ under conditions that fractionate the DNA from 50 kb to 1,000 kb range as recommended by the vendor. Lambda ladders from Bio-Rad were used as size standards on all gels. For blotting, the gels were denatured for 20 min and transferred to Hybond-N+ membranes (Amersham) for at least 8 hr and then neutralized for 15 min before hybridization. Filters were prehybridized in 10% dextran sulfate, 1 M NaCl and 1% SDS at 65°C for 2 to 12 hr. For repetitive probes, 25 mg/ml sheared and denatured human placental DNA was included in the prehybridization. Probes were prepared by random-priming with $\alpha^{32}$P-dATP and boiled for 5 min with 1 mg sonicated salmon sperm DNA. Hybridization was carried out at 65°C for 16 to 24 hr in a Model 310 Hybridization Incubator (Robbins Scientific). Filters were washed twice in 0.2×SSC, 0.1%SDS at 65°C for 30 min, and exposed to Kodak X-OMAT film at −70°C for 2 hr to 5 d. After autoradiography, radioactive probe was removed from the filters by treating with TE buffer (10 mM Tris HCl, pH 7.5, 1 mM EDTA) containing 0.5%SDS at 95°C for 5 min and rehybridized.

### STS Development

DNA sequences for STS development were obtained from published reports, GenBank, or by sequencing cosmids directly or plasmid subclones of markers. All PCR assays were performed either in a Perkin Elmer-Cetus 9600 machine or in a robot that transfers reaction tubes to three water baths set at the appropriate temperatures. We performed 35 cycles as follows: a denaturation step at 94°C for 60 sec; an annealing step at 55°C, 60°C or 65°C for 120 sec; and an elongation step at 72°C for 120 sec. Buffer TNK50 is described in Blanchard et al (Blanchard, M. M., Taillon-Miller, P., Nowotny, P. and Nowotny, V., in preparation). Reactions in Buffer B contained 10 mM Tris-HCl, pH8.3, 50 mM KCl; 1.5 mM $MgCl_2$; 0.2 mM dNTPs; 0.4 μM each primer. Buffer A is described in Green and Olson [17]. STSs were performed in 5 μl reaction volumes containing 30 ng genomic or yeast DNA with 0.5 units AmpliTaq polymerase (Perkin Elmer-Cetus). Inverse PCR assays were performed in 100 μl reaction volumes with 1 unit AmpliTaq polymerase in either Buffer TNK50 or Buffer B.

### YAC Library Screening and YAC Characterization

The YAC libraries used in this study were constructed at the Center for Genetics in Medicine at Washington University, St. Louis and are described in the Results and in references 22 and 39. A combination of PCR and hybridization was used to screen the A−D YAC libraries [17]. In these cases, PCR of DNA from pools of YACs was used to identify a single pool of 384 colonies (96×4) from four microtiter plates of the gridded libraries, and this was followed by hybridization of the DNA probe to a membrane containing DNA from the colonies to identify the individual microtiter well carrying the positive YAC clone. The E library was screened by using PCR on DNA from pools of microtiter plates, followed by PCR on DNA from pools of wells from the positive microtiter plate to identify the positive clone. After positive clones were identified, at least twelve streaked colony subclones from each putative positive YAC clone were analyzed by colony-lysis hybridization to confirm that only a single yeast strain was present in the microtiter well. In addition, each subclone was analyzed by PFGE to compare its insert size and hybridization pattern with other subclones.

Yeast strains carrying YACs were maintained under selective media throughout this study [14]. Colony lysis assays, YAC agarose block preparation and restriction enzyme digestions were performed as described[16, 23, 40]. PFGE analysis was performed as described above for human genomic DNA under Physical Mapping.

### YAC End Isolation

We used the protocol described by Silverman et al.[26] to isolate YAC ends by inverse PCR, with some modifications. PFGE-purified YAC clones were digested with six different frequent cutting enzymes for left arms and seven enzymes for right arms. We inactivated TaqI by phenol extraction; all other restriction enzymes were inactivated by heating at 65°C. The digests were then circularized at a low DNA concentration ( <0.5 ng/ml ). We used primer sets that were in the inverse orientation on the vector. After PCR amplification of circularized vector ends, we usually found PCR products in several different digestions from each YAC end. In many cases, we observed two bands differing in size of about 200 to 300 bp in a single digest, which we interpreted as the result of partial digestion prior to circularization. The sizes of the PCR products varied from 150 bp to 900 bp. Each product contained a genomic end fragment from the insert and two small pieces of vector fragments at each end (60 to 218 bps). In cases where we had several successful PCR amplifications from a YAC end, we used the fragment that contained the minimal length of vector sequence and had the most amplification product for further analysis. Several YAC end fragments amplified by inverse PCR were cloned into a plasmid vector and their DNA sequence was determined. In all cases, the presence of new DNA sequences adjacent to YAC vector sequences at the EcoRI cloning site and at the site used to digest for circularization were observed. A few right ends of YAC clones could not be amplified by inverse PCR, while in most cases the left ends were easily amplified. F11RH3, a YAC end from D42F11, was amplified by inverse PCR but did not give unique signals when used as a probe in all three steps of characterization, indicating that it may contain repetitive sequences.

Alu-vector PCR was performed by using two Alu primers TC65 and 278 [41] and two different vector primers, which are derived from the sup4 region of pYAC4 and overlap the EcoRI cloning sites: ODC333 (5'-TAG CTC GAG GAC TTT AAT TTA TCA CTA CGG AAT TC-3') for the left arm; ODC334 (5'-TAG CTC GAG CGC CCG ATC TCA AGA TTA CGG AAT TC-3') for the right arm. The two ends of YAC A109E7 were isolated by this approach. Alu-vector PCR products were compared with those of Alu primers alone and were hybridized with vector oligos. The specific bands were cut out from 1% agarose gel and ligated into the SmaI site of plasmid vector pSP72. These two YAC ends (desginated pE7L12h and pE7R14e) were sequenced and their identity was verified by the presence of vector oligo sequences, EcoRI cloning sites and Alu sequences. These amplified fragments were shown to be YAC ends by comparing their hybridization patterns with SalI and XbaI digests of the YAC DNA to the patterns obtained when the left and right vector fragments of pYAC4 were used as probes [42] (data not shown).

We used cosmids that we had isolated by hybridization to probes from the HD region as a means of identifying the ends of some YAC clones. Cosmids were directly labeled by random-priming and hybridized to blots containing EcoRI-digested YAC DNAs. Prior to hybridization, the blots were prehybridized with 25 μg/ml human placental DNA as described above. In addition, the EcoRI fragment sizes present in each cosmid were determined by agarose gel electrophoresis. Those YACs that hybridized to some but not all of the EcoRI bands of a cosmid identified the cosmid as likely overlapping one of the ends of the YAC clone.

## ACKNOWLEDGEMENTS

This work has involved a close and extensive collaboration with Dr. David Schlessinger at the Center for Genetics in Medicine, Washington University School of Medicine, and we are grateful for his support and contributions. We thank Dr. Marcy MacDonald (Harvard University) for probes XP500, C4H and 337, Dr. Jim Gusella (Harvard University) for probe pK082, Dr. John Wasmuth (University of California, Irvine) for probes pBS674 and pBS731, and Drs. Francis Collins (University of Michigan) and John Wasmuth for YAC clone B214B9 isolated from Washington University YAC library. Some probes used in this study were obtained from a chromosome 4-specific cosmid library constructed at the Human Genome Center, Los Alamos National Laboratory, Los Alamos, NM 87545 under the auspices of the U. S. Department of Energy and kindly provided by Dr. Larry Deaven. We thank Robert Lagace, Richard Goold, Achilles Dugaiczyk and Guy DiSibio in the UCSF Human Genome Mapping Center for their help in STS development, Dr. Gary Silverman (Washington University School of Medicine, now at Harvard University) for his advice about inverse PCR, and Dr. Russell Snell (University Hospital of Wales, Cardiff) for communicating his results prior to publication. We acknowledge colleagues in our laboratories for thoughtful discussions, sharing materials, and encouragement, particularly Laura Bull, Rosalind John, Yuh-Shan Jou, Catrin Pritchard, Ning Zhu, Andy Peterson and Grant Hartzog. This work was supported by grants from the Wills Foundation and NIH RO1 # NS26237 to R. M. M. and D. R. C., and NIH Grant # HG00201 to David Schlessinger.

## REFERENCES

1. Hayden, M.R. (1981) *Huntington's chorea* (Springer-Verlag, Berlin).
2. Martin, J.B., Gusella, J.F. (1986) *New Eng. J. Med.* **315**, 1267–1276.
3. Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E. et al. (1983) *Nature* **306**, 234–238.
4. Gilliam, T.C., Tanzi, R.E., Haines, J.L., Bonner, T.I., Faryniarz, A.G., Hobbs, W.J. et al. (1987) *Cell* **50**, 565–571.
5. Bates, G.P., MacDonald, M.E., Baxendale, S., Youngman, S., Lin, C., Whaley, W.L. et al. (1991) *Am. J. Hum. Genet.* **49**, 7–16.
6. Wasmuth, J.J., Hewitt, J., Smith, B., Allard, D., Haines, J.L., Skarecky, D., Partlow, E., Hayden, M.R. (1988) *Nature* **332**, 734–736.
7. Barron, L., Curtis, A., Shrimpton, A.E., Holloway, S., May, H., Snell, R.G. et al. (1991) *J. Med. Genet.* 28, 520–522.
8. MacDonald, M.E., Haines, J.L., Zimmer, M., Cheng, S.V., Youngman, S., Whaley, W.L. et al. (1989) *Neuron* **3**, 183–190.
9. MacDonald, M.E., Lin, C., Srinidhi, L., Bates, G., Altherr, M., Whaley, W.L. et al. (1991) *Am. J. Hum. Genet.* **49**, 723–734.
10. Whaley, W.L., Bates, G.P., Novelletto, A., Sedlacek, Z., Cheng, S., Romano, D., et al. (1991) *Som. Cell Mol. Genet.* **17**, 83–91.
11. Adam, S., Theilmann, J., Buetow, K., Hedrick, A., Collins, C., Weber, B., et al. (1991) *Am. J. Hum. Genet.* **48**, 595–603.
12. Snell, R.G., Lazarou, L., Youngman, S., Quarrell, O.W.J., Wasmuth, J.J., Shaw, D.J., et al. (1989) *Med. Genet.* **26**, 673–675.
13. Theilmann, J., Kanani, S., Shiang, R., Robbins, C., Huggins, M., et al. (1989) *Med. Genet.* **26**, 676–681.
14. Burke, D.T., Carle, G.F., Olson, M.V. (1987)*Science* **236**, 806–812.
15. Murray, A.W., Szostak, J.W. (1983) *Nature* **305**, 189–193.
16. Brownstein, B.H., Silverman, G.A., Little, R.D., Burke, D.T., Korsmeyer, S.J., Schlessinger, D., Olson, M.V. (1989) *Science* **244**, 1348–1351.
17. Green, E.D., Olson, M.V. (1990) *Proc.Natl.Acad.Sci.USA* **87**, 1213–1217.
18. Bucan, M., Zimmer, M., Whaley, W.L., Poustka, A., Youngman, S., Allitto, B.A., et al. (1990) *Genomics* **6**, 1–15.
19. Pritchard, C.A., Casher, D., Uglum, E., Cox, D.R., Myers, R.M. (1989) *Genomics* **4**, 408–18.
20. Cox, D.R., Pritchard, C.A., Uglum, E., Casher, D., Kobori, J., Myers, R.M. (1989) *Genomics* **4**, 397–407.
21. Lin, C., Altherr, M., Bates, G., Whaley, W.L., Read, A.P., Harris, R., Lehrach, H., Wasmuth, J.J., Gusella, J.F., MacDonld, M.E. (1991) *Somat. Cell Mol.Genet.* **17**, 481–488.
22. Bronson, S.K., Pei, J., Taillon-Miller, P., Chorney, M.J., Geraghty, D.E., Chaplin, D. (1991) *Proc.Natl.Acad.Sci.USA* **88**, 1676–1680.
23. Schlessinger, D., Little, R.D., Freije, D., Abidi, F., Zucchi, I., Porta, G., Pilia, G., Nagaraja, R. et al. (1991) *Genomics* **11**, 783–793.
24. Monaco, A., Walker, A.P., Millwood, I., Larin, Z., Lehrach, H. (1992) *Genomics* **12**, 465–473.
25. Palmieri, G., Capra, V., Romano, G., D'Urso, M., Johnson, S., Schlessinger, D., Morris, P., Hopwood, J., Natale, P.D., Gatti, R., Ballabio, A. (1992) *Genomics* **12**, 52–57.
26. Silverman, G.A., Jockel, J.I., Domer, P.H., Mohr, R.M., Taillon-Miller, P., Korsmeyer, S.J. (1991) *Genomics* **9**, 219–228.
27. Nelson, D.L. (1990) *Genet. Anal.Tech.Appl.* **7**, 100–106.
28. Stanley, W., Chu, E.H.Y. (1978) *Cytogenet.Cell Genet.* **22**, 228–231.
29. Green, E.D., Riethman, H.C., Dutchik, J.E., Olson, M.V. (1991) *Genomics* **11**, 658–669.
30. Gusella, J.F., Altherr, M.R., McClatchey, A.I., Doucette-Stamm, L.A., Tagle, D., Plummer, S., Groot, N., Barnes, G., Hummerich, H., Collins, F.S., Housman, D.E., Lehrach, H., MacDonld, M.E., Bates, G., Wasmuth, J.J. (1992) *Genomics* **13**, 75–80.
31. Neil, D.L., Villasante, A., Fisher, R.B., Vetrie, D., Cox, B., Tyler-Smith, C. (1990) *Nucleic Acids Research* **18**, 1421–1428.
32. Allitto, B.A., MacDonald, M.E., Bucan, M., Richards, J., Romano, D., Whaley, W.L. et al. (1991) *Genomics* **9**, 104–112.
33. Lovett, M., Kere, J., Hinton, L.M. (1991) *Proc.Natl.Acad.Sci.USA* **88**, 9628–9632.
34. Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D., Weissman, S.M. (1991) *Proc.Natl.Acad.Sci.USA* **88**, 9623–9627.
35. Buckler, A.J., Chang, D.D., Graw, S.L., Brook, J.D., Haber, D.A., Sharp, P.A., Housman, D.E. (1991) *Proc.Natl.Acad.Sci.USA* **88**, 4005–4009.
36. Duyk, G.M., Kim, S., Myers, R.M., Cox, D.R. (1990) *Proc.Natl.Acad.Sci.USA* **87**, 8995–8999.

37. Smith, B., Skarecky, D., Bengtsson, U., Magenis, R.E., Carpenter, N., Wasmuth, J.J. (1988) *Am.J.Hum.Genet.* **42**, 335–344.
38. Whaley, W.L., Michiels, F., MacDonld, M.E., Romano, D., Zimmer, M., Smith, B., Leavitt, J., Bucan, M., Haines, J.L., Gilliam, T.C., Zehetner, G., Smith, C., Cantor, C.R., Frischauf, A., Wasmuth, J.J., Lehrach, H., Gusella, J.F. (1988) *Nucl.Acids Res.* **16**, 11769–11780.
39. Abidi, F.E., Wada, M., Little, R.D., Schlessinger, D. (1990) *Genomics* **7**, 363–376.
40. Carle, G.F., Olson, M.V. (1985) *Proc.Natl.Acad.Sci.USA* **82**, 3756–3760.
41. Nelson, D.L., Ledbetter, S.A., Corbo, L., Victoria, M.F., Ramirez-Solis, R., Webster, T.D., Ledbetter, D.H., Caskey, C.T. (1989) *Proc.Natl.Acad.Sci.USA* **86**, 6686–6690.
42. Silverman, G.A., Ye, R.D., Pollack, K.M., Sadler, J.E., Korsmeyer, S.J. (1989) *Proc.Natl.Acad.Sci.USA* **86**, 7485–7489.
43. Shampay, J., Blackburn, E.H. (1988) *Proc.Natl.Acad.Sci.USA* **85**, 534–538.
44. Abbott, C., Povey, S. (1991) *Genomics*, **9**, 73–77.
45. Richards, B., Horn, G.T., Merrill, J.J., Klinger, K.W. (1991) *Genomics* **9**, 235–240.

# CHAPTER TWO:

## Construction of cosmid contigs and high-resolution restriction mapping of the Huntington disease region on human chromosome 4

The text of this chapter is a reprint of the material as it will appear in Human Molecular Genetics Vol. 2. July 1993 issue.

The gene responsible for Huntington disease (HD) has been localized to a 2.2 million base pair (Mbp) region between the loci *D4S10* and *D4S98* on the short arm of human chromosome 4. As part of a strategy to clone the gene based on its chromosomal location, we and others previously identified overlapping yeast artificial chromosome (YAC) clones covering most of this region. While these YAC clones were useful for initially obtaining long-range clone continuity, a number of features of the YACs indicated that smaller clones are generally more useful in the subsequent steps of the positional cloning strategy. In this paper, we use these YAC clones to generate sets of overlapping cosmid clones covering most of the HD region. We isolated a large number of cosmids by screening a chromosome 4-specific cosmid library with labeled DNA from a minimal overlapping set of YAC clones. These cosmid clones were further analyzed by restriction mapping and hybridization experiments, leading to the assembly of 185 cosmids into eleven contigs covering more than 1.65 Mbp and to a fine-structure restriction map of the region. Nine of these contigs cover 90 percent of the 1.7 Mbp subregion between loci *D4S125* and *D4S98* where the HD gene is now known to lie. The detailed restriction map and the cosmid clones should facilitate the detailed analysis of the putative HD gene, the identification and localization of other cDNAs and polymorphic markers, and they provide reagents for large scale DNA sequencing of this region of the human genome. Our results suggest that this strategy should be generally useful for converting YAC clones into cosmid contigs and generating high-resolution restriction maps of genomic regions of interest.

REVISED, APRIL 26, 1993

# Construction of cosmid contigs and high-resolution restriction mapping of the Huntington disease region of human chromosome 4

Jian Zuo, Carolyn Robbins, Siamak Baharloo, David R. Cox[1] and Richard M. Myers

Departments of Physiology and Psychiatry[1], The University of California at San Francisco, 513 Parnassus Avenue, San Francisco, CA 94143-0444

**ABSTRACT**

The gene responsible for Huntington disease (HD) has been localized to a 2.2 million base pair (Mbp) region between the loci *D4S10* and *D4S98* on the short arm of human chromosome 4. As part of a strategy to clone the gene based on its chromosomal location, we and others previously identified overlapping yeast artificial chromosome (YAC) clones covering most of this region. While these YAC clones were useful for initially obtaining long-range clone continuity, a number of features of the YACs indicated that smaller clones are generally more useful in the subsequent steps of the positional cloning strategy. In this paper, we use these YAC clones to generate sets of overlapping cosmid clones covering most of the HD region. We isolated a large number of cosmids by screening a chromosome 4-specific cosmid library with labeled DNA from a minimal overlapping set of YAC clones. These cosmid clones were further analyzed by restriction mapping and hybridization experiments, leading to the assembly of 185 cosmids into eleven contigs covering more than 1.65 Mbp and to a fine-structure restriction map of the region. Nine of these contigs cover 90 percent of the 1.7 Mbp subregion between loci *D4S125* and *D4S98* where the HD gene is now known to lie. The detailed restriction map and the cosmid clones should facilitate the identification and localization of cDNAs and polymorphic markers, and they provide reagents for large scale DNA sequencing of this region of the human genome. Our results suggest that this strategy should be generally useful for converting YAC clones into cosmid contigs and generating high-resolution restriction maps of genomic regions of interest.

## INTRODUCTION

Huntington disease (HD) is a neurodegenerative disorder inherited in an autosomal dominant manner that has an average onset of 38 years of age and is characterized by uncontrolled choreiform movements and psychiatric problems, frequently including severe depression and progressive dementia [1]. While postmortem studies on brains of HD patients suggest that the initial neuronal degeneration is in the caudate nucleus in the basal ganglia, the role that the mutated gene plays in this degeneration is unknown. A number of groups have applied meiotic linkage mapping and genomic cloning strategies to search for the gene responsible for the disease based on its chromosomal location [2]. Linkage studies and analyses of recombination events provided strong evidence that the HD gene lies in a region of the short arm of chromosome 4 located a few Mbp from the telomere [3]. This region is distal to a set of recombination crossovers between *D4S10* and *D4S126* and proximal to a single recombination breakpoint between *D4S43* and *D4S98* [4, 5]. Long range restriction mapping has determined that this distance is less than 2.2 Mbp [4, 6, 7, 8]. Linkage disequilibrium data from several groups suggested that the gene most likely lies between *D4S180* and *D4S182*, a 0.7 Mbp subfragment of the 2.2 Mbp region [9, 10, 11].

The identification of a disease gene based on its chromosomal location can be greatly facilitated by cloning the genomic region between the two flanking loci. From the cloned material, more polymorphic markers can be identified to further narrow the candidate region if possible; cDNAs can be isolated from the region and causative mutations for the disease can be finally identified. As part of efforts to identify the HD gene and to begin constructing cloned contigs of the human genome, we and others previously identified overlapping yeast artificial chromosome (YAC) clones covering most of this region [8, 12]. These YACs have provided the necessary cloned materials for further characterizing this region of chromosome 4 in detail. However, several features of YAC clones are problematic. First, about half of the inserts in YAC clones from most libraries are chimeric, meaning that they contain non-contiguous segments from different parts of the genome [13, 14]. Second, many YACs have been shown to contain sizable internal deletions, suggesting that other YACs likely contain undetected internal rearrangements or small deletions [8, 12]. Third, obtaining DNA from YAC clones is time-consuming and the yields of DNA from such preparations are low.

To circumvent these problems, we chose to convert the overlapping set (contig) of YACs into smaller and more easily manipulated cosmid clones. One approach for such a conversion would be to make small cosmid libraries from the YACs, which would allow a high representation of cosmids for the corresponding genomic region [15]. However, there are several problems with such an approach. Chimerism and rearrangements, especially small undetected rearrangements, of the YACs would be present in the cosmid clones. In addition, cosmid clones from the ends of the YACs would contain the YAC vector sequences. For these reasons, and because several groups have had success in using YACs as labeled DNA probes on cDNA and cosmid libraries [16, 17], we decided to employ an alternative approach in which YACs are used to screen a chromosome-specific cosmid library. With this approach, chimerism is not a problem because the portions of chimeric YACs from other chromosomes are not represented in the chromosome-specific cosmid library. In addition, because they are derived from a different cloning system, cosmids are unlikely to contain the same deletions and rearrangements that are present in YACs. In this paper, we describe our application of this approach to obtain a large number of cosmid clones, assemble them into contigs and construct a fine-structure restriction map of most of the 2.2 Mbp region. Our results suggest that this approach provides a general strategy for cloning and finely mapping other regions of complex genomes.

While this paper was under review, the gene responsible for Huntington disease was reported [34]. The gene, which covers about 200 kbp of the chromosome, is indeed located in the region predicted based on linkage disequilibrium, and encodes a cDNA that is 11 kbp in length. Thus, while the cosmid contigs reported in our paper are not necessary for initial identification of the HD gene, they provide reagents for studying the function of the gene and for studying other genes and chromosomal elements in the region.

## RESULTS

### Strategy

We based most of our efforts to build cosmid contigs on seven YAC clones representing a continuous, minimum overlapping set in the 1.7 Mbp region between markers *D4S125* and *D4S98*. This segment contains the 0.7 Mbp region of linkage disequilibrium that most likely contains the gene [9]. We also used an eighth YAC clone at

locus *D4S126* to isolate cosmids, but we did not attempt to connect these cosmids with the 1.7 Mbp region, as this segment contained a gap in the YAC map.

Our strategy was as follows. We first screened a chromosome 4-specific cosmid library by using the eight YACs as hybridization probes and identified a large number of cosmid clones that gave positive signals. We transferred these cosmid clones in a gridded array to another nylon filter, and did secondary screens with 26 YAC clones, including the original eight YACs, from the HD region as hybridization probes. Based on knowledge of the physical maps of these YACs, we ordered these cosmids into "bins", defined as groups of cosmids that have common YAC hybridization patterns (Figure 1). DNA was prepared from each cosmid, digested with EcoRI, electrophoresed in agarose gels and analyzed with a combination of ethidium bromide staining and Southern blot hybridization with cosmid vector ends, a large number of previously characterized unique DNA probes from the region, and probes made from many of the cosmids themselves. This analysis allowed us to order the cosmids into several contigs and simultaneously to construct a fine restriction map of the region. Each of the steps of this strategy and the results obtained are described below.

Primary screening of the cosmid library

The cosmids we used for this study are from a flow-sorted human chromosome 4-specific library constructed in the vector sCos1 [18] by the Los Alamos National Laboratory (Materials and Methods). To facilitate the screening of this library, we replicated it and stamped the colonies onto 68 nylon filters, such that each filter contained colonies from four different 96-well microtiter plates. We estimate that the number of colonies used in our screens represents about four equivalents of the 200 Mbp of human chromosome 4 (Materials and Methods).

DNA from the minimum overlapping set of seven YAC clones as well as the YAC for *D4S126* (Figure 1) was radiolabeled, pre-annealed with a large excess of unlabeled human placental DNA to block repetitive sequences, and hybridized to the filters. In general, we observed two types of positive signals, which we designated "strong" and "weak", for each YAC probe (Figure 2A). The primary screening with the seven YAC probes in the minimum YAC contig identified 498 cosmid clones, including 277 with strong signals and 221 with weak signals. In addition, primary screening with the YAC from *D4S126* yielded seven strong signals.

In addition to screening the cosmid library with labeled YAC probes, we employed 48 cosmids that we had previously isolated from the same library by using single-copy probes from loci scattered throughout the region. Forty of these 48 cosmids were present in the set of the 498 cosmid clones that we identified with YAC probes above. The remaining eight of these 48 cosmid clones did not yield positive signals in the primary YAC screen. However, as these eight cosmids did hybridize to YAC probes in the secondary screen, which is described below, it is likely that we initially missed them because of poor growth of the cosmid colonies on the primary nylon filters.

Our combination of primary YAC and single-copy probe screening yielded a total of 506 different cosmids. These include the 277 strong hybridizing clones identified in the YAC screen, the 8 new strong hybridizing clones identified with single-copy probes, and the 221 weak hybridizing clones from the YAC screen.

In addition to these 506 cosmid clones, we isolated 12 overlapping cosmids that represent a contig at locus *D4S125* (Figure 1) [8]. Because we were able to determine early in our efforts that this contig represented a continuous overlapping set that extends to the YAC, yWST57, at the proximal end of the seven overlapping YACs, we did not use the YAC from *D4S125* (yWST55 in Figure 1) to screen the cosmid library.

Secondary screening of the cosmid library

To simplify our subsequent analysis, we transferred the 277 cosmid clones that yielded strong signals in the primary screen, the eight new cosmids that were identified by single-copy probes but missed in the primary screen, and the twelve cosmids we isolated with single-copy probes around *D4S125* to several copies of two new filters. These filters were then hybridized with labeled probes made with DNA from each of 26 YACs from the 1.7 Mbp region, including the original set of seven YACs (Figure 1). In these secondary hybridizations, clearly identified signals were observed for 194 (65%) of the cosmid clones, whereas 103 (35%) cosmid clones gave no signals (Figure 2B).

To determine whether any of the 103 cosmid clones that gave no signals in the secondary screens or the 221 cosmid clones that gave weak signals in the primary screens were true positive clones, we decided to analyze them further. We made small amounts of DNA from each of these clones, digested the DNA with EcoRI, electrophoresed the fragments on agarose gels, transferred the DNA to nylon blots, and hybridized the blots

with labeled DNA probes made from each of the seven YACs in the minimum overlapping set. Through this additional analysis, we determined that only four cosmid clones from the 103 cosmids giving no signal in the secondary screen and only 21 of the 221 cosmids from the "weak" signal category gave positive signals in this secondary YAC screen. The combination of primary and secondary colony screening, as well as the EcoRI-digested cosmid DNA screening, resulted in a final number of 219 cosmid clones that gave positive signals with YAC probes. Our results suggest that the remainder of the cosmid clones with positive signals were the result of screening artifacts, most likely related to vector or low-copy repeat cross-hybridization.

Sublocalization of cosmids by a "binning" strategy

We analyzed the hybridization results from the secondary screens in a manner that allowed us to determine a more refined location for each cosmid, which simplified the subsequent fingerprinting analysis for determining cosmid overlaps. Previous characterization of the 26 YAC clones had identified their order and had provided some information about their degree of overlap. By hybridizing all 26 YAC probes to the cosmids, we could localize the cosmids into "bins", which are defined as groups of cosmids that have common YAC hybridization patterns. For example, cosmids that gave a positive signal with yWST61 and yWST62 (Figure 1) would be assigned to a bin defined by the segments of those two YACs that overlap. These cosmids could be further subdivided into smaller bins defined by their hybridization results with yWST18, yWST65, yWST63, yWST64 and yWST66 (Figure 1).

From the 219 cosmid clones from the primary and secondary YAC screens, we were able to group 174 into 30 bins, each containing from one to fourteen cosmids (Table 1), that were easily reconciled with the known overlapping patterns of the YACs. All 174 of these clones were confirmed to be in contigs from this region of chromosome 4 by subsequent fingerprinting and hybridization experiments (described below). The remaining 45 cosmid clones either hybridized to a single YAC or to multiple non-overlapping YACs. This set of clones defined a large number of small or ambiguous bins that could not be reconciled with the known YAC overlapping patterns. As described below, we subsequently showed that none of these 45 cosmids overlap with the 174 cosmid clones.

## Fine restriction mapping of cosmids by fingerprinting

To determine the overlaps between cosmids within each bin and between adjacent bins, as well as to construct a fine-structure restriction map of the 1.7 Mbp region, we determined the EcoRI restriction patterns for the 219 cosmids. After digestion with the enzyme, the cosmid DNAs were loaded onto agarose gels in an order corresponding to the order of the bins. Following electrophoresis, we blotted the gels and hybridized the blots to oligonucleotide probes corresponding to the phage T7 and T3 promoters. This hybridization allowed us to determine which two EcoRI fragments correspond to the ends of the human DNA insert in each cosmid [18]. We also compared the ethidium-stained banding pattern for each set of cosmids in a bin and in adjacent bins, and cosmids containing two or more bands of equal size were considered to be overlapping. Because cosmids in each portion of the gels were derived from the same small region of the genome, it was much easier to identify overlapping fragments by this visual inspection than if we had loaded the digests on the gels randomly (see Figure 3A for example). The prior mapping of the cosmids to the 1.7 Mbp region, as well as to even more refined locations defined by the bins, increases the likelihood that two EcoRI bands of the same apparent size in two cosmids are due to true overlap. These data allowed us to establish several sets, or contigs, of overlapping cosmids covering the 1.7 Mbp region. In addition, analysis of these data allowed us to determine order information for many of the EcoRI fragments, due to their locations at insert endpoints [19]. The contigs were analyzed further by hybridization and additional restriction enzyme digestion, as described below, to confirm their validity and to determine the orders of additional EcoRI fragments.

## Confirmation of the contigs

We used several additional types of analysis to confirm that the contigs we established by binning and fingerprinting were correct. First, we labeled 52 of the cosmids and, after pre-annealing the probes with human placental DNA, hybridized them to blots containing EcoRI digests of the cosmids in the same and adjacent bins (Table 1). This set of probes included all of the cosmids that had fewer than two overlapping bands on the EcoRI digests, as well as a large number of cosmids spread throughout the region that did not overlap with each other. This analysis allowed several discrepancies to be resolved and provided a large amount of confirmatory data. In addition, many of the EcoRI

fragments could be ordered within the cosmids based on this hybridization data that could not be ordered by using the ethidium-stained fingerprinting data alone (Figure 4).

Second, we hybridized the EcoRI digested cosmids with 37 DNA probes that had been previously shown to be from the HD region. These hybridization results, which are summarized in Figures 1 and 4, not only provided confirmation of the locations and orientations of the contigs, but also established the precise localization of these probes, most of which had only been mapped at lower resolution by pulsed-field gel electrophoresis.

Third, we performed BamH1 digestions on many of the cosmids and analyzed them in the same way we did the EcoRI digestions (data not shown). These results were in agreement with the EcoRI results, and in one region, near *D4S135*, the BamH1 data allowed us to determine unequivocally overlap between cosmids that shared only portions of two very large EcoRI fragments.

Fourth, we compared the EcoRI maps in our contigs with results from maps of two small segments in this region of chromosome 4 that were published by other groups [9, 20]. One of these segments is around *D4S95* and the other is around *D4S43*. In both cases, our results were consistent with the published reports.

Finally, we isolated single-copy probes from near the ends of most of the cosmid contigs, and hybridized these probes to blots containing DNA from the 26 YACs and genomic DNA from somatic cell hybrids and radiation hybrids containing all or part of human chromosome 4 [7, 21]. In every case, the results confirmed that the ends of the contigs are derived from the 1.7 Mbp region of chromosome 4 (data not shown).

Considering only the cosmids that we have definitively mapped to the 1.7 Mbp region of chromosome 4 between *D4S125* and *D4S98*, where the HD gene most likely lies, we now represent the region as a set of 174 cosmids in nine contigs (Table 1). In addition to these cosmids, we identified and restriction mapped 12 cosmids around loci *D4S10* and *D4S126* (Figures 1). Figure 4 shows a composite map of the entire 2.2 Mbp Huntington disease region, with most of these cosmids and their EcoRI restriction sites shown below the map. For convenience, a number of the cosmids in Table 1 that were almost identical in restriction patterns to other cosmids in their vicinity are not included in Figure 4. These cosmids cover over 1.6 Mbp in ordered EcoRI fragments, and the region between *D4S125* and *D4S98* can be represented with a minimal set of 51 cosmids (Figure 4). Restriction patterns and hybridization results for these 51 cosmids are shown in Figure 3.

The 45 cosmid clones that we could not easily place into bins from the YAC screening experiments were used as probes to hybridize to three sets of digests: EcoRI-digested DNA from the 174 cosmids in the nine contigs, DNA from the 26 YACs, and genomic DNA from somatic and radiation cell hybrids containing portions of human chromosome 4. None of these 45 clones hybridized to any of the 174 mapped cosmids. Therefore, these clones either are not derived from this region of chromosome 4 or are derived from the gaps between the contigs. Three of these cosmids (47B5, 65C1, and 234A4) were found to comprise a contig that covers about 50 kbp. These cosmids hybridized to the YACs and hybrid cell lines in a manner that indicates that they are from the 1.7 Mbp region and lie in either Gap 3 or Gap 4 (Figure 1). Thirteen of the 45 cosmids did not hybridize to any of the 26 YACs and gave hybridization results with the hybrid cell lines inconsistent with a location in this region of 4p16.3, and nine of the 45 hybridized to multiple non-overlapping YACs. It is likely that these 22 cosmids were obtained in the primary and secondary screens due to the presence of low-copy repetitive sequences recognized by the YAC probes. The remaining 20 cosmids hybridized to single YACs, and are either false positive clones or are derived from gaps between the contigs.

## NotI and MluI restriction map of the region

In an effort to identify the locations of potential genes in the cosmid contig, we tested the cosmids for the presence of restriction enzyme cleavage sites containing the dinucleotide CpG, which often occur in clusters adjacent to and within genes. We determined the locations of the NotI and MluI sites in the minimum overlapping set of 51 cosmids by a combination of single and double digests (data not shown). These experiments not only allowed us to locate these rare-cutter sites, but also ordered several of the EcoRI sites that were ambiguous from the fingerprinting results. We identified 21 NotI and 21 MluI restriction sites, which include all of the sites defined by pulsed-field gel electrophoresis [4, 6, 7, 8], as well as additional sites that were missed by PFGE (Figure 4).

## Refinement of the orders of EcoRI fragments by partial restriction digests

Analysis of the complete EcoRI digests, the hybridization results in which cosmids were used as probes, and the NotI/EcoRI double digests of the cosmid clones allowed us to establish the order of about half of the EcoRI fragments in the 1.7 Mbp region. To confirm

these orders and to determine the order of most of the remaining EcoRI fragments, we used partial digestion and "indirect end-labeling" [18, 22] (Materials and Methods). These experiments allowed us to determine unequivocally the orders of almost all of the EcoRI fragments in the cosmids tested in this way. These results confirmed the fragment orders that we had determined above, and also established orders for many of the EcoRI fragments that were not obtained by the complete digestions. The results of all the analyses for determining EcoRI fragment order are summarized in Figure 4.

Estimation of the sizes of the gaps between the cosmid contigs

By determining the total length of all the EcoRI fragments, we estimate that less than 200 kbp of the 1.7 Mbp region is not represented in the cosmid contigs. A comparison of the distances between pairs of single copy DNA probes in the cosmid contig versus the long range restriction map of the region were also consistent with this estimation. These results indicate that the average size of the eight gaps in the map is about 25 kbp. We were able to determine the sizes of some gaps more accurately by performing additional hybridizations and comparing our results to published data. We found that Gap #1 is less than 15 kbp, as end probes from the two flanking cosmids recognize the same 15 kbp SphI band, 20 kbp ClaI band and 50 kbp SalI band in yWST56 and yWST57 (Figure 1). In addition, we found that Gap #2 is less than 20 kbp, as the EcoRI fragments at the insert ends of the two cosmids flanking the gap share the same 20 kbp XhoI fragment in yWST56 and yWST58 (Figure 1). In agreement with these results, the recent publication of the isolation of the HD gene identified a single cosmid that bridges our Gap #2 (ref). Finally, from a published EcoRI restriction map of the region around *D4S43*, we were able to determine that the size of Gap #6 is less than 10 kbp (Figure 1) [20].

To determine whether the labeled YAC probes missed some cosmids in the region in our initial hybridizations, we isolated single copy probes from near the ends of each of the cosmids flanking each of the eight gaps, and used them to screen the cosmid library. None of these probes identified any new cosmids, indicating that our initial screenings probably identified all the cosmids in the library that are homologous to the YACs. We do not know whether the segments are missing in the library because of statistical variation in distribution or because they are difficult to clone in cosmid vectors. The fraction of the 1.7 Mbp region that we did not obtain in cosmid clones is about 10% of the total length, and the fraction expected to be missing in a four-fold redundant library is about 2% [23]. In

preliminary experiments, we have identified bacteriophage λ clones that contain DNA present in at least some of the gaps (data not shown), suggesting that screening genomic clones in an alternative vector cloning system provides one way to fill gaps in this type of contig map.

## DISCUSSION

We have isolated more than 1.6 Mbp of the human genome as a set of overlapping cosmid clones and established a high-resolution restriction map of the region. Both the clones and the mapping information provide potentially valuable tools for more detailed analysis of this region of the genome, including searches for genes and polymorphic probes. Indeed, during the construction of the contigs, we had begun to use the cosmids to isolate and map a number of candidate cDNAs for Huntington disease, as well as to identify new probes that recognize polymorphic variation for use in genetic linkage analysis of the disease. These clones provide manageable DNA segments for studying the genomic structure and control regions of the HD gene now that it has been identified [34]. In addition, the clones provide reagents for identifying and studying the apparently large number of other genes in the region [35].

In addition to their uses specific to HD, the cosmid clones and restriction map will likely be useful as reagents for determining the complete nucleotide sequence of the region. While sequencing of large regions of the human genome could proceed by using YAC clones as the source of DNA template, we believe that smaller clones are more attractive for this use. Cosmid DNA is much easier to isolate in large quantities than is YAC DNA, and bacterial episomal DNA preparations are more amenable to automation. It is easier to determine high-resolution restriction mapping information and to confirm whether clones represent the genome faithfully with cosmids than with YACs. This type of mapping information is critical for most sequencing strategies. While we do not have definitive evidence that cosmids are more stable than YACs, the cosmids we studied here showed no evidence for chimerism and little suggestion of rearrangements, in sharp contrast to the YAC clones. Thus, it seems likely that large-scale efforts to sequence Mbp segments of the human genome will utilize "sequence-ready" maps and clones, such as those described here.

Other reports have described the construction of cosmid or bacteriophage clone contigs and high-resolution restriction maps of entire genomes [19, 24, 25, 26] or human

chromosomes [27]. These studies have employed "bottom-up" strategies, in which fingerprinting or other restriction mapping data from each clone is compared with that from every other clone. This type of approach requires many pairwise comparisons, very precise restriction digestion measurements, and a high level of redundancy and overlap between clones to obtain accurate contigs. By contrast, the "top-down" strategy we used focuses on cosmid clones that are derived from a specified region of the genome, and utilizes previously determined, highly accurate mapping information to sublocalize the cosmids into bins within the region. This information dramatically decreases the number of pairwise comparisons that must be made, allows the use of simplified restriction enzyme analysis, and decreases the amount of redundancy and overlap required. Although we have not performed statistical analysis to estimate likelihoods of true overlap based on the minimum number of two EcoRI fragments of the same size, we tested a large number of our putative overlapping clones by independent hybridization experiments, and in every case, the overlaps were confirmed. While we found that visual examination of the EcoRI fragment sizes was simple and accurate enough to analyze overlaps of cosmids derived from the same region, the efficiency of this type of analysis could be improved by using computerized image analysis for contig assembly.

An additional feature of using a top-down approach to build cosmid contigs is that the prior mapping information can be used to determine the positions and orientations of contigs that are not continuous with one another. The sizes of gaps remaining after localizing and orienting the cosmid contigs are usually easy to determine, which increases the immediate usefulness of the clones and provides information for directing efforts to fill the gaps.

Our results indicate that the efficacy of screening chromosome-specific cosmid libraries with labeled YAC probes that are pre-annealed with excess unlabeled human DNA to block repeat sequences is very high. Among a set of 48 cosmids that we isolated from the HD region with single-copy probes, all were also identified in our primary or secondary YAC screens. These results suggest that, with proper growth of cosmid colonies on the screening filters, this procedure very likely identifies all or almost all the cosmids in a library that are complementary to the YACs.

While there was some subjectivity in assigning cosmid colonies as "strong" and "weak" in the YAC screens, it is instructive to consider our final results with these two types of colonies. More than half of the strong signal cosmids were true positives in that they were eventually shown to be derived from the 1.7 Mbp region of chromosome 4.

Therefore, a significant number of strong signal cosmids are false positives or fall into gaps in the contigs. However, the results for cosmids that gave weak signals were in striking contrast. Of 221 weak signal clones identified in the primary screen, 21 were positive in the secondary YAC screen. Of these, only seven were found by subsequent analysis to be derived from the 1.7 Mbp region. The entire length of DNA covered by these seven clones was present in cosmids that we identified as strong signals in the YAC screen, so no additional information was derived from the weak signal clones. These results, and the fact that the strong hybridization signals appear to identify all the cosmids in the library complementary to the YACs, indicate that, for future applications of this procedure, weak signal clones should not be considered further and that effort should be spent only on strong signal clones. Subsequent work in our laboratory with other regions of the genome has corroborated this conclusion.

Besides obtaining binning information and EcoRI fingerprints for each of the cosmids, we performed several additional tests during the building of the contigs. These included using whole cosmids and ends of cosmids as hybridization probes, partial restriction digestion mapping, and fingerprinting with an additional restriction enzyme. While the results of these tests were important for determining the orders of the EcoRI restriction fragments, much of the information was not necessary for determining the amount of overlap between and the orders of the cosmids. Thus, in cases where contigs of cosmids are desirable but a complete restriction map is not critical, it should be possible to streamline the process and use only those tests that provide enough information to build accurate contigs.

With the approach we used here, it was possible to generate cosmid contigs covering over a million base pairs of a human chromosome. We estimate that the amount of labor required to do this work was less than one person-year, and that the lessons we learned during this effort will likely considerably decrease the amount of work for subsequent efforts. Some regions of the genome may be difficult to obtain in cosmids by using this method in which YACs are used as hybridization probes. For instance, cosmids from regions of the genome that contain a large number of low-copy repeats may not be easily scored by YAC hybridization due to difficulties in competing the repetitive sequences with an large excess of unlabeled human DNA. Even if this is the case, it is likely that this or similar approaches will be successful for much of the genome, and that other methods could be applied to recalcitrant regions. Indeed, our recent experience in applying this approach successfully to several other regions of the human genome suggests that it is generally applicable.

## MATERIALS AND METHODS

### Cosmid library

The chromosome 4-specific cosmid library that we used in this work was constructed at the Human Genome Center at the Los Alamos National Laboratory (Los Alamos, NM 87545) under the auspices of the U.S. Department of Energy, and was kindly provided by Drs. Larry Deaven and Jon Longmire. The library was constructed from partial Sau3A digests of chromosome 4 DNA flow-sorted from hamster-human hybrid cell line UV20HL21-27, which contains human chromosomes 4, 8 and 21. The inserts were cloned into the BamH1 site of the sCos-1 cosmid vector [18] and propagated in *E. coli* strain HB101. The original library consists of 25,920 independent colonies in 270 microtiter plates, and about 94% of the clones contain human DNA inserts. The entire library was replicated onto 68 nylon filters such that each filter contained a total of 384 colonies derived from four microtiter plates. We used four copies of this set of 68 filters for our studies. Based on an average insert size of 38 kbp, and a loss of about 10% of the colonies during our replication process, we estimate that the number of colonies that we screened represents about four equivalents of the 200 Mbp chromosome 4.

### Cosmid library screening with YAC probes

YACs were electrophoresed on low-melting agarose pulsed-field gels and excised for use as probes. About 20-50 ng of gel-purified YAC DNA was labeled by random-priming and pre-annealed in 250 µl 100 mM $NaH_2PO_4$ containing 0.5 to 1 mg unlabeled human placental DNA at 65°C for 10 to 20 hours. With this treatment, only those sequences in the YACs that are single-copy or very low copy number in the human genome remain single stranded and are available to hybridize to the cosmid DNA on the filters [28]. These pre-annealed YAC probes were hybridized to the 68 cosmid library filters at 65°C for 16 to 24 hours in Hybridization Solution (1 M NaCl, 10% dextran sulfate, 1% SDS). Filters were washed at 65°C with 0.2X SSC and 0.1% SDS for two times 30 minutes each and exposed to Kodak XAR film at -70°C for three hours to two days.

Although three of the seven YACs that we used for primary screening migrated with endogenous yeast chromosomes on the pulsed-field gel used to purify the YACs, we

found that the inclusion of yeast DNA in the hybridization probes did not affect the quality of the screening results.

Cosmid DNA preparation, Southern blotting and single-copy DNA probes

Small preparations of cosmid DNAs were made from 1.5 ml cultures by alkaline lysis [29]. A portion of this DNA was digested with restriction enzymes, electrophoresed on 0.5% to 1% agarose gels, ethidium-stained, and transferred to nylon membranes for Southern blot analysis. Ethidium images of the gels were recorded and the sizes of the DNA fragments in each of the restriction digests were determined by comparing their mobilities to standards.

To prepare probes from cosmids, we labeled 50 ng of cosmid DNA by random priming. These probes were pre-annealed and hybridized to blots prepared by transferring DNA from gels to Hybond-N+ nylon membranes with the alkaline transfer method or to the 68 cosmid library filters. All hybridizations were performed as described for YAC probes.

Most of the single-copy DNA probes used in this study are described in references 8, 33, 9 and 4. Probes D4S181*1, and D4S10*2 were PCR products that we generated from human genomic DNA based on published sequences [30]. Probes L19ps11, L19ps14, Y12Eco2.3, R4ps17, S1.5, LCD2, p309, p359, p363, p358 and C102/2.7 were kindly provided by Dr. Marcy MacDonald and are described in references 33, 9 and 4. Probe 270a is described in reference 7, and probes STS4-487 and STS4-558 are described in reference 32.

Partial restriction digestion analysis to determine EcoRI site order

We used the "indirect end-labeling" method [18, 22] to establish orders of EcoRI fragments in the cosmid contig. We digested DNA from half of the minimal set of 51 cosmids with NotI, which cleaves in the sCos1 vector about 100 bp from both sides of the insert cloning site. We then treated these NotI digestions with low amounts of EcoRI for various lengths of time to generate sets of partially digested inserts. These digests were electrophoresed on 0.5% agarose gels, transferred to nylon membranes, and hybridized sequentially with 32P end-labeled T3 and T7 oligonucleotide probes, which recognize each of the segments of the cosmid vector near the insert cloning site [18]. Hybridizations were performed at 42°C in Church Buffer [31], and the blots were washed two times sequentially

with 6X SSC/0.1% SDS, 2X SSC/0.1% SDS and 0.2X SSC/0.1% SDS at 42°C for 30 minutes. Blots were exposed to X-ray film for several hours.

## ACKNOWLEDGMENTS

We thank Larry Deaven and Jon Longmire of Los Alamos National Laboratory for providing the chromosome 4 cosmid library and for their advice, Marcy MacDonald, Jim Gusella, John Wasmuth, and members of the UCSF Human Genome Mapping Center for DNA probes, Francis Collins and John Wasmuth for providing us with several of the YAC clones, Guy diSibio and Kristen Smith for help in replicating the cosmid library, and members of our laboratory for their support and useful discussions. An expanded version of Figure 4 and the cosmid clones reported in this work are available upon request. This work was supported by grants from the NIH (2RO1NS26237) and the Wills Foundation.

## ABBREVIATIONS

YAC: yeast artificial chromosome; PFGE: pulsed-field gel electrophoresis; PCR: polymerase chain reaction; HD: Huntington disease; Mbp: million base pairs; kbp: kilobase pairs.

## REFERENCES

1. Hayden, M.R. (1981) Huntington's chorea. Springer-Verlag, Berlin.

2. Martin, J.B. and Gusella, J.F. (1986) New Eng. J. Med., **315**, 1267-1276.

3. Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E. and et al. (1983) Nature, **306**, 234-238.

4. Bates, G.P., MacDonald, M.E., Baxendale, S., Youngman, S., Lin, C., Whaley, W.L. and et al. (1991) Am. J. Hum. Genet., **49**, 7-16.

5. Snell, R.G., Thompson, L.M., Tagle, D.A., Holloway, T.L., Barnes, G., Harley, H.G., Sandkuijl, L.A., MacDonald, M.E., Collins, F.S., Gusella, J.F., Harper, P.S. and Shaw, D.J. (1992) Am. J. Hum. Genet., **51**, 357-362.

6. Bucan, M., Zimmer, M., Whaley, W.L., Poustka, A., Youngman, S., Allitto, B.A. and et al. (1990) Genomics, **6**, 1-15.

7. Pritchard, C.A., Casher, D., Uglum, E., Cox, D.R. and Myers, R.M. (1989) Genomics, **4**, 408-418.

8. Zuo, J., Robbins, C., Taillon-Miller, P., Cox, D. and Myers, R.M. (1992) Hum. Molec. Genet., **1**, 149-159.

9. MacDonald, M.E., Lin, C., Srinidhi, L., Bates, G., Altherr, M., Whaley, W.L. and et al. (1991) Am. J. Hum. Genet., **49**, 723-734.

10. Snell, R.G., Lazarou, L., Youngman, S., Quarrell, O.W.J., Wasmuth, J.J., Shaw, D.J. et al. (1989) Med. Genet., **26**, 673-675.

11. Theilmann, J., Kanani, S., Shiang, R., Robbins, C., Huggins, M. et al. (1989) Med. Genet., **26**, 676-681.

12. Bates, G.P., Valdes, J., Hummerich, H., Baxendale, S., Le Paslier, D.L., Monaco, A.P., Tagle, D., MacDonald, M.E., Altherr, M., Ross, M., Brownstein, B.H., Bentley, D., Wasmuth, J.J., Gusella, J.F., Cohen, D., Collins, F. and Lehrach, H. (1992) Nature Genet., **1**, 180-187.

13. Green, E.D., Riethman, H.C., Dutchik, J.E. and Olson, M.V. (1991) Genomics, **11**, 658-669.

14. Schlessinger, D., Little, R.D., Freije, D., Abidi, F., Zucchi, I., Porta, G., Pilia, G., Nagaraja, R. et al. (1991) Genomics, **11**, 783-793.

15. Bellanne-Chantelot, C., Barillot, E., Lacroix, B., Le Paslier, D. and Cohen, D. (1991) Nucleic Acids Res., **19**, 505-510.

16. Baxendale, S., Bates, G.P., MacDonald, M.E., Gusella, J.F. and Lehrach, H. (1991) Nucleic Acid Res., **19**, 6651.

17. Elvin, P., Slynn, G., Black, D., Graham, A., Butler, R., Riley, J., Anand, R. and Markham, A.F. (1990) Nucleic Acid Res., **18**, 3913-3917.

18. Evans, G.A., Lewis, K. and Rothenberg, B.E. (1989) Gene, **79**, 9-20.

19. Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. and Frank, T. (1986) Proc.Natl.Acad.Sci.USA, **83**, 7826-7830.

20. Gillam, T.C., Bucan, M., MacDonald, M.E., Zimmer, M., Haines, J.L., Cheng, S.V., Pohl, T.M., Myers, R.H., Whaley, W.L., Allitto, B.A., Faryniarz, A., Wasmuth, J.J., Frischauf, A., Conneally, P.M., Lehrach, H. and Gusella, J.F. (1987) Science, **238**, 950-952.

21. Cox, D.R., Pritchard, C.A., Uglum, E., Casher, D., Kobori, J. and Myers, R.M. (1989) Genomics, **4**, 397-407.

22. Smith, H.O. and Birnstiel, M.L. (1976) Nucleic Acid Research, **3**, 2387-2398.

23. Abidi, F.E., Wada, M., Little, R.D. and Schlessinger, D. (1990) Genomics, **7**, 363-376.

24. Kohara, Y., Akiyama, K. and Isono, K. (1987) Cell, **50**, 495-508.

25. Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986) Proc.Natl.Acad.Sci.USA, **83**, 7821-7825.

26. Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J. and Waterston, R. (1991) BioEssays, **13**, 413-417.

27. Tynan, K., Olsen, A., Trask, B., de Jong, P., Thompson, J., Zimmermann, W., Carrano, A. and Mohrenweiser, H. (1992) Nucleic Acids Res., **20**, 1629-1636.

28. Lewin, B.M. (1980) Gene Expression. John Wiley & Sons, Inc. Vol. 2, Second Edition, pp. 503-530.

29. Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) In: Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, pp. 1.25-1.28.

30. Gusella, J.F., Altherr, M.R., McClatchey, A.I., Doucette-Stamm, L.A., Tagle, D., Plummer, S., Groot, N., Barnes, G., Hummerich, H., Collins, F.S., Housman, D.E., Lehrach, H., MacDonald, M.E., Bates, G. and Wasmuth, J.J. (1992) Genomics, **13**, 75-80.

31. Church, G.M. and Gilbert, W. (1984) Proc.Natl.Acad.Sci.USA, **81**, 1991-1995.

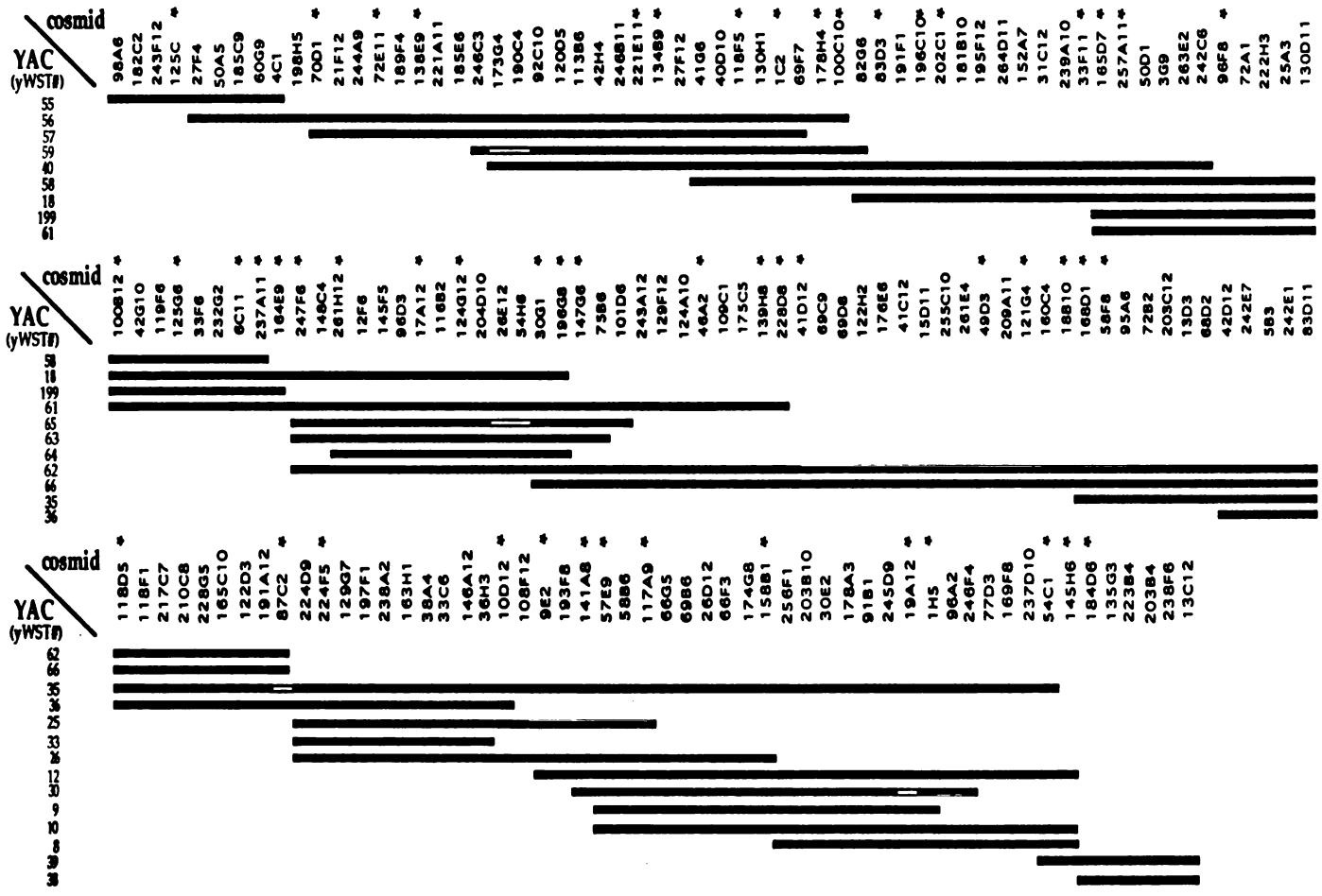
32. Goold, R.D., diSibio, G.L., Xu, H., Lang, D.B., Dadgar, J.B., Magrane, G.G., Dugaiczyk, A., Smith, K.A., Cox, D.R., Masters, S.B. and Myers, R.M. (1993) Genomics, submitted.

33. Whaley, W.L., Bates, G.P., Novelletto, A., Sedlacek, Z., Cheng, S., Romano, D., Ormondroyd, E., et al. (1991) Somat. Cell Mol. Genet., **17**, 83-91.

34. The Huntington's Disease Collaborative Research Group. (1993) Cell. **72**, 971-983.

35. Carlock, L., Wisniewski, D., Lorincz, M., Pandrangi, A., and Vo, T. (1992) Genomics **13**, 1108-1118

**Table 1: Binning Strategy**




Note:
1. YACs are labeled according to nomenclature designated by the Center for Genetics in Medicine of Washington University (e.g. yWST64) and their sizes are indicated in references 8 and 12.
2. Cosmids are labeled according to their microtiter well positions in the Los Alamos National Laboratory arrayed chromosome 4 cosmid library.
3. Filled bars indicate positive signals obtained when cosmids were hybridized during primary and secondary screening with corresponding YAC probes; Blank bars indicate negative signals which are likely due to either deletions in the YACs or false negative hybridization in the YAC screens.
4. Cosmids designated by asterisks were used as labeled probes to confirm their overlapping patterns to the adjacent cosmids.

## FIGURE LEGENDS

**FIGURE 1:** Composite map of the HD region on the short arm of human chromosome 4. The open square indicates the 4p telomere. At the top of the figure, several key loci are labeled by their D4S numbers, followed by the names of the probes corresponding to the loci (see text). The HD region is indicated by an arrowed line below the composite map. The thick hatched line within the arrowed line indicates a 0.7 Mbp region that is in linkage disequilibrium with the HD mutation. YACs used for primary screens and for binning are drawn as lines according to their sizes from the HD region on chromosome 4 [8] and are labeled with numbers that correspond to their yWST (Washington University Center for Genetics in Medicine) names (see reference 8). Cosmid contigs and gaps in the *D4S125* to *D4S98* region, which is where we concentrated most of our efforts, are shown at the bottom of the figure. The contigs are named according to numbers shown above the lines, and the gaps between the contigs are numbered below the lines. Most of these cosmids were obtained by screening a cosmid library with YAC probes; however, the contig around *D4S125* was obtained by screening the library with single copy plasmid probes. Additional cosmids were obtained around loci *D4S10* and *D4S126* by screening with single copy probes and yWST47, although no attempt was made to connect these contigs with cosmid contigs 1-9 due to the lack of YAC clones covering this region.

**FIGURE 2:** Cosmid library screening with YAC probes.

A. Representative autoradiogram from a primary screening experiment of the arrayed cosmid library with yeast artificial chromosome (YAC) inserts as labeled DNA probes. The filter contained 384 cosmid colonies in an array from four different microtiter plates (plates 145 to 148) of the Los Alamos National Laboratory chromosome 4-specific cosmid library. The filter was screened with YAC yWST62 (D42F11), which is one of the seven minimum overlapping YACs from the HD region. A strong positive signal from position 145F5, a strong positive signal from 147G6, and several weak positive signals from 145C1, 147C1, 148D1 and 146E1 are observed on this autoradiogram. Further analysis showed that the two cosmids with strong signals were located within the HD region where the YAC was mapped, while other cosmids with weak signals were not derived from the region.

B. Autoradiogram from a secondary screen in which a YAC was used as a labeled DNA probe on a nylon filter containing cosmids that showed strong positive signals in the

primary screen. 192 cosmid colonies from primary screens and from single copy DNA probe hybridization results were replicated onto this filter in a 2 X 96 array. This filter was hybridized with a YAC (yWST59/B134B4) from the HD region that was not used in the primary screen. Nineteen positive signals were observed on this autoradiogram. Among them, three (positions 1A5, 1E9 and 1C10 corresponding to positions 64C3, 178H4 and 244A9 in the primary cosmid library) were relatively weak, but further analysis showed that 178H4 and 244A9 are located within the HD region. We determined that the weak signal between columns 10 and 11 and rows D and E does not align with a colony and is therefore a hybridization artefact.

**FIGURE 3:**

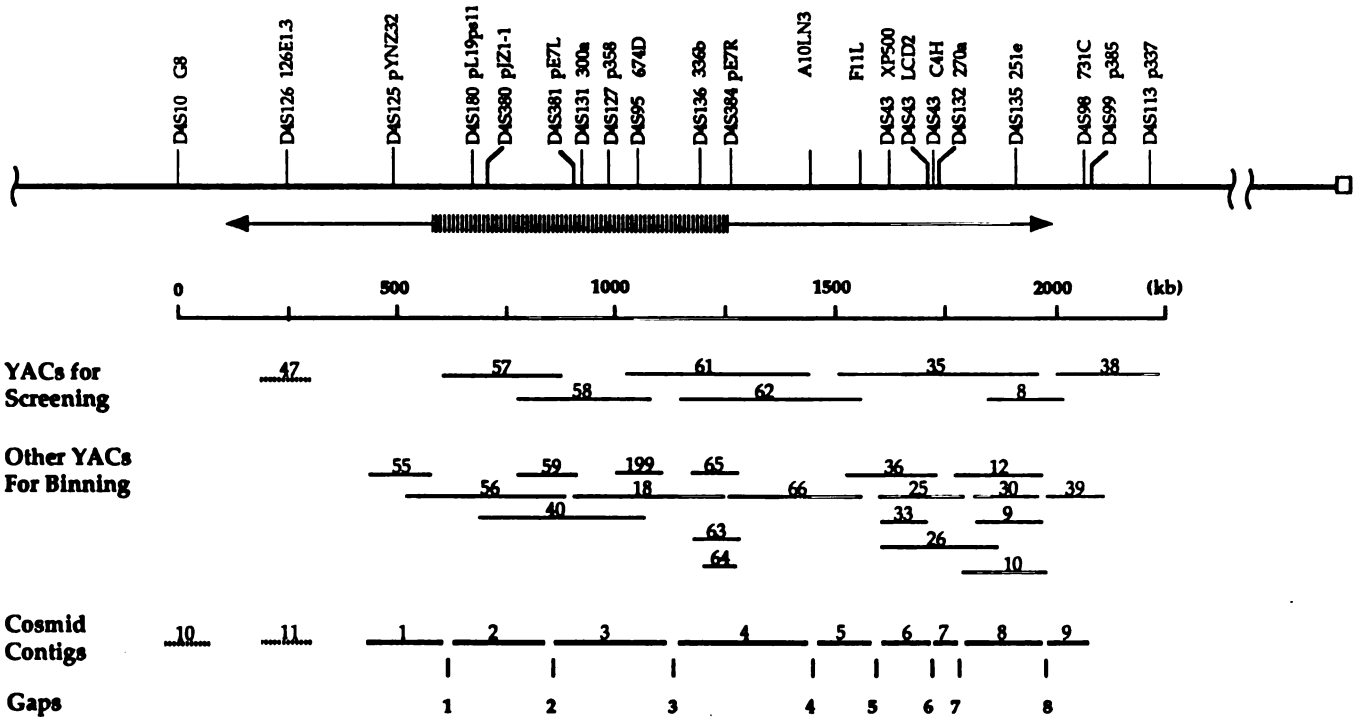
A. Ethidium image of an agarose gel containing EcoRI-digested DNAs from the 51 cosmids representing the minimum overlapping set between loci *D4S125* to *D4S98*. Each cosmid DNA was digested with EcoRI to completion and electrophoresed in an 0.8% agarose gel. The size markers are the 1 kbp ladder from Bethesda Research Laboratories. The EcoRI maps of these cosmids are shown in Figure 4.

B. Autoradiogram showing results of a labeled cosmid probe hybridized to a blot of the gel shown in (A). DNA in the gel was transferred to a nylon filter and the filter was hybridized with a pre-annealed probe (see Materials and Methods) of labeled cosmid 165D7. Only the two adjacent overlapping cosmids (33F11 and 257A11) hybridized with this cosmid probe. The 6.7 kbp band in each lane is the result of hybridization of the cosmid vector in the probe that was not blocked by the pre-annealing treatment.

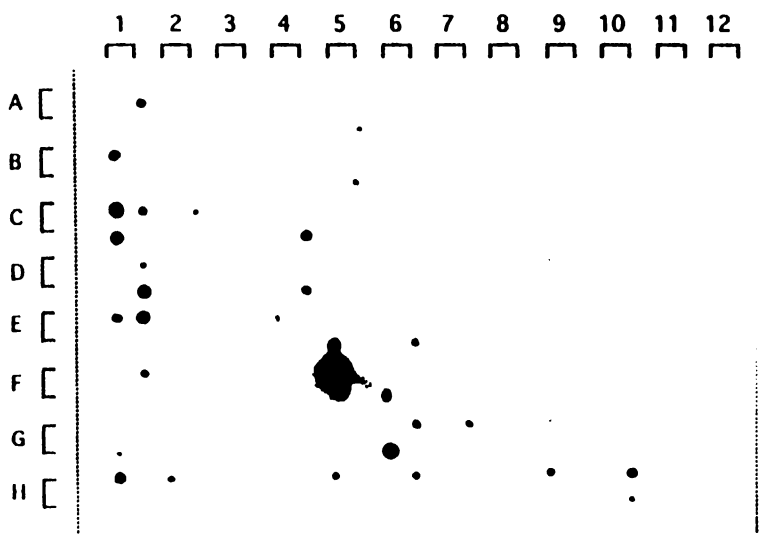
**FIGURE 4:**

EcoRI, NotI, MluI restriction map of cosmid contigs in the HD region. Short vertical lines at the top of the composite map represent EcoRI sites and the size in kbp of each EcoRI fragment is indicated between two adjacent sites. Vertical lines with open circles indicate NotI sites, and vertical lines with open diamonds indicate MluI sites. Thick vertical lines represent precise locations of single copy probes, including D4S numbers in bold and probe names in parentheses, used in this study. Hatched horizontal lines below the composite map indicate positions where the orders of the EcoRI fragments were not

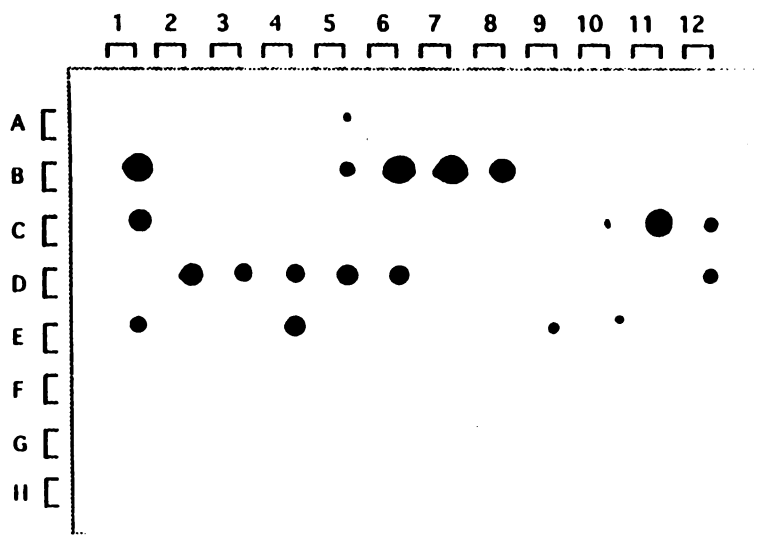
determined. Wavy vertical lines on the composite map indicate gaps and the ends of the cosmid contigs. Below the composite map, each horizontal line represents a single cosmid from the chromosome 4-specific cosmid library with the cosmid microtiter plate and well position on the left. One cosmid, 125C, which hybridizes to probe *D4S125*, was isolated from a pooled version of the arrayed chromosome 4 cosmid library, and we did not determine the microtiter plate position of this cosmid. Short lines above the cosmids represent EcoRI restriction sites. End fragments of cosmid inserts at the T3 side of the cosmid vector are labeled as open squares, and T7 end fragments are designated as filled squares. Horizontal lines lacking squares at both ends represent those cosmids for which the locations of the end fragments were not determined, whereas the lines with a square at only one end represent those cosmids for which data was not obtained for the other end. The minimum set of 51 overlapping cosmids are indicated as bold horizontal lines. About 20 of the cosmids that we characterized from the region were found to have restriction patterns almost identical to other cosmids in their vicinity, and are not included in this figure.

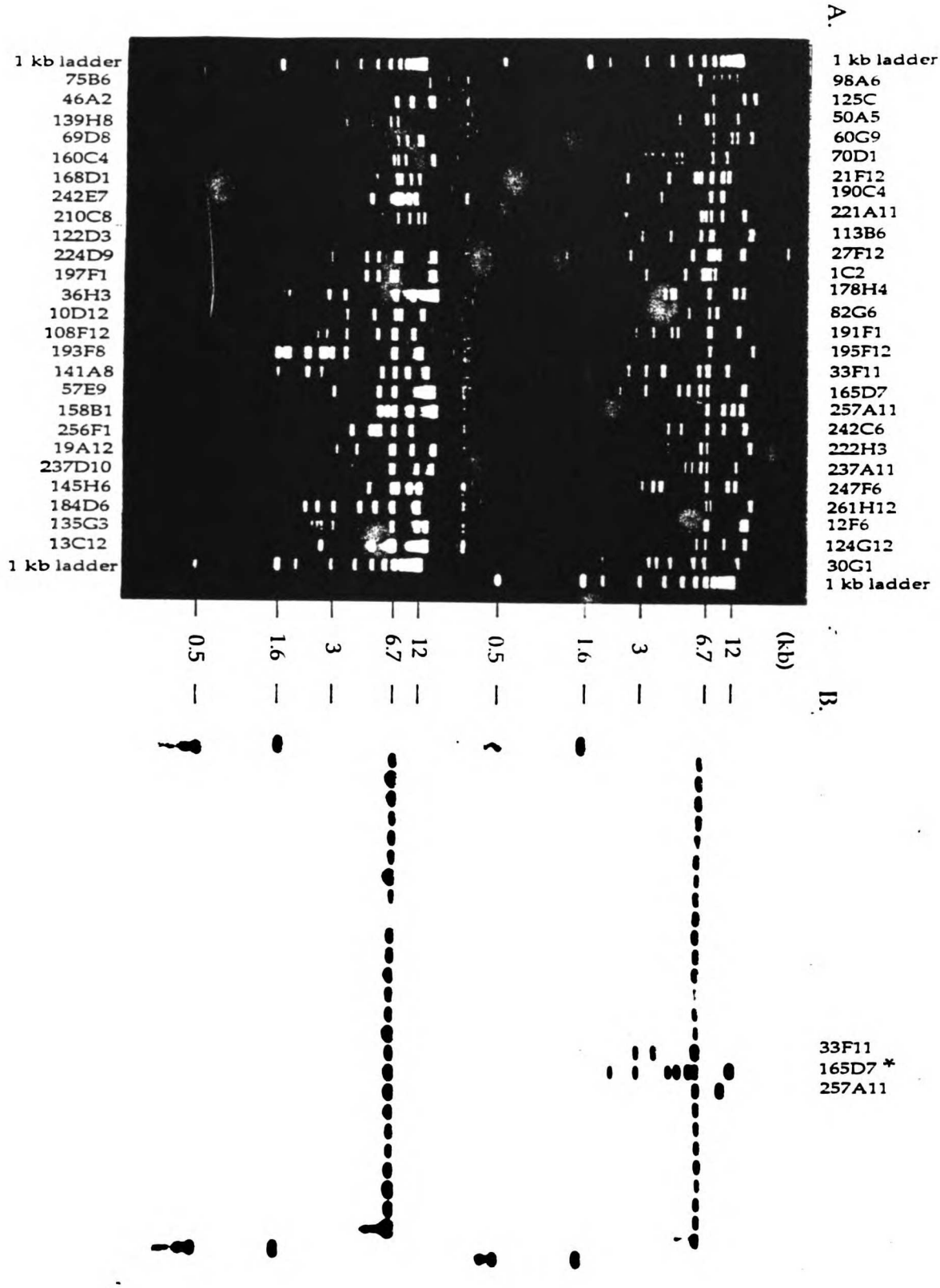

DMS10 G8
DMS126 126E1.3
DMS125 pYNZ32
DMS180 pL19ps11
DMS380 pJZ1-1
DMS381 pE7L
DMS131 300a
DMS127 p358
DMS95 674D
DMS136 336b
DMS384 pE7R
A10LN3
F11L
DMS43 XP500
DMS43 LCD2
DMS43 C4H
DMS132 270a
DMS135 251e
DMS98 731C
DMS99 p385
DMS113 p337
0
500
1000
1500
2000
(kb)
YACs for Screening
47
57
61
35
38
58
62
8
Other YACs For Binning
55
59
199
65
36
12
56
18
66
25
30
39
40
33
9
63
26
64
10
Cosmid Contigs
10
11
1
2
3
4
5
6
7
8
9
Gaps
1
2
3
4
5
6
7
8

A.




B.




A.
1 kb ladder
75B6
46A2
139H8
69D8
160C4
168D1
242E7
210C8
122D3
224D9
197F1
36H3
10D12
108F12
193F8
141A8
57E9
158B1
256F1
19A12
237D10
145H6
184D6
135G3
13C12
1 kb ladder
1 kb ladder
98A6
125C
50A5
60G9
70D1
21F12
190C4
221A11
113B6
27F12
1C2
178H4
82G6
191F1
195F12
33F11
165D7
257A11
242C6
222H3
237A11
247F6
261H12
12F6
124G12
30G1
1 kb ladder
B.
(kb)
12
6.7
3
1.6
0.5
33F11
165D7 *
257A11


Contig 1
Contig 2
Gap 1
Contig 2
Contig 3
Gap 2
Contig 3
Contig 4
Gap 3
10 kb


Contig 4
Contig 5
Contig 6
Contig 7
Contig 8
Contig 9
Contig 10
Contig 11
Gap 4
Gap 5
Gap 6
Gap 7
Gap 8
(A16L)
(F11L)
160C4
121C4
148D1
228C5
224D9
129C7
197F1
166C10
122D3
191A12
87C2
163H1
238A2
38A4
33C6
36H3
46A2
179C5
139H3
118F1
122H2
217C7
210C8
16D12
148F12
193F8
982
141A8
38A4
33C6
36H3
57E9
245D9
19A12
96A2
246F4
169F8
77D8
237D10
145H16
256F1
203B10
30E2
178A3
91B1
104D6
139C3
203B4
223B4
230F6
13C12
145H16
90A4
51C4
192G7
889
220H7
239C10
220D11
27G11
40D11
243A1
226F1

## SUMMARY AND PERSPECTIVES

This work describes the detailed physical mapping of an important genomic region of the human genome, the Huntington disease (HD) region. As part of a group effort to clone the HD gene by a positional cloning strategy, I primarily worked on: 1) cloning of the HD region in yeast artificial chromosomes and 2) construction of cosmid contigs and a high-resolution restriction map of the HD region.

Cloning of the HD region required isolation of a large number of DNA markers, long range restriction mapping of these probes, utilization of the yeast artificial chromosome (YAC) cloning system, isolation and characterization of a large number of YACs from this over 2 Mbp region.

To isolate DNA markers, we utilized a somatic hybrid cell line, C25, which contained a small portion of the chromosome 4 short arm, where the HD gene is located. A Lambda phage library was constructed from this cell line. The phage clones that hybridized to total human DNA probes were separated from hamster background. I further mapped many of these clones by utilizing several different somatic hybrid cell lines, which contained various portions of the telomeric portions of the short arm of chromosome 4. Two of these probes were therefore identified to be from the HD region.

The prerequisite for physical mapping of the HD region was to localize all of the DNA probes in a long range restriction map. To directly map these probes, I utilized the pulsed-field gel electrophoresis (PFGE) technique. Genomic DNA was first digested with rare cutting enzymes, such as NotI, MluI, NruI and CspI, by a combination of single and double digests, and was then separated by

PFGE. The hybridization patterns of these DNA probes were then compared to each other, and a long range restriction map of the HD region was constructed.

After the locations of these DNA markers were known, we developed a large number of PCR based sequence-tagged-sites (STS). These STSs were then used to screen the total human YAC libraries at Washington University. We obtained 28 YACs from these screens, and further characterized them by determining their lengths of inserts, their probe-content, and by isolating and characterizing their ends. This maneuver resulted in a complete coverage of the 2 Mb HD region between D4S125 to D4S98. These 28 YACs provided 2.3 fold redundant coverage of the HD region and were the necessary material for identification of polymorphic markers, cDNA clones and further characterization of this region.

Since the YAC clones are subject to rearrangements, including deletions and inversions, and about half of the YACs contained non-continuous segments from different parts of genome, it was necessary to convert the YAC clones into smaller clones, such as cosmids and phages. We used cosmids to construct a high-resolution restriction map of the entire HD region, and to provide a sequence-ready clone source for the future sequencing of this portion of the human genome. We developed a method to isolate cosmids from the HD region by screening a chromosome specific cosmid library with YAC probes. Based on the overlapping pattern of the characterized YACs, we were able to put the cosmids into bins and further determine their restriction maps. An EcoRI restriction map of 90% of the HD region was constructed, and all the available DNA markers from the HD region were mapped to the corresponding EcoRI fragments. Furthermore the sites of NotI and MluI were determined by double restriction digestion in combination with EcoRI. This high resolution map of the

D region provided easily manipulated cloned DNA material for characterizing e candidate genes for HD.

One goal of the human genome project is to determine the sequence of the ntire human genome. With present sequencing technology, it is desirable to rst obtain sequencible clones to cover the genome. The method we have eveloped in this study should be generally usable to convert YAC contigs into osmid contigs with a simultaneous construction of a high resolution restriction ap of genomic regions of interest.

**uture directions**

. Huntington disease

Understanding the mechanism of the mutation causing HD and providing ure for the HD patients will represent the future challenges for this devastating sease.

In the near future, the survey over large number of HD families and ormal individuals will establish the distribution of the copy number of the CAG peat. This information will provide clues for understanding the nature of this enetic defect.

Furthermore careful studies on the linkage disequilibrium and haplotypes the large number of HD patients may pinpoint the exact locations of the under mutations in respect to the CAG repeat.

Studies on the HD chromosomes without CAG expansion will potentially ovide genetic evidence to prove if IT15 is the true transcript for HD. Searching r point mutations in this gene, genotyping in its vicinity and linkage studies on

those "inconsistent" HD families will represent the next challenge in genetic studies of HD.

Creating mouse models for HD by homologous recombination or gene knock-out will potentially provide proof for the HD gene, and will be potentially beneficial to the development of cure for the disease. However, the success in creating mouse model for HD will rely on the presence of the putative HD gene in mouse. Unfortunately this project could potentially be hampered by the very instability of repeats in mouse, such as in the case of FAX and MD.

Understanding the neuropathophysiology of the disease requires both commitment and intellectual creativity. Several approaches can be taken to study the cell biology of the putative HD gene. First, the full length cDNA of IT15 can be isolated and its gene structure can then be determined. Polyclonal antisera against different parts of the IT15 product can be raised to study the tissue and subcellular distribution of the gene product. The uncertainty of whether the CAG repeat is in the coding region of the gene can be resolved by raising antibodies against the polyglutamine stretch as well as the flanking region of the gene. The antisera will also determine whether the CAG repeat expansion prevents the production of the mature protein.

Different in vitro assays can be developed to address how the CAG repeat affects the expression pattern of the putative HD gene and why the CAG repeat is unstable. The expanded version of the repeat can be compared to the normal one on the expression of the putative HD gene. The putative premutation in the vicinity of the CAG repeat can be searched by careful analysis of the genomic region surrounding the CAG repeat. Consensus sequence can be searched for by comparing genomic sequences in the vicinity of the four disease loci. Further analysis of the potential DNA binding property of this consensus sequence will test the hypothesis describe above.

B. Genomic analysis

Assembling cosmid contigs and generating high resolution map are necessary before a complete characterization and sequencying the entire human genome. The method we have developed here to achieve these goals is appealing. It not only utilizes results of the YAC mapping that has been the center for the human genome project for the past few years, but also takes the advantage of chromosome specific genomic libraries that are currently available through the collaborative genome project. In theory, this approach should be applicable to the entire human genome. Our work on the HD region has shown some promises in this respect. Since the HD region is one of the most characterized regions of the genome, one may expect some difficulty in analyzing other regions of the genome. An immediate test of this approach to genome analysis could be to apply it to other regions of the genome of interest and to the entire chromosome 4. Preliminary results from our laboratory on other regions of the genome with 1 or 2 megabase pairs in length have been promising, while a large scale approach on chromosome 4 is underway in the Human Genome Mapping Center at Stanford University.

The future challenges for the genome project, in my opinion, are: 1) to develop efficient methods to sequence genomic clones such as cosmid in a large scale; 2) to develop an efficient method to store and analyze the information that has been generated by the mapping analyses. In theory, these two objectives are achievable. Therefore the genome project can, in theory, be done in a short period of time, or much shorter than everyone initially expected.

## APPENDIX ONE:

Molecular analysis of an unusual family with the Huntington disease.

The text of this chapter is a reprint of the material as it appears in American Journal of Human Genetics Vol.50. 1218-1230, 1992. As a separate project to analyze an unusual family with Huntington disease to pinpoint the disease locus, I participated in the initial studies of genotype analysis, fingerprint analysis and making hybrids from this family. I contributed to Figures 1, 2, and Table 1.

Am. J. Hum. Genet. 50:1218–1230, 1992

# Recombination of 4p16 DNA Markers in an Unusual Family with Huntington Disease

Catrin Pritchard,*,[1] Ning Zhu,*,[2] Jian Zuo,* Laura Bull,‡ Margaret A. Pericak-Vance,§,‖ Jeffery M. Vance,§ Allen D. Roses,§,‖ Athena Milatovich,# Uta Francke,#,** David R. Cox,†,‡ and Richard M. Myers*,‡

Departments of *Physiology, †Psychiatry, and ‡Biochemistry and Biophysics, University of California, San Francisco; §Department of Neurology and ‖Joseph and Kathleen Bryan Alzheimer Disease Research Center, Duke University Medical Center, Durham, NC; #Department of Genetics and **Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford

## Summary

The Huntington disease (HD) mutation has been localized to human chromosome 4p16, in a 6-Mb region between the *D4S10* locus and the 4p telomere. In a report by Robbins et al., a family was identified in which an affected individual failed to inherit three alleles within the 6-Mb region originating from the parental HD chromosome. To explain these results, it was suggested that the HD locus (*HD*) lies close to the telomere and that a recombination event took place between *HD* and the most telomeric marker examined, *D4S90*. As a test of this telomere hypothesis, we examined six members of this family, five of whom are affected with HD, for the segregation of 12 polymorphic markers from 4p16, including *D4S169*, which lies within 80 kb of the 4p telomere. We separated, in somatic cell hybrids, the chromosomes 4 from each family member, to determine the phase of marker alleles on each chromosome. We excluded nonpaternity by performing DNA fingerprint analyses on all six family members, and we found no evidence for chromosomal rearrangements when we used high-resolution karyotype analysis. We found that two affected siblings, including one of the patients originally described by Robbins et al., inherited alleles from the non-HD chromosome 4 of their affected parents, throughout the 6-Mb region. We found that a third affected sibling, also studied by Robbins et al., inherited alleles from the HD chromosome 4 of the affected parent, throughout the 6-Mb region. Finally, we found that a fourth sibling, who is likely affected with HD, has both a recombination event within the 6-Mb region and an additional recombination event in a more centromeric region of the short arm of chromosome 4. Our results argue against a telomeric location for HD and suggest that the HD mutation in this family is either associated with DNA predisposed to double recombination and/or gene conversion within the 6-Mb region or is in a gene that is outside this region and that is different from that mutated in most other families with HD.

## Introduction

Despite intensive effort, the gene responsible for Huntington disease (HD) has not yet been identified. The disease gene was first localized to chromosome 4p16 by genetic linkage analysis with the polymorphic DNA marker *D4S10* (Gusella et al. 1983), and subsequent studies narrowed the gene region to the 6-Mb interval between *D4S10* and the 4p telomere (Gilliam et al. 1987). The analysis of recombination events in the 6-Mb interval has since failed to provide a unique location for the mutation. In three families (MacDonald et al. 1989; Robbins et al. 1989), the HD locus (*HD*) appears to recombine with each of three markers tested in the 6-Mb interval. If it is assumed that only a single recombination event occurred in each of these cases, these results suggest that *HD* is very close to the

Received August 8, 1991; final revision received February 3, 1992.

Address for correspondence and reprints: Dr. R. M. Myers, Department of Physiology, Room S-762, University of California, 513 Parnassus Avenue, San Francisco, CA 94143-0444.

1. Present address: Institute of Molecular Medicine, John Radcliffe Hospital, Oxford.

2. Present address: Cancer Research Institute, University of California, San Francisco.


0002-9297/92/5006-0009$2.00

telomere. By contrast, genetic studies of three other families place *HD* in a more centromeric part of the 6-Mb segment, if it is assumed that only one recombination event occurred in the 6-Mb region in each recombinant chromosome (MacDonald et al. 1989, 1991; Barron et al. 1991). In agreement with these results, several DNA markers in a 2.5-Mb segment in this second region between the DNA markers *D4S10* and *D4S168* show linkage disequilibrium with the mutation, whereas markers in the telomeric region do not (Snell et al. 1989; Theilmann et al. 1989; Adam et al. 1991; MacDonald et al. 1991).

In the family (family 217) reported by Robbins et al. (1989), an individual with HD was discovered who failed to inherit alleles of three DNA markers in 4p16 from the parental *HD* chromosome, including the *HD* allele at *D4S90,* which maps approximately 300 kb from the telomere. This result was interpreted as evidence for a telomeric location for *HD* and implied that an as yet unidentified crossover event had taken place between *D4S90* and the 4p telomere. Here we have performed a more detailed genetic analysis of this family, in search of the apparent telomeric recombination event. We determined genotypes for six individuals, including two additional offspring who were not analyzed in the previous publication, with 12 polymorphic DNA markers from 4p16, one of which is located within 80 kb of the 4p telomere. We separated the chromosome 4 homologues from each family member in somatic cell hybrids to confirm the phase of chromosome 4 marker alleles, which allowed for the unambiguous determination of recombination events. Our results failed to identify telomeric recombination events in this family, reducing the likelihood that *HD* is located near the telomere. Our analysis suggests that this family either (*a*) is associated with DNA predisposed to double recombination and/or to gene conversion within the 6-Mb region or (*b*) carries a mutation in a gene that is located outside the 6-Mb region and that is different from that responsible for HD in most other families.

## Material and Methods

### DNA Fingerprint Analysis

For M13 fingerprinting, the method described by Vassart et al. (1987) was followed. Ten micrograms of genomic DNA was digested with *Hae*III and electrophoresed through a 1% agarose gel until fragments less than 1 kb had run off. The DNA was transferred to a Hybond N-Plus membrane (Amersham), which was then hybridized overnight at 42°C, in a buffer containing 40% formamide, 6 × SSC, 5 mM EDTA, and 0.25% dried skimmed milk, to M13 phage DNA radiolabeled by random priming (Feinberg and Vogelstein 1984). The filter was washed twice for 15 min each in 2 × SSC and 0.1% SDS at room temperature, four times for 15 min each in 2 × SSC and 0.1% SDS at 65°C, and twice for 30 min each in 1 × SSC at 65°C. The filter was exposed to X-ray film at room temperature for 12–72 h.

The dinucleotide repeat polymorphism at *D21S120* was analyzed by the method described by Burmeister et al. (1990). The VNTR polymorphism at *D17S5* was detected by digesting the genomic DNAs with *Hin*fI and hybridizing with the probe pYNZ22 (Odelberg et al. 1989). The VNTR at *D2S44* was analyzed by digesting genomic DNAs with *Bam*HI and hybridizing with the pYNH24 probe (Odelberg et al. 1989).

### Cytogenetic Analysis

Lymphoblastoid cell lines from the six members of this nuclear family were coded for cytogenetic analysis. The cells were cultured as described below in Derivation of Somatic Cell Hybrids. The cultures were split either 1:2 or 1:3 the day prior to harvesting. To each 10 ml of culture, 0.2 ml of $10^{-5}$ M methotrexate was added, and the cultures were incubated at 37°C for 17 h. The block was released with fresh medium that had 0.2 ml each of $10^{-3}$ M thymidine and 1 mg adenosine/ml added. These cultures were reincubated for 5 h with 0.1 ml of 10 μg colcemid/ml added for the final 10 min of incubation time. In some cultures, after 3 h, 0.1 ml ethidium bromide (1 mg/ml) and 30 μl colcemid (same concentration as above) were added, and they were placed at 37°C for an additional 2 h. When the cultures were not synchronized, 0.2 ml ethidium bromide and 0.1 ml colcemid were added, and they were kept at 37°C for 2 h. Cultures were harvested by standard procedures for 20 min in hypotonic 0.075 M KCl and were fixed with Carnoy's fixative (3:1 methanol:glacial acetic acid). Slides were made as soon as possible after harvest and were G-banded with trypsin and Giemsa stain. Metaphases were photographed by using Kodak Technical Pan film 2415 at ASA 80.

### Order of the Markers on 4p

The order of 11 of the 12 4p DNA markers used in our study has been deduced previously by meiotic linkage mapping (Buetow et al. 1991) and/or by

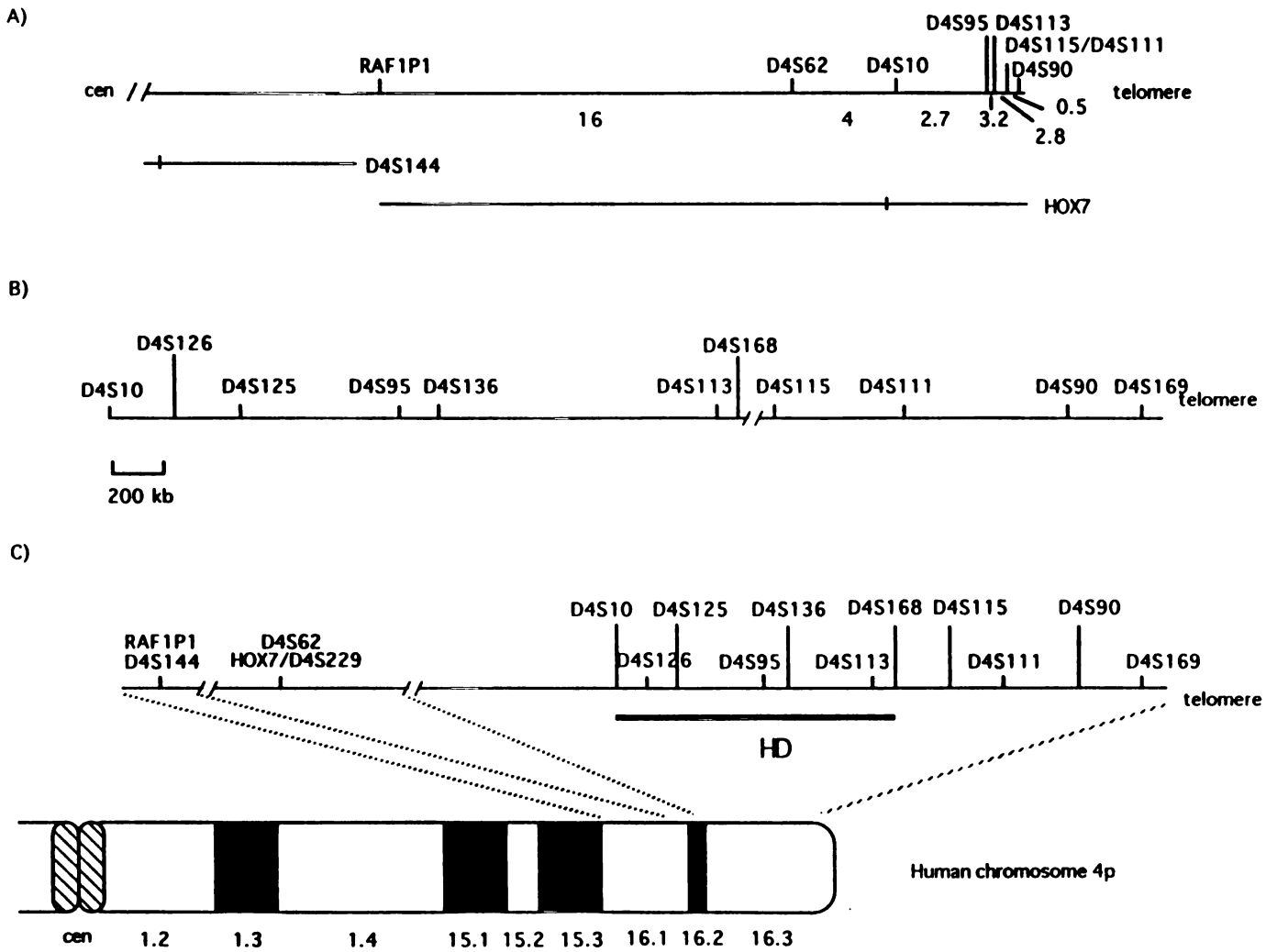



**Figure 1** Map of 4p markers analyzed in present study. *A*), Genetic map. This sex-averaged map of the region between *D4S10* and *D4S90* is derived from the data of Buetow et al. (1991) for the CEPH reference pedigree. Interlocus intervals are given in centimorgans (cM). The map spans approximately 30 cM. The precise locations of *D4S144* and *HOX7* are not known. However, the crossbar indicates their best-fit locations within the multipoint map. We did not score the *RAF1P1* locus in our study, as we were unable to identify a polymorphism informative in this family. Its location is included on this map as a reference for the location of *D4S144*, *HOX7*, and *D4S229*. *B*), Physical map of the region between *D4S10* and the telomere. These mapping data have been described by Bucan et al. (1990), Pritchard et al. (1989, 1990), and Bates et al. (1991). The precise physical distance of the region is not known, as *D4S113* and *D4S115* have not yet been linked by physical mapping techniques. *C*), Composite map. This map is not drawn to scale, as it represents a compilation of the genetic and physical mapping data. Hybrid mapping results place *D4S144* in the same interval as *RAF1P1* in 4p16.1. *HOX7* and *D4S229* are localized in the same interval as *D4S62* in 4p16.2, although the relative order of these three markers is not yet known. The 2.5-Mb region thought to contain the *HD* mutation, as defined by Bates et al. (1991), is indicated by the black horizontal bar.

pulsed-field gel electrophoresis (PFGE) restriction mapping (Bucan et al. 1990; Pritchard et al. 1989, 1990; Bates et al. 1991). The genetic mapping data are illustrated in figure 1*A*. While *D4S144* and *HOX7* could not be uniquely placed on the multilocus map, their best-fit locations are shown. A PFGE map has so far been reported only for the region between *D4S10* and the 4p telomere; this map is depicted in figure 1*B*. *D4S169* is the most terminal of the markers, mapping to within 80 kb of the 4p telomere (Pritchard et al. 1990, 1991). *D4S229* is the only locus not yet mapped by genetic or physical means. To establish a position for this locus and to define more accurate locations for *D4S144* and *HOX7*, we hybridized probes for each locus to the mapping panel of somatic cell hybrids containing chromosomes with various translocations and deletions of 4p (Smith et al. 1988; Cox et al. 1989). This was performed by the Southern blot/hybridization techniques described by Pritchard et al. (1989).

#### Derivation of Somatic Cell Hybrids

Lymphoblastoid cell lines established for each family member were used as donor cells for generation of somatic cell hybrids. These cells were grown in RPMI medium with 15% FCS, penicillin, streptomycin, and L-glutamine. In initial fusion experiments, the HPRT-deficient hamster line 380-6 was utilized as the recipient cell type. However, we subsequently switched to using the HPRT-deficient mouse cell RAG because hybrids constructed with it retained human chromosome 4 in a higher fraction of hybrids. Both recipient cell lines were cultured in DME medium with 10% FCS, penicillin, streptomycin, and 6 μg 6-thioguanine/ml.

Somatic cell hybrids were constructed according to the method described by Puck et al. (1989). Hybrid clones were picked and transferred to the wells of a 24-well plate. After 7 d, half of the cells in each clone were harvested and used in a PCR assay (Kim and Smithies 1988) to test for the retention of two loci on human chromosome 4: *D4S229* and *D4S136*. The cells were washed with PBS, transferred to microfuge tubes, and resuspended in 50 μl water. They were then incubated on dry ice for 1–2 h, after which they were boiled for 8 min. One microgram of Proteinase K was added, and the mixture was incubated at 55°C for 1–2 h. The PCR assays were performed as specified above for each locus. Amplification products were detected by electrophoresis through 3% agarose (Nusieve; Seakem) gels and by ethidium bromide staining. The hybrid clones giving positive amplification signals were identified and expanded, and genomic DNA was isolated from them (Hofker et al. 1985). Clones containing one or both human chromosomes 4 were distinguished by assaying the genomic DNAs with the 12 polymorphic markers as described below.

#### Genotype Analysis

*RFLP analysis.*—Table 1 provides both a description of the enzymes used and allele assignments for the eight loci for which RFLPs were detected. For this analysis, the Southern blot-hybridization technique described by Pritchard et al. (1989) was employed.

The VNTR polymorphism at *D4S125* was detected by following the method described by Richards et al. (1991). After amplification by PCR, the products were phenol extracted, ethanol precipitated, and digested with *Bgl*I. They were then electrophoresed through a 10% nondenaturing polyacrylamide gel containing 15% glycerol and were visualized by ethidium bromide staining.

*GT repeat polymorphisms.*—The primer sequences and allele assignments of the four loci for which dinucleo-

**Table 1**

**RFLPs Detected in Present Study**

| Locus (probe), Enzyme | Allele Sizes (designations) (kb) | Reference |
|---|---|---|
| *D4S90* (D5), *Pvu*II | 5.6 (A) and 5.0 (B) | Youngman et al. 1988 |
| *D4S111* (p157.9), *Pst*I | 1.9 (A) and 1.7 (B) | MacDonald et al. 1989 |
| *D4S115* (p252-3), *Pst*I | 2.4 (A), 2.3 (B), and 2.1 (C) | MacDonald et al. 1989 |
| *D4S113* (p337), *Stu*I | 6.3 (A) and 5.6 (B) | Whaley et al. 1988 |
| *D4S95* (pBS674), *Taq*I | 2.3 (E1) and 1.7 (E2) | Wasmuth et al. 1988 |
| *D4S125* (YNZ32), *Bgl*I | .605 (A), .51 (B), .5 (C), and .475 (D) | Richards et al. 1991 |
| *D4S10* (pKO83), *Eco*RI | 15.0 (A) and 10.0 (B) | Gusella et al. 1984 |
| *D4S144* (pIFM19-1), *Hin*dIII | 4.5 (A) and 3.7 and .7 (B) | Buetow et al. 1991 |

tide repeat polymorphisms were detected are shown in table 2. The polymorphism at *D4S169* was detected as described by Pritchard et al. (1991). For *D4S229*, *HOX7*, and *D4S136*, the PCR products were detected by using an end-labeled primer. End-labeling was performed with T4 polynucleotide kinase and $\gamma^{32}$P-ATP (170 µCi/50 pmol primer). PCR assays were performed in 10–50-µl reaction volumes with 100–500 ng genomic DNA, 200 mM dNTPs, 1 pmol each primer/µl reaction mix, 1 U AmpliTaq (Perkin Elmer Cetus), and PCR buffer (67 mM Tris-HCl pH 8.8, 1.5 mM $MgCl_2$, 16.6 mM $NH_4SO_4$, 10 mM β-mercaptoethanol, and 6.7 mM EDTA). *D4S229* and *D4S136* reactions were cycled 35 times through 1 min at 94°C, 30 s at 60°C, and 1 min at 72°C. *HOX7* reactions were cycled 35 times through 1 min at 94°C, 30 s at 55°C, and 1 min at 72°C. After PCR, 6 µl of the reactions was mixed with 4 µl loading buffer (80% formamide, 0.1% bromophenol blue, and 0.1% xylene cyanol), and the samples were electrophoresed on denaturing 6.5% polyacrylamide DNA sequencing gels. Gels were dried and exposed to X-ray film at room temperature for 12–72 h.

### Lod Score Analysis

Two-point lod score analysis was performed to test linkage between *D4S10* and *HD* and between *D4S95* and *HD* by using the program LINKAGE (Lathrop et al. 1984). At-risk individuals were assigned affectation status by using an age-at-onset curve generated from the pedigree data. The analysis included genotypes for *D4S10* and *D4S95* for all the individuals reported by Robbins et al. (1989), as well as genotypes for individuals II-1 and II-4, who were not included in the previous publication.

## Results

### Family History and Clinical Features

Family 217 has been reported elsewhere and is part of an extended pedigree with autopsy-proven HD (Robbins et al. 1989). Eighteen family members with polymorphisms at *D4S10*, *D4S95*, and *D4S90* were analyzed, and one affected individual was discovered who inherited alleles from the non-HD chromosome of the affected parent. In our study, we focused on the nuclear family containing this unusual individual (fig. 2). The family comprises two parents (I-1 and I-2), of which I-1 was affected with HD. There were four children and one miscarriage. II-2 and II-3 are the two offspring described by Robbins et al. (1989), and II-2 is the affected individual with the unusual genotype. We analyzed two additional offspring, II-1 and II-4,

**Table 2**

**GT Repeat Polymorphisms Analyzed in Present Study**

| Locus and Primer Sequences | Size of PCR Product | No. of Dinucleotides (allele designations) |
|---|---|---|
| *D4S229:* | | |
| 5′-CTACCTGTCATATTCAGGAATCACC-3′ | 200 bp | $n+7$ (A), $n+3$ (B), $n+1$ (C), and $n$ (D)[a] |
| 5′-AGGCCATTGCTGATGGCAGGAAACA-3′ | | |
| *HOX7:* | | |
| 5′-TTAGATTGTCATCAGTCCTC-3′ | 130 bp | $n+2$ (A), $n+1$ (B), and $n$ (C)[a] |
| 5′-GGGCATGTTGATGTCTGCTGAC-3′ | | |
| *D4S136:* | | |
| 5′-CTGACTTGATCCAATCCAAAGGAAAG-3′ | 225 bp | $n+4$ (A) and $n$ (B)[a] |
| 5′-TTGAACCTAGTAGGCGGAAGTTGCAC-3′ | | |
| *D4S169:* | | |
| 5′-GAATTCAGTTTTAGCTGAGCTAAGG-3′ | 1.6 kb (200-bp *Rsa*I fragment) | 19 (A), 18 (B), and 15 (C) |
| 5′-GAATTCAGTCGACTGAGAATCCTTT-3′ | | |

[a] The number of repeats relative to the sequenced, cloned DNA was not determined. Therefore, alleles were assigned relative to the smallest allele (referred to as "$n$").

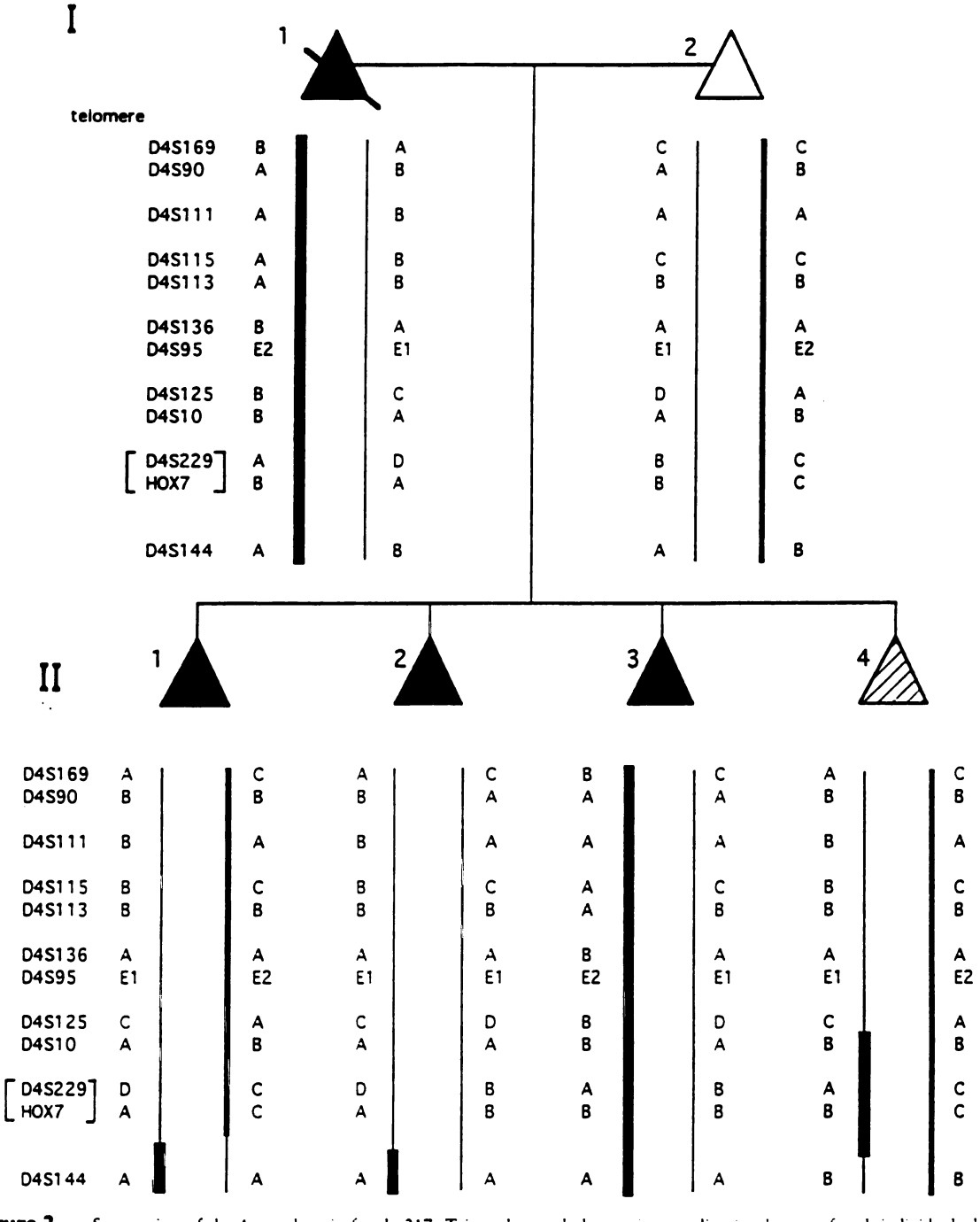



**Figure 2** Segregation of the 4p markers in family 217. Triangular symbols are given to disguise the sex of each individual, thus maintaining the confidentiality of the family. Blackened triangles represent individuals affected with HD, and unblackened triangles indicate unaffected individuals. The hatched triangle indicates that this individual may be affected with HD but that the symptoms are not developed to the point that a definitive diagnosis can be given. II-2 is the affected offspring with the unusual genotype previously described by Robbins et al. (1989). The chromosome 4 genotypes are indicated for each family member. The *HD* chromosome of I-1 is represented by the thicker black vertical line in each panel. This same chromosome is found in all affected members of the extended pedigree. The order of the markers is based on the map shown in fig. 1C. *HOX7* and *D4S229* are bracketed into one haplotype, as their relative orientation is not known.

who were not reported in the Robbins et al. paper. Clinical examinations were performed by several neurologists, and the following symptoms characteristic of HD were observed:

Individual I-1 is now deceased but was diagnosed with HD before death at age 65 years. The patient had mental status changes before age 45 years, with violent outbursts requiring psychiatric hospitalization. The patient later developed prominent choreic movements (before age 50 years).

Individual II-1 is 49 years old and has recently been diagnosed with HD. II-1 is described by the family to be irritable, with violent outbursts. The patient admits to being restless and has difficulty with motor/movement skills. In 1989, II-1 was observed to have rare choreic movements of the extremities. Results of neuropsychiatric testing at that time were within normal limits. On examination, there was protrusion of the tongue for more than 20 s but with flagrant and uncontrollable writhing movements. II-1 had fine movements of the fingers of both hands when extended and had occasional spontaneous jerking of the lower extremity. Gait and eye movements are normal. The patient exhibits slight asymmetric dysmetria on coordination testing.

Individual II-2 is 44 years old and has been followed clinically for several years, with gradual progression in symptoms and confirmed diagnosis of HD. Neuropsychiatric testing in 1989 suggested impairment in verbal organization, constructional praxis, and spatial judgment, along with mild memory impairment. Choreic movements were noted occasionally in 1989 and now are more frequent in outstretched hands, trunk, and legs. Gait, saccades, and results of coordination testing were unremarkable.

Individual II-3 is 42 years old and has been diagnosed with HD. The patient has had multiple bouts of violent behavior, including threats of suicide and threats to the family's safety. II-3 complains of difficulty concentrating, irritability, and sleeplessness. Examination revealed rare spontaneous choreoathetosis movements of the fingers on extension. The patient has occasional shifts of weight, with characteristic veering while walking, as well as moderate dysdiadochokinesia and dysmetria on heel-to-shin testing.

Individual II-4 is 33 years old and was previously considered normal and was felt to be in good health with no history of emotional lability. However, on recent examination (in July 1991), II-4 was unable to sustain tongue protrusion for more than 10 s on repeated tests and exhibited moderate spontaneous choreoathetoid movements in extended fingers. Gait and results of coordination testing were normal. Although these symptoms are highly suggestive of HD, they are presently not definitive enough to conclude that the patient is affected with the disease. This individual is being followed longitudinally to see whether further characteristic symptoms develop.

### *DNA Fingerprint Analysis*

To rule out sampling errors and nonpaternity, we performed fingerprint analysis of DNA isolated from lymphoblastoid cell lines of each of the six family members of our study. The probe we used for this analysis was labeled M13 phage DNA, which has been shown to detect multiple polymorphic minisatellite loci in the human genome (Vassart et al. 1987). These results indicated that the two parents, I-1 and I-2, had different fingerprint patterns, such that none of the approximately 15 DNA fragments resolved on the gel were the same size in these two individuals (fig. 3). All of the DNA fragments in each of the four offspring were observed in the fingerprint patterns of either the mother or the father, except for a single novel fragment approximately 11.5 kb in length in II-3 (indicated by the arrowhead in fig. 3, see Discussion). Comparison of the DNA fragment patterns between II-1, II-2, II-3, and II-4 indicated that approximately 50% of the DNA fragments were shared between each pair of these siblings. We also derived M13 fingerprints for three siblings of I-1 (data not shown), and in all cases approximately 50% of fragments were shared in common with I-1, and even fewer were shared with generation II.

As a further test that the individuals whom we studied are the offspring of I-1 and I-2, we determined the genotypes of the six family members with 12 polymorphic DNA markers on 4p16 (see below), one dinucleotide repeat locus on 21q (*D21S120*), one VNTR locus on 17p (*D17S5*), and one VNTR locus on 2q (*D2S44*) (data not shown). Allele frequencies are available for eight of these loci (*D4S10, D4S95, D4S90, D4S113, D4S144, D21S120, D17S5,* and *D2S44*), allowing us to estimate a probability of >.99997 that I-1 and I-2 are the true parents of II-1, II-2, II-3, and II-4 (Odelberg et al. 1989).

### *Cytogenetic Analysis*

The chromosomes from lymphoblastoid cell lines derived from the six individuals depicted in figure 2 were analyzed by high-resolution cytogenetic methods, without knowledge of pedigree status. For each

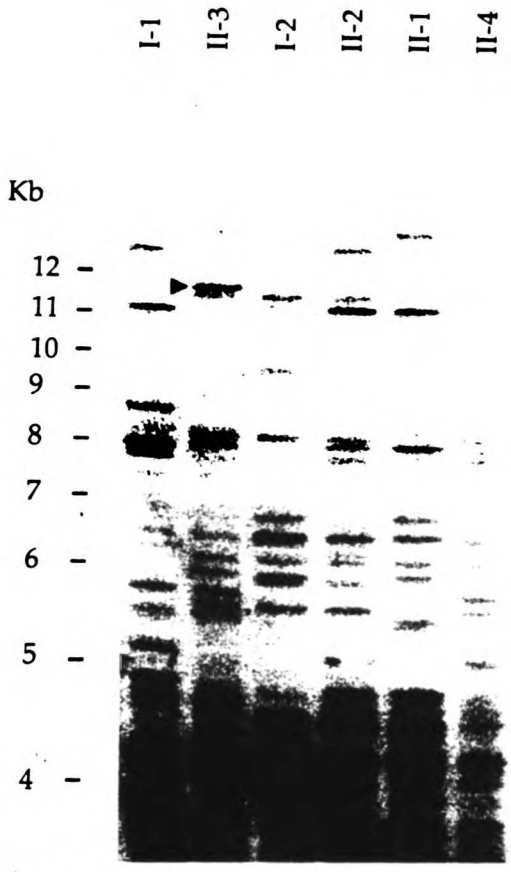



**Figure 3** DNA fingerprint of family 217. The fingerprint was generated by digesting each genomic DNA sample with *Hae*III, fractionating the products through an agarose gel, and hybridizing the DNA to radiolabeled phage M13. Approximately 15 polymorphic DNA fragments are observed in each individual. All of the fragments present in generation II can be traced to generation I, except for the novel fragment, in II-3, indicated by the arrowhead.

individual, a minimum of 30 metaphases were completely cytogenetically analyzed. All six cell lines appeared to have normal karyotypes (data not shown). Additional metaphase cells were examined that contained chromosomes 4 at high-resolution, 750–850-band stage. At that stage, band 4p16 is subdivided, and the relative position of grey subband 4p16.2 was carefully examined. At this level of resolution, the relative distances between 4p16.2 and 4p15.3, i.e., the size of band 4p16.1, were identical in both chromosome 4 homologues in cells from all six individuals. While these data suggest that the chromosomes 4 of all six individuals are structurally normal, they do not rule out (*a*) the presence of a small inversion or small duplication entirely within lightly staining bands 4p16.1 and 4p16.3 or (*b*) one breakpoint in each of these bands that are equidistant from 4p16.2 (see fig. 1C).

#### *Order of the Markers on 4p*

The order of the 12 DNA markers on 4p that were used in the present study was deduced mainly from previously published genetic and physical maps (see Material and Methods and fig. 1). In addition, we obtained map positions for *D4S144*, *HOX7*, and *D4S229* by hybridizing probes for these loci to a panel of somatic cell hybrids containing defined deletions and translocations of chromosome 4p (Smith et al. 1988; Cox et al. 1989). Our results (fig. 1C) placed *D4S144* in the vicinity of *RAF1P1*, either proximal or distal to it. *HOX7* and *D4S229* were located in the same interval as was *D4S62*. A composite map of the 12 markers was constructed by combining the previously published data with the hybrid mapping results (fig. 1C).

#### *Determination of Genotypes of the 4p Markers in Family 217*

In our initial studies with the six members of the family reported here, we determined the genotypes of the 12 DNA markers shown in figure 2. While it was possible to deduce the chromosomal phase of the alleles unambiguously for most of these markers, the phase had to be inferred for *D4S10*, *D4S95*, and *D4S90* in those cases where both parents and offspring were heterozygous for the same alleles at these loci. To determine the phase of these markers unequivocally, we separated the chromosome 4 homologues of each individual by generating somatic cell hybrids that retained single copies of human chromosome 4 (see Material and Methods). Each such hybrid cell line was then typed with the 12 polymorphic markers. An example of the results obtained for *D4S10*, *D4S95*, *D4S169*, and *HOX7* is shown in figure 4. In this figure, DNA from a hybrid containing one chromosome 4 homologue (lanes L), from a hybrid containing the other chromosome 4 homologue (lanes R), and from the lymphoblastoid cell line containing both chromosome 4 homologues (lanes T) is included for each individual. At least six independent hybrids containing a single copy of human chromosome 4 were analyzed for each individual. In no case did we observe phase or segregation of alleles in the hybrid cell lines that was inconsistent with that determined by using lymphoblastoid cell lines from these six individuals and other members of the extended family.

We found that for all markers that we tested, with

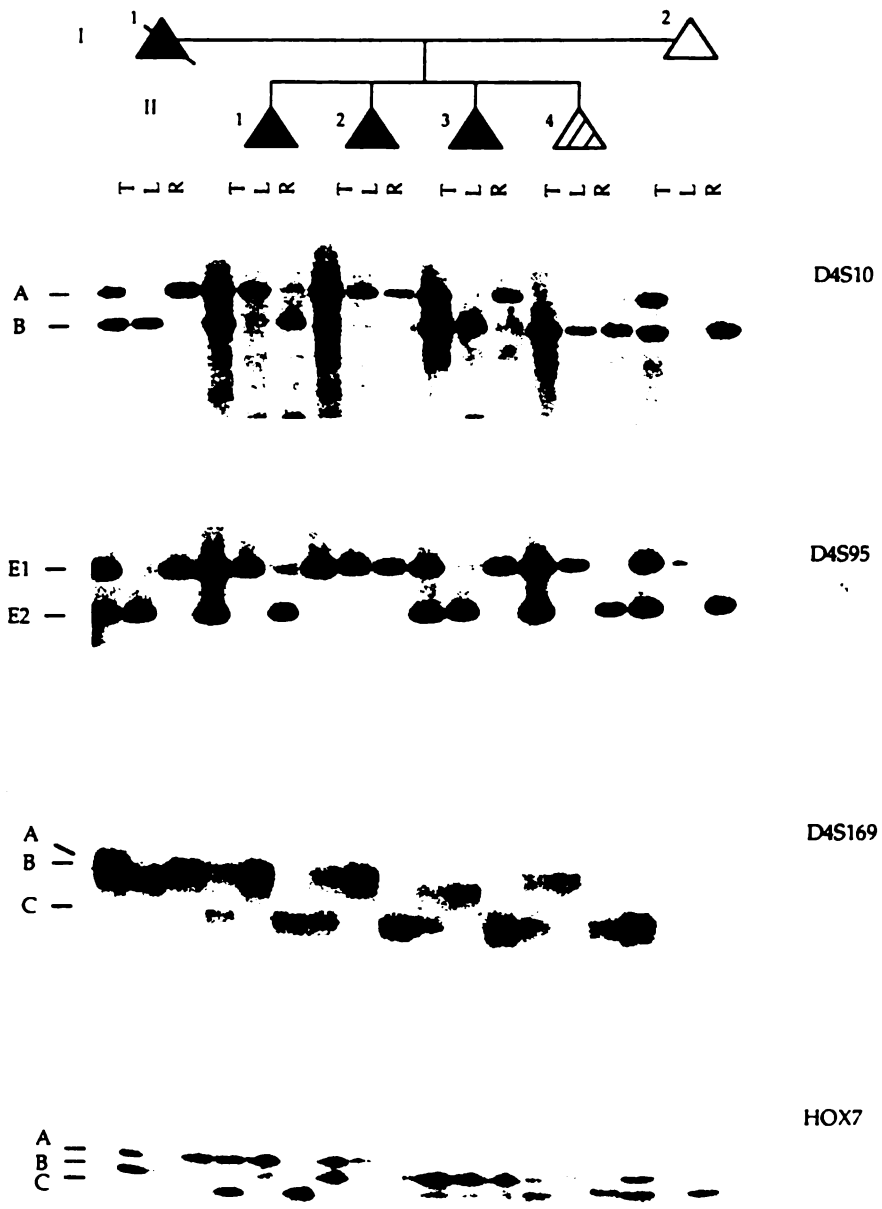



**Figure 4** Analysis of hybrids with 4p polymorphic markers. Genomic DNA derived from each individual and from representative hybrids containing either the left or right human chromosome 4 (as depicted in fig. 2) was analyzed with *D4S10*, *D4S95*, *D4S169*, and *HOX7*. The pedigree structure of family 217 is shown at the top of the figure. The lanes of each autoradiogram correspond to the family member directly above in the pedigree. T = total genomic DNA isolated from the lymphoblastoid cell lines; and L and R = genomic DNA isolated from somatic cell hybrids containing only the left or right human chromosome 4, respectively. The alleles detected at each locus (see nomenclature in tables 1 and 2) are indicated on the left-hand side of each autoradiogram. Separation of the GT-repeat alleles was greatly aided by detection of only one of the two DNA strands. Although the A and B alleles at *D4S169* migrate close together, analysis of the L and R hybrids of I-1 shows that they are clearly distinct. The hybrids allowed the phase of the chromosomes 4 to be unequivocally set. For example, alleles B at *D4S10*, E2 at *D4S95*, B at *D4S169*, and B at *HOX7* are observed in the L hybrid of I-1, thus confirming the phase of these alleles that is shown in fig. 2. The low signal obtained with all markers in the L hybrid of I-2 indicates that the chromosome 4 in this hybrid is unstable. Consequently, we were unable to amplify *D4S169* sequences in this hybrid, despite several attempts. In some hybrids, residual levels of the homologous chromosome 4 were detectable, e.g., for the R hybrid of II-1.

the exception of *D4S144*, II-2 carried alleles from the non-*HD* chromosome of the affected parent (fig. 2). These results confirm those reported by Robbins et al. (1989), in which individual II-2 was shown to carry alleles of three markers—*D4S10*, *D4S95*, and *D4S90*—derived from the affected parent's non-*HD* chromosome. In light of the fact that *D4S169* is located within 80 kb of the 4p telomere, these data provide evidence against the hypothesis that a recombination event occurred between *D4S90* and the telomere on the chromosome 4 homologue inherited by II-2 from the affected parent. However, these data indicate that a recombination event occurred in the 16-cM interval between *HOX7/D4S229* and *D4S144*, outside the *HD* region, on the chromosome 4 homologue inherited by II-2 from the affected parent. The fact that individual II-2 did not inherit alleles from the affected parent's *HD* chromosome 4 homologue in the 6-Mb *HD* region is an unexpected result. Even more unexpected is the finding that individual II-1, who is also affected with HD, inherited alleles from the affected parent that were identical to those observed in II-2 (fig. 2). By contrast, individual II-3, who is also clearly affected with HD, inherited alleles from the *HD* chromosome of the affected parent, for all 12 chromosome 4p markers (fig. 2). Of the subset of these markers that have been tested in the extended family, the alleles found on the *HD* chromosome 4 of II-3 are found in all other affected individuals reported by Robbins et al. (1989). Finally, individual II-4, who has symptoms that are highly suggestive of HD, inherited from the affected parent a chromosome 4 that appears to be recombined at two locations in 4p. This individual inherited the *D4S144* allele from the non-*HD* chromosome, the *HOX7/D4S229* and *D4S10* alleles from the *HD* chromosome, and the *D4S125*, *D4S95*, *D4S136*, *D4S113*, *D4S115*, *D4S111*, *D4S90*, and *D4S169* alleles from the non-*HD* chromosome of the affected parent (fig. 2). Thus, *D4S10* is the only marker of those tested in the 6-Mb *HD* region that was inherited from the affected parent's *HD* chromosome by II-4.

## Discussion

Although the chromosomal location of *HD* has been known from linkage studies for several years, refinement of the location of the gene on chromosome 4p has been hampered by the failure to observe even a single 4p chromosomal rearrangement that is associated with the disease. Therefore, the only two approaches that have narrowed down the location of the *HD* gene have been linkage disequilibrium and analysis of rare recombination events between 4p DNA markers and *HD*. Linkage-disequilibrium studies suggest that the *HD* gene is in a 4p segment near *D4S95*, although the evidence for disequilibrium is weak. The strongest evidence for localizing the *HD* gene has come from studies of individuals with recombination events in 4p16, but unfortunately these results do not lead to a single consistent hypothesis: some recombination events place the gene in a 2.5-Mb region between *D4S10* and *D4S168*, while others suggest a more telomeric location for *HD*. The combination of linkage disequilibrium and recombination results points to the 2.5-Mb proximal region as the most likely location for the *HD* gene.

In an effort to resolve this inconsistent localization that is shown by studies of rare recombination events, we have studied one of the families that suggested a telomeric location for the gene. In family 217 we studied the two parents and two affected offspring who were initially described by Robbins et al. (1989), as well as two additional offspring who had not been studied before. Prior to extensive genotype analysis of this family, we asked whether the apparent telomeric location could be explained by misdiagnosis, nonpaternity, inadvertent sample mix-ups, or chromosomal rearrangements. Our analysis suggests that it is unlikely that any of these possibilities has occurred. The chance that this family has a degenerative neurological disease other than HD is unlikely, as postmorten studies of an affected member of the extended pedigree showed, in the caudate nucleus and putamen, neuronal degeneration characteristic of HD. In addition, clinical neurological evaluations of many family members are completely consistent with a diagnosis of HD in these individuals.

Our DNA fingerprinting results with an M13 minisatellite probe, 12 loci on 4p, and three other highly polymorphic DNA markers located on different human chromosomes are entirely consistent with the four children being the true offspring of the two parents tested. These results essentially exclude the possibility that a random individual or even a close relative of I-1 or I-2 is the parent of any of the offspring. The novel M13 fragment observed in individual II-3 (see fig. 3 and Results) does not change this conclusion, as all other alleles observed in this individual can be traced to the two parents. On repeated digestion of the same genomic DNA sample, we did not observe changes in the intensity of this novel band, a result suggesting that it is unlikely to have arisen by partial digestion. It is more likely that the fragment resulted from mutation of an M13 allele, either in II-3 or in

the lymphoblastoid cell line derived from II-3. Such mutations occur fairly frequently and have been detected for several other minisatellite loci (Jeffreys et al. 1988; Amour et al. 1989).

Our high-resolution karyotype analysis of this family revealed no gross rearrangements of either 4p16 or any other portion of the genome, suggesting that any rearrangement, if it exists, must be small. In addition, all of the 4p16 markers segregated concordantly in the somatic cell hybrids derived from each patient, arguing against the possibility of a small translocation between these markers.

Other possibilities that might explain the apparent telomeric location in this family are that I-1 is homozygous for the disease locus or that I-2 carries a mutated *HD* allele. Both of these suggestions are highly unlikely, as (*a*) HD is not observed in the extended family of the nonaffected parent of I-1 or in the extended family of I-2 and (*b*) I-2, at age 70 years, shows no symptoms of the disease. In addition, II-1 and II-2 inherited different chromosomes 4 from I-2, both of which would have to carry the mutant gene if the latter hypothesis were true.

As there was no trivial explanation for the apparent localization of *HD* to a telomere in family 217, we attempted to identify the obligatory recombination event, between *D4S90* and the telomere in individual II-2, that is predicted by a telomeric location. This study required an informative polymorphic DNA probe from a locus very near the telomere. Although a number of DNA probes from this region have been isolated (Bates et al. 1990; Pritchard et al. 1990), it has been difficult to identify polymorphisms with these markers. Nevertheless, we identified a simple sequence repeat of the GT type in a DNA fragment that maps less than 80 kb from the telomere, and we found that it detects variable alleles in family 217, allowing us to test the telomere hypothesis. Our analysis of individual II-2, the HD patient whose genotype with *D4S90* suggested a telomeric location for *HD* in the paper by Robbins et al., indicated that there was no recombination between *D4S90* and *D4S169*, significantly decreasing the likelihood that *HD* is at the telomere in this family.

In light of the fact that a telomeric location for the *HD* gene is unlikely, how can we explain that two affected individuals carry alleles of 4p16 markers from their affected parent's non-*HD* chromosome? Two alternative explanations seem most likely. The first is that gene conversion or other multiple recombination events that we have failed to detect have occurred in a small region of 4p16, transferring the mutant version of the *HD* gene to the non-*HD* chromosome 4 in both of these individuals. Although multiple recombination events in such a small chromosomal region would seem very unlikely, evidence suggesting that recombination in the *HD* region is unusual (MacDonald et al. 1991) is consistent with this multiple-recombination hypothesis. Robbins et al. (1989) did not identify, in one of these individuals (II-2), multiple recombination events in this region with three DNA markers spaced about 3 Mb apart. In the present study, we tested six additional DNA markers in this 6-Mb region, at an average distance of about 700 kb, with the largest interval about 1,000 kb (between *D4S136* and *D4S113*). If multiple recombination events occurred in the 6-Mb region in individuals II-1 and II-2, they would have to have occurred in one of the intervals between the DNA markers that we tested.

A second explanation of our results is that HD is genetically heterogeneous and that the neurological features observed in affected individuals from family 217 are caused by a mutation in a gene located outside 4p16. Although all four offspring potentially share a segment of the affected parent's *HD* chromosome, in the region between *HOX7* and *D4S144*, there is no reason to suspect that this region of the genome is more likely than any other region to be the location of the *HD* gene in these individuals. If the disease phenotype in family 217 is due to a mutation elsewhere in the genome, then studying the chromosomes 4p16 from individuals II-1 and II-2 will not provide useful information about *HD*. Although genetic heterogeneity studies provide no evidence for a separate locus for *HD*, the confidence limits on the estimate of the fraction of families that are heterogeneous are large, such that as many as 12% of families with HD could have a mutation at a separate locus (Conneally et al. 1989). Therefore, it is not unreasonable to consider the possibility that family 217 has HD due to a mutation in a gene different from that responsible for the disease in most families. While the genotypes of offspring II-1 and II-2 are inconsistent with linkage of *HD* to 4p16, the genotypes of the remaining individuals in the extended family reported by Robbins et al. (1989) are consistent with linkage. Combined lod score analysis of these extended family members and the six individuals reported in the present study (some of whom were not reported by Robbins et al.) fails to provide conclusive evidence for or against linkage of two 4p16 markers—*D4S10* and *D4S95*—to *HD* in the extended family (table 3).

It is difficult to determine which, if either, of these two alternative explanations of our results is more

**Table 3**

**Two-Point Lod Score Analysis of Family 217 (Males and Females)**

| | Lod Score at Recombination Fraction of | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Loci | .000 | .010 | .050 | .100 | .150 | .200 | .300 | .400 |
| *HD/D4S10*.... | −99.000 | −1.0027 | −.3647 | −.1402 | −.0403 | .0097 | .0423 | .0331 |
| *HD/D4S95*.... | −99.000 | −1.3628 | −.7020 | −.4476 | −.3158 | −.2322 | −.1271 | −.0565 |

likely, as there is no evidence strongly supporting one hypothesis over the other. However, if the multiple-recombination hypothesis is correct, individual II-4 could provide information for localizing the *HD* gene on 4p16. The chromosome 4 that this individual inherited from the affected parent appears to contain a single crossover event between *D4S10* and *D4S125*, two DNA markers that are less than 500 kb apart. If II-4 is indeed affected with HD, then one explanation for these crossover data, combined with other single crossovers in this region reported in other families, is that the *HD* gene lies in the small interval between these two DNA markers. However, an unequivocal HD diagnosis for II-4 remains to be established, and this individual is from an unusual family. Nevertheless, if this hypothesis is correct, not only should a mutation responsible for HD map in this region, but also alleles from the affected parent's *HD* chromosome should be present in individuals II-1 and II-2, in the vicinity of the mutation.

## Acknowledgments

We thank Dr. Marcy MacDonald for providing p337, p252-3, and p157.9; Dr. John Wasmuth for providing pBS674 and the somatic cell hybrids used for mapping; Dr. Jim Gusella for providing pKO83; Dr. Duncan Shaw for providing D5; Drs. Maxine Singer and Wesley McBride for providing P8; Dr. Ray White for providing pYNH24 and pYNZ22; Dr. Jeff Murray for providing pIFM19-1 and the information regarding the *HOX7* GT repeat polymorphism; and Dr. Sue Povey for providing the RAG cells. We also thank our colleagues for encouragement and support. This work was funded by the Wills Foundation and by NIH grants to U.F., R.M.M., D.R.C., J.M.V., M.A.P.-V, and A.D.R. U.F. is an Investigator of the Howard Hughes Medical Institute.

## References

Adam S, Theilmann J, Buetow K, Hedrick A, Collins C, Weber B, Huggins M, et al (1991) Linkage disequilibrium and modification of risk for Huntington disease. Am J Hum Genet 48:595–603

Armour JA, Patel I, Thein SL, Fey MF, Jeffreys AJ (1989) Analysis of somatic mutations at human minisatellite loci in tumors and cell lines. Genomics 4:328–334

Barron L, Curtis A, Shrimpton AE, Holloway S, May H, Snell RG, Brock DJH (1991) Linkage disequilibrium and recombination make a telomeric site for the Huntington's disease gene unlikely. J Med Genet 28:520–522

Bates GP, MacDonald ME, Baxendale S, Sedlacek Z, Youngman S, Romano D, Whaley WL, et al (1990) A yeast artificial chromosome telomere clone spanning a possible location of the Huntington disease gene. Am J Hum Genet 46:762–775

Bates GP, MacDonald ME, Baxendale S, Youngman S, Lin C, Whaley WL, Wasmuth JJ, et al (1991) Defined physical limits of the Huntington disease gene candidate region. Am J Hum Genet 49:7–16

Bucan M, Zimmer M, Whaley WL, Poustka A, Youngman S, Allitto BA, Ormondroyd E, et al (1990) Physical maps of 4p16.3, the area expected to contain the Huntington disease mutation. Genomics 6:1–15

Buetow KH, Shiang R, Yang P, Nakamura Y, Lathrop GM, White R, Wasmuth JJ, et al (1991) A detailed multipoint map of human chromosome 4 provides evidence for linkage heterogeneity and postition-specific recombination rates. Am J Hum Genet 48:911–925

Burmesiter M, Cox DR, Myers RM (1990) Dinucleotide repeat polymorphism located at D21S120. Nucleic Acids Res 18:4969

Conneally PM, Haines J, Tanzi R, Wexler N, Penchaszadeh C, Harper P, Folstein S, et al (1989) No evidence of linkage heterogeneity between Huntington disease (HD) and G8 (D4S10). Genomics 5:304–308

Cox DR, Pritchard CA, Uglum E, Casher D, Kobori J, Myers RM (1989) Segregation of the Huntington disease region of human chromosome 4 in a somatic cell hybrid. Genomics 4:397–407

Feinberg AP, Vogelstein B (1984) Addendum: a technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 137:266–267

Gilliam TC, Tanzi RE, Haines JL, Bonner TI, Faryniarz AG, Hobbs WJ, MacDonald ME, et al (1987) Localization of the Huntington's disease gene to a small segment of

chromosome 4 flanked by D4S10 and the telomere. Cell 50:565–571

Gusella JF, Wexler N, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234–238

Gusella JF, Gibbons K, Hobbs W, Heft R, Anderson M, Rashtchian R, Folstein S, et al (1984) The G8 locus linked to Huntington's disease. Am J Hum Genet 36 [Suppl]: 139S

Hayden MR (1981) Huntington's chorea. Springer, Berlin

Hofker MH, Wapenaar MC, Goor N, Bakker E, Pearson PL (1985) Isolation of probes detecting restriction fragment length polymorphisms from X chromosome-specific libraries: potential use for diagnosis of Duchenne muscular dystrophy. Hum Genet 70:148–156

Jeffreys AJ, Royle NJ, Wilson V, Wong Z (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332:278–281

Kim H-S, Smithies O (1988) Recombinant fragment assay for gene targeting based on the polymerase chain reaction. Nucleic Acids Res 16:8887–8903

Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446

MacDonald ME, Haines JL, Zimmer M, Cheng SV, Youngman S, Whaley WL, Bucan M, et al (1989) Recombination events suggest potential sites for the Huntington's disease gene. Neuron 3:183–190

MacDonald ME, Lin C, Srinidhi L, Bates G, Altherr M, Whaley WL, Lehrach H, et al (1991) Complex patterns of linkage disequilibrium in the Huntington disease region. Am J Hum Genet 49:723–734

Odelberg SJ, Plaetka R, Eldridge JR, Ballard L, O'Connell P, Nakamura Y, Leppert M, et al (1989) Characterization of eight VNTR loci by agarose gel electrophoresis. Genomics 5:915–924

Pritchard C, Casher D, Bull L, Cox DR, Myers RM (1990) A cloned DNA segment from the telomeric region of chromosome 4p is not detectably rearranged in Huntington disease patients. Proc Natl Acad Sci USA 87:7309–7313

Pritchard CA, Casher D, Uglum E, Cox DR, Myers RM (1989) Isolation and field-inversion gel electrophoresis analysis of DNA markers located close to the Huntington disease gene. Genomics 4:408–418

Pritchard C, Cox DR, Myers RM (1991) Dinucleotide repeat polymorphism located at D4S169. Nucleic Acids Res 19:6347

Puck JM, Nussbaum RL, Smead DL, Conley ME (1989) X-linked severe combined immunodeficiency: localization within the region Xq13.1-q21.1 by linkage and deletion analysis. Am J Hum Genet 44:724–730

Richards B, Horn GT, Merrill JJ, Klinger KW (1991) Characterization and rapid analysis of the highly polymorphic VNTR locus D4S125 (YNZ32), closely linked to the Huntington disease gene. Genomics 9:235–240

Robbins C, Theilmann J, Youngman S, Haines J, Altherr MJ, Harper PS, Payne C, et al (1989) Evidence from family studies that the gene causing Huntington disease is telomeric to D4S95 and D4S90. Am J Hum Genet 44: 422–425

Shaw M, Caro A (1982) The mutation rate to Huntington disease. J Med Genet 19:161–167

Smith B, Skarecky D, Bengtsson U, Magenis RE, Carpenter N, Wasmuth JJ (1988) Isolation of DNA markers in the direction of the Huntington disease gene from the G8 locus. Am J Hum Genet 42:335–344

Snell RG, Lazarou L, Youngman S, Quarrell OWJ, Wasmuth JJ, Shaw DJ, Harper PS (1989) Linkage disequilibrium in Huntington's disease: an improved localization for the gene. J Med Genet 26:673–675

Theilmann J, Kanani S, Shiang R, Robbins C, Quarrell O, Huggins M, Hedrick A, et al (1989) Non-random association between alleles detected at D4S95 and D4S98 and the Huntington's disease gene. J Med Genet 26:676–681

Vassart G, Georges M, Monsieur R, Brocas H, Lequarre AS, Cristophe D (1987) A sequence in M13 phage detects hypervariable minisatellites in human and animal DNA. Science 235:683–684

Wasmuth JJ, Hewitt J, Smith B, Allard D, Haines JL, Skarecky D, Partlow E, et al (1988) A highly polymorphic locus very tightly linked to the Huntington's disease gene. Nature 322:734–736

Whaley WL, Michiels F, MacDonald ME, Romano D, Zimmer M, Smith B, Leavitt J, et al (1988) Mapping of D4S98/S114/S113 confines the Huntington's defect to a reduced physical region at the telomere of chromosome 4. Nucleic Acids Res 16:11769–11780

Whaley WL, Bates GP, Novelletto A, Sedlacek Z, Cheng S, Romano D, Ormondroyd E, et al (1991) Mapping of cosmid clones in Huntington's disease region of chromosome 4. Somatic Cell Mol Genet 17:83–91

Youngman S, Shaw DJ, Gusella JF, MacDonald M, Stanbridge EJ, Wasmuth J, Harper PS (1988) A DNA probe, D5 (D4S90) mapping to human chromosome 4p16.3. Nucleic Acids Res 16:1648