DOE/ER/60692-5

MAPPING AND ORDERED CLONING OF THE HUMAN X CHROMOSOME

Progress Report

September 1991 — November 1992

C. Thomas Caskey, M.D.
David L. Nelson, Ph.D.

Institute for Molecular Genetics

Baylor College of Medicine

Houston, Texas 77030

December 1992

PREPARED FOR THE U.S. DEPARTMENT OF ENERGY
UNDER GRANT NUMBER DE-FG05-88ER60692

A. Progress 9/1/91-11/30/92

Progress in the past year has been quite good. Significant advances have been made toward several goals. In particular, the accumulation of additional large insert clones of known location on the X, identification of disease loci and associated genes, and generation of new cDNAs, polymorphic markers and sequence-tagged sites.

YAC libraries

We have obtained a new X chromosome specific YAC library from Dr. Robert Nussbaum of U.Penn. This library consists of 3300 clones with average insert size of 300 kb each. We also have available the total human YAC libraries from Washington University and from the CEPH in Paris. These two combined encompass over 100,000 clones and represent ~6x coverage of the X chromosome. These libraries are in addition to those produced locally and described in last year's report.

In the coming months, the new Genethon "MegaYAC" library is anticipated to arrive. This library contains clones of >800 kb average insert size and will greatly facilitate isolation of large contigs.

Cosmid library

A flow-sorted human X chromosome cosmid library prepared by Dr. Pieter de Jong of Lawrence Livermore has recently been obtained. This 24,000 clone arrayed library has been gridded onto dense filters (1536 each) for screening and provides substantial support for YAC cloning efforts either in filling "holes" or isolating "subclones" of the regions covered by the YAC contigs. These activities will become extremely important as the effort shifts to the MegaYACs, which may have a greater propensity to rearrangement. Also, as efforts to identify genes intensify, the use of cosmids will become more important. Cosmids isolated from this library thus far have been numerous, and of high quality with large (38-44 kb) inserts and little or no rearrangement detectable.

YAC assignment

The majority of YAC isolation and assignment during this period has come from the CEPH YAC library. Over 250 X chromosome YACs have been isolated from this and the St. Louis and Nussbaum libraries in the past year, bringing the total number of YACs of defined location isolated from the X to over 500.

Random YAC assignment

We have determined the feasibility of subdividing YACs from our small insert X-specific libraries into sub-libraries derived from regions using *Alu* PCR product probes derived from somatic cell hybrids retaining those regions. This approach has not been successful due to the small insert sizes of the YACs, and is no longer being pursued. Encouraging results have been published by the Genethon group suggesting that this approach is feasible in larger insert YACs.

Specific loci

Fragile X Syndrome One of the major successes of the past two years has been the identification and characterization of the fragile X site and an associated gene, FMR-1. One of the randomly assigned clones from the our X library, RS46, was found to be close to the fragile X mutation site as defined by a series of fragile X associated X chromosome breakpoints. This led eventually to isolation of the fragile X mutation and associated gene.

The fragile site has turned out to be one of the most interesting loci in the human genome from the point of view of mutational mechanisms. The fragile site consists of a repetitive sequence present in an exon of the FMR-1 gene. The repeat is $(CGG)_n$ and is present in the 5' untranslated portion of the cDNA for FMR-1. We have determined by PCR that the number of CGG repeats is polymorphic in the human population, varying from a low of 6 to a high of 54 repeats, with an average length of 29 repeats (Fu *et al.*, 1992). Females are heterozygous for allele length ~65% of the time, and we have observed 28 different alleles thus far. Premutation alleles observed in fragile X families range from a low of 43 to a high of ~200 repeats. These alleles confer no disease on carrier males or females, but do predispose to full mutations in the offspring of female carriers. Full mutations are large (200-3000 repeats) and become methylated, depressing transcription of the FMR-1 mRNA. Premutations are found to mutate with a frequency approaching 1 in fragile X families. Some mutations involve amplification to full mutation, while others remain in the premutation size range. However, all alleles differ between parent and offspring. This degree of instability in DNA is unprecedented.

Continued work on the fragile X locus has focussed on the structure of the FMR-1 gene, with a large-scale sequencing project in conjunction with the DNA sequencing core of the Baylor Human Genome Center. To date, over 35 kb of this gene has been sequenced, and all exon intron boundaries have been determined. The remainder of the gene as well as flanking sequences will be sequenced in the coming 6 months; it is anticipated that over 60 kb of sequence data will be developed from the region. A large number (~40) of oligonucleotides defining characterized STSs have been generated from this effort. Extension of the YAC and cosmid contig into those surrounding the fragile X site is ongoing.

Myotonic dystrophy. An obvious candidate for a mutation similar to that found in the fragile X syndrome was myotonic dystrophy (DM). This disorder displays "anticipation", a phenomenon whereby the disease is found to be progressively more severe as the mutant gene is passed to subsequent generations. Although this gene is not X-linked, it was felt that the importance of study of this mutational mechanism warranted a deviation from the strictly defined aims of this project. Through identification of a triplet repeat in a cosmid clone derived from a YAC known previously to be located in the critical region of the DM gene, a CTG repeat in the 3' untranslated region of the myotonin protein kinase gene (Fu et al., 1993) was discovered and characterized. This triplet is polymorphic in length, with an average of 5 repeats, but ranging in the population to 30. In DM families, the repeat is found to vary from 35 to several thousand copies, with larger numbers conferring a more severe phenotype. This repeat is similarly unstable in meiotic transmission, and has a mutation rate that is also close to unity in the >50 copy range. Sequence analysis of the DM gene reveals a complicated gene with many alternatively spliced products. The mechanism of mutation leading to phenotype is unclear, since the expanded repeat is present in the 3' untranslated region, does not interrupt the coding sequence and is un-methylatable. Work is ongoing to understand the basis of the disease and how the mutation can lead to the phenotype.

Lowe's syndrome. Of the 97 YAC clones assigned to sub-regions of Xq24-qter, one clone was identified which crosses the breakpoint in the X chromosome retained in Lowe-3. This chromosome was isolated from a female patient with the oculocerebrorenal syndrome of Lowe, and is a translocation of X to chromosome 3 with a breakpoint in Xq25. In collaboration with Robert Nussbaum of U. Penn, we have identified two genes in this region. One is the putative Lowe's syndrome gene, showing multiple mutations in a number of Lowe's patients. This gene has been designated OCRL-1, and its predicted amino acid sequence shows significant homology with an enzyme involved in inositol phosphate metabolism. Dr. Nussbaum is continuing research to identify possible defect in inositol phosphate in Lowe's syndrome patients. The second gene, which is not found mutant in any of the Lowe's patients has been designated SNF2l, due to its homology to the SNF2 gene of yeast, a global activator of transcription. This is the first human member of this family to be identified which includes genes with homeotic mutations in *Drosophila*a.

Other X-linked loci of interest at Baylor. Contig assembly is proceeding rapidly in several other regions surrounding disease loci of interest to Baylor researchers. Among these are the XIST, KAL, XLA, LYP, GK/AHC, IDS, and Rett syndrome-translocation loci. Additionally, efforts to span Xq28 and identify novel sequence based polymorphisms will provide materials for the numerous disease loci in this popular region of the genome. Genes for the X-inactivation center and for the KAL (Kalmann) syndrome defect have been identified, and significant YAC contigs have been defined surrounding these areas by Andrea Ballabio's group at Baylor.

A collaborative effort with Dr. Ballabio and Dr. Huda Zogbhi to finish the contig assembly from KAL proximal to the DMD gene has begun. Thus far, many contigs exceeding 1-2 Mb have been identified, along with a deletion panel composed of nearly 50 intervals defined by patients with deletions and translocations. It is anticipated that production of 150 STSs in this 40 Mb region will be sufficient to define the contig, and that this effort will be complete in 6-9 months.

Collaboration with Bakary Sylla and Gilbert Lenoir of IARC in Lyon has allowed construction of a 1 Mbp contig of YACs near the X-linked lymphoproliferative syndrome locus (XLP) in Xq25. Significant progress in bridging the scarce markers in the Xp21.2-21.3 regions distal to DMD has been made by the efforts of Huda Zoghbi and Ed McCabe's groups at Baylor (~5-7 Mb in contig). These focussed efforts provide highly detailed description of X chromosome regions of great interest. The NIH funded Human Genome Center at Baylor provides further support of X chromosome efforts, with core facilities for YAC isolation, sequencing and DNA synthesis, computation and mapping support.

B. Objectives 12/1/92-11/30/93

No major alterations from the proposed studies are contemplated. Major efforts will concentrate on improving the physical maps of the X chromosome, and support of the regional mapping efforts ongoing at Baylor and collaboratively in various centers worldwide will continue. It is anticipated that as physical maps of the chromosome improve that more emphasis will be placed on transcript identification and characterization, and on the development of genomic sequence data.

A slight direction shift involving identification of additional genes containing triplet repeats with high GC content has been undertaken by Dr. Caskey in a further effort to characterize this novel mutational mechanism. In this effort, over 100 cDNA clones have been identified as containing either CGG or CTG repeat sequences, and many of these have been characterized for polymorphism. Clones exhibiting polymorphism will be characterized further for possible large expansions. These clones will also be assigned to chromosomes. While this direction deviates from the X chromosome focus of the grant, it follows logically from the fragile X and myotonic dystrophy results and is well within the mission of the OHER in understanding mutational mechanisms. It is possible that a separate (or supplemental) application may be sought for these studies at the time of renewal of this application.

It is expected that the next year of funding will generate significant new insights into the organization of the human X chromosome and associated disease loci, while continuing to produce cloned materials useful for these analyses.

C. Publications 9/1/91-12/1/92

Attree O, Olivos IM, Okabe I, Bailey LC, Nelson DL, Lewis RA, McInnes RR and Nussbaum RL (1992) The Lowe's oculocerebrorenal syndrome gene encodes a protein highly homologous to inositol polyphosphate-5-phosphatase. *Nature* 358:239-242.

Caskey CT, Pizzuti A, Fu YH, Fenwick RG Jr. and Nelson DL (1992) Triplet repeat mutations in human disease. *Science* **256**:784-789.

Edwards A and Caskey CT (1991) Closure strategies for random DNA sequencing. *Methods* **3**:41-47.

Edwards A and Caskey CT (1991) Genetic marker technology. *Curr. Opin. Biotechnol.* **2**:818-822.

Fu, Y-H, Kuhl, DPA, Pizzuti, A, Pieretti, M, Sutcliffe, JS, Richards, S, Verkerk, AJMH, Holden, JJA, Fenwick Jr., RG, Warren, ST, Oostra, BA, Nelson, DL, Caskey, CT (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox. *Cell* **67**:1047-1058.

Fu, Y-H, Pizzuti, A, Fenwick, RG, Jr., King, J, Rajnarayan, S, Dunne, PW, Dubel, J, Nasser, GA, Ashizawa, T, de Jong, P, Wieringa, B, Korneluk, R, Perryman, MB, Epstein, HF, Caskey, CT (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255**:1256-1258.

Kilimann MW, Pizzuti A, Grompe M and Caskey CT (1992) Point mutations and polymorphisms in the human dystrophin gene identified in genomic DNA sequences amplified by multiplex PCR. *Hum. Genet.* **89**:253-258.

Lee CC and Caskey CT Direct cDNA cloning using PCR. *Methods Enzymol.*, in press.

Mandel JL, Monaco AP, Nelson DL, Schlessinger D and Willard H (1992) Genome analysis and the human X chromosome. *Science* **258**:103-109.

Macpherson, JN, Nelson, DL, Jacobs, PA (1992) Frequent small amplifications in the FMR-1 gene in Fra(X) families: Limits to the diagnosis of 'premutations'. *J Med Genet* **29**:802-806.

Nelson DL (1991) Applications of polymerase chain reactions in genome mapping. *Curr. Opin. Genet. Dev.* **1**:62-68.

Nelson DL (1991) Interspersed repetitive sequence polymerase chain reaction (IRS PCR) for generation of human DNA fragments from complex sources. *Methods* **2**:60-74.

Nelson DL, Ballabio A, Victoria MF, Pieretti M, Bies RD, Gibbs RA, Maley JA, Chinault AC, Webster TD and Caskey CT (1991) *Alu*-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: identification of clones associated with a disease locus. *Proc. Natl. Acad. Sci. USA* **88**:6157-6161.

Okabe I, Attree O, Bailey LC, Nelson DL and Nussbaum RL Isolation of cDNA sequences around the chromosomal breakpoint in a female with Lowe syndrome by direct sequencing of cDNA libraries with yeast artificial chromosomes. *J. Inherited Metab. Dis.*, in press.

Okabe I, Bailey LC, Attree O, Srinivasan S, Perkel JM, Laurent BC, Carlson M, Nelson DL and Nussbaum RL (1992) Cloning of human and bovine homologs of SNF2/SWI2: a global activator of transcription in yeast *S.cerevisiae*. *Nucleic Acids Res.* **20**:4649-4655.

Pieretti, M, Zhang, F, Fu, Y-H, Warren, ST, Oostra, BA, Caskey, CT, Nelson, DL (1991) Absence of expression of the FMR-1 gene in Fragile X syndrome. *Cell* **66**:817-822.

Pizzuti A, Pieretti M, Fenwick RG, Gibbs RA and Caskey CT (1992) A transposon-like element in the deletion-prone region of the dystrophin gene. *Genomics* **13**:594-600.

Riggins GJ, Sherman SL, Oostra BA, Sutcliffe JS, Feitell D, Nelson DL, van Oost BA, Smits APT, Ramos FJ, Pfendner E, Kuhl DPA, Caskey CT and Warren ST (1992) Characterization of a highly polymorphic dinucleotide repeat 150 kb proximal to the fragile X site. *Am. J. Med. Genet.* **43**:237-243.

Rossiter BJF and Caskey CT (1992) The Human Genome Project and clinical medicine. *Oncology* **6**:61-71.

Rossiter BJF and Caskey CT The Human Genome Project. *Clin. Obstet. Gynecol.*, in press.

Sutcliffe JS, Nelson DL, Zhang F, Pieretti M, Caskey CT, Saxe D and Warren ST (1992) DNA methylation represses *FMR-1* transcription in fragile X syndrome. *Hum. Mol. Genet.* **1**:397-400.

Sutcliffe JS, Zhang F, Caskey CT, Nelson DL and Warren ST (1992) PCR amplification and analysis of yeast artificial chromosomes. *Genomics* **13**:1303-1306.

Verkerk AJMH, deVries BBA, Niermeijer MF, Nelson DL, Fu YH, Warren ST, Majoor-Krakauer DF, Halley DJJ and Oostra BA (1992) Intragenic probe used for diagnostics in fragile X families. *Am. J. Med. Genet.* **43**:192-196.

Verkerk, AJMH, Pieretti, M, Sutcliffe, JS, Fu, Y-H, Kuhl, DPA, Pizzuti, A, Reiner, O, Richards, S, Victoria, MF, Zhang, F, Eussen, BE, van Ommen, G-JB, Blonden, LAJ, Riggins, GJ, Chastain, JL, Kunst, CB, Galjaard, H, Caskey, CT, Nelson, DL, Oostra, BA, and Warren, ST (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in Fragile X syndrome. *Cell* **65**:905-914.

Wang, Q, Ishikawa-Brush, Y, Monaco, AP, Nelson, DL, Caskey, CT, Pauly, SP, Lenoir, GM, Sylla, BS (1993) Physical mapping of the Xq24-q25 around loci closely linked to the X-linked lymphoproliferative syndrome locus: an overlapping YAC map and linkage between DXS12, DXS42 and DXS37. *Eur J Hum Genet*, **1**:64-71.

END

DATE
FILMED
7/13/93