ISOLATION OF cDNAs FROM THE HUMAN X CHROMOSOME AND DERIVATION OF RELATED STSs

Final Progress Report

RECEIVED

JAN 04 1996

OSTI

April 1992 — March 1995

David L. Nelson, Ph.D.

Department of Molecular and Human Genetics

Baylor College of Medicine

Houston, Texas 77030

September 1995

PREPARED FOR THE U.S. DEPARTMENT OF ENERGY
UNDER GRANT NUMBER DE-FG05-92ER61401

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Progress 4/1/92-3/28/95

Over the course of this funding period, the number of genes assigned to the human X chromosome has approximately tripled from less than one hundred to nearly three hundred characterized, cloned genes assigned to it. The aims of this project were to develop methods for gene identification and to identify and characterize expressed sequences from the X chromosome.

The rapidly changing environment of the human genome project provided abundant resources for gene characterization, and since methods for gene identification became rather robust over this period, these aims were de-emphasized during the project. Among the methods developed was a local one (reciprocal probing) that was developed by Drs. Cheng Chi Lee and C. Thomas Caskey, with emphasis on the human X chromosome (Lee et al., *Hum Mol Genet* 4:1373, 1995). The development of this method offered significant expressed sequence resources for this project, particularly when coupled with the efforts to identify cosmid clones from specific X chromosome locations, as the reciprocal probing process results in paired genomic (cosmid) and cDNA materials. Attention, then has been paid to characterization of genes rather than to their identification.

EST mapping:

A major aim of this application involved the further characterization of ESTs generated by the Venter group and assigned to the X chromosome. Toward this aim, a number of cDNAs have been studied. A manuscript describing sublocalization of these clones has been published (Parrish and Nelson, *Hum Mol Genet* 2:1901, 1993). In this paper, 19 additional X-linked genes were assigned to regions as well as to YAC and cosmid contigs in many cases.

The Lee et al. paper cited above placed several dozen novel ESTs onto the chromosome.

A paper in press in Human Molecular Genetics (Ferrero et al., *Hum Mol Genet* 4, in press, 1995) describes additional fine mapping of ESTs in a 35 Mbp region of distal Xp in the course of constructing a complete YAC contig of this region.

EST mapping has become a major endeavor in a number of groups, although the emphasis has turned to mapping random ESTs with genome-wide resources such as radiation hybrids.

Numerous specific genes from the distal long arm of the X chromosome have been characterized. These are discussed below in pter-qter order:

FMR1/FRAXA (Xq27.3)

Significant effort has been expended in characterization of this gene involved in fragile X syndrome. Most of this effort has been funded by either an NIH RO1 for characterizing the gene and disease, or by a separate DOE grant (ER61830) that supports mapping and sequencing of the X chromosome. Sequence of 152 kb surrounding this gene has been deposited in the public databases. This sequence was developed in collaboration with Dr. Richard Gibbs at BCM. An interesting feature of this sequence is the apparent absence of neighboring genes.

FRAXE (Xq27.3)

This chromosomal fragile site is associated with mental retardation that is less severe than that found in classical fragile X syndrome. Despite over two years of knowledge of the fragile site, expansion and disease, no gene has been identified. We have devoted significant effort to characterizing this region at the genomic level (cosmid contigs, extensive genomic sequence) in order to identify a gene which may be affected by this mutation. Exon amplification, cDNA

selection, direct screening of cDNA libraries, and other methods failed to provide bona fide cDNAs. We uncovered two pseudogenes from these efforts. In recent months, we have identified through GRAIL searches of genomic sequence developed over 150 kb distal to the fragile site an exon which provided access to a cDNA from the region. This is a large mRNA, and a quite large gene, but a key feature is that its 5' end is encoded by an exon carrying the FRAXE GCC repeat. This is almost certainly the gene affected by FRAXE expansion and methylation, and we anticipate submission of a manuscript regarding this gene by the end of the year. Dr. Yanghong Gu was supported by the present grant for some of this work.

TH4 (Xq27.3-q28)

This gene was originally identified as an EST by Shoji Tsuji in Niigata using a method we developed to clone human cDNAs specifically from somatic cell hybrids. The gene is located immediately proximal to the IDS gene, as revealed by cosmid-based sequencing carried out in collaboration with Dr. Gibbs. Indeed, TH4 represents an alternatively poly adenylated version of the IDS transcript, thus it is not a novel gene.

IDS (Xq28)

This gene, iduronate sulfatase, results in Hunter syndrome when defective. In collaboration with Dr. Gibbs, we have determined the sequence of 130 kb of DNA surrounding the IDS gene. This has resulted in the identification of two novel genes (termed X and X') and a potential third (Y). This work was recently published (Timms et al., *Genome Res.* 5:71, 1995) and points to the utility of genomic sequence for gene identification and characterization. The sequence resolved the highly similar and closely mapping X and X' genes, and also cleared up confusing issues regarding the TH4 sequence (see above) and a very close by IDS pseudogene.

FRAXF (Xq28)

This fragile site was identified by Dr. Julia Parrish in the group through identification of cosmids from the flow-sorted X chromosome library carrying high-GC content sequences. The probe used for this (a cDNA termed H10) was obtained from Ben Oostra in Rotterdam as related to the fragile X gene, FMR1. On further study, the H10 sequence unlikely to be expressed. A paper describing the FRAXF mutation was published (Parrish et al., *Nature Genet* 8:229, 1994). This, like all other folate sensitive fragile sites, is caused by expanded numbers of CGG repeats which are methylated in their expanded versions. We were able to determine that FRAXF does not confer a phenotype on individuals carrying the expanded version. However, access to the cDNAs developed by Dr. Lee and coworkers determined that a member of a large gene family was present adjacent to the FRAXF site. This family is known as the MAGE genes, and they were first identified as tumor specific antigens present on the cell surfaces of melanoma and other neuroectodermally derived tumors. Subsequent analysis by us and others has placed at least 12 MAGE genes in the interval of the X chromosome between FRAXF and GABRA3.

Ku autoantigen (Xq28)

Reciprocal probing efforts by Dr. Lee positioned this gene in Xq28, and we have sublocalized it to a region just distal to FRAXF. As this gene has been assigned to an autosome by FISH analysis, we were curious to determine whether the sequences on X corresponded to the authentic gene encoding this protein, antibodies to which are often found in patients with lupus. Genomic sequence analysis of cosmid subclones revealed that this is a pseudogene.

Na⁺-dependent creatine transporter and 3A1/CDM (Xq28).

In the course of searching for additional CGG trinucleotide repeats present on the X chromosome, we identified a cosmid located in Xq28, which contains three genes identified through the use of

local arrayed cDNA libraries and hybridization with the cosmid clone. One of these genes is involved in the human disorder adrenoleukodystrophy (ALD), which was recently identified. Another, 3A1/CDM was characterized by us, and was also characterized by the group of J-L Mandel in the course of studying ALD. The third gene is the Na⁺-dependent creatine transporter (CRT), which was recently cloned from rabbit and human.

The most remarkable aspect of these genes (CRT and CDM) is that they have undergone a very recent (in evolutionary time) duplication and insertion into 16p11. A complete copy of the CRT gene is found on chromosome 16, while a truncated portion of CDM is also present. These have retained order and orientation. Cosmids from chromosome 16 (obtained from LANL) provided the ability to sequence each region completely and characterize the duplication event. Approximately 26 kb is in duplicate, with overall conservation of greater than 95% identity. A single stop codon in the CRT, and the inability to identify transcripts from this locus suggest it is a pseudogene. A pentameric, GC rich repeat is found near the junctions on both X and 16. This sequence is similar to pentamers involved in recombination of chromosome rearrangements in the immune system. A manuscript describing this extensive data is in preparation.

HCF1 (VP16 accessory protein) (Xq28).

This gene was described by Dr. Winship Herr as an accessory factor for herpes virus infection. Dr. Barbara Trask assigned the gene to Xq28 by FISH, and we were contacted to provide a more refined location. The gene is adjacent to the L1-cell adhesion molecule gene and close to the renin binding protein gene. Cosmids and YACs have been identified for this sequence in reference libraries. This work was recently published (Wilson et al., *Genomics* 25:462, 1995).

DXS1012E (Xq28).

This is one of the ESTs we assigned to the region, and we have followed this with complete cDNA characterization.

DNL1L (Xq28).

In the course of characterizing genomic clones at the DXS1010E locus, we identified this gene and have determined the complete cDNA sequence, along with exon-intron boundaries and genomic sequence. The gene is similar to Dnase I, is expressed abundantly in muscle and heart, and formed a candidate for Emery-Dreifuss muscular dystrophy, but was ruled out by identification by others of a neighboring gene exhibiting defects in these patients. An intriguing insertion/deletion polymorphism was found in DNL1L, where 190 bp of the 3' untranslated region is absent in some individuals. This work has appeared in a publication (Parrish et al. *Hum Mol Genet* 4:1557, 1995).

DXS1010E (Xq28).

This is one of the ESTs we assigned to the region, and we have followed this with complete cDNA characterization.

Identification of over one thousand cosmids from this region of the chromosome using YAC probes to identify and "bin" the cosmids has allowed cross-comparison with the reciprocal probing efforts of Dr. Lee. This has resulted in dozens of additional potentially expressed sequences from this 10 Mbp interval. Identification of additional cDNAs has made use of cloned genomic material from Xq28 to isolate cDNAs directly from local arrayed cDNA libraries. This method is robust and provides a convenient short-cut around other methods.

Efforts to make use of 3' terminal exon trapping for identification of novel genes from cloned genomic sources have been uniformly unsuccessful.

Rapid Assessment of Gene Content by Directed Sample Sequencing (Xq28)

We have developed and tested a strategy to assess the gene content over a large span of genomic DNA by directed sample sequencing across the region. The strategy utilizes physical mapping data to restrict a region for analysis followed by construction of a M13 shotgun library from a group of cosmids known to reside within the region. The M13 subclones are arrayed and hybridized with the cosmids used to construct the library as well as other cosmids from the region to localize the M13 subclones to intervals or "bins" defined by the cosmids. Multiple M13 subclones from each of the bins are subjected to sequence analysis, and the sequence information is used to search the public databases for interesting features. Sequence is derived in an ordered fashion and a reserve of well mapped sequence ready clones are available for further analysis of specific regions. We selected a well characterized gene rich interval in Xq28 from the color vision (*CV*) genes to glucose 6-phosphate dehydrogenase (*G6PD*) gene spanning 260 kb. 384 bacteriophage plaques were picked and arrayed for analysis. 71% of the M13 subclones contained Xq28 specific inserts and sequence analysis was performed on 115 of the Xq28 specific M13 subclones. Assuming a mean sequencing read of 400 bp, sequence information was obtained for 18% of the interval. Homologues were found to all but one of the reported genes from the interval. Our results indicate that this technique can support gene finding expeditions across large regions of genomic DNA with low levels of sequence coverage and will become more useful with the rapidly expanding cDNA sequence information available in the public databases.

Students trained:

Evan E. Eichler, Ph.D. (Currently a postdoctoral fellow with H. Moherenweiser, LLNL)
Kellie A. Lugenbeel (in training)
Andrea Peier (in training)

Postdoctoral fellows supported:

Julia E. Parrish, Ph.D. (in training)
Angela Scheuerle, M.D. (Currently on faculty Univ. Texas at Houston, Dept. Of Genetics)
Yanghong Gu, M.D. (in training)

Publications (*indicates specific citation of support)

*Firulli, BA, Parrish, JE, Nelson, DL (1996) Rapid assessment of gene content by directed sample sequencing. in preparation.

*Richards, S, Eichler, EE, King, J, Pizutti, A, Nelson, DL, Gibbs, RA (1995) Complete sequence of the human FMR1 locus. in preparation.

*Eichler, EE, Lu, F, Shen, Y, Antonucci, R, Doggett, NA, Moyzis, RK, Baldini, A, Gibbs, RA, Nelson, DL (1996) Duplication of the Xq28 CDM -CTR region to 16p11.1: A novel pericentromeric-directed mechanism for paralogous genome evolution. in preparation.

Ferrero, GB, Franco, B, Roth, EJ, Firulli, BA, Borsani, G, Delmas-Mata, J, Weissenbach, J, Halley, G, Schlessinger, D, Chinault, AC, Zoghbi, HY, Nelson, DL, Ballabio, A (1995) An integrated physical and genetic map of a 35 Mb region on chromosome Xp22.3-Xp21.3. *Hum Mol Genet*, in press.

Timms, KM, Lu, F, Shen, Y, Pierson, CA, Gu, Y, Nelson, DL, Gibbs, RA (1995) 130 kb of DNA sequence reveals two new genes and a regional duplication distal to the human iduronate-2 sulfatase locus. *Genome Research*, 5:71-78.

*Parrish, JE, Ciccodicola, A, Wehnert, M, Cox, GF, Chen, E, Nelson, DL (1995) A muscle-specific DNase I-like gene in human Xq28. *Hum Mol Genet*, 4:1557-1564.

*Wilson, AC, Parrish, JE, Massa, HF, Nelson, DL, Trask, BJ, Herr, W (1995) The gene encoding the VP16-accessory protein HCF resides in human Xq28 and is highly expressed in fetal tissues and the adult kidney. *Genomics*, 25:462-468.

*Parrish, JE, Oostra, B, Verkerk, AJMH, Richards, CS, Reynolds, J, Spikes, AS, Shaffer, LG and Nelson, DL (1994) Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE. *Nature Genetics*, 8:229-235.

*Parrish, JE, Nelson, DL. (1993) Regional assignment of 19 X-linked ESTs. *Hum. Mol. Genet.*, 2:1901-1905.

*Parrish, JE and Nelson, DL (1993) Methods for finding genes: A major rate limiting step in positional cloning, *Genetic Analysis: Techniques and Applications* 10:29-41.