DOE/ER60692-1

MAPPING AND ORDERED CLONING OF THE HUMAN X CHROMOSOME

Final Progress Report

March 1991 — February 1995

C. Thomas Caskey, M.D., PI

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-FG03-88ER60692

Progress 3/1/94-2/28/95

The final year of funding provided to Dr. Caskey was used to develop and further elaborate the reciprocal probing method devised by Dr. Caskey and Dr. Cheng Chi Lee of BCM. A publication describing the results of this pilot study appeared in a recent issue of Human Molecular Genetics (Lee et al., 1995).

Briefly, the method entails the use of pooled cDNA probes to arrayed chromosome-specific cosmid libraries (prepared by the national gene library project of LANL or LLNL) to identify cosmids containing sequences capable of hybridizing with a cDNA clone in the pool. Cosmid clones are isolated and used as probes to identify specific cDNA clones, and the cDNA and cosmid are paired. Sequence of the cDNA insert and FISH-based mapping of the cosmid clone provide a possible identity and location for each.

In this pilot study, placental cDNA clones from a library prepared by Dr. Lee were used to identify cosmids from both chromosomes X and 17. Sixty unique cDNAs were identified, of which 22 were novel when compared with the sequence databases while another four cDNAs showed homologies to ESTs. Thirty-four cDNAs showed significant homology to previously described eukaryotic genes. Of the described gene matches, 26 were from human while the remainder were homologous to rodent, chicken and nematode sequences. The average size of the cDNA inserts was 1.5 kb. Nine cDNA clones (xp161, xp22G3, 6B10, 2D9, 2E8, 9F3, 12B5, 15C8, 24F3) contained both the published 5' and 3' untranslated regions of the corresponding genes, demonstrating that they contain the complete protein coding region of their respective mRNAs. This pilot study identified eight genes previously mapped to the X chromosome and eight mapped to chromosome 17, suggesting that it is a viable technique for identification of chromosome-specific genes (Tables 1 and 2). Some items of interest were found in these data. Four cDNAs (5G12 and xp278 in Table 1 and 3G12 and 8D1 in Table 2) identified cosmids mapping to two or more locations on the chromosome. These may represent members of gene families, pseudogenes, etc., and provide a measure of the selectivity of the method. As this is a hybridization-based approach, it provides the ability to identify imperfect matches between cDNA and genomic clone. While this generates uncertainty regarding the derivation of the cDNA, it offers the opportunity to identify chromosome-specific genomic clones containing sequences related to the cDNA, which may be interesting in themselves. This is also seen in the identification of cosmids by cDNAs that are clearly not X-linked (cytokeratins 8 and 18, for example), but identify homologous X clones (An X-linked locus homologous with cytokeratin 18 has been previously reported. The method offers a rich source of genomic clones for analysis, and the cDNAs identified are also of interest, but with caution regarding authentic map position.

The efficiency of the reciprocal probing strategy was estimated through analysis of the 32 chromosome 17 genes identified in the pilot study. The 2592 randomly isolated placental cDNA clones were pooled as probes and identified 278 cosmids from the 20,000 clones in the Los Alamos chromosome 17 (LA17NC01) cosmid library. Of these, 75 could not be associated with a cDNA and appear to have been false positive clones. The remaining 203 (73%) cosmids were true positive signals as identified by pooled cDNA probes. Hybridization with individual cDNAs separated the 203 cosmids into 35 non-overlapping groups associated with a specific cDNA. Thirty-two unique genes were isolated, sequenced and mapped with the 35 cosmid groups. Three of these cosmid groups generated cDNAs that were problematic during growth or sequencing. EST/STS primers generated from 24 of 27 clones (88%—only three of the nine genes known to map to 17 were used) were found to amplify both the associated cosmid and members of a chromosome 17 somatic cell hybrid panel. These data suggest that three of four cosmids identified by the pooled cDNA probes will be real positives, and the associated cDNA will map to the expected chromosome at a frequency of 90%.

cDNA	Cosmid	Similarity	ACC#	Position
9H4	65G1	Human ubiquitin	M26880	Xp22.2
2A11/2E8	78H10	Human HMG-17	X13546	Xp22.1
xp22G3	250A1	**Human pyruvate dehydrogenase**	HUMPYDH	Xp22.1
xh95C10	49H4	**Human dystrophin**	HUMDYS	Xp21.1
13G8	152B1	**Human monoamine oxidase**	X60819	Xp11.3
12B5	11G7	Mouse initiation factor 4AII	X56953	Xp11.2
xp519	210G5	Mouse MOPC gene	M11515	Xp11.2
11B6	22E7	EST03649	T03649	Xp11.2
2D9	150G11	Human cytoskeletal gamma actin	X04098	Xp11.2
18G2/9G12	230B5	Human cytokeratin 8	X74929	Xp11.1-q11
14C1	131A10	**Human moesin**	M69066	Xq12
10C7	7F4	Human haptoglobin related protein	M69197	Xq12-q13.1
9F3	185G3	Human homolog of mouse P21	X64899	Xq13
5G12/xp1515	147A10 133D1 22D2 27D5 90G4	Human elongation factor 1 alpha	X03558	Xq21.1-21.2 Xq21 + q25 + q27.3 Xq22 Xq23 Xq27.3-q28
2G6	128G2	Human initiation factor 4AI	D13748	Xq21.3 + Yp11.3
xp22F9	111C9	**Human alpha galactosidase**	HUMGALX	Xq22
3B4	197E11	Human vacuolar proton ATPase	X71490	Xq24
6B10	127E9	**Human ADP/ATP translocase**	J03591	Xq24-q25
xp278	236D6	Human Ku autoantigen	S38729	Xq27.3-q28
xp161	231E6	**Human iduronate 2-sulphatase**	M58342	Xq27.3-q28
1F11/xp22F2	155E7	ESTHSAAACDCK	Z20438	Xq27.3-q28
xp428	92G1	Human pregnancy specific beta glycoprotein	M93705	Xq27.3-q28
6G6	189D3	Human cytokeratin 18	X12883	Xq27.3-q28
3A1	52G10	**Human DXS1357E**	Z31696	Xq27.3-q28
7B11	4B4	Gallus filamin	U00147	Xq27.3-q28
6G3	78B10	**Human MPP1**	M64925	Xq27.3-q28

Table 1. X-associated cDNAs exhibiting significant similarity with sequence database entries. cDNAs and cosmid names are provided along with the highest match in the database and accession number. Locations of cosmid clones determined by FISH are provided. Genes in bold type were previously known to map to the X chromosome.

The sensitivity of hybridization with a pooled cDNA probe is an important parameter. If a single clone in a large pool is incapable of identifying its corresponding cosmids, this transcript will be missed by the method. Clone redundancy was measured for positive cDNAs in the chromosome 17 pilot. Of the 32 positive cDNAs, 15 were found in single copy in their respective pools of 864, while 11 were in two copies, two were in three copies, one had four copies and one each had seven and eight copies. Clearly, cDNAs in single copy in the pool are capable of identifying corresponding cosmids.

cDNA	Cosmid	Location
7D8	93B6	Xp22.3 + Yp11.3
3G12	93B6 152G7 127E9	Xp22.3 + Yp11.3 Xp11.2 Xq24
xp664	14G3	Xp11.3-p11.2
8D1	47G2 27C7	Xp11.2 Xq23
xp587	2C11	Xq22
6A11	122H2	Xq22
12D8	70B7	Xq23-24
7E7	197E11	Xq24
xp464	33B2	Xq24-25

Table 3. X-associated cDNAs exhibiting no similarity with sequence database entries. cDNAs and cosmid names are provided with locations of cosmid clones determined by FISH are provided.

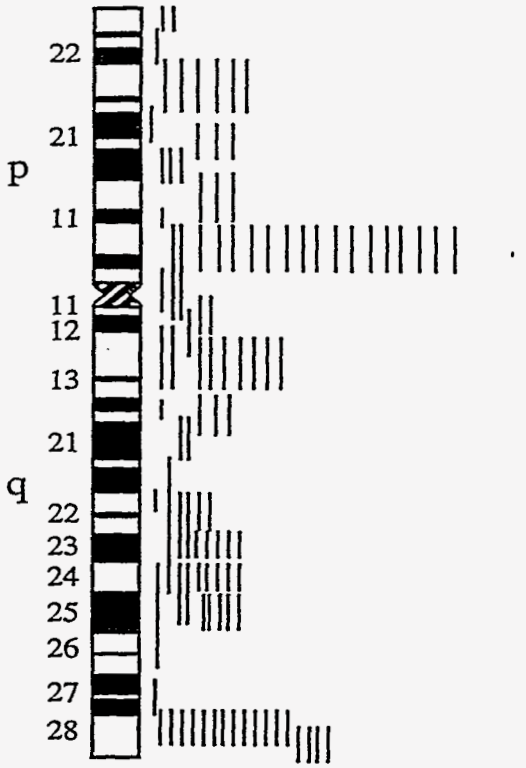

Figure 1. Positions of gene-associated cosmids mapped by FISH (103 total).

A scaled-up series of experiments is in progress, with a goal of 60,000 cDNAs to be analyzed in the same two genomic libraries. Two-thirds of these cDNAs (40,000 placental and heart clones) have been grouped into 46 pools of 864 clones each, and used as probes on the LLNL library. A total of 2848 cosmids were identified and isolated as individual clones. This represents over 10% of the clones in the library—a substantial proportion of the chromosome. Cosmids associated with each cDNA pool were re-arrayed for further manipulation. Individual back hybridization of the cosmids is impractical, so cosmids have been grouped into pools of ten (pilot studies demonstrated the feasibility of this approach with known positives) for hybridization to the cDNA filters. To date 130 cosmid pools have been used, identifying nearly 400 potential X-specific cDNAs. To identify individual cDNA-cosmid associations, cDNA clones identified will be individually hybridized to the rearrayed cosmid clones. The total number of hybridizations required by this method to identify the cDNAs with X chromosome association from the initial 40,000 cDNA clones is estimated to be ~1000 (46 initial cDNA pools; 285 cosmid pools; 800 individual cDNAs). This is a large number of hybridizations, but is manageable, especially when considering that the targets for hybridization are successively smaller in subsequent experiments. For example, the initial cDNA pool probes are applied to 25,000 clones, the cosmid probes are applied to an 864 clone array, and the individual cDNA probes are used against the cosmids identified by their parent pool (average of ~60 cosmids). Thus the scale and difficulty of each hybridization step is successively reduced.

At the present time over 100 unique cDNA-cosmid associations have been made (including those described in the pilot studies) for the X chromosome. The majority of the cosmids have been mapped by FISH (Figure 1), and cluster into three major regions already known to be gene rich (Xp11, Xq13 and Xq28). EST/STS primers have been generated from 52 of the cDNA sequences. These primers amplified the associated cosmid for 28/52 (54%) cDNAs. Not included in this analysis were nine genes previously assigned to the X chromosome. Taking these nine into account, 37 of 61 (60%) cDNAs isolated clearly derive from the X chromosome. This number is

somewhat lower than that found for chromosome 17, which may reflect a difference in the number of gene families and/or pseudogenes present on the X chromosome. Nevertheless, these data suggest that six of ten cDNAs isolated by the reciprocal probing method will be derived from the X chromosome, a significant enrichment over mapping cDNAs at random.

If the 800 potential cDNAs identified in these studies follow this pattern, then ~350 will truly be X-specific after the redundancy in the cDNA library is taken into account. This represents a two-fold increase over currently known genes. To generate this number of unique cDNAs mapped to the X through random assignment of cDNAs to chromosomes would require mapping of 7000 perfectly normalized cDNAs, an effort whose scale has not been approached. Another major advantage of the method is that it directly provides one or more genomic clones for the locus of interest. Since a reference cosmid library is used, it offers the ability to immediately check for information regarding the clone under study, even at the stage of initial identification of the cosmid with the cDNA pool.

Parallel efforts in the laboratories of Drs. David Nelson, Huda Zoghbi, Andrea Ballabio, and A. Craig Chinualt have aimed to establish physical map data for the LLNL flow sorted X cosmids. Through the NIH funded BCM genome center, informatics systems for managing the coordination of these data have been developed. Grid positions of cosmids identified either by YAC hybridization or reciprocal probing is entered and maintained in a Sybase database, the Cosmid Relational Data Base (CRDB). The LLNL library is composed of ~24,000 clones, and data regarding over 8000 of these has been developed thus far. The physical map data is largely found for clones in the distal Xq and Xp regions, with some efforts in proximal Xq as well. Cross-referencing of these data offers ready identification of fine-scale map positions for genomic clones associated with cDNA clones, simplifying and refining the mapping of the genes giving rise to the cDNAs.

From about 1300 of these cosmids, 206 unique cDNA-cosmid associations have been established, approximately doubling the number described in the preliminary publication. Of these 206 cDNAs-cosmid associations, 78 were from a heart cDNA library and 128 were from the placenta library. YAC probing of the cosmid library has revealed over 1000 "common" cosmids also identified by the cDNAs pool probes by querying the CRDB. In the majority of cases, several cosmids found to be in a common physical mapping interval (bin) are positive with the same cDNA. This offers substantial evidence of the validity of both the bin assignment and the cDNA assignment. Numerous examples in both Xq and Xp have been found, and many of these have been taken to paired genomic/cDNA sequence to establish the identity of the coding sequence.

Continuing efforts in these areas are focused on continued YAC-association of cosmid clones in preparation for long-range sequencing, and further characterization of the associated cDNAs to define the gene content of these regions of the X chromosome. No additional cDNA libraries are being used in these efforts due to inadequate funding.

Students trained:

Matt Mulloy (summer undergraduate student)

Postdoctoral fellows supported:

Ali Yazdani, M.D. (in clinincal training)
Zhouyang Zhao, Ph.D. (in postdoctoral training)

Publications (*indicates specific citation of support)

*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.

*Lee, CC, Yazdani, A, Wehnert, M, Zhao, Z, Lindsay, EA, Bailey, J, Coolbaugh, M, Couch, L, Xiong, M, Chinault, AC, Baldini, A, Caskey, CT. (1995) Isolation of chromosome-specific genes by reciprocal probing of arrayed cDNA and cosmid libraries. *Hum Mol Genet*, 4:1373-1380.

Zhao, Z, Lee, CC, Baldini, A, Caskey, CT. (1995) A human homologue of the *Drosophila* polarity gene *frizzled* has been identified and mapped to 17q21.1. *Genomics*, 27:370-373.

Zhao, Z, Lee, CC, Jiralerspong, S, Juyal, RC, Lu, F, Baldini, A, Greenberg, F, Caskey, CT, Patel, PI. (1995) The gene for a human microfibral-associated glycoprotein is commonly deleted in Smith-Magenis syndrome patients. *Hum Mol Genet*, 4:589-592.