This proposal will generate a "whole genome" data resource to enable rapid and reliable sequencing of genomic DNA by the definition and characterization of the more than 52 regions of high homology now known to be distributed within unrelated genomic regions and cloned in BACs and PACs. To do this, we will:

1. Define regions of true homology in the human genome by characterizing subsets of the 4,700 BAC/PACs that generate multiple hybridization signals using fluorescence in situ hybridization (FISH). Of the 1,200 sites of multiple signals, more than 52 regions contain repeats as defined by 600 BAC/PACs. The chimerism rate, multiple clone wells, and chromosome of origin will be defined by re-streaking each clone, followed by fingerprint, FISH and PCR-based end-sequence analyses on hybrid panels and radiation hybrids.

Data will be shared with large sequencing efforts, deposited in the 4D database, available with annotation on ftp server and through GDB.

2. Generate contigs of BACs and PACs in regions of complex genome organization. Using STS, EST analyses, fingerprinting, BAC/PAC to BAC/PAC Southerns, end sequence walking in 3.5-20X libraries, and metaphase/interphase FISH, contigs will be seeded in 2-5 of the regions of known genome complexity, each of which is estimated as 2-5 Mb. These data will be used to evaluate and provide independent quality assurance of the STS and Radiation hybrid, and genetic maps in these regions. The most significant of these include 1p36/1q; 2p/q; multiple sites; 8p23 and 8 further sites; 9p/q.

3. Define additional regions of complex genomic structure. Library screening using known members of multiple member retro-transposon and other known repeated sequences defined by the ncbi database, followed by FISH analyzes to determine structure and potential large regions of associated homologies.

Collaboration with other genome and sequencing centers will provide quality control in the generation of sequence-ready maps for sequencing templates.

We believe that this effort is important since 1) it will provide a critical mapping tool necessary for the generation of sequence ready maps; 2) if initiated now, the problem areas could be delineated before scale ups to full production occur in major genome centers; 3) represents a modest cost such that the cost of these data would comprise only a small fraction of the cost of the entire genome sequence and would vastly decrease the cost of sequencing errors 4) and could be completed in a, short time (2 to 3 years) so as to be of maximum benefit to sequencing centers. The Principal Investigator in this project is ideally suited for this effort because the group has developed the technology and initiated FISH and genome analyses of over 4000 clones.

We believe that this project represents a critical and timely effort to enable rapid and cost effective human genome sequencing.

Subcontract under Glen Evans' DOE Grant No. DE-FC03-96ER62294.

Mapping and Sequencing of the Human X Chromosome

D. L. Nelson, E.E. Eichler, B.A. Firulli, Y. Gu, J. Wu, E. Brundage, A.C. Chinault, M. Graves, A. Arenson, R. Smith, E.J. Roth, H.Y. Zoghbi, Y. Shen, M.A. Wentland, D.M. Muzny. J. Lu, K Timms, M. Metzger, and R.A. Gibbs
Department of Molecular and Human Genetics and Human Genome Center; Baylor College of Medicine; Houston, TX 77030
713/798-4787, Fax: -6370 or -5386, *nelson@bcm.tmc.edu*
http://www.bcm.tmc.edu/molgen

The human X chromosome is significant from both medical and evolutionary perspectives. It is the location of several hundred genes involved in human genetic disease, and has maintained synteny among mammals; both of these aspects are due to its role in sex determination and the haploid nature of the chromosome in males. We have addressed the mapping of this chromosome through a number of efforts, ranging from long-range YAC-based mapping to genomic sequence determination.

YAC mapping. The YAC-based map of the X is essentially complete. We have constructed a 40 Mb physical map of the Xp22.3-Xp21.3 region, spanning an interval from the pseudoautosomal boundary (PABX) to the Duchenne muscular dystrophy gene. This region is highly annotated, with 85 breakpoints defining 53 deletion intervals, 175 STSs (20 of which are highly polymorphic), and 19 genes.

Cosmid binning. The YAC-based physical is being used in a systematic effort to identify and sort cosmids prepared at LLNL from flow sorted X chromosomes into intervals. Gene identification through use of a common database for cDNA pool hybridization data is continuing. Over 50 YACs have been utilized as probes to the gridded cosmic arrays. These have identified over 9000 cosmids from the 24,000 member library. An additional 4000 cosmids have been identified using a variety of probes, with the bulk coming from cDNA pool probes. More recent emphasis has been placed on BAC clones as their identity for sequencing has been established. These have been identified using the usual methods.

Cosmid contig construction. Creation of long-range continuity in cosmids and BACs proceeds from clones identified by the YAC-based binning experiments. Identification of STS carrying clones is carried out by a combined PCR/