

EST Sequencing for Gene Discovery in Chinese Hamster Ovary Cells

Katie Fraass Wlaschin,¹ Peter Morin Nissom,² Marcela de Leon Gatti,¹ Peh Fern Ong,² Sanny Arleen,² Kher Shing Tan,² Anette Rink,³ Breana Cham,² Kathy Wong,² Miranda Yap,² Wei-Shou Hu¹

¹University of Minnesota Department of Chemical Engineering and Materials Science, 421 Washington Avenue SE, Minneapolis, Minnesota 55455-0132, telephone: (612) 626-7630; fax: (612) 626-7246; e-mail: acre@cems.umn.edu

²Bioprocessing Technology Institute A*STAR, #06-01 Centros, 20 Biopolis Way, Singapore

³Animal Disease and Food Safety Laboratory, 350 Capitol Hill Avenue, Reno, Nevada

Received 20 August 2004; accepted 24 February 2005

DOI: 10.1002/bit.20511

Abstract: Chinese hamster ovary (CHO) cells are one of the most important cell lines in biological research, and are the most widely used host for industrial production of recombinant therapeutic proteins. Despite their extensive applications, little sequence information is available for molecular based research. To facilitate gene discovery and genetic engineering, two cDNA libraries were constructed from three CHO cell lines grown under various conditions. The average insert size for both libraries is approximately 800–850 bp, and each library has comparable redundancy levels of 36%–38% for the sequences isolated. Random sequencing of 4,608 ESTs yielded 2,602 unique assemblies, 76% of which were annotated as orthologs of sequences in the GenBank database. A high abundance of mitochondrial genome transcripts facilitated the assembly of the complete mitochondrial genome by PCR walking. Comparative analysis of sequences from both mitochondrial and nuclear genomes with orthologous genes from other species shows that CHO sequences are generally most similar to mouse; however, examples with highest similarity to rat or human are common. A cDNA microarray, including all 4,608 ESTs, was constructed. The microarray results reveal a high level of consistency between transcript abundance in the libraries and fluorescence intensities. Inclusion of redundant clones in the microarray, additionally, allows small changes in abundant mRNAs to be discerned with a high degree of confidence. The information and tools generated provide access to genomic technology for this important cell line. © 2005 Wiley Periodicals, Inc.

Keywords: Chinese hamster ovary cells (CHO cells); EST sequencing; mitochondrial genome; cDNA microarray

INTRODUCTION

Established mammalian cell lines are invaluable research tools, providing in vitro models for a wide range of biological investigations. CHO cells (derived from Chinese hamster ovary) (Puck et al., 1958), HeLa cells (derived from human cervix carcinoma) (Gey et al., 1952), and NIH 3T3 cells (derived from mouse embryo) (Todaro and Green, 1963) are among the most widely used cell lines in published literature. CHO cells, in addition to their importance in research, are, arguably, the most utilized host for large-scale production of pharmaceutically important proteins, accounting for billions of dollars in therapeutic protein products annually (Andersen and Krummen, 2002; Chu and Robinson, 2001). Despite their value in biomedical research and industrial pharmaceutical protein production, genome-based resources for CHO cells have not been developed, and less than 700 Chinese hamster DNA sequences can be found in the GenBank database (March, 2004), as opposed to both HeLa and 3T3 cells, for which the entire human and mouse genome sequences are available.

CHO cells have many attractive characteristics that contribute to their popularity as a research tool. They have been used extensively to obtain several nutritionally deficient mutant cell lines (Kao and Puck, 1968). In these experiments, the frequency of isolation of recessive mutants did not decrease, as expected, when cell lines of abnormal ploidy (triploid, tetraploid, aneuploid) were used (Chasin, 1973; Chasin and Urlaub, 1975; Harris, 1971). Through such observations, CHO cells are regarded as functionally haploid at many genetic loci, making them useful in elucidating molecular functions through isolation of mutants. The relative ease of genetic manipulation earned them popularity as a research tool. Genetically modified cell lines are used in research of signal transduction, cytoskeletal structure, cell

Correspondence to: Wei-Shou Hu

Contract grant Sponsor: University of Minnesota Supercomputing Institute

Contract grant Sponsor: NSF fellowship

Contract grant Sponsor: NIH Biotechnology Training

Contract grant number: GM08347

cycle control, DNA repair, toxicology, and even drug addiction. Additionally, CHO cells are highly adaptable to a variety of growth conditions. The molecular nature of their adaptability is not completely understood, but it has been suggested that epigenetic factors may play a role (Holliday and Ho, 1998; Holliday et al., 1996; Paulin et al., 1998; Siminovitch, 1976).

The isolation of a CHO cell mutant with functional deficiencies at the dihydrofolate reductase (*dhfr*) locus (Urlaub and Chasin, 1980) led to the DHFR selection system, which is widely used for expressing and amplifying heterologous proteins (Kaufman and Sharp, 1982; Kaufman et al., 1987). These foreign proteins are produced and secreted into the medium fully glycosylated and biologically active (Geisse et al., 1996). Additionally, the glycoforms produced by CHO cells are very similar to human glycoproteins (Warner, 1999). These properties have distinguished CHO cells as one of the principle cell lines of choice for expression of heterologous genes for therapeutic applications (Chu and Robinson, 2001). CHO cells have been engineered to produce a variety of commercial products including factor VIII (Kaufman et al., 1988; Wood et al., 1984), tissue plasminogen activator (TPA) (Kaufman et al., 1985), erythropoietin (EPO) (Lin et al., 1985), tumor necrosis factor (TNF) (Korn et al., 1988), human interferon- γ (Scahill et al., 1983), and numerous recombinant antibodies (Korke et al., 2002).

As a host for heterologous protein expression in biotechnological applications, CHO cells have exceeded expectations; yet, the genetic and physiological properties that render them such capable protein producers and secretors are not well characterized. The approach for obtaining a high-producing clone for a particular product is laborious, and the results vary greatly for different heterologous proteins. There is little understanding of how much further CHO cells can be engineered to become better and more consistent producers.

It is clear that high-throughput, hybridization-based techniques, especially cDNA microarrays, which are now common practice in biological and biotechnological research (Levy-Nissenbaum et al., 2003; Marcotte et al., 2003; Russo et al., 2003; Seta and Millhorn, 2004), will greatly facilitate gene discovery and cell engineering research in this important cell line. These tools have not been extensively employed in CHO cell research because little sequence information is publicly available. EST sequencing, beginning with the human genome project (Adams et al., 1991), has provided a useful means for obtaining sequence data for gene discovery and functional sequence annotation. The potential for gene discovery and elucidation of gene regulation through EST sequencing provides the motivation for our efforts to develop CHO cell sequence data and a cDNA microarray. Our effort is one of the first, large-scale sequencing efforts specifically for cultured mammalian cell lines designed for application in industrial recombinant protein production. In this communication, we report the initial results of our efforts in EST isolation, sequencing, and annotation.

MATERIALS AND METHODS

Cell Culture

DXB-11 cells (Urlaub and Chasin, 1980) were maintained in a modified DMEM:F12 (1:1) media with the following components added or adjusted: glucose (17.5 mM), glutamine (4 mM), sodium bicarbonate (29 mM), ascorbic acid (0.11 mM), putrecine (6.2 μ M), penicillin G (0.17 mM), streptomycin (68.6 μ M), pluronic F68 (0.01%), phenol red (19.9 μ M), apotransferrin (63.7 nM). In addition, the medium was supplemented with thymidine (41.3 μ M) and FBS (5% v/v) (Atlas Biologicals, Ft. Collins, CO). Recombinant cells, derived from DXB-11, were maintained in the same modified DMEM:F12 (1:1) media, and were additionally supplemented with 0.15 μ M methotrexate and Intralipid (0.01%) (Sigma Aldrich, St. Louis, MO). For RNA isolation, DXB-11 cells were grown in roller bottles in a 37°C, 5% CO₂ environment. Cells were harvested at the late exponential phase of growth. The recombinant cells were grown in two 250-mL spinner flasks (100-mL working volume). Cells were isolated on the second day of culture at a cell density of $\sim 1.1 \times 10^6$ cells/ml and 91% viability.

CHO-IFN- γ cells were derived from *dhfr*⁻ DXB-11 cells (Urlaub and Chasin, 1980). They have been further adapted for growth in suspension culture. CHO-IFN- γ cells are maintained in HyQ CHO MPS media (Hyclone, Logan, UT) supplemented with 4 mM glutamine and 0.25 μ M methotrexate in a 37°C, 5% CO₂ environment. For RNA isolation, cells were grown in a 2-L bioreactor (B. Braun, Melsungen, Germany) with a working volume of 1.2 L. 1×10^8 cells were harvested in exponential phase (48 h at 0.75×10^6 cells/mL and 98% viability), early stationary phase (96 h at 2.40×10^6 cells/mL and 97% viability), and late stationary phase (120 h at 2.20×10^6 cells/mL and 85% viability).

Library Construction

Library A

Total RNA was extracted from DXB-11 and recombinant cell samples using Trizol reagent according to the manufacturer's protocol (Invitrogen, Carlsbad, CA). mRNA was purified from total RNA using Oligotex[®] mRNA isolation kits (Qiagen, Valencia, CA). 2.5 μ g of mRNA from each of the cell samples was used to construct a phage library using the Stratagene ZAP Express cDNA synthesis kit and the ZAP Express cDNA Gigapack III Gold Cloning Kit (Stratagene, La Jolla, CA), according to the manufacturer's protocols.

Library B

Total RNA from CHO-IFN- γ cells were extracted from each sample individually using Trizol reagent according to the manufacturer's protocol (Invitrogen). Total RNA from various growth stages was diluted in nuclease-free water to a final concentration of 5 μ g/ μ L. Three individual libraries

were constructed from the total RNA using Clontech SMART cDNA synthesis kits, followed by directional ligation into the pDNR-Lib vector (BD Biosciences, Palo Alto, CA), according to the manufacturer's protocol. The three libraries were pooled in a 1:1:1 ratio to give the final working library.

Sequencing

In vivo mass excisions of plasmids from the phage library (library A) were carried out and used to transform competent *E. coli*. Library B plasmids were directly transformed to competent *E. coli* and plated similarly. Individual colonies were picked into growth and freezer compatible media of the following composition: 90% (v/v) 2 × YT (pH 7.4), 4.4% (v/v) glycerol, 13 mM KH₂PO₄, 36 mM K₂HPO₄, 1.5 mM sodium citrate, 4.1 mM MgSO₄ · 7H₂O, 7.1 mM (NH₄)₂SO₄, and 50 µg/mL kanamycin (library A) or 34 µg/mL chloramphenicol (library B). The liquid cultures were incubated for 24–36 h at 37°C, replicated, and frozen at –80°C.

Frozen stocks were used to inoculate 200 µL liquid cultures in LB + kanamycin (50 µg/mL) or chloramphenicol (34 µg/mL) media and were grown overnight at 37°C. Ten µL of the fresh cultures were used to inoculate 1.2 mL shake-cultures in 96-well deep-well grow blocks, which were grown for 24–36 h at 37°C and 250 rpm. Cells were pelleted by centrifugation and plasmids were isolated using QiaPrep 96-well Miniprep kits (Qiagen). Plasmid DNA (100–200 ng) was sequenced (5' end, single pass) with 5 pmol of primer (T3 for Library A, M13 for Library B) at the University of Minnesota Advanced Genetic Analysis Center (AGAC). The sequencing reactions were prepared with 5 × Sequencing Buffer and Big Dye Terminator Version 2.0 (Applied BioSystems, Foster City, CA), PCR amplified for 25 cycles in ABI 9700 thermocyclers (Applied BioSystems), purified using DyeEx 96 plates (Qiagen), and run in ABI 3700 96-well capillary sequence analyzers (Applied Biosystems).

Mitochondrial Genome Sequencing

Cell pellets of DXB-11 and the recombinant clone were washed and re-suspended in Buffer M (210 mM mannitol, 5 mM EDTA, 5 mM HEPES, 70 mM sucrose, pH 7.4) and then disrupted with a cell disruption bomb (N₂ cavitator) at 600 psi. Pellets containing mitochondria were collected after two centrifugation steps at 1,340g for 5 min and re-suspended in WB buffer (0.35M sorbitol, 50 mM Tris pH 7.6, 0.1% BSA).

Eight contigs containing likely mitochondria sequences from the cDNA libraries were aligned to the mouse mitochondrial genome (NC_001569) to identify un-sequenced gaps. The 5' and 3' ends of contigs near the gaps were used for primer design using the Primer 3 algorithm (http://frodo.wi.mit.edu/primer3/primer3_code.html). PCR reactions were performed on mitochondria samples from DXB-11 and the recombinant clone separately, using nine sets of primers. PCR products were purified using the QIAquick

PCR purification kit (Qiagen) and sequenced on an ABI 377 DNA fragment analyzer (Applied BioSystems) at AGAC. Phred/Phrap/Consed (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998) was used to assemble the mitochondrial genome sequence.

Sequence Analysis

EST Assembly and Finishing

Sequence chromatograms were manually screened for quality, then read, screened for vector and low quality sequence, and assembled into contigs using *msi_trim_phred-Phrap*, a modified version of Phred/Phrap/Consed, created at the University of Minnesota Supercomputing Institute (Dr. Zheng Jin Tu, personal communication). Default alignment parameters were utilized except *minmatch* = 50 and *minscore* = 100. The modified parameters decreased the occurrence of misassemblies by comparing larger fragments.

Comparison to Public Databases and Annotation

Sequences were compared to a local version of the GenBank nucleotide database using the basic local alignment search tool (BLAST) algorithm (blastall 2.2.5). BLAST results were obtained in XML format, and results were extracted and uploaded into an ORACLE database. Each sequence was manually annotated using a custom interface that displays BLAST information and allows users to select the annotation and indicate confidence level. Annotations, along with corresponding data, classifications, and comments, were also archived. Sequences that did not show significant homology to a cDNA were compared to the human and mouse genome databases using the NCBI genome blast web interface (<http://www.ncbi.nlm.nih.gov/BLAST>). Location was determined relative to known transcribed regions in both genomes.

Multiple Sequence Alignments and Dendograms

ClustalX (1.81) (Thompson et al., 1997) was used to perform multiple sequence alignments using default alignment parameters. Default bootstrapped N-J dendograms were drawn for each alignment and visualized with TreeView (v1.6.6) (Page, 1996).

Microarray

Microarray Printing

Inserts were amplified from CHO EST libraries in 96-well format using primers specific to the cloning vectors (T3/T7 or M13 forward/reverse). PCR reactions were performed in standard 50 µL reactions containing 0.5 U Taq polymerase (Fermentas, Hanover, MD), 100 µM primers, 100 µM each of dNTPs, and 1.5 mM MgCl₂. The cycling conditions were 95°C, 1 min, 60°C, 1 min, and 72°C, 2 min for 35 cycles.

Following amplification, PCR products were purified using 96-well Multiscreen[®] PCR purification plates (Millipore, Billerica, MA). Aliquots of each purified probe were analyzed by gel electrophoresis and quantified on a Tecan GENios spectrophotometer (Tecan, Maennedorf, Switzerland). The remainder was desiccated in a speed vac (Savant Instruments, Holbrook, NY) and re-suspended in phosphate printing buffer (100 mM potassium phosphate, pH 7.5, 10% glycerol, 1.5M betaine, 1 × SSC) at a final DNA concentration of at least 150 ng/μL.

Probes were spotted on polylysine-coated slides using a Biorad Chipwriter (BioRad, Hercules, CA) equipped with quill-type steel pins (Telechem, Sunnyvale, CA). Spots were printed at a nominal center-to-center spacing of 200 μM. Probes were immobilized and blocked as described previously (Diehl et al., 2001).

Microarray Hybridization

Total RNA was isolated from CHO DXB-11 and its recombinant derivative, grown under conditions identical to those used to obtain the library mRNA, except that the DXB-11 cells were grown in T flasks. Trizol was used for RNA extraction according to the manufacturer's protocol.

Five technical replicate microarray hybridizations were performed (two of the five with swapped dyes). For each microarray, cDNA was synthesized from 40 μg of total RNA using SuperScriptII (Invitrogen), according to the manufacturer's recommendations. The cDNA synthesis reaction was stopped by addition of NaOH and EDTA to 0.2M and 0.1M concentrations, respectively, and additional incubation at 65°C for 15 min. The reaction mixture then was neutralized with 1M Tris-HCl (pH 7.4). cDNA was purified using Microcon 30 concentrator columns (Millipore) by three washes with 500 μL of water and dried in a speed vac.

All subsequent manipulations were carried out in darkness to avoid photobleaching of Cy dyes: Indirect labeling with Cy3 and Cy5 dyes (Amersham Biosciences, Piscataway, NJ) was carried out by re-suspension of the cDNA and Cy dyes in 9 μL of 0.1M (pH 9) sodium bicarbonate buffer, and incubation for 1 h at room temperature. Following dye coupling, the unreacted dye was quenched in 4.5 μL of 4M hydroxylamine for 15 min at room temperature. The cDNA samples were cleaned using QIAQuick PCR Purification kits according to the manufacturer's protocol (Qiagen), and concentrated in a speed vac to a final volume of 16.8 μL.

The purified, labeled cDNA samples were combined with blockers and dye spikes in the following amounts: 1 μL mouse Cot1 DNA (25 μg/μL), 2 μL polyA (10 μg/μL), 2 μL yeast tRNA (4 μg/μL), 1 μL each of cloning vector (1 μg/μL), and 0.2 μL Cy3 labeled *B. subtilis* (X17013) oligonucleotide spike (1 μM). The probe/blocker mixture was heated to 96°C for 5 min and snap cooled on ice for 5 min. Twenty-five μL of pre-warmed (42°C) hybridization buffer (50% formamide, 10 × SSC, 0.2% SDS) was added to the above mixture and applied to the microarrays under lifter slips. The microarrays were hybridized for 12.5 h in a 42°C water bath. The

microarrays were removed from the hybridization chambers, submerged repeatedly for 1 min in 0.57 × SSC and 0.02% SDS, transferred into 0.04 × SSC and, again, submerged repeatedly for 1 min. Arrays were dried by centrifugation at 600 rpm for 5 min and stored in darkness until scanning.

Microarray Data Processing and Analysis

Microarrays were scanned using ScanArray[®] Express (Packard BioScience Company, Meriden, CT) Microarray Analysis System. Intensity data were extracted using GenePix[®] Pro 4.1 analysis software (Axon Instruments, Union City, CA). The data normalization was carried out with GeneSpring[®] (Silicon Genetics, Redwood City, CA) and median intensity values were normalized for each spot using the Loess method (Yang et al., 2001). Average median intensities, average natural logarithmic ratios, and *P* values were calculated for each clone on the array, excluding spots having natural logarithmic ratios that lie more than 1.5 times the standard deviation from the mean natural log ratio. Only data where four or more spots were used to perform calculations were considered (>98% of all spotted sequences). The filtered, normalized data was then uploaded into Spotfire v 6.3 (Spotfire, Somerville, MA) for visualization and further analysis.

RESULTS

Sequence Assembly, Characterization, and Annotation

Two CHO cDNA libraries were constructed from three CHO cell lines grown under a number of conditions exhibiting characteristics that are important to recombinant protein production. A phage library (library A) was constructed from a 50:50 pool of purified mRNA from two cell lines isolated in exponential growth phase: DXB-11 (parental clone, grown as adherent monolayers in 5% FBS) and a recombinant clone (derived from DXB-11, grown as suspension cells in serum-free media). A plasmid library (library B) was constructed by pooling three separate plasmid libraries (1:1:1) constructed using mRNA isolated from CHO cells producing recombinant human interferon-γ (IFN-γ) in different phases of growth (exponential, stationary, and late stationary phases) in serum-free, suspension cultures. Library A yielded 1 × 10⁶ primary transformants with an average insert size of 850 bp. Library B plasmids had an average insert size of 800 bp and were directly transformed into *E. coli* for random isolation.

Four thousand six hundred and eight randomly isolated clones from the two libraries were sequenced. Chromatograms were read, scanned for low quality and vector sequences, trimmed, and assembled into contigs using a modified version of Phred/Phrap/Consed to group sequences isolated more than once from the libraries. Single pass, 5' end sequencing yielded 4,219 high-quality sequences longer than 200 bp (excluding vector sequences). Two thousand eight hundred and eighteen sequences were isolated from library A

Table I. Summary of assembly results.

Sequence source	Library A	Library B	ALL
Total number of sequences analyzed	2,818	1,401	4,219
Number of ESTs in contigs	1,400	623	2,024
Number of contigs	313	111	407
Number of singlets	1,418	778	2,195
Number of unique sequences	1,731	889	2,602
Redundancy (%)	38.6	36.5	38.3

Sequences obtained from the libraries were aligned separately and as a group using a modified version of the Phred/Phrap scripts (msi_trim_phred-Phrap) with minmatch = 50 and minscore = 100. Redundancy is calculated as $100 \times (1 - \text{unique sequences}/\text{total sequences})$.

and 1,401 were isolated from library B. Redundancy and representation were assessed for each library throughout the sequencing effort (every four to five 96-well plates) to ensure that unique sequence information was being isolated at a reasonable rate. More sequences were obtained from library A because the redundancy was observed to be slightly lower for a similar number of clones (data not shown). Datasets representing all sequences isolated were assembled into contigs of overlapping sequence regions using Phred/Phrap/Consed for each individual library, and for all sequences obtained from both libraries. Data for each assembly are shown in Table I. Library A yielded 1,731 unique sequences with a redundancy of 38.6%, and library B yielded 889 unique sequences with a redundancy of 36.5%. A total of 2,602 unique sequences were obtained from the assembly of both library datasets together.

Analysis of the size and composition of the assembled contigs reveals that less than 30 cDNA species contribute significantly to the observed redundancy. Figure 1 is a graphical representation of the assembly data for all sequenced ESTs from both libraries. The figure shows the distribution of ESTs among the assembled contigs, illustrating the relative contributions of a particular transcript to the overall redundancy. The *x*-axis represents the different sizes of contigs generated by the assembly. The tallest bar at the far left of the graph shows 1,418 ESTs were not assembled into contigs. The next tallest bar represents 322 contigs containing only two ESTs each. The ESTs represented by the tall bars at the left of the graph do not contribute significantly to the observed redundancy of the libraries. The short bars on the right of the graph are formed from frequently isolated ESTs. The right-most bar (100+) represents 112 ESTs with significantly overlapping sequence regions that were assembled into a single contig (size = 112 ESTs). The twenty-one largest contigs are contained in the last seven bars of the graph. These bars comprise 919 ESTs. These highly present ESTs account for 65% of the observed redundancy. In other words, 22% of all isolated ESTs only contribute 1% of the unique sequence information.

The distribution of ESTs among contigs in Figure 1 is very similar for the individual assemblies of both library A and library B. Additionally, the identities of the ESTs that comprise the large contigs are similar in both cDNA libraries. The most significantly abundant ESTs that are common between the two libraries are transcripts from the mitochondrial genome. The frequency of isolation of mitochondrial

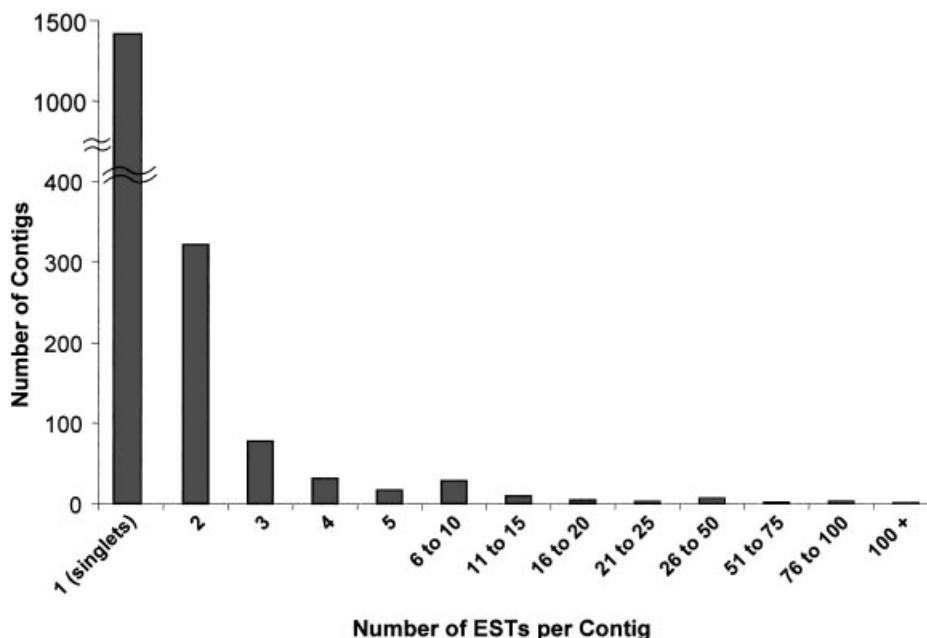


Figure 1. Distribution of redundantly isolated species into assembled contigs. Data summarizing the results of multiple sequence alignment of the ESTs using Phred/Phrap/Consed are plotted to illustrate the contributions of ESTs to unique sequence information. Each bar represents a contig size or range of sizes. The size of a contig refers to the number of ESTs assembled into a contig. The height of each bar represents the number of contigs that are formed, which are of those sizes. Each contig likely represents a unique sequence. The tall bars at the left of the graph illustrate that the majority of ESTs assemble into smaller contigs, containing fewer than 10 ESTs. These ESTs do not contribute as significantly to library redundancy as those ESTs and contigs represented at the right side of the graph. One contig contains more than 100 ESTs (112) and was assembled from transcripts from the mitochondrial genome.

genome encoded ESTs is shown in upper half of Table II. The number of times a protein-coding sequence from the mitochondrial genome was isolated differs from the contig numbers described in Figure 1. This occurs because mitochondrial genome transcripts are longer and sometimes represent more than one gene product. The mitochondrial genome is circular, similar to a bacterial genome, with continuous, sometimes overlapping, genes that are transcribed as a single mRNA transcript. Contig assemblies do not accurately represent repetitive isolation of a single protein-coding mRNA species. To give a more accurate representation of gene abundance, each EST from the mitochondrial genome was individually annotated.

Table II. Abundant ESTs.

Gene description for contig	I: Total number of clones	II: Number	III: Number
		of clones from library A	of clones from library B
Mitochondrial genome transcripts			
ND4/DN4L	137	103	34
ATP6/ATP8	101	34	67
CYTB	82	51	31
ND1	79	28	51
12S rRNA/16S rRNA	65	4	61
ND2	59	48	11
COX2	55	19	36
COX1	37	21	16
COX3	34	21	13
ND5	14	11	3
ND6	9	7	2
D-loop	4	2	2
Total	676	349	327
Novel sequence	31	0	31
DHFR	27	26	1
Immune (γ) interferon	24	0	24
Ribosomal protein L23a	23	19	4
Ribosomal protein L9	22	21	1
Cyclophilin	20	20	0
Ribosomal protein S3a	19	17	2
EF-1 alpha	16	13	3
Ribosomal protein L19	14	10	4
Ribosomal protein S27a	12	12	0
Ribosomal protein L6	12	11	1
Heat-shock 70 kDa protein 5/glucose regulated protein 78 kDa (GRP78)	12	3	9
Ribosomal protein L7a	11	11	0
Ribosomal protein L5	10	9	1
Ribosomal protein S17	9	7	2
Ribosomal protein L21	9	6	3
Ribosomal protein L35	9	9	0

The top 25 contigs containing the most ESTs were annotated using the BLAST algorithm. The most abundant species obtained from the libraries are listed along with the number of clones obtained from each library. The contigs containing the most ESTs comprised transcripts from the mitochondrial genome. These contigs were split to obtain the number of clones and percentages for each mitochondrial mRNA transcript. Column I shows the number of times each sequence was isolated in total (both cDNA libraries). Columns II and III show the number of times that a sequence was isolated from each individual library.

The relative abundance of each mRNA transcript varies slightly between the two libraries, although the four most abundant gene products, ND4/ND4L, ATP8/ATP6, CYTB, and ND1, were consistently abundant in both libraries, relative to the other protein-coding transcripts. Transcripts covering 12 of the 13 protein-coding regions and both mitochondrial rRNAs were obtained as ESTs. NADH dehydrogenase subunit 3 (ND3) is the only protein-coding sequence of the mitochondrial genome that was not isolated. A few transcripts corresponding to the mitochondrial D-loop region were also found. The combined mRNA sequences covered approximately 70% of the entire mitochondrial genome, as determined by alignment of the corresponding contigs to the mouse mitochondrial genome.

The bottom half of Table II lists the most frequently sequenced ESTs that are transcribed from the nuclear genome. Among these genes, several ribosomal protein transcripts were isolated from both libraries. Other abundant transcripts include cyclophilin, elongation factor 1- α (EF1- α), heat-shock protein GRP78, and an uncharacterized novel transcript. Genes related to amplification and expression of recombinant proteins (dhfr, interferon- γ) were also found at high levels from both libraries. Most of transcripts listed in Table II were found in both libraries, although in different relative abundance. Two ribosomal proteins and cyclophilin were isolated exclusively from library A, and one novel transcript and the recombinant INF- γ were isolated only in library B.

In addition to identifying and annotating the most redundantly isolated ESTs, sequences for all assembled contigs and for each individual clone were compared to the GenBank non-redundant nucleotide database using the BLAST algorithm. The results were manually reviewed and annotations were assigned based on a combination of score, *e*-value, and alignment length. Sequences showing significant alignments (*e*-value $< 10^{-10}$) were assigned as CHO orthologs of the most appropriate match. Sequences that aligned to a single locus of genomic DNA, but did not match any cDNA species, were annotated as genes of unknown function that map to a chromosomal locus. High-quality sequences, with an *e*-value greater than 10^{-10} , were annotated as novel. Sequences with significant alignments to several positions on multiple chromosomes, or to many non-orthologous ESTs, were assigned as repetitive element containing sequences.

A summary of the BLAST results for the complete unique gene dataset (2,602 sequences) is displayed as a pie chart in Figure 2. Seventy-four percent of the CHO sequences align in a meaningful way to a cDNA sequence in the GenBank database, which is represented by the segment labeled expressed sequences. Approximately 80% of these alignments ($\sim 1,550$ sequences) are to sequences that have been assigned meaningful biological descriptions through sequence similarity or experimental confirmation (data not shown). The remaining alignments are to ESTs of unknown function or to ESTs with predicted functions based on motif-scanning algorithms. Two percent of the unique sequences

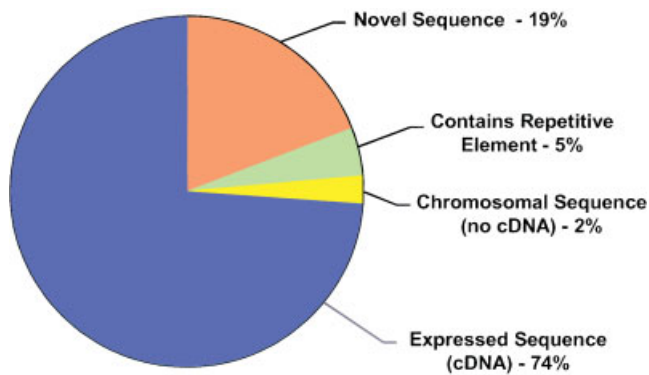


Figure 2. BLAST comparison of 2,602 unique Chinese hamster ovary (CHO) sequences to the GenBank non-redundant nucleotide database. Sequences for each assembled contig were compared to the GenBank non-redundant nucleotide database using the BLAST algorithm. The results were manually reviewed and annotations were assigned based on a combination of score, *e*-value, and alignment length. Sequences showing significant alignments (*e*-value < 10⁻¹⁰) were assigned as CHO orthologs of the most appropriate match. Sequences that aligned to a single locus of genomic DNA but did not match any cDNA species were annotated as genes of unknown function that map to a chromosomal locus. High-quality sequences, with an *e*-value greater than 10⁻¹⁰, were annotated as novel. Sequences with significant alignments to several positions on multiple chromosomes, or to many non-orthologous ESTs, were assigned as repetitive-element containing sequences. The pie chart shows the distribution of the BLAST results according to the ability to assign an annotation based on homology to GenBank cDNA sequences.

align significantly to a single chromosomal locus, but have not been isolated in cDNA form. Most of those significant chromosomal DNA alignments are to mouse chromosomes, but approximately 25% are more similar to a locus on a human chromosome. After identifying the relative positioning of these ESTs on the mouse and human chromosomes, 90% of them were shown to fall within or near a mapped EST or an open reading frame (ORF). The remaining ESTs in Figure 2 are novel (19%) or contain a conserved repetitive element (5%). Translation of a portion of the novel sequences using the online version of ESTScan (<http://www.ch.embnet.org/software/ESTScan.html>) (Iseli et al., 1999) yielded either no ORFs or only short ORFs that do not significantly match known protein sequences (data not shown). Among ESTs containing repetitive elements, two primary families were identified. Each element is approximately 100–130 bp in length and was found in multiple non-redundant ESTs. One of the repetitive elements accounts for 40% of the repetitive-element containing clones and is a likely Alu-type element, showing similarity to the mouse Alu sequence. The other sequence, which is present in 30% of the repetitive element containing clones, does not show homology to any other known repetitive elements.

org/software/ESTScan.html) (Iseli et al., 1999) yielded either no ORFs or only short ORFs that do not significantly match known protein sequences (data not shown). Among ESTs containing repetitive elements, two primary families were identified. Each element is approximately 100–130 bp in length and was found in multiple non-redundant ESTs. One of the repetitive elements accounts for 40% of the repetitive-element containing clones and is a likely Alu-type element, showing similarity to the mouse Alu sequence. The other sequence, which is present in 30% of the repetitive element containing clones, does not show homology to any other known repetitive elements.

Comparative Sequence Analysis

The GenBank non-redundant nucleotide BLAST results were examined to determine the overall level of similarity of CHO cell sequences to sequences isolated from other organisms. Since there is little information available for hamster species, it is important to determine which organism(s) with large amounts of sequence information available will be useful in further CHO genome analysis and for creating probes to isolate genes that are not obtained in random sequence isolation. Table III lists the primary species providing significant BLAST hits for the CHO ESTs along with the number of CHO sequences which were annotated based on having the most significant similarity to a DNA sequence from each particular species. The total number of nucleotide sequences in the GenBank database for each species (at the time of the BLAST comparison) is also listed.

Only 286 of the 2,602 unique CHO sequences were previously identified in any hamster species; thus, the vast majority of the information generated by this study (2,316 sequences) is new. Species that provide the majority of significant BLAST alignments are mouse, rat, and human. The largest fraction of the BLAST results show that mouse sequences provide the highest level of similarity. Rat is the next most similar species, followed by human. Since a similarly large amount of information is available in the

Table III. Distribution of BLAST results among species.

Species	I: GenBank nucleotide entries	II: Number of Chinese hamster ovary (CHO) sequences	III: Percent of CHO sequences (%)
Chinese hamster (<i>Cricetulus griseus</i>)	648	248	8.1
Golden hamster (<i>Mesocricetus auratus</i>)	1,485	31	1.2
Long-tailed hamster (<i>Cricetulus longicaudatus</i>)	118	7	0.3
Mouse (<i>Mus musculus</i>)	5,842,417	1,551	59.6
Rat (<i>Rattus norvegicus</i>)	1,051,971	469	18.0
Human (<i>Homo sapiens</i>)	8,343,251	223	8.6
Other species	n/a	110	4.2

Unique assemblies were compared to the GenBank non-redundant nucleotide database using the BLAST algorithm. The first hit from each result was extracted and the number of CHO sequences that match to each species is tabulated in Column II (Column III shows the percent distribution of CHO sequences among the listed species and is calculated based on numbers in Column II). The number of annotations based upon a particular species depends on both sequence similarity and presence of the orthologous sequence in the database; thus, Column I shows the number of nucleotide entries for each species in the GenBank database at the time of the BLAST comparison.

GenBank databases for mouse, rat, and human (entire genome sequences are available for all three), the larger percentage of mouse BLAST hits suggest that mouse is the closest relative among species for which a large amount of data is available. This is not the case for all genes. Some of the CHO sequences are more similar to rat sequences than to mouse sequences. Putative ribosomal protein encoding sequences from CHO cells are typically most similar to rat sequences. Less frequently, a human sequence provided the closest match, even when orthologous mouse and rat sequences were available. Mouse sequence information will provide a sufficient basis for comparative analysis of CHO ESTs; however, integration of information from human and rat will be useful and, in some cases, better than mouse data.

To further examine the cases where mouse sequences do not provide the most significant alignment, ClustalX was used to obtain multiple sequence alignments and construct

dendrograms for a few selected ESTs. Identity matrices from the alignments and their corresponding dendrograms for cytochrome C oxidase subunit Vic (COX6C), ribosomal protein S19 (RPS19), lactate dehydrogenase A (LDHA), and calmodulin 2 (CALM2) are shown in Figure 3a–d. The dendrograms generated from the alignments of sequences for COX6C, RPS19, and LDHA (Fig. 3A–C) are very similar in structure. For all three of these alignments, the mouse and rat sequences are most similar to one another. Additionally, the Mouse:CHO and Rat:CHO alignment identities are approximately equal, although slight identity differences are illustrative of the cases where rat sequences provide more significant BLAST alignments.

The dendrogram generated for the CALM2 sequence alignment of Figure 3d has very different structure than the other three dendrograms described. This is an example of a CHO sequence where similarity to human provides the best

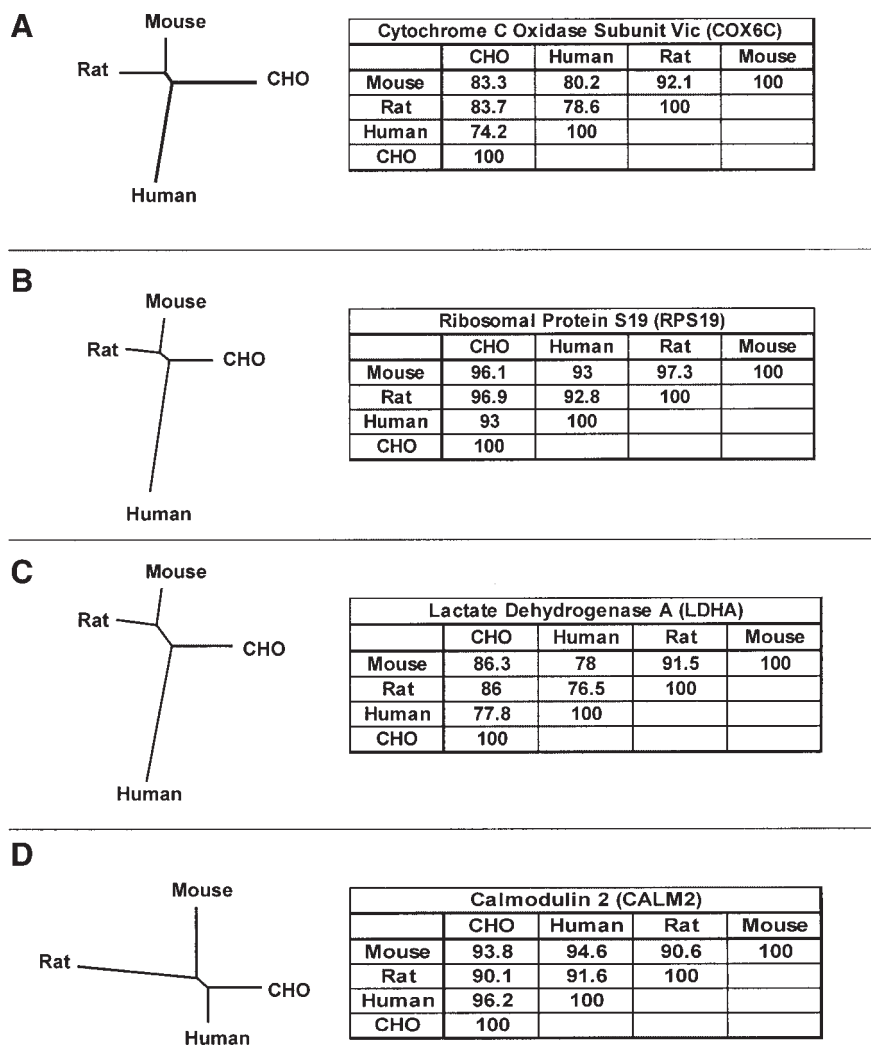


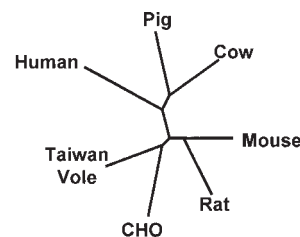
Figure 3. Multiple sequence alignments using ClustalX. ClustalX (version 1.81) was used to align sequences for (A) cytochrome C oxidase subunit 6 (COX6C) (GenBank sequences: Mouse/BC024666, Rat/NM_019360, Human/NM_004374); (B) ribosomal protein S19 (RPS19) (GenBank sequences: Mouse/NM_023133, Rat/XM_214822, Human/NM_001022); (C) (Theodorakis et al., 1996) lactate dehydrogenase A (LDHA) (GenBank sequences: Mouse/BC005509, Rat/NM_017025, Human/X02152), and (D) calmodulin 2 (CALM2) (GenBank sequences: Mouse/NM_001743, Rat/NM_017326, Human/NM_001743). ClustalX was also used to create the unrooted dendrograms using the bootstrapped neighbor-joining method. The dendrogram was viewed using the TreeView program.

alignment. The percent identities and dendrogram show that human and CHO sequences are the most similar to one another among all four species (96.2%), while, in most cases (including those in Fig. 3a–c), mouse and rat sequences are most similar to one another. The percent identity data also shows that for this sequence segment, the mouse:CHO alignment (93.8%) has a higher percent identity than the mouse:rat alignment (90.6%). Twenty-three CHO sequences were found to have higher identity to a human sequence over an orthologous mouse or rat sequence (similar to the CALM2 alignment). The other CHO ESTs where a human sequence provided the best alignment are cases where the known orthologs in mouse or rat have not yet been isolated. The alignments illustrate that the relative homology of CHO sequences with sequences from other organisms is not consistent for the entire genome, and using only mouse as a model for comparison may not provide the best similarity information for all DNA segments. Additionally, these data show that relative sequence identity levels vary for different genes. For example, the tables in Figure 3 show that ribosomal proteins and CALM2 are more highly conserved sequences (identity between all chosen species is greater than 90%). COX6C and LDHA sequences are both less conserved than the ribosomal proteins and CALM2, with similarities averaging between 80% and 90%. The variability between species and among different genes must always be considered when examining CHO EST data.

Mitochondrial Genome Sequencing and Assembly

The relative positioning of genes in the mitochondrial genome is known to be highly conserved among mammals (Wolstenholme, 1992). Using Phred/Phrap/Consed, the CHO EST mtRNA segments were aligned to the mouse mitochondrial genome to obtain the relative positioning of the ESTs. The alignment showed that approximately 70% of the entire CHO mitochondrial genome sequence was isolated as ESTs, with only nine gaps. Primers were designed using CHO sequences near the gaps of the CHO/mouse alignment for PCR walking. Mitochondria were isolated from both DXB-11 cells and the DXB-11 derived recombinant cells, and the primers were used to PCR amplify sequence gaps from both clones. The sequences of the PCR products were identical between the two cell lines and the sizes were consistent with the expected lengths according to the alignment with the mouse mitochondrial genome. The PCR amplified sequences were added to the mitochondrial EST sequences and assembled using Phred/Phrap/Consed. The CHO mitochondrial genome is 16,285 bp long, with 36.7% GC content. The relative gene arrangement is identical to that found in other mammals.

The complete CHO mitochondrial genome sequence was aligned with other available mammalian mitochondrial genomes using ClustalX, and a dendrogram, based on the alignment, was generated (Fig. 4). This alignment provides further insight into the evolutionary relation of the Chinese



Mitochondrial Genome							
	CHO	Vole	Cow	Pig	Human	Rat	Mouse
Mouse	78	78	73	75	70	82	100
Rat	77	77	73	75	71	100	
Human	69	70	72	73	100		
Pig	74	74	80	100			
Cow	72	73	100				
Vole	78	100					
CHO	100						

Figure 4. CHO mitochondrial genome sequence comparison to other mammalian mitochondrial genomes. ClustalX (version 1.81) was used to align the complete mitochondrial genome sequences for the following species: Mouse/NC_001569, Rat/NC_001665, Human/NC_001807, Pig/NC_000845, Cow/NC_001567, Taiwan Vole/NC_003041. The alignment output files provide the percent identity matrix for similarity between each species shown in the above tables. ClustalX was also used to create the unrooted dendrogram using the bootstrapped neighbor-joining method. The dendrogram was viewed using the TreeView program.

hamster to other species. The percent identities in the table are an average value for the entire genome alignment; however, more detailed examination of the alignment shows that certain regions (some corresponding to proteins) have higher or lower similarity than these average values for segments longer than 500 bp (data not shown). The whole mitochondrial genome alignment is less conserved than the nuclear transcript alignments shown in Figure 3. The CHO mitochondrial genome aligns with the highest identity (78%) to both the mouse and the Taiwan vole genomes, but the dendrogram shows that the Taiwan vole is likely the closest evolutionary neighbor among the species compared. Similarity of the CHO sequence to the rat sequence is almost as high as its similarity to mouse and Taiwan vole sequences (77%). Generally, the mitochondrial genome dendrogram is similar to the multiple sequence alignments in Figure 3a–c, as the rat and mouse mitochondrial genome sequences are more similar to one another than any other two species are to one another.

cDNA Microarray

All cDNA species isolated were spotted onto a microarray, including all redundant clones, to test library representation. RNAs from the two CHO cell lines used to construct library A (DXB-11 and its recombinant derivative) were compared by hybridization to five replicate arrays. Shown in Figure 5A is a linear plot of the normalized fluorescence intensity data for each probe on the microarray. The *x*-axis represents the intensities for the parental cell line (DXB-11), and the *y*-axis represents intensities for the recombinant cells. The vast majority of the data points lie near the diagonal, as expected with properly normalized data.

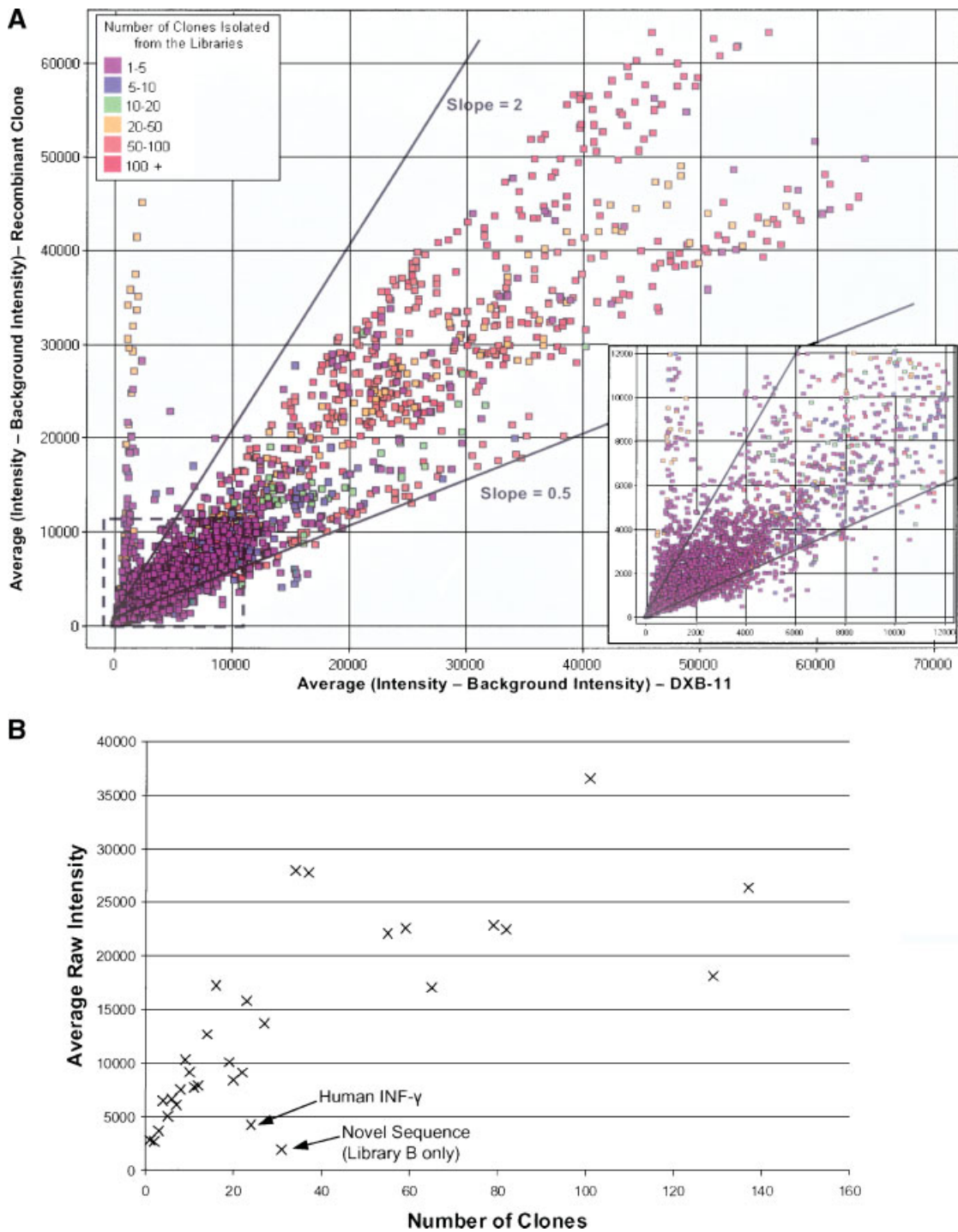


Figure 5. Normalized microarray intensities for CHO DXB-11 and the DXB-11 derived recombinant clone. **A:** Average intensities for the two CHO clones, CHO DXB-11 and the recombinant producing CHO cell line are plotted linearly using Spotfire. A color along a gradient is assigned to each spot according to the number of times the sequence is represented on the microarray (this data was obtained from the Phred/Phrap/Consed assembly of both libraries). Red spots are represented the most on the microarray and purple spots are present less than five times on the microarray. The plot shows that many high-intensity spots are red and represent clones that were isolated redundantly from the libraries. The inlaid plot is a closer view of the boxed region where intensity is less than 12,000 intensity units. **B:** Average intensity is plotted against transcript isolation abundance to show the level of agreement between random transcript isolation data and microarray intensity data. The y-axis shows the average intensity for transcripts isolated “x” times, where the x-axis is the number of times a particular transcript was isolated from the libraries. The trend shows that intensity is in agreement with the redundancy of isolation of a particular sequence.

To better represent the redundancy information, the probes on the microarray were grouped according to their representation in the library. In Figure 5A, the markers for the most frequently isolated cDNA species (isolated more than 100 times) are colored red. Purple markers represent cDNA species isolated less than five times from the libraries. Other colors are assigned to transcripts with isolation frequencies between these two limits, as shown in the legend. The transcript species that are more abundant in the transcriptome are expected to be isolated from the library more frequently. Although many factors affect the hybridization fluorescence intensity on microarrays, a semi-quantitative relationship between the intensity and the transcript level does exist. The more abundant species, which are also represented several times as probes on the microarray, have higher fluorescence intensities. Such a general trend is seen in Figure 5A: the purple markers have low fluorescence intensities, and the intensity increases progressively as the degree of abundance of those species increases.

To describe this general trend more quantitatively, a box (dotted-line) was drawn in Figure 5A to divide the intensity data into two regions, representing different ranges of intensity (greater or less than approximately 12,000 intensity units). The inlaid plot in Figure 5A is an enlargement of the boxed region. Approximately 80% of the spots on the microarray had fluorescence intensities within the boxed region. The identities of the cDNAs in the high fluorescence intensity region (outside of the box) are in agreement with redundantly isolated ESTs. Among the markers lying outside of the boxed area (20% of the total spots), 96% represent clones that were isolated more than once from the cDNA libraries. Eighty-two percent of the spots outside the boxed region are among the 25 most redundantly isolated sequences listed in Table II.

Figure 5B more clearly shows the level of agreement between random transcript isolation data and microarray intensity data by taking averages of fluorescence intensity for ESTs that were assembled into the same contigs and plotting intensity data against transcript isolation frequency. The y-axis shows the average fluorescence intensity and the x-axis is the number of times a particular transcript was isolated from the libraries. The trend clearly shows that average overall intensity increases in agreement with the frequency of isolation of a sequence. Two data points, as indicated by arrows, appear to contradict this trend. These two points correspond to sequences isolated only in library B: human IFN- γ and a novel sequence.

The two black lines drawn in Figure 5A represent slopes of 2.0 and 0.5, respectively. Spots outside the region bounded by these two lines are species whose intensities differ by more than twofold between the two RNA samples. These sequences are likely to be differentially expressed between the two CHO clones. A few of those sequences are rather prominent by deviating significantly from the diagonal (slope = 1). Those lying close to the y-axis, in the high-intensity region, are mRNA species for either dhfr or components of the cloning vector used to introduce the recombinant protein into

the parental cell line. Since these mRNA species were artificially introduced into the recombinant producing cell line, they should not be present in the DXB-11 parental clone.

For many genes on the microarray, there are multiple probes with sequences assembling into the same contig. When possible, the intensity ratio of these multiple sequences should be considered for a better determination of whether a gene is differentially expressed. By using data from multiple probes, smaller fold changes can be confidently discerned. This is shown by plotting intensity data in combination with EST assembly information and assigning different colors and shapes to gene products represented by more than one spot on the microarray. Figure 6 illustrates this by examining transcripts from the mitochondrial genome (mtRNAs), the most abundant mRNA species represented in the libraries. The x-axis is the \log_2 of the ratio of normalized average median intensities for the recombinant cells to the normalized average median intensities for DXB-11 cells. The y-axis represents the normalized average median fluorescence intensity for the recombinant clone. Each protein-coding cDNA from the mitochondrial genome is represented by a different color and shaped marker. The solid markers correspond to points for which the Student's *t*-test/*P*-value confidence for the \log_2 of the intensity ratio is less than 0.01. The empty markers are spots with *P* values greater than 0.01, and are regarded as statistically insignificant.

Plotted in this manner, one can view clusters of points in the data, corresponding to each mtDNA transcript. Those spots of the same transcript scatter over an area, and this area becomes smaller when *P* values higher than 0.01 are excluded. With the multiple spots on the array for these sequences, even a low level of differential expression (the \log_2 intensity ratios are less than 2) can be discerned with a high level of confidence. A few genes are more highly differentially expressed, for example, the D-loop transcript, and the 12S and 16S rRNA segments are significantly more present in the parental clone. Figure 6 also shows agreement between mtRNA intensity and Table II. The two most abundantly isolated transcripts are the ND4/ND4L and ATP6/ATP8 transcripts. These two species correspondingly fluoresce with the highest intensities in Figure 6.

DISCUSSION

cDNA Library Design

In this study, we report the isolation and sequencing of 4,608 ESTs from CHO cells. The information generated greatly expanded that available for Chinese hamster in the GenBank database. Since cDNA libraries contain only the actively transcribed regions of DNA at a single point in time, the cell lines and growth conditions for mRNA isolation were carefully chosen to include genes important to many aspects of cell line development and bioprocessing. Pools of different CHO cell clones under various growth conditions provided the transcripts for the construction of the libraries. The CHO clones selected include both host cell lines for expression of

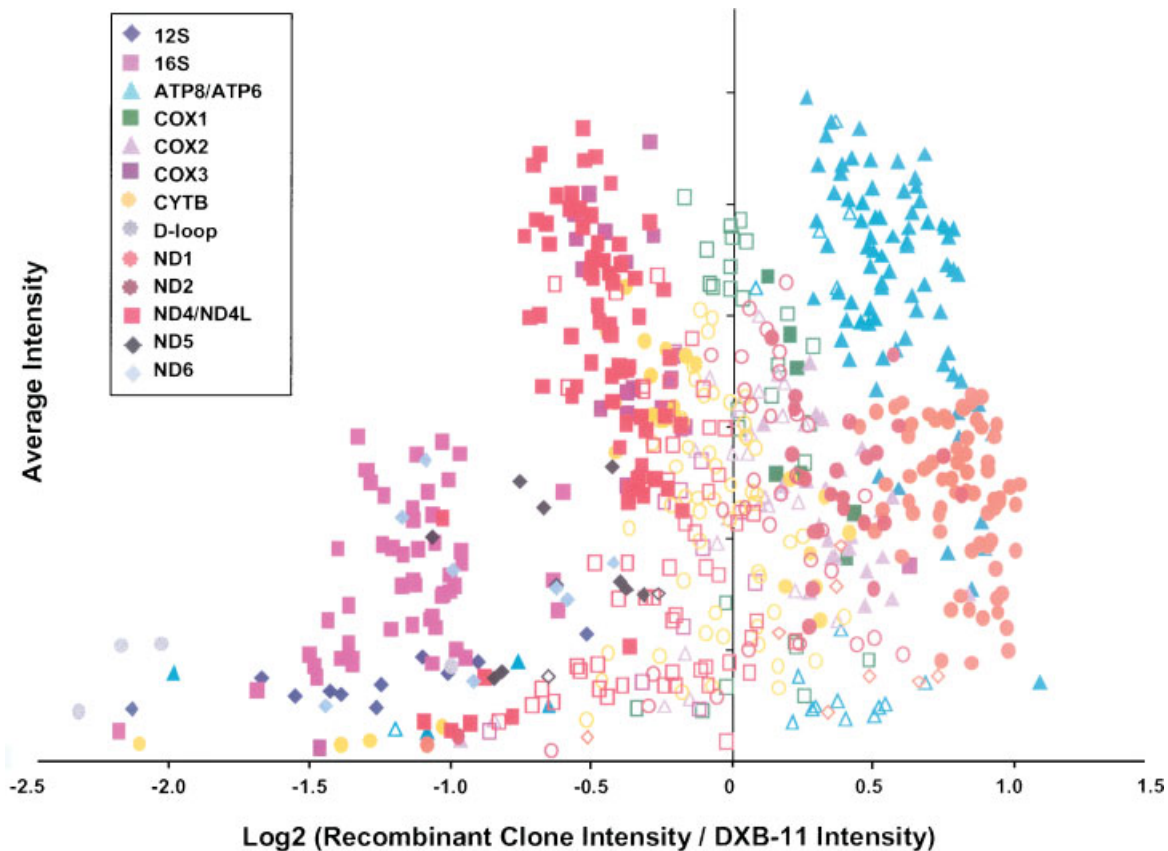


Figure 6. Fluorescence intensity and intensity ratio microarray data for mitochondrial DNA transcripts in DXB-11 and its derived recombinant producing clone. Fluorescence intensity and expression ratio data are plotted for the subset of highly abundant transcripts from the mitochondrial genome. Each transcript product is represented by a different color and/or shaped marker. Solid markers are those where the Student's *t*-test *P*-value for the log ratio is less than 0.01. Different transcript products cluster in this graph in a range of intensity and log-ratio values. The clustering of replicate spots gives confidence to less than twofold differences in mtRNA transcript abundance.

heterologous genes (CHO DXB-11), and recombinant CHO cell lines expressing different heterologous proteins introduced and amplified using the DHFR system. Additionally, the cell lines were subjected to a variety of culture conditions which encompass those typically used in cell line development and in processing, including the use of serum or serum-free medium, growth in suspension or adherently, and growth in early, exponential, and stationary phases.

Random sequencing from the two libraries allowed for the isolation of unique sequences at a reasonable frequency. Analysis of the assembled contigs shows that only a few transcripts, primarily from the mitochondrial genome, were highly abundant in both libraries. The redundancy levels for the two libraries are not prohibitive to further sequencing, and combining data from both libraries did not increase the overall isolated clone redundancy, rather it served to increase transcript diversity. Further sequencing from either library would likely provide new and relevant information; however, it is expected that eventually the prospect of isolating new genes from these libraries will face diminished returns.

In a typical EST project for human, animal, or plant, new transcript isolation is accomplished by generating new libraries from different tissues or from the same tissue in different disease or physiological states (Dempsey et al.,

2000; Rudd, 2003). The objective of this CHO EST project is to profile transcript abundance and identify genes important in CHO cells for recombinant protein production; thus, isolation and sequencing ESTs from Chinese hamster tissues may not expand the relevant gene pool. For future gene discovery, imposing strong environmental or genetic perturbations to CHO cells may elicit more transcript diversity to warrant the construction of new libraries for discovery of new and relevant sequences.

Annotation and Similarity to Other Species

BLAST comparison of the unique dataset with the non-redundant nucleotide database allowed for annotations, based on sequence similarity, to be assigned for almost 75% of the sequences. Less than 250 of the isolated sequences were previously deposited in the GenBank database for CHO cells; thus, our efforts provided approximately 2,300 new CHO transcripts, including nearly 600 novel sequences that do not show significant similarity to any other ESTs. While these novel sequences do not currently provide significant insight into intracellular regulation, they are an opportunity for discovery of new genes with potential importance. ESTs with significant similarity only to a single

chromosomal locus were also discovered and are a source of novel genes, or alternatively spliced isoforms. These sequences are not currently known to be expressed in other well-studied organisms, and may have an important, currently undiscovered role in CHO cells.

Among the CHO transcripts with significant sequence similarity to an EST in the GenBank database, most show the best alignment to an orthologous mouse sequence; however, a significant portion are most similar to a sequence isolated from rat or human tissues. Multiple sequence alignments for a few illustrative sequences (described in Fig. 3) confirm that for some transcripts, CHO cell sequences have more similarity to human and rat sequences than to the orthologous mouse sequence. The mitochondrial genome has been reported to be less conserved between species than the nuclear genome (Wolstenholme, 1992). Additionally, mitochondrial genome comparisons are considered good indicators of phylogenetic relationships, as accumulation of mutations is faster than that seen in transcribed regions of the nuclear genome. Mouse and vole mitochondrial sequences have the highest similarity to the CHO sequence, with an average alignment identity across the entire genome of 78% (this number is even lower in some regions). The mitochondrial sequence comparison results are consistent with the alignment data shown in Figure 3. With continued contributions of nucleotide data to the GenBank database and other databases, future comparisons will give a clearer picture of the actual phylogenetic relationship between CHO cells and other organisms.

Even though some CHO sequences are more similar to genes isolated from rat and human, our results suggest that mouse is currently the best model organism with a complete set of genetic information available. Establishing such a model for comparative analysis of CHO EST data is important because it is not likely the entire CHO genome sequence will be available in the near future. Structurally, these two species are very different in terms of the number of chromosomes typically found in each species (mouse $2n = 40$, Chinese hamster $2n = 22$). The chromosomal structural differences, however, are due to large chromosome segment rearrangements, leaving the relative positioning of most genes with respect to one another intact. Reciprocal chromosome painting of mouse chromosomes onto Chinese hamster chromosomes yielded 47 regions of homology (Yang et al., 2000). Using the mouse genome as a backbone for positioning CHO cell ESTs, in combination with chromosome mapping data, a genome context can be inferred for the CHO EST dataset. Providing some chromosomal context is important, as gene regulation is dictated by upstream and downstream regulatory elements that are not captured in EST sequencing projects.

CHO Mitochondrial Genome

A large number of mtRNA clones were isolated from both CHO cDNA libraries. Several studies report an increase in transcript abundance for mtRNAs in immortalized cell lines (Duncan et al., 2000; Kim et al., 2001). In addition to 12 of the

13 (excluding ND3) protein products and 2 ribosomal RNAs (12S and 16S) encoded in the mitochondrial genome, a portion of the D-loop was isolated as a transcribed species from both libraries. D-loop transcripts were also observed in a comparison of pre- and post-transformed human breast fibroblast cells (Duncan et al., 2000). Isolation of this regulatory element as a transcribed species may be indicative of the CHO cell's transformed phenotype.

Mitochondria are integral to cell survival, playing key roles in energy generation and regulation of apoptosis (Fernandez-Silva et al., 2003). Differences in the regulation of genes encoded in the mitochondrial genome may provide insight into the physiology of CHO cell lines, as changes in mitochondrial genome transcription have been observed in studies of mitochondrial cytopathies and changes in the metabolic properties of oxidative phosphorylation (Kunz, 2003). It was also reported that CHO cells grown for 2 weeks in the absence of serum have a measurable increase in the number of mitochondria per cell (Meents et al., 2002). A better understanding of the regulation of mitochondria number, energy production, and the relationship between mitochondria and apoptosis may provide opportunities for engineering more robust cell lines. Including redundant mitochondrial genome sequences on CHO microarrays will provide the opportunity to monitor differences in mtDNA transcription.

CHO cDNA Microarray

The major product of our EST sequencing effort is the development of the homology information needed for furthering DNA based research in CHO cells. Additionally, a DNA microarray for large-scale analysis of transcription profiles has been constructed and tested. Sequence information from the CHO EST libraries provided the required information regarding the level of similarity between the CHO genome and the genomes of other heavily sequenced mammalian species (e.g., mouse, rat, human). The multiple sequence alignments show that the level of homology between orthologous sequences vary for each different gene (Fig. 3). These few examples show a range of 83%–96% sequence identity conservation between CHO and mouse sequences. Prior to this sequencing effort, one would have to resort to using mouse or rat DNA microarrays for cross-species hybridization to CHO cDNAs for transcript profiling experiments. At this level of homology and with the variation between different genes observed in this study, ambiguity inevitably would arise in the interpretation of cross-species results. This is shown when the transcription of CHO cells and mouse cells subjected to identical sodium butyrate treatment was analyzed using both mouse and CHO cDNA microarrays (results to be published; manuscript in preparation). The availability of CHO EST sequence and DNA microarrays will greatly facilitate transcriptome analysis of CHO cells cultured under conditions important in bioprocessing.

The cDNA microarray was used to compare transcript levels between the CHO clones used to construct library A.

The results nicely illustrate the clone representativeness of the libraries and the utility of this tool for discerning differences in transcript regulation. Looking at the identities of some of the differentially expressed genes between the recombinant and parental clone further support the utility and representation of sequences included in the microarray. Genes involved in the introduction and amplification of a recombinant protein (i.e., *dhfr*) are significantly more abundant in the recombinant clone.

The fluorescence intensity for a spot on the array is not an exact quantitative measure of transcript abundance. The amount, orientation, conformation, and accessibility of the probe DNA immobilized on the surface may differ among spots, contributing to differences in fluorescence intensity, even when the same concentration of cDNA is present in the sample. Nevertheless, the intensity data can be used for order of magnitude estimates of the transcript abundance. The hybridization results revealed the strong correlation between transcript isolation frequency and fluorescence intensity.

The cDNA microarray constructed for this study included all sequenced clones. Redundantly isolated sequences are present as multiple probes for a single mRNA transcript. Some of the probes have significant overlapping sequences, while some are from non-overlapping segments of the same contig (gene). The presence of multiple probes for more abundant transcripts provided additional advantages for identifying differentially expressed genes. In comparing an abundant transcript species between two samples using a microarray assay, a small difference in expression may reflect a large change in the quantity of transcripts (in terms of total number of transcripts). The quantity of the transcript changed between two samples is often more than that observed for a transcript that is expressed at a lower level, but has a larger difference in expression. It is probable that in some cases, such smaller changes in the level of an abundant transcript will give rise to profound changes in cellular physiology. Small changes are difficult to discern with statistical confidence when only one or two probes of a given gene are present on a microarray. As a result, such small differences in expression are often disregarded in microarray analysis. With multiple probes on the microarray, smaller differences can be discerned as illustrated in Figure 6.

We anticipate that a primary use of the CHO cDNA microarray will be in bioprocess research, especially for the investigation of the effects of culture conditions and cell line development on gene expression. Unlike the gene expression patterns in microbial organisms or in the developmental processes of higher organisms, physiological perturbations of established cell lines in culture often result in only relatively small changes in the transcriptome (Seth et al., 2005). An ability to see smaller transcript changes for more abundant species will enhance the utility of this versatile research tool.

CONCLUDING REMARKS

Isolation and sequencing of ESTs from CHO cell cDNA libraries has provided novel information and highly reliable

tools for genome-based research in CHO cells. As whole genome sequencing of the Chinese hamster is unlikely to be pursued in the near future, EST sequencing will likely be the major tool for genomic exploration in this industrially important cell line. Based on current redundancy levels, more EST sequencing will contribute significantly to the quantity of CHO sequence data. Additionally, further gene discovery in other organisms and functional annotation based on protein motif prediction will reveal potential roles of unannotated (novel) sequences. Modern biotechnological exploration is largely based on DNA sequence information. Our understanding of the genetic traits that affect the growth and production characteristics of CHO cells, and our ability to exploit these understandings to engineer superior producing cell lines will greatly benefit from the availability of their sequence information. This work represents an important first step towards genomic exploration of CHO cells.

This work was supported in part by funding from the agency for Science, Technology, and Research (A*STAR), Singapore. The bioinformatics support was provided by the University of Minnesota Supercomputing Institute. The technical assistance of Dr. Zheng Jin Tu and Wen Dong are gratefully acknowledged. We thank Dr. Craig Beattie for stimulating discussions. K.F.W. was supported by an NSF fellowship. M.L.G. and K.F.W. were supported by a NIH Biotechnology Training Grant (GM08347).

References

- Adams M, Kelley J, Gocayne J, Dubnick M, Polyneropoulos M, Xiao H, Merril C, Wu A, Olde B, Moreno R. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252:1651–1656.
- Andersen DC, Krummen L. 2002. Recombinant protein expression for therapeutic applications. *Curr Opin Biotechnol* 13:117–123.
- Chasin LA. 1973. The effect of ploidy on chemical mutagenesis in cultured Chinese hamster cells. *J Cell Physiol* 82:299–308.
- Chasin LA, Urlaub G. 1975. Chromosome-wide event accompanies the expression of recessive mutations in tetraploid cells. *Science* 187:1091–1093.
- Chu L, Robinson DK. 2001. Industrial choices for protein production by large-scale cell culture. *Curr Opin Biotechnol* 12(2):180–187.
- Dempsey AA, Ton C, Liew CC. 2000. A cardiovascular EST repertoire: Progress and promise for understanding cardiovascular disease. *Mol Med Today* 6(6):231–237.
- Diehl F, Grahlmann S, Beier M, Hoheisel JD. 2001. Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Res* 29(7):E38.
- Duncan EL, Perrem K, Reddel RR. 2000. Identification of a novel human mitochondrial D-loop RNA species which exhibits upregulated expression following cellular immortalization. *Biochem Biophys Res Commun* 276:439–446.
- Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred II. Error probabilities. *Genome Res* 8(3):186–194.
- Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175–185.
- Fernandez-Silva P, Enriquez JA, Montoya J. 2003. Replication and transcription of mammalian mitochondrial DNA. *Exp Physiol* 88(1): 41–56.
- Geisse S, Gram H, Kleuser B, Kocher HP. 1996. Eukaryotic expression systems: A comparison. *Protein Expr Purif* 8(3):271–282.

- Gey GO, Coffman WD, Kubicek MT. 1952. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *J Exp Med* 12:264–265.
- Gordon D, Abajian C, Green P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res* 8(3):195–202.
- Harris M. 1971. Mutation rates in cells at different ploidy levels. *J Cell Physiol* 78:117–184.
- Holliday R, Ho T. 1998. Evidence for gene silencing by endogenous DNA methylation. *Proc Natl Acad Sci USA* 95(15):8727–8732.
- Holliday R, Ho T, Paulin R. 1996. Gene silencing in mammalian cells. In: Russo VEA, Martienssen R, Riggs AD, editors. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. p 5–27.
- Iseli C, Jongeneel CV, Bucher P. 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol* 1999:138–148.
- Kao FT, Puck T. 1968. Genetics of somatic mammalian cells, VII. Induction and isolation of nutritional mutants in Chinese hamster cells. *Proc Natl Acad Sci USA* 60(4):1275–1281.
- Kaufman RJ, Murtha P, Davies MV. 1987. Translational efficiency of polycistronic mRNAs and their utilization to express heterologous genes in mammalian cells. *EMBO J* 6:187–193.
- Kaufman RJ, Sharp PA. 1982. Amplification and expression of sequences cotransfected with a modular dihydrofolate reductase complementary DNA gene. *J Mol Biol* 159:601–621.
- Kaufman RJ, Wasley LC, Dorner AJ. 1988. Synthesis processing and secretion of recombinant human factor VIII expressed in mammalian cells. *J Biol Chem* 263(13):6352–6362.
- Kaufman RJ, Wasley LC, Spiliotes AJ, Gossels SD, Latt SA, Larsen GR, Kay RM. 1985. Coamplification and coexpression of human tissue-type plasminogen activator and murine dihydrofolate reductase sequences in Chinese hamster ovary cells. *Mol Cell Biol* 5:1750–1759.
- Kim H, You S, Kim I-J, Farris J, Foster LK, Foster DN. 2001. Increased mitochondrial-encoded gene transcription in immortal DF-1 cells. *Exp Cell Res* 265:339–347.
- Korke R, Rink A, Seow TK, Chung M, Beattie CW, Hu W-S. 2002. Genomic and proteomic perspectives in cell culture engineering. *J Biotechnol* 94:73–92.
- Korn JH, Mory Y, Ziberstein A, Holtmann H, Revel M, Wallach D. 1988. Cloning of genomic DNA for tumor necrosis factor and efficient expression in CHO cells. *Lymphokine Res* 7(4):349–358.
- Kunz WS. 2003. Different metabolic properties of mitochondrial oxidative phosphorylation in different cell types—Important implications for mitochondrial cytopathies. *Exp Physiol* 88(1):149–154.
- Levy-Nissenbaum O, Sagi-Assif O, Raanani P, Avigdor A, Ben-Bassat I, Witz IP. 2003. cDNA microarray analysis reveals an overexpression of the dual-specificity MAPK phosphatase PYST2 in acute leukemia. *Protein Phosphatases* 366:103–113.
- Lin FK, Suggs S, Lin CH, Browne JK, Smalling R, Egrie JC, Chen KK, Fox GM, Martin F, Stabinsky Z. 1985. Cloning and expression of the human erythropoietin gene. *Proc Natl Acad Sci USA* 82(22):7580–7584.
- Marcotte ER, Srivastava LK, Quirion R. 2003. cDNA microarray and proteomic approaches in the study of brain diseases: Focus on schizophrenia and Alzheimer's disease. *Pharmacol Therap* 100(1):63–74.
- Meents H, Enenkel B, Eppenberger HM, Werner RG, Fussenegger M. 2002. Impact of coexpression and coamplification of sICAM and antiapoptosis determinants bcl-2/bcl-xL on productivity, cell survival, and mitochondria number in CHO-DG44 grown in suspension and serum-free media. *Biotechnol Bioeng* 80(6):706–716.
- Page RDM. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 12:357–358.
- Paulin RP, Ho T, Balzer HJ, Holliday R. 1998. Gene silencing by DNA methylation and dual inheritance in Chinese hamster ovary cells. *Genetics* 149(2):1081–1088.
- Puck TT, Ciecura SJ, Robinson A. 1958. Genetics of somatic mammalian cells III. Long-term cultivation of euploid cells from human and animal subjects. *J Exp Med* 108:945–956.
- Rudd S. 2003. Expressed sequence tags: Alternative or complement to whole genome sequences? *Trends Plant Sci* 8(7):321–329.
- Russo G, Zegar C, Giordano A. 2003. Advantages and limitations of microarray technology in human cancer. *Oncogene* 22(42):6497–6507.
- Scahill SJ, Devos R, Van Der Heyden J, Fiers W. 1983. Expression and characterization of the product of a human immune interferon complementary DNA gene in Chinese hamster ovary cells. *Proc Natl Acad Sci USA* 80(15):4654–4658.
- Seta KA, Millhorn DE. 2004. Functional genomics approach to hypoxia signaling. *J Appl Physiol* 96(2):765–773.
- Seth G, Philp RJ, Denoya CD, McGrath K, Stutzmann-Engwall KJ, Yap M, Hu W-S. 2005. Large-scale gene expression analysis of cholesterol dependence in NS0 cells. *Biotechnol Bioeng* 90:552–567.
- Siminovitch L. 1976. On the nature of heritable variation in cultured somatic cells. *Cell* 7:1–11.
- Theodorakis P, D'Sa-Eipper C, Subramanian T, Chinnadurai G. 1996. Unmasking of a proliferation-restraining activity of the anti-apoptosis protein EBV BHRF1. *Oncogene* 12(8):1707–1713.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25:4876–4882.
- Todaro GJ, Green H. 1963. Quantitative Studies of the Growth of Mouse Embryo Cells in culture and their development into established lines. *J Cell Biol* 17:299–313.
- Urlaub G, Chasin LA. 1980. Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proc Natl Acad Sci USA* 77(7):4216–4220.
- Warner TG. 1999. Enhancing therapeutic glycoprotein production in Chinese hamster ovary cells by metabolic engineering endogenous gene control with antisense DNA and gene targeting. *Glycobiology* 9(9):841–850.
- Wolstenholme DR. 1992. Animal mitochondrial DNA, structure and evolution. *Int Rev Cytol* 141:173–216.
- Wood WI, Capon DJ, Simonsen CC, Eaton DL, Gitschier J, Keyt B, Seeburg PH, Smith DH, Hollingshead P, Wion KL. 1984. Expression of active human factor VIII from recombinant DNA cloned. *Nature* 312: 337–342.
- Yang F, O'Brien PC, Ferguson-Smith MA. 2000. Comparative chromosome map of the laboratory mouse and Chinese hamster defined by reciprocal chromosome painting. *Chromosome Res* 8(3):219–227.
- Yang YH, Dudoit S, Luu P, Speed TP. 2001. Normalization for cDNA microarray data. San Jose, California: SPIE BiOS.