Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 10645–10649, October 1994
Applied Biological Sciences

A catalogue of genes in the cardiovascular system as identified by expressed sequence tags

C. C. LIEW*†, D. M. HWANG*, Y. W. FUNG*, C. LAURENSSEN*, E. CUKERMAN*, S. TSUI‡, AND C. Y. LEE‡

*Laboratory of Molecular Cardiology, Departments of Clinical Biochemistry and Medicine, The Center for Cardiovascular Research, The Toronto Hospital, University of Toronto, Canada; and ‡Department of Biochemistry, The Chinese University of Hong Kong, Hong Kong

Communicated by C. C. Tan, June 27, 1994

ABSTRACT **The heart, which is composed of all the cellular components of the circulatory system, is a representative organ for obtaining genes expressed in the cardiovascular system in normal and disease states. We used partial sequences of cDNA clones, or expressed sequence tags, to identify and tag genes expressed in this organ. More than 3500 partial sequences representing >3000 cDNA clones have been obtained from either the 5′ or 3′ end of inserts derived from human heart cDNA libraries. Of 3132 cDNA clones analyzed by sequence similarity searching against the GenBank/EMBL data bases, 1485 (47.4%) were found to represent additional, previously undiscovered genes, whereas 267 clones were matched to human brain expressed sequence tags. Clones matching to known genes were catalogued according to their putative structural and cellular functions. cDNA probes from reverse-transcribed mRNAs of fetal and adult hearts were used to study differential expression of selected clones in cardiac development. Cataloguing genes expressed in the heart may provide insight into the genes involved in health and cardiovascular disease.**

The heart is a complex organ consisting of many different cell types working in concert to propel blood through the circulatory system. While much progress has been made in understanding the macroscopic, physiological function of the heart, considerably less is known about the molecular basis of cardiac function. For example, genes expressed in the processes of ontogeny and growth remain largely unknown, whereas the genetic and molecular basis of a broad spectrum of cardiovascular diseases such as hypertension, atherosclerosis, coronary artery disease, and heart failure also remains to be determined.

Partial sequencing of clones from cDNA libraries of specific tissues or cell types to generate expressed sequence tags (ESTs) has proven to be a rapid and efficient means of discovering genes on a large scale and of providing both quantitative and qualitative information regarding gene expression in a variety of tissues and cells such as brain, liver, and lymphocyte (1–5). Information from single-pass sequencing of cDNA clones has also been used in many other applications, including the generation of physical maps of chromosomes (6, 7). For these reasons, we have implemented an efficient and cost-effective procedure to generate ESTs from human heart cDNA libraries (8, 9).

Using ESTs matching genes of known sequence, we have initiated the categorization of genes expressed in the heart during normal growth as well as in disease states. Here, we report the extensive sequencing of cDNA clones derived from human fetal and adult heart cDNA libraries and the systematic classification of the genes expressed in the cardiovascular system, as viewed through the heart.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Summary of sequences and clones represented

	Forward	Reverse	Clones represented
New sequences			
No match	1466 (47.6%)	406 (51.2%)	1485 (47.4%)
Match to human ESTs	266 (8.6%)	69 (8.7%)	267 (8.5%)
Known genes	1349 (43.8%)	318 (40.1%)	1380 (44.1%)
Grand total	3081	793	3132

Forward and reverse sequences (ESTs) were obtained using the forward and reverse sequencing primers, respectively. In some cases, one cDNA clone was sequenced from both forward and reverse directions; hence, the total number of clones represented (3132) is less than the total number of ESTs obtained (3874). Sequences matching to human ESTs in the GenBank/EMBL data bases were classified as new sequences; the percentages given were calculated based on the total of each column.

MATERIALS AND METHODS

Chemical reagents, *Taq* polymerase, and reverse transcriptase were purchased from Pharmacia. A directionally cloned human fetal heart cDNA library was constructed in λ gt22 expression vector (9). The human adult heart cDNA library was purchased from Clontech. Partial sequencing of cDNA clones proceeded as described (8, 9). Sequence comparisons against the GenBank/EMBL nucleotide and protein data bases were done by using the BLAST network server (10, 11) at the National Center for Biotechnology Information.

Human fetal and adult heart total RNAs for dot-blot analysis were obtained by the guanidinium thiocyanate/phenol extraction method (12). Poly(A)$^+$-enriched mRNAs were used to prepare radiolabeled double-stranded cDNA probes in the presence of [α-^{32}P]dATP, [α-^{32}P]dCTP, and oligo(dT) primers, using a modified Gubler and Hoffman protocol (13). The PCR products generated from EST-tagged clones representing additional and known transcripts were dot-blotted on Nytran (Schleicher & Schuell). After hybridization for 24 hr at 65°C (1 M NaCl/1% SDS/10% dextran sulfate), the membranes were washed in 0.1× standard saline/citrate (SSC) at 42°C for 30 min. Autoradiographs were obtained by exposing films at −70°C for 24 hr, after which differential expression patterns were revealed by comparison of dot intensities.

RESULTS

Analysis of cDNA Clones Sequenced. The 3874 partial cDNA sequences, or ESTs, representing 3132 cDNA clones from a human adult heart cDNA library, have been obtained by using either forward or reverse primers derived from the λ gt11 vector (Table 1). Approximately 47% of the ESTs (1485

Abbreviation: EST, expressed sequence tag.

†To whom reprint requests should be addressed at: Department of Clinical Biochemistry, Banting Institute, University of Toronto, Toronto, ON Canada M5G 1L5.

Table 2. Human cardiac ESTs matched to known genes in the GenBank/EMBL data bases

Contractile Elements
- α-Actin (4)
- α-Cardiac actin (13)
- Fetal skeletal muscle actin (mouse) (2)
- C protein, skeletal muscle (chicken)
- α-Cardiac myosin heavy chain
- β-Myosin heavy chain (29)
- Nonmuscle myosin heavy chain (2)
- Myosin heavy chain (rabbit)
- Myosin IB (cow)
- Myosin alkali light chain (6)
- Myosin alkali light chain (mouse)
- Myosin light chain (7)
- Myosin light chain 1V/Sb isoform
- Myosin regulatory light chain
- 20-kDa myosin light chain
- Nonmuscle myosin alkali light chain
- Smooth muscle myosin
- Ventricular myosin light chain 1
- Ventricular myosin light chain 2 (6)
- Skeletal muscle α-tropomyosin (14)
- Tropomyosin
- Cardiac troponin C
- Slow skeletal troponin C
- Cardiac troponin I (3)
- Slow twitch skeletal troponin I
- Cardiac troponin T (rat) (8)

Cytoskeletal
- Actin (nonexact human)
- α-Actin, vascular (rat)
- β-Actin (4)
- Actin-binding protein (2)
- Actin-related protein (nonexact) (2)
- Centrosome-associated actin homologue (dog)
- α-Actinin (2)
- Skeletal muscle α2 actinin (3)
- Assembly protein AP50 (rat) (2)
- Cofilin
- Cytokeratin
- Desmin (3)
- Dynein-associated polypeptide (rat)
- Filamin (chicken)
- Hemopoietic proteoglycan core protein (nonexact)
- Microfibril-associated glycoprotein (cow)
- Microtubule-associated protein
- Microtubule-assembly protein (rat)
- Mitotic kinesin-like protein
- Nestin
- Non-erythroid band-3-like protein
- Skelemin (mouse) (3)
- α-Spectrin (2)
- Talin (mouse)
- Tensin (chicken)
- Epithelial tropomyosin
- α-Tubulin (7)
- β-Tubulin (5)
- Vimentin (3)

Extracellular Matrix
- Biglycan
- Collagen α1 (I) (8)
- Collagen α2 (I) (4)
- Collagen pro-α1 (III) (4)
- Collagen α1 (IV)
- Collagen α2 (IV)
- Collagen pro-α1 (V) (nonexact)
- Collagen α1 (VI)
- Collagen α1 (XVI)
- Collagen α1 (XVIII)
- Colligin (3)
- Connectin (chicken) (2)
- Elastin (5)
- Extracellular matrix protein BM-40
- Fibronectin, cellular
- Fibronectin
- Laminin B2 chain
- S laminin
- Laminin-binding protein
- Osteonectin (2)
- Nidogen

Energy Metabolism
- Aconitase (pig) (2)
- ADP/ATP translocase
- Aldolase
- Aldolase A (3)
- Aldolase A, fibroblast (2)
- Aspartate aminotransferase
- Carnitine palmitoyltransferase I (rat) (2)
- Citrate synthase (pig)
- Creatine kinase M (5)
- Cytochrome *bc*-1 complex core protein
- Cytochrome *bc*-1 complex core protein II
- 2,4-Dienoyl-CoA reductase (rat)
- Dihydrolipoamide dehydrogenase
- α-Enolase (4)
- Glyceraldehyde-3-phosphate dehydrogenase (7)
- Glycogen synthase kinase 3α (rat)
- Glycogen phosphorylase, brain
- H^+-ATP synthase subunit b
- Inosine-5′-monophosphate dehydrogenase
- Ketoacid dehydrogenase kinase (rat)
- Lactate dehydrogenase A (3)
- Lactate dehydrogenase B (nonexact)
- Lipoprotein lipase
- Malate dehydrogenase (pig) (2)
- Mitochondrial ATP synthase
- Mitochondrial malate dehydrogenase (mouse)
- NADH-cytochrome *b5* reductase (2)
- NADH-ubiquinone oxidoreductase
- Neuroleukin (glucose phosphate isomerase) (3)
- Phosphogluconate dehydrogenase (sheep) (3)
- Phosphoglycerate mutase
- Phosphofructokinase
- Pyruvate kinase M2-type (3)
- Transglutaminase (5)
- Triose-phosphate isomerase (3)
- Ubiquinone oxidoreductase (cow)

Hormones and Hormonal Regulation
- Atrial natriuretic factor (9)
- Bone morphogenetic protein 1
- Glutathione-insulin transhydrogenase (2)
- Inhibin β (A) subunit
- Insulin-like growth factor II
- Preproenkephalin
- Prothymosin-α
- Retinoic acid receptor γ 1
- Retinoic acid-binding protein
- Steroid hormone receptor
- Thyroid hormone-binding protein (2)

Signal Transduction and Cell Regulation
- Adenylyl cyclase (dog)
- Calcium-dependent protein kinase I (rat)
- cAMP-dependent protein kinase
- CAP protein
- Casein kinase I-delta (rat)
- cdc2/CDC28-like protein kinase (nonexact)
- CDC21 homolog (*Xenopus*)
- Epsilon 14-3-3 isoform (mouse)
- G_s α subunit
- G_s GTP-binding protein
- p190-GAP-associated protein (rat)
- GTP-binding protein (*Discopyge*)
- GTPase
- Guanylate cyclase
- Modifier 3 protein (mouse)
- NAD-ADP ribosyltransferase (nonexact)
- Nuclear protein p47 (rat)
- Nucleic acid-binding protein (mouse)
- p78 protein
- PLA-X (nonexact)
- 80-kDa protein kinase C substrate (2)
- Protein phosphatase 2A catalytic subunit β
- Protein-tyrosine phosphatase HPTPβ
- RAB13 GTP-binding protein (rat)
- Rab GDP-disassociation inhibitor (rat)
- A-*raf*-1 oncogene
- RecA-like protein
- Rho-GAP protein
- Serine-threonine protein kinase
- Serine-threonine protein kinase
- Serine-threonine protein kinase
- Serine-threonine protein kinase
- Siah-1B protein (mouse)
- Stathmin
- c-*syn* protooncogene
- *tre* oncogene (nonexact)
- Transducin-like enhancer protein
- TSE1 protein kinase A regulatory subunit

Transcription and Translation
- (A+U)-rich element RNA-binding protein AUF-1
- CAAT-box binding transcription factor
- Chaperonin
- Chaperonin-like protein
- DNA-binding protein (nonexact)
- DNA-binding protein A
- Elongation factor 1α (6)
- Elongation factor 1γ (2)
- Elongation factor 1-Δ (nonexact)
- Elongation factor 2 (3)
- Ro ribonucleoprotein autoantigen
- HnRNP type A/B protein
- H19 RNA (6)
- 71-kDa heat shock cognate protein (3)
- Heat shock protein (neurospora)
- Heat shock protein Hsp70 (nonexact)
- Heat shock protein Hsp89-α
- 90-kDa heat shock protein (5)
- Helix–loop–helix protein (Id-2)
- HnRNP core protein A1 (3)
- Novel hnRNP protein
- Initiation factor 4AI
- Initiation factor 4B (3)
- Late upstream transcription factor
- Liver expressed protein
- α-Palindromic binding protein
- Poly(A)-binding protein
- Ribophorin I
- Acidic ribosomal phosphoprotein
- Acidic ribosomal phosphoprotein PO (nonexact)
- Large ribosomal subunit protein (mouse)
- Ribosomal protein L3 (3)
- Ribosomal protein L4 (3)
- Ribosomal protein L5
- Ribosomal protein L6
- Ribosomal protein L7
- Ribosomal protein L8 (2)
- Ribosomal protein L13 homologue
- Ribosomal protein L18 (2)
- Ribosomal protein L19
- Ribosomal protein L23 (2)
- Ribosomal protein L23A (rat)
- Ribosomal protein L29 (rat)
- Ribosomal protein L37a (2)
- Ribosomal protein S3
- Ribosomal protein S3a
- Ribosomal protein S4 (7)
- Ribosomal protein S6
- Ribosomal protein S8
- Ribosomal protein S9 (rat)
- Ribosomal protein S19 (2)
- Ribosomal protein S20
- SnRNP protein B
- TAF gene
- Transcription factor ISGF-3
- Zinc finger protein
- Zinc finger protein
- Zinc finger protein 42
- Zinc finger protein Kox5

Membrane-Associated
- Amyloid protein
- Amyloid β/A4
- Anion exchange protein 3 (nonexact)
- Ca^{2+}-ATPase (2)
- Cardiac Ca^{2+}-release channel (ryanodine receptor)
- *CD34* gene (2)
- Chloride channel protein (cow)
- ClC-K1 chloride channel (rat)
- Cysteine-rich FGF receptor (chicken)
- Fibronectin receptor α subunit
- Fibronectin receptor β subunit
- Formyl-peptide receptor
- Heparin-binding growth factor receptor
- HLA-associated transcript 3 (bat3) (5)
- HLA-DR associated protein I
- Insulin-like growth factor-binding protein 5
- Integrin β5 subunit (nonexact)
- Integrin α6
- Interleukin 5 receptor
- Junctional sarcoplasmic reticulum glycoprotein (rabbit)
- Laminin receptor (nonexact)
- Laminin receptor homolog
- Lectin (14 kDa)
- Lysosomal membrane glycoprotein CD63 (2)
- Minimal change nephritis glycoprotein (rat)
- MUC 18 glycoprotein
- Myasthenic syndrome antigen B
- Na^+/Ca^{2+} exchanger
- Na^+/K^+ ATPase (4)
- P-glycoprotein (*Drosophila*)
- Ror1 transmembrane receptor tyrosine kinase
- Signal sequence receptor β subunit (dog)

Table 2. (*Continued*)

Tetrodotoxin-insensitive Na^+ channel
Tie-2 (cow)
Voltage-dependent anion channel
Voltage-dependent anion channel isoform 1

Miscellaneous

49-kDa protein (nonexact human)
78-kDa glucose-regulated protein
85-kDa protein
A5 protein (*Xenopus*)
α1 acid glycoprotein (mouse)
Actin polymerization inhibitor protein (chicken)
Apolipoprotein J
B1 protein (*Xenopus*)
D16S4447 (BBC1)
Binding protein (2)
Ca^{2+}-dependent protease (3)
Calpastatin
Calphobindin II
Calsequestrin (rabbit)
can gene (2)
Carboxypeptidase N small subunit (nonexact)
Carcinoembryonic antigen (rat)
β-Catenin
Cellular resistance protein
Choline kinase
Complement factor B
Complement factor C1r
Cyclophilin (4)
Cyclophilin-like protein
Cysteine-rich peptide
Cysteine-rich intestinal protein (mouse)
Cytosolic epoxide hydrolase
Δ-Aminolevulinate synthase
Diff6 protein homolog (*Drosophila*)
DNA repair helicase (nonexact)
DNA topoisomerase II
Drosophila female sterile homeotic homologue
Epididymal apical protein 1 (rat)
Farnesyl protein transferase
Globin
β-Globin precursor (nonexact)
H5 brain protein (mouse)
hAES-2 gene
α-Hemoglobin
High-mobility group box SSRP1
Histamine *N*-methyltransferase (rat)
Histidine-rich Ca^{2+}-binding protein
Histone H1
HIV-1 TAR RNA-binding protein
Human open reading frame
Immunoglobulin S(u)-like sequence (nonexact)
Insulinoma rig-analog DNA-binding protein
Liver-expressed protein
Lysosomal serine carboxypeptidase
MAC30
M-Phase phosphoprotein 2
α2 macroglobulin (nonexact)
Myoglobin
N-Acetylglucosaminyltransferase I
Neuronal protein (rat)
Neutrophil oxidase (nonexact)
Nuclear mitotic apparatus gene (nonexact)
Olfactory neuron-specific clone (nonexact)
P311 gene (mouse) (2)
P5 protein (mouse)
PBX2 gene (nonexact)
Placental ribonuclease inhibitor (nonexact)
Poly(ADP-ribose) polymerase (2)
Polyubiquitin (pea)
Porphobilinogen synthase (nonexact)
Pre-B-cell enhancing factor
Pregnancy-specific β1 glycoprotein (nonexact)
Prosaposin (nonexact)
Protein upstream of N-Ras
Protocadherin 43
Putative homeotic protein
7S L gene
Rapamycin-binding protein
RGH2 gene
RGH2 gene (nonexact)
Sarcolumenin (rabbit)
Skeletal muscle 165-kDa protein
Small cellular 7SK RNA
Sorcin CP 22
Syntrophin 1 (mouse)
Thioltransferase (glutaredoxin) (pig)
Ubiquitin (5)
Ubiquitin-activating enzyme E1
Uracil-DNA glycosylase
UV-damaged DNA-binding protein (*Cercopithecus aethiops*)
Valosin-containing protein (pig)

Presented are genes to which cardiac ESTs matched. Numbers in parentheses indicate the frequency clones with ESTs matched to these genes. In cases where ESTs matched to nonhuman sequences, the organism from which the matching sequence was derived is also indicated in parentheses. Also indicated are ESTs matched incompletely with known human sequences (nonexact). PLA-X, GeneBank accession no. X06705; HPTPβ, human protein-tyrosine phosphatase β; Rab GDI, Rab GTP-dissociation inhibitor; HnRNP, heterogeneous nuclear ribonucleoprotein; SnRNP, small nuclear ribonucleoprotein; TAF, trans-activating factor; ClC-K1, chloride channel-kidney; FGF, fibroblast growth factor; hAES-2, human protein exhibiting similarity to N terminal of *Drosophila* enhancer of split groucho protein; HIV-1, human immunodeficiency virus type 1; MAC30, meningioma-associated clone; RGH2 gene, human endogenous retrovirus-like element (clone RGH2); Tie-2, tyrosine kinase with immunoglobulin and epidermal growth factor homology domains.

clones) demonstrated no matches with entries in the GenBank/EMBL data bases using the BLAST network server at the National Center for Biotechnology Information (10, 11) and were defined to be different, previously uncharacterized transcripts present in the cardiovascular system. Another 8.5% of these transcripts matched to other ESTs currently deposited in the public data bases, although not to any other known sequences. Of the 3132 cDNA clones sequenced, 1380 sequences, or ≈44% of the partial sequences, exhibited significant identity to known genes, among which 12% were similar to repeated sequences (e.g., *Alu*, LINE-1), whereas 12% represented mitochondrial transcripts (data not shown).

Accuracy of our single-pass sequencing technique was assessed by using ESTs matching to the human mitochondrial consensus sequence. Of 19,259 bp analyzed, the average accuracy over the first 300 bp of sequence was 97.8%. The average accuracy for portions of sequence compiled beyond 300 bp was slightly lower (96.2%; data not shown). This level of ambiguity did not significantly affect the identification of known and previously unknown transcripts by data base search.

Catalogue of Cardiovascular ESTs. Clones corresponding to a total of 342 specific known genes (excluding mitochondrial genes and repetitive elements) were classified according to distribution and function (see Table 2). Aside from mitochondrial transcripts and repetitive elements, which were also found at high levels in other EST projects, no single transcript represented >1% of all ESTs generated by this project. The most frequently occurring gene was β-myosin heavy chain, which was represented by 29 ESTs (0.87%), followed by α-tropomyosin, which was recorded 14 times (0.42%) (Table 2).

In some applications, redundant sequencing of cDNA clones representing a single transcript proved beneficial, by facilitating the assembly of distinctive, full-length human cDNA sequences. Fig. 1 shows that the complete cDNA sequence of the human cardiac troponin T sequence was determined by overlapping ESTs similar to the rat cardiac troponin T transcript. This result confirmed the full-length cDNA sequence of the human cardiac troponin T gene, which was recently published (14).

Among the 342 distinctive genes, 73 were genes sequenced in other organisms, but for which the complete

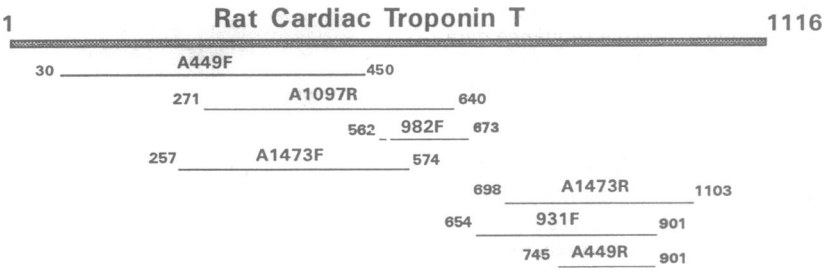

FIG. 1. Redundant sequencing of cDNA clones achieved a complete contiguous sequence of the human cardiac troponin T. Seven cDNA clones were partially sequenced using either λ gt11 forward (F) or reverse (R) primers by cycle sequencing.

Table 3. Distribution of human heart, brain, and hepatocyte ESTs with data base matches by functional categories

Category	Heart	Brain	Liver
Contractile	21.0 (109)	0.0 (0)	0.0 (0)
Cytoskeletal/structural	10.6 (55)	26.3 (227)	3.1 (6)
Extracellular matrix	8.1 (42)	0.0 (0)	0.0 (0)
Energy metabolism	13.5 (70)	6.1 (53)	9.2 (18)
Hormones/hormonal regulation	4.2 (22)	0.7 (6)	1.0 (2)
Signal transduction/cell regulation	7.9 (41)	19.9 (172)	8.2 (16)
Transcription/translation	18.7 (97)	18.8 (162)	31.8 (62)
Membrane-associated	8.8 (46)	15.3 (132)	2.1 (4)
Other—metabolism	6.2 (32)	11.7 (101)	14.4 (28)
Other—secreted protein	1.2 (6)	1.3 (11)	30.3 (59)
Total	100 (n = 520)	100 (n = 864)	100 (n = 195)

Presented are percentages of ESTs in each category, with actual number of ESTs represented in parentheses. Figures from brain and liver were obtained from Adams et al. (3), though slightly different classifications are used. Data for Other—metabolism and Other—secreted protein for heart were derived from the Miscellaneous section of Table 2.

human sequence remains unknown. Tagged clones representing the putative human homologues of such genes are denoted in Table 2 by the organism from which the homologous sequence was derived (indicated in parentheses, Table 2). Also indicated are 31 ESTs matching incompletely (<90%) with known human sequences; these may represent additional members of gene families.

The distribution of ESTs into each of the categories listed in Table 2 was compared against similar distributions for human brain (3) and human hepatocyte (5) ESTs (Table 3). Contractile elements and extracellular matrix proteins were far more abundant in heart than in hepatocyte and brain, where they were absent. However, cytoskeletal, regulatory, and membrane-associated proteins were more abundant in brain than in either heart or hepatocyte, whereas secretory proteins and transcription and translation machinery were most abundant in the hepatocyte. Also of some note, relative amounts of general metabolic enzymes (energy plus other) were roughly similar for heart (19.7%), brain (17.8%), and liver (23.6%).

The identities of the 1485 cDNA clones that were not matched to sequences in the GenBank/EMBL data bases, as well as the identities of the 267 clones matching only with other ESTs, have yet to be determined. To further characterize these clones, we performed dot-blot analyses to elucidate their level of expression in the cardiovascular system and their involvement in myocardial development. Most dots exhibited identical intensity for both the fetal and adult mRNA probes, regardless of whether they represented known or novel cDNA clones (Fig. 2). However, a few striking differences in the levels of expression between the fetal and adult heart are indicated by arrows. These differences may reflect the involvement of the transcripts in the course of myocardial development.

DISCUSSION

The partial sequencing of randomly selected clones from tissue-specific cDNA libraries to generate ESTs has been demonstrated to be an efficient approach to examine tissue expression patterns while compiling extensive sequence data (1–5). Our group has initiated the use of ESTs to catalogue the genes expressed in the human heart and has developed a cost-effective approach (≈U.S. $2.50 per EST) for the cDNA sequencing of this organ (8, 9). Our results, consistent with those of other groups, have shown that 50–60% of ESTs sequenced (including those matching solely to other ESTs) represent additional, previously uncharacterized human transcripts, showing no match to any known sequences in the GenBank/EMBL data bases. We have also identified many genes that may represent additional members of gene families or that may be human homologues of genes previously characterized only in other species. EST sequencing is therefore an effective means of discovering and tagging new genes of the human genome.

ESTs corresponding to known sequences were used to compare broad patterns of gene expression in the human heart, brain, and liver (Table 3). These data correlated well to expected patterns, based on histological characteristics

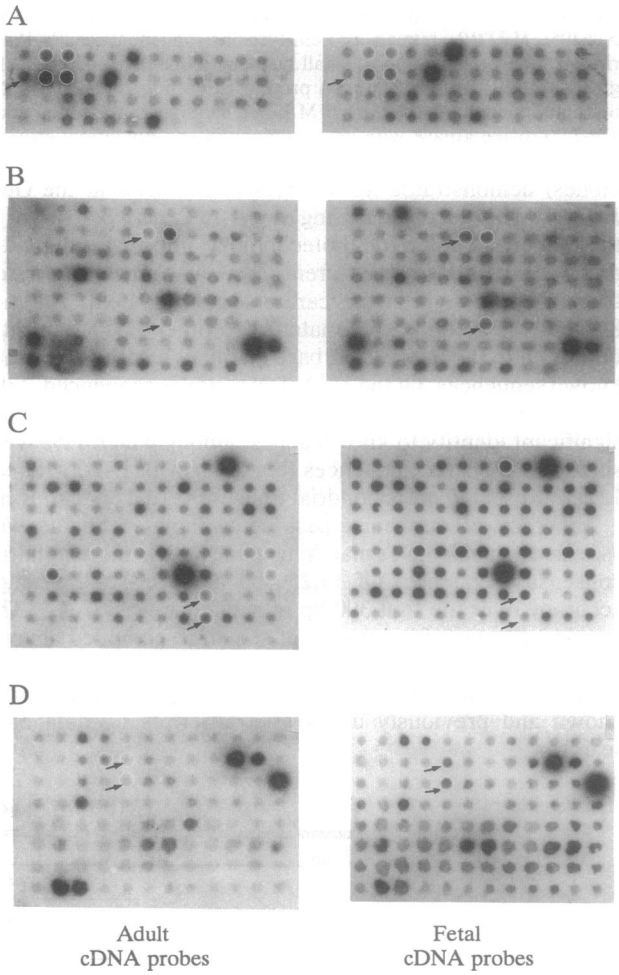

FIG. 2. Dot-blotted PCR products of known (A and B) and novel (C and D) transcripts were used for hybridization. cDNA probes were prepared from human fetal and adult heart mRNAs using [α-^{32}P]dATP and [α-^{32}P]dCTP (refer to *Materials and Methods*). The results indicated the differential expressions during myocardial development, as highlighted by the white circles and arrows.

and physiological function of the individual organs. For example, the high abundance of ESTs representing contractile proteins in the heart is likely associated with its contractile function, whereas the relatively abundant extracellular matrix proteins presumably compose the fibrous skeleton of the heart, which functions to transduce force generated by contractile components to produce useful mechanical work.

In contrast, the brain and liver, organs that neither grossly contract nor contain copious quantities of connective tissue, might be expected to express much less, if any, contractile or extracellular matrix proteins. Rather, much of the support in the brain appears derived from intracellular structural proteins such as actin, tubulin, and glial fibrillary acidic protein, as evidenced by the abundance of ESTs from the human brain representing such transcripts (3); and while the liver as an organ does contain small amounts of connective tissue, the elaboration of such tissue is generally attributed to fibroblasts and, hence, the absence of ESTs representing extracellular matrix components in the hepatocyte cell line (5).

The physiological implications of differences in gene expression patterns between brain and liver have been discussed (3). Introduction of cardiac EST data broadens the scope of this discussion and has permitted several new observations. One additional point of interest was that the proportions of transcripts dedicated to general metabolic processes (i.e., energy plus other metabolism) were approximately equal in each of the three data sets analyzed. This result would seem to indicate that different cell types, regardless of their specific function, need to sustain certain basal activities associated with upkeep and maintenance of general cellular function. Although this concept is intuitively sensible, large-scale identification of genes important to such basal activities would nevertheless prove extremely difficult to perform by using conventional approaches. However, as EST data are compiled from a variety of organs, comparison of such data will no doubt permit a detailed understanding of how differential gene regulation and expression impact on the structure and function of specialized tissues and organs in the human body, while also elucidating novel genes that are ubiquitously expressed and that may therefore be important in the general maintenance of cell function.

While the value of EST sequencing in the identification and rapid sequencing of new transcripts has been well-established, other potential applications of EST information and EST-tagged clones are only now beginning to be explored. One such application is the assembly of full-length cDNA sequences from redundant EST data. It was initially believed that the high redundancy of sequencing of abundant transcripts, such as those of housekeeping genes and contractile elements, would be a critical drawback of the EST approach in the heart. However, our data have demonstrated that such redundancy has not proven prohibitively high; rather, redundancy of sequencing of cDNA clones representing a single transcript allowed for the construction of the complete cDNA sequence of the putative human cardiac troponin T gene (Fig. 1). Similarly, international collaborations to align overlapping ESTs from various projects should expedite the assembly of the full-length cDNA sequences of the complete set of human genes, perhaps as early as 1998.§

§Venter, J. C., Fifth Genome Sequencing and Analysis Conference, Oct. 23–27, 1993, Hilton Head Island, SC.

We have also begun to use EST-tagged clones in dot-blot studies to identify additional genes potentially involved in myocardial development. Although our approach is similar in theory to other differential hybridization techniques, it does possess the advantages that the identities of the spots on the filter are known in advance (if only by the EST) and that differentially hybridizing clones are readily available for use in further characterization. Although this method has been used to study developmental processes, it can also be broadly applied to the study of various pathological processes, such as hypertrophic cardiomyopathy, atherosclerosis, and hypertension, simply by using radiolabeled cDNA probes generated from tissue sample mRNA obtained from patients with the disorders.

The power and utility of EST data in the study of human disease have also been recently manifested in the discovery by Papadopoulos *et al.* (15) of a human homologue to the bacterial *mutL* gene putatively involved in hereditary colon cancer, the isolation and sequencing of which were facilitated and expedited by the prior availability of EST-tagged clones corresponding to the gene of interest. It is only reasonable to assume that as the number of ESTs grows, so too will their use in the identification and isolation of genes putatively involved in a spectrum of human diseases.

The sequencing of randomly selected cDNA clones to generate ESTs is therefore a powerful technique that holds tremendous potential, as existing applications of EST data and tagged clones are further expanded and as new, more powerful applications are discovered.

The technical assistance of B. Kellam, P. Chang, and S.-H. Ng is greatly appreciated. The study was supported by the Medical Research Council, The Ontario Heart and Stroke Foundation, and the Chinese University of Hong Kong. D.M.H. is supported by a Medical Research Council Studentship.

1. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. & Venter, J. C. (1991) *Science* **252**, 1651–1656.
2. Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. & Venter, J. C. (1992) *Nature (London)* **355**, 632–634.
3. Adams, M. D., Kerlavage, A. R., Fields, C. & Venter, J. C. (1993) *Nat. Genet.* **4**, 256–267.
4. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C. & Venter, J. C. (1993) *Nat. Genet.* **4**, 373–380.
5. Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. & Matsubara, K. (1992) *Nat. Genet.* **2**, 173–179.
6. Khan, A. S., Wilcox, A. S., Polymeropoulos, M. H., Hopkins, J. A., Stevens, T. J., Robinson, M., Orpana, A. K. & Sikela, J. M. (1992) *Nat. Genet.* **2**, 180–185.
7. Wilcox, A. S., Khan, A. S., Hopkins, J. A. & Sikela, J. M. (1991) *Nucleic Acids Res.* **19**, 1837–1843.
8. Liew, C. C. (1993) *J. Mol. Cell. Cardiol.* **25**, 891–894.
9. Hwang, D. M., Hwang, W. S. & Liew, C. C. (1994) *J. Mol. Cell. Cardiol.*, in press.
10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
11. Gish, W. & States, D. J. (1993) *Nat. Genet.* **3**, 266–272.
12. Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* **162**, 156–159.
13. D'Alessio, J. M., Noon, M. C., Ley, H. L., III, & Gerard, G. F. (1987) *Focus* **9**, 1–4.
14. Mesnard, L., Samson, F., Espinasse, I., Durand, J., Neveux, J. Y. & Mercadier, J. J. (1993) *FEBS Lett.* **328**, 139–144.
15. Papadopoulos, N., Nicolaides, N. C., Wei, Y.-F., Ruben, S. M., Carter, K. C., *et al.* (1994) *Science* **263**, 1625–1629.