Transcriptional Program of the Endocrine Pancreas in Mice and Humans

Klaus H. Kaestner,[1] Catherine S. Lee,[1] L. Marie Scearce,[1] John E. Brestelli,[1] Athanasios Arsenlis,[1] Phillip Phuc Le,[1] Kristen A. Lantz,[1] Jonathan Crabtree,[2] Angel Pizarro,[2] Joan Mazzarelli,[2] Deborah Pinney,[2] Steve Fischer,[2] Elisabetta Manduchi,[2] Christian J. Stoeckert, Jr.,[1,2] Gerard Gradwohl,[3] Sandra W. Clifton,[4] Juliana R. Brown,[5] Hiroshi Inoue,[6] Corentin Cras-Méneur,[6] and M. Alan Permutt[6]

The Endocrine Pancreas Consortium was formed in late 1999 to derive and sequence cDNA libraries enriched for rare transcripts expressed in the mammalian endocrine pancreas. Over the past 3 years, the Consortium has generated 20 cDNA libraries from mouse and human pancreatic tissues and deposited >150,000 sequences into the public expressed sequence tag databases. A special effort was made to enrich for cDNAs from the endocrine pancreas by constructing libraries from isolated islets. In addition, we constructed a library in which fetal pancreas from *Neurogenin 3* null mice, which consists of only exocrine and duct cells, was subtracted from fetal wild-type pancreas to enrich for the transcripts from the endocrine compartment. Sequence analysis showed that these clones cluster into 9,464 assembly groups (approximating unique transcripts) for the mouse and 13,910 for the human sequences. Of these, >4,300 were unique to Consortium libraries. We have assembled a core clone set containing one cDNA for each assembly group for the mouse and have constructed the corresponding microarray, termed "PancChip 4.0," which contains >9,000 nonredundant elements. We show that this PancChip is highly enriched for genes expressed in the endocrine pancreas. The mouse and human clone sets and corresponding arrays will be important resources for diabetes research.
***Diabetes* 52:1604–1610, 2003**

Despite recent progress in β-cell biology and diabetes research, tools for the treatment of diabetes have not changed fundamentally. Although it is now clear that islet transplantation is a valuable therapeutic approach, this solution is severely limited by the shortage of islet tissue. Over the past decade, significant advances have been made toward identifying the hierarchy of transcription factors that govern pancreatic development (1). In addition, it has been shown that embryonic stem cells can be differentiated in vitro toward insulin-producing cells, although the issue remains controversial (2–4). Despite these discoveries, major obstacles to the isolation, expansion, and differentiation of pancreatic endocrine stem and/or progenitor cells exist, including a lack of appropriate cell surface antibodies for sorting of progenitor cell populations and an only rudimentary understanding of the lineage of β-cells during development and regeneration of the pancreas.

To accelerate the progress toward the identification of endocrine precursor cells and factors that regulate the development and differentiation of β-cells, the National Institute of Diabetes and Digestive and Kidney Diseases sponsored a program entitled "Functional Genomics of the Developing Endocrine Pancreas" in 1999. The Endocrine Pancreas Consortium was created in response to this program to construct and sequence cDNA libraries derived from multiple stages of pancreatic development. Its purpose was to provide the public expressed sequence tag (EST) databases with sequences from mouse and human endocrine pancreas to discover novel transcripts that could be incorporated into custom microarrays to enhance research in diabetes and other metabolic diseases. A limited pancreas microarray or "PancChip," based on a combination of expression analysis and database mining but not on any novel cDNA libraries, has been described previously and made available to the diabetes research community (5).

Here we summarize the efforts of the Endocrine Pancreas Consortium over the past 3 years. We have constructed and sequenced 20 cDNA libraries from a variety of pancreatic sources from mice and humans, yielding >150,000 EST sequences to date. All EST clones have been submitted to the IMAGE (Integrated Molecular Analysis of Genomes and their Expression) Consortium for distribution. The sequences provided by the Endocrine Pancreas Consortium have allowed the identification of thousands of transcripts that have not yet been described in any other library, illustrating the usefulness of targeted EST projects that allow for in-depth sequencing of high-quality libraries. The Consortium ESTs were also clustered to derive non-

From the [1]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania; the [2]Center for Bioinformatics, University of Pennsylvania, Philadelphia, Pennsylvania; [3]INSERM 381, Strasbourg, France; the [4]Genome Sequencing Center, Washington University, St. Louis, Missouri; the [5]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts; and the [6]Department of Internal Medicine, Washington University, St. Louis, Missouri.

Address correspondence and reprint requests to Klaus H. Kaestner, Department of Genetics, University of Pennsylvania, 415 Curie Blvd., Philadelphia, PA 19104. E-mail: kaestner@mail.med.upenn.edu.

Received for publication 13 January 2003 and accepted in revised form 31 March 2003.

DoTS, Database of Transcribed Sequences; EST, expressed sequence tag; GO, Gene Ontology; nrdb, nonredundant database; SSC, sodium chloride–sodium citrate.

redundant clone sets that represent >10,000 assembly groups approximating transcripts expressed in the pancreas of mice and humans. We illustrate one use of these clone sets by building a cDNA microarray that captures a vast proportion of the pancreatic and islet gene expression profile. These nonredundant pancreas clone sets derived from our cDNA libraries promise to become a useful tool for the diabetes research community.

RESEARCH DESIGN AND METHODS

cDNA library construction. Embryonic day (E) 10.5 and 12.5 pancreatic buds, E16.5 pancreas, newborn pancreas, adult pancreas, and adult islets of Langerhans were isolated from ICR mice. Adult pancreas and adult islets were isolated from male mice; embryonic and newborn samples were obtained from both sexes. Islets were isolated by distending the pancreas via the bile duct with 0.38 mg/ml Liberase RI (Roche) in Dulbecco's modified Eagle's medium (Life Technologies) and digestion at 37°C for 22 min, followed by density separation over Histopaque 1077 (Sigma). Human islet RNA was isolated from male and female human cadaver tissue (pooled) obtained from Jack O'Neil at the Joslin Diabetes Research Center, Boston, Massachusetts. Total RNA was isolated with the RNeasy kit (Qiagen), and poly-A^+ was purified by two rounds of selection with the Oligotex kit (Qiagen).

cDNA libraries were constructed with the SuperScript Plasmid system kit (Invitrogen Life Technologies). To increase RT processivity and thereby increase the yield of long cDNAs, oligo-dT primed first-strand cDNA synthesis was performed at 42°C in a 20-μl reaction containing 35 units RNA Guard Porcine RNase inhibitor (Amersham Pharmacia Biotech) and 4.15 μg T4 Gene 32 protein (United States Biochemical). Remaining steps were performed according to the SuperScript kit instructions. Libraries are directional, with a *Not*I site in the oligo-dT primer adapter used to prime the cDNA synthesis and a *Sal*I site at the 5′ end (with the T3 and T7 promoters at the 3′ and 5′ ends, respectively). Double-stranded cDNA was size-fractionated over Sephacryl S-500 (Sigma), cloned into the pSPORT1 vector, and electroporated into DH10B *Escherichia coli*. After preliminary analysis to assess quality, the libraries were amplified once on solid support, and plasmid DNA from each library was prepared.

Library normalization was performed by method number four from Bonaldo et al. (6). Briefly, single-stranded DNA was prepared by phage F1 endonuclease (Gene II protein; Invitrogen Life Technologies) and Exonuclease III digestion and purified by hydroxyapatite chromatography. An aliquot of the single-stranded DNA was used as a template to generate PCR products representing library inserts. Single-stranded library plasmid DNA (0.5 μg) was mixed with 5 μg PCR product and hybridized to an Ecot of 6 (for N1-MMS1; 48 h) or 20 (for N4-HIS1; 160 h) at 30°C in 0.12 mol/l NaCl and 50% formamide. Ecot was calculated using a γ of 0.41 (7) and a formamide correction factor of 0.45 (8). Single-stranded (unhybridized) plasmids were isolated by hydroxyapatite chromatography, repaired to double-stranded DNA using random hexamers, and electroporated into DH10B *E. coli*.

The SMART PCR cDNA Synthesis Kit (Clontech) was used to obtain full-length cDNA from E14.5 control mRNA from wild-type mouse pancreas. A modified 3′-oligo (dT) containing a *Not*I site was used to prime the first-strand synthesis reaction to allow directional cloning. This full-length cDNA was used for PCR amplification to generate double-stranded cDNA. After size selection (>1 kb), *Sal*I adapters were ligated onto both ends of the control cDNA, digested with *Not*I, and cloned into pSPORT1 vector (LTI). The library was electroporated into DH12S cells (Invitrogen Life Technologies). A total of 22,080 clones were picked and sequenced. The same procedure was used to construct a E14.5 neurogenin 3 (*Ngn3*) mutant library from pancreata dissected from *Ngn3*$^{-/-}$ embryos (9), with the exception that this library was cloned into pSPORT2 (LTI). A total of 5,376 clones were picked and sequenced from this library. For the construction of the *Ngn3* wild-type/mutant subtracted library, both control (target) and *Ngn3*$^{-/-}$ (driver) cDNA libraries were amplified once. The ssDNA from the control cDNA library was prepared by infecting with M13KO7 and blocked with 5′-GCGGCCGCT$_{15}$ oligo before hybridization. *Sal*I-digested double-stranded DNA from the *Ngn3*$^{-/-}$ cDNA library was used to synthesize biotinylated RNA with T7 RNA polymerase. Subtractive hybridization was performed at 42°C in 80% formamide, 100 mmol/l HEPES (pH 7.5), 2 mmol/l EDTA, and 0.2% SDS. Streptavidin was used to remove clones common between the two libraries. After repair of the subtracted ssDNA with 5′-GCGGCCGCT$_{15}$ oligo using *Taq* polymerase, the subtracted cDNA library was electroporated into DH12S cells (LTI), and 2,880 clones were picked and sequenced.

For the human islet cDNA library (HR85), poly-A^+ RNA was extracted from adult human pancreatic islets that had been isolated and purified by Dr. Barbara J. Olack (Washington University School of Medicine). cDNA was made by oligo-dT priming (Superscript Plasmid System, Life Technologies), size-selected on agarose gels, and cloned into the *Not*I/*Xho*I site of pBluescript SK (Stratagene). The 5′ *Xho*I site was destroyed after directional cloning. The library was electroporated into DH10B cells and amplified once on solid medium. The average insert size was ~1 kb.

TABLE 1
Distribution of the assembly groups found in the mouse and human Consortium libraries

	Mouse (%)	Human (%)
Total	9,464	13,910
Exact matches	2,461 (26)	3,454 (24.8)
Homologs	1,703 (18)	2,134 (15.4)
Novel family members	2,951 (31.2)	4,900 (35.3)
EST matches only	1,578 (16.7)	2,488 (17.9)
No match	771 (8.1)	934 (6.7)

Exact matches have BLASTX nrdb *P* values of e-200 for the same species in nrdb. Homologs have BLASTX nrdb *P* values of e-100 for other species. Novel family members have BLASTX nrdb *P* values of e-10 in any species. "EST matches only" are not in the previous categories but have BLASTN dbEST *P* values of e-10. Assemblies that do not fit in any of these categories are classified as "no match." Note the similar distribution in both sets.

Sequencing. The EST sequence was generated by the Darwin EST Production Team, Genome Sequencing Center, Washington University, in the St. Louis School of Medicine and submitted to the database of ESTs.

Cluster analysis and microarray construction. All mouse (cutoff date 19 August 2002) and human (2 October 2002) ESTs were obtained from dbEST (10), including those generated as described above. These were used along with identifiable mRNAs in GenBank (10) as part of the Database of Transcribed Sequences (DoTS) v.5 build. Identifiable mRNAs include RefSeq, RIKEN, and GenBank records of type "mRNA" or "RNA" with annotated coding sequences. Details can be found at www.allgenes.org/statistics.html. A manuscript describing the DoTS build process is in progress and will be published elsewhere. Briefly, vector sequences were detected and removed using cross_match (www.phrap.org) and the GenBank vector database. Also removed were ribosomal and mitochondrial sequences, trailing poly A and leading poly T sequences, and poor quality ends where the percentage of unknown bases in a 20-bp window exceeded 20%. Sequences were then blocked for repeats using RepeatMasker (www.phrap.org), with the relevant libraries of repeats depending on the organism used as the source for library construction. If <50 bp of informative sequence remained, then these sequences were not used further. This process eliminated 10,169 of the starting 49,829 mouse sequences and 3,989 of the starting 55,798 human sequences. The remaining blocked sequences were clustered by running a BLASTN (11) matrix with parameters of N = 10 and M = 5. Clusters were formed by a connected components analysis of all the BLASTN matches with minimum cutoff values of 92% identity and 40 bp length. Very large clusters (>10,000 members) were broken up by incrementally increasing cutoff thresholds as needed to 95% identity and 50 bp overlap, then 98% identity and 100 bp overlap. The clusters were assembled to form consensus sequences using the CAP4 algorithm packaged with the Paracel Transcript Assembler (www.paracel.com). The resulting consensus sequences were then blocked with RepeatMasker, clustered with BLASTN (95% identity, 75 bp overlap) and incrementally assembled with CAP4 to complete the DoTS build. The assembly consensus sequences are assigned a stable DoTS identifier and subjected to a number of automated annotations including assignment of Gene Ontology (GO) functions and BLASTX query against the nonredundant database (nrdb) of protein sequences at the National Center for Biotechnology Information. Homology to the nonredundant database was used to generate Table 1. Upon completion of the DoTS build, all assemblies containing an EST from one of the Consortium cDNA libraries were identified to generate the data presented here. These assemblies were used for BLASTN analysis to dbEST for the "EST matches only" category in Table 1. Clone information was not used during the DoTS build process because errors in clone assignments lead to generating chimeric assemblies. However, clone information was used to form assembly groups to compress those DoTS assemblies containing at least one clone from the Consortium libraries to better represent transcripts.

Bacterial clones selected for the array were grown to confluence overnight in flat-bottom 96-well square block plates with a volume of 1 ml LB-Amp medium per well. Plasmid DNA was prepared for each clone using a Qiagen 3000 robot, and the purified DNA was used as a template to amplify the inserts

with PCR. The clones were assigned a PCR score of "pass" indicating a single strong band and "fail" if there were no bands or multiple bands. All the PCR products were purified, eluted with deionized sterile water, diluted with an equal volume of DMSO (Sigma), and printed on poly-L-lysine–coated slides with a Biorobotics Microgrid II. The PancChip 4.0 contains the 3,840 elements of PancChip 2.0 (5) plus 7,226 clones derived from the mouse Consortium libraries plus various control spots. Information about the PancChip 4.0 is available at www.cbil.upenn.edu/EPConDB.

Preparing and labeling RNA. Ten adult CD1 male mice were killed, and the pancreata were immediately homogenized in 10 ml denaturing solution (4 mol/l guanidium thiocyanate, 0.1 mol/l Tris-Cl, pH 7.5, 1% β-mercaptoethanol) per organ. Total RNA was extracted using the acid-phenol extraction method (12). Mouse islets were isolated after collagenase treatment and purification over a Ficoll gradient (13) and immediately homogenized in 1 ml TRIzol Reagent (Life Technologies). RNA was purified following the manufacturer's protocols with the exception that 20 μg glycogen (Roche) was added to each sample. Subsequently, RNA pellets were washed with 75% ethanol and resuspended in 300 μl TES (10 mmol/l Tris, pH 7.5, 1 mmol/l EDTA, 0.1% SDS). The RNA was re-extracted with 600 μl phenol:chloroform:isoamyl alcohol (25:24:1) and precipitated with 1/10 volume 3 mol/l sodium acetate and three volumes ethanol and stored at −80°C until use.

cDNAs were labeled with a modified indirect labeling protocol. Total RNA (20 μg) and 0.4 pmol oligo-dT_{21} were brought to 25 μl with diethyl pyrocarbonate–treated water and incubated for 5 min at 70°C. The RNA mixture was then cooled to 42°C. An equal volume of reaction mix (2× first-strand buffer [Invitrogen], 0.5 mmol/l dATP, 0.5 mmol/l dGTP, 0.5 mmol/l dCTP, 0.3 mmol/l dTTP, 0.2 mmol/l amino-allyl dUTP, 10 mmol/l dithiothreitol, 20 units RNasin [Promega], and 400 units Superscript II Reverse Transcriptase [Invitrogen]) was added, and the reaction was incubated for 2 h at 42°C. The reaction was terminated by bringing it to 0.202 N NaOH and 0.101 mol/l EDTA and incubated at 70°C for 10 min. The reaction was neutralized by adjusting it to 0.334 mol/l Tris-HCl, pH 7.5. After purification with a Micron YM-30 Concentrator (Millipore), the reactions were dried in a vacuum centrifuge. The cDNA was brought up in 15 μl sterile deionized water. A CyScribe Post Labeling Kit containing Cy5 or Cy3 dye (40,000 pmol each) (Amersham Pharmacia RPN5661) were brought up in 15 μl freshly prepared 0.1 mol/l Na bicarbonate buffer, pH 9.0, and added to the cDNA. The reactions were incubated 1 h at 25°C. The reactions were terminated with the addition of 15 μl hydroxylamine and 15 min incubation at 25°C. The coupling reactions were combined, purified with Qiaquick PCR Purification Kit (Qiagen), and precipitated with 1 μl polyacryl carrier (Molecular Research Center), 0.1 volume 1 mol/l Na acetate (pH 5.2), and three volumes ethanol at −20°C overnight. Following precipitation, the pellets were air-dried.

Hybridization. In preparation for hybridization, the cDNA pellets were resuspended in 15 μl sterile deionized water. Then, 5 μl oligo-dT_{21} blocker (0.5 μg/μl) and 2.5 μl Mouse Cot1 DNA (1 μg/μl; Invitrogen) were added to the cDNA and incubated at 95°C for 5 min. An equal volume of prewarmed (42°C) 2× hybridization buffer (50% Formamide, 10× sodium chloride–sodium citrate [SSC], 0.2% SDS) was added, and the sample was transferred to a prehybridized glass array, covered with a coverslip (22 × 60 mm), and incubated overnight in a Corning hybridization chamber at 42°C. The coverslip was removed from the labeled array in 2× SSC, 0.1% SDS. The arrays were then washed two times for 5 min each with agitation: once at 40°C in 0.2× SSC, 0.1% SDS, and once in 0.2× SSC at room temperature and then dried by centrifugation in a slide rack for 3 min at 1,000 rpm.

Scanning and image analysis. All slides were scanned immediately after hybridization and washing using a Genepix 400b scanner. The laser power was set to 100%, and the gain of the photomultiplier tube was varied to avoid signal saturation in any spots. The image analysis was performed with GenepixPro 3.0. Signal and background intensities were determined by the median pixel values. All of the array data, as well as more detailed descriptions of the methods used, are available through www.cbil.upenn.edu/EPConDB.

RESULTS

cDNA libraries from mouse and human pancreas. One of the goals defined in the request for applications that led to the formation of the Endocrine Pancreas Consortium was the cloning and identification of rare transcripts expressed in the mammalian endocrine pancreas. We reasoned that some of these transcripts are not only expressed at low copy number, but also transiently during the development of the endocrine pancreas. One example for this class of genes is *Neurogenin 3* (*Ngn3*), which is expressed in endocrine precursors of the pancreas and gastrointestinal tract but transcriptionally extinguished in the mature β-cell (9,14). The importance of this class of genes is highlighted by the fact that mice homozygous for null alleles of *Ngn3* lack all pancreatic endocrine cells and most gastric enteroendocrine cells (9,15). Therefore, we decided to develop a multitude of cDNA libraries derived from pancreatic tissues of various developmental stages.

TABLE 2
Libraries used in the EST project

Mouse	
E10.5 to 12.5 pancreas	1,115
Amplified E10.5 to 12.5 pancreas 1 M10S1-A	3,672
E14.5 pancreas	9,344
E14.5 pancreas amplified	1,252
E14.5 *Ngn3*$^{-/-}$ pancreas	2,237
E14.5 wild-type/*Ngn3*$^{-/-}$ pancreas subtracted	1,159
E16.5 pancreas library M16Z1	622
Amplified E16.5 pancreas 3 M16S1-A	4,484
E16.5 pancreas library 2 M16B2	9,143
Newborn pancreas	880
Adult pancreas 1	1,036
Adult pancreas 2	592
Normalized mixed mouse pancreas 1 N1-MMS1	23,357
Mouse islets MIZ1	431
Amplified mouse islets 1 MIS1-A	3,018
Total	62,342
Human	
Human fetal pancreas 1A	444
Human fetal pancreas 1B	4,430
Human islets HIZ1	560
Human pancreatic islets	1,204
Normalized human islet 4 N4-HIS 1	14,109
HR85 islet	36,814
Human insulinoma	30,688
Total	88,249

The source of the tissue is indicated as well as the number of EST deposited as of 21 November 2002.

Given that the pancreas contains only ~2% endocrine cells, it was of course desirable to construct cDNA libraries only from pancreatic islets. This was possible for both adult mouse and human islets, where we used isolated islets as starting material. In the developing pancreas, endocrine cells and precursors are interspersed among the mesenchymal and exocrine cells; thus, the endocrine compartment cannot be isolated by physical means. Therefore, we used a genetic approach to enrich for transcripts expressed in the fetal endocrine compartment, making use of mice lacking *Ngn3*. In these mice, the entire endocrine compartment of the pancreas fails to develop (9,15). Thus, by constructing a subtracted library from wild-type and *Ngn3* null (consisting of only mesenchymal and exocrine cells) fetal pancreas, we were able to obtain a cDNA library enriched in transcripts expressed in the endocrine compartment, including those that are only expressed transiently during development. All libraries and source tissues are detailed in Table 2.

A great advantage of the Consortium approach was that library construction and EST sequencing were tightly coordinated. For each library constructed, we initially sequenced 500–1,000 clones to assess the quality of the library in terms of percentage of successful sequencing runs, contamination by ribosomal or mitochondrial sequences, and the complexity of the library. Only good-

quality libraries were sequenced further. In total, >150,000 ESTs were obtained and immediately made available to the public through dbEST. The libraries themselves will be distributed through the IMAGE consortium.

The pancreas transcriptome. In any cDNA sequencing project, even when using normalized and/or subtracted libraries, there is significant redundancy. To assess how many unique transcripts were represented in our cDNA libraries, and to thus obtain a first pass description of the entire pancreatic transcriptome, we generated cDNA assemblies using a computational approach. In this process, all mouse and human sequences contained in dbEST and mRNAs from GenBank are organized as DoTS assemblies as part of the Genomics Unified Schema data system (16), accessible through AllGenes (www.allgenes.org). Each entry or assembly represents a consensus of overlapping, confirmed, and putative transcribed sequences. The complete human and mouse DoTS assemblies, derived from >4 million human and ~2.5 million mouse ESTs and mRNAs, were then queried for those that contain at least one entry from Consortium libraries. The ~50,000 Consortium mouse ESTs available at the time of the last DoTS built were distributed among 13,484 assemblies (approximating unique transcripts). Clone information was used to group assemblies that did not contain overlapping sequences but were from the same transcript, reducing the number to 9,464 assembly groups. Similarly, for ~50,000 Consortium human ESTs, we obtained 20,411 assemblies and 13,910 assembly groups. Of these assembly groups, 1,821 were unique to Consortium mouse libraries, whereas 2,529 were novel in our human libraries. Most of these assembly groups consist of only one cDNA clone (1,604/1,821 for the mouse, 2,210/2,529 for the human sequences). Because most of these assemblies have significant similarity to known genes or other ESTs, it is likely that they represent rare transcripts and not genomics contamination of the libraries. Estimates of the number of transcripts expressed in a given cell range from 10,000 to 20,000. For an organ made up of multiple cell types such as the pancreas, the number of transcripts is likely to be higher. Nonetheless, the 9,464 mouse and 13,910 human assembly groups represent a major fraction of the pancreatic transcriptome.

Despite the fact that millions of ESTs and cDNAs existed before we began our project, this library construction and sequencing effort has yielded several thousand candidates for novel and rare transcripts. This is due in large part to very deep sequencing of our best libraries. Whereas some apparently novel transcripts may be cloning artifacts such as genomic contaminants, assembly groups consisting of sequences from more than one clone and/or having significant homology to another gene are likely to be real transcripts. Expression analysis with our mouse pancreas cDNA microarray derived from these assembly groups also supports the notion that the vast majority of these clones are derived from true transcripts (see below).

Characterization of pancreatic transcripts. The pancreas assembly groups described above were categorized in two ways. Using BLAST similarity to the GenBank/European Molecular Biology Laboratories/DNA Data Bank of Japan nonredundant protein database, we determined the fraction of assembly groups that came from known mouse genes, those that were known genes from other species (but not previously found in mice), and those that were only homologous to known genes and represent novel family members. In addition, we used BLAST similarity to dbEST to assess which of the remaining assembly groups had at least been found as an EST in another species versus those that are completely novel. The result of this analysis is provided in Table 1. For both the human and mouse transcriptomes, the fraction of exact matches to known genes was ~25%, whereas ~25% had no match or only matched previous ESTs.

We also categorized the assembly groups according to GO functions. The GO is a database of universally accepted terms for molecular functions, biological processes, and cellular components (17). The GO also includes the hierarchical relationships between the terms. The GO functions were predicted using a method we recently developed that has an estimated accuracy of ~85% (18). We were able to provide a GO function prediction for approximately half of the assembly groups containing Consortium ESTs. The percentage was much lower, however, for those assembly groups containing ESTs unique to Consortium libraries, which is not surprising given that these sequences are novel and/or rare transcripts. The breakdown of top-level GO functions is illustrated in Fig. 1. The largest group in both the human and mouse transcriptome were genes encoding enzymes, nucleic acid binding proteins, proteins involved in ligand binding, and proteins involved in signal transduction. These were also the top categories found when including all DoTS assemblies (www.allgenes.org/statistics.html). Examples of pancreatic transcripts that appear to be novel but for which functions can be predicted by homology to a portion of a known protein are provided in Table 3.

A microarray based on Consortium libraries. In addition to the discovery of novel genes expressed in the endocrine pancreas that provide the substrate for future functional studies, one obvious application of the clone sets derived from the Consortium libraries is the assembly of cDNA microarrays. For each of the mouse assembly groups described above, we selected one cDNA clone, with preference given to the clone containing the most 3′ sequence. The cDNA inserts were amplified by PCR and analyzed individually by agarose gel electrophoresis. The overall PCR success rate was over 90%. The PCR products were purified, diluted appropriately, and spotted as high-density microarray on glass slides. This mouse PancChip 4.0 combines the clones derived from Consortium libraries described above with those on the previous PancChip 2.0 (5) and contains 13,848 elements representing >10,000 unique genes. To illustrate the enrichment of the PancChip 4.0 for genes expressed in the endocrine pancreas, we performed hybridization experiments with liver, total pancreas, and pancreatic islets. We used the indirect labeling technique for the generation of fluorescently labeled cDNAs. In this approach, RNA is reverse-transcribed in the presence of amino-allyl–modified nucleotides, and the fluorescent moiety is coupled with the cDNA in a second step. We had shown previously that this method combined good sensitivity and reproducibility with accurate representation of signal intensity over a wide range of input

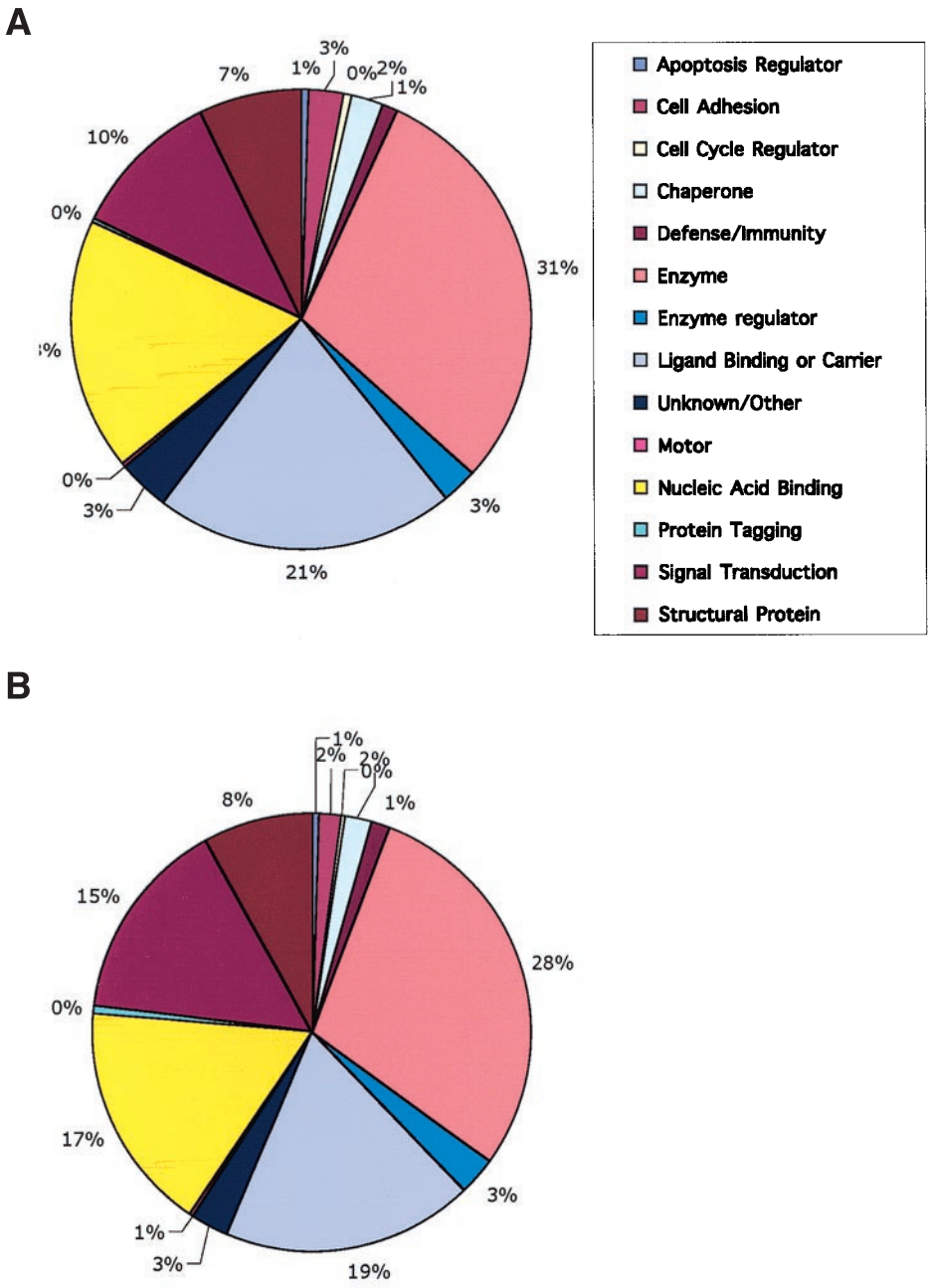

FIG. 1. Primary GO functions of the gene products encoded by the cDNAs found in Consortium libraries. Note that the functional groups are represented in roughly equal proportions in the mouse (*A*) and human (*B*) sets.

RNA (19). When comparing liver cDNA labeled in the red channel (Cy5) with total pancreas RNA labeled in the green channel (Cy3), the vast majority of the spots appear green, indicating higher or exclusive expression in the pancreas, as would be expected for a microarray enriched in targets expressed in the pancreas (Fig. 2*A*). Furthermore, >70% of the spots were also present when hybridized with RNA obtained from isolated pancreatic islets, indicating the usefulness of this array for studies aimed at the endocrine compartment of the pancreas (Fig. 2*B*). The corresponding human PancChip is currently under construction, and the rearrayed clone sets will be made available to the diabetes research community via the IMAGE Consortium. Details on the availability and distribution of both the human and mouse PancChip will be posted on the EPCon website (www.cbil.upenn.edu/EPConDB).

DISCUSSION

A goal of the Endocrine Consortium has been to identify the genes expressed in the developing endocrine pancreas and, in particular, novel ones that may provide cell surface markers or act as key regulators of islet cell development. The different assembly groups resulting from the analysis of the ESTs generated from the Consortium libraries provide an estimate of the number of different mRNAs expressed in the pancreas. The number of mouse and human genes represented in the Consortium ESTs based on this approach is 9,464 and 13,910, respectively. However, this is likely to be an overestimate due to fragmen-

TABLE 3
Examples for functional assignments made for novel transcripts found in the human Consortium libraries

DoTS identifier	GO	Description	Protein domain
DT.95216000	Cell adhesion	90% identity to lithostathine 1 β precursor	N-CAM transmembrane
DT.95264275	Cell adhesion	90% identity to 9% of AL031588	EGF-like protocadherin
DT.97411316	Chaperone	93% identity to 22% of peptidyl-prolyl *cis-trans* isomerase B precursor)	Zn-finger
DT.94156620	Enzyme	58% identity to 68% of pancreatic lipase-related protein 1	Lipase
DT.94309574	Enzyme	85% identity to 40% of TESP4	Peptidase
DT.97353778	Enzyme	73% identity to 28% of trypsinogen 16	Peptidase
DT.97386400	Enzyme regulator	67% identity to 10% of calcineurin-binding protein cabin 1	Protein phosphatase inhibitor
DT.94292455	Ligand binding	86% identity to 80% similar to GTP-binding protein	GTP binding
DT.97347758	Ligand binding	66% identity to 16% of DNA repair protein RAD50	ATP binding
DT.94327513	Signal transducer	93% identity to 21% of frizzled homolog 10	Transmembrane
DT.94126502	Transporter	76% identity to 21% of NAD-dependent formate dehydrogenase α-subunit	Cation transporter
DT.97776412	Nucleic acid binding	75% identity to 17% of SRE-ZBP	Zn-finger
DT.97777897	Nucleic acid binding	60% identity to zinc finger protein 133	Zn-finger
DT.97805496	Nucleic acid binding	55% identity to 65% of XM_116105	Zn-finger
DT.91889905	Nucleic acid binding	78% identity to 79% of XM_085788 similar to HNRP A1	Single-strand RNA binding protein
DT.92325855	Nucleic acid binding	89% identity to 11% of hypothetical protein KIAA0017	Suppressor of hairless
DT.91638544	Nucleic acid binding	87% identity to 35% of nuclear autoantigen Sp-100	Bromodomain
DT.97822149	Nucleic acid binding	77% identity to 51% of DAX-1 protein-human	Nuclear receptor
DT.95264017	Nucleic acid binding	97% identity to 30% of XM_009529 hypothetical protein	Leucine zipper

tation of genes from nonoverlapping 5′ and 3′ segments. The overlap between the mouse and human clone sets based on best reciprocal BLAST hits is ~25%. This is likely an underestimate due to the nature of DoTS assemblies, which are often small or fragmented. A more accurate assessment of the genes represented by the assembly groups and the overlap between the two sets based on genomic alignments to the human and mouse genomes is in progress. Although >150,000 ESTs were generated for this project, no single library was sequenced to completion, and therefore it is possible that very rare transcripts specific to each library remain to be discovered.

The discovery of known and novel genes was also consistent between the mouse and human sets. Known genes (exact matches and homologues) comprised 40–44% of the transcripts found. Transcripts that could be

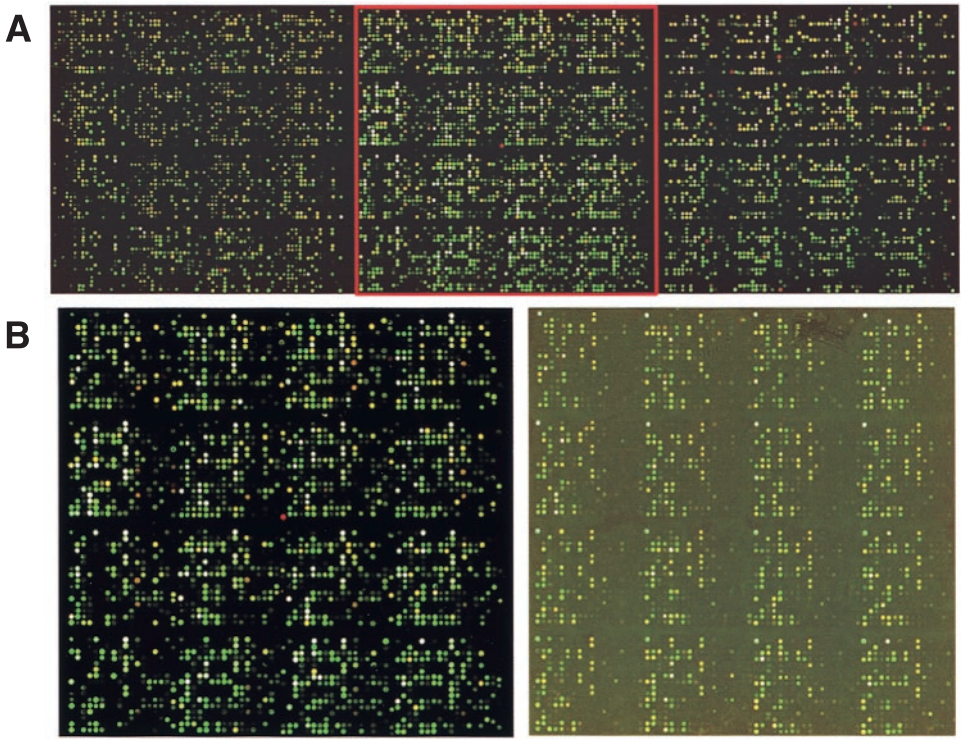

FIG. 2. PancChip 4.0: a cDNA microarray derived from the mouse Consortium libraries. *A*: The PancChip 4.0 with 13,848 elements was hybridized to total pancreas (green) or total liver cDNA (red). Note that most spots appear green or yellow, indicating expression in the pancreas. *B*: Comparison of total pancreas to islet RNA. Part of the image shown in *A* (indicated by the red frame) was enlarged (left panel) and is compared with the corresponding portion of the PancChip 4.0 hybridized to RNA isolated from purified islets (right panel). Note that the majority of the spots shows signal with both probes.

identified based on homology to other species constituted 30–35%. This category provides much promise because the homology to known proteins increases our ability to predict their function and, consequently, identify novel candidates for cellular processes of interest. Approximately one-quarter of the transcripts could not be matched to any known sequence or only to ESTs from other species. These will require further information from other sources (such as expression profiles from microarray analysis) to gain insight into their role in the pancreas. Recently, similar analyses of hematopoietic stem cells (20) and stromal cells (21) were published. Over half of the transcripts found in these studies were of the unknown or "EST only" category, whereas only 5–6% represented new family members. Thus, the pancreatic transcripts provide a richer resource for analysis of novel genes related to known proteins.

Sequence analysis of the Consortium libraries also revealed the surprising fact that >4,000 assembly groups only contained clones from our libraries, thus representing completely novel transcripts. These sequences represent a rich resource for future genetic and biochemical investigations. In addition, all the sequences obtained and characterized by the Consortium are valuable for further annotation of the mouse and human genomes. In many cases, alignment of our assemblies to genomic sequence revealed that there were no gene predictions matching the assembly; thus, our cDNA sequencing project will lead to an increase in the number of known genes.

The classes of gene function represented in the Consortium clone sets roughly mirrors what was observed for the set of DoTS assemblies for all tissues. The consistency of the categories for mouse and human GO function assignments (Fig. 1) suggests that the observed distribution reflects the true transcript classes of normal mammalian pancreatic tissues. One striking difference between the predicted pancreatic GO functions and those predicted for all tissues is the underrepresentation of signal transduction genes in pancreatic tissues (10–15% of predicted functions) compared with all tissues (16–26%). Defense/immunity predictions were also underrepresented (1% for pancreatic vs. 4–8% for all tissues), although this is not surprising because the normal pancreas is not involved in the immune response.

Perhaps the most immediate impact of this work is the development of tissue-specific microarrays derived from our clone sets. These clone sets allow for the production of large numbers of cDNA microarrays at low cost, facilitating the use of large numbers of replicates in gene expression profiling experiments, which leads to increased data quality. Also, because many of the cDNAs contained on our microarrays are not contained on commercial platforms at present, they provide a unique tool to diabetes researchers.

ACKNOWLEDGMENTS

We gratefully acknowledge support through National Institute of Diabetes and Digestive and Kidney Diseases Grant 56947 to K.H.K. and Grant 56954 to M.A.P.

We are grateful to Dr. Doug Melton for his valuable input during the entire project. We also thank Brian Brunk for helping generate the DoTS build and thank members of the Computational Biology and Informatics Laboratory in the Penn Center for Bioinformatics for helpful discussions.

REFERENCES

1. Edlund H: Transcribing pancreas. *Diabetes* 47:1817–1823, 1998
2. Lumelsky N, Blondel O, Laeng P, Velasco I, Ravin R, McKay R: Differentiation of embryonic stem cells to insulin-secreting structures similar to pancreatic islets. *Science* 292:1389–1394, 2001
3. Rajagopal J, Anderson WJ, Kume S, Martinez OI, Melton DA: Insulin staining of ES cell progeny from insulin uptake. *Science* 299:363, 2003
4. Edlund H: Pancreatic organogenesis: developmental mechanisms and implications for therapy. *Nat Rev Genet* 3:524–532, 2002
5. Scearce LM, Brestelli JE, McWeeney SK, Lee CS, Mazzarelli J, Pinney DF, Pizarro A, Stoeckert CJ Jr, Clifton SW, Permutt MA, Brown J, Melton DA, Kaestner KH: Functional genomics of the endocrine pancreas: the pancreas clone set and PancChip, new resources for diabetes research. *Diabetes* 51:1997–2004, 2002
6. Bonaldo MF, Lennon G, Soares MB: Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 6:791–806, 1996
7. Britten RJ, Graham DE, Neufeld BR: Analysis of repeating DNA sequences by reassociation. *Methods Enzymol* 29:363–418, 1974
8. Soares MB, Bonaldo MF: Constructing and screening normalized DNA libraries. In *Genome Analysis: A Laboratory Manual.* Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press, 1998, p. 49–157
9. Gradwohl G, Dierich A, LeMeur M, Guillemot F: Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc Natl Acad Sci U S A* 97:1607–1611, 2000
10. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 28:10–14, 2000
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 215:403–410, 1990
12. Chomczynski P, Sacchi N: Single-step method of RNA isolation. *Anal Biochem* 162:156–159, 1987
13. Scharp DW, Kemp CB, Knight MJ, Ballinger WF, Lacy PE: The use of ficoll in the preparation of viable islets of Langerhans from the rat pancreas. *Transplantation* 16:686–689, 1973
14. Schwitzgebel VM, Scheel DW, Conners JR, Kalamaras J, Lee JE, Anderson DJ, Sussel L, Johnson JD, German MS: Expression of neurogenin3 reveals an islet cell precursor population in the pancreas. *Development* 127:3533–3542, 2000
15. Lee CS, Perreault N, Brestelli JE, Kaestner KH: Neurogenin 3 is essential for the proper specification of gastric enteroendocrine cells and the maintenance of gastric epithelial cell identity. *Genes Dev* 16:1488–1497, 2002
16. Davidson S, Crabtree J, Brunk B, Schug J, Tannen V, Overton G, Stoeckert C: K2/Kleisli and GUS: experiments in integrated access to genomic data sources. *IBM Systems Journal* 40:113–117, 2001
17. Gene Ontology Consortium: Creating the gene ontology resource: design and implementation. *Genome Res* 11:1425–1433, 2001
18. Schug J, Diskin S, Mazzarelli J, Brunk BP, Stoeckert CJ Jr: Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res* 12:648–655, 2002
19. Manduchi E, Scearce LM, Brestelli JE, Grant GR, Kaestner KH, Stoeckert CJ: Comparison of different labeling methods for 2-channel high-density microarray experiments. *Physiol Genomics.* 10:169–179, 2002
20. Phillips RL, Ernst RE, Brunk B, Ivanova N, Mahan MA, Deanehan JK, Moore KA, Overton GC, Lemischka IR: The genetic program of hematopoietic stem cells. *Science* 288:1635–1640, 2000
21. Hackney JA, Charbord P, Brunk BP, Stoeckert CJ, Lemischka IR, Moore KA: A molecular profile of a hematopoietic stem cell niche. *Proc Natl Acad Sci U S A* 99:13061–13066, 2002