Biochem Genet (2009) 47:117–125
DOI 10.1007/s10528-008-9211-5

Analysis of Expressed Sequence Tags in Porcine Uterus Tissue

Hui Chai · Dong-Liang Yu · Bing Zhang · Yan Fu · Song-Nian Hu

Received: 15 February 2008 / Accepted: 15 August 2008 / Published online: 24 January 2009

Abstract Two non-normalized cDNA libraries of uteri from Danish Landrace and Chinese Erhualian pigs were constructed, and 13,756 expressed sequence tags (ESTs) were randomly sequenced. The ESTs were clustered by Phrap software, and 6,139 distinct tentative consensus sequences were produced, including 2,730 contigs and 3,409 singlets. Using Blast tools, these 6,139 candidate genes were compared to the nr and nt databases; 5,210 of them were assigned putative functions, whereas 929 potentially represent new genes. Highly expressed genes appear to be associated with basic energy metabolism, transferase activity, localization, cellular physiological process, protein binding, and nucleic acid binding. Antileukoproteinase was the most highly expressed gene, corresponding to endometrial differentiation and conceptus or fetal development.

Keywords Porcine uterus · Porcine ESTs · Gene expression · Pig reproduction

H. Chai · D.-L. Yu
College of Life Sciences, Zhejiang University, Hangzhou 310058, People's Republic of China

B. Zhang · S.-N. Hu
Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, People's Republic of China

Y. Fu (✉)
College of Animal Sciences, Zhejiang University, Hangzhou 310029, People's Republic of China
e-mail: fuyan@zju.edu.cn

Y. Fu · S.-N. Hu
James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310008, People's Republic of China

Introduction

The reproductive process is central to pig production efficiency, and the ovary and the uterus are two important organs for pig reproduction (Caetano et al. 2003). The uterus holds the embryo and enables it to derive nutrients from and deliver wastes to the maternal blood. It is widely accepted that uterine capacity is the greatest restraint on litter size in pigs (Jiang et al. 2003). The mechanisms underlying ovarian follicular development have been extensively studied (Caetano et al. 2003). There is, however, no information available about gene expression in the uterus.

An expressed sequence tag (EST) is a partial sequence of a transcribed spliced nucleotide sequence; ESTs can be used in gene identification and can give information about a gene acting under different conditions. The number of EST hits per gene is used as an indicator for gene expression in the study of gene expression profiles. Currently, ESTs have become an effective resource to investigate gene expression in specific tissues and developmental stages (Zhang et al. 2004).

In this study, we present sequencing and analysis of 10,879 high-quality 5′-end ESTs generated from two porcine cDNA libraries from the uterus. By using the Phrap, Phred, and Consed packages (http://www.phrap.com/), all of the ESTs were assembled into 6,139 clusters, which were then annotated based on homolog searching and were also classified by Gene Ontology (GO) (http://www.geneontology.org/).

Materials and Methods

Library Construction and EST Sequencing

The total RNA of two different uterus tissues collected from Danish Landrace and Chinese Erhualian pigs was isolated using Trizol reagent (Invitrogen, Carlsbad, USA). The uterus stage used for library construction was "pregnant." They were both at the same stage from animals of the same age. The mRNA was purified with Oligotex mRNA kits (Qiagen, Hilden, Germany). Synthesis of cDNA was performed using Superscipt II-RT (Invitrogen) and DNA polymerase I (Promega, Madison, USA). The cDNA fragment that was flanked with *Eco*RI adapter (Stratagene, La Jolla, USA) and digested by *Xho*I (Stratagene) was cloned into the vector treated by *Eco*RI (Promoga) and *Xho*I (Promega). The cDNA clone was transferred into *Escherichia coli* to be amplified. The plasmids were extracted according to the alkaline lysis protocol and used for capillary sequencing (MegaBACE 1000).

Data Processing and Bioinformatics Analysis

The chromatogram files as raw data were processed for base-calling and quality assessment by Phred software. The low-quality sequences were trimmed off with Q20 (99% accuracy). The vector sequences were screened with the Cross Match program (version 0.990319). The simple repeat sequences were also masked by Repeat Masker Perl script (http://repeatmasker.genome.washington.edu/). All ESTs were compared with linker sequences, and those that included linker were

interrupted on the linker locus. The ESTs that were longer than 100 bp were retained for later analysis. All high-quality and clean ESTs were assembled by Phrap software, with 40 minmatch and 0.95 repeat stringency. Contigs and singlets were called clusters. All clusters were compared to the nonredundant nucleic acid (nt) and protein (nr) database provided by GenBank with Blast tools. The best hit for each query was used for function assignment and subsequently manually checked. Function category was performed with the GO database. All clusters were also compared with the human EST database for homologous sequences by Blast.

Significant Differentiation Statistic Test

The ESTs from porcine uterus cDNA libraries were divided into two groups, one from Danish Landrace and the other from Chinese Erhualian. The Web tool IDEG6 (http://telethon.bio.unipd.it/bioinfo/IDEG6_form/) was used to detect differentially expressed gene categories with $P < 0.05$.

Results

Overview of cDNA Libraries and Clustering

In order to get an overview of porcine genes in the uterus, two nonnormalized cDNA libraries were constructed from different breeds (Danish Landrace and Chinese Erhualian). In total, 13,756 cDNA clones were randomly selected (6,905 from Danish Landrace and 6,851 from Chinese Erhualian) and partially sequenced from cDNA 5′ ends to generate ESTs (Table 1). The initial EST sequences were screened to mask vector sequences and porcine mitochondrial sequences. After comparison to linker sequences, a few chimeric sequences were discarded. Finally, we gained 10,879 high-quality ESTs (5,576 from Danish Landrace and 5,303 from Chinese Erhualian) with a minimum length of 100 bp and an average read length of 349 bp. These high-quality ESTs were assembled with the Phrap to identify those representing nonredundant transcripts. As a result, the total ESTs were assembled into 2,730 contigs (1,412 from Danish Landrace and 1,318 from Chinese Erhualian). The remaining 3,409 ESTs (1,776 from Danish Landrace and 1,633 from Chinese Erhualian) could not be assembled into contigs (so-called singlets). Therefore, a total of 6,139 assembly sequences (clusters) were generated (3,188 from Danish Landrace and 2,951 from Chinese Erhualian). Among them, 14 clusters were larger than 100, which contained 2,201 ESTs. The maximal cluster size was 365.

Annotation and Gene Identification

All of the clusters were compared with the nt and nr database. A total of 1,876 clusters (28.7%) had homologies in the nr database with more than 30% alignment length and 25% identity (E-value $<$ 1e-5), and 4,759 clusters (72.8%) had matches with the nt database (E-value $<$ 1e-10). Although 1,327 clusters (20.3%) could not find any homologous sequence in these two databases, 37 contigs had 50 or more

Table 1 Number of ESTs from two porcine uterus cDNA libraries

Pig breed	Total ESTs	>Q20	>100 bp	Clusters	Contigs	Singlets
Danish Landrace	6,905	6,083	5,576	3,188	1,412	1,776
Chinese Erhualian	6,851	5,749	5,303	2,951	1,318	1,633
Total	13,756	11,832	10,879	6,139	2,730	3,409

Table 2 Annotation of porcine uterus contigs (50 or more ESTs)

Gene name	Total ESTs	ESTs from Danish Landrace	ESTs from Chinese Erhualian
ALP	1072	523	549
Predicted: similar to NADH dehydrogenase	300	149	151
SPPI	247	121	126
Predicted: similar to EF-1 alpha	246	117	129
Predicted: similar to RPS8	220	148	72
UABP-2	199	135	64
GedBS044P	175	88	87
COX subunit I	168	38	130
Rps12 protein	160	79	81
COIII	132	63	69
Predicted: similar to vimentin	121	59	62
Predicted: similar to RPP0	118	85	33
Similar to granulin isoform 1 precursor	110	53	57
Procollagen, type III, alpha 1	104	48	56
3-beta-hydroxysteroid dehydrogenase/delta-5-delta-4 isomerase	86	27	59
Predicted: similar to RPL9	83	58	25
Cytochrome P450 11A1	77	19	58
RPS11 protein	68	35	33
Selenium binding protein 1	65	32	33
Predicted: similar to RPP2	60	48	12
Predicted: similar to RPL6	60	43	17
Predicted: similar to ribosomal protein S16	59	27	32
Beta 2-microglobulin	56	25	31
Predicted: similar to RPS23	55	43	12
Predicted: similar to hypothetical protein FLJ22175	53	29	24
Predicted: similar to RPL21	51	38	13
Ribosomal protein L10a	51	25	26
Predicted: similar to guanine nucleotide binding protein	50	22	28
Ribosomal protein L28	50	27	23

EST sequences (Table 2). More than half of the largest contigs had consensus sequences that were homologous to genes involved in protein synthesis (initiation factors, elongation factors, and ribosomal proteins). There was a tissue-specific

gene, uteroferrin-associated basic protein-2 (UABP-2, NM_213845), which is also abundantly expressed.

As revealed by the cDNA frequency, antileukoproteinase (ALP) was the most abundantly expressed gene in porcine uterus. Other highly expressed genes were NADH dehydrogenase (NP_008644), secreted phosphoprotein-I (SPPI, NM_214023), and elongation factor 1α (EF-1 alpha, NM_001097418).

The gene expression profiles of the two breeds were different. The genes that were more highly expressed in Chinese Erhualian pigs were cytochrome c oxidase (COX) subunit I, 3-beta-hydroxysteroid dehydrogenase/delta-5-delta-4 isomerase (3 beta-HSD), and cytochrome P450 11A1; in Danish Landrace the highly expressed genes were ribosomal protein S8 (RPS8), UABP-2, 60S acidic ribosomal protein P0 (RPP0), ribosomal protein L9 (RPL9), 60S acidic ribosomal protein P2 (RPP2), 60S ribosomal protein L6 (RPL6), ribosomal protein S23 (RPS23), and ribosomal protein L21 (RPL21).

The cDNAs were classified according to the GO index, with categories for cellular component, molecular function, and biological process. There were 83 contigs (432 individual cDNA clones) with GO cellular component annotations. Each contig contained more than six cDNA clones, and the numbers were similar in the two pig breeds. Based on the number of cDNA clones, the majority of cellular mRNA-encoded component was ribonucleoprotein complex (Table 3). There were 602 contigs (2,370 individual cDNA clones) clustered into the group involved in molecular function. The genes in this group expressed higher in Chinese Erhualian than in Danish Landrace, and were associated with nucleic acid binding, nucleotide binding, protein binding, and hydrolase activity. The consensus sequences for most contigs were homologous to genes whose products were involved in transferase activity (catalytic activity). There were 727 contigs (2,833 individual cDNA clones) with GO biological process annotations. More than half of the genes were involved in cellular metabolism. The genes expressed higher in Chinese Erhualian were associated with cellular physiological processes and localization.

Discussion

A primary object of EST sequencing is gene identification (Jiang et al. 2003). One method to identify the factors that control ovarian function is to characterize the genes that are expressed in the uterus (Caetano et al. 2003). The random sampling strategies resulted in highly expressed genes represented by many EST sequences. The frequency of cDNA within each tissue could be determined as each clone was sequenced from its original library (Zhang et al. 2004). We analyzed 10,879 high-quality ESTs generated from two nonnormalized porcine uterus cDNA libraries. The sequences clustered into 2,730 contigs, and the contig sequences were blasted against the nr or nt databases in Genbank. The genes were associated with common cell functions, such as energy metabolism, protein synthesis, signal transduction, cell communication, transport, development, and cell-cycle regulation. Genes associated with uterus-specific functions, such as *UABP-2*, were also identified (Table 2).

Table 3 Number of cDNAs within categories of the GO index

GO index	Total cDNAs	cDNAs from Danish Landrace	cDNAs from Chinese Erhualian
Cellular component			
Protein complex, respiratory chain complex I	20	11	9
Protein complex, ribonucleoprotein complex	40	18	22
Protein complex, transcription factor complex	6	2	4
virion, viral capsid	17	11	6
Molecular function			
Binding	4	1	3
Binding, ion binding	31	17	14
Binding, lipid binding	81	41	40
Binding, nucleic acid binding	88	23	65
Binding, nucleotide binding	14	3	11
Binding, pattern binding	1	1	0
Binding, peptide binding	1	1	0
Binding, protein binding	92	35	57
Binding, ribonucleoprotein binding	1	0	1
Binding, selenium binding	5	2	3
Binding, steroid binding	4	1	3
Binding, vitamin binding	2	1	1
Catalytic activity	15	7	8
Catalytic activity, hydrolase activity	33	13	20
Catalytic activity, isomerase activity	1	1	0
Catalytic activity, ligase activity	1	1	0
catalytic activity, lyase activity	5	2	3
Catalytic activity, oxidoreductase activity	25	11	14
Catalytic activity, small protein conjugating enzyme activity	4	1	3
Catalytic activity, transferase activity	146	90	56
Enzyme regulator activity, GTPase regulator activity	4	2	2
Enzyme regulator activity, enzyme activator activity	10	4	6
Enzyme regulator activity, enzyme inhibitor activity	3	1	2
Enzyme regulator activity, kinase regulator activity	6	2	4
Enzyme regulator activity, ornithine decarboxylase regulator activity	1	0	1
Obsolete molecular function, FK506-sensitive peptidyl-prolyl cis-trans isomerase	1	1	0
Obsolete molecular function, Rho small monomeric GTPase activity	10	3	7
Obsolete molecular function, ba3-type COX	1	0	1
Obsolete molecular function, barbed-end actin capping/severing activity	2	1	1
Obsolete molecular function, cell adhesion molecule activity	10	2	8

Table 3 continued

GO index	Total cDNAs	cDNAs from Danish Landrace	cDNAs from Chinese Erhualian
Biological process			
Obsolete biological process, peroxidase reaction	3	1	2
Physiological process	3	0	3
Physiological process, cellular physiological process	112	43	69
Physiological process, homeostasis	24	11	13
Physiological process, localization	127	41	86
physiological process, metabolism	372	189	183
Physiological process, organismal physiological process	5	2	3
Physiological process, response to stimulus	47	21	26
Regulation of biological process, regulation of development	2	1	1
Regulation of biological process, regulation of enzyme activity	3	1	2
Regulation of biological process, regulation of physiological process	29	13	16

As revealed by the cDNA frequency, ALP cDNA was the most abundant cDNA (Table 2). ALP is a physiologic inhibitor of granulocytic serine proteases. Other highly expressed genes in porcine uterus were NADH dehydrogenase and SPPI. NADH dehydrogenase, known as the NADH:Ubiquinone oxidoreductase, is complex I of the mitochondrial electron transfer chain, and it catalyzes the transfer of electrons from NADH to coenzyme Q (Malathy et al. 1990). It is well known that SPPI is a highly phosphorylated form that has been associated with cell transformation (Roberts and Bazer 1988). The fourth highest expressed gene is EF-1 alpha, an essential component of the eukaryotic translational apparatus, which is a GTP-binding protein that catalyzes the binding of aminoacyl-transfer RNAs to the ribosome (Li et al. 2002). Other genes involved in protein synthesis (including ribosomal proteins) were also highly expressed. We found 12 ribosomal protein genes for all 80 components of the ribosome among 6,139 clusters, which indicates that we have found significant expression information in the porcine uterus.

The gene expression profiles were different for the two porcine species. The genes with higher expression in Chinese Erhualian than in Danish Landrace were COX subunit I, 3 beta-HSD, and cytochrome P450 11A1. COX is one of a superfamily of proteins that act as the terminal enzymes of respiratory chains. The two main classes are COXs and quinol oxidases. Mitochondrial COX and its bacterial homologs catalyze electron transfer and proton translocation reactions across membranes (Mammalian Gene Collection Program Team 2002). Three-beta-HSD catalyzes the oxidative conversion of delta 5-3-beta-hydroxysteroids to the delta 4-3-keto configuration, and is therefore essential for the biosynthesis of all classes of hormonal steroids, namely, progesterone, glucocorticoids, mineralocorticoids, androgens, and estrogens (Adams et al. 1995). This may be associated with the larger litter size of the Chinese Erhualian, compared with the Danish Landrace. The function of the mammalian P450 system concerns its role as the rate-limiting

enzyme in the synthesis of all steroid hormones and many prostaglandins and leucotrienes. As such, the P450s play a central role in mineral balance, sugar regulation, reproduction, water balance, digestion of lipids, vascular tone, pain, and inflammation (Fahrenkrug et al. 2002).

In summary, our study provides a catalog of 2,730 contigs derived from 10,879 cDNA sequences obtained from porcine uterus tissues. For most contigs, the frequency of the sequenced genes was too low to study the gene expression reliably across tissues. Almost a quarter of the EST clusters did not have any match with nt or nr databases. Why are there so many anonymous ESTs? Cirera et al. (2000) reviewed various possible reasons, but here we propose two main explanations. First, the pig genome project is ongoing, and many genes expressed in the uterus still have not been identified. Second, a number of ESTs sequenced from uterus cDNA libraries are too short to be identified, and they may represent untranslated regions of the gene; 3′ UTR sequences vary more than the coding regions between organisms. Some of the ESTs may represent transcripts that have diverged to the extent that they are not recognized as orthologs; others may be inaccurate sequence data (Cirera et al. 2000). Thus, sequencing cDNA from the mammalian uterus is a good strategy for novel gene identification.

At present, the strategies of gene prediction include two basic approaches (Wang et al. 2003; Rogic et al. 2001). One is the ab initial computational prediction using statistic information, and the other is the integrated method of computational and sequence similarity search among species. The latter relies largely on cDNA resources and homologous comparison among relatively close organisms. Thus, these ESTs from the porcine uterus should be a useful resource for the pig genome sequencing project and its annotation.

Acknowledgment This study was supported by the Sino-Danish Pig Genome Project.

References

Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, Sutton G, Blake JA, Brandon RC, Chiu M-W, Clayton RA, Cline RT, Cotton MD, Hughes JE, Fine LD, Fitzgerald LM, FitzHugh WM, Fritchman JL, Geoghagen NSM, Glodek A, Gnehm CL, Hanna MC, Hedblom E, Hinkle PS Jr, Kelley JM, Klimek KM, Kelley JC, Liu L-I, Marmaros SM, Merrick JM, Moreno-Palanques RF, McDonald LA, Nguyen DT, Pellegrino SM, Phillips CA, Ryder SE, Scott JL, Saudek DM, Shirley R, Small KV, Spriggs TA, Utterback TR, Weidman JF, Li Y, Barthlow R, Bednarik DP, Cao L, Cepeda MA, Coleman TA, Collins E-J, Dimke D, Feng P, Ferrie A, Fischer C, Hastings GA, He W-W, Hu J-S, Huddleston KA, Greene JM, Gruber J, Hudson P, Kim A, Kozak DL, Kunsch C, Ji H, Li H, Meissner PS, Olsen H, Raymond L, Wei Y-F, Wing J, Xu C, Yu G-L, Ruben SM, Dillon PJ, Fannon MR, Rosen CA, Haseltine WA, Fields C, Fraser CM, Venter JC (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377(Suppl.):3–16

Caetano AR, Johnson RK, Pomp D (2003) Generation and sequence characterization of a normalized cDNA library from swine ovarian follicles. Mamm Genome 14:65–70

Cirera S, Winter AK, Fredholm M (2000) Why do we still find anonymous ESTs? Mamm Genome 11:689–693

Fahrenkrug SC, Smith TPL, Freking BA, Cho J, White J, Vallet J, Wise T, Rohrer G, Pertea G, Sultana R, Quackenbush J, Keele JW (2002) Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. Mamm Genome 13:475–478

Jiang Z, Zhang M, Wasem VD, Michal JJ, Zhang H, Wright RW (2003) Census of genes expressed in porcine embryos and reproductive tissues by mining an expressed sequence tag database based on human genes. Biol Reprod 69:1177–1182

Li N, Zhao ZH, Liu ZL, Zhao XB, Lian ZX, Wu CX (2002) Analysis of expressed sequence tags from porcine liver organ. Sci Agric Sinica 35(12):1525–1528

Malathy PV, Imakawa K, Simmen RC, Roberts RM (1990) Molecular cloning of the uteroferrin-associated protein, a major progesterone-induced serpin secreted by the porcine uterus, and the expression of its mRNA during pregnancy. Mol Endocrinol 4(3):428–440

Mammalian Gene Collection (MGC) Program Team (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. PNAS 99(26):16899–16903

Roberts RM, Bazer FW (1988) The functions of uterine secretions. J Reprod Fertil 82(2):875–892

Rogic S, Mackworth AK, Ouellette FBF (2001) Evaluation of gene-finding programs on mammalian sequences. Genome Res 11:817–832

Wang J, Li ST, Zhang Y, Zheng HK, Xu Z, Ye J, Yu J, Wong GK-S (2003) Vertebrate gene predictions and the problem of large genes. Nat Rev Genet 4:741–749

Zhang B, Jin W, Zeng YW, Su ZX, Hu SN, Yu J (2004) EST-based analysis of gene expression in the porcine brain. Genomics Proteomics Bioinformatics 2:237–244