EST-based Analysis of Gene Expression in the Porcine Brain

Bing Zhang[1], Wu Jin[2], Yanwu Zeng[1], Zhixi Su[1], Songnian Hu[1,3]*, and Jun Yu[1,3]*

[1] *James D. Watson Institute of Genome Sciences, Zhejiang University/Hangzhou Genomics Institute, Hangzhou 310008, China;* [2] *The Second Affiliated Hospital of China Medical University, Shenyang 110004, China;* [3] *Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China.*

Since pig is an important livestock species worldwide, its gene expression has been investigated intensively, but rarely in brain. In order to study gene expression profiles in the pig central nervous system, we sequenced and analyzed 43,122 high-quality 5′ end expressed sequence tags (ESTs) from porcine cerebellum, cortex cerebrum, and brain stem cDNA libraries, involving several different prenatal and postnatal developmental stages. The initial ESTs were assembled into 16,101 clusters and compared to protein and nucleic acid databases in GenBank. Of these sequences, 30.6% clusters matched protein databases and represented function known sequences; 75.1% had significant hits to nucleic acid databases and partial represented known function; 73.3% matched known porcine ESTs; and 21.5% had no matches to any known sequences in GenBank. We used the categories defined by the Gene Ontology to survey gene expression in the porcine brain.

Key words: pig, brain, expressed sequence tags, gene identification

Introduction

Pig represents a clade between the orders of primate and rodent, providing an evolutionary model with factitious selected interference, and becomes a potential donor for xenotransplantation (*1*). Analyses of the pig genome and its gene expression would help understanding mammalian organogenesis, development, and evolution. Currently, expressed sequence tags (ESTs; ref. *2*) have become an effective resource to investigate gene expression in specific tissues and developmental stages (*3*, *4*). Quite a few ESTs derived from various porcine tissues and physiological conditions have been deposited in the database (*4–6*), but those generated from libraries of porcine brain are less represented (*4*). There are more than 150,000 porcine ESTs that have been submitted to GenBank, most of which were generated from two projects, the USDA-MARC EST project (*7*) and the Pig EST Coordination Program led by Iowa State University (*8*). The majority of the ESTs were produced in normalized porcine cDNA libraries. We could obtain important information of gene expression profiles in various tissues and developmental stages from these projects. Particularly, the gene expression profiles in muscle (*6*),

embryo (*5*, *9*), and reproductive glands (*5*) were investigated intensively. The Porcine Gene Index (PGI), containing 24,820 tentative consensus sequences and 38,107 singletons (Release 8.0, www.tigr.org/tdb/ssgi/), has been integrated to facilitate further gene identification in pig and comparative genomics research. As more porcine ESTs have been generated, they have become a significant resource for porcine physical and genetic mapping, and also a preliminary task for completing the pig genome sequencing (*10*).

Research for the gene expression in the brain promotes understanding its function and mechanism. At least one-third of all genes in the whole genome are expressed in mammalian brain (*10*, *11*), many of which are large genes (*12–14*). The porcine brain is interesting both for its applicability to agriculture and for its significance as a medical model (*4*). In this report, we present sequencing and analysis of 43,122 high-quality 5′ end ESTs generated from seven different porcine cDNA libraries of cortex cerebrum, brain stem, and cerebellum. By Phrap (http://www.phrap.com/) clustering, all of the ESTs were assembled into 16,101 clusters, which were annotated based on matches to protein and nucleotide acid databases in GenBank, and were also classified by Gene Ontology (http://www.geneontology.org/).

* Corresponding authors.
E-mail: husn@genomics.org.cn;
junyu@genomics.org.cn

Results

Overview of cDNA libraries and clustering

To get an overview of the repertoire and the temporal expression profile of porcine genes in brain, seven non-normalized cDNA libraries were constructed from different regions of porcine brain (Table 1). It is well known that the most proliferation of mammalian neurons occurs during the prenatal period; and the outgrowth and differentiation of neurons occurs in the neonatal period, followed by synaptogenesis in the first postnatal week (*15*). There-fore, the tissues during the five time points were used to construct cDNA library (See Materials and Methods). A total of 53,993 random selected cDNA clones were partially sequenced from cDNA 5′ ends to generate ESTs. The initial EST sequences were screened to mask vector sequences and porcine mitochondria genome sequences by CROSS_MATCH (version 0.990319). The simple repeat sequences were masked by the RepeatMasker program (http://repeatmasker.genome.washington.edu/). Comparing to linker sequences, a few chimera sequences were discarded. Finally, we gained 43,122 high-quality EST sequences, with minimum length of 100 bp and average read length of 388 bp.

Table 1 Summary of Porcine Brain ESTs

	Developmental stage	Brain region	High-quality ESTs
Prenatal	Foetus 50 d*/100 d	Cerebellum	9,829
	Foetus 50 d	Cortex cerebrum	8,650
	Foetus 50 d	Brain stem	8,423
Postnatal	Adult	Whole brain	4,742
	Earlyborn 107 d	Cortex cerebrum	5,941
	Newborn 115 d	Brain stem	5,537

*d denotes the days of pregnancy.

The 43,122 high-quality ESTs were analyzed with the Phrap assembly program to identify those representing redundant transcripts. As a result, a total of 32,360 ESTs were assembled into 5,339 contigs, while the remaining 10,762 ESTs could not be assembled into contigs. Therefore, a total of 16,101 assembly sequences (hereafter referred to "clusters") were generated. Among them, 25 clusters' sizes were larger than 100, which contained 3,997 ESTs. The maximal cluster size was 343. Most of the genes were represented by only one EST.

Annotation and gene identification

All of clusters were searched with the non-redundant (nr) protein database and the nucleic acid (nt) database in GenBank by BLAST. A total of 4,929 (30.6%) clusters had homologies to the nr database with more than 30% alignment length and 25% identity (E-value<1e–5), and 12,098 (75.1%) clusters had matches with the nt database (E-value<1e–10), while 3,462 (21.5%) clusters could not find homologous sequences with the nt or nr database. Furthermore, the BLAST search of porcine EST database indicated that 73.3% clusters had significant matches (E-value<1e–5). In all, 21,461 EST sequences (4,929 clusters) that displayed homologies with the nr database were annotated according to the best matches.

In this report, we defined highly expressed gene as the one that was represented by more than 30 (>0.06%) ESTs (Table 2). The most abundant gene family is the ribosomal protein family involving 4,247 EST sequences (not shown in Table 2). Other highly expressed genes mainly involve two classes of genes: the first is housekeeping genes encoding cytoskeleton, catalysis, cell defense, energy metabolism, signal transmit, substance degradation, and cell cycle proteins; and the second is tissue-specific genes encoding nervous system specific functional proteins. The heat shock protein family (HSP27, HSPJ2, HSP10, HSP70, HSP90) concerned with a variety of cell stress response was also highly expressed.

Among these highly expressed genes, the genes encoding for cytoskeletal proteins, tubulin, and actin are necessary to intracellular protein trafficking, cell division and substance transport. As the main component of spindle, abundant expressed tubulin genes indicate that cell mitosis occurs at high frequency in the porcine brain. Apolipoprotein E (APO E), a con-

Table 2 High Expression Gene Collection (the number of ESTs>30) in porcine brain ESTs*

Accession number	Definition	Accession number	Definition
NP_035783	tubulin alpha 1	NP_036168	visinin-like 1
NP_003395	14-3-3 protein beta	NP_006588	heat shock 70kDa protein 8
NP_076205	tubulin beta	AAR24462	neuron growth-associated protein 43
AAH70494	FTH1 protein	NP_034937	MARCKS-like protein
NP_001393	translation elongation factor alpha	P13696	Phosphatidylethanolamine-binding protein
NP_999377	similar to transthyretin	NP_112252	proteolipid protein 1
Q02399	cell division protein kinase 5	NP_035868	14-3-3 protein eta
AAA36746	thymosin beta-10	AAP36938	G protein beta polypeptide
NP_005554	stathmin 1	XP_293312	similar to H3 histone
NP_001605	actin gamma 1 propeptide	AAH10469	PEA15 protein
NP_003286	translationally-controlled 1	NP_733779	S-phase kinase-associated protein 1A
AAA57047	ubiquitin	NP_996734	reticulon 1 isoform C
NP_035784	tubulin alpha 2	NP_001437	fatty acid binding protein
NP_006817	14-3-3 protein tau	NP_068832	thymosin beta
NP_035785	tubulin beta 5	NP_004763	neuronal protein 3.1
P00355	glyceraldehyde-3-phosphate dehydrogenase	CAA40268	protein synthesis initiation factor 4A
AAB92373	polyubiquitin	NP_999473	apolipoprotein E
NP_847893	neuronatin	Q95KL4	selenoprotein W
A44367	tumor-specific transplantation antigen p23	NP_085048	Nedd4 family interacting protein
CAC17762	organic cation transporter	P01965	hemoglobin alpha chain
AAP36911	*Homo sapiens* GNAS complex locus	P05124	creatine kinase b chain
NP_001060	tubulin beta polypeptide	O18882	vacuolar ATP synthase
NP_999518	peptidyl-prolyl cis-trans isomerase A	0408231A	protein S100
AAP36156	calmodulin	A24327	carboxypeptidase E
NP_998928	carboxyl-terminal hydrolase L1	NP_035558	synaptosomal-associated protein
AAH66941	laminin receptor	O97555	rab GDP dissociation inhibitor alpha
AAL99919	CLL-associated antigen KW-12	2GSRA	porcine class Pi glutathione S-Transferase
NP_999136	complement cytolysis inhibitor	AAH61581	PP protein
XP_213821	similar to alpha NAC/1.9.2. Protein	1F3WA	muscle pyruvate kinase
NP_999138	90-kDa heat shock protein	AAB06444	extracellular matrix associated protein

*The accession number and definition are assigned by the best match to protein database in GenBank.

stituent of very-low-density lipoproteins and high-density lipoproteins, participates in many biological processes, including lipid metabolism, cholesterol transport, tissue repair, immune response, regulation, as well as cell growth and differentiation. In human, APO E is related to cardiovascular disease, Alzheimer's disease (*16*), neuromuscular disease, multiple sclerosis, stroke, diabetes, cancer (*17*), and renal disease (*18*).

Eta, theta, and beta subunits of the most abundant brain specific protein 14-3-3 (511 ESTs) were found in the porcine brain. Eta subunit shows 100% identity to that in mouse; beta subunit has uniform 232 out of 246 amino acid residues with human beta isoform; and theta subunit has 229 identical amino acids out of 245 with human. Recently, 8 subunits of the 14-3-3 protein family have been isolated, which participated in various regulations of target protein function, including the inhibition of apoptosis, change of nuclear export and import rate, change of the intrinsic catalytic activity of the target protein, and protection of the target protein from proteolysis and dephosphorylation (*19*). Another brain-specific protein, neuronatin, is also expressed abundantly in the porcine brain (*20*). It is mainly expressed at the porcine foetal stage with a few found in the newborn

porcine brain. In our experiment, we found no neuronatin expressed in the adult porcine brain, which indicates that neuronatin participates in neurogenesis and nervous system development.

The ESTs encoding complement cytolysis inhibitor (21), heat shock protein, ubiquitin, thymosin, and selenoprotein W (22) were represented at high level in the porcine brain. These genes involve functions in cell defense, immunity regulating, substance degradation, and antioxidation.

Genes encoding peptide hormones

As a vital neuroendocrine organ, mammalian brain secretes various hormones, and a number of hormones are the encoding product of genes. Forty ESTs were found to represent encoding peptide hormones, including cholecystokinin (CCK), neurotenssin (NT), insulin-like-growth factor 2 (IGF2), neuropeptide Y precursor (NPY), and calcitonin receptor-stimulating peptide (CRSP). Among them, CCK has the complete CDS and the polymorphism in its 3′ UTR is identical to the previous report (23). CCK has been found to play important roles in endocrine and neural systems in the periphery as well as in the central nervous system, such as function in satiety, anxiety, intestinal transit, and inhibitory influence on GnRH-1 neuronal migration, contributing to the appropriate entrance of these neuroendocrine cells into the brain (24). According to the EST distribution in different brain regions, CCK was found expressed in porcine cerebrum and cerebellum, but not in brain stem. This implies that it has different endocritic roles among different brain regions.

Another important peptide hormone is neurotensin, a member of a structurally related peptide family that has a variety of central nervous system (CNS) effects. The neurotensin system in behavior is related to the effects of antipsychotic drugs (25). Neurotensin serves as an endogenous antipsychotic-like compound; neurotensin neurotransmission is integral to the mechanism of action of antipsychotic drugs. Comparison with human neurotensin preproprotein reveals that more than 90% amino acid residues are conserved. So pig may serve as an animal model for studying the function of neurotensin.
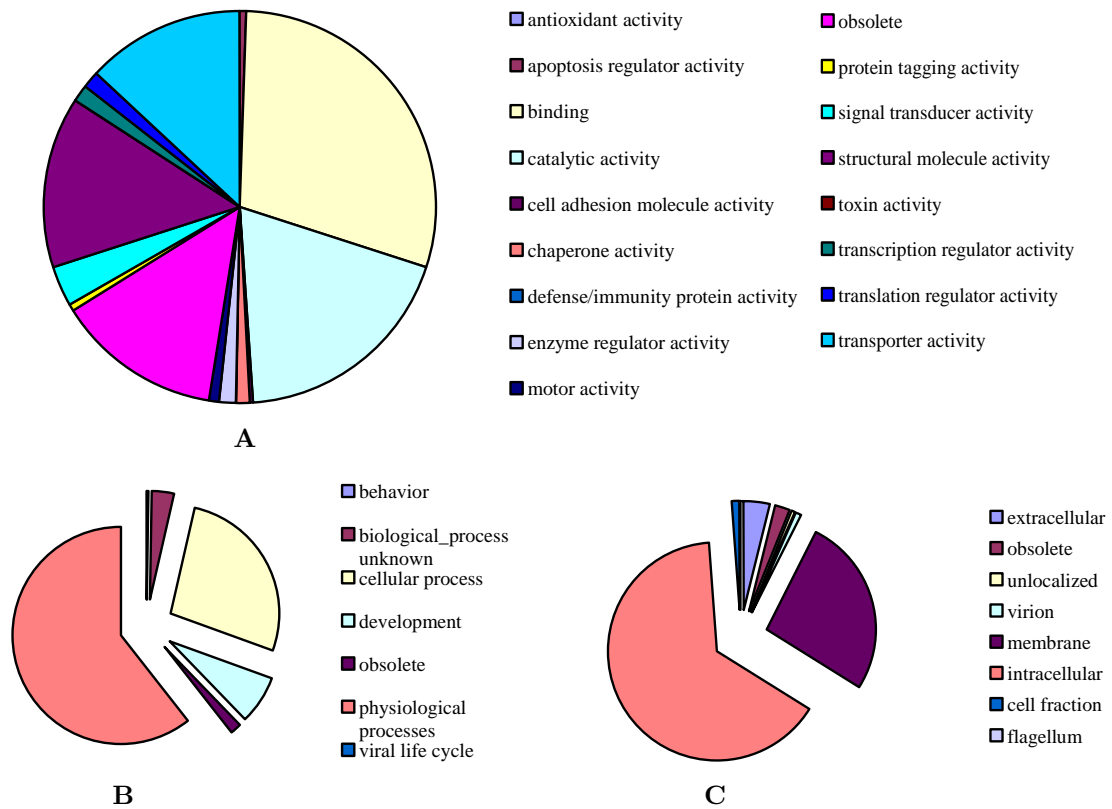
Gene classification based on Gene Ontology

To investigate the functional profile of genes expressed in the porcine brain, 4,390 (26.6%) clusters with similarities to known proteins were classified into different categories based on Gene Ontology (26, 27). We assigned 4,306 clusters to 17 main molecular functional categories, 4,331 clusters to 7 main biological processes, and 4,309 clusters to 8 main cellular components (Figure 1). Of the molecular function categories, binding proteins and catalytic activity proteins (enzymes) contained a majority of ESTs. In the molecular functional subcategory, proteins for nucleic acid binding, protein binding, oxidoreductase activity, hydrolase activity, structural constituent of ribosome, carrier activity, and ion transporter activity accounted for overwhelming ESTs; and each of them composed of more than 2,000 EST sequences. It is remarkable that the high expression of carrier activity and ion transporter proteins is necessary to nerve system activity in vertebrate brain. As far as biological process is concerned, proteins in the subcategories of metabolism, cell growth/maintenance, cell communication, morphogenesis, and response to external stimulus and stress are highly expressed.

Differentially expressed gene groups at different developmental stages

Because brain has different cell cyclic, motoric and cognitive functions at various developmental stages, we supposed that gene expression differences in prenatal and postnatal porcine brains should be found. In the present data, 26,902 and 16,220 ESTs were generated from the prenatal and postnatal porcine brain tissues, respectively. On Gene Ontology categories, the abundance of a few gene expression groups displayed significant differences ($P<0.05$) at different developmental stages (Figure 2). Relative to the postnatal stages, the gene classifications that were significantly up-regulated at the prenatal stages include nucleic acid binding, nucleotide binding, transferase activity, kinase activity, hydrolase activity, motor activity, RNA-directed DNA polymerase, globins, signal transducer activity, structural constituent of ribosome, structural constituent of muscle, structural constituent of cytoskeleton, transcription regulator activity, translation regulator activity, channel class transporter activity, and auxiliary transport protein activity. On the other hand, expression levels of other gene categories at the postnatal stages displayed significant increase, which involve steroid binding, isoprenoid binding, vitamin binding, heavy metal binding, chaperone activity, peptide hormone, and transposase activity.

Fig. 1 Classification of porcine brain ESTs based on Gene Ontology. A. molecular function. B. biological process. C. cellular component. Predicted functions of porcine brain ESTs were assigned by sequence homology with the GenBank non-redundant protein database.
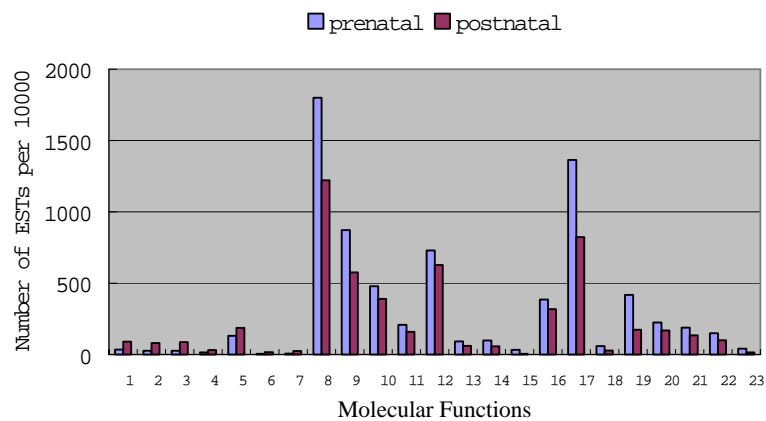
Fig. 2 Significant difference (P<0.05) of gene expression groups from prenatal and postnatal porcine brains. The chi-square test was used to detect the significant difference. The molecular functions from 1-23 are as follows: 1. steroid binding; 2. isoprenoid binding; 3. vitamin binding; 4. transposase activity; 5. chaperone activity; 6. peptide hormone; 7. heavy metal binding; 8. nucleic acid binding; 9. nucleotide binding; 10. transferase activity; 11. kinase activity; 12. hydrolase activity; 13. motor activity; 14. RNA-directed DNA polymerase; 15. globin; 16. signal transducer activity; 17. structural constituent of ribosome; 18. structural constituent of muscle; 19. structural constituent of cytoskeleton; 20. transcription regulator activity; 21. translation regulator activity; 22. channel class transporter activity; 23. auxiliary transport protein activity.

Discussion

A primary object of EST sequencing is gene identification. The universality in sampling is vital to gain those low expressed transcripts. In this report, we found 69 different ribosomal protein genes of all 80 components of the ribosome among 16,101 clusters, which indicates that we have found significant expression information in the porcine brain. The random sampling strategies result in highly expressed genes represented by many EST sequences. Hence, the cDNA library normalization is a necessary step to reduce redundancy of highly expressed genes and to gain as much expression information as possible. The overwhelming genes expressed in the porcine brain are few except those from mitochondrial, ribosome, and cytoskeleton proteins, such as tubulin and actin. We analyzed 43,122 high-quality ESTs generated from seven non-normalized porcine brain cDNA libraries. Almost a quarter of EST clusters did not have any matches with nt or nr databases in GenBank, much less with functional known genes. Why are there so many anonymous ESTs? Cirera *et al* (*28*) reviewed various possible reasons, but here we suppose that it could be explained by two main reasons: firstly, there are still a lot of genes expressed in mammalian brain that have not been identified so far; and secondly, a number of ESTs sequenced from brain cDNA libraries are too short to be identified and they represent 5′ untranslated regions of gene. Thus, sequencing cDNA from mammalian brain is a good strategy for novel gene identification.

At present, the strategies of gene prediction include two basic approaches (*14*, *29*), one is the *ab initio* computational prediction using statistic information; the other is the integrated method of computational and sequence similarity search among species. The latter relies largely on cDNA resources and homologous comparison amongst relative close organisms. Thus, these ESTs from the porcine brain would be a wonderful resource for the coming pig genome sequencing project and genome annotation.

A great number of differentially expressed genes between prenatal and postnatal porcine brains were found in this report. Accordingly, neurogenesis and development of brain can be characterized by the large up-regulated expression of genes for biosynthetic enzymes, cytoskeleton, energy metabolism, and signal transduction. At the postnatal developmental stage, the up-regulated gene categories of steroid, isoprenoid, and vitamin binding proteins mainly attribute to transthyretin, sterol carrier protein 2, and steroid membrane binding protein. Transthyretin is the major transporter of thyroid hormone and vitamin A in cerebrospinal fluid and binds Alzheimer β-peptide; and this binding protects brain against neurodegeneration and mediates clearance of the brain β-peptide. Furthermore, the up-regulated chaperone activity attributes to HSP90 and peptidyl-prolyl cis-trans isomerase A, which help protein folding and refolding. This probably indicates the differentiation of protein degradation in prenatal and postnatal brains.

Materials and Methods

Tissue collection

Porcine brain tissues were collected from embryo, piglet, and adult pigs for cDNA library construction. The cerebellum, cortex cerebrum, and brain stem were collected from 50-day-old embryo, 100-day-oldembryo, 107-day-old early-born piglet, and 115-day-old newborn piglet. The whole brain was collected from adult pig (Table 1). All tissues were snap-frozen in liquid nitrogen and stored at –80°C before use.

Library construction and EST sequencing

For each tissue, total RNA was isolated using TRIzol Reagent (Invitrogen, Carlsbad, USA). The mRNA was purified with Oligotex mRNA Kits (QIAGEN, Hilden, Germany). The cDNA synthesis was performed using Superscipt II-RT (Invotrigen) and DNA polymerase I (Promega, Madison, USA). The cDNA fragment that was flanked with *Eco*R I adaptor (Stratagene, La Jolla, USA) and digested by *Xho* I (Stratagene) was cloned into the vector treated by *Eco*R I (Promoga) and *Xho* I (Promega). The cDNA clone was transferred into *E. coli* to be amplified. The plasmids were extracted according to 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 assessing by Phred software (Phred-Phrap-Consed package). The low-quality sequences were trimmed off with Q20 (99% accuracy). The vector sequences were screened with

CROSS_MATCH program. The simple repeat sequences were also masked by RepeatMasker perl script. All ESTs were compared to linker sequences and those ESTs that included linker were interrupted on locus of linker. 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. As assembled results, contigs and singlets were called clusters. All clusters were compared to the nucleic acid database and protein database in GenBank by BLAST (*30*). The clusters that had hits with protein database were assigned to gene classification based on Gene Ontology. The clusters were annotated according to the best hit against known protein sequences or nucleic acid sequences. All clusters were also searched for homologies against human EST database by BLAST.

Significant differentiation statistic test

All ESTs from porcine brain cDNA libraries were divided into two groups, one from prenatal brain and the other from postnatal brain. A web tool of IDEG6 (*31*) was used to detect differentially expressed gene categories with $P<0.05$.

Acknowledgements

References

1. Onishi, A., *et al.* 2000. Pig cloning by microinjection of fetal fibroblast nuclei. *Science* 289: 1188-1190.

2. Adams, M.D., *et al.* 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

3. Adams, M.D., *et al.* 1992. Sequence identification of 2,375 human brain genes. *Nature* 355: 632-634.

4. Nobis, W., *et al.* 2003. Development of a porcine brain cDNA library, EST database, and microarray resourse. *Physiol. Genomics* 16: 153-159.

5. Jiang, Z., *et al.* 2003. Census of genes expressed in porcine embryos and reproductive tissues by mining an expressed sequences tag database based on human genes. *Biol. Reprod.* 69: 1177-1182.

6. Zhao, S.H., *et al.* 2003. Complementary DNA macroarray analyses of differential gene expression in porcine fetal and postnatal muscle. *J. Anim. Sci.* 81: 2179-2188.

7. Fahrenkrug, S.C., *et al.* 2002. Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. *Mamm. Genome* 13: 475-478.

8. Tuggle, C.K., *et al.* 2003. EST-based gene discovery in pig: virtual expression patterns and comparative mapping to human. *Mamm. Genome* 14: 565-579.

9. Blomberg, L.A. and Zuelke, K.A. 2004. Serial analysis of gene expression (SAGE) during porcine embryo development. *Reprod. Fertil. Dev.* 16: 87-92.

10. Hoog, C. 1991. Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. *Nucleic Acids Res.* 19: 6123-6127.

11. Sutcliffe, J.G. 1988. mRNA in the mammalian central nervous system. *Annu. Rev. Neurosci.* 11: 157-198.

12. Nagase, T., *et al.* 2001. Prediction of the coding sequences of unidentified human genes. XXI. The complete sequences of 60 new cDNA clones from brain which code for large proteins. *DNA Res.* 8: 179-187.

13. Ohara, O., *et al.* 1997. Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins. *DNA Res.* 4: 53-59.

14. Wang, J., *et al.* 2003. Vertebrate gene predictions and the problem of large genes. *Nat. Rev. Genet.* 4: 741-749.

15. Mody, M., *et al.* 2001. Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc. Natl. Acad. Sci. USA* 98: 8862-8867.

16. Poirier, J. 2003. Apolipoprotein E and cholesterol metabolism in the pathogenesis and treatment of Alzheimer's disease. *Trends Mol. Med.* 9: 94-101.

17. Moore, R.J., *et al.* 2004. Apolipoprotein E and the risk of breast cancer in African-American and non-Hispanic white women. *Oncology* 66: 79-93.

18. Liberopoulos, E., *et al.* 2004. Apolipoprotein E and renal disease. *Am. J. Kidney Dis.* 43: 223-233.

19. Takahashi, Y. 2003. The 14-3-3 proteins: gene, gene expression, and function. *Neurochem. Res.* 28: 1265-1273.

20. Joseph, R., *et al.* 1995. Neuronatin mRNA: alternatively spliced forms of a novel brain-specific mammalian developmental gene. *Brain Res.* 690: 92-98.

21. Diemer, V., *et al.* 1992. Expression of porcine complement cytolysis inhibitor mRNA in cultured aortic smooth muscle cells. Changes during differentiation in vitro. *J. Biol. Chem.* 267: 5257-5264.

22. Whanger, P.D. 2000. Selenoprotein W: a review. *Cell. Mol. Life Sci.* 57: 1846-1852.

23. Gubler, U., *et al.* 1984. Cloned cDNA to cholecystokinin mRNA predicts an identical preprocholecystokinin in pig brain and gut. *Proc. Natl. Acad. Sci. USA* 81: 4307-4310.

24. Giacobini, P., *et al.* 2004. Cholecystokinin modulates migration of gonadotropin-releasing hormone-1 neurons. *J. Neurosci.* 24: 4737-4748.

25. Kinkead, B. and Nemeroff, C.B. 2002. Neurotensin: an endogenous antipsychotic? *Curr. Opin. Pharmacol.* 2: 99-103.

26. The Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. *Genome Res.* 11: 1425-1433.

27. Ashburner, M., *et al.* 2000. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25: 25-29.

28. Cirera, S., *et al.* 2000. Why do we still find anonymous ESTs? *Mamm. Genome* 11: 689-693.

29. Rogic, S., *et al.* 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* 11: 817-832.

30. Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

31. Romualdi, C., *et al.* 2003. IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. *Physiol. Genomics* 12: 159-162.