DOI: 10.1007/s11427-005-0181-0

Gene expression profiling in porcine mammary gland during lactation and identification of breed- and developmental-stage-specific genes

SU Zhixi[1], DONG Xinjiao[2], ZHANG Bing[1], ZENG Yanwu[1], FU Yan[1], YU Jun[1,3] & HU Songnian[1,3]

1. James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310008, China;
2. School of Life and Environmental Sciences, Wenzhou Normal College, Wenzhou 325027, China;
3. Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China

Correspondence should be addressed to Hu Songnian (email: husn@genomics.org.cn) or Yu Jun (email: junyu@genomics.org.cn)

Received 31 May, 2005; accepted July 11, 2005

Abstract A total of 28941 ESTs were sequenced from five 5′-directed non-normalized cDNA libraries, which were assembled into 2212 contigs and 5642 singlets using CAP3. These sequences were annotated and clustered into 6857 unique genes, 2072 of which having no functional annotations were considered as novel genes. These genes were further classified into Gene Ontology categories. By comparing the expression profiles, we identified some breed- and developmental-stage-specific gene groups. These genes may be relative to reproductive performance or play important roles in milk synthesis, secretion and mammary involution. The unknown EST sequences and expression profiles at different developmental stages and breeds are very important resources for further research.

Keywords: expressed sequence tags, expression profile, domestic pig, mammary gland.

The mammary gland provides an excellent system to study questions pertaining to organogenesis, cell differentiation and oncogenesis. Intensive efforts have been made to understand the development of the mammary gland, particularly in terms of lactogenesis, cells apoptosis and tissue remodeling. The regulation mechanism of the mammary gland development is complex. Briefly speaking, systemic hormones and inductive epithelial — stromal interactions control are two major pathways to regulate mammary gland development[1]. Many genes that may play a role during developmental stages of the mammary gland have been identified. However, many intimate mechanisms of the mammary gland function remain unknown[2].

To further understand the molecular mechanisms and mammary gland development, it is necessary to obtain a high-coverage collection of genes relevant to different physiological stages and construct the physiological-stage specific gene expression profile. Single-pass, partial sequencing of randomly selected cDNA clones from cDNA libraries to generate expressed sequence tags (EST), combined with bioinformatics analysis, has proved powerful for the discovery of alternative splicing (AS)[3,4], novel genes[5] and the characterization of gene function[6,7], the differential and quantitative analysis of expression patterns[8,9], as well as the evaluation of the gene expression profile in a given tissue[10−12]. Previous efforts to generate ESTs from mammary gland cDNA libraries were limited. Ethical considerations have limited the feasibility of constructing cDNA libraries from high-quality human mammary tissues that represent different developmental stages. Only 3655 human ESTs have been generated from normal mammary tissue[13]. Porcine mammary gland is not only an alternative model organ used to study human mammary gland, but also a good system for bioreactor developing[14]. A further research of the porcine mammary gland may help us understand the development of mammary gland. At the same time, it will improve the development of pig production. EST generated from porcine mammary gland is not currently available in Unigene database yet. Therefore, a catalog of mammary-derived sequences from pig could obviously provide researchers with more comparable sequence resources.

The lactation performance of sow is closely correlated with its reproductive performance. Previous studies showed that the production and composition of milk are critical for survival of the suckling pig, and milk production is one of the most important factors limiting neonatal pig growth[15]. The polymorphism of some high molecular weight proteins (HMWP) in sow milk are correlated with sow's reproductive performance[16].

The present report is on the establishment of a gene expression profile of mammary gland based on the analysis of 28941 ESTs as well as preliminary results of comparison of the expression profiles at different breeds and developmental stages. As a result, some breed- or developmental-stage-specific gene groups were identified. These genes may be relative to reproductive performance or play important roles in milk synthesis, secretion and mammary involution. This result will facilitate the use of functional genomics approaches (i.e. cDNA microarrays) aiming at unraveling the molecular mechanisms underlying mammary gland development.

1 Material and methods

1.1 Library construction

In order to provide a rough estimate of the gene expression levels, all the libraries were not normalized. Five non-normalized cDNA libraries were generated from different breeds and different developmental stages (Table 1). The gilt was killed at specific developmental stages, and mammary tissues were recovered. All tissues were snap frozen in liquid nitrogen. The total RNA was isolated with TRIzol Reagent (Invitrogen) and the mRNA was purified with Oligotex mRNA Kits (Qiagen). The cDNA synthesized with Superscipt II-RT (Invitrogen) and DNA Polymerase I (Promega) was franked by *Eco*R I adaptor (Stratagene) and treated by *Xho* I (Stratagene), and then cloned into *Eco*R I and Xho I digested vector. The plasmid was transformed into the *E. coli* (DH10B) to amplify the cDNA.

1.2 EST sequencing and sequence pre-processing

The cDNA clones were picked randomly. Plasmids were isolated according to a standard alkaline lysis protocol, and sequenced with MegaBACE 1000 (Amersham Pharmacia). Base-Calling was performed with PHRED, the cutoff Phred score was 20[17]. Original sequences were generated by removing vector and *E. coli* DNA sequences using Cross-match. Repeat sequences in these original ESTs were then masked by RepeatMasker program[18]. Using Blastn, sequences were further screened for contaminants such as bacterial chromosomal DNA, RNA, viral DNA, rRNA[19].

Table 1 Description of cDNA libraries

Library ID	Physiological stage	Breed	Raised place	Vector
MGP	7 d before parturition	DLY[a)]	Danmark	pTrueBlue
MGM	14 d after parturition	DLY	Danmark	pTrueBlue
MGA	7 d after weaning	DLY	Danmark	pTrueBlue
CGA	14 d after parturition	Erhualian	China	pBlueskript II SK (+)
MCP	1 h after parturition	LY[b)]	Danmark	pBlueskript II SK (+)

a) DLY, Duroc ♂×(Landrace♂×Yorkshire ♀) b) LY, Landrace ♂ ×Yorkshire♀

The sequences came from porcine rRNA and mitochondrial DNA are not contaminants, but these sequences are not necessary for this experiment. So, they were also screened by Blastn. Sequences less than 100 bp were filtered. Further screens for possible contaminants were conducted by Blastn searches of the GenBank Non-Redundant Nucleotide Sequences (nt), EST_human, EST_mouse, and EST_others databases (Release Apr, 2003). Sequences with >95% identity over at least 100 bp to other species were identified as contaminants. ESTs were then analyzed to identify chimeric inserts using the Perl scripts. Chimeric clones could be detected by back-to-back poly(A)+ tails or linker-to-linker sequences within ESTs. Blastn and Blastx searches for nt and GenBank Non-Redundant Protein database (nr) (Release Apr, 2003) were also performed to identify chimera. If the 3′ and 5′ sequences from one EST had matches (E values < 10^{-20}) to different gene, that EST was identified as chimera. After the above procedures, high-quality ESTs were generated for subsequent assembly and clustering.

1.3 ESTs assembly and annotation

High-quality ESTs were assembled into contigs using TGICL[20]. Default settings were used except minimum overlap was 40 bp and 95% identity. To assign annotation to the assembled ESTs (contigs), these sequences were searched against the nr and nt (E values < 10^{-15}) for homology comparison using Blastx and Blastn. Sequence with >90% identity over at least 100 bp to a known porcine gene was considered identical with that gene, while EST sequence with significant similarity (E values < 10^{-15}) to orthologous gene was assigned a putative function, the others were considered novel. All assembled sequences having the same annotation were further clustered into a unique gene.

1.4 Gene function classification, expression profiles construction and analysis

If the cluster was assigned an annotation, it proceeded to be classified into Gene Ontology (GO) categories (http://www.geneontology.org/)[21]. Using GO molecular function items as criteria, we constructed three (MGP, MGM and MGA) gene expression profiles. Gene expression difference was evaluated by Chi-squared test (χ^2)[22]. Comparison of gene expression level among libraries was performed with IDEG6[23]. The gene expression levels among three developmental stages in DLY sow and between Erhualian sow and DLY sow were also compared. The annotation and classification were based on data collected and extracted from the following public databases and files: GenBank (release Apr 11, 2003), gene_association.compugen.GenBank (version: 0.5.1 03/07/2002); component.ontology.txt (version: 2.663), process.ontology.txt (version: 2.663) and function. ontology.txt (version: 2.663). The last four files are downlooded from Gene Ontology Consortium (http://www.geneontology.org/).

2 Results

2.1 Generation and assembly of ESTs

Five non-normalized cDNA libraries were generated from different breeds and different developmental stages. The insert size of the cDNA libraries is 0.6-6kb and the average insert size is about 1 kb. Clones from each library were randomly picked and then sequenced partially from the 5′ end. A total of 30266 original ESTs were generated. After trimming of vector, low-quality, contaminated sequences, ribosomal RNA, mitochondria DNA sequences and chimeric sequences and filtering for minimum length (100 bp) from the original sequences, 28941 high-quality ESTs with 374 bp average length were generated. About 95% of the original ESTs are high-quality ESTs (Table 2), which indicates that all cDNA libraries in this experiment are reliable for the following analysis. The high-quality ESTs were then assembled into 7854 tentative consensus sequences (TCs), including 2212 contiguous sequences (contigs) and 5642 singletons.

2.2 EST annotation, functional classification and expression profile construction

After Blastn and Blastx search, a total of 5782 TCs were assigned function or putative function. The remaining 2072 sequences were considered novel ones. We further clustered all assembled sequences according to the results of the annotation. The clusters with the same annotation were considered to derive from the same unique gene. In total, 6857 clusters were

Table 2 ESTs pre-processing summary

Library ID	Original	Repeats (%)	Chimera (%)	< 100 bp (%)	High-quality (%)
MGP	7150	76 (1.06)	4 (0.06)	34 (0.48)	6832 (95.6)
MGM	5516	44 (0.80)	10 (0.18)	29 (0.53)	5354 (97.1)
MGA	6155	72 (1.17)	3 (0.05)	72 (1.17)	5899 (95.8)
MCP	7732	86 (1.11)	10 (0.13)	32 (0.41)	7289 (94.3)
CGA	3713	25 (0.67)	8 (0.22)	42 (1.13)	3567 (96.1)
Total	30266	303 (1.00)	35 (0.12)	209 (0.69)	28941 (95.6)

the same annotation were considered to derive from the same unique gene. In total, 6857 clusters were generated, 4785 (69.8%) of which have significant homology to known genes. These 4785 gene clusters roughly present the majority of function known genes expressed in porcine mammary gland during lactation (Supplementary table 1) (Electronic supplementary materials of this article are available at: http://www.wigs.zju.edu.cn/PorcineEST/MG.html). Distribution of ESTs among gene clusters was evaluated (Fig. 1). The majority (88%) of these clusters contain less than 10 ESTs, while remaining 17 clusters assembled by more than 100 ESTs. Most of these large clusters came from milk protein genes, indicating that milk protein genes are highly expressed in mammary gland at these developmental stages. The most abundant 10 genes are beta casein (*Csnb*), alpha-S1 casein (*CSN1S1*), beta-lactoglobulin, alpha-S2 casein (*CSN1S2*), histamine-releasing factor (*HRF*), kappa casein, ribosomal protein S27a, serum amyloid A-2, eukaryotic translation elongation factor 1 alpha 1 (*EEF1A1*) and alpha-lactalbumin respectively. According to the gene nominating system provided by Gene Ontology, the clusters with assigned annotations were classified into GO categories. A total of 2379 clusters were assigned the GO terms (Supplementary Table 2). Using the molecular functional items of GO as criteria, the gene expression profiles of three different cDNA libraries (MGP, MGM and MGA) were constructed and analyzed (Fig. 2). As a result, we found the diversity of porcine mammary gene expression is remarkable at different developmental stages. Genes of ribosome structural constituent are abundantly expressed in the porcine mammary gland 7 d before parturition. At the same time, some other categories are also abundantly expressed such as transcription regulator, translation regulator, nucleic acid binding, receptor binding and antioxidant. At the mid-lactation, metal ion binding genes are highly expressed. To the porcine mammary gland 7 d after weaning, some functional categories such as cell adhesion molecule, extracellular matrix (ECM) and structural constituent of cytoskeleton are abundantly expressed.

2.3 *Identification of genes differentially expressed in Erhualian and DLY sow mammary gland during lactation*

Using IDEG6, the genes expressed in Erhualian and DLY sow mammary gland 14 d after parturition were compared, and the differentially expressed genes were identified (p<0.01). Many genes were found differentially expressed, indicating the gene expression profile of mammary gland during lactation between Erhualian and DLY sow is quite different. A total of 64 genes

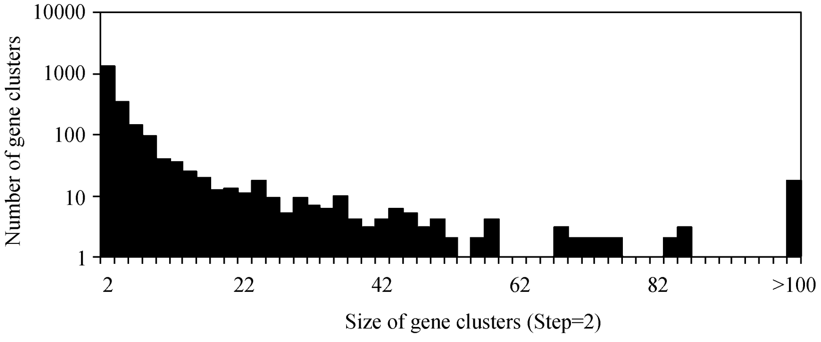

Fig. 1. Histogram of the size distribution of gene clusters. X-axis shows the number of ESTs contained in each gene clusters. Y-axis represents the number of gene clusters of a particular size in log scale.

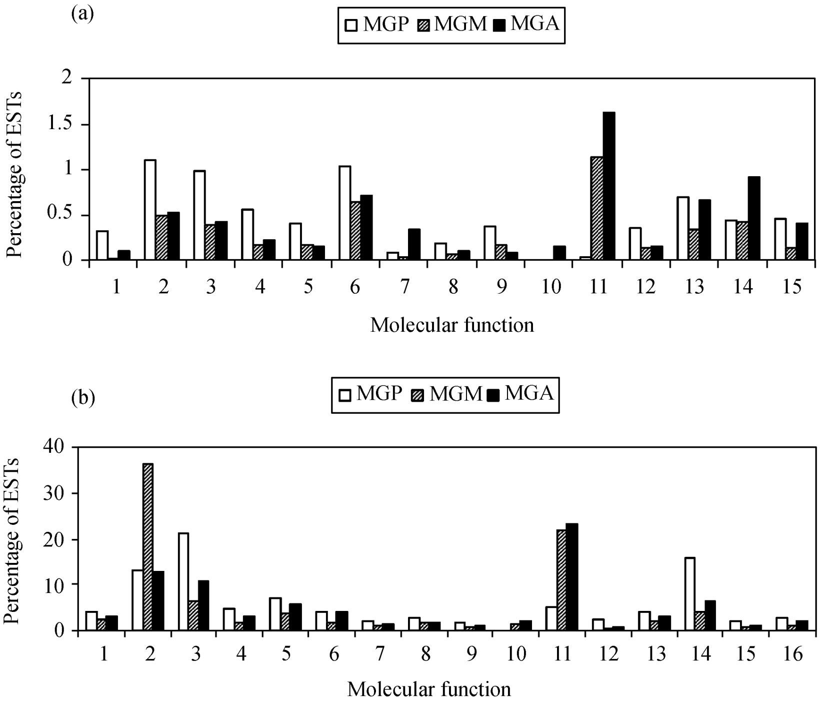

Fig. 2. EST abundance in each major GO molecular function category. X-axis represents the molecular function terms based on GO. Y-axis is the percentage of ESTs of each library. (a) 1, Antioxidant; 2, receptor binding; 3, glycosaminoglycan binding; 4, helicase; 5, lyase; 6, kinase; 7, cell adhesion molecule; 8, DNA repair protein; 9, electron transfer flavoprotein; 10, galactose binding lectin; 11, heterotrimeric G-protein GTPase; 12, high-density lipoprotein; 13, ECM structural constituent; 14, structural constituent of cytoskeleton; 15, structural constituent of muscle. (b) 1, Apoptosis inhibitor; 2, metal ion binding; 3, nucleic acid binding; 4, nucleotide binding; 5, protein binding; 6, hydrolase; 7, oxidoreductase; 8, transferase; 9, chaperone; 10, defense/Immunity protein; 11, enzyme inhibitor; 12, protein degradation tagging; 13, signal transducer; 14, constituent of ribosome; 15, transcription regulator; 16, translation regulator.

have higher expression level in Erhualian mammary gland than in DLY, and 36 of them are genes of ribosome structural constituent. Most of the rest genes are immunoglobulin genes and genes coding antimicrobial proteins (such as beta-lactoglobulin, complement component C3, whey acidic protein, ferritin, lactoferrin). On the contrary, only 16 genes are more abundant in DLY mammary at this stage. Most of them are casein protein genes and genes involved in the fatty acid synthesis pathway such as alpha-S1 casein, alpha-S2 casein, kappa casein, stearoyl-CoA desaturase and fatty-acid-Coenzyme A ligase (Table 3).

2.4 Identification of genes differentially expressed in different development stages in DLY sow mammary gland

Based on the expression profile comparison, we screened some genes showing significantly different expression abundance among the three developmental stages of DLY sow mammary gland (p<0.01). The porcine mammary gland 7 d before parturition had the most genes highly expressed, from which the genes of ribosome structural constituent were the most abundant, and some regulators controlling the transcription and translation processes were highly expressed such as transcription factor *Nrf2*, eukaryotic translation initiation factor and eukaryotic translation elongation factor (Table 4). In the porcine mammary gland 14 d after parturition, the genes coding milk proteins were extraordinarily highly expressed. More than 35% ESTs were derived from milk protein genes such as alpha-s1 casein, alpha-s2 casein and alpha-lactalbumin. Some genes involved in fatty acid synthesis pathway were also abundant in this stage such as stearoyl-CoA desaturase, fatty-acid-Coenzyme A ligase, fatty acid-binding protein and lipoprotein lipase (Table 5). While in the porcine mammary gland 7 d after weaning, the following genes showed significant abundance: cathepsin gene family, such as cathepsin C and cathepsin L; membrane protein genes, such as galectin 3 and odor-

Table 3 Breed-specific differentially expressed genes

CGA (%)[a)]	MGM (%)	Annotation	CGA (%)	MGM (%)	Annotation
Genes significantly more abundant in CGA than MGM[b)]			0.2	0	Ch4 and secrete domains of swine IgM
1.49	0	Ig lambda light chain VLJ region	0.2	0	Haptoglobin alpha 1S
2.38	1.14	Serum amyloid A-2 protein	0.14	0	Interferon, alpha-inducible protein
12.64	0.04	Beta-lactoglobulin	0.17	0	*Col1A1* gene for pro alpha 1(I) collagen
0.2	0	Complement component C3	0.17	0	Poly-Ig receptor
0.87	0	Ig alpha heavy chain	0.2	0	Proline-rich protein BstNI subfamily 2
2.07	0.02	Whey acidic protein	Genes significantly more abundant in MGM than CGA		
0.67	0.13	Fatty acid-binding protein, heart	0.53	11.08	Alpha-S2 casein
0.22	0.02	Similar to tumor-specific transplantation antigen P198 homolog p23	0.45	20.66	Alpha-S1 casein
0.45	0	Ferritin heavy chain	1.51	3.55	Kappa casein
0.17	0	Actin, gamma 1	0	0.3	60S ribosomal protein L22
0.25	0	Ferritin light subunit	0.25	2.28	Tumor protein, fortilin
0.22	0.04	Laminin receptor 1	0	0.35	60S ribosomal protein L9
0.34	0.02	G protein, beta polypeptide	0.03	0.37	Ribosomal protein S24
0.25	0	Cystatin C	0	0.57	Acyl-CoA synthetase, long-chain 3
0.17	0	*SLA-2/2*	0	0.22	Microsomal glutathione S-transferase 1
0.48	0	Lactoferrin	0	0.19	Phenylalanine hydroxylase
0.5	0	Peptidoglycan recognition protein	0	0.19	Small inducible cytokine A28 precursor
0.36	0	Collagen alpha 1(i) chain	0	0.26	Lipoprotein lipase
0.14	0	Collagen type 1, alpha-1	0	0.19	Insulin-like growth factor binding protein 7
0.42	0	Interferon-induced protein 1-8U	0	0.26	Acyl-CoA synthetase, long-chain 6
0.2	0	*MFGE8*	0	0.24	Similar to human cDNA FLJ38658 fis
0.14	0	Small EDRK-rich factor 2	0.03	0.28	Stearoyl-CoA desaturase

a) EST abundance in this cDNA library; b) 36 ribosomal protein genes are not shown.

Table 4 Genes significantly highly expressed in MGP[a)]

MGA (%)[b)]	MGM (%)	MGP (%)	Annotation	MGA (%)	MGM (%)	MGP (%)	Annotation
0.97	0.24	1.19	*EEF1A*	0.05	0	0.18	Tetraspanin 13
0	0	0.07	Protein kinase C	0	0	0.1	Scrapie responsive protein 1
0.02	0.04	0.15	Proteasome alpha 2 subunit	0	0.02	0.15	Thyroid receptor interactor
2.9	2.28	3.91	Tumor protein, fortilin	0	0.09	0.09	MacGAP protein (*MacGAP*)
0.22	0.07	0.4	*EEF1B2*	0.07	0	0.07	Signal peptidase 12 kDa subunit
0.02	0	0.83	Fc gamma (IgG) receptor IIa	0	0.04	0.09	*HNRPH1*
0.17	0.09	0.25	Collagen alpha 2(I) chain	0	0.02	0.09	Enthoprotin (*ENTH*)
0.37	0.24	0.51	*E25B*	0.02	0.02	0.12	Acylphosphatase, muscle type isozyme
0.36	0.13	0.42	Thymosin beta-4	0.03	0	0.09	Dendritic cell protein
0.15	0	0.16	Matrix Gla protein	0	0	0.12	*NUMB*
0.15	0.09	0.32	*NACA*	0	0.02	0.12	Myostatin; *GDF-8*
0.31	0.13	0.31	Epithelial glycoprotein (EGP)	0.03	0	0.09	*PRKAR1*
0	0.04	0.09	Laminin receptor 1	0.02	0.02	0.12	Lipoamide dehydrogenase
0.12	0.19	0.28	Small inducible cytokine A28	0.02	0	0.09	DNA topoisomerase II, beta isozyme
0.2	0.06	0.18	Diamine acetyltransferase	0	0	0.09	Methionine Aminopeptidase-2
0.07	0.04	0.22	*EIF4A2*	0.02	0	0.09	*E2IG5*
0.07	0.02	0.32	Peroxiredoxin 3	0.02	0	0.07	Protein phosphatase-1 gamma
0.19	0.04	0.19	Thioredoxin	0	0	0.06	RNA binding motif protein 3
0.08	0	0.28	Heat shock 10kDa protein 1	0	0	0.09	*RABGGTB*
0.05	0.04	0.2	*EIF3S3*	0.03	0	0.07	DEK oncogene (DNA binding)
0.08	0.04	0.16	Superoxide dismutase	0	0	0.06	Gamma-glutamyl hydrolase
0	0	0.06	*EIF4G2*	0	0	0.06	*HPRT1*
0.08	0.02	0.2	*ATP5F1*	0	0	0.07	*TIMM9*
0.07	0.04	0.2	*PTS*	0	0	0.07	Transcription factor *Nrf2*
0.03	0.07	0.15	B-cell translocation gene 3	0	0	0.07	*EIF2S1*
0.02	0.04	0.13	*HNRPA2B1*	0	0	0.06	*ATP5H*

a) 23 ribosomal protein genes and 28 unknown genes are not shown; b) EST abundance in this cDNA library.

Table 5 Genes significantly highly expressed in MGM[a)]

MGA (%)[b)]	MGM (%)	MGP (%)	Annotation	MGA (%)	MGM (%)	MGP (%)	Annotation
1.56	11.08	0.13	Alpha-s2 casein	0.02	0.15	0.04	Immunoglobulin J-chain
0.42	1.03	0.03	Alpha-lactalbumin	0.05	0.26	0.09	Lipoprotein lipase
5.95	20.66	7.22	Alpha-s1 casein	0.08	0.26	0.03	Acyl-CoA synthetase, long-chain 6
0.64	3.55	0.48	Kappa casein	0	0.28	0.03	Stearoyl-CoA desaturase
0	0.13	0.09	*FABP3*	0.15	0.17	0.01	*PAS-4*
0.29	0.57	0.13	Acyl-CoA synthetase, long-chain 3	0	0.07	0.01	Ovostatin
0.05	0.19	0.09	Phenylalanine hydroxylase	0	0.07	0.01	Ubiquitin-conjugating enzyme E2A

a) 4 unknown genes are not shown; b) EST abundance in this cDNA library.

Table 6 Genes significantly highly expressed in MGA[a)]

MGA (%)[b)]	MGM (%)	MGP (%)	Annotation	MGA (%)	MGM (%)	MGP (%)	Annotation
22.8	21.74	4.61	Beta casein	0.12	0	0.04	Epithelial membrane protein 1
1.63	1.14	0.03	Serum amyloid A-2 protein	0.08	0	0.04	Integrin beta-1 subunit (CD29)
0.2	0.02	0.09	Cathepsin S preproprotein	0.14	0	0	Tetraspanin TM4-A
0.81	0.3	0.2	Beta-2-microglobulin (Lactollin)	0.14	0	0.01	Intramembrane cleaving protease
0.31	0.13	0.09	Vimentin	0.15	0	0	Galectin 3
0.36	0.09	0.22	Tropomyosin 3	0.1	0	0.04	*PDK4*
0.12	0	0.04	Odorant binding protein	0.05	0	0	Decorin (Bone proteoglycan II)
0.2	0.06	0.18	Diamine acetyltransferase	0.1	0	0.03	*RAB2*
0.22	0.02	0.06	Thioredoxin interacting protein	0.08	0	0.04	F1-Atpase, alpha subunit
0.25	0.04	0.18	Nogo-A protein short form	0.12	0.02	0	Haptocorrin (R protein)
0.29	0.06	0.12	Cathepsin L	0.1	0	0.01	Proteoglycan 1, secretory granule
0.15	0.02	0.12	Apolipoprotein R	0.07	0	0	Haptoglobin alpha 1S
0.31	0.04	0.06	*CHRNA3*	0.1	0	0.01	Manganous superoxide dismutase
0.15	0.17	0.01	*PAS-4*	0.07	0.02	0	RNase PL3
0.25	0.02	0	Osteopontin (Bone sialoprotein 1)	0.08	0	0.03	CGI-86 protein
0.17	0.04	0.03	Cathepsin C	0.08	0	0	MD-2 protein
0.05	0	0	SLA-DQ alpha chain	0.07	0.02	0	Ubiquinone-binding protein
0.14	0.02	0.03	Phosphoprotein phosphatase	0.08	0.06	0	*TRAM*
0.15	0	0.04	*ARPC5*	0.07	0	0	Inorganic pyrophosphatase 2
0.15	0	0.03	Syntenin(*SDCBP*)	0.07	0	0	Lysozyme

a) 2 ribosomal protein genes and 9 unknown genes are not shown; b) EST abundance in this cDNA library.

ant binding protein; ECM structural constituent, such as osteopontin and haptoglobin; cell adhesion molecule, such as integrin; structural constituent of cytoskeleton, such as vimentin and tropomyosin 3 (Table 6). The expression of most milk protein genes reached peak after parturition and had a descending trend after weaning. However, *Csnb* gene had a different expression profile. The EST abundance of *Csnb* was 4.61% 7 d before parturition. It rose to 21.74% 14 d after parturition, and kept climbing after weaning, up to 22.8% 7 d after weaning.

3 Discussion

EST sequences are generated in a single pass sequencing, and they have a higher error rate than sequences verified by multiple sequencing runs, on the order of 3%, so the pre-processing of ESTs is especially important[24]. Per-base quality scores are sometimes available to mitigate the effects of sequence errors. Sequences can also be aligned to the genome and errors can be corrected based on the genomic sequences. In addition to sequencing errors, ESTs may contain various kinds of contaminations such as bacterial DNA/RNA, virus DNA/RNA, repeat sequence, linker and vector sequences[25]. These contaminated sequences can be generally identified by sequence similarity search such as RepeatMasker, Cross-match and BLASTN. Some ESTs may derive from chimeric clones. A chimera is a concatenation of two or more expressed sequences from different genomic loci. These kinds of artifacts can be identified by alignment to genome sequence. In this experiment, for lack of porcine genome sequences, these kinds of artifacts were difficult to identify. Some new methods were adopted to identify the chimera. ESTs with back-to-back poly(A)+ tails or linker-to-linker sequences in it were discarded as chimera. Sequence similarity searching was performed to known gene database. If the 3′ and 5′ sequences from one EST match to different genes, they were discarded, too.

In order to obtain more EST annotation information, 5′ directional sequencing strategy was chosen. However, because transcripts coming from the same gene may have different length in the process of cDNA library construction, some genes may have different consensus sequences after assembly. Otherwise, many mammalian genes are alternative spliced. This may also the ESTs coming from the same gene to be assembled to different consensus sequences. These phenomena may cause false positive in the TCs annotation and remarkably influence the construction of gene expression profile. So, we further clustered the TCs which have been assigned the annotation. The TCs with the same annotation were clustered to a gene cluster, with each gene cluster representing a unique gene.

Because the ESTs are generated by random cDNA clone sequencing, large-scale sequencing of non-normalized cDNA library can be used to evaluate the gene expression levels. However, EST abundance is an imperfect approximation of gene expression level, so we only look for genes for which the relative EST abundance is highly varied between the libraries, in which case the true gene expression levels are more likely to be different. Genes that are expressed at low levels or have smaller changes may also contribute to the phenotypic differences, but in this experiment, we did not analyze these genes because of the influences of cDNA library construction and the quantity of the sequenced ESTs.

It is significant to study the gene expression of mammary gland at different developmental stages, which will improve the study of some physiological processes such as organogenesis, cell differentiation and oncogenesis and so on. Development of the mammary gland occurs in defined stages that are connected to sexual development and reproduction. These are embryonic, prepubertal, pubertal, pregnancy, lactation, and involution. In newborn piglet a rudimentary system of small ducts is present which grows slowly until the onset of puberty when pronounced ductal growth occurs. Development of the ducts continues in cycling virgins leading to the formation of a ductal tree that fills the entire mammary fat pad. Extensive ductal branching and alveolar growth occurs during pregnancy and is largely completed at parturition. Terminal differentiation of the alveolar epithelium is completed at the end of gestation with the onset of milk secretion at parturition. After weaning the entire alveolar epithelium apoptosis, the gland is being remodeled. Within a few weeks the gland has the appearance of that of a mature virgin. The regulatory mechanisms of mammary epithelial cells' proliferation, differentiation and apoptosis are quite complicated. Although many factors which regulate many physiological processes of mammary gland have been found, such as some hormones and some genes which may play a role in cell to cell or cell to stroma interactions, many intimate mechanisms of the mammary gland development remain to be deciphered[1,2].

Whey acidic protein (*WAP*) and α-lactalbumin were abundantly expressed near the end of gestation, which can be used as a symbol of the mature of mammary epithelial cells[26]. According to the gene expression profiles of DLY sow mammary gland, they were weakly expressed a week before parturition, indicating that only a few mammary epithelial cells were differentiated at that time. In late gestation, gilt's mammary gland was prepared for lactating and its gene expression was active. The number of genes significantly highly expressed at that time was quite more than other two developmental stages. To prepare for largely synthesize protein at lactation; genes of ribosome structural constituent were abundantly expressed in this stage. At the same time, some proteins involved in transcriptional and translational regulation were also abundantly expressed, such as myostatin, thyroid receptor interactor and translation initiation factor.

In the middle of lactation, a great deal of milk proteins, fatty acids and lactose were synthesized in mammary epithelial cells. So, besides the milk protein genes, the genes involved in the pathways of the biosynthesis of these compounds were also highly expressed in this developmental stage. For example, the stearoyl-CoA desaturase and fatty-acid-Coenzyme A ligase which take charge of the fatty acid biosynthesis; the phenylalanine hydroxylase which was one of the enzymes in the tyrosine metabolism; the α-lactalbumin which participated in the biosynthesis of lactose. Many of primary milk proteins were in the peak of expression in this stage, but the expression profile of β-casein gene differed from other primary milk proteins. The expression level of other proteins was decreased significantly after weaning, but the β-casein gene was highly expressed at this stage, even higher than in the middle of lactation. This indicated that the gene regulation system of β-casein was different from other primary milk proteins.

The mammary gland underwent involution after cessation of milking or weaning of the young. After involution, the structure of the involuting gland became almost identical to that of the resting virgin gland. Mammary gland involution went through two distinct stages. In the first stage, alveolar cells underwent programmed cell death (PCD), but there was no remodeling of the lobular-alveolar structure. During the second stage, the lobular-alveolar structure of the gland was obliterated as proteinases degraded basement membrane and extracellular matrix (ECM)[27]. Based on the expression profile, we can infer that the gilt's mammary gland a week after weaning was in the late involution. At that time, lysozyme, some kinds of hydrolases and cathepsin were abundantly expressed in mammary gland. It indicated that during the second stage of involution, the lysosomes might be involved in degrading the apoptotic cell fragments as well as the milk proteins. As the lobular-alveolar structure was destroyed gradually, on the contrary, some membrane protein genes, ECM structural constituent and cell adhesion molecule genes were abundantly expressed. It would improve the rebuilding of the new basement membrane and ECM.

We found that the gene of serum amyloid A-2 protein and haptoglobin gene were extremely highly expressed 7 d after weaning. At the same time, some immunoreaction related genes such as SLA class II histocompatibility antigen and beta-2-microglobulin were also found. It suggested that the acute-phase response and some other immunoreactions were involved in the process of mammary gland involution. This result was consistent with Clarkson and Stein's previous study in the process of mouse mammary gland involution[28,29].

The previous researches suggested that the lactation performance of sows is closely correlated with its reproductive performance. Furthermore, the production and composition of milk may affect sows' reproductive performance. Porcine milk contains many kinds of bioactive molecules, which may be important for piglets, affecting its development and immunity. Chinese Meishan sow and US Yokshire sow have different reproductive performance. Based on the comparisons between these two breeds, Zou and his colleagues[15] found the colostrum and milk composition of them are significantly different. Scientists have reported that the polymorphism of HMWP in sow milk is correlated with sow's reproductive performance[16]. However, the previous researches mainly focused on analysis of a single protein or whole milk composition, and no reports derived from comparison of the diversity of gene expression were found. So it is important to study the physiological and biochemical characteristic of porcine galactopoiesis from gene expression profile level. The Chinese Erhualian breed is known for its ability to bear and raise large litters. In addition, Erhualian pigs were reported to have an earlier age of maturity than US and European breeds and have a relatively high body fat and low lean body mass[30]. DLY breed is the very reverse of Erhualian breed. Through the comparison of these two breeds' gene expression profile of mammary gland in the middle of lactation, we screened 64 genes highly expressed in Erhualian breed. These genes were mainly whey protein genes such as immunoglobulin, complement component C3, whey acid protein and lactoferrin. The primary functions of these proteins were providing immune protection, and preventing pathogenic bacteria and viruses from infecting piglets. The Ehualian pigs were raised in the region of Taihu Lake, an area with high average air temperature and humidity, which was suitable for microorganism growth. So we can infer that the gene expression characteristic of the mammary gland of Erhualian breed in lactation was not only consist with its good reproductive performance, but also had a little correlation with its environmental adaptability. A total of 16 genes were found highly expressed in the mammary gland of DLY breed. Most of them were casein protein genes and the genes involved in the adipose biosynthesis pathways. We guess that the gene expression characteristic of the mammary gland of DLY breed might contribute to the fast growth rate of DLY piglets, but further biological experiments are needed to verify this hypothesis.

Oligonucleotide DNA chip, cDNA microarray, serial analysis of gene expression (SAGE) and large-scale ESTs sequencing all are efficient methods for studying gene expression profile[31,32]. Although the method of large-scale ESTs sequencing is not sensitive enough for genes expressed at a low level and cannot analyze its expression level, we can get sequence resources and identify more novel genes through this method. A total of 2072 novel ESTs were identified in this work (data not shown). It will be helpful for further new gene cloning.

From this experiment, using large-scale EST sequencing as strategy, we have constructed DLY sows mammary gland gene expression profiles in different developmental stages, and found some developmental-stage-specific genes and some novel genes (its function needs to be verified by experiments). The results will help us understand the molecular mechanism of porcine mammary gland development. Furthermore, the comparison of mammary gland gene expression profile between two porcine breeds provided new clues for the study of the relationship between sow's lactation and reproductive performance. This work will improve the development and research of swine germplasm resources. It will also provide efficient criteria for porcine excellent breed selection, as well as for the breeding and management of these two porcine breeds.

Acknowledgements We thank Danish Institute of Agricultural Sciences for providing the cDNA libraries of Danish porcine breeds in Sino-Danish Pig Genome Project. This work was supported by the National High-Tech Research and Development Program (863) (Grant No. 2002AA229061) and the Knowledge Innovation Program Important Project of Chinese Academy of Sciences (Grant No. KSCX1-01).

References

1. Rosen, J. M., Wyszomierski, S. L., Hadsell, D., Regulation of milk protein gene expression, Annu. Rev. Nutr., 1999, 19: 407－436.
2. Jaggi, R., Marti, A., Guo, K. *et al.*, Regulation of a physiological apoptosis: Mouse mammary involution, J. Dairy Sci., 1996, 79: 1074－1084.
3. Croft, L., Schandorff, S., Clark, F. *et al.*, ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome, Nat. Genet., 2000, 24: 340－341.
4. Modrek, B., Resch, A., Grasso, C. *et al.*, Genome-wide detection of alternative splicing in expressed sequences of human genes, Nucleic Acids Res., 2001, 29: 2850－2859.
5. Adams, M. D., Kerlavage, A. R., Fleischmann, R. D. *et al.*, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence, Nature, 1995, 377: 3－174.
6. Papadopoulos, N., Nicolaides, N. C., Wei, Y. F. *et al.*, Mutation of a mutL homolog in hereditary colon cancer, Science, 1994, 263: 1625－1629.
7. Jiang, F. B., Chen, C., Deng, Y. J. *et al.*, Analysis of porcine MHC expression profile., Chinese Science Bulletin, 2005, 50(9): 880－890
8. Zhong, B. X., Yu, Y. P., Xu. Y. S. *et al.*, Analysis of ESTs and gene expression patterns of the poste-rior silkgland in the fifth instar larvae of silkworm, *Bom-byx mori* L., Science in China Ser. C, 2005, 48(1): 25－33.
9. Ryo, A., Kondoh, N., Wakatsuki, T. *et al.*, A method for analyzing the qualitative and quantitative aspects of gene expression: a transcriptional profile revealed for HeLa cells, Nucleic Acids Res., 1998, 26: 2586－2592.
10. Wang, L. L., Ma, L., Leng, W. C. *et al.*, Analysis of part of the *Trichophyton rubrum* ESTs, Science in China, Ser. C, 2004, 47(5): 389—395.
11. Mao, M., Fu, G., Wu, J. S. *et al.*, Identification of genes expressed in human CD34(+) hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning, Proc. Natl. Acad. Sci. USA, 1998, 95: 8175－8180.
12. Yu, Y., Zhang, C., Zhou, G. *et al.*, Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs, Genome Res., 2001, 11: 1392－1403.
13. http://ncbi.nlm.nih.gov/UniGene.
14. Wheeler, M. B., Production of transgenic livestock: promise fulfilled, J. Anim. Sci., 2003, 81(Suppl 3): 32－37.
15. Zou, S. X., Mclaren, D. G., Hurley, W. L., Pig colostrum and milk composition: Comparisons between Chinese meishan and US breeds, Livestock Production Sci., 1992, 30: 115－127.
16. Qin, Y. D., Xu, Y. X., Zou, S. X. *et al.*, The polymorphism of high molecular weight protein in sow milk and its relatiomship with reproductive performance, Acta Veterinaria et Zootechnica Sinica (in Chinese), 2002, 33(5): 429－432.
17. Ewing, B., Hillier, L., Wendl, M. C. *et al.*, Base-calling of automated sequencer traces using phred. II. Error Probabilities, Ge-

nome Res., 1998, 8: 175－185.

18. http://www.genome.washington.edu/UWGC.
19. Altschul, S., Madden, T., Schaffer, A. *et al.*, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res., 1997, 25: 3389－3402.
20. Pertea, G., Huang, X., Liang, F. *et al.*, TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets, Bioinformatics, 2003, 19: 651－652.
21. Consortium TGO., Creating the gene ontology resource: Design and implementation, Genome Res., 2001, 11: 1425－1433.
22. Romualdi, C., Bortoluzzi, S., Danieli, G. A., Detecting differentially expressed genes in multiple tag sampling experiments: Comparative evaluation of statistical tests, Hum. Mol. Genet., 2001, 10: 2133－2141.
23. Romualdi, C., Bortoluzzi, S., D'Alessi, F. *et al.*, IDEG6: A web tool for detection of differentially expressed genes in multiple tag sampling experiments, Physiol. Genomics, 2003, 12: 159－162.
24. Boguski, M. S., Lowe, T. M., Tolstoshev, C. M., dbEST——Database for "expressed sequence tags", Nat. Genet., 1993, 4: 332—333.
25. Hillier, L. D., Lennon, G.., Becker, M. *et al.*, Generation and analysis of 280000 human expressed sequence tags, Genome Res., 1996, 6: 807－828.
26. Robinson, G. W., McKnight, R. A., Smith, G. H. *et al.*, Mammary epithelial cells undergo secretory differentiation in cycling virgins but require pregnancy for the establishment of terminal differentiation, Development, 1995, 121: 2079－2090.
27. Li, M., Liu, X., Robinson, G.. *et al.*, Mammary-derived signals activate programmed cell death during the first stage of mammary gland involution, Proc. Natl. Acad. Sci. USA, 1997, 94: 3425－3430.
28. Clarkson, R. W., Wayland, M. T., Lee, J. *et al.*, Gene expression profiling of mammary gland development reveals putative roles for death receptors and immune mediators in post-lactational regression, Breast Cancer Res., 2004, 6: R92－R109.
29. Stein, T., Morris, J. S., Davies, C. R. *et al.*, Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving LBP, CD14 and STAT3, Breast Cancer Res., 2004, 6: R75－R91.
30. Chinese Livestock Records Compiling Committee, ed., Chinese Pig Breeds Record (in Chinese), Shanghai: Publish House of Science and Tech-nology of Shanghai, 1986.
31. Service, R. F., DNA chips survey an entire genome, Science, 1998, 281: 1122.
32. Powell, J., SAGE. The serial analysis of gene expression, Methods Mol. Biol., 2000, 99: 297－319.