In silico Discovery of Genes Expressed in Liver, Kidney, Spleen and Small Intestine of Pigs

Zengxiang Pan, Honglin Liu*, Jie Chen, Dan Xu, Zhihua Jiang[1] and Zhuang Xie

College of Animal Science and Technology, Nanjing Agricultural University, Nanjing 210095, P. R. China

ABSTRACT : An *in silico* approach was developed to survey the genes expressed in four internal organs of pig: liver, kidney, spleen and small intestine. The major procedures of the approach included: (1) BLAST searching against GenBank "est_others" database using human cDNA sequences as queries to screen the porcine orthologous expressed sequence tags (ESTs), (2) classifying the porcine ESTs records by resources according to certain criteria and (3) analyzing data for ESTs specifically expressed in each organ. In order to do so, four Java programs were developed. Based on the ESTs available in the GenBank database, it was found that there were at least 2,100 genes expressed in these four organs, including 128 in the liver, 81 in the kidney, 780 in the spleen, and 1,423 in the small intestine respectively (a few genes co-expressed in these tissues). Gene expression patterns, such as co-expressed genes, preferentially expressed genes and basic active genes were also compared and characterized among these organs. This study provides a comprehensive model on how to use the bioinformatics approach and Genbank databases to facilitate the discovery of new genes in livestock species. ***(Asian-Aust. J. Anim. Sci. 2005. Vol 18, No. 2 : 170-178)***

Key Words : *In silico*, Internal Organs, Gene Expression Profiling, Ests, Pig

INTRODUCTION

With the advances in molecular biology technology and the genomic projects in various organisms, a number of bioinformatic resources have rapidly accumulated worldwide. So an innovative technology is coming of age, i.e., *in silico* study. Based on experimental data from developing organisms, it uses the power of computational methods to explore the properties of real gene networks (Dearden et al., 2000). It can significantly impact the laboratory study process, based on the full use of genomic data and resources, and it would provide candidate information for related studies more easily. By mining comprehensive sequence data and integrating diverse information, it is possible to characterize a locus or gene, *in silico*, to a considerable level of detail (Barnes, 2002). *In silico* studies have become increasingly important, because the various sequencing projects are leading to an accumulating amount of raw sequence data, and the database is expanding rapidly (Pruess et al., 2003), especially for expression sequence tags (ESTs). Currently about 5,000,000 human expressed sequence tags (ESTs) are available through GenBank. As for pigs, more than 100,000 ESTs have been released to the public domain. Meanwhile, with the successful progress of the human genome project and the release of draft sequences, it was demonstrated that there are approximately 34,000 genes that span a total of 3,030 Mb of sequences in the human genome. No doubt, these resources will provide the base for *in silico* studies, and so bring forth the revolutionary changes in identifying and cloning new genes. Now ESTs have been utilized as useful tools for analyzing comparative genome organization in livestock species, further enabling accurate transfer of valuable information from one species to another. (Adams et al., 1991; Rohrer et al., 2002; Jiang et al., 2003; Lee et al., 2003).

To detect the gene expression profiles in the major internal organs of the pig, we have developed an *in silico* approach for mining the NCBI (The National Center for Biotechnology Information) porcine EST sequence resources using the human genome cDNA sequences as references. The information reported in this paper should be useful for researchers in the field to analyze genes and proteins of their own interest, and to study comparative and functional genomics.

MATERIALS AND METHODS

Database resources

NCBI, being one of the largest biological information systems available in the world, is our major data resource library for this study. Here, about 34,000 human gene cDNA sequences were collected. Of the sequences identified, 33,308 are coding genes distributed in the following Homo sapiens chromosomes (HSC): 3,000 on HSC1, 2,461 on HSC2, 1,872 on HSC3, 1,582 on HSC4, 1,766 on HSC5, 1,847 on HSC6, 1,766 on HSC7, 1,395 on HSC8, 1,368 on HSC9, 1,425 on HSC10, 1,904 on HSC11, 1,557 on HSC12, 777 on HSC13, 1,057 on HSC14, 1,152 on HSC15, 1,258 on HSC16, 1,439 on HSC17, 640 on HSC18, 1,644 on HSC19, 828 on HSC20, 386 on HSC21,

* Corresponding Author: Honglin Liu. Tel: +86-25-84395278, Fax: +86-25-84395314, E-mail: liuhonglin@263.net

[1] Department of Animal Sciences, Washington State University, Pullman, USA.

Received April 9, 2004; Accepted August 23, 2004

Table 1. ESTs libraries of the porcine major internal organs (retrived in NCBI)

Tissue name	Library name of EST
Liver	① CSEQFXL11 pig liver; ② Porcine liver cDNA; ③ heterologous RT-PCR; ④ Subtracted Porcine hepatocyte; ⑤pig pUC18 Library
Kidney	CSEQFXL12 pig kidney
Spleen	① Porcine Spleen cDNA library, Mini, Cot 30; ② Infected Porcine Spleen cDNA library; ③ Pig Spleen lambda gt 11 Library (Clontech Cat # PL1006b)
Small Intestine	① Porcine small intestine cDNA library; ② directionally cloned cDNA in XL1-blue MRF'

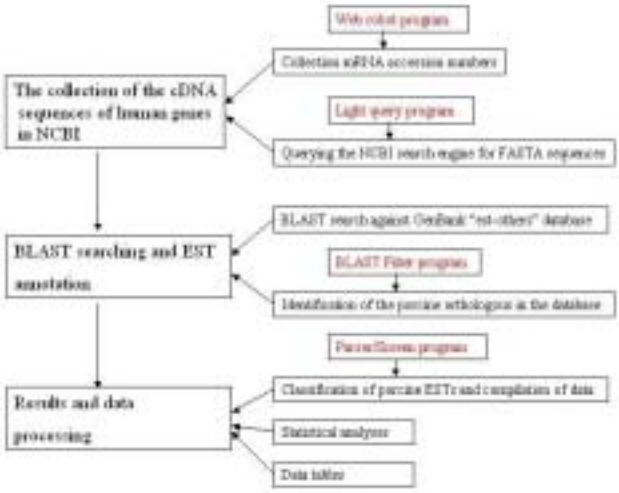


Figure 1. Flowchart of *in silico* study.

675 on HSC22, 1,315 on HSCX and 195 on HSCY. As for the ESTs sequence resources in the major internal organs of the pig, currently there are a certain amount of cDNA sequences derived from individual or pooled cDNA libraries of liver, kidney, spleen, small intestine and so on. According to the different methods to acquire the cDNA libraries or the different vector used constructing the libraries, a few types of ESTs sequences expressed in the major organs of the pig were found in the GenBank. By a standalone BLAST search, 11 porcine ESTs libraries were retrieved, five libraries were constructed from liver tissue, one from kidney, three from spleen, and two from small intestine respectively (Table 1). All these ESTs records are available in the "est-others" database at NCBI.

Sequences analysis

In this study, we built up an *in silico* approach to identify the homologies between cDNA sequences of human genes and porcine ESTs sequences by utilizing the present software tools and developing Java programs. First, a standalone BLAST searching program was installed and used to perform BLAST searches to annotate porcine ESTs expressed in the major internal organs. The major procedures included firstly: (1) BLAST searching against GenBank "est_others" database using human cDNA sequences as queries, (2) identifying and screening the porcine orthologous ESTs, (3) classifying the porcine ESTs records by resources according to certain criteria, (4) collecting and arranging data for ESTs specifically expressed in the major internal organs. Secondly, four Java programs were developed for sequences collection, sequences alignment and data processing, which included: (1) Web robot, used for collecting mRNA accession numbers by detecting all human chromosome information web pages, (2) Light query, used for automatically querying the NCBI search engine (http://www.ncbi.nlm.nih.gov/entrez/) for the FASTA sequences using mRNA accession numbers as search fields, (3) BlastFilter, used to filter out the BLAST matches that do not represent the porcine ESTs or do not meet the requirements with sequence identity by higher than 80% within a continuous alignment of sequences longer than 100 bp, (4) Parser/Screen, used to process Plain text in BLAST research result pages so as to analyze more conveniently and easily, and ultimately form the Microsoft Excel format tables.

Statistical analysis

For the homologous match records between human and pig, we performed statistical analyses after classification and arrangement, which included:

(1) counting the human genes and their homologous porcine ESTs expressed in the liver, kidney, spleen, and small intestine according to the human chromosome number, and identifying the extremely significant match records in the respective tissues (represented by the top 10).

(2) comparing and detecting gene expression differences among liver, kidney, spleen and small intestine by paired-samples T test according to the human chromosome number.

(3) investigating the basic active genes (BAGs) expressed in the major internal organs of pig by setting certain criteria (records were arranged by the number of the porcine EST hit(s) per human gene, and then the cut-off criteria to choose the basic active genes were selected at 3 EST hits (about 0.1% of their total hits) per gene.

To understand the procedure more easily, a flowchart was presented in Figure 1.

RESULTS

Gene expression profiles

In this study, sequence similarities were identified by BLAST using 33,308 human coding genes as references. It

Table 2. Genes expressed or co- expressed in the major internal organs of pig

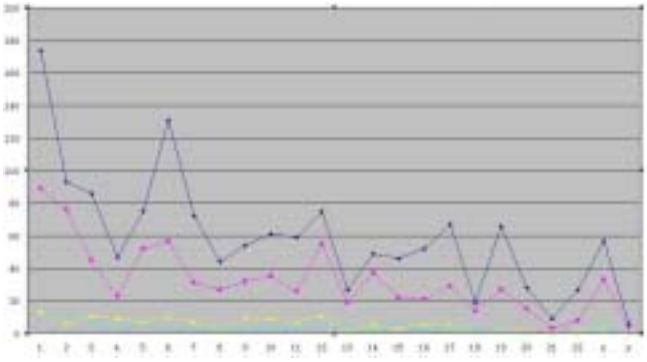
Tissues	Four tissues	Three tissues				Two tissues						Single tissue			
	SILK	SIL	SIK	SLK	ILK	SI	SL	SK	IL	IK	LK	S	I	L	K
Number of genes	1	2	3	1	2	251	4	9	18	9	2	508	1,137	98	55
Total (2100)	1	8				293						1,798			

S, I, L, K represents spleen, small intestine, liver and kidney respectively.
SILK represents the four tissues which have the same gene expressed, and so on.

Table 3. The highly homologous records between human genes and ESTs sequences in the major internal organs of pig (top 10)

Small intestine	Gene NO.	NM_153334	NM_003093	NM_015178	NM_002948	NM_000975	NM_006196	NM_018120	NM_000484	NM_001743	NM_001892
	Est NO.	AJ241168	AJ241172	Z84099	Z84172	F14532	F14489	AJ241144	Z84022	F14583	F22872
	Score	979	771	743	741	702	680	658	658	644	636
	E value	0	0	0	0	0	0	0	0	0	e-179
Spleen	Gene NO.	NM_007217	XM_018473	NM_015646	NM_006924	NM_001731	NM_003347	NM_004725	NM_002107	NM_014263	NM_002137
	Est NO.	BI336540	BI326846	BI326846	BI326930	BI326871	BI326900	BI119118	BI326822	BI326853	AA056904
	Score	896	841	841	825	793	773	745	743	735	733
	E value	0	0	0	0	0	0	0	0	0	0
Liver	Gene NO.	NM_000546	NM_001457	NM_002520	XM_170477	NM_004236	NM_006007	NM_006265	NM_021957	XM_084770	XM_059067
	Est NO.	BG695753	BG695798	BM659212	BM659228	BM659103	BM659189	BM659201	BM659188	BM659212	BM659212
	Score	1,600	1,031	609	599	595	593	573	567	557	533
	E value	0	0	e-171	e-168	e-167	e-166	e-160	e-158	e-156	e-149
Kidney	Gene NO.	NM_015640	NM_001904	XM_001389	NM_006141	NM_001686	XM_170477	NM_012207	NM_024835	NM_001271	NM_006292
	Est NO.	BM658979	BM658994	BM658979	BM658945	BM675757	BM658973	BM659073	BM675759	BM659065	BM658997
	Score	706	654	613	591	589	587	557	533	525	505
	E value	0	0	e-172	e-166	e-165	e-165	e-155	e-148	e-1545	e-140

The gene NO. and the Est NO. represent their accession number in GenBank; the score and the E value refer to the BLAST alignment results.



The human chromosome number

Figure 2. Genes expressed in the major internal of pig. The blue, red, yellow and green lines represent the small intestine, the spleen, the liver and the kidney respectively. The ordinate represents the numbers of genes expressed and the abscissa the human chromosome number.

was found that there were at least 2,100 genes (about 6.3% of the total number of human genes) which had high similarities to the ESTs sequences expressed in the major internal organs of pig. Of these 2,100 genes, 128 genes were expressed in liver (about 6.10% of the total 2,100 genes), 81 in kidney (3.86%), 780 in spleen (37.14%) and 1,423 in small intestine (67.76%) respectively. There were genes co-expressed in these tissues (Table 2). Among these co-expressed genes, one was expressed in the four tissues (NM_005410), 8 in three tissues (NM_000064, NM_021999, NM_000035, XM_065611, XM_171283, XM_087349, NM_000282, NM_006667), 293 in two tissues, and the remaining 1,798 genes were expressed only in one of the four tissues. It was found that there were 70,896 high homologous ESTs matches in the "est-others" database by BLAST search. Among them, 1,400 expressed in the major internal organs of pig, 139 in liver, 66 in kidney, 428 in spleen, and 767 in small intestine respectively. On the other hand, the total hits were 2,819 in the four tissues (many ESTs hits appeared more than once), 185 in liver, 83 in kidney, 967 in spleen, 1,584 in small intestine respectively. Therefore, the average ESTs match per human gene was 1.34 (2,819/2,100) in the major internal organs of pig. In addition, the extremely significant matches to human genes were counted for the four tissues respectively, represented by the top 10 records (Table 3).

Gene expression comparison in the major internal organs of pig

The data were classified and summarized in each organ of the pig according to the human chromosome number in the major internal organs of the pig. It was found that there were significant differences in gene expression among the four tissues by paired-samples T test. Among them, the difference between liver and kidney reached the 0.05 significance level ($p<0.05$, $t=2.806$), and all the remainder reached 0.01 significance level ($p<0.01$, $t=7.246$ between liver and spleen, 7.769 between liver and small intestine, 7.492 between kidney and spleen, 7.740 between kidney and small intestine, and 6.647 between spleen and small intestine respectively). The gene expression profiles were shown in Figure 2.

Statistical analysis for basic active genes (BAGs)

As mentioned above, the number of porcine EST hit(s)

Table 4. The basic active genes (BAGs) expressed in the major internal organs of pig

Tissues	Gene symbol	Accession NO.	Number of BAGs	Gene symbol	Accession NO.	Number of BAGs
Liver	APOH	NM_000042	3	AKR1D1	NM_005989	3
	CYP2C18	NM_000772	3	CYP2E1	NM_000773	4
	HP	NM_005143	3	ALB	NM_000477	35
Kidney	-	-	0			
Spleen	FTL	NM_000146	3	LOC165061	XM_092346	3
	HBB	NM_000518	3	LOC169986	XM_093029	3
	C7	NM_000587	3	LOC170278	XM_093223	3
	RPS19	NM_001022	3	LOC167204	XM_094341	3
	H3F3A	NM_002107	3	LOC204275	XM_115290	3
	HLA-DQA1	NM_002122	3	LOC202144	XM_116349	3
	NPM1	NM_002520	3	LOC257429	XM_171133	3
	HLA-DQA2	NM_020056	3	LOC255637	XM_172564	3
	NFKBIA	NM_020529	3	LOC253947	XM_172666	3
	LOC148356	XM_059067	3	LOC254835	XM_173056	3
	LOC132070	XM_067649	3	CAP1	NM_006367	4
	LOC136875	XM_070114	3	C1QG	XM_031238	4
	LOC144202	XM_084770	3	LOC152059	XM_087372	4
	LOC151561	XM_087240	3	LOC219398	XM_166705	4
	LOC148370	XM_088808	3	HLA-DRB3	NM_022555	5
	LOC142932	XM_089560	3	HLA-DPB1	XM_165807	5
	LOC145799	XM_090998	3	LOC221518	XM_165806	6
	LOC164789	XM_092163	3	WBP4	NM_007187	12
Small intestine	RPS16	NM_001020	3	LOC131123	XM_067215	3
	GBP1	NM_002053	3	LOC133702	XM_068486	3
	PTMA	NM_002823	3	LOC253823	XM_172714	3
	TEGT	NM_003217	3	RCP9	NM_014478	4
	DMBT1	NM_004406	3	LOC121994	XM_058602	4
	ATP5H	NM_006356	3	LOC122818	XM_063371	4
	LOC145123	XM_041473	3	LOC133363	XM_068306	4
	LOC129160	XM_066330	3	LOC167160	XM_094318	4
	LOC139057	XM_066449	3	LOC255727	XM_171365	4

per human gene was used as an indicator for gene expression activity analysis by setting a certain criterion. By the criterion 3 EST hits per gene (approximately 2,819×0.001), it was found there were 60 BAGs expressed in the major internal organs of the pig, including 6 in the liver, 0 in the kidney, 36 in the spleen, and 18 in the small intestine respectively (Table 4).

DISCUSSION

Application of database resources: *in silico* study

Database resources are bases for *in silico* study. The release of the first draft of the human genome, often referred to as the "golden path", promises to provide a wealth of data to develop diverse biological studies (Barnes, 2002). On the other hand, a vast amount of EST sequences have been accumulated in the public domain from different organisms. It is expected that more of these types of sequence resources will be released, as many sequencing projects in different species are still under way (Jiang et al., 2003). Currently, a large number of single-pass ESTs sequence data of cDNA clones from numerous swine tissues have been collected by many laboratories. More than 100,000 porcine EST sequences now exist in public databases and they are assembled into 47,540 unique sequences (Rohrer et al., 2002). Doubtlessly, these sequences will provide information and assistance for the discovery of novel genes, identification of homologous genes, analysis of alternative splicing, chromosomal localization of gene, and detection of polymorphisms (Pandey and Lewitter, 1999). Faced with so vast and chaotic data, laboratory scientists will be at wit's end as it is notoriously slow, labor intensive and expensive to utilize them. Yet, it is clear that *in silico* study, with its battery of computer simulation tools and massive databases, can use resources efficiently. More importantly, it has gathered momentum as, worldwide, scientists have united in a common quest to sequence, store and analyze complete genomes. It provides radical new ways of looking at problems in biology; it offers new ways of incorporating huge amounts of disparate but relevant data into a discovery pipeline and offers the hope of integrated solutions. In addition, a number of excellent, freely available, continuously improving and increasing web sites can give us kits-like tools similar in the molecular biology and make research more efficient (Attwood and Miller, 2002; Farber

and Medrano, 2003; Lynn et al., 2003). Realizing these, we developed an *in silico* study here. In this study, retrieving homologous sequences between human genes and porcine ESTs sequences by BLAST, allowed us to annotate the porcine ESTs derived from the major internal organs by a fresh gene oriented approach which characterized genes expressed in some tissues from genes to ESTs. This approach provided tools and information for rapid discovery of genes of interest. In addition, this methodology can be utilized by researchers engaged in detecting of gene expression in other species. In addition, by combining other bioinformatics analysis, we can understand the nucleotide sequences, genes and biological events of interest more profoundly, for instance, we can further analyze or validate our in silico study results by exploiting the Gene Indices database in TIGR and the Unigenes database in NCBI. In fact, this will be our follow up work.

In the genome sequences of two species, usually, the functional parts will be conserved and over time the nonfunctional regions of the sequences will slowly diverge. So if you align two syntenic regions of the genome, the conserved regions will highlight things like the exons and, possibly, the regulatory binding sites. Thus, you can predict genes without the need of protein or cDNA sequences already in the database (Birney et al., 2002). ESTs, as short single-pass DNA sequences obtained from either end of cDNA clones, are powerful resources for this syntenic study. They have been used primarily for the identification of new genes, physical map construction, identification of disease-causing mutations and other polymorphisms, and for the annotation of genomic sequences. They can be used to determine expression profiles of genes, compare expression patterns in different tissues or disease states etc. For the exponentially growing ESTs data, various information strategies have been devised to query EST databases. Since most of the analysis was performed with a computer, the term "*in silico*" study had been coined (Gill and Sanseau, 2000; Skrabanek et al., 2001). These strategies already have had a great impact on biological research, and the approach is providing researchers with new tools to address complex molecular events. Obviously they are contributing to the current shift in research directions from the classical emphasis on individual genes and molecules to the investigation of patterns of gene expression and the elucidation of comprehensive functional networks (Strausberg et al., 1999; Martin et al., 2000). So for database miners or users, there are three challenges to confront: (1) making full and efficient use of the present software tools resources on-line or off-line in sequence analysis, genetic analysis, and data processing etc. (2) developing a certain program for large-scale *in silico* study, e.g., four programs ‘Web robot’, ‘Light query’, ‘BlastFilter’, and ‘Parser/Screen’ were developed by Java in this study. In addition, we could exploit the present software further, e.g., Farber et al. (2003) wrote a SAS program, SAS-UCD1, to transform blastcl3 text files to a Microsoft Excel table format that was easily utilized. (3) improving and optimizing present software and algorithms, which is the greatest challenge for *in silico* study, and in fact, it is under way all the time.

ESTs have provided insights into transcribed genes in a variety of organisms and are widely used for gene discovery and expression analysis. However, identifying encoded genes from ESTs presents a number of challenges, e.g., a large numbers of redundancy, relatively frequent chimaerism, and a moderate rate of vector and adapter contamination (Pertea et al., 2003). So inevitably, they can lead to some irrelevant or incorrect conclusions *in silico* studies. On the other hand, the first draft of the human genome, as the ‘golden path’, may also bring forth problems if used without proper quality checks. In other words, it is important to keep in mind some caveats regarding *in silico* experiments. Researchers must be cautious of *in silico* results and perform necessary biological experiments to confirm such predictions and conclusions.

Genes expressed in the major internal organs of pig

Genes co-expressed in tissues : By performing a BLAST search to mine the porcine dbESTs, it was found that there were at least 2,100 human genes expressed in the major internal organs of the pig. As mentioned above, there was 1 gene SEPP1 co-expressed in the four tissues. This gene encodes Selenoprotein P which is an extracellular glycoprotein known to occur in mammalian plasma and is the only selenoprotein known to contain multiple selenocysteine residues. This protein is ubiquitously expressed in mammalian tissues and functions as an oxidant defense in the extracellular space and in the transport of selenium (Mostert, 2000; Saito and Takahashi, 2002); there were 2 genes co-expressed in the liver, spleen and small intestine, i.e., C3 and ITM2B. The former encodes complement component 3 precursor. Complement component C3 is the central component of the complement cascade system, and people with C3 deficiency are susceptable to bacterial infection (Wimmers et al., 2001; Meijssen et al., 2002). The latter codes for a protein of 266 amino acids in mouse and human, and it is homologous to another integral membrane protein, Itm2A. Itm2A and Itm2B belong to a family of type II integral membrane proteins, which contains a third member, Itm2C. *Itm2b* is ubiquitously expressed in mouse but there had not been related reports in pigs (Pittois et al., 1998; Austen et al., 2002); there were 3 genes co-expressed in the kidney, spleen and small intestine, including ALDOB and two function-unknown genes (predicted by automated

computational analysis using gene prediction methods such as GenomeScan, BLAST etc.). ALDOB encodes fructose-1,6-bisphosphate aldolase B which was studied extensively and had been known in great detail. It is a tetrameric glycolytic enzyme that catalyzes the reversible cleavage of fructose-1,6-bisphosphate (FBP) into glyceraldehyde 3-phosphate (G3P) and dihydroxyacetone phosphate (DHAP), and is thus important in glycolysis and gluconeogenesis. For humans, aldolase A is the major product in the developing embryo, and in adult liver, kidney and intestine, aldolase A expression is repressed and aldolase B is produced (Tolan and Penhoet, 1986; Esposito et al., 2002); There were 2 genes co-expressed in the liver, kidney and small intestine, PCCA and PGRMC1. Both are function-known genes. The former encodes the alpha subunit of the heterodimeric mitochondrial enzyme propionyl-CoA carboxylase which is a mitochondrial, biotin-dependent enzyme involved in the catabolism of branched chain amino acids, odd chain fatty acids, and other metabolites (Campeau et al., 2001; Rodriguez-Pombo et al., 2002). The latter encodes the progesterone binding protein which is a putative steroid membrane receptor. Progesterone induces the acrosome reaction in the mammalian sperm through an increase of intracellular (Ca^{2+}). This protein is expressed predominantly in the liver and kidney in humans and recently the amino acid sequence of a steroid membrane receptor protein from porcine vascular smooth muscle cells was published (Gerdes et al., 1998; Buddhikot et al., 1999). The results here showed that an *in silico* study could be credible and efficient for detecting gene expression profiles in a certain tissue, and it also further confirmed that there were many common aspects in gene expression between humans and pigs. Thus there was relative high conservation in phylogenesis between them. In fact, we could predict or discover genes co-expressed in many tissues of a certain organism by this method, and so study some ubiquitously expressed genes. This would probably be our later work.

Highly homologous genes (top-ranking 10) : By statistical analysis, we found high homologous matches between human genes and porcine ESTs in the major internal organs, represented by the top-ranking 10 records here (Table 3). They were TP53, FLNB, NPM1, TRIP15, ZNF216, RAD21, GYS2, LOC254635, LOC144202, and LOC148356 in the liver (the latter 3 are function-unknown genes), PAI-RBP1, CTNNB1, DNCLI2, ATP5B, HNRPH3, LZK1, CHD2, TSG101, LOC152502, and LOC254635 in the kidney (the latter 2 are function-unknown genes), PDCD10, RAP1B, SFRS1, BTG1, UBE2L3, BUB3, H3F3A, YME1L1, HNRPA2B1, and LOC153416 in the spleen (the latter one is a function-unknown gene), and SCARF2, SNRPC, RHOBTB2, RPL15, RPL11, PCBP1, FLJ10511, APP, CALM2, and CSNK1A1 in the small intestine (all are function-known genes). These function-known genes had been characterized particularly in extensive studies of various organisms. We describe only the top one in every tissue here. In the liver, TP53, as one of the recently well-studied genes, encodes the transcription factor p53 which regulates the expression of different genes related to cell cycle control, apoptosis, specifically in the transition from G0 to G1. As a nuclear protein, It is the most frequently mutated gene in human cancers, and therefore, p53 appears to be an appealing target for gene therapy. It is postulated to bind as a tetramer to a p53-binding site and activates expression of downstream genes that inhibit growth and/or invasion, and thus functions as a tumor suppressor. So it is used widely in human cancer studies. But in pigs, reports or studies for this gene are few (Guimaraes and Hainaut, 2002; Dumont et al., 2003; Narayanan et al., 2003); PAI-RBP1 expressed in the kidney, encodes PAI-1 mRNA-binding protein which binds to an RNA sequence (cyclic nucleotide responsive sequence, CRS) in the plasminogen activator-inhibitor (PAI-1) mRNA that confers cAMP regulation of mRNA stability (Heaton et al., 2001); in the spleen, PDCD10, as a programmed cell death gene, encodes a protein similar to proteins that participate in apoptosis (Scarr et al., 2002); and in the small intestine, SCARF2, as a member of the scavenger receptor class F, codes for a protein similar to SCARF1/SREC-I which is a scavenger receptor protein that mediates the binding and degradation of acetylated low density lipoprotein (Ac-LDL). Multiple functions of scavenger receptors, including endocytosis, phagocytosis and adhesion and signal transduction triggered by the binding and uptake of modified LDL (Yamada et al., 1998; Ishii et al., 2002). The results above showed that there were certain similar events in gene expression between humans and pigs, and so it could provide some relevant information for phylogenesis and even xenotransplantation studies. These sequences were highly homologous between humans and pigs, and so these human genes could be used as candidate genes for porcine gene prediction or discovery. Naturally, these porcine ESTs would be useful for the identification of new genes, physical map construction and so on. This would be confirmed in our discussion and analysis for BAGs later.

Gene expression patterns in the major internal organs of pig : The genes expressed in the major internal organs of the pig are arranged with human chromosome number in Figure 2. By a statistical test, it was found that the gene expression patterns were significantly different in the major internal organs of the pig. The quantitative distribution of tissue expression becomes harder to determine because of the different ways of preparing EST libraries and therefore the frequency of representation of a gene in dbEST should not be used to predict its expression level. However, the presence in a tissue-specific library of an EST attributed to a certain gene implies the given gene is expressed in that

specific tissue (or set of tissues). This reasoning can be used to calculate not only general expression profiles, but also to identify genes that are specifically expressed in one particular tissue or organ (Skrabanek et al., 2001). The results here may provide some information on gene expression patterns in the major internal organs of the pig. The gene expression tendencies or profiles between humans and pigs were demonstrated here. The amount of genes gradually decreased with the increase of the human chromosome number, and there were relatively more homologous porcine genes in human chromosome 1, but fewer in the sex chromosomes. Probably, this was due to the differences in the quantitative distributions of gene expression among the human chromosomes.

Basic active genes (BAGs, represented by the top-ranking 10) : 60 BAGs were screened from the matched records by the selected criteria as described previously (Table 4), and here we analyzed the 16 top-ranking 10 BAGs (7 genes had the same ESTs hits 4), including 2 in the liver (NM_000773, NM_000477), 0 in the kidney, 8 in the spleen (NM_006367, XM_031238, XM_087372, XM_166705, NM_022555, XM_165807, XM_165806, NM_007187), and 6 in the small intestine (NM_014478, XM_058602, XM_063371, XM_068306, XM_094318, XM_171365) respectively. Among these genes, 6 were function-known ones, and the remaining 10 genes were function-unknown genes. The top-ranking BAGs in the liver were identified as CYP2E1 and ALB. The former encodes a member of the cytochrome P450 superfamily of enzymes. Cytochrome (CYP) P450 2E1 is clinically and toxicologically important and it is constitutively expressed in the liver and many other tissues (Lejus et al., 2002; Chalasani et al., 2003). The latter codes for the protein albumin which is synthesized in the liver. Albumin functions primarily as a carrier protein for steroids, fatty acids, and thyroid hormones and plays a role in stabilizing extracellular fluid volume (Gorinstein et al., 2002). In pigs, Shito et al. (2001) used this protein for microchannel bioreactor study and this doubtlessly gave us some inspiration in exploiting these genes expressed in the internal organs. The top-ranking BAGs expressed in the spleen included CAP1, HLA-DRB3, WBP4, C1QG, LOC152059, LOC219398, HLA-DPB1, and LOC221518 (the latter 5 are function-unknown genes). CAP1 encodes adenylate cyclase-associated protein 1. This protein is related to the S.cerevisiae CAP protein, which is involved in the cyclic AMP pathway (Moriyama and Yahara, 2002). HLA-DRB3 codes for a major histocompatibility complex protein, class II, DR beta 3, which belongs to the HLA class II beta chain paralogues. It plays a central role in the immune system by presenting peptides derived from extracellular proteins (Balas et al., 2000). WBP4 encodes WW domain-containing binding protein 4. It is a general spliceosomal protein that may play a role in cross-intron bridging of U1 and U2 snRNPs in the spliceosomal complex A (Bedford et al., 1998). The top-ranking BAGs in the small intestine were recognized as RCP9 and 5 other function-unknown genes. RCP9 encodes calcitonin gene-related peptide-receptor component protein. Calcitonin gene-related peptide (CGRP) is a 37-amino acid neuropeptide which induces increased intracellular cAMP levels. So far, only a single form of CGRP has been found in pigs, despite an extensive search for an additional form. Much lab data indicated that CGRP might act as a neuromodulator of gastric function. To investigate the effect of CGRP on antrum motility, Rasmussen et al. (2001b) used synthetic porcine CGRP to study the effect of calcitonin gene-related peptide (CGRP) on motility and on the release of substance P, neurokinin A, somatostatin and gastrin in the isolated perfused porcine antrum and demonstrated that CGRP is a strong stimulator of antral motility in pigs. By studying of the localisation and neural control of the release of calcitonin gene-related peptide (CGRP) from isolated perfused porcine ileum, it was discovered that CGRP functioned essentially in the intestine of pig (Rasmussen et al., 2001a,b). The expression information of BAGs showed that the expression activities of human genes were apparently different in the major internal organs of the pig. The results indicated it was a efficient way to discover genes preferentially expressed in some tissues by detecting the basic active genes (BAGs), and it was also a brand-new way to discover new genes. On the other hand, we could rapidly acquire the full length cDNA sequence of a certain gene by overlapping fragments assembling of the corresponding ESTs sequences. Then, if necessary, we could combine some other ways of cloning full length cDNA sequence such as RACE (Rapid Amplification of cDNA End), LFS (Ligase-free Subcloning) and so on. In the meantime, we could somewhat know some functions of these sequences in advance by their homologous human genes. Obviously, this method deserves more attention.

Until the complete sequences of the porcine genome are available, a reasonable accomplishment will be the identification of a large proportion of the gene content in the porcine genome. This might be done by EST or genomic sequencing in the near future. The next step, in the short run, would be the analysis of the data and the development of new approaches for *in silico* studies so as to mine and utilize the stocked resources efficiently and fast.

REFERENCES

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde and R. F. Moreno. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Sci. 252:1651-1656.

Attwood, T. K. and C. J. Miller. 2002. Progress in bioinformatics

and the importance of being earnest. Biotechnol. Annu. Rev. 8:1-54.

Austen, B., O. el-Agnaf, S. Nagala, B. Patel, N. Gunasekera, M. Lee and V. Lelyveld. 2002. Properties of neurotoxic peptides related to the BRI gene. Biochem. Soc. Trans. 30:557-559.

Balas, A., S. Santos, M. J. Aviles, F. Garcia-Sanchez, R. Lillo and J. L. Vicario. 2000. Identification by sequencing based typing and complete coding region analysis of three new HLA class II alleles: DRB3×0210, DRB3×0211 and DQB1×0310. Tissue Antigens. 56:380-384.

Barnes, M. R. 2002. Psychiatric genetics *in silico*: databases and tools for psychiatric geneticists. Psychiatric Genetics 12:67-73.

Bedford, M. T., R. Reed and P. Leder. 1998. WW domain-mediated interactions reveal a spliceosome-associated protein that binds a third class of proline-rich motif: the proline glycine and methionine-rich motif. Proc. Natl. Acad. Sci. USA 95:10602-10607.

Birney, E., M. Clamp and T. Hubbard. 2002. Databases and tools for browsing genomes. Annu. Rev. Genomics Hum. Genet. 3:293-310.

Buddhikot, M., E. Falkenstein, M. Wehling and S. Meizel. 1999. Recognition of a human sperm surface protein involved in the progesterone-initiated acrosome reaction by antisera against an endomembrane progesterone binding protein from porcine liver. Molecular and Cellular Endocrinology 158:187-193.

Campeau, E., L. R. Desviat, D. Leclerc, X. Wu, B. Perez, M. Ugarte and R. A. Gravel. 2001. Structure of the PCCA gene and distribution of mutations causing propionic academia. Mol. Genet. Metab. 74:238-247.

Chalasani, N., J. C. Gorski, M. S. Asghar, R. A. Asgha, B. Foresman, S. D. Hall and D. W. Crabb. 2003. Hepatic cytochrome P450 2E1 activity in nondiabetic patients with nonalcoholic steatohepatitis. Hepatology 37:544-550.

Dearden, P. and M. Akam. 2000. Computational biology: Segmentation *in silico*. Nature 406:131-132.

Dumont, P., J. I. Leu, A. C. Della Pietra 3rd, D. L. George and M. Murphy. 2003. The codon 72 polymorphic variants of p53 have markedly different apoptotic potential. Nat. Genet. 33:357-365.

Esposito, G., L. Vitagliano, R. Santamaria, A. Viola, A. Zagari and F. Salvatore. 2002. Structural and functional analysis of aldolase B mutants related to hereditary fructose intolerance. FEBS Lett. 531:152-156.

Farber, C. R. and J. F. Medrano. 2003. Putative *in silico* mapping of DNA sequences to livestock genome maps using SSLP flanking sequences. Animal Genetics 34:11-18.

Gerdes, D., M. Wehling, B. Leube and E. Falkenstein. 1998. Cloning and tissue expression of two putative steroid membrane receptors. Biol. Chem. 379:907-911.

Gill, R. W. and P. Sanseau. 2000. Rapid *in silico* cloning of genes using expressed sequence tags (ESTs). Biotechnol. Ann. Rev. 5:25-44.

Gorinstein, S., A. Caspi, A. Rosen, I. Goshev, M. Zemser, M. Weisz, M. C. Anon, I. Libman, H. T. Lerner and S. Trakhtenberg. 2002. Structure characterization of human serum proteins in solution and dry state. J. Pept. Res. 59:71-78.

Guimaraes, D. P. and P. Hainaut. 2002. TP53: a key gene in human cancer. Biochimie. 84:83-93.

Heaton, J. H., W. M. Dlakic, M. Dlakic and T. D. Gelehrter. 2001. Identification and cDNA cloning of a novel RNA-binding protein that interacts with the cyclic nucleotide-responsive sequence in the Type-1 plasminogen activator inhibitor mRNA. J. Biol. Chem. 276:3341-3347.

Ishii, J., H. Adachi, J. Aoki, H. Koizumi, S. Tomita, T. Suzuki, M. Tsujimoto, K. Inoue and H. Arai. 2002. SREC-II, a new member of the scavenger receptor type F family, trans-interacts with SREC-I through its extracellular domain. J. Biol. Chem. 277:39696-39702.

Jiang, Z., M. Zhang, V. D. Wasem, J. J. Michal, H. Zhang and R. W. Wright, Jr. 2003. Census of Genes Expressed in Porcine Embryos and Reproductive Tissues by Mining EST Database Based on the Human Genes. Biol. Reprod. 69:1177-1182.

Lee, J. H., C. Moran1 and C. S. Park. 2003. Current Status of Comparative Mapping in Livestock. Asian-Aust. J. Anim. Sci. 16:1411-1420.

Lejus, C., A. Fautrel, Y. Malledant and A. Guillouzo. 2002. Inhibition of the cytochrome P450 2E1 by propofol in human and porcine liver microsomes. Biochemical pharmacology. 64:1151-1156.

Lynn, D. J., A. T. Lloyd and C. O'Farrelly. 2003. Bioinformatics: implications for medical research and clinical practice. Clin. Invest. Med. 26:70-74.

Martin, K. J. and A. B. Pardee. 2000. Identifying expressed genes. PNAS. 97:3789-3791.

Meijssen, S., H. van Dijk, C. Verseyden, D. W. Erkelens and M. C. Cabezas. 2002. Delayed and exaggerated postprandial complement component 3 response in familial combined hyperlipidemia. Arterioscler. Thromb. Vasc. Biol. 22:811-816.

Moriyama, K. and I. Yahara. 2002. Human CAP1 is a key factor in the recycling of cofilin and actin for rapid actin turnover. J. Cell. Sci. 115:1591-1601.

Mostert, V. 2000. Selenoprotein P: properties, functions, and regulation. Arch. Biochem. Biophys. 376:433-438.

Narayanan, B. A., N. K. Narayanan, G. G. Re and D. W. Nixon. 2003. Differential expression of genes induced by resveratrol in LNCaP cells: P53-mediated molecular targets. Int. J. Cancer. 104:204-212.

Nieto, N., S. L. Friedman and A. I. Cederbaum. 2002. Cytochrome P450 2E1-derived reactive oxygen species mediate paracrine stimulation of collagen I protein synthesis by hepatic stellate cells. J. Biol. Chem. 277:9853-9864.

Pandey, A. and F. Lewitter. 1999. Nucleotide sequence databases: a gold mine for biologists. Trends Biochem. Sci. 24:276-280.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai and J. Quackenbush. 2003. Bioinformatics 19:651-652.

Pittois, K., W. Deleersnijder and J. Merregaert. 1998. cDNA sequence analysis, chromosomal assignment and expression pattern of the gene coding for integral membrane protein 2B. Gene. 217:141-149.

Pruess, M., W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. Kriventseva, V. Mittard, N. Mulder, I. Phan, F. Servant and R. Apweiler. 2003. The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. Nucleic Acids Res. 31:414-417.

Rasmussen, T. N., P. Schmidt, S. S. Poulsen and J. J. Holst. 2001a.

Localisation and neural control of the release of calcitonin gene-related peptide (CGRP) from the isolated perfused porcine ileum. Regulatory Peptides. 98:137-143.

Rasmussen, T. N., P. Schmidt, S. S. Poulsen and J. J. Holst. 2001b. Effect of calcitonin gene-related peptide (CGRP) on motility and on the release of substance P, neurokinin A, somatostatin and gastrin in the isolated perfused porcine antrum. Neurogastroenterol. Mot. 13:353-359.

Rodriguez-Pombo, P., C. Perez-Cerda, L. R. Desviat, B. Perez and M. Ugarte. 2002. Transfection Screening for Defects in the *PCCA* and *PCCB* Genes Encoding Propionyl-CoA Carboxylase Subunits. Mol. Genet. Metab. 75:276-279.

Rohrer, G. A., S. C. Fahrenkrug, D. Nonneman, N. Tao and W. C. Warren. 2002. Mapping microsatellite markers identified in porcine EST sequences. Animal Genetics 33:372-376.

Saito, Y. and K. Takahashi. 2002. Characterization of selenoprotein P as a selenium supply protein. Eur. J. Biochem. 269:5746-5751.

Scarr, R. B. 2002. PDCD2 is a negative regulator of HCF1 (C1). Oncogene. 21:5245-5254.

Shito, M., N. H. Kim, H. Baskaran, A. W. Tilles, R. G. Tompkins, M. L. Yarmush and M. Toner. 2001. *In vitro* and *in vivo* evaluation of the Albumin synthesis rate of porcine hepatocytes in a flat-plate bioreactor. Artificial Organs. 25:571-578.

Skrabanek, L. and F. Campagne. 2001. Tissueinfo: high-throughput identification of tissues expression profiles and specificity. Nucleic Acids Res. 29:21-102.

Strausberg, R. L., E. A. Feingold, R. D. Klausner and F. S. Collins. 1999. The Mammalian Gene Collection. Science 286:455-457.

Tolan, D. R. and E. E. Penhoet. 1986. Characterization of the human aldolase B gene. Mol. Biol. Med. 3:245-264.

Wimmers, K., S. Mekchay, S. Ponsuksili, T. Hardge, M. Yerle and K. Schellander. 2001. Polymorphic sites in exon 15 and 30 of the porcine C3 gene. Animal Genetics 32:40-53.

Yamada, Y., T. Doi, T. Hamakubo and T. Kodama. 1998. Scavenger receptor family proteins: roles for atherosclerosis, host defence and disorders of the central nervous system. Cell Mol. Life Sci. 54:628-640.