A Thesis

For The Degree Of Doctor Of Philosophy

Analysis of Korean Native Pig Genome Using Full-Length Enriched cDNA Libraries

Vijaya Ramu Dirisala

Department of Animal Biotechnology

GRADUATE SCHOOL

CHEJU NATIONAL UNIVERSITY

TABLE OF CONTENTS

요약문

기능 유전체학 분야에서 full-length cDNA는 genome의 정확한 해석 및 유전자의 구조와 기능 분석을 위하여 필수적이다. 한국재래돼지의 11가지의 각기 다른 조직(대뇌의 신피질, 소뇌, 비장, 간장, 신장, 폐, 뇌간, 정소, 눈, 정자, 근육)으로부터 3가지의 다른 실험 방법(SMART, modified oligo-capping, cap-trapping)을 통하여 15개의 full-length enriched cDNA library가 성공적으로 제작되었다.

제작된 library의 특성과 유용성을 시험해보기 위해 SMART 방법과 modified oligo-capping 방법을 이용하여 제작된 비장, 신피질, 뇌간, 간장의 full-length enriched cDNA library로부터 생성된 총 3,390개 (3,210개의 5’ sequence와 180개의 3’ sequence)의 sequence가 단일염기다형성(SNP)과 전사시작위치(Transcriptional Start Site) 확인을 위하여 분석되었다. 뇌간, 간장, 비장, 신피질 library의 경우, 제한효소 처리 후 확인한 클론들의 평균 insert 길이는 각각 2 Kb, 1.8 Kb, 1.1 Kb, 1.1 Kb였다. 이들 library의 plaque forming unit 또는 colony forming unit은 평균적으로 1x 10^6을 나타내었다. 비장, 뇌간, 간장 library에서 가장 많이 발현된 유전자는 각각 β-globin, tubulin, albumin이었다. Full-length 클론들은 시작코돈을 포함하고 있으며, 비장, 신피질, 뇌간, 간장 library에서 full-length 클론이 차지하는 비율은 각각 60%, 40%, 80%, 70%로 추정되었다. 대뇌의 신피질과 간장 library에서 210개의 클론들을 임의적으로 선택하여 3’ end 부분을 sequencing 하였을 때, 모든 클론들이 poly-A tail을 포함하고 있었다. 네 종류의 library에서 100개의 클론을 BLAST 분석하였을 때, 93%의 sequence (E value < 10^{-100})가 돼지나 다른 종과 일치되었다. 본 연구에서 제작된 library는 full-length 클론들을 생성하기 위해 제작되었으므로 85%의 클론이 현재 NCBI EST에 등록된 sequence보다 더 긴 5’ end를 포함하는 sequence를 가지고 있으며, 돼지 유전자의 5’ UTR 지역을 분석하는데 매우 유용함을 나타낸다.

In silico 분석 방법에 의한 SNP 분석을 위하여 Genbank trace file

archive에서 50,000개의 돼지 EST(Expressed Sequence Tag) 크로마토그램을 검색하였고 본 연구에서 생성된 3,210개의 5’ EST sequence와 병합하였다. Phred quality value가 30 이상인 sequence들을 선택하여 Phrap sequence assembly 프로그램을 이용하여 assemble하였다. Assembly 과정에서 8,118개의 contig가 생성되었다. 49개의 contig는 각 contig 내에 최소 두 개의 한국재래돼지 sequence와 두 개의 EST sequence를 포함하였다. 이들 중 Phrap 분석 후 최소 하나의 candidate SNP를 포함하는 7개의 contig가 선택되었다. 집단분석을 통한 확인 실험을 진행하기 위하여 8개의 candidate cSNP가 채택되었다. 한국재래돼지 sequence를 제거하였을 때 7개의 contig 중에서 3개의 contig만이 cSNP로 추정되었고, 이는 유전적 변이를 증대시키기 위하여 유전적 자원으로써 한국재래돼지의 중요성을 나타낸다. 돼지 유전자의 genomic sequence 정보는 불충분하므로 이들과 밀접하게 연관된 종(*Homo sapiens, Bos taurus, Mus musculus, Canis familiaris*)으로부터 7개의 유전좌위에서 exon-intron 구조에 근거하여 PCR 프라이머를 제작하였고, 4종류의 돼지 품종(듀록, 요크셔, 랜드레이스, 한국재래돼지)에서 각각 다른 5마리의 동물로부터 PCR 산물을 얻었다. 140개의 샘플을 direct sequencing하여 분석한 결과 *in silico* SNP detection으로부터 발굴된 7개의 SNP에서 6개의 SNP가 확인(86% 정확도)되었고, 614bp 당 하나의 SNP꼴로 나타났다.

6개의 확인된 SNP와 더불어 네 가지 품종으로 구성된 20마리의 돼지로부터의 sequence 분석을 통하여 *in silico* 과정에서 확인할 수 없었던 7개의 SNP를 추가로 더 발견하였다. 이들 13개의 확인된 SNP는 338bp 당 하나의 SNP꼴로 나타났다. 동종 염기 간 변환(transition)이 92%로 8%가 나타난 이종 염기 간 변환(transversion)보다 더 많이 나타났다. 4종류의 돼지 품종의 다형성 변이(polymorphic variation)를 분석함으로써 13개의 확인된 SNP에서 대립 유전자 빈도가 측정되었다. 고정된 대립 유전자의 빈도가 듀록과 한국재래돼지보다 요크셔와 랜드레이스에서 상당히 높게 나타났으며 이는 요크셔와 랜드레이스의 경우 이 지역에서 유전적 homozygosity가 더 높음을 의미한다. Hyaluronidase(NM_213953)에서 position 717과 730의 두 SNP는 네 가지의 모든 종에서 모두 보여 지며 매우 polymorphic한 SNP임을 나타낸다.

Vitronectin (D61396)의 SNP(536번 위치)는 한국재래돼지를 제외한 듀록과 요크셔, 랜드레이스에서 monomorphic 하였다. 이에 반하여 β-globin (AY610360)의 SNP 15의 경우 한국재래돼지에서만 monomorphic 하였다. 이 결과는 비록 제한된 개체(20마리)에서 대립 유전자 빈도가 측정되었으나, 한국재래돼지 집단 내에서 특이적인 유전적 다양성이 존재함을 의미한다.

Neuronal and endocrine protein (M23654)의 경우, 한국재래돼지의 아미노산 126번 위치에서 3개의 뉴클레오타이드 결실이 확인되었다. 이 결실은 encode된 단백질의 secretogranin 도메인의 아미노산 valine을 제거한다. 코돈의 alternative splicing 때문에 일어난 이 결실은 본 연구에서 처음으로 확인되었다. 본 연구는 한국재래돼지에 있어서 방대한 규모의 EST sequencing이 본래의 polymorphism을 바탕으로 한 high-resolution gene-function 연구에 효과적으로 응용될 수 있음을 나타낸다.

전사시작위치를 분석하기 위하여 4개의 full-length enriched cDNA library에서 3,390개의 EST sequence를 CAP3 프로그램을 이용하여 assembly 하였고 420개의 contig를 얻었다. 4개의 포유종물 종(*Homo sapiens, Bos Taurus, Mus musculus, Sus scrofa*)에서 각각 최소 5개 이상의 전사시작위치가 알려진 유전자 sequence를 검색하였으며, 이 중 E-value > 100 인 141개의 contig가 선택되었고 이 contig들의 유전자 sequence가 분석되었다. 40개의 contig sequence가 기준을 충족하였고 4개의 각기 다른 포유동물 종들과 비교하여 전사시작위치를 분석하였다. 예비 연구로써 Human T-cell leukemia virus type-1 binding protein (Tax1BP3) (NM_014604), NDRG family member 3 (NM_032013), Serine incorporator 1 (NM_020755), Thiosulfate sulfurtransferase (NM_003312), Polyubiquitin (M18159)의 다섯 가지 유전자의 전사시작위치 분석을 진행하였다.

본 연구에서 제작된 full-length enriched cDNA library의 대다수의 sequence의 5’ 말단이 5’ end 쪽으로 확장되었음에도 불구하고, sequence의 수가 제한되어 각각의 유전자의 전사시작위치로 결정짓기는 불가능하였다. 그러므로 Genbank에서 수집된 4개의 다른 종에서 5번 이상 나타난 유전자의 full-length sequence를 본 연구에서 제작된 full-length sequence와 align하였다.

Human T-cell leukemiavirus type I binding protein 3 (Tax1BP3)의 경우 10개의 *Sus scrofa* sequence, 20개의 *Bos taurus* sequence, 20개의 *Mus musculus* sequence와 같은 위치에서 전사가 시작한다. NDRG family member 3 (NDRG3) 와 serine incorporator 1 (SERINC1) 유전자는 20개의 *Mus musculus* sequence와 같은 위치에서 전사가 시작하고, polyubiquitin (UBC)는 15개의 *Homo sapiens* sequence와 같은 위치에서 전사가 시작한다. 4개의 종으로부터 선택된 5개의 유전자들의 염기서열 중에서 *Mus musculus* sequence가 5' end 방향쪽에 더 많은 variation을 가지고 있음을 보여준다. thiosulfate sulfurtransferase (TST) 유전자의 경우, 모든 종에서 5' end 방향 쪽으로 가장 많은 variation을 보였다. 예비 실험에서 *Mus musculus*의 총 5 유전자 중 두 유전자(TAX1BP3, NDRG3)가 분석되었는데, 이 sequence들은 다른 종들의 sequence와 비교해 볼 때 약간 더 긴 것으로 나타났다. *Bos taurus* sequence는 다른 종의 sequence와 비교해 보았을 때 세 유전자(TAX1BP3, NDRG3, UBC) 에서 더 짧은 것으로 나타났다. *Homo sapiens* 의 NDRG3의 경우 *Homo sapiens*의 다른 유전자 sequence와 비교하였을 때 조금 더 짧았으나 *Bos taurus*의 다른 유전자 sequence와 비교하였을 때 조금 더 긴 것으로 나타났다. *Homo sapiens*의 다섯 유전자들 중에서 이 유전자만이 짧게 나타났다. 기능적 분석을 위하여 transcription factor binding site와 promoter 길이의 차이에 의한 유전자 발현의 상이성 분석에 대한 구체적인 결과의 도출을 위해서는 더 많은 분석이 요구된다. 결론적으로 본 연구에서 제작된 library의 대규모 분석은 돼지의 유전체 분석과 주석달기 연구에 유용한 정보로 이용될 것이다.

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Full-length cDNAs are essential for the correct annotation of genomic sequences and the analysis of the structure and function of the genes in functional genomics era. Fifteen full-length enriched cDNA libraries were successfully constructed from 11 different tissues (neocortex, cerebellum, spleen, liver, kidney, lung, brainstem, testis, eye, sperm, and muscle) of Korean native pig employing three different methodologies (SMART, Modified oligo-capping, Cap-trapping).

To test the quality and usefulness of these libraries, a total of 3390 sequences (3210 5' sequences and 180 3' sequences) generated from sequences of four full-length enriched cDNA libraries of spleen, neocortex, brainstem and liver constructed by SMART and modified oligo-capping methods were analyzed for SNP identification and characterization of Transcriptional Start Sites (TSS). Average length of insert in the clones evaluated by restriction analysis was 2 Kb, 1.8 Kb, 1.1 Kb and 1.1 Kb for the brainstem, liver, spleen and neocortex libraries respectively. The plaque forming unit or colony forming unit of these libraries was found to be 1x 10^{6} on an average. When the start codon containing clones were considered as full-length clones, the percentage of full-length clones from the spleen, neocortex, brainstem and liver libraries were estimated to be 60%, 40%, 80% and 70% respectively. 93% of sequences from the four libraries were matched with sequences of either pig or other species with E value < 10^{-100} based on evaluation of 100 clones. Eighty five percent of clones from the four libraries had longer 5' end sequences than currently available NCBI EST sequences, suggesting that these libraries are very useful for characterization of 5' UTR regions of porcine genes.

For SNP identification by *in silico* analysis, chromatograms of 50,000 pig Expressed sequence tags (ESTs) retrieved from the Genbank trace file archive were combined with 3210 5' EST sequences from four libraries. Sequences with Phred quality value higher than 30 were chosen and assembled using the Phrap sequence assembly program. The assembly process generated 8118 contigs. Forty nine contigs were consisted of both a minimum of two Korean native pig sequences and two public EST sequences within each contig. Among these, seven contigs containing a minimum of one putative SNP from Phrap analysis were selected. Finally eight putative cSNPs were chosen for confirmation through population analysis. Only three of seven contigs remained putative cSNPs when Korean native pig sequences were removed, indicating the importance of Korean native pigs as a genetic resource to increase genetic variation. To confirm the putative cSNPs, PCR primers were designed for the 7 loci based on exon-intron structures of closely related species (*Homo sapiens, Bos Taurus, Mus musculus and Canis familiaris*), since genomic sequence information of porcine genes is hardly available and PCR products were generated from five different animals each from four different pig breeds (Duroc, Yorkshire, Landrace, Korean native pig). All primers successfully amplified, producing specific bands. Analysis of direct sequencing result from 140 samples revealed the confirmation of 6 out of 7 SNPs identified (86 % accuracy) from *in silico* SNP detection which yielded 1 SNP per 614 bp.

In addition to 6 confirmed SNPs, we identified the presence of 7 additional SNPs which were unidentifiable from the *in silico* process through the sequence analysis using four breeds of 20 pigs. The SNP detection frequency from these 13 confirmed SNPs was 1 SNP per 338 bp. Allele frequencies were calculated for 13 confirmed SNPs by analyzing polymorphic variations from four pig breeds. The frequency of fixed alleles was significantly higher (6 of 13, 46.1%) in Yorkshire than Duroc, Landrace and Korean native pig breeds, indicating genetic homozygosity is higher in Yorkshire for these regions.

In the neuronal and endocrine protein (M23654), the three nucleotide deletion was identified from an allele of the gene of Korean native pigs at amino acid position 126. This deletion removes the amino acid Valine from the Secretogranin domain of the encoded protein. We confirmed that the deletion was caused by alternative splicing due to NAGNAG motif.

Our study showed that a large scale EST sequencing from the Korean native pig can be effectively employed for high-resolution gene-function studies based on natural polymorphisms.

For the characterization of TSS, 3390 EST sequences from four full-length enriched cDNA libraries were assembled by CAP3 which yielded 420 contigs. Of these, 141 contigs with E-value<100 were selected and the gene sequences of these contigs were analyzed for the presence of minimum 5 gene sequences in each of the 4 mammalian species (*Homo sapiens*, *Bos Taurus*, *Mus musculus* and *Sus scrofa*). 40 of the contig sequences satisfied the criteria and were analyzed for the presence of TSS in comparison with four different mammalian species. Transcriptional start site analysis was performed for 5 genes, human T-cell leukemia virus type-1 binding protein gene (Tax1BP3) (NM_014604), NDRG family member 3 (NM_032013), Serine incorporator 1 (NM_020755), Thiosulfate sulfur transferase (NM_003312) and Poly ubiquitin (M18159) as a pilot study and similarities and differences between species were analyzed.

The full-length sequence of single gene appearing more than 5 times from four different species was aligned with our full-length sequence and the similarities the differences in the transcription start site of four mammalian species were analyzed. Among the sequences for 5 genes from 4 species, *Mus musculus* sequences showed more variation while reaching the 5’ end. All the species showed maximum variation towards the 5’ end for thiosulfate sulfur transferase (TST) gene. Sequences in *Mus musculus* for two of the genes (TAX1BP3, NDRG3) from the total of five genes analyzed in this pilot study are slightly longer in comparison with sequences from other species. The *Bos taurus* sequences are shorter for 3 genes (TAX1BP3, NDRG3, UBC) in comparison with sequences from other species. The *Homo sapiens* sequences for NDRG3 is little shorter when compared with the *Homo sapiens* sequences from other genes but is slightly longer than that of *Bos taurus* for that gene in comparison with other sequences. This is the only gene among 5 genes where the *Homo sapiens* sequences are short. Further analysis is required in determining transcription factor binding sites for functional analysis and effects due to difference in promoter length.

In conclusion large scale analysis of our libraries will provide useful information for pig genome analysis and annotation.

Chapter I.

I LITERATURE REVIEW

1. PORCINE GENOMICS

1.1 Importance of pigs. The pig, *Sus scrofa*, was domesticated from wild boar subspecies in Asia and Europe (Giuffra et al., 2000; Okumura et al., 2001) over 7000 years ago. By analysis of mitochondrial DNA sequences, it was revealed that the domestication of pig took place at multiple centers across Eurasia Larson et al., (2005). Pigs are widely important in agriculture as one of the world's most important livestock and pigs have vast geographic distribution and are represented by nearly 500 breeds worldwide (Rothschild, 2004). Pork is the major meat consumed (43%) worldwide (Rothschild and Ruvinsky, 1998).

More recently, the pigs role has expanded beyond just being a food source to potentially serving as an important model system for human health and representing a significant future source of organs for transplantation (Rothschild, 2003; 2004). Increased emphasis on understanding animal behavior and welfare in production agriculture has made it extremely important to examine how common agricultural practices affect cognitive development and welfare of the piglet (Gonyou et al., 1998; Worobec et al., 1999; Gaines et al., 2003). Research with the pig falls into two major categories: production of pork to supply human food and gaining basic knowledge applicable to human health (Douglas, 1972; Pond and Houpt, 1978 Mikkelsen et al., 1999; Grate et al., 2003; Traystman., 2003; Schook et al., 2005a).

During domestication, the pig has undergone intense selection pressures for various phenotypes. Intense selection and breeding have provided distinct phenotypes differing in metabolism, fecundity, disease resistance and the products that are used by humans. These selective pressures have differentiated subpopulations and produced phenotypes extremely relevant to current and future biomedical research for humans. The selection of the "mini" and "micro" pigs with respect to size, independently by investigators throughout the world, attests to the global relevance of this experimental animal in biomedical research (Smith et al., 1990).

Clear understanding of genetic interactions with environmental factors will be a major focus of future biomedical research. The pig model system is also relevant to human health research priorities such as obesity, female health, infection, cardiovascular disease, nutritional studies with respect to the pig being an omnivore (Tumbleson and Schook, 1996; Cooper and Keogh, 2001). The vast amount of research has been conducted with respect to interactions of genetic and environmental factors associated with complex and polygenic physiological traits. The pig has also played an extensive role as a source of biological material in physiological and biochemical research. Use of the pig for biochemistry, enzymology, endocrinology, reproduction, and nutrition research has contributed significantly to the continual improvement of human health (Rohrer et al., 2003).

1.2 Porcine genome. The pig *Sus scrofa*, has 38 somatic chromosomes (19 pairs) (Jimenez-Martin et al., 1962). According to the improvement of a reproducible G-banding technique, standardized karyotype of the domestic pig could be established by Committee for the Standardized Karyotype of the Domestic Pig (Gustavsson, 1988). The pig chromosomes consist of 5 pairs of metacentric, 7 pairs of submetacentric and 6 pairs of acrocentric chromosome pairs plus the X and Y (Schmitz et al., 1992).

Schmitz et al (1992) have used flow karyotyping to estimate the size of all pig chromosomes, yielding a cumulative total of 2.72 X 10^9 bp for an X chromosome and 2.62 X 10^9 bp for a Y chromosome.

From the Sino-Danish Porcine Genome Project, it was reported that the GC content of the pig genome was 42%, which was approximately 1% different from that of the human (Werenersson et al, 2005).

1.3 Pig genome sequencing project. Efforts to sequence the pig genome has been initiated by "Swine Genome Sequencing Consortium" (SGSC) comprising of 8 countries including Korea. SGSC was formed in September 2003 by academic, government and industry representatives to provide international coordination for sequencing the pig genome. The SGSC mission is to advance biomedical research for animal production and health by the development of DNA based tools and products resulting from the sequencing of the swine genome. Their objective is to use a hybrid sequencing approach in which 3X coverage of Bacterial Artificial Chromosomes (BACs) comprising the Minimal Tilling Path (MTP) and 3X of the whole genome shotgun libraries will be used to develop a draft 6X coverage of the pig genome (Rothschild, 2005, Schook et al., 2005 (b)). As of June 2006, the SGSC has selected over 50% of the BAC clones of the minimal tilling path. Shotgun libraries are being constructed from these BAC clones and being prepared to enter the sequencing pipeline. The Korean national Livestock Research Institute (NLRI), a member of this project is the first to begin depositing whole genome shot gun reads in the Ensembl/NCBI trace repository with a running total of over 320,000 reads (The International Swine Genome Sequencing Consortium, 2006).

2. PIG GENOME MAPPING

2.1 Genetic mapping. Genome mapping can be described as recording the location of gene or markers of interest that can be classified into two types, genetic mapping and physical mapping. Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequence features on a genome. Alfred Sturtevant developed the first genetic map for Drosophila X chromosome by using genes as markers (Sturtevant, 1913). The greater the frequency of recombination (segregation) between two genetic markers, the further apart they are assumed to be.

By working out the number of recombinants it became possible to obtain a measure for the distance between the genes. This distance is called a genetic map unit (m.u), or a centimorgan (cM) and is defined as the distance between genes for which one product of meiosis in hundred is recombinant. A recombinant frequency (Θ) of 1 % is equivalent to 1 m.u. Genetic maps help researchers to locate other markers, such as other genes by testing for genetic linkage of the already known markers.

2.1.1 *Linkage map of porcine genome.* The linkage map of porcine genome has been mainly developed by 3 groups. European pig Gene Mapping project (PiGMaP) initiated in 1989 was the first one among them. The members of this group developed a three-generation F2 intercross pedigree (Archibald et al., 1995). They set their objective of developing a genetic map with markers spaced at approximately 20 cM intervals over at least 90% of the pig genome, to map loci affecting traits of economic and biological significance in the pig and to develop molecular tools to allow the future identification and cloning of mapped loci (Haley et al., 1990).The members of this project have developed many DNA markers including polymorphic microsatellite markers and developed the genetic (linkage) map.

The second comprehensive genetic linkage map was developed by scientists from Sweden and Denmark. It has been developed by typing 128 genetic markers in a cross between the European wild boar and a domestic breed (Large White). Novel multi-point assignments were provided for 54 of the markers. The map covered about 1800 cM and the average spacing between the markers is 11cM (Marklund et al., 1996). The third map was developed by USDA-MARC. They have linked 376 microsatellite loci with seven restriction fragment length polymorphic loci in a backcross (commercial white composite and a Duroc or a phenotypically different Chinese breed) reference population. Linkage groups are assigned to 13 autosomes and the X-chromosome (Rohrer et al., 1994).

2.1.2 *QTL mapping.* QTL mapping is the statistical study of the alleles which occur in a locus and the phenotypes they produce. As most traits of interest are governed by more than one gene, defining and studying the entire locus of genes related to a trait gives hope of understanding what effect the genotype of an individual might have. QTL analysis serves important practical applications such as marker-assisted selection (MAS) of breeding animals (Hospital et al., 1992; Spelman and Garrick, 1997; Kadarmideen et al., 2006).

Statistical analysis is required to demonstrate that different genes interact with one another and to determine whether they produce a significant effect on the phenotype. QTL mapping identifies particular regions of the genome as containing a gene that is associated with the trait being assayed or measured. They are shown as intervals across a chromosome, where the probability of association is plotted for each marker used in the mapping experiment. Mapping efforts of quantitative trait loci (QTL) in pigs during the past decade have resulted in hundreds of QTLs reported for growth, meat quality, reproduction, disease resistance and other traits and this number is continuously growing with many researchers reporting new QTL (Amayasi et al., 2006; Ramos et al., 2006 Stratil et al., 2006; Sato et al., 2006). Since it's a challenge to correctly locate, interpret and compare QTL results from different studies, Rothschild group reported a relational database (PigQTLdb) to integrate all available pig QTL data in the public domain to facilitate the use of QTL data in further studies (Hu et al., 2005; 2006). The PigQTLdb includes data representing major genes and markers associated with large effect on economically important traits. As of November 2006, over 1287 QTLs from 94 publications have been curated into the database (http://www.animalgenome.org/QTLdb/pig.html) which represents 246 different traits.

2.1.3 *Comparative mapping.* O'Brien et al. (1993) proposed to construct comparative maps in mammals by using a set of anchored reference loci to find evolutionary break points between species. Comparative mapping helps in understanding genome evolution, i.e the evolution of species. Maps constructed in one species can be compared with closely related species by means of common markers (or single gene traits). Development and refinement of comparative maps and combining both chromosome-painting and gene-mapping approaches in a broader number of species enable us to estimate lineage-specific rates of chromosomal change and to make more accurate reconstruction of ancestral genomes (Murphy et al., 2001). Meyers group has constructed a fine resolution comparative map of pig and human (Meyers et al., 2005). Scientists from the European community developed a tool for comparative mapping in pig- The GENETPIG: Identification of genes controlling economic traits in pigs. The GENETPIG database collects the mapping results and links them to other sources of mapping data such as pig maps and comparative mapping results from the Mouse Genome Database and from human-pig bi-directional chromosome painting or Zoo-FISH.

2.2 Physical mapping. Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features (Brown, 1999). Radiation hybrid mapping, *In situ* hybridization and contig building using large insert clones fall in the category of physical mapping.

2.2.1 *Radiation hybrid mapping.* For the generation of radiation hybrid panels (RH), the donor cell line is irradiated with lethal dose of χ rays or ɣ rays, and fused with the recipient cell line (MC Carthy et al., 1996). Non recombinant donor cells will die within a week after irradiation and the recombinants or hybrid colonies will survive and grow for which DNA can be isolated. The hybrid DNAs and control DNAs are then screened for genetic markers generally using PCR. The retention pattern of markers for each hybrid is compared to determine linkage and map distance between markers. The recipient cell line will contain a selectable marker. The distance between genes or markers is expressed in centirays. RH panel exists for several farm animals such as pig (IMpRH7000, Yerle et al., 1998; SSRH, Hamashima et al., 2003), cattle (Womack et al., 1997), and horse (Kiguwa et al., 2000). The first generation porcine whole genome radiation hybrid mapping analysis has been performed on 699 microsatellite markers and 58 ESTs/genes, resulting in 128 linkage groups for the 19 swine chromosomes using the INRA-University of Minnesota porcine Radiation hybrid (IMpRH7000; Hawken et al., 1999).

2.2.2 *In situ hybridization*. *In situ* hybridization is a technique that provides information about the intact chromosomal location of the DNA sequence used as a labeled hybridization probe. In general, metaphase chromosomes are used as the template substrate of *in situ* hybridization. In the early days of its discovery, the probe was radioactively labeled (RISH), but this was not enough to achieve better sensitivity and resolution. To overcome these problems in RISH, the non-radioactive probe labeling system using biotin and digoxigenin has been used and visualized by fluorescein, this procedure is called the fluorescent *in situ* hybridization (FISH; Pinkel et al., 1981). One of advantages in FISH is that several loci are simultaneously detectable with different colored emissions by using fluorolabels (Trask, 1991). Fish on metaphase chromosomes has approximately 1 Mb resolution. Furthermore, the resolution of FISH could be narrowed down to 25 kb using specific probes and interphase chromosomes (fibre-FISH; Heiskanen et al., 1996). Comparative chromosome painting (ZOO-FISH) is a powerful technique, which presents synteny between chromosomes from different species (Scherthan et al., 1994). The comparative cytogenic maps between human and porcine chromosomes were determined by chromosomes painting (Goureau et al., 1996).

3. FULL-LENGTH ENRICHED cDNA LIBRARIES

3.1 Importance of full-length enriched cDNA libraries in comparison with conventional cDNA libraries. A pool of complementary DNA clones produced by cDNA cloning of total messenger RNA from a single source (cell type, tissue, and embryo) constitutes a cDNA library. Majority of transcripts in cDNA library are truncated i.e. they are not extended towards the 5' end, whereas in a full-length enriched cDNA library, majority of transcripts are extended towards the 5' end (Figure 1-1). Full-length cDNAs represent a valuable resource for functional gene studies. Unfortunately cDNA clones constructed according to conventional methods contain low percentage of full-length clones due to the premature stop of reverse transcription of the template mRNA, especially when the mRNA presents a stable secondary structure. Hence full-length cDNAs are strongly underrepresented in conventional libraries. Due to the reduced representation of full-length clones, several rounds of screening are needed to select the cDNAs carrying the complete sequence (full-length cDNA).

To overcome this, several approaches have been developed to determine the 5'end of cDNA. One such method is 5'RACE that provides the Cap site sequence when focusing on particular cDNAs (Schaefer, 1995; Frohman 1998) and primer extension (Mcknight et al., 1982). The drawback of this approach is that it's not suitable for large scale sequencing projects. Other efforts have been made to establish a system for constructing full-length cDNA libraries. Most of these methods are based on either RNA oligo ligation to the 5'end of mRNA (Kato et al., 1994; Suzuki et al., 1997), 5' cap affinity selection via eukaryotic initiation factor 4E (Edery et al., 1995) or 5' cap biotinylation followed by biotin affinity selection (Carninci et al., 1997). Common to all these methods is they are based on selecting the full-length cDNA by the cap structure, which is the specific structure of the 5' end mRNA in eukaryotic cells (Figure 2-1, 2-2, 2-3).

Many conventional cDNA libraries have been constructed in porcine

(Tuggle and Schmitz, 1994; Wintero et al., 1996; Tosser-Klopp et al., 1997; Davoli et al., 1999; Ponsuksili et al., 2001; Smith et al., 2001; Davoli et al., 2002; Fahrenkrug et al., 2002; Rink et al., 2002; Yao et al., 2002; Bertani et al., 2003; Caetano et al., 2003; Chen et al., 2003; Nobis et al., 2003; Tuggle et al., 2003; Jiang et al., 2004; Mikawa et al., 2004; Whitworth et al., 2004; Dvorak et al., 2005 Zhang et al., 2005) which are described in Table 1-1. As of December 2006, 641857 ESTs are available in the Genbank porcine EST database. Recently, full-length enriched cDNA libraries have also been produced (Fujisaki et al., 2004; Uenishi et al., 2004; Dirisala et al., 2005; Kim et al., 2006, Chen et al., 2006) which are described in Table 1-2. Despite the importance of full-length enriched cDNA clonesin porcine functional genomics, the number of full-length enriched clones in the public database is significantly low (Dirisala et al., 2005; Kim et al., 2006). Thus, there is a need for full-length enriched cDNA libraries to be constructed that can serve as valuable resource for porcine functional genomics.

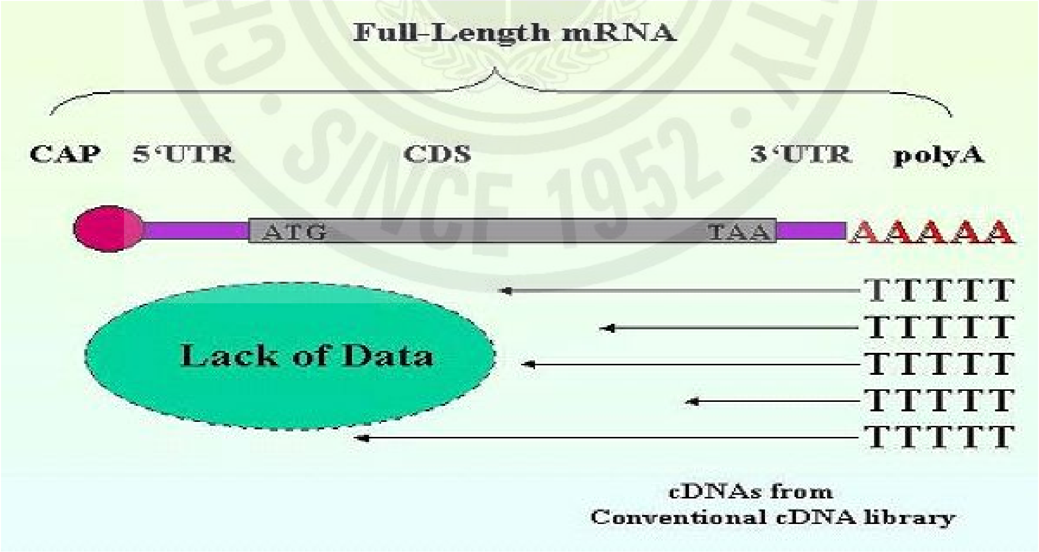


Figure 1-1. Difference between a conventional cDNA library and a full-length enriched cDNA library.

3.2 Applications of full-length enriched cDNA libraries. Full-length cDNAs have allowed the characterization of "microexons" which are usually ignored by existing gene prediction algorithms. Full-length enriched cDNA libraries have also permitted the accurate identification of 5'untranslated regions within full-length cDNAs and mapping of transcriptional start sites thus correctly positioning the respective promoters. This becomes important considering the recent discovery that a great number of transcription factor binding sites (TFBS) on chromosomes 21 and 22 are located in noncanonical regions, which have been poorly examined until now (Cawley et al., 2004). A recently developed technique called CAGE (cap analysis of gene expression) based on the generation of sequence tags of the cap region of full-length molecules, allows the identification of a huge number of novel transcriptional start points (TSP) located far upstream or downstream (both in exons and introns), thus allowing new targets for the identification of new promoter elements and for gene discovery to be identified (Shiraki et al., 2003). Accumulated evidence has shown that multiple transcriptional start sites are more frequent than what was previously thought and that the availability of full-length cDNAs can greatly improve the study of alternative splicing events frequency. Construction of full-length enriched cDNA libraries is an essential step towards the generation of highly informative ESTs for cDNA microarray experiments (Kim et al., 2006). Sequencing of full-length enriched cDNAs from 5' and 3' end gives complete information about the gene and its annotation by connecting the 5' and 3'ESTs separately existing in the public databases. Therefore, availability of full-length cDNA libraries and their sequences represents a critical look for improving the quality of many genomic annotation parameters.

3.3 Identification of SNPs. Much of the EST data generated from porcine cDNA libraries and full-length enriched cDNA libraries can serve as good resource for finding SNPs that can serve as valuable markers for the porcine genome. With the objective to generate SNPs from porcine cDNA libraries and full-length enriched cDNA libraries, many researchers constructed cDNA libraries from tissues of different developmental stages (Fahrenkrug et al., 2002) by pooling RNA in equal amount of the same tissue or cell from different breeds for cDNA library construction (Fahrenkrug et al., 2002; Kim et al., 2006), using cross-bred pigs for cDNA library construction (Uenishi et al., 2004). Since the porcine EST sequences deposited in Genbank are from different breeds, analysis of sequence variations within the Porcine ESTs of Genbank database can help in finding SNPs. There is only one extent native pig breed in Korea ie. the Korean native pig breed (Porter et al., 1993). Analysis of sequence variation within the Korean native pig population and comparison with sequence information from other breeds can provide new information on the current status of the genetic makeup of Korean native pigs (Dirisala et al., 2005; 2006).

3.4 Transcriptional start site analysis. Transcriptional start sites (TSS) marks the 5’ end limit of cDNA and majority of the sequences from the full-length enriched cDNA libraries are extended towards the 5’ end (Suzuki et al., 2001). Analysis of sequences from full-length enriched cDNA libraries of Korean native pigs helps to analyze the TSS which is important for identifying the promoter region.

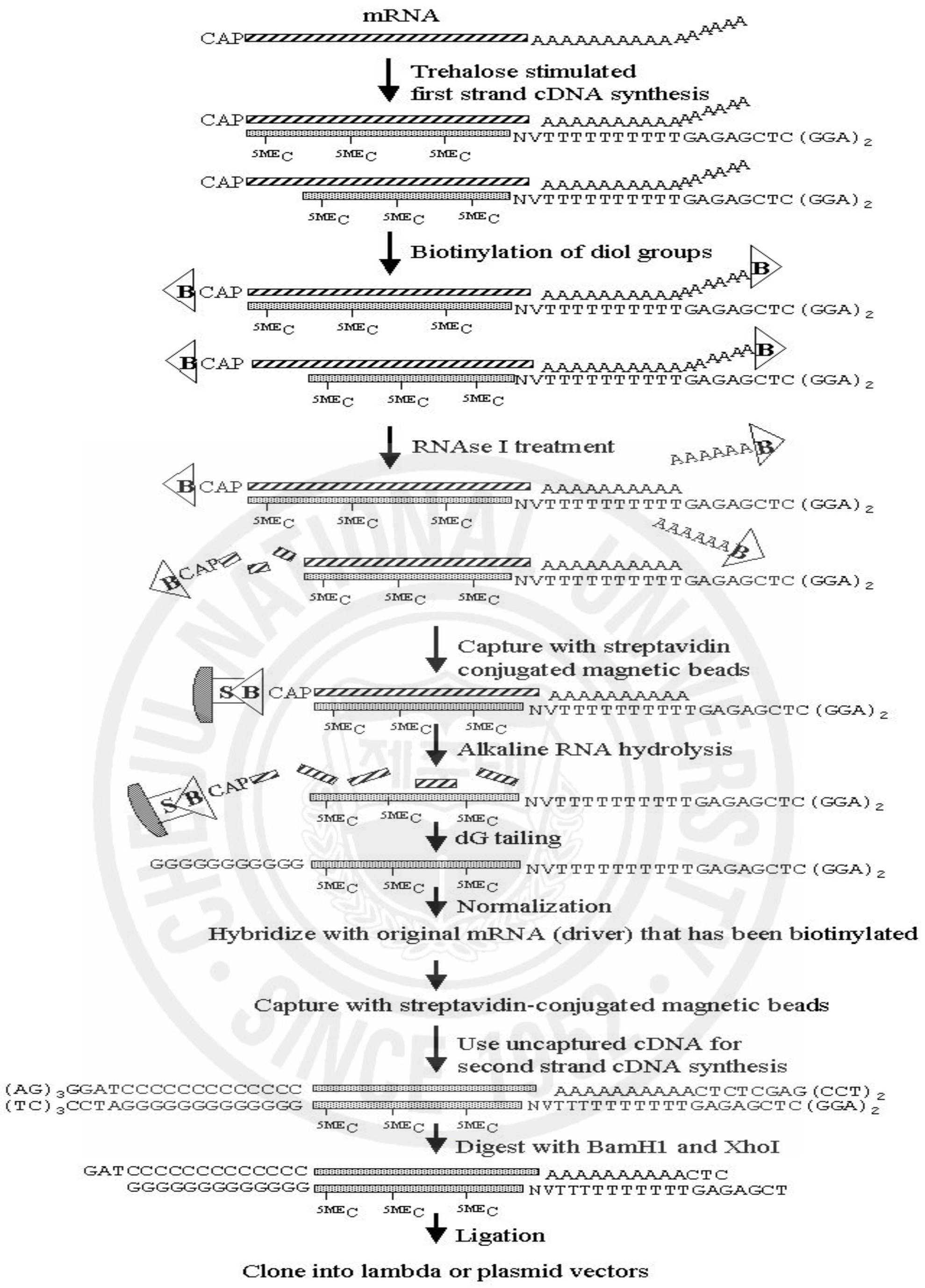

mRNA
CAP
AAAAAAAAAAAAAAAA
Trehalose stimulated
first strand cDNA synthesis
NVTTTTTTTTTTGAGAGCTC (GGA) 2
5ME C
Biotinylation of diol groups
B
RNAse I treatment
AAAAAAB
AAAAAAAAAA
Capture with streptavidin
conjugated magnetic beads
S
Alkaline RNA hydrolysis
dG tailing
GGGGGGGGGGG
Normalization
Hybridize with original mRNA (driver) that has been biotinylated
Capture with streptavidin-conjugated magnetic beads
Use uncaptured cDNA for
second strand cDNA synthesis
(AG) 3GGATCCCCCCCCCCCCC
(TC) 3CCTAGGGGGGGGGGGGG
AAAAAAAAAAACTCTCGAG (CCT) 2
NVTTTTTTTTTTGAGAGCTC (GGA) 2
Digest with BamH1 and XhoI
GATCCCCCCCCCCCCC
GGGGGGGGGGGGG
AAAAAAAAAAACTC
NVTTTTTTTTTTGAGAGCT
Ligation
Clone into lambda or plasmid vectors

Figure 1-2. Strategy for preparation of a full-length enriched cDNA library by biotinylated Cap-trapping procedure (Adapted from: Carninci et al., 1999; Methods in Enzymology).

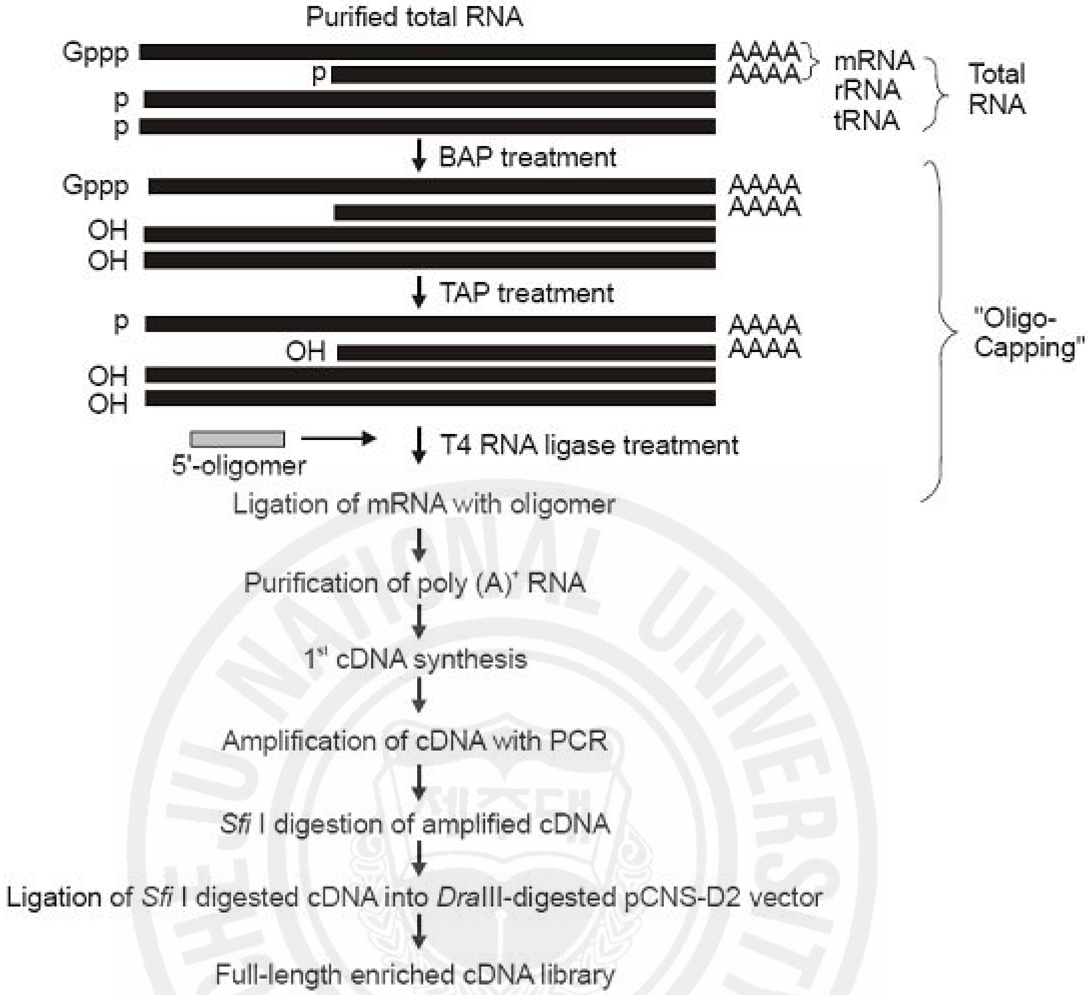

Purified total RNA
Gppp
p
p
p
AAAA
AAAA
mRNA
rRNA
tRNA
Total RNA
BAP treatment
Gppp
OH
OH
AAAA
AAAA
TAP treatment
p
OH
OH
OH
AAAA
AAAA
"Oligo-Capping"
5'-oligomer
T4 RNA ligase treatment
Ligation of mRNA with oligomer
Purification of poly (A)+ RNA
1st cDNA synthesis
Amplification of cDNA with PCR
Sfi I digestion of amplified cDNA
Ligation of Sfi I digested cDNA into DraIII-digested pCNS-D2 vector
Full-length enriched cDNA library

Figure 1-3. Strategy for constructing a full-length enriched cDNA library by modified oligo-capping method (RNA molecules are represented by solid lines and 5' oligo is represented by gray box. Gppp-cap structure; P-Phosphate; OH-Hydroxyl; BAP-Bacterial Alakaline Phosphatase; TAP-Tobacco Acid Pyrophosphate). (Adapted from Oh et al., 2003; Experimental and Molecular Medicine)

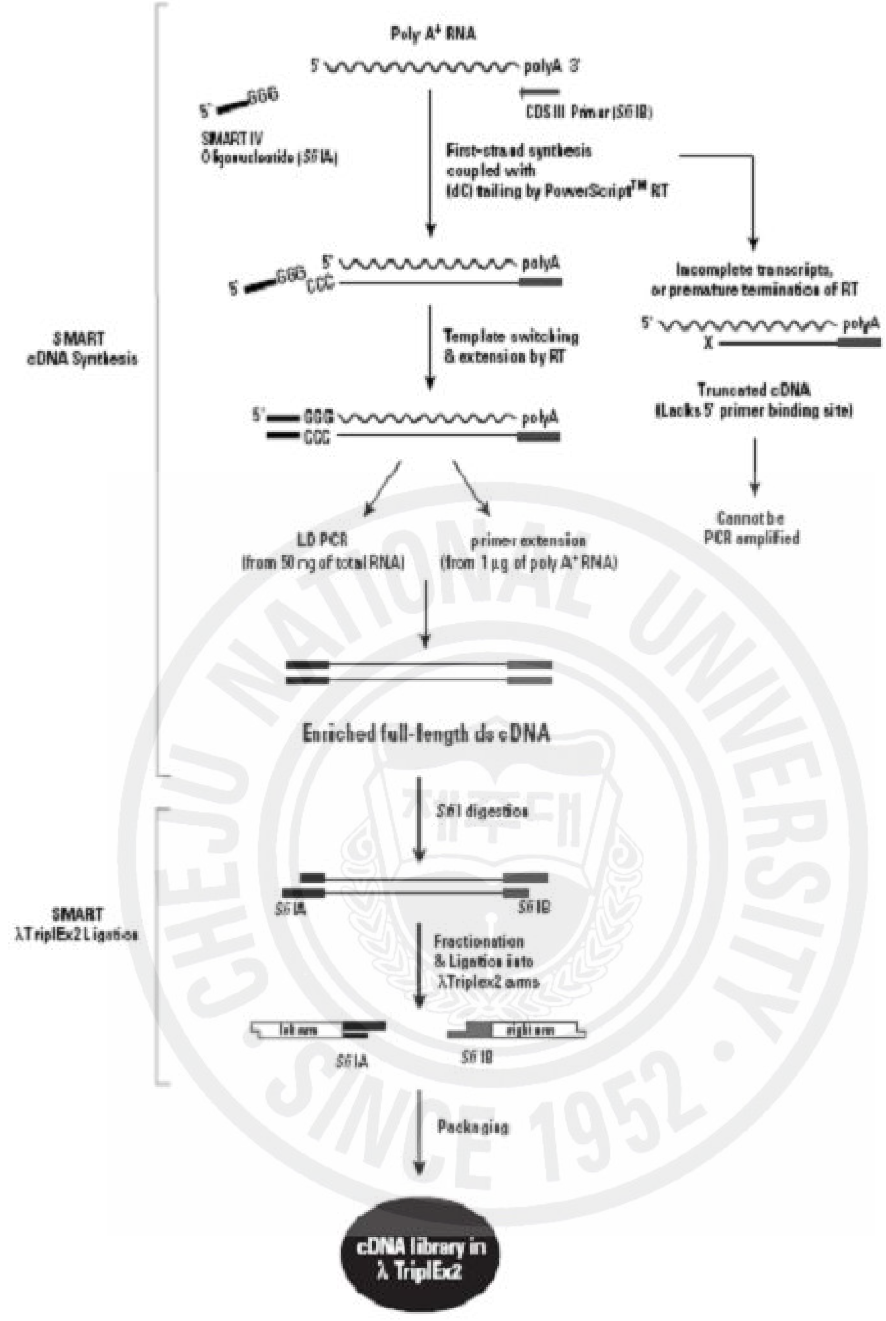


Figure 1-4. Schematic representation of SMART methodology for full-length enriched cDNA library construction. The right side of the flow chart shows the fate of incomplete transcripts caused by RNA degradation or premature termination of reverse transcription. (Adapted from Clontech SMART kit user manual).

Table 1-1. List of porcine cDNA libraries constructed by conventional methods.

Tissue	No. of ESTs deposited	Country	References
Skeletal Muscle		USA	Tuggle and Schmitz, 1994.
Small intestine	839	Denmark	Wintero et al., 1996
Pig ovaries	238	France	Tosser-Klopp et al., 1997.
Skeletal muscle	111	Italy	Davoli et al., 1999.
Embryonic and reproductive tissues	66,245	USA	Fahrenkrug et al., 2001
Skeletal muscle	510	Italy & France	Davoli et al., 2001.
Liver	240	Germany	Ponsuksili et al., 2001.
Early embryonic	781	USA	Smith et al., 2001.
Orthopedic implant-associated infection	7620	USA	Rink et al., 2002.
Skeletal muscle	782	USA	Yao et al., 2002.
Anterior pituitary	168	France	Bertani et al., 2003
Ovarian follicles	5231	USA	Caetano et al., 2003
Brain	965	USA	Nobis et al., 2003

Fetal thymus	7,071	China	Chen et al., 2003
Anterior pituitary, placenta, uterus, embryo, conceptus, hypothalamus, Ovary	21,499	USA	Tuggle et al., 2003.
Germinal vesicle-stage oocytes, in vivo and invitro produced four-cell- and blastocyst-stage embryos	8066	USA	Whitworth et al., 2004
Backfat tissue	3577	Japan	Mikawa et al., 2004
Ovary	15,613	USA	Jiang et al., 2004
Brain (Cerebellum, cortex cerebrum and brainstem)	43,122	China	Zhang et al., 2004

Table 1-2. List of porcine full-length enriched cDNA libraries.

Tissue	Breed	Method	No. of sequences	Country	Average insert size	Full-length%	Journal
Thymus, Spleen, Uterus, Lung, Liver, Ovarian tissues, Peripheral blood mononuclear cells	Crossbred Pigs	oligo-capping	68076	Japan	1.5 Kb	70%	Uenishi et al., 2004; Nucleic Acids Research
Olfactory bulb	Landrace Pigs	oligo-capping	883	Japan	1.7 Kb	80%	Fujisaki et al., 2004; Journal of Veterinary Medical Sciences
Brainstem, Spleen	Korean Native Pigs	oligo-capping, SMART	1205	Korea	2.0 Kb 1.1 Kb	80%, 60%	Dirisala et al., 2005; Korean Journal of Genetics
Back fat tissue	Crossbred Pigs	oligo-capping	16110	Korea	1.7 Kb	70%	Kim et al., 2006; BMC Genomics
Adipose tissue	Lee-Sung Pigs	SMART	2880	Taiwan	0.8 Kb	16%	Chen et al., 2006; Journal of Animal Science
Neocortex, Spleen Liver, Brainstem	Korean Native Pigs	SMART oligo-capping	3390	Korea	1.5 Kb	60%	Dirisala et al., 2006; (Communicated)

REFERENCES

Amayasi, M., Grindflek, E., Javor, A. and Lien, S. 2006. Investigation of two candidate genes for meat quality traits in a quantitative trait locus region on SSC6: the porcine short heterodimer partner and heart fatty acid binding protein genes. J. Anim. Breed. Genet. 123:198-203.

Archibald, A. L., Haley, C. S., Brown, J. F., Couperwhite, S., McQueen, H. A., Nicholson, D., Coppieters, W., Vande Weghe, A., Stratil, A., Wintero, A. K., et al. 1995. The PiGMaP consortium linkage map of the pig (*Sus scrofa*). Mamm. Genome 6:157-175.

Bertani, G. R., Johnson, R. K., Robic, A. and Pomp, D. 2003. Mapping of porcine ESTs obtained from anterior pituitary. Anim. Genet. 34:132-134.

Brown, T.A. 2002. Genomes Willey-Liss 2nd, pp128-129.

Caetano, A. R., Johnson, R. K. and Pomp, D. 2003. Generation and sequence characterization of a normalized cDNA library from swine ovarian follicles. Mamm. Genome 14:65-70.

Carninci, P., Westover, A., Nishiyama, Y., Ohsumi, T., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Schneider, C. and Hayashizaki, Y. 1997. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. DNA Res. 4:61–66.

Cawley, S., Bekiranov, S., Huck, H. N. G., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K. and Gingeras, T. R. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116:499–509.

Chen Y, Li S, Ye L, Geng J, Deng Y and Hu S. 2003. Gene expression profiling in porcine fetal thymus. Geno. Prot. Bioinfo. 1:171-172.

Chen, C. H., Lin, E. C., Cheng, W. T. K., Sun, H. S., Mersmann, H. J. and Ding S.T. 2006. Abundantly expressed genes in pig adipose tissue: An expressed sequence tag approach. J. Anim. Sci. 84:2673-2683.

Cooper, D. K. and Keogh, A. M. 2001. The potential role of xenotransplanatation in treating endstage cardiac disease: a summary of the report of Xenotransplantation Advisory Committee of the International Society for Heart and Lung Transplantation. Curr. Opin. Cardiol. 16:105-109.

Davoli, R., Zambonelli, P., Bigi, D., Fontanesi, L. and Russo, V. 1999. Analysis of expressed sequence tags of porcine skeletal muscle. Gene 233:181-188.

Davoli, R., Fontanesi, L., Zambonelli, P., Bigi, D., Gellin, J., Yerle, M., Milc, J., Braglia, S., Cenci, V., Cagnazzo, M. and Russo, V. 2002. Isolation of porcine expressed tags for the construction of first genomic transcript map of the skeletal muscle in pig. Anim. Genet. 33:3-18.

Dirisala, V. R., Kim, J., Park, K., Kim, N., Lee, K. T., Oh, S. J., Oh, J. H., Kim, N. S., Um, S. J., Lee, H. T., Kim, K. I. and Park, C. 2005. cSNP mining from full-length enriched cDNA libraries of the Korean native pig. Korean J. Genetics 27:329-335.

Dirisala, V. R., Kim, J., Park, K., Lee, H. T and Park, C. 2006. cSNP mining from full-length enriched cDNA libraries of the Korean native pig. 30th International Conference on Animal Genetics, Porto Seguro, BA., Brazil.

Douglas W. R. 1972. Of pigs and men in research: a review of applications and analogies of the pig, S*us scrofa*, in human medical research. Space Life Sci. 3:226-234.

Dvorak, C. M., Hyland, K. A., Machado, J. G., Zhang, Y., Fahrenkrug, S. C. and Murtaugh, M. P. 2005. Gene discovery and expression profiling in porcine peyer's patch. Vet. Immunol. Immunopathol. 105:301-315.

Edery, I., Chu, L. L., Sonenberg, N. and Pelletier, J. 1995. An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). Mol. Cell. Biol. 15:3363–3371.

Fahrenkrug, S. C., Smith, T. P., Freking, B. A., Cho, J., White, J., Vallet, J., Wise, T., Rohrer, G., Petea, G., Sultana, R., Quackenbush, J. and Keele, J. W. 2002. Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. Mamm. Genome 13:475-478.

Frohman, M. A., Dush, M. K. and Martin, G. R. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. 85:8998–9002.

Fujisaki, S., Sugiyama, A., Eguchi, T., Watanabe, Y., Hiraiwa, H., Honma, D., Saito, T. and Yasue, H. 2004. Analysis of a full-length cDNA library constructed from swine olfactory bulb for elucidation of expressed genes and their transcription initiation sites. J. Vet. Med. Sci. 66:15-23.

Gaines, A. M., Carroll, J. A., Yi, G. F., Allee, G. L. and Zannelli, M. E.2003. Effect of menhaden fish oil supplementation and lipopolysaccharide exposure on nursery pigs. Domest. Anim. Endocrinol. 24:353–365.

Giuffra, J. M., Kijas, J. M. H., Amarger, V., Carlborg, O., Jeon, J. T. and Andersson, L. 2000. The origin of the domesticated pig: Independent domestication and subsequent introgression. Genetics 154:1785-1791.

Gonyou, H. W., Beltranena, E., Whittington, D. L. and Patinence, J.

F.1998. The behaviour of pigs weaned at 12 and 21 days of age from weaning to market. Can. J. Anim. Sci.78:517–523.

Goureau, A., Yerle, M., Schmitz, A., Riquet, J., Milan,D., Pinton, P., Frelat, G. and Gellin, J.1996. Human and porcine correspondence of chromosome segments using bindirectional chromosome painting. Genomics 36:252-262.

Grate, L. L., Golden, J. A., Hoopes, P. J., Hunter, J. V. and Dunhaime, A. C. 2003. Traumatic brain injury in piglets of different ages: techniques for lesion analysis using histology and magnetic resonance imaging. J. Neurosci. 123:201–206.

Gustavsson, I. 1998. Standard karyotype of the domestic pig. Committee for the Standardized Karyotype of the Domestic Pig. Hereditas 109:151-157.

Haley, C. S., Archibald, A. L., Andersson, L., Bosma, A. A., Davies, W., Fredholm, M., Geldermann, H., Groenen, M., Gustavsson, I., Ollivier, L., Tucker, E. M. and Van de Weghe, A. 1990. The Pig Gene Mapping Project – PiGMaP. 4thWorld Congress on Genetics Applied to Livestock Production XⅢ. p. 67-70.

Hamashima, N., Suzuki, H., Mikawa, A., Morozumi, T., Plastow, G. and Mitsuhashi, T. 2003. Construction of a new porcine whole-genome framework map using a radiation hybrid panel. Anim. Genet. 34:216-220.

Hawken, R. J., Murtaugh, J., Flickinger, G. H., Yerle, M., Robic, A., Milan, D., Gellin, J., Beattie, C. W., Schook, L. B. and Alexander, L. J. 1999. A first-generation porcine whole-genome radiation hybrid map. Mamm. Genome 10:824-830.

Heiskanen, M., Peltonen, L. and Palotif, A. 1996. Visual mapping by high resolution FISH. Trends Genet. 12:379-382.

Hospital, F., Chevalet, C. and Mulsant, P. 1992. Using markers in gene introgression breeding programs. Genetics 132:1199-1210.

Hu, Z., Dracheva, S., Jang, W., Maglott, D., Bastiaansen, J., Rothschild, M. F. and Reecy, J. M. 2005. A QTL resource and comparison tool for pigs: PigQTLDB. Mamm. Genome 16:792-800.

Hu, Z., Humphray, S., Scott, C., Meyers, S. N., Rogers, J., Rothschild, M. F. and Reecy, J. M. 2006. Extension of PigQTLdb: Genome-wide Alignment of BAC FPC Maps and RH Maps for QTL Positional Gene Mining Plant & Animal Genome XIV Conference, San Diego, CA, U.S.A. p. 14-18.

Jiang, H., Whitworth, K. M., Bivens, N. J., Ries, J. E., Woods, R. J., Forrester, L. J., Springer, G. K., Mathialagan, N., Agca, C., Prather, R. S. and Lucy, M. C. 2004. Large-scale generation and analysis of expressed sequence tags from porcine ovary. Biol. Reprod. 71:991-2002.

Jimenez-Marin, G., Lopez-Saez, J. F. and Monge, E. G. 1962. Somatic chromosomes of the pig. J. Hered. 53:281.

Kadarmideen, H. N., Rohr, P. V. and Jans, L. L. G. 2006. From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. Mamm. Genome 17:548-564.

Kato, S., Sekine, S., Oh, S. W., Kim, N. S., Umezawa, Y., Abe, N., Yokoyama Kobayashi M. and Aoki, T. 1994. Construction of human full-length cDNA bank. Gene 150:243-250.

Kiguwa, S. L., Hextall, P., Smith, A. L., Critcher, R., Swinburne, J., Millon, L., Binns, M. M., Goodfellow, P. N., McCarthy, L. C., Farr, C. J. and Oakenfull., E. A. 2000. A horse wholegenome-radiation hybrid panel: Chromosome 1 and 10 preliminary maps. Mamm. Genome 11: 803-805.

Kim, T. H., Kim, N. S., Lim D., Lee, K. T., Oh, J. H., Park, H. S., Jang, G. W., Kim, H. Y., Jeon, M, Choi, B. H., Lee, H. Y., Chung, H. Y. and

Kim, H. 2006. Generation and analysis of large-scale expressed sequence tags (ESTs) from a full-length enriched cDNA library of porcine backfat tissue. BMC Genomics 7:36.

Larson, G., Dobney, K., Albarella, U., Fang, M., Smith, E.M., Robins, J., Lowden, S., Finlayson, H., Brand T., Willerslev, E., Rowley-Conwy. P., Andersson, L. and Cooper. A. 2005. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science 307:1618-1621.

Marklund, L., Johansson, M. M., Hoyheim, B., Davies, W., Fredholm, M., Juneja, R.K., Mariani, P., Coppieters, W., Ellegren, H. and Andersson, L. 1996. A comprehensive linkage map of the pig based on a wild pig-Large White intercross. Anim. Genet. 27:255-269.

McCarthy, L. 1996. Whole genome radiation hybrid mapping. Trends Genet. 12:491-493.

McKnight, S. L. and Kingsbury, R. (1982) Transcriptional control signals of a eukaryotic protein-coding gene. Science 217:316–324.

Meyers, S. N., Rogatcheva, M. B., Larkin, D. M., Yerle, M., Milan, D., Hawken, R. J., Schook, L. B., Beever, J. E. 2005. Piggy-BACing the human genome Ⅱ. A high-resolution, physically anchored, comparative map of the porcine aurosomes. Genomics 86:739-754.

Mikawa, A., Suzuki, H., Suzuki, K., Toki, D., Uenishi, H., Awata, T. and Hamashima, N. 2004. Characterization of 298 ESTs from porcine back fat tissue and their assignment to the SSRH radiation hybrid map. Mamm. Genome 15:315-322.

Mikkelsen, M., Moller, A., Jensen, L. H., Pedersen, A., Harajehi, J. B. and Pakkenberg, H. 1999. MPTP-induced Parkinsonism in minipigs: a behavioral, biochemical, and histological study. Neurotoxicol. Teratol. 21:169–175.

Murphy, W. J., Stanyon, R. and O'Brien, S. J. 2001 Evolution of mammalian genome organization inferred from comparative gene mapping. Genome Biol. 2:REVIEWS0005.

Nobis, W., Ren, X., Suchyta, S. P., Suchyta, T. R., Zanella, A. J. and Coussens, P. M. 2003. Development of a porcine brain cDNA library, EST database and microarray resource. Physiol. Genomics 16:153-159.

O'brien, S. J., Womack, J. E., Lyons, L. A., Moore, K. J., Jenkins, N. A. and Copeland, N. G. 1993. Anchored reference loci for comparative genome mapping in mammals. Nat. Genet. 3:103-112.

Okumura, N., Kurosawa, Y., Kobayashi, E., Watanobe, T., Ishiguro, N., Yasue, H. and Mitsuhashi, T. 2001. Genetic relationship amongst the major non-coding regions of mitochondrial DNAs in wild boars and several breeds of domesticated pigs. Anim. Genet. 32:139-147.

Pinkel, D., Straume, T. and Gray, J. W. 1981. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. Proc. Natl. Acad. Sci. 83:2934-2938.

Pond, W.G and Houpt, K. A. 1978. The biology of the pig. Ithaca Comstock Pub. Associates p. 13-15.

Ponsuksili, S., Wimmers, K. and Schellander, K. 2001. Application of differential display RT-PCR to identify porcine liver ESTs. Gene 280:75-85.

Porter, V. 1993. Pigs-a handbook to the breeds of the world. Helm information Ltd, UK.

Ramos, A. M., Helm, J., Sherwood, J., Rocha, D. and Rothschild, M. F. 2006. Mapping of 21 genetic markers to a QTL region for meat quality on pig chromosome 17. Anim. Genet. 37:296-297.

Rink, A., Santschi, E. M. and Beattie, C. W. 2002. Normalized cDNA libraries from a porcine model of orthopedic implant-associated infection. Mamm. Genome 13:198-205.

Rohrer, G. A., Alexander, L. J., Hu, Z., Smith, T. P., Keele, J. W. and Beattie, C. W. 1996. A comprehensive map of the porcine genome. Genome Research 6:371-391.

Rohrer, G., Beever, J. E., Rothschild, M. F. and Schook, L. B. 2003. Porcine genome sequencing initiative. p. 1-2.

Rothschild, M. F. 2003. Advances in pig genomics and functional gene discovery. Comp. Funct. Genom. 4:266-270.

Rothschild, M. F. 2004. Porcine genomics delivers new tools and results: this little piggy did more than just go to market. Genet. Res. 83:1-6.

Rothschild, M.F. 2005. Sequencing the pig genome. Iowa State University Animal Industry Report.

Rothschild, M. F. and Ruvinsky, A. 1998. The genetics of the pigs. CAB international. p. 313-344.

Sato, S., Atsuji, K., Saito, N., Okitsu, M., Sato, S., Komatsuda, A., Mitsuhashi, T., Nirasawa, K., Hayashi, T., Sugimoto, Y. and Kobayashi, E. 2006. Identification of quantitative trait loci affecting corpora lutea and number of teats in a Meishan x Duroc F2 resource population. J. Anim. Sci. 84:2895-2901.

Schaefer, B. C. 1995. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. Anal. Biochem. 227:255–273.

Scherthan, H., Cremer, T., Arnason, U., Weier, H. U., Lima-de-Faria, A. and Fronicke, L. 1994. Comparative chromosome painting discloses homologous segments in distantly related mammals. Nature Genet. 6:342-347.

Schmitz, A., Chaput, B., Fouchet, P., Guilly, M.N., Frelat, G.. and Vaiman, M. 1992. Swine chromosomal DNA quantitation by bivariate flow karyotype interpretation. Cytometry 13:703-710.

Schook, L., Beattie, C., Beever, J., Donovan, S., Jamison, R., Zuckermann, F., Steven Niemi, S., Rothschild, M. F., Rutherford, M. and Smith, D. 2005 (a). Swine in Biomedical Research: Creating the Building Blocks of Animal Models. Animal Biotechnology. 16:183-190.

Schook, L.B., Beever, J.E., Rogers, J., Humphray, S., Archibald, A., Chardon, P., Milan, D., Rohre, G.A. and Eversole, K. 2005 (b). Swine genomic sequencing consortium (SGSC): A strategic roadmap for sequencing the pig genome. Comparative and functional genomics 6: 251-255.

Shiraki, T., Kondo, S. Katayama, K., Waki, T., Kasukawa, H., Kawaji., Kodzius, R., Watahik, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. 100:15776–15781.

Smith, T. P., Fahrenkrug, S. C., Rohrer, G. A., Simmen, F. A., Rexroad, C. E. and Keele, J. W. 2001. Mapping of expressed sequence tags from a porcine early embryonic cDNA library. Anim. Genet. 32:66-72.

Smith, A.C., Spinale, F.G. and Swindle, M.M. 1990. Cardiac function and morphology of Hanford miniature swine and Yucatan miniature and micro swine. Lab. Anim. Sci. 40:47-50.

Spelman, R. J. and Garrick, D. J. 1998. Genetic and economic responses for within-family markers-assisted selection in dairy cattle breeding schemes. J. Dairy Sci. 81:2942–2950.

Stratil, A., Van Poucke, M., Bartenschlager, H., Knoll, A., Yerle, M., Peelman, L. J., Kopecny, M. and Geldermann, H. 2006. Porcine OGN and ASPN: mapping, polymorphisms and use for quantitative trait loci identification for growth and carcass traits in a Meishan x Pietrain intercross. Anim. Genet. 37:415-418.

Sturtevant, A. H. 1913. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J. Exp. Zoology 14:43-59.

Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. and Sugano, S. 1997. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. Gene 200:149–156.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T. and Morishita, S. 2001. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep. 2:388–393.

The International Swine Genome Sequencing Consortium. 2006. Pig Tales News letter. 1:1-2.

Tosser-Klopp, G., Benne, F., Bonnet, A., Mulsant, P., Gasser, F. and Atey, F. 1997. A first catalog of genes involved in pig ovarian follicular differentiation. Mamm. Genome. 8:250-254.

Trask, B. J. 1991. Fluorescence in situ hybridization: applications in cytogenetics and gene mapping. Trends Genet. 7:149-154.

Traystman, R. J. 2003. Animal models of focal and global cerebral ischemia. ILAR J. 44:85–95.

Tuggle, C. K. and Schmitz, C. B. 1994. Cloning and characterization of pig muscle cDNAs by an expressed sequence tag approach. Anim. Biotechnol. 5:1-13.

Tumbleson, M. E. and Schook, L. B. 1996. Advances in Swine Biomedical Research Plenum Press. p. 905.

Uenishi, H., Eguchi, T., Suzuki, K., Sawazaki, T., Toki, D., Shinkai, H., Okumura, N. and Awata, T. 2004. PEDE (Pig EST Data Explorer): Construction of a database for ESTs derived from porcine full-length cDNA libraries. Nucleic Acids Res. 32:484-488.

Wernersson, R., Schierup, M. H., Jorgensen, F. G., Gorodkin, J., Panitz, F., Staerfeldt, H H., Christensen, O. F., Mailund, T., Hornshoj, H., Klein, A., Wang, J., Liu, B., Hu, S., Dong W., Li, W., Wong, G. K., Yu, J., Wang, J., Bendixen, C., Fredholm, M., Brunak, S., Yang, H. and Bolund, L. 2005. Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing. BMC Genomics 10:70.

Whitworth, K., Springer, G. K., Forrester, L. J., Spollen, W. G., Ries, J., Lamberson, W. R., Bivens, N., Murphy, C. N., Mathialigan, N., Green, J. A. and Prather, R. S. 2004. Developmental expression of 2489 gene clusters during pig embryogenesis: an expressed sequence tag project. Biol. Reprod. 71:1230-1243.

Wintero, A. K., Fredholm, M. and Davies, W. 1996. Evaluation and characterization of porcine small intestine cDNA library: analysis of 839 clones. Mamm. Genome 7:509-517.

Womack, J. E., Johnson, J. S., Owens, E. K., Rexroad, C. E., Schlapfer, J. and Yang, Y. P. 1997. A whole genome radiation hybrid panel for bovine gene mapping. Mamm. Genome 8:854-856.

Worobec, E. K., Duncan, I. J. H. and Widowski, T. M.1999. The effects of weaning 7, 14, and 28 days on piglet behaviour. Appl. Anim. Behav. Sci. 62:173–182.

Yerle, M., Pinton, P., Robic, A., Alfonso, A., Palvadeau, Y., Delcros, C., Hawken, R, Alexander, L., Beattice, C., Schook, L., Milan, D. and Gellin, J. 1998. Construction of whole genome radiation hybrid panel for high resolution gene mapping in pigs. Cytogen cell Genet. 82:182-188.

Yao, J., Coussens, P. M., Saama, P., Suchyta, S. and Ernst, C. W. 2002. Generation of expressed sequence tags from a normalized porcine skeletal muscle cDNA library. Anim. Biotechnol. 13:211-222.

Zhang, B., Jin, W., Zeng, Y., Su, Z., Hu, S. and Yu, J. 2004. EST-based analysis of gene expression in the porcine brain. Geno. Prot. Bioinfo. 2: 237-244.

Chapter 2.

CONSTRUCTION AND ANALYSIS OF FULL-LENGTH ENRICHED cDNA LIBRARIES FROM KOREAN NATIVE PIGS

1. INTRODUCTION

A full-length enriched cDNA library is advantageous over a conventional cDNA library since it allows cloning of complete sequence in a single step. However the representation of full-length cDNA clones has been low in cDNA libraries prepared using standard techniques. To overcome this problem, researchers devised various methods for the construction of full-length enriched cDNA libraries (Frohman et al., 1988; Edery et al., 1995; Carninci et al., 1996; Zhu et al., 2001; Clepet et al., 2004; Kato et al., 2005). Of these methods, only three methods i.e oligo-capping, cap-trapping and SMART procedure have been widely employed for construction of full-length enriched cDNA libraries (Sugahara et al., 2001) (Figures 1-2, 1-3, 1-4). All these methods have their own advantages and disadvantages. We constructed full-length enriched cDNA libraries from tissues of Korean native pigs using these three methods.

2. MATERIALS AND METHODS

2.1 Construction of full-length enriched cDNA libraries by modified oligo-capping method.

2.1.1 *Tissue collection and RNA isolation.* Tissue samples were collected from Korean native pigs in Jeju. Neocortex, liver, spleen, brainstem, testis, kidney tissues were dissected from Korean native pig at 5 days of age, liver was dissected from Korean native pig at 9 days of age and cerebellum was dissected from Korean native pig at 24 days of age. The tissues were snap frozen in liquid nitrogen and stored at −80°C until use. Total RNA was prepared from one gram of tissue by using RNeasy Maxi kit (Qiagen, Hilden, Germany) following manufacturers instructions. The quality of RNA was examined for integrity by formaldehyde denaturing agarose gel electrophoresis. The mRNA was prepared from the total RNA using a µMACS mRNA isolation kit (Miltenyi biotech, Bergisch Gladbach, Germany).

2.1.2 *Brief description of method employed.* RNA (200 µg) from brainstem and liver tissues was used for the construction of cDNA library by modified oligo-capping method at the Genome Research Centre, KRIBB (Oh et al., 2003). The RNA sample was treated with bacterial alkaline phosphatase (TaKaRa, Shiga, Japan) and then with 100 units of tobacco acid pyrophosphatase (Waco, Osaka, Japan). The pretreated total RNA was ligated with 0.4 µg of5'-oligoribonucleotide (5'-AGCAUCGAGUCGGCCUUGUUGGCCUACUGG-3'). After completing the oligo capping reactions, mRNA was isolated using an Oligotex mini kit (Qiagen, Hilden, Germany). The synthesis of first-strand cDNA from the purified mRNA and the cDNA amplification were performed as previously described (Maruyama et al., 1994). The amplified PCR products were then digested with *Sfi*I, and cDNAs longer than 1.3 kb were ligated into *Dra*III-digested pCNS-D2 (Oh et al., 2003) which is described in Figure 2-1 in an orientation-defined manner. The ligated cDNA was then transformed into *Escherichia coli* Top 10F' (Invitrogen, Carlsbad, CA, USA) by electroporation (BioRad, Hercules, USA) and plated onto ALb plates (50 μg/ml ampicillin). The size of cDNA inserts was evaluated by restriction analysis of clones. Plasmid DNA was prepared in 96-well plate format using multi-well filter plates (Pall Corporation, Ann Arbor, MI, USA). Sequencing of cloned cDNA inserts was performed with 200ng of plasmid DNA as template and T7 (AATACGACTCACTATAG) as primer using ABI PRISM Bigdye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems, Foster City, CA, USA) according to manufacturer's instructions and analyzed on ABI 3700 automated sequencers (Applied Biosystems, Foster City, CA, USA).

2.2 Construction of full-length enriched cDNA libraries by cap-trapping procedure.

2.2.1 ***First strand cDNA synthesis.*** RNA (50 μg) isolated from spleen tissue was used for the construction of full-length enriched cDNA library by cap-trapping (Carninci et al., 1997). Reverse transcription reaction was carried out in 150μl volume using 2000 U of Superscript^{TM}III reverse transcriptase (Invitrogen, Carlsbad, CA, USA). First, 50μg RNA, 10mM each of dATP, dTTP, dGTP, 5-methyl DCTP and 10μg of primer adapter containing the *Xho*I site (5'AGATTGGTCTCCTCGAGT$_{(18)}$VN-3' (V as degenerate base in synthesis as G, A or C and N as G, A, T or C) were added into one tube and incubated at 65°c for 10 minutes. Then, 30μl of first-strand buffer, 7.5μl of 0.1M DTT and 1μl RNase inhibitor were added in a separate tube and both the tubes were pre-incubated at 42°c for 2followed by negative ramp of -17°C, with slope of 0.1°C/second, followed by 25°C for 10 minutes and 50°C for 1 hour. Superscript reverse transcriptase (2000 U) (Invitrogen, Carlsbad, CA, USA) was added during the start of 50°C for 1 hour step. To stop the reaction, 2*μ*l of 0.5m EDTA, 2*μ*l of 10% SDS and 5*μ*l of 10ml proteinase K were added, and the reaction mixture was further incubated at 45°C for 1 hour. Subsequently, the cDNA/RNA was extracted once with phenol–chloroform, precipitated with ethanol, washed with 70% ethanol and resuspended in RNase-free water.

2.2.2 *RNA oxidation, biotinylation and full-length cDNA capture.* The resuspended cDNA/RNA was oxidized in 66m sodium acetate (pH 4.5) and 5m $NaIO_4$. The oxidation was carried out on ice in the dark for 45 minutes. To precipitate the oxidized RNA, 10% SDS to a final concentration of 0.1% SDS, 5M NaCl to a final volume of 0.5 M NaCl were added. To the total reaction mixture, 1 volume of isopropanol was added, and the mixture was centrifuged at 12000xg for 30 min at 4°C. The pellet was washed with 70% ethanol and resuspended in RNase-free water. Then 1M sodium acetate (pH 6.1), 10% SDS and 10m biotin hydrazide long-arm were added to the oxidized cDNA/RNA for biotinylation overnight at room temperature. Subsequently, the cDNA/RNA was precipitated at −80°C for 1h by adding 1M sodium acetate (pH 6.1), 5M NaCl and 2.5 volumes of ethanol to the total of the volume was added. Finally, the pellet was washed twice with 80% ethanol and resuspended in RNase-free water. RNase digestion of the first-strand cDNA reaction was performed using RNase I at 37°C for 1 MPG–streptavidin beads (500μg, PureBiotech LLC,, Middlesex, NJ, USA) and DNase-free tRNA (100μg) were mixed and incubated on ice for 30 min. The beads were separated using a magnetic stand and washed three times with 500 µl of washing/binding buffer (4.5M NaCl and 50mM EDTA, pH 8.0). Finally, the beads were resuspended in washing/binding buffer and mixed with the cDNA/RNA sample at room temperature for 30min with gentle mixing. After removal of unbound cDNA/RNA, the beads were washed three times with 4.5M NaCl and 50mm EDTA (pH 8.0). To release the cDNA from the beads, 100µl of 50mM NaOH/1mM EDTA (pH 8.0) were added to the cDNA/RNA mixture and eluted. Eluted cDNA was added to a tube containing 100μl of 1m Tris-HCl (pH 7.5). Subsequently, the cDNA was extracted once with phenol-chloroform, precipitated with ethanol and resuspended in RNase-free water.

2.2.3 *Double strand DNA synthesis and sequencing of the clones.* The primer 5'-GAGAGAGAG AGAGAGAGAGAGCTCACTAGTCCCCCCCCCCC-3' was used for the dsDNA synthesis. To 40μl of cDNA, 600 ng of primer-adapter, 6μl of 2.5m dNTP, 6μl 10x LA PCR buffer and 15 units of LA Taq (Takara, Shiga, Japan) were added in a final volume of 60 μl. The reaction was performed on a MJ thermal cycler (MJ Research, Waltham, MA, USA) by initially denaturing for 5min at 65°C followed by negative ramp of -20°C, with slope of 0.1°C/second, two cycles of 45°C for 10 min, 68°C for 20 min and 72°C for 10min. dsDNA was then restricted with 25 of the restriction enzymes *Sst*I and *Xho*I (for 1 microgram of ds cDNA, scaled down based on the quantity) at 37°C for 1h. The digested ds DNA was passed through a MicroSpin S-400 HR Column (Amersham Biosciences, Buckingham shire, UK), extracted once with phenol-chloroform, precipitated with ethanol, washed with 70% ethanol and resuspended in RNase-free water. The double stranded DNA was cloned into a *Xho*I and *Sst*I digested λ ZAPII vector (Figure 2-2). Packaging of ligated phage was done by phage packaging extract (Epicenter, Madison, USA) and mass excision was performed following manufacturers instructions (Stratagene, La Jolla, CA, USA). Colonies were selected from LB agar plates containing ampicillin and transferred to 96-well plates containing LB medium with 50 µg/ml ampicillin followed by cDNA sequencing. The size of cDNA inserts was evaluated by restriction analysis of clones. Plasmid DNA was prepared in 96-well plate format using multi-well filter plates (Pall Corporation, Ann Arbor, MI, USA). Sequencing of cloned cDNA inserts was performed with 200of plasmid DNA as template and M 13 forward primer (GTAAAACGACGGCCAGT) using ABI PRISM Bigdye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems, Foster City, CA, USA) according to manufacturer's instructions and analyzed on ABI 3700 automated sequencers (Applied Biosystems, Foster City, CA, USA).

2.3 <u>Construction of full-length enriched cDNA libraries by SMART method.</u>

2.3.1 *First-strand cDNA synthesis and amplification of cDNA by LD PCR.* Messenger RNA (5 µg) from spleen, neocortex, liver, kidney, cerebellum, testis, eye tissues was used for the construction of cDNA library using SMART^{TM} kit (Clontech, Mountain View, CA, USA) following manufacturer's instructions with slight modification (Chenchik et al., 1998). To 5 µg of RNA sample, SMART IV oligonucleotide and CDS III/3´ PCR Primer were incubated at 72°C for 2 min. After cooling for short time in ice, 5X first-Strand buffer, 0.1 M DTT, 10mM dNTP Mix and 150 units of PowerScript Reverse Transcriptase were added and incubated at 42°C for 1 h and the final volume was made to 20 µl with H_2O. For the preparation of second strand cDNA, 2 µl of first-strand cDNA, Advantage 2 PCR Buffer, dNTP Mix, 5´ PCR primer, CDS III/3´ PCR primer and Advantage 2 polymerase mix were added into a new pre-chilled 0.5 ml eppendorf tube and the final volume was made to 100 µl with H_2O. The PCR was done by the following program: 95°C 20S; 24 cycles of 95°C 5 S; 68°C 6 min. The PCR product (5 µl) was taken and analyzed by running on 1.1% agarose/EtBr gel alongside 1 Kb DNA marker (G & P, Korea) and the concentration of the double strand (ds) cDNA was roughly estimated.

2.3.2 *PCR product purification & SfiI digestion.* The amplified ds cDNA was purified by Strataprep PCR Purification Kit (Stratagene, La Jolla, CA, USA) following manufacturers instructions. The purified ds cDNA was restricted using 300 units of *Sfi*I enzyme (New England Biolabs, UK) and incubated at 50°C for 2 h followed by purification with MicroSpin S-400 HR Column (Amersham Biosciences, Buckingham shire, UK) and precipitated with ethanol. The pellet was finally dissolved in 5 µl deionized H_2O. One µl of solution was run on 1.1% agarose/EtBr gel alongside 0.1 µg of 1 Kb DNA size marker (G & P, Korea) at 80 V for 40 min and the concentration was determined.

2.3.3 *Ligation of cDNA to λTriplEx2 vector, packaging and titration.* Complementary DNA and λTriplEx2 vector (Clontech, Mountain View, CA, USA) (Figure 2-3) restricted with *Sfi*I were ligated at a ratio of 2:1 using T4 DNA ligase (New England Biolabs, UK). The ligated product was packaged using MaxPlax™ Packaging Extract (Epicenter, Madison, USA) following manufacturers instructions. XL1-Blue working stock plate was prepared using LB/tetracycline at a concentration of 15 µg/ml, from which a single colony was used to inoculate LB/$MgSO_4$/maltose broth to prepare the XL1-Blue overnight culture. The culture was centrifuged, the pellet was resuspended in 10mM $MgSO_4$ and the OD was brought to 0.5. Made a 1:10 dilution of each of the packaging products, from which 1 µl of the diluted phage was taken and added to 200 µl of the XL1-Blue overnight culture, and the phage was allowed to be preabsorbed at 37°c for 15 min. Three ml of melted LB/$MgSO_4$ top agar was added to each mixture of XL1-Blue and phage, and were poured onto 90 mm LB-agar/$MgSO_4$ plates pre-warmed to 37°C. The plates were inverted after solidification and incubated at 37°C for 12h. The plaques were counted and the titer of the phage (pfu/ml) was calculated. Three ml of melted LB/$MgSO_4$ top agar were added to each mixture of 200 µl of the XL1-Blue overnight culture, 1 µl of the diluted phage, 50 µl IPTG stock solutions (0.1 mol/L) and 50 µl X-gal stock solutions (0.1 mol/L) into a sterilized tube and then poured onto 90 MM LB/$MgSO_4$ plates pre-warmed to 37°c. The plates were inverted after solidification and were incubated at 37 °c for 12h. The ratio of white plaques (recombinants) to blue plaques (non-recombinants) was calculated to determine recombination efficiency.

2.3.4 *Library amplification, titration and determining the percentage of recombinant clones.* The λ lysate-packaged product was transferred into 15 ml sterilized tube with 500 µl of XL1-Blue overnight culture and incubated in a 37°C water bath for 15 min. After 4.5 ml of melted LB/$MgSO_4$ top soft agar were added into each tube, the mixture was poured onto LB/$MgSO_4$ agar plate. The plate was inverted after solidification and incubated at 37°C for 12 h and 8ml of 1x lambda dilution buffer was added to each plate. The plates were stored at 4°C overnight, and then incubated on a platform shaker at 50 rpm at room temperature for 1 h. Each of the phage lysates was mixed well and then poured into a sterile 50 ml polypropylene screw-cap micro centrifuge tube containing 10 ml of chloroform, vortexed for 2 min and centrifuged (Beckman, Germany) at 7000 rpm for 10min. The supernatant was transferred into another sterilized 50 ml centrifuge tube, and DMSO was added (final concentration 7%). The amplified library was transferred into 1.5 ml sterilized micro centrifuge tubes and stored at −70°C as stock. Ten µl of 1:100000 diluted phage and 200 µl of the XL1-Blue overnight culture were added into a 5 ml sterilized tube, and the phage was allowed to adsorb at 37°C for 15 min. Three ml of melted LB/$MgSO_4$ top soft agar were added into the tube, and the mixture was poured onto a

90 mm LB/$MgSO_4$ plate preheated to 37°C. After solidification, the plate was inverted and incubated at 37°C for 12 h. The plaques were counted and the titer of the phage (pfu/ml) was calculated. The percentage of recombinant clones in amplified library was determined in the same way as that used for unamplified library.

2.3.5 ***Cre-lox mediated excision and retransformation.*** The clones were excised by Cre-lox mediated excision followed by mass plasmid isolation. The plasmids were retransformed into *Escherichia coli* DH10B cells (Invitrogen, Carlsbad, CA, USA) and plated on to LB plates containing ampicillin (50μg/ml). The recombinant clones were determined by x-gal staining. The white colonies from the plates were picked and transferred to 96-well plates and grown for 24 h in LB containing 50 μg/ml ampicillin.

2.3.6 ***Complementary DNA sequencing of the clones from the SMART library with λ triplex2 vector.*** Plasmid DNA was prepared in 96-well plate format using multi-well filter plates (Pall Corporation, Ann Arbor, MI, USA). Sequencing of cloned cDNA inserts was performed with 200of plasmid DNA as template and (TCCGAGATCTGGACGAGC) Tex2 primer using ABI PRISM Bigdye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems, Foster City, CA, USA) according to manufacturer's instructions and analyzed on ABI 3700 automated sequencers (Applied Biosystems, Foster City, CA, USA).

2.3.7 ***Ligation of double strand cDNA to pDNR-Lib.*** Double strand cDNA obtained from SMART method and pDNRLib vector restricted with *Sfi*I (Clontech, USA) (Figure 2-4) were ligated at a molar ratio of 2:1 using Ligation Mighty Mix (Takara, Shiga, Japan). Sequencing of clones was performed in a similar manner as for pTriplex and the primer used for sequencing from 5' end was M13 F (GTAAAACGACGGCCAGT)

2.3.8 *complementary DNA sequencing of the clones from the SMART library with pDNR-Lib vector.* Plasmid DNA was prepared in 96-well plate format using multi-well filter plates (Pall Corporation, Ann Arbor, MI, USA). Sequencing of cloned cDNA inserts was performed with 200of plasmid DNA as template and Tex2 primer (TCCGAGATCTGGACGAGC) using ABI PRISM Bigdye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems, Foster City, USA) according to manufacturer's instructions and analyzed on ABI 3700 automated sequencers (Applied Biosystems, Foster City, CA, USA).

3. RESULTS

3.1 Characteristics of the libraries constructed by modified oligo-capping method. The primary titer for the liver and brainstem libraries were 1.0x10^6 cfu (colony forming unit), respectively. The average cDNA insert size estimated by restriction analyses of 18 clones each was 1.8and 2 Kb respectively (Figure 2-5), which is larger than that of conventional libraries (Yao et al., 2004) and similar to that of full-length enriched cDNA libraries reported (Carninci et al., 1999). Based on the analysis of 92 clones from the brainstem library, the clones without inserts were only 2.2% (Table 2-2). The results of clone description by NCBI blastn analysis against GenBank nonredundant nucleotide database (nr) showed that 94% of clones from brainstem libraries showed significant homology to known genes (Table 2-2). In the brainstem library, many ESTs did not show match to the pig genes but to the other species, indicating significant portions of genes expressed from the brainstem have not been analyzed in porcine (Table 2-2). Seventy five percent of the sequences from the brainstem library had shown Expect value (E-value) "0" indicating highest significance of the match. The remaining 25% had E-values ranging from e < 10^{-20} to e < 10^{-178}. The most abundantly expressed gene for the brainstem and liver libraries were tubulin and albumin respectively. When the start codon-containing clones were considered as full length clones, the percentage of full-length clones was estimated to be 70 and 80% from the liver and brainstem libraries respectively. Majority of clones in the liver library were represented by albumin gene indicating redundancy is very high.

3.2 Characteristics of the libraries constructed by SMART method. All the libraries constructed by SMART procedure using λTriplex2 has primary library titer varying from $5x10^5$-$1.1x10^6$ pfu (plaque forming unit) or $5x10^5$ cfu when pDNR Lib vector was used respectively (Table 2-1). The most abundantly expressed gene for the spleen library was β-globin. The results of clone description by NCBI blastn analysis against GenBank non-redundant nucleotide database (nr) showed that 88 % of clones from spleen library showed significant homology to known genes (Table 2-2). Based on the presence or absence of the translation initiation site in the 5' sequences, the percentage of full-length clones was estimated to be 60% on an average. Although this is a little lower than the results obtained using the cap-trapping method (Beisel et al., 2004), the ratio of full-length clones was significantly higher than the conventional libraries and full-length cDNA libraries reported (Otsuka et al., 2003; Chen et al., 2006). The average cDNA insert size estimated by restriction analysis of 20 clones from the libraries constructed by SMART procedure varied from 0.9 Kb to 1.8 Kb (Figure 2-6) which is higher than that of recent report (Chen et al., 2006). Eighty four clones were sequenced from both the 5' and 3' end of neocortex library and 200 clones from 3' end of spleen library to find the presence of poly A tail in the 3' end. All the sequences contained poly A tail indicating that contamination of genomic DNA was almost absent.

3.3 Characteristics of the libraries constructed by cap-trapping procedure.

The average cDNA insert size estimated by PCR amplification of 18 clones from neocortex library was 1.1 Kb (Figure 2-7) by cap-trapping which is smaller than that of other full-length enriched cDNA libraries constructed by cap-trapping (Beisel et al., 2005) although our result was based on only one library. The primary library titer for neocortex library was 5.0x10^5 pfu.

4. DISCUSSION

A total of 11 full-length enriched cDNA libraries were constructed from Korean native pig tissues using the three widely employed methods for full-length enriched cDNA libraries construction (Table 2-1). Successful SNP detection was reported by Dirisala et al., 2005; Dirisala et al., 2006, using 3,390 sequences obtained from spleen, neocortex, brainstem and liver libraries. These 3,390 sequences are also being used for Transcriptional start site determination (Manuscript in preparation) and Radiation hybrid mapping (Manuscript in preparation). The number of sequences is likely to increase from 3,390 to 10,000 by the addition of sequences from the libraries that are being constructed and sequenced. Construction of four new libraries from four different tissues is in progress (Table 2-1). Cap-trapping method is being standardized for construction of high quality full-length enriched cDNA libraries as this method proved to be superior in comparison with other methods despite its complexity (Sugahara et al., 2001). 93% of sequences were matched with sequences of either pig or other species with E value < 10^{-100} (Table 2-3) and majority of clones from our libraries are extended towards the 5' end (Table 2-4) which are useful is the identification of 5’untranslated region (UTR).

cDNA libraries for the porcine genome have not been reported for many tissues such as eye, tongue, stomach, large intestine, kidney, lung, sperm, skin, testis, ear, etc. Since no cDNA library is reported for these tissues, construction of cDNA libraries from these tissues will improve our knowledge about expression of genes in these tissues. Also full-length enriched cDNA libraries need to be constructed from tissues that are not reported to date, so that the information generated from these full-length enriched cDNA libraries can serve as valuable resource for pig functional genomics.

Table 2-1. List of full-length enriched cDNA libraries constructed from various tissues of Korean native pigs.

S.No	Organ	Tissue Source	Methodology	Vector used	Average insert size
1	Neocortex*	P5	SMART	λTriplEx2	1.2 Kb
2	Spleen*	P5	SMART	λTriplEx2	1.1 Kb
3	Cerebellum	P24	SMART	λTriplEx2	1.7 Kb
4	Kidney	P5	SMART	λTriplEx2	1.0 Kb
5	Testis	P5	SMART	λTriplEx2	0.9 Kb
6	Liver	P9	SMART	λTriplEx2	1.6 Kb
7	Lung	P9	SMART	λTriplEx2	In progress
8	Sperm		SMART	λTriplEx2	In progress
9	Eye	P5	SMART	λTriplEx2	In progress
10	Muscle	P5	SMART	λTriplEx2	In progress
11	Brainstem*	P5	Modified oligo-capping	pCNS-D2	2.0 Kb
12	Liver*	P5	Modified oligo-capping	pCNS-D2	1.8 Kb
13	Neocortex	P5	SMART	PDNR-Lib	1.1 Kb
14	Cerebellum	P24	SMART	PDNR-Lib	0.9 Kb
15	Neocortex	P5	Cap-trapping	λZAPII	1.1 Kb

*3390 clones from these four libraries analyzed for SNP identification, transcriptional start site elucidation and radiation hybrid mapping.

Table 2-2. The characteristics of spleen and brainstem libraries constructed by SMART and modified oligo-capping methods based on blast analysis.

Category (%)	Tissues	
	Spleen	Brainstem
Annotated transcripts	278 (88%)	87 (94%)
Pig	173 (55%)	19 (22%)
Other Species	105 (33%)	68 (72%)
Homology with pig EST	6 (1.9%)	0 (0%)
No match	4 (1.2%)	2 (2.2%)
No insert	1 (0.3%)	2 (2.2%)
Mitochondrial transcripts	26 (8.2%)	1 (1.0%)
Total	315 (100%)	92 (100%)

Table 2-3. Blast analysis results of clones from brainstem, liver, neocortex and spleen full-length enriched cDNA libraries.

Category by BLAST analysis	No. of clones (%)	Extension of 5’-ends to NCBI 5’-EST sequences	No. of clones (%)
EST Match	93 (93)	Longer	79 (85)
Pig	81 (87)	Same	2 (2)
Other species	12 (13)	Shorter	12 (13)
No match	7 (7)		
Total	100 (100)	Total	100 (100)

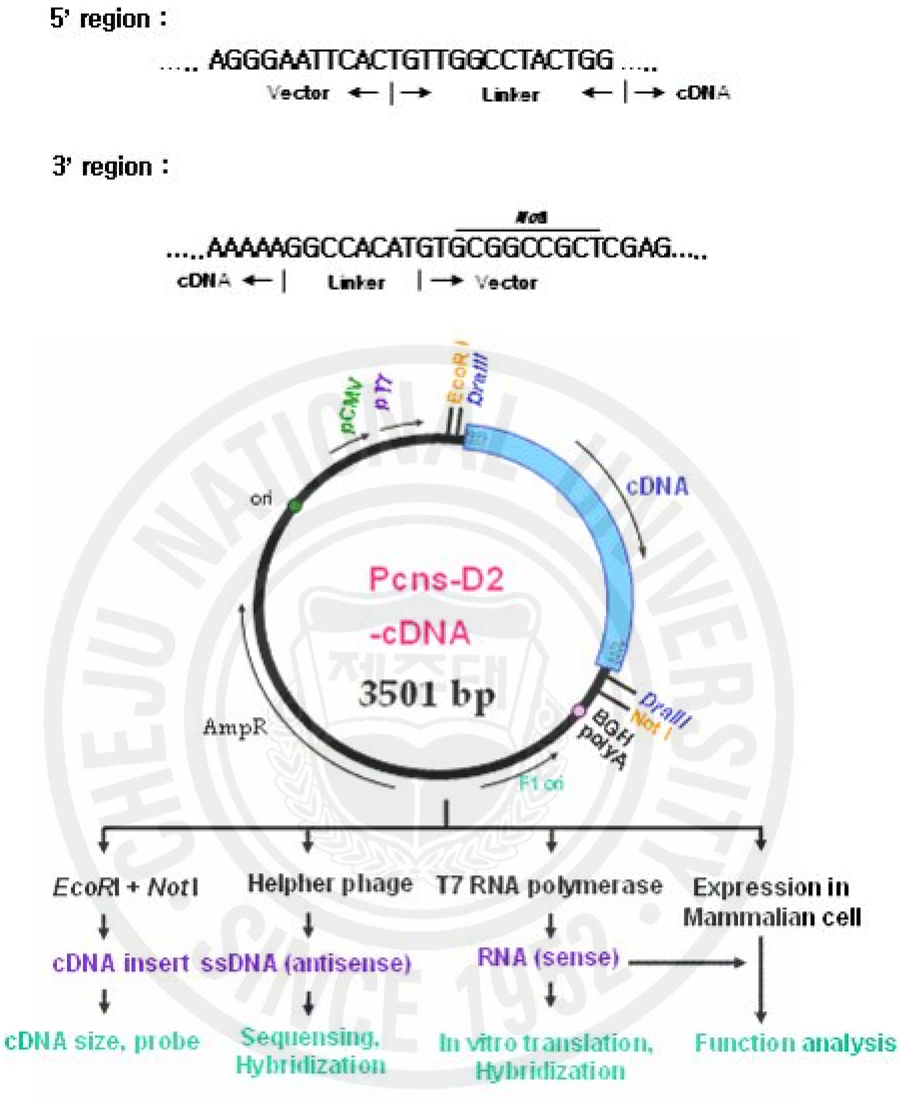
Table 2-4. Description of five newly identified 5' untranslated region sequences from the analysis of full-length enriched cDNA libraries.

GenBank accession	Definition	Sequence of 5'UTR
AY550066	Myosin regulatory light chain	5′**GAAGTGCCGGCGTCGCCGCTGTTGCTCCCGCAGTTCCTCCCGCAGCTCCGCACTCGTAGCCTCCGCTCGTTTCGCTCAGGAAGTCCGGGTTCTGGTTCTGGTATTTGGCCGCAAATTAAACTGCCACCATGTCGAGCAAAAAGGCAAAGACCAAGACCACCAAGAAGCGCCCCCAGCGCGCAACTTCCAATGTGTTTGCCATGTTTGACCAGTCACAGATTCAGGAGTTCAAGGAGGCCTTCAACATGATCGATCAG**AACAGAGATGGTT
NM_214211	Ubiquitin/ ribosomal fusion protein	5′**GGGGCTTTTTCTC**TTCAACGAGGCGGCCGAGCAGACGCAGAGATG
AY610343	Peroxiredoxin gene	5′**ACGGCCGGGGGGCCCCGAGAACGCAAGTACCT**GAGTCTTCTCGTCGGTGCGTCCCGCCCTTGCCCACGCAGCTTTCAGTCATGGCCTC
L21163	T-Cell receptor delta chain	5′**CTGTGACAGCTACGTGGACGGTGGGATACGGGACGTATCGATAAACTCATCTTTGGAAAAGGGACTCAGCTGGTTGTGGAACCACGA**AGTCAGCCTAATTCCAAACCATCCGTTTTTGTCATGAAAAATGGAACAAATGTTG
NM_214276	Citrate synthase	5′**TGGGGCAGCGGCGGCGGCAGCTCCCGTTCCT**GCCGCATTTCTCTTCCCTCCTTCCCTCCCCGCCAGATCTCCGAATTCGCCTGCCATGGCC

Note:

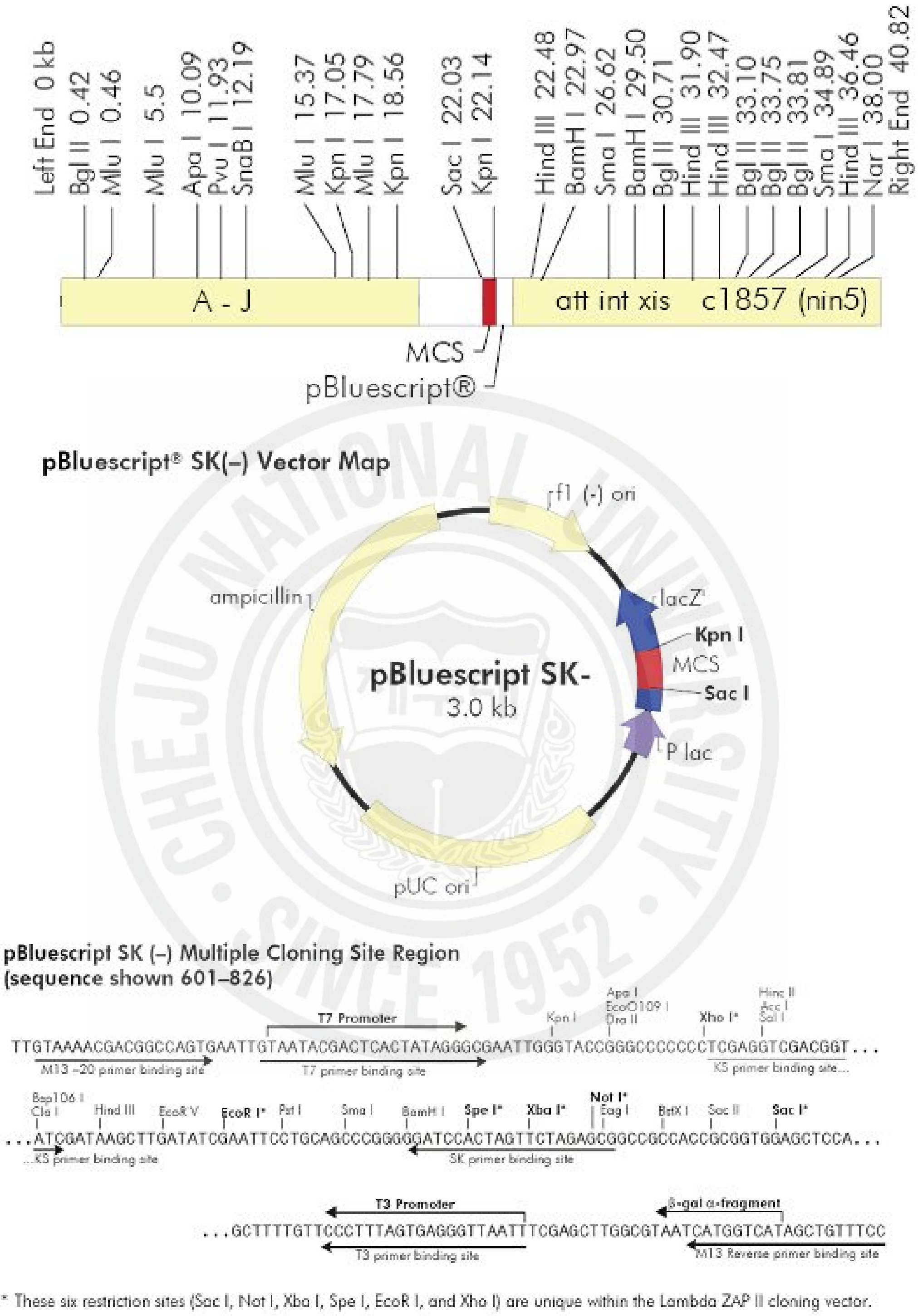
[1] Bold letters indicate the newly identified sequences comparing to the existing sequences in the GenBank.

[2] ATG, the translation initiation signal.



Multi-functional pCNS-D2 vector

Figure 2-1. Vector map describing pCNS-D2 vector (Adapted from Oh et al., 2004; Plasmid).



pBluescript® SK(–) Vector Map

pBluescript SK (–) Multiple Cloning Site Region
(sequence shown 601–826)

T7 Promoter
Kpn I
Apa I
EcoO109 I
Dra II
Xho I*
Hinc II
Acc I
Sal I
TTGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGGTACCGGGCCCCCCCTCGAGGTCGACGGT . . .
M13 –20 primer binding site
T7 primer binding site
KS primer binding site...
Bsp106 I
Cla I
Hind III
EcoR V
EcoR I*
Pst I
Sma I
BamH I
Spe I*
Xba I*
Not I*
Eag I
BstX I
Sac II
Sac I*
. . . ATCGATAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATCCACTAGTTCTAGAGCGGCCGCCACCGCGGTGGAGCTCCA. . .
...KS primer binding site
SK primer binding site
T3 Promoter
β-gal α-fragment
. . . GCTTTTGTTCCCTTTAGTGAGGGTTAATTTCGAGCTTGGCGTAATCATGGTCATAGCTGTTTCC
T3 primer binding site
M13 Reverse primer binding site

* These six restriction sites (Sac I, Not I, Xba I, Spe I, EcoR I, and Xho I) are unique within the Lambda ZAP II cloning vector.

Figure 2-2. Vector map of λ ZAPII vector and excised pBluescript SK- vector (Adapted from λ ZAPII undigested vector kit user manual from Stratagene).

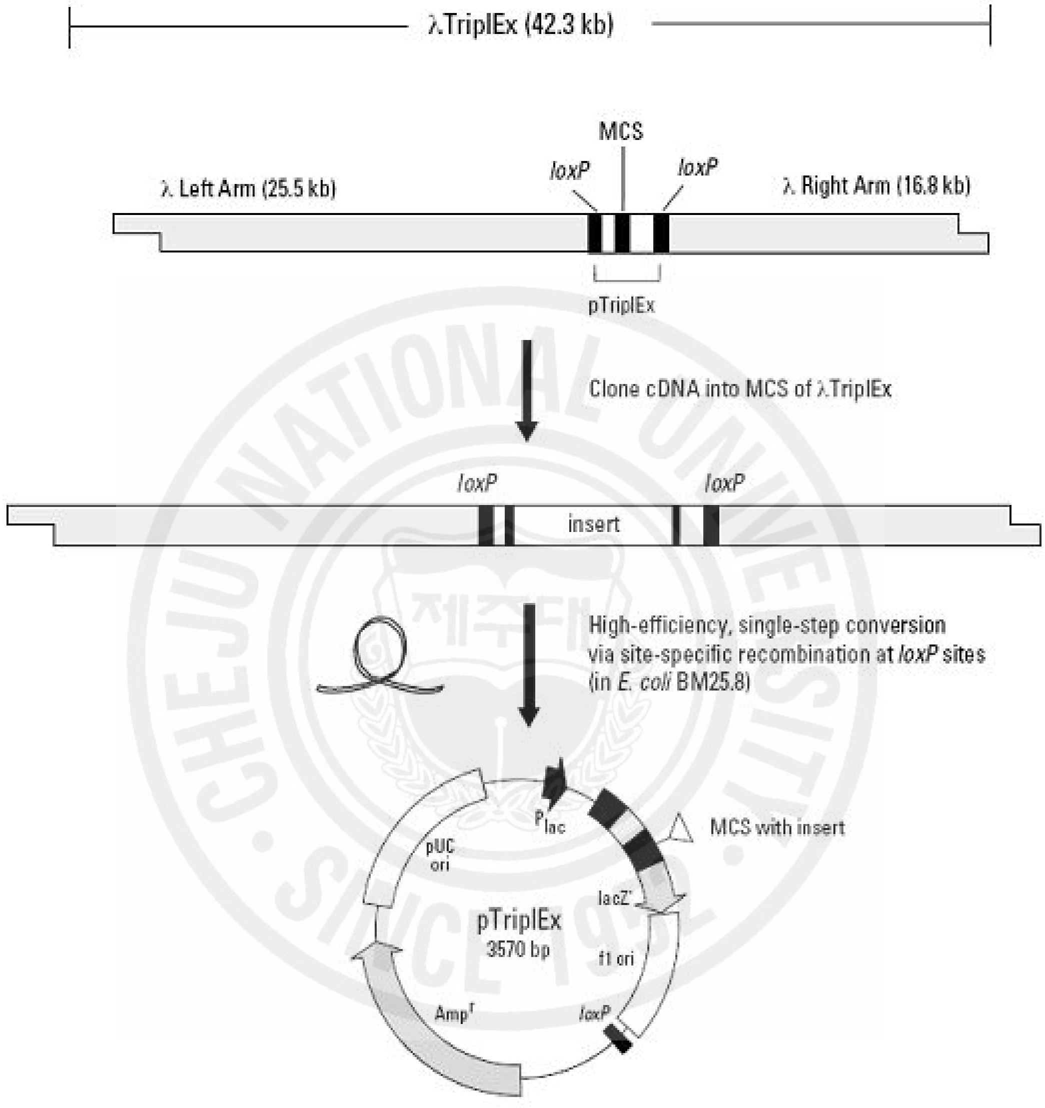


Figure 2-3. Conversion of a recombinant λ TriplEx2 to the corresponding pTriplEx2. λ TriplEx2 restricted with *Sfi*I was used as vector for cloning double stranded cDNA obtained by SMART procedure.(Adapted from Clontech SMART cDNA library construction kit user manual).

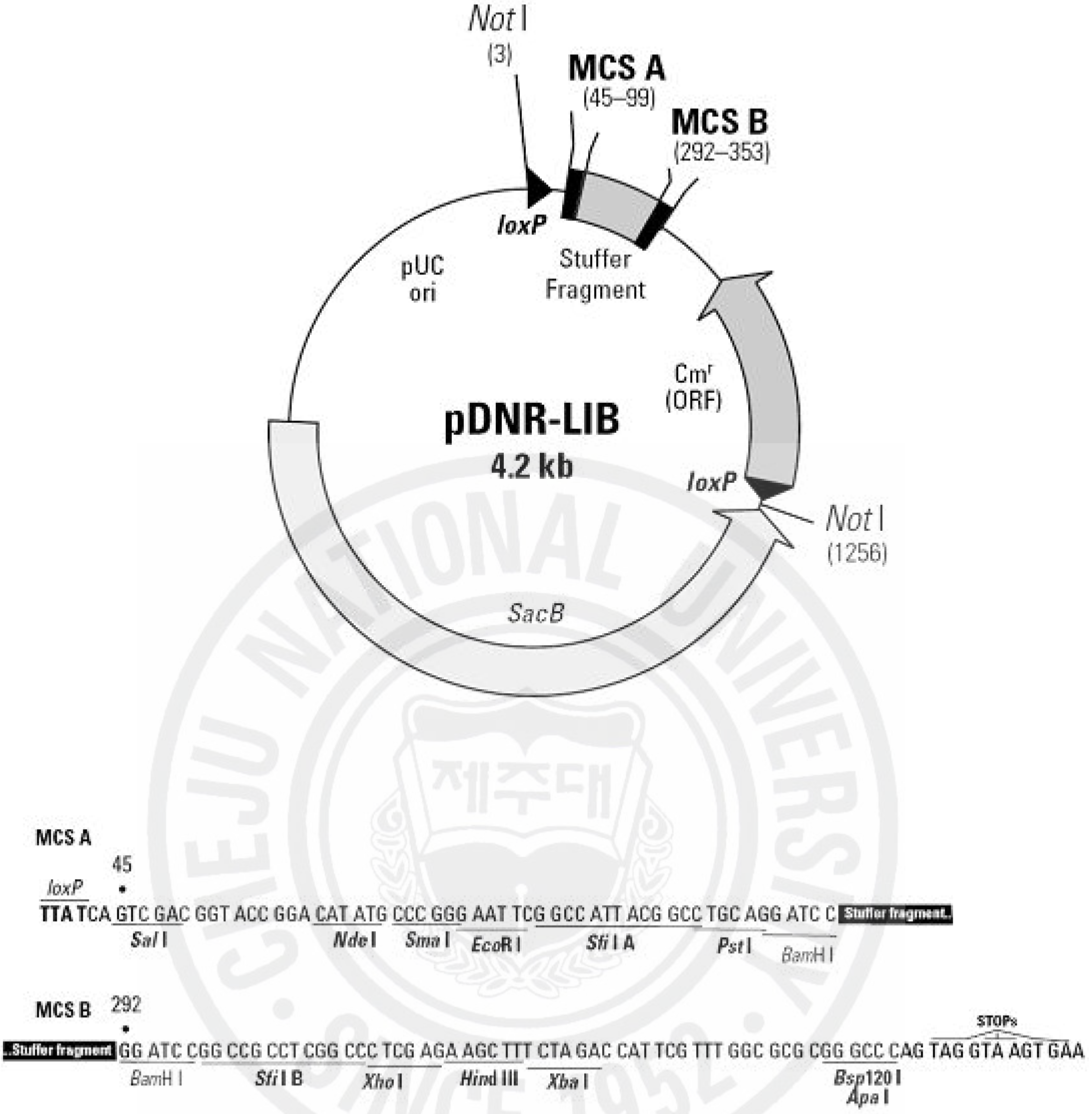

Not I
(3)
MCS A
(45–99)
MCS B
(292–353)
loxP
pUC
ori
Stuffer
Fragment
Cm^r
(ORF)
pDNR-LIB
4.2 kb
loxP
Not I
(1256)
SacB
MCS A
45
loxP
TTA TCA GTC GAC GGT ACC GGA CAT ATG CCC GGG AAT TCG GCC ATT ACG GCC TGC AGG ATC C Stuffer fragment..
Sal I
Nde I
Sma I
EcoR I
Sfi I A
Pst I
BamH I
MCS B
292
..Stuffer fragment GG ATC CGG CCG CCT CGG CCC TCG AGA AGC TTT CTA GAC CAT TCG TTT GGC GCG CGG GCC CAG TAG GTA AGT GAA
STOPs
BamH I
Sfi I B
Xho I
Hind III
Xba I
Bsp120 I
Apa I

Figure 2-4. Restriction map and MCS of pDNR-LIB vector (pDNR-Lib vector restricted with *SfiI* was used for cloning double stranded cDNA obtained by SMART procedure) (Adapted from pDNR-LIB vector kit user manual from Clontech).

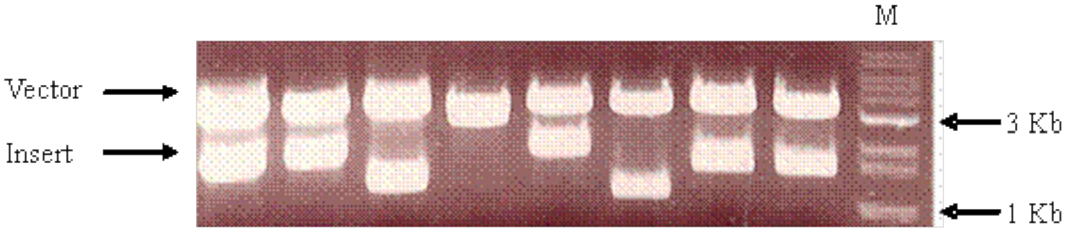


Figure 2-5. Restriction analysis of clones from full-length enriched brainstem library constructed by modified oligo-capping method (M=Molecular size marker).

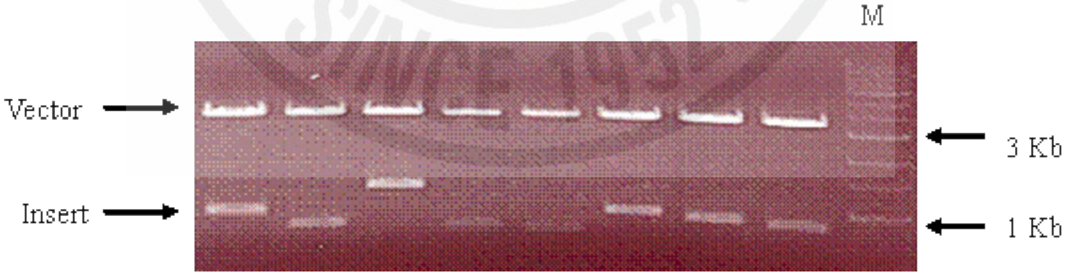


Figure 2-6. Restriction analysis of clones from full-length enriched cerebellum cDNA library constructed by SMART method (M=Molecular size marker).

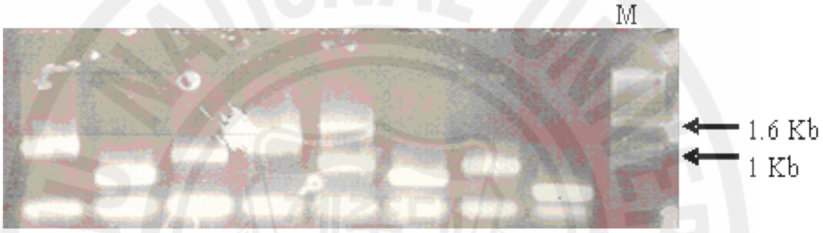


Figure 2-7. PCR amplification of inserts from clones of full-length enriched neocortex cDNA library constructed by cap-trapping (M=Molecular size marker).

REFERENCES

Beisel, K. W.,Shiraki, T., Morris, K. A., Pompeia, C., Kachar, B., Arakawa, T., Bono, H., Kawai, J., Hayashizaki, Y. and Carninci, P 2004. Identification of unique transcripts from a mouse full-length subtracted inner ear cDNA library. Genomics 83:1012-1023.

Carninci, P., Westover, A., Nishiyama, Y., Ohsumi, T., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Schneider, C. and Hayashizaki, Y. 1997. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. DNA Res. 4:61–66.

Clepet, C., Clainche, I. L. and Caboche, M. 2004. Improved full-length cDNA production based on RNA tagging by T4 DNA ligase. Nucleic Acids Res. 32:e6.

Chen, C. H., Lin, E. C, Cheng, W. T. K., Sun, H. S., Mersmann, H. J. and Ding, S. T. 2006. Abundantly expressed genes in pig adipose tissue: An expressed sequence tag approach. J. Anim. Sci. 84:2673-2683.

Chenchik, A., Zhu, Y. Y., Diatchenko, L., Li, R., Hill, J. and Siebert, P. D. 1998. Generation and use of high-quality cDNA from small amounts of total RNA by SMART PCR. Bio Techniques Books, Natrick, MA. p. 305-320.

Dirisala, V. R., Kim, J., Park, K., Kim, N., Lee, K. T., Oh, S. J., Oh, J. H., Kim, N. S., Um, S. J., Lee, H. T., Kim, K. I. and Park, C. 2005. cSNP mining from full-length enrichedcDNA libraries of the Korean native pig. Korean J. Genetics 27:329-335.

Dirisala, V. R., Kim, J., Park, K., Lee, H. T and Park, C. 2006. cSNP mining from full-length enriched cDNA libraries of the Korean native pig. 30th International Conference on Animal Genetics, Porto Seguro, BA., Brazil.

Edery, I., Chu, L. L., Sonenberg, N. and Pelletier, J. 1995. An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). Mol. Cell. Biol. 15:3363–3371.

Frohman, M. A., Dush, M. K. and Martin G. R. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. 85:8998-9002.

Kato, S., Ohtoko, K., Ohtake, H. and Kimura, T. 2005. Vector-capping: a simple method for preparing a high-quality full-length cDNA library. DNA Res. 12:53-62.

Kim, T. H., Kim, N. S., Lim D., Lee, K. T., Oh, J. H., Park, H. S., Jang, G. W., Kim, H. Y., Jeon, M., Choi, B. H., Lee, H. Y., Chung, H. Y. and Kim H. 2006. Generation and analysis of large-scale expressed sequence tags (ESTs) from a full-length enriched cDNA library of porcine backfat tissue. BMC Genomics 7:36.

Mikawa, A., Suzuki, H., Suzuki, K., Toki, D., Uenishi, H., Awata, T. and Hamashima, N. 2004.Characterization of 298 ESTs from porcine back fat tissue and their assignment to the SSRH radiation hybrid map. Mamm. Genome 15:315-322.

Oh, J. H., Kim, Y. S. and Kim, N. S. 2003. An improved method for constructing a full-length enriched cDNA library using small amounts of total RNA as a starting material. Exp. Mol. Med. 35:586-590.

Otsuka, M., Arai, M., Mori, M., Kato, M., Kato, N., Yokosuka, O., Ochiai, T., Takiguchi, M., Omata, M. and Seki, N. 2003. Comparing gene expression profiles in human liver, gastric and pancreatic tissues using full-length enriched cDNA libraries. Hepatol. Res. 27:76-82.

Sugahara,Y., Carninci, P., Itoh, M., Shibata, K., Konno, H., Endo, T., Muramatsu, M. and Hayashizaki, Y. 2001. Comparative evaluation of

5′-end sequence quality of clones in CAP trapper and other full-length cDNA libraries. Gene 263:93-102.

VanWijk, H. J., Arts, D. J., Matthews, J. O., Webster, M., Ducro, B. J. and Knol, E. F. 2005. Genetic parameters for carcass composition and pork quality estimated in a commercial production chain. J. Anim. Sci. 83:324-333.

Yao, J., Ren, X., Ireland, J. J., Coussens, P. M., Smith, T. P. L. and Smith G. W. 2004. Generation of a bovine oocyte cDNA library and microarray: resources for identification of genes important for follicular development and early embryogenesis. Physiol. Genomics 19:84-92.

Zhu, Y. Y., Machleder, E. M., Chenchik, A. C., Li, R. and Siebert, P. D. 2001. Reverse transcriptase template-switching: A SMART approach for full-length cDNA library construction. BioTechniques 30:892-897.

Chapter III.

PIG cSNP DISCOVERY USING FULL-LENGTH ENRICHED CDNA LIBRARIES

1. INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most frequent form of DNA variations in the genome of any organism (Brookes, 1999). Owing to their high abundance and stability in populations, they emerged as the marker of choice in several applications that include physical mapping, evolutionary studies and association genetics (Syvanen, 2001; Vignal et al., 2002; Dimmic et al., 2005). The most classical way to identify SNPs is by direct sequencing of amplicons of candidate genes from a set of individuals that represent the diversity in the population of interest. The drawback of this approach is time and money required (Useche et al., 2001).

An alternative method takes advantage of the redundancy of gene sequences generated by expressed sequence tags (ESTs) at minimal costs (Gu et al., 1998; Elahi et al., 2004). EST sequence data provides the richest source of biologically useful SNPs due to the high redundancy of gene sequences and the diversity of genotypes represented within databases (Barker et al., 2002). Each SNP would also be associated with an expressed gene. The use of EST information for the detection of SNPs in the genomes of mammals has been carried out by several groups (Kwok et al., 1994; Gu et al., 1998; Buetow et al., 1999; Garg et al., 1999; Marth et al., 1999; Picoult-Newberg et al., 1999; Irizarry et al., 2000; Cox et al., 2001 Fahrenkrug et al., 2002; Hu et al., 2002; Kim et al., 2003; Fitzsimmons et al., 2004; Guryev et al., 2004; Hawken et al., 2004; Zimdahl et al., 2004; Lee et al., 2006).

Many conventional and full-length enriched cDNA libraries have been constructed in porcine and as of December 2006, 641857 porcine ESTs are available in the Genbank porcine EST database. Although the number of ESTs has been significantly increased, the analysis made less progress comparing to the amount of data. We evaluated the process of cSNP discovery using full-length enriched cDNA libraries as a method and Korean native pigs as a divergent genetic resource. We showed that the result of SNP identification using our approach was relatively accurate. We also listed 13 new cSNPs with their allele frequencies for four breeds of pigs.

2. MATERIALS AND METHODS

2.1 SNP identification. The chromatograms of 3390 EST sequences obtained by sequencing spleen, neocortex, brainstem and liver libraries which have been described in chapter 2 were transferred to a Linux workstation. EST sequence trace files were base-called using Phred (Ewing and Green, 1998; Ewing et al., 1998) and the trace files were assessed for having the Phred quality score higher than 30 in 20 bp windows. Sequences that met the criteria were subjected to Cross_match (P. Green, Unpublished) for vector trimming. RepeatMasker was used for removing repeat sequences (http://repeatmasker.genome.washington.edu/; A.F.A. Smit and P. Green, Unpublished). The processed sequences were assembled by Phrap (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html) and viewed with Consed program (Gordon et al., 1998). The amino acid change at the SNP site is determined by NCBI blastx (http://www.ncbi.nlm.nih.gov/BLAST/) and Nucleotide Amino Acid Alignment Program (NAP) (http://athena.bioc.uvic.ca).

2.2 PCR primer design. SNPs were chosen from the contigs that had BLAST (Altschul et al., 1990) matches to known genes with E < 10-100. Since genomic sequence information of porcine genes is hardly available, exon-intron structures were determined from those of closely related species (*Homo sapiens, Bos taurus, Mus musculus, Canis familiaris*) with available genomic sequences using Spidey, (www.ncbi.nlm.nih.gov/IEB/Research/Ostell/ Spidey/). Primers were designed based on the expected exon-intron structures to generate amplicons between 300 and 1500 bp long.

2.3 PCR and sequencing. Genomic DNA was isolated from the blood samples of five randomly selected individuals of each of four pig breeds (Korean native pig, Yorkshire, Duroc, Landrace) by phenol/chloroform extraction and ethanol precipitation procedures (Sambrook et al., 1989). PCR conditions and primers used for genomic DNA amplification are listed in Table 3-1. PCR amplification was carried out in 15 µl reaction mixtures containing 50 ng of genomic DNA, PCR buffer (10mM Tris-HCl pH8.3, 50 mM KCl, 1.5mM $MgCl_2$), 0.5 µM Primers, 200 µM dNTP and 0.75 units of Taq DNA polymerase (Takara Bio Inc, Shiga, Japan). The PCR conditions were 35 cycles of 94˚C for 30 sec, specific annealing temperature for each primer for 30 sec and 72˚C for 30 sec with an initial denaturation step at 94˚ C for 3 min plus a final extension at 72˚ C for 4 min. The PCR products were directly sequenced as indicated above.

2.4 Accession numbers of Korean native pig alleles. Accession numbers of sequences reported to Genbank from the current report are endozepine (DQ885192), peroxisomal enoyl coenzyme A hydratase 1 (DQ885193), vitronectin (DQ885194), neuronal and endocrine protein (DQ885195), antithrombin III (DQ885196), microsomal glutathione S-transferase (DQ885197) and β-globin gene (DQ885198).

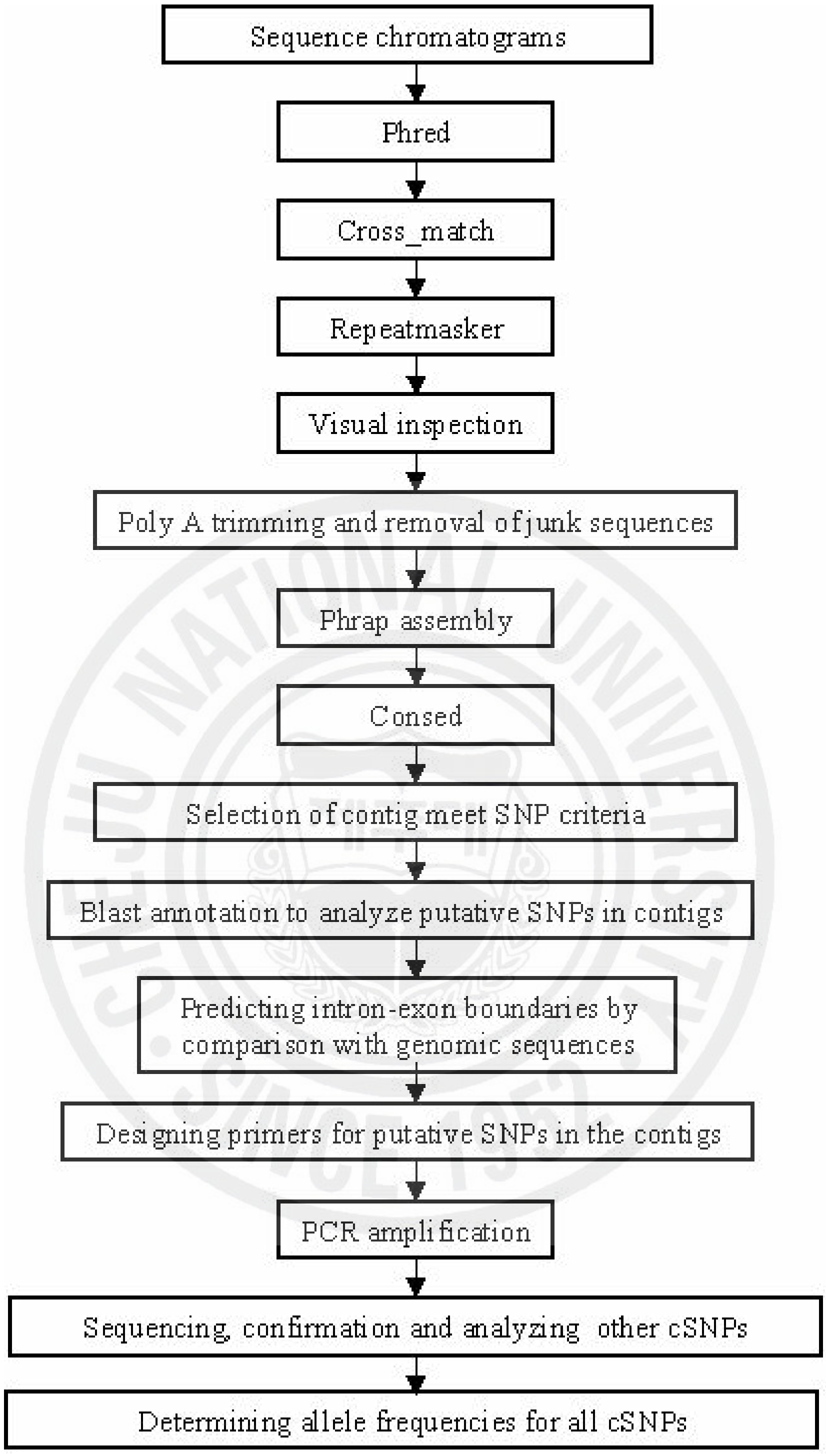

Sequence chromatograms
Phred
Cross_match
Repeatmasker
Visual inspection
Poly A trimming and removal of junk sequences
Phrap assembly
Consed
Selection of contig meet SNP criteria
Blast annotation to analyze putative SNPs in contigs
Predicting intron-exon boundaries by comparison with genomic sequences
Designing primers for putative SNPs in the contigs
PCR amplification
Sequencing, confirmation and analyzing other cSNPs
Determining allele frequencies for all cSNPs

Figure 3-1. Overview of the sequence analysis process employed for SNP detection.

Table 3-1. Summary of seven primer pairs utilized for SNP confirmation.

Gene description for contig	Genbank accession number	Forward primer reverse primer	Product size (bp)	SNP Confirmed
Peroxisomal enoyl coenzyme A	DQ157552	5’-GACATGGCTTCGGACATCTT-3’ 5’-CACAGTACCGGATGTCACAG-3’	442	Yes
Hyaluronidase	NM_213953	5’-GATGGATCAGCCGCTACTAC-3’ 5’-GTAGGTGGCACCGTGGTTGT-3’	470	Yes
Vitronectin	D61396	5’-CCTTCAGCCAGATGATGAGT-3’ 5’-GACATCTTGGATGAGCTTGG-3’	399	Yes
		5’-AGGAAGTGTCAGTGTGACGA-3’ 5’-GGTGAAGGCGGCATCAATGG-3’	611	No
Microsomal glutathione S-transferase	AY609810	5’-CTGGATTGTTGGACGAGTTC-3’ 5’-CTGGATTGTACGTGGAGTTC-3’	1200	Yes
ß-globin	AY610360	5’-GGTGTGGATTCGTCTGTATG-3’ 5’-CTCGAAGAACCTCTGAGTCC-3’	694	Yes
Neuronal and endocrine protein	M23654	5’-NNTGTGAAGTCCTGCCAGAG-3’ 5’-CCAAGCCTGGATAGTCATGT-3’	640	Yes

3. RESULTS

3.1 Identification of putative cSNPs using a bioinformatics approach. As an effort to characterize the Korean native pig genome, cSNP identification was performed between Korean native pigs and other breeds. A total of 3,390 chromatograms (3,210 5'-end sequences and 180 3'-end sequences) from four full-length enriched cDNA libraries of the Korean native pig and 50,000 chromatograms of porcine EST sequences retrieved from the GenBank trace file archive were combined and analyzed for presence of SNPs using the procedures shown in Figure 3-1.

The process yielded 8,118 contigs. Forty nine contigs were consisted of both a minimum of two Korean native pig sequences and two public EST sequences within each contig. Among these, seven contigs containing a minimum of one putative SNP from Phrap analysis were selected. Finally eight putative cSNPs were chosen for confirmation through population analysis (Table 3-2 and 3-4). Interestingly, only three of seven contigs remained putative cSNPs when Korean native pig sequences were removed (data not shown), indicating the importance of Korean native pigs as a genetic resource to increase genetic variation.

3.2 Evaluation and confirmation of cSNPs. To confirm the putative cSNPs, PCR primers were designed for the 7 loci (Table 3-1) and PCR products were generated using five different individuals from each of four different pig breeds (Duroc, Yorkshire, Landrace, Korean native pig). All primers successfully amplified, producing specific bands. Analysis of direct sequencing result from 140 samples revealed the confirmation of 6 out of 7 SNPs identified (86 % accuracy) from *in silico* SNP detection (Table 3-2).

In addition to 6 confirmed SNPs, we identified the presence of 7 additional SNPs which were unidentifiable from the *in silico* process through the sequence analysis using four breeds of 20 pigs. The SNP detection frequency from these 13 confirmed SNPs was 1 SNP per 338 bp. Transitions (92%) were more than transversions (8 %).

3.3 Characterization of breed specific differences using 13 SNPs. Allele frequencies were calculated for 13 confirmed SNPs by analyzing polymorphic variations from four pig breeds (Table 3-3). There was variation in allele frequencies of each SNP among different breeds. Interestingly, the frequency of fixed alleles was significantly higher (6of 13, 46.1%) in Yorkshire than Duroc, Landrace and Korean native pig breeds, indicating genetic homozygosity is higher in Yorkshire for these regions. In two SNPs (positions 717 and 730) from hyaluronidase, all four breeds showed polymorphism, showing that these are very polymorphic SNPs.

One SNP (position 536) from vitronectin was monomorphic in Duroc, Yorkshire and Landrace except for Korean native pigs. Contrarily, SNP 15 of β-globin was monomorphic only in Korean native pigs. These results suggest that the presence of unique genetic diversity within Korean native pig population although the allele frequency was calculated from a limited number of individuals (n=20).

In the neuronal and endocrine protein, the three nucleotide deletion was identified from an allele of the gene of Korean native pigs. This deletion removes the amino acid valine from the secretogranin domain of the encoded protein (Figure 3-2). It is interesting to more precisely know the frequency of deletion allele in the Korean native pig population since the finding was only based on five animals.

4. DISCUSSION

Our primary goal for this study was to test the efficiency of cSNP identification using Korean native pigs as a resource for genetic diversity. Since the genetic relationship between Korean native pigs and other commonly used pig breeds in genetic analysis are expected to be distantly related, the possibility of identifying genetic differences between them could be higher than comparing the sequences among commonly used breeds.

In humans, the SNP detection frequency was reported to be 1 SNP per 700 bp on an average across the whole genome (The International SNP Map Working Group, 2001) and in some regions it was as high as 1 SNP per 300 bp (Wang et al., 1998; Dawson, 1999). In chicken the frequency of putative SNPs was reported to be 1 SNP per 2,119 bp (Kim et al., 2003) or 1 SNP per 1,900 bp (Fitzsimmons et al., 2004). In Hanwoo cattle, the SNP detection frequency was 1 SNP per 787 bp (Lee et al., 2006). The difference in SNP detection frequency from the reports might come from the genetic diversity within species or applied methods.

In porcine, the SNP detection frequency was 1 SNP per 525 bp when the *in silico* method was used (Grapes et al., 2006). When the sequencing method was used for SNP identification, 1 SNP per 183 bp was identified from MARC porcine reference population, which consisted of multiple pig breeds including Chinese breeds (Fahrenkrug et al., 2002). In our study, the *in silico* method yielded 1 SNP per 614 bp. When the amplicons were sequenced to confirm the putative SNPs, the SNP detection frequency was increased to 1 SNP per 338 bp due to the discovery of new SNPs, suggesting that use of Korean native pigs for discovery of SNPs is an efficient method. Increasing the number of sequences from Korean native pig cDNA library should greatly increase the finding of polymorphisms associated with Korean native pigs.

The SNP confirmation percentage in humans varied between 50-82% (Gu et al., 1998; Buetowet al., 1999; Picoult-Newberg et al., 1999; Irizarry et al., 2000). When our putative SNPs from *in silico* analysis were analyzed for confirmation using genomic DNA, the validation percentage was higher (83%) than other studies using pigs which ranged from59~64 % (Kollers et al., 2005; Grapes et al., 2006). This shows that the accuracy of our analysis method is in good agreement with others and even better. These differences in validation percentage may be due to the different data filtering methods used to discern sequence errors from putative SNPs. In the EST-SNP approach, some mismatches are probably attributed to base calling errors or errors produced during cDNA synthesis and propagation in E.coli (Cooper and Krawczak, 1995).

To reduce these potential errors, we have only taken sequences with the Phred quality value higher than 30, which was more stringent than the value chosen for earlier reports (Picoult-Newberg et al., 1999; Fahrenkrug et al., 2002; Cheng et al., 2004; Zimdahl et al., 2004, Lee et al., 2006). For SNP confirmation we had chosen only those contigs with more than two Genbank sequences and two Korean native pig sequences containing SNPs in them. These factors might have contributed to high confirmation percentage in comparison with that of previous reports. However, use of higher Phred quality value may also increase the possibility of missing valid SNPs by applying more stringent filtering criteria than usual.

Our data suggest that the SNP detecting strategy utilizing the current NCBI pig trace archive only has some limitation since we found 7 additional SNPs from 6 loci. through sequence analysis from 4 breeds of 20 animals. However, with the progress of recently initiated pig genome sequencing project, the pig sequence trace archive will be increased significantly in short period of time. Thus, we expect that the power of SNP detecting analysis using the future pig sequence data in NCBI will be much improved.

The amino acid valine deletion which was identified in the sequence of pig neuronal and endocrine protein was also found from other species including human (Mbikay et al., 2001). The three nucleotide-deletion is due to alternative-splicing mechanism by NAGNAG acceptor. It has been reported that splice acceptors with the genomic NAGNAG motif may cause NAG insertion-deletion in transcripts (Hiller et al., 2004, 2006; Akerman and Mandel-Gutfreund, 2006). According to the analysis of the human orthologue, it was proved to be the case (Figure 3-2) and we could conform that the similar mechanism is operating in pigs.

Unlike humans and mice, pigs do not have a large repository of identified SNPs. The availability of such information is still in progress. Although we tried to confirm all the putative SNPs identified in our study, it was not possible to design primers for three putative SNPs (Table 4) due to the location of cSNPs at the start of the exon or the presence of a very long intron (>1500 bp). Although some cSNPs identified from our pipeline were not confirmed, we expect that they represent true SNPs considering our high validation percentage from the confirmed SNPs.

Several researchers suggested the necessity of more genetic markers for the dissection of economically important traits in pigs (Fahrenkrug et al., 2002; Kollers et al. 2005; Jeon et al., 2006). Our libraries can serve as an excellent genetic resource to identify cSNPs and can contributE significantly to pig genome analysis.

Table 3-2. Description of 13 confirmed cSNPs.

Description	Acc. No.	Location[1]	Sequence variation	Bp Pos.[2]
Peroxisomal Enoyl Coenzyme A hydratase 1	DQ157552	Exon 3	*TCCAT**AAC**^{asn}CTCAT → TCCAT**AAT**^{asn}CTCAT	422
		Exon 3	TCAGC**ATC**^{ile}ATCGA → TCAGC**GTC**^{val}ATCGA	453
Hyaluronidase	NM_213953	Exon 9	ACTAC**TTC**^{phe}ATGTC →ACTAC**TTT**^{phe}ATGTC	717
		Exon 9	GCCCG**GGC**^{gly}AGACC → GCCCG**GAC**^{asp}AGACC	730
		Exon 11	*CCCAC**CAT**^{his}GGGGA → CCCAC**CGT**^{arg}GGGGA	764
Vitronectin	D61396	Exon 3	GCTGC**ACT**^{thr}GACTA → GCTGC**GCT**^{ala}GACTA	228
		Exon 3	GCTGC**ACT**^{thr}GACTA → GCTGC**ACC**^{thr}GACTA	230
		Exon 4	AAGTG**ACT**^{thr}CGCGG → AAGTG**ACC**^{thr}CGCGG	263
		Exon 7	*TCCGA**GGG**^{gly}CTGTA → TCCGA**GGA**^{gly}CTGTA	536
Microsomal glutathione S-transferase	AY609810	Exon 5	*GCAAG**CGA**^{arg}AGTCG → GCAAG**CGG**^{arg}AGTCG	620
ß-globin	AY610360	Exon 1 (5' UTR)	AACTG**C**ACAAACAT →AACTG**A**ACAAACAT	15
		Exon 1	*CTGCT**GAG**^{glu}GAGAA → CTGCT**GAA**^{glu}GAGAA	49
Neuronal and endocrine protein	M23654	Exon 4	*AACAG**TA**^{val}**G**ATGGAT→AACAG---ATGAT	486

[1] The position was based on human genomic sequences except for ß-globin, which is based on the porcine genomic sequence.

[2] Bp position was based on the nucleotide position of cDNA sequences.

*SNPs identified from the *in silico* analysis.

Table 3-3. Allele frequencies of 13 confirmed cSNPs in four pig breeds.

Description	Bp position[1]	SNP	No. of animal for each genotype[2]				Allele Frequency[3]			
			D	L	Y	K	D	L	Y	K
Peroxisomal enoyl coenzyme A	422	T/C	2, 2, 1	2, 2, 1	2, 3, 0	2, 1, 2	0.6	0.6	0.7	0.5
	453	G/A	1, 3, 1	4, 1, 0	5, 0, 0	2, 1, 2	0.5	0.9	1.0	0.5
Hyaluronidase	717	T/C	3, 0, 2	2, 1, 2	3, 0, 2	4, 0, 1	0.6	0.5	0.6	0.8
	730	G/A	0, 2, 3	1, 3, 1	4, 1, 0	1, 4, 0	0.2	0.5	0.9	0.6
	764	G/A	5, 0, 0	1, 3, 1	1, 3, 1	0, 5, 0	1.0	0.5	0.5	0.5
Vitronectin	228	G/A	5, 0, 0	1, 4, 0	0, 5, 0	0, 5, 0	1.0	0.6	0.5	0.5
	230	T/C	4, 0, 1	3, 1, 1	5, 0, 0	5, 0, 0	0.8	0.7	1.0	1.0
	263	T/C	2, 3, 0	5, 0, 0	5, 0, 0	5, 0, 0	0.7	1.0	1.0	1.0
	536	G/A	5, 0, 0	5, 0, 0	5, 0, 0	2, 1, 2	1.0	1.0	1.0	0.5
Microsomal glutathione S-transferase	620	C/T	1, 2, 2	5, 0, 0	0, 2, 3	5, 0, 0	0.4	1.0	0.2	1.0
ß-globin	15	C/A	4, 1, 0	1, 2, 2	4, 0, 1	5, 0, 0	0.9	0.4	0.9	1.0
	49	G/A	3, 0, 2	5, 0, 0	5, 0, 0	1, 0, 4	0.6	1.0	1.0	0.2
Neuronal and endocrine protein	486	TAG/---	5, 0, 0	5, 0, 0	5, 0, 0	4, 0, 1	1.0	1.0	1.0	1.0
Frequency of fixed allele (%)							30.8	38.5	46.1	30.8

[1]Bp position was based on the nucleotide position of cDNA sequences.

[2]In each breed, the first (left allele) and third (right allele) are homozygous genotypes. The middle is for the heterozygous genotype.

[3]The values were calculated for the left nucleotide of SNP. D, Duroc; L, Landrace; Y, Yorkshire; K, Korean native pig.

Table 3-4. The list of putative cSNPs identified from our *in silico* analysis without validation.

Locus	Genbank accession number	Contig length	Bp pos. in cDNA	Location*	Sequence variation
Endozepine	NM_214119	514	225	Exon 2	GGAAT**GGG**^{gly}CTGAA→GGAAT**GGC**^{gly}CTGAA
Antithrombin III	AF281653	1055	43 105	Exon 1 (5'UTR) Exon 1	GCCCA**G**ACCTG→GCCCA**C**ACCTG AAAGG**ACG**^{thr}GAGT→AAAGG**AGG**^{arg}GAGTG

*The position was based on human genomic sequences.

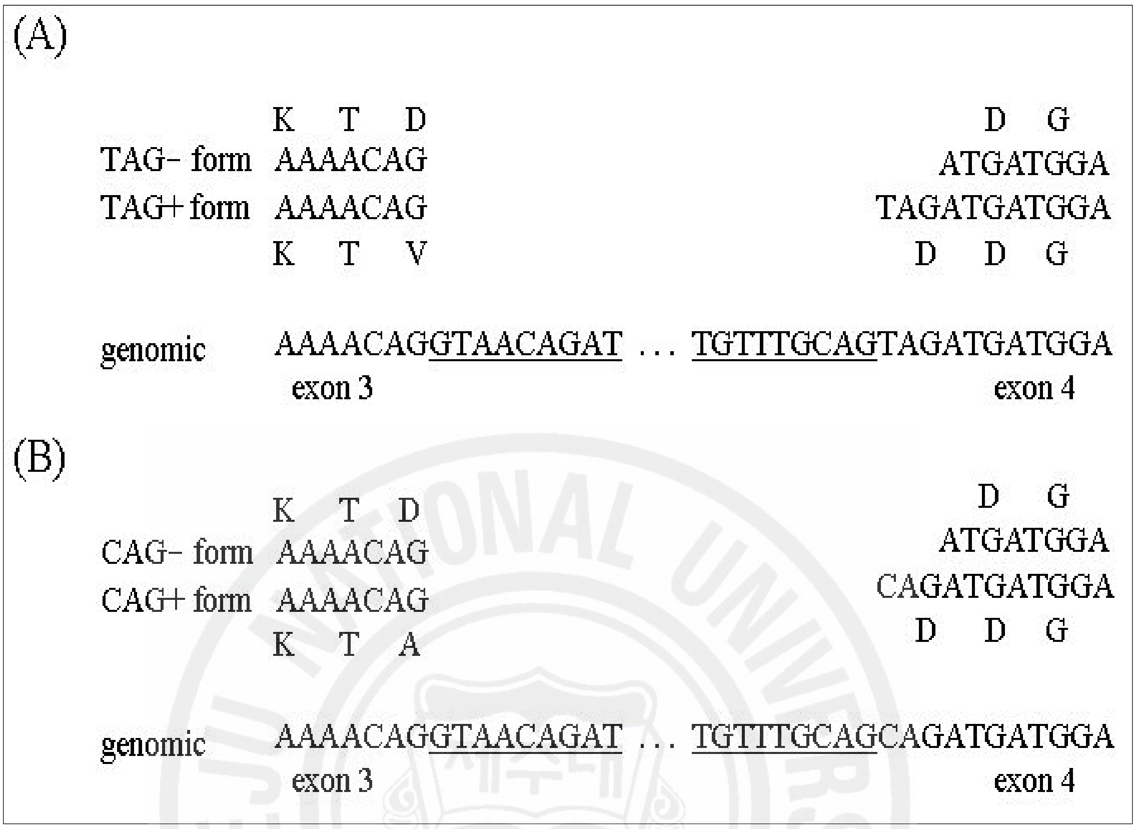


Figure 3-2. Comparison of the valine and alanine deletion alleles of porcine neuronal and endocrine protein. The TAG- form (ACC. No. CJ011171) and the TAG+ form (Acc. No. BW971409) indicate the valine deletion and the wildtype alleles, respectively. (A) The deleted three nucleotides in pig were TAG at the beginning of exon 4, resulting in a deletion of valine. (B) The deleted three nucleotides in human were CAG., resulting in a deletion of alanine. The Genbank accession numbers for CAG- form and CAG+ are NM_003020 and BC005349, respectively. The accession number for the human genomic sequence is NC_000015.

REFERENCES

Akerman, M. and Mandel-Gutfreund, Y. 2006. Alternative splicing regulation at tandem 3' splice sites. Nucleic Acids Res. 34:23-31.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Barker, G., Batley, J., Sullivan, H. O., Edwards, K. J. and Edwards, D. 2002. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421-422.

Brookes, A. J. 1999. The essence of SNPs. Gene. 234:177-186.

Buetow, K. H., Edmonson, M. N. and Cassidy, A. B. 1999. Reliable identification of large number of candidate SNPs from public EST data. Nat. Genet. 21:323-325.

Cooper, D. N. and Krawczak, M. 1995. *Human gene mutation,* Bios scientific publishers, UK.

Cox, D. G., Boillot, C. and Canzian, F. 2001. Data mining: Efficiency of using sequence databases for polymorphic discovery. Hum. Mutat. 17:141-150.

Dawson, E. 1999. More molecular markers needed? Mol. Med. Today 5:419-420.

Dimmic, M. W., Sunyaev, S. and Bustamante, C. 2005. Inferring SNP function using evolutionary, structural and computational methods. Pac. Symp. Biocomput. 10:382-384.

Dirisala, V. R., Kim, J., Park, K., Kim, N., Lee, K. T., Oh, S. J., Oh, J. H., Kim, N. S., Um, S. J., Lee, H. T., Kim, K. I. and Park, C. 2005. cSNP mining from full-length enriched cDNA libraries of the

Korean native pig. Korean J. Genetics 27:329-335.

Elahi, E., Jochen, K. and Mostafa, R. 2004. Global genetic analysis. J. Biochem. Mol. Biol. 37:11-27.

Ewing, B. and Green, P. 1998. Base calling of automated sequencing tracers using phred. II. Error probabilities. Genome Res. 8:186-194.

Ewing, B., Hillier, L., Wendl, M. and Green, P. 1998. Base calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175-185.

Fahrenkrug, S. C., Freking, B. A., Smith, T. P. L., Rohrer, G. A. and Keele, J. W. 2002a. Single nucleotide polymorphism (SNP) discovery in porcine expressed genes. Anim. Genet. 33:186-195.

Fahrenkrug, S. C., Smith, T. P., Freking, B. A., Cho, J., White, J., Vallet, J., Wise, T., Rohrer, G., Petea, G., Sultana, R., Quackenbush, J. and Keele, J. W. 2002b. Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. Mamm. Genome 13:475-478.

Fitzsimmons, C. J., Savolainen, P., Amini, B., Hjalm, G., Lunderberg, J. and Andersson, L. 2004. Detection of sequence polymorphisms in redjungle fowl and white leghorn ESTs. Anim. Genet. 35:391-396.

Garg, K., Green, P. and Nickerson, D. A. 1999. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res. 9:1087-1092.

Gordon, D., Abajian, C. and Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195-202.

Grapes, L., Rudd, S., Fernando, R. L., Megy, K., Rocha, D. and Rothschild, M. F. 2006. Prospecting for pig single nucleotide polymorphisms in the human genome: Have we struck gold. J. Anim. Breed. Genet. 123:145-151.

Gu, Z., Hillier, L. and Kwok, P. Y. 1998. Single-nucleotide polymorphism hunting in cyberspace. Hum. Mutat. 12:221-225.

Guryev, V., Berezikov, E., Malik, R., Plasterk, R. H. and Cuppen, E. 2004. Single nucleotide polymorphisms associated with rat expressed sequences. Genome Res. 14:1438-1443.

Hawken, R. J., Barris, W. C., McWilliam, S. M. and Dalrymple, B. P. 2004. An interactive bovine *in silico* SNP database (IBISS). Mamm. Genome 15:819-827.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backonfen, R. and Platzer, M. 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nature Genet. 36:1255-1257.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backonfen, R. and Platzer, M. 2006. Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. Am. J. Hum. Genet. 78:291-302.

Hu, G., Modrek, B., Stensland, H. M. F. R., Saarela, J., Pajukanta, P., Kustanovich, V., Peltonen L., Nelson, S. F. and Lee, C. 2002. Efficient discovery of single nucleotide polymorphisms in coding regions of human genes. Pharmacogenomics J. 2:236-242.

Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W. and Lee, C. J. 2000. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat. Genet. 26:233-236.Jeon, J. T., Lee, J. H., Kim, K. S., Park, C. K. and Oh, S. J. 2006. Application of DNA markers in animal industries. Aust. J. Exp. Agric. 46:173-182.

Kim, H., Shmidt, C. J., Decker, K. S. and Emara, M. G. 2003. A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. Anim. Genet.

34:249-254.

Kollers, S., Megy, K. and Rocha, D. 2005. Analysis of public single nucleotide polymorphisms in commercial pig populations. Anim. Genet. 36:426-431.

Kwok, P. Y., Carlson, C., Yager, T. D., Ankener, W. and Nickerson, D. A. 1994. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. Genomics. 23:138-144.

Lee, S. H., Park, E. W., Cho, Y. M., Lee J. W., Kim, H. Y., Lee, J. H., Oh, S. J., Cheon, I. C. and Yoon, D. H. 2006. Confirming single nuclotide polymorphisms from expressed sequence tag datasets derived from three cattle cDNA libraries. J. Biochem. Mol. Biol. 39:183-188.

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z. J., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P. Y. and Gish, W. R. 1999. A general approach to single-nucleotide polymorphism discovery. Nat. Genet. 23:452-456.

Mbikay, M., Seidah, N. G. and Chretien, M. 2001. Neuroendocrine secretory protein 7 B2: structure, expression and functions. Biochem. J. 357:329-342.

Picoult-Newberg, L., Idekar, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A. and Boyce-Jacino, M. 1999. Mining SNPs from EST databases. Genome Res. 9:167-174.

Sambrook, J., Fritsch, E. F. and Maniatis, T. 1989. *Molecular cloning: A laboratory manual*, 2nd edition Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NewYork.

Syvanen, A. C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat. Rev. Genet. 2:930-942.

The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933.

Useche, F. J., Gao, G., Harafey, M. and Rafalski, A. 2001. High-throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform. Ser. 12:194-203.

Uenishi, H., Eguchi, T., Suzuki, K., Sawazaki, T., Toki, D., Shinkai, H., Okumura, N. and Awata, T. 2004. PEDE (Pig EST Data Explorer): Construction of a database for ESTs derived from porcine full-length cDNA libraries. Nucleic Acids Res. 32:484-488.

Vignal, A., Milan, D., SanCristobal, M. and Eggen, A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34:275-305.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Krugylyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J. and Lander, E. S. 1998. Large-scale identification, mapping, genotyping of single nucleotide polymorphisms in the human genome. Science 280:1077-1082.

Zimdahl, H., Nyakatura, G., Brandt, P., Schulz, H., Hummel, O., Fatmann, B., Brett, D., Droege, M., Monti, J., Lee, Y. A., Sun, Y., Zhao, S., Winter, E. E., Pontig, C. P., Chen, Y., Kasprzyk, A., Birney, E., Ganten, D. and Hubner, N. 2004. A SNP map of rat genome generated from cDNA sequences. Science 303:807.

Chapter IV.

TRANSCRIPTIONAL START SITE ELUCIDATION FROM FULL-LENGTH ENRICHED cDNA SEQUENCES

1. INTRODUCTION

Although many researchers have deposited ESTs in the public database, the 5' end termini of most of these sequences deposited in public databases are incomplete and conventional methods for determining exact transcription start sites (TSS) such as 5'RACE (Frohman et al., 1988; Schaefer, 1995) or primer extension (Mcknight et al., 1982) are laborious. In addition, computational methods for predicting the location of promoters are immature (Suzuki et al., 2002). This difficulty can be overcome by the sequences from full-length enriched cDNA libraries where majority of sequences are extended towards the 5' end. Exact determination of TSS is important for identifying the promoter region, which is located just proximal to TSS in many cases. When multiple TSS is observed, it enables us to examine the dynamic nature of the transcriptional initiation events (Suzuki et al., 2001). Since TSS marks the 5'end limit of cDNA, the amino acid sequence of the exact N-terminus of the encoding protein could be deduced, which is essential to examine the presence of protein sorting signals. Database of transcriptional start sites for human and mouse have been reported (Suzuki et al., 2001; Suzuki et al., 2002; Suzuki et al., 2004). There is only one report regarding transcriptional start site identification from the full-length enriched sequences of porcine i.e. from olfactory bulb cDNA library (Fujisaki et al., 2004). We tried to analyze transcriptional start sites from the sequences of our full-length enriched libraries by comparison with *Mus musculus*, *Bos taurus* and *Homo sapiens* sequences.

2. MATERIALS AND METHODS

2.1 Bioinformatic approach for transcriptional start site identification. For analysis of EST sequences, 3,390 trace files obtained from the spleen, neocortex, brainstem and liver libraries were transferred to a Linux work station and base called using Phred (Ewing and Green, 1998; Ewing *et al.*, 1998), assessed for having the quality score higher than 20 in 20 bp windows. Sequences that met the criteria were subjected to Cross_match (P. Green, Unpublished) for vector trimming. RepeatMasker was used for removing repeat sequences (http://repeatmasker.genome.washington.edu/; A.F.A. Smit and P. Green, Unpublished). The processed sequences were assembled by Phrap (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html). The assembled sequences were analyzed by NCBI blastall with the option E-value>100. Museq box was used for parsing the sequences and analyzed for the presence of minimum 5 sequences and a maximum of 20 sequences from 4 species (*Homo sapiens*, *Bos taurus*, *Mus musculus* and *Sus scrofa*).

2.2 Comparison of transcription start sites among different species. Precise transcription start sites for the sequences from each of the 4 species were determined by alternative splicing using Spidey program and analyzed to find similarities and differences among species. The TSS identified and analyzed from these four mammalian species were compared with chicken sequences.

3. RESULTS

3.1 Identification of transcriptional start sites. Assembly of 3390 EST sequences from four full-length enriched cDNA libraries by bioinformatics approach yielded 420 contigs. Of these, the contigs with E-value>100 were selected. Contigs that met the criteria were 141 and the gene sequences of these contigs were analyzed for the presence of minimum 5 gene sequences in each of the 4 mammalian species (*Homo sapiens*, *Bos Taurus*, *Mus musculus* and *Sus scrofa*). 40 of the contig sequences satisfied the criteria and were analyzed for the presence of transcriptional start sites in comparison with four different mammalian species. We have shown transcriptional start site analysis for 5 genes human T-cell leukemia virus type-1 binding protein (Tax1BP3) (NM_014604), NDRG family member 3 (NM_032013), serine incorporator 1 (NM_020755), thiosulfate sulfurtransferase (NM_003312) and polyubiquitin (M18159) for a pilot study.

4. DISCUSSION

Although the 5'terminus for majority of sequences from our full-length enriched cDNA libraries were extended towards the 5' end, it is not possible to determine the transcription initiation site of each gene, if the number of sequences are limited. Thats why the full-length sequence of single gene appearing more than 5 times from four different species was aligned with our full-length sequence. For Human T-cell leukemia virus type I binding protein 3 (Tax1BP3) gene, 10 *Sus srofa* sequences were starting at the same point and 20 *Bos taurus* sequences, 20 *Mus musculus* sequences were starting at same point. For NDRG family member 3 (NDRG3) and Serine incorporator 1 (SERINC1) gene, 20 sequences from *Mus musculus* were starting at same point. For Polyubiquitin (UBC) gene, 15 sequences from *Homo sapiens* were starting at the same point. Among the sequences for 5 genes from 4 species, *Mus musculus* sequences showed more variation while reaching the 5' end. All the species showed maximum variation towards the 5' end for Thiosulfate sulfurtransferase (TST) gene. Sequences in *Mus musculus* for two of the genes (TAX1BP3, NDRG3) from the total of five genes analyzed in this pilot study are slightly longer in comparison with sequences from other species. The *Bos taurus* sequences are shorter for 3 genes (TAX1BP3, NDRG3, UBC) in comparison with sequences from other species. The *Homo sapiens* sequences for NDRG3 is little shorter when compared with the *Homo sapiens* sequences from other genes but is slightly longer than that of*Bos taurus*for that gene.in comparison with other sequences. This is the only gene among 5 genes where the Homo sapiens sequences are short.Further analysis are required in determining transcription factor binding sites for functional analysis and effects due to difference in promoter length.

Our methodology was based on analyzing the transcription start sites by comparison with *Mus musculus*, *Homo sapiens*, *Bos taurus* and *Sus scrofa* sequences from Genbank EST database. Comparison of sequences of human and other organisms helps to extract biologically meaningful information as to which parts of the genomic sequence are likely to have functional relevance. It is expected that the functionally important regions such as exons and promoter elements are evolutionary conserved and could be discriminated from non-conserved ones, which are supposed to be subject to fewer functional constraints (Hardison, 2000; Boguski, 2002). We expect that the full-length enriched cDNA libraries constructed in our study will enable to identify promoter regions and helps in better understanding of complex transcription process operating in species.

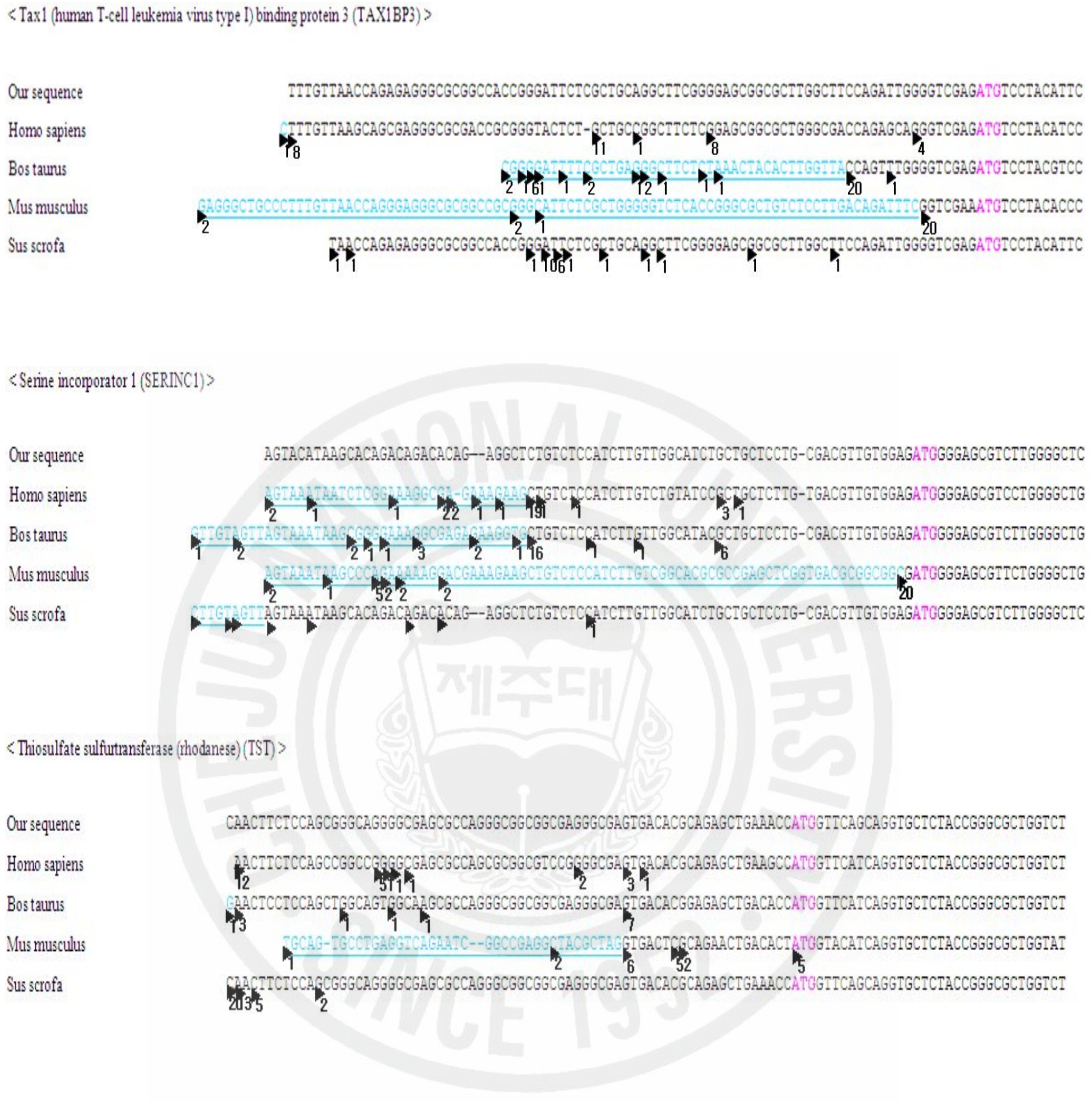
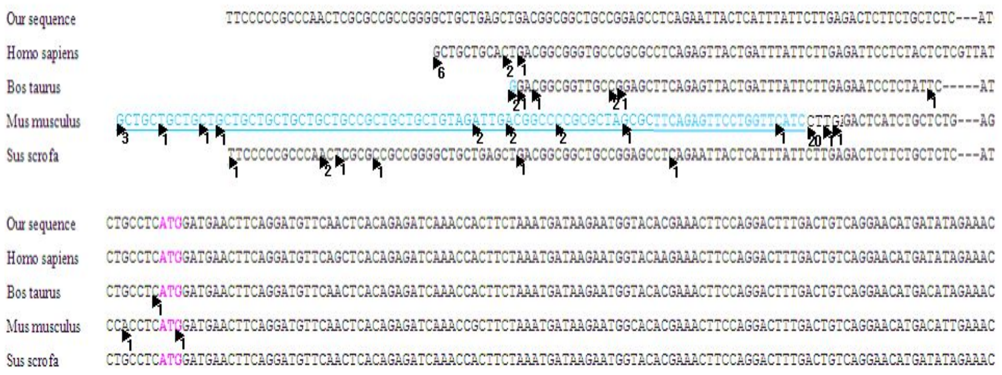


Figure 4-1. Transcriptional start site analysis from 5 genes as examples. ATG is the start codon. Triangled arrows indicate the number of sequences from the same species starting at same point.

<NDRG family member 3 (NDRG3)>



<Polyubiquitin (UBC)>

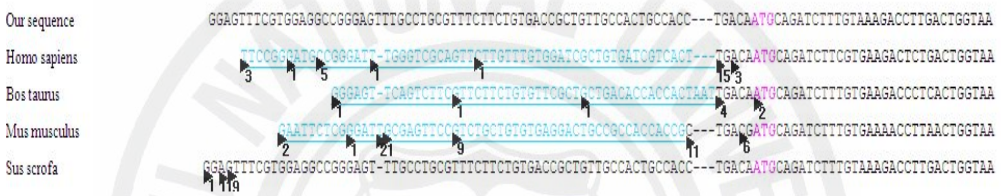


Figure 4-1. (continued).

REFERENCES

Boguski, M.S. 2002. Comparative genomics: the mouse that roared. Nature 420:520-562.

Ewing, B. and Green, P. 1998. Base calling of automated sequencing tracers using phred. II. Error probabilities. Genome Res. 8:186-194.

Ewing, B., Hillier, L., Wendl, M. and Green, P. 1998. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175-185.

Frohman, M. A., Dush, M. K. and Martin, G. R. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl Acad. Sci. 85:8998–9002.

Hardison, R. C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. 16:369-372.

McKnight, S. L. and Kingsbury, R. 1982. Transcriptional control signals of a eukaryotic protein-coding gene. Science 217:316–324.

Schaefer, B. C. 1995. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. Anal. Biochem. 227:255–273.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T. and Morishita, S. 2001. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep. 2:388–393.

Suzuki, Y., Yamashita, R., Nakai, K. and Sugano, S. 2002. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. 30:328–331.

Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. 2004. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004.

Nucleic Acids Res. 32:D78-D81

Suzuki, Y., Yamashita, R., Nakai, K. and Sugano S. 2002. DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. Nucleic Acids Res. 30:328-331.

Uenishi, H., Eguchi, T., Suzuki, K., Sawazaki, T., Toki, D., Shinkai, H., Okumura, N. and Awata, T. 2004. PEDE (Pig EST Data Explorer): Construction of a database for ESTs derived from porcine full-length cDNA libraries. Nucleic Acids Res. 32:484-488.

Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K. and Sugano, S. 2006. DBTSS: DataBase of Human Transcription Start Sites, progress report. Nucleic Acids Res. 34:d86-89.

ACKNOWLEDGEMENT

I praise and thank God who has been an unfailing source of strength and inspiration in the completion of this work. I would like to express my profound gratitude to my advisor, Dr. Chankyu Park, for his support, patience, and encouragement throughout my graduate studies. It is not often that one finds an advisor and colleague that always find the time for listening to the little problems and roadblocks that unavoidably crop up in the course of performing research. His valuable advices were essential for the completion of this study.

I am deeply indebted to my other advisor Dr. Kyu-Il Kim and I also express my heartfelt thanks to my committee members Drs.Dong-Ki Jeong, Won-Geun Son and Sunjoo Lee. I also thank all other professors in Department of Animal Biotechnology. I wish to thank all my lab members Chanjin, Juhyun, Nameun, Rui, Yunshin, Hojun, Inyeop andKwangha for their constant encouragement and support in different ways during my study. Without them, this piece of work would not have been completed.

I must express my heartfelt gratitude and thanks to my friends at Cheju National University Anil, Reddy and all other foreign students who made my stay in Korea a pleasant one. I would also like to acknowledge Dr. Kiran K. Sharma, Principal Scientist, ICRISAT who gave me the opportunity to work with international community during my Masters and made me think about the prospects of doing Ph.D.

My heartfelt thanks to my friend Sridhar who was then at Goettingen, Germany and now moved to Bristol, UK helped me to encounter difficulties. Also, I greatly acknowledge visiting professors at Konkuk University, Drs. Sai and Panda, and other Indian friends Dhandapani, Gupta and Ziban.

Finally, my English knowledge is not mere sufficient to find words with which I express the gratitude that I owe to my parents, sister's and my fiancé's families. My parent's tender love and affection have always been the cementing force for building the blocks of my academic career. I can never ignore the all round support rendered by them which provided the much needed stimulant to overcome difficulties. I would like to thank all my relatives, friends, teachers, well wishers, so numerous to list here helped me in many ways to pass through the phase of difficulties to gain wisdom.