Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Sequence Analysis of a Porcine Normalized Full-length cDNA Library

Tsedeke Kocho Ketema

Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics

464

Uppsala 2011

Master's Thesis, 30 hp

Erasmus Mundus Programme – European Master in Animal Breeding and Genetics

Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Sequence Analysis of a Porcine Normalized Full-length cDNA Library

Tsedeke Kocho Ketema

Supervisors:

Richard Crooijmans, WUR, Animal Breeding and Genetics Group
Göran Andersson, SLU, Department of Animal Breeding and Genetics

Examiners:

Erling Strandberg, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: EX0556
Programme: Erasmus Mundus programme
– European Master in Animal Breeding and Genetics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2011
Name of series: Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics, 464
On-line publication: http://epsilon.slu.se

Key words: DNA sequence, structural genes, RNA, pigs

Education and Culture
Erasmus Mundus

Sequence Analysis of a Porcine Normalized Full-length cDNA Library

By

Tsedeke Kocho Ketema
780825-451-120

MAJOR MSc. THESIS ANIMAL BREEDING AND GENETICS (ABG 80430)
July 2011

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Animal Breeding and Genomics Centre (ABGC)

SUPERVISORS:

1. Dr. Richard Crooijmans (Wageningen University)
2. Prof Dr.Göran Andersson (Swedish University of Agricultural Sciences)

Education and Culture
Erasmus Mundus

Sequence Analysis of a Porcine Normalized Full-length cDNA Library

By

Tsedeke Kocho Ketema

A thesis submitted in partial fulfillment of the requirements for the degree

of

MASTER OF SCIENCE

in

Animal Breeding and Genetics

_____	_____
Dr. Richard Crooijmans (WUR)	Prof. Dr. Martien Groenen (WUR)
Supervisor	Examiner
_____	_____
Prof. Dr. Göran Andersson (SLU)	Prof. Dr. Erling Strandberg (SLU)
Supervisor	Examiner

Wageningen University
Wageningen, The Netherlands
July 2011

TABLE OF CONTENT

PREFACE

My deepest acknowledgment goes to my supervisor Dr. Richard Crooijmans for his scientific guidance and personal encouragements throughout the research period. I'm also thankful for Prof. Dr. Martien Groenen for offering me the opportunity to undertake this research and also assisting data analysis. My second supervisor, Prof. Dr. Göran Andersson of SLU is acknowledged for his inputs in the research. The Erasmus Mundus European Master in Animal Breeding and Genetics Consortium is duly acknowledged for this wonderful study program. I'm thankful to Prof. Dr. Ir. Johan van Arendonk for his vision and pioneer this study programme which offered a unique prospect especially for many of us from developing countries with rare opportunity to receive such intense and integrated training in animal breeding and genetics. The financial support from European Union made this study possible and highly appreciated. Both of my host universities, SLU and WUR, are sincerely acknowledged for their supports under their extraordinary educational and research systems. Dieuwertje Lont of WUR was always understandable and helpful throughout the study period and highly acknowledged. Patricia Huijbers and Milou Toetenel of WUR have made a warm reception and provided many supports during the Introductory EM-ABG period at Wageningen. I'm thankful to Dr. Birgitta Malmfors of SLU for her warm reception and many supports during my stay in Uppsala, Sweden. Sofia Folestam, Helena Eken and Maria Starkenberg of SLU have provided various supports and acknowledged. Bert Dibbits and Tineke Veenendaal provided help during the lab work and Yogesh Paudel assisted in processing data and I'm grateful for them. I'm thankful to Sandra Bernal for her valuable information for this research. Dr. Ir. Hans Komen, our study coordinator at WUR has provided guidance with the course and research works and duly acknowledged.

SUMMARY

The pig is besides an important livestock species also a model organism for human biomedical research. Knowledge of the porcine genome is essential for improving product quality, animal welfare and also the biomedical research. This is accomplished through investigating the transcribed regions of the genome by collecting, sequencing and analyzing transcribed sequences (mRNA) converted into a complimentary DNA (cDNA) providing a complete sets of expressed genes. Therefore, the objective of this study was to sequence and analyze 10,000 porcine normalized full-length cDNA clones. Total RNA was extracted from 11 tissues of a fetal clone of pig and a normalized full-length cDNA library was constructed by a commercial company. The cDNA clones were cultured in 384-well plates and sequenced using Sanger sequencing method. The sequence similarity search was performed using Basic Local Alignment Search Tool (BLAST) against the porcine genome, porcine cDNA, human cDNA and mouse cDNA databases. Combining sequences from this study and the dataset generated earlier, a total of 13,989 sequences of at least 50 bps or more were generated from an overall of 19,968 cDNA clones processed. From the overall clone sequences, a total of 12,220 sequences provided hit in one or more of the pig, human or mouse databases. Blasting against the pig genome provided larger hits of 10,857. On the other hand, the pig cDNA database has provided total hits of 6,597. The human and mouse cDNA provided a total hits of 4,786 and 2,801, respectively, that enable comparative analyses to identify the homologous pig genes. Only 52 sequences have the same start-site with their respective pig transcripts and the majority of sequences shown variation. A total of 3,164 genes were identified from the library. A large-scale collection and characterization of the normalized cDNA library using direct sequencing on 384-well plates provides a valuable tool for understanding and investigation of the pig genome.

1. INTRODUCION

The pig is an important livestock species with pork being the leading source of animal protein worldwide (Archibald *et al*., 2010). In addition, the pig is also an important model organism for human biomedical research due to its anatomic, physiological, biochemical and metabolic similarity with human (Kim *et al.*, 2010). Domestic animal breeding for improving product quality and animal welfare and also biomedical research requires a better understanding of major physiological functions and their interactions which largely regulated by the genome of the animal (Ota *et al*., 2004; Imanishi et al., 2004). To help achieve this goal, the draft genome sequence of the domestic pig (*Sus scrofa*) is available and a revised assembly (Sscrofa10) is under construction which incorporate the whole genome shotgun sequence (WGS) data providing >30× genome coverage (Archibald *et al*., 2010). Understanding the genome sequence will be greatly enhanced by the availability of all expressed genes allowing the transcription analyses of any function of the organism (Ota *et al*., 2004; Hayashizaki *et al*., 2002; Maeda *et al*., 2006).

Coding regions of eukaryotic genome are interspersed with noncoding DNA presenting major challenge in interpreting the function of the genome sequences (Kawai *et al*., 2001). Different prediction programs may lead to different predicted gene sets and gene products and a true number of protein-coding genes remains uncertain (Gorodkin *et al*., 2007; Seki *et a*l., 2002). The cDNA cloning identifies exon-intron structure and promoter region of the gene with higher accurately than the bioinformatic predictions based on genome sequence (Kato *et al*., 2005; Shcheglov *et al*., 2007; Harbers, 2008). Expressed sequence tags (ESTs) are commonly used for gene discovery and investigation but their usefulness is limited because many EST clones lack the complete sequences of mRNAs (Kim *et al*, 2006). They represent only a fragment of a highly expressed gene but not full transcript and rare ESTs are often hard to find among the large numbers of highly expressed ESTs. Similarly, a standard or non-normalized cDNA library also provides redundant information of the most abundant transcripts in which intermediately and highly expressed cDNAs would be sequenced redundantly and ineffective for discovering rarely expressed genes (Carninci *et al*, 2000).

A major challenge related with generating full-length cDNA libraries especially for discovering novel genes stemmed from the nature of distribution of eukaryotic cellular mRNA (Carninci *et al*, 2000). The mRNA constitutes approximately 1–5% of total RNA mass and depending on their expression level, they could be either superprevalent/abundant,

intermediate or rare representing a total mRNA of 20%, 40%–60% and 20%–40%, respectively (Shcheglov *et al*., 2007). Thus, it is essential to obtain a pool of mRNA that is representative of all expressed sequences. This requires normalizing the frequencies of full-length cDNAs from mRNAs belonging to the three different classes of expression (Carninci *et al*, 2000). Normalization is performed through hybridization in which the most abundant cDNA species hybridize faster than rare ones and the double-stranded hybrids formed by abundant transcripts can be removed from the remaining single-stranded cDNAs of less abundant transcripts. Normalization of the cDNA library greatly increases the efficiency and economy of sequencing through reducing frequency of abundant genes and enriching the library with rare genes (Natarajan *et al*., 2010).

Therefore, it is essential for all genomic studies to investigate the transcribed regions in the genome, typically performed by sequencing mRNA transcripts converted into complimentary DNA (cDNA) and mapping each sequence on the genome (Kawai *et al*., 2001, Gorodkin *et al*., 2007; Kato *et al*., 2005). The cDNA libraries only present parts of genes while having a full-length cDNA library provide the complete transcript of a gene. Sequencing of the different clones will also identify the different transcripts present in a certain gene (Bonnet *et al*., 2008; Fahrenkrug *et al*., 2002). Accordingly, a genome-scale collections and characterization of normalized full-length cDNA libraries become crucial in structural and functional analyses of genes (Fahrenkrug *et al*., 2002; Nguyen *et al*., 2010; Oshikawa *et al*., 2008).

Large-scale cDNA sequencing requires constructing cDNA libraries that contain large proportion of full-length cDNA clones. The messenger RNAs of eukaryotes have a cap structure at its 5′-end and a 3′-poly(A) stretch and the sequence information between both ends is important for identifying the coding region and the non-coding regions. Isolation of a ‘full-length’ cDNA which consists a copy of all the sequences between the cap and the poly(A) of a mRNA is an indispensable step for the analysis of gene structure and function. A full-length cDNA library provides a full-length cDNA clones that contain the mRNA start site of the gene (Suzuki *et al*., 1997). However, this was often challenged by the reduced efficiency of reverse transcriptome reaction and lack of efficient techniques for selecting the complete recombination sequences. A cap-tapper technique was reported to be efficient and commonly employed for constructing full-length cDNA library (Carninci *et al*., 1996). The full-length of the clone sequences is often evaluated using the relative positions of 5’-end of

the oligo-capped cDNA with reference to the reference databases. Accordingly, the sequences that covered the annotated coding-sequence start sites are categorized as 'full or near-full' (Ota, *et al*., 2004).

A large-scale normalized full-length cDNA library is sequenced using Sanger sequencing technique (Sanger *et al*., 1997). The single stranded clone DNA template are annealed with universal primer (T3) at plasmid DNA adjacent to the insert DNA and that synthesise a new DNA strand complementary to the template DNA. The strand synthesis reaction is catalyzed by DNA polymerase enzymes with four deoxyribonucleotide triphosphates (dNTPs-dATP, dCTP, dGTP, and dTTP) as a substrate. A small amount of four dideoxynucleotide triphosphates (ddNTPs-ddATP, ddCTP, ddGTP, and ddTTP) labelled with different fluorescent markers is used for termination reaction. The clone sequences obtained from large-scale sequencing of full-length cDNA can be quickly searched for sequence similarity against the reference database using a Basic Local Alignment Search Tool (BLAST) tool that calculates the statistical significance of matches (Altschul *et al*., 1990). BLAST programs of NCBI detect similarities between a query sequence and database sequences and it is the most fundamental and frequent type of analysis performed on GenBank data (Besnon, *et al*., 2011).

A major task following genome sequencing is the transcriptome analyses (Gorodkin *et al*., 2007). Several genomic studies to elucidate biological processes have focused on the investigation of transcribed regions of the genome by collection and characterization a pool of cDNA libraries (Kuroshu *et al*., 2010). So far, the pig transcriptome has been analyzed by many groups involving large-scale collection, sequencing and characterization of expressed sequence tags (ESTs) traversing from cells and tissues of interest and developmental stages. These databases were deposited in public databases for free use.

Recently, Lee *et al* (2009) has constructed both normalized and non-normalized cDNA libraries (SUSFLECKs) from tissues related to energy metabolism in pig including abdominal fat, induced fat cells, loin muscle, liver, and pituitary gland. Accordingly, they have generated a total of 71,100 ESTs sequenced once from the 5′-ends of the clones. They have deposited a total of 55,658 sequences in the public database. They have reported that the analyses of the libraries provided an important insight to discover the functional pathways in gene networks and to expand understanding of energy metabolism in the pig. Similarly, an earlier study by Kim *et al* (2006) has deposited a total of 16,110 ESTs sequences generated from full-length enriched cDNA library of the porcine backfat tissue.

Gorodkin *et al* (2006) have generated large-scale ESTs sequences from 97 non-normalized and 1 normalized cDNA libraries covering 35 different tissues of pig. Accordingly, they have generated pig EST resource comprising over one million ESTs. The same study has also shown the differential expression of genes between different tissues and developmental stages of the pig. Similarly, Bonnet *et al* (2008) has also reported the production and deposition of 24,449 ESTs of pig in the EMBL database. They have constructed six different normalised and subtracted multi-tissue cDNA libraries from 38 porcine tissues. The libraries are pooled as brain, digestive function, glands, heart and muscle, male reproductive organ and female reproductive organ.

Related to the profile of gene expression at various developmental stages, Whitworth *et al* (2004) has identified mRNAs that are present at early stages of embryogenesis in pig providing a catalog of expressed genes during the porcine early embryo developmental stage. Accordingly, they have generated a total of 8,066 3'-end sequenced ESTs from cDNA libraries constructed from porcine embryos. In this line, Fahrenkrug *et al* (2002) has also constructed two normalized porcine cDNA libraries; one from pig embryo at several developmental stages associated with muscle development and organogenesis and the other from tissues involved in porcine reproductive physiology and generated and deposited a total of 66,245 5-end one-pass sequences.

In addition, a valuable on-line porcine expressed sequence tag database (Pig EST Data Explorer, PEDE) (http://pede.dna.affrc.go.jp/) has been established and further upgraded (Pig Expression Data Explorer, PEDE) that provide a catalogue of pig ESTs generated from full-length cDNA libraries constructed from several important pig tissues (Uenishi *et al*., 2004; Uenishi *et al*., 2007). Accordingly, a total of 190, 370 ESTs (July 2011) were sequenced and available from sequencing of clones in seventeen full-length-enriched cDNA libraries derived from fifteen different pig cells and tissues. The database also comprises a putative single nucleotide polymorphisms (SNPs) in EST assemblies grouped in breed specificity and their effect on coding amino acids providing useful information to explore genes that may be responsible for traits of interest. This resource provide a crucial support for further analyses of the functions of specific genes through picking the physical cDNA clones and also contributes to the sequencing integrity of the pig genome. Generally, the profile of ESTs from various databases are deposited together at GenBank of NCBI and currently 1,631,972 ESTs (July 2011) are available in the database (Besnon, *et al*., 2011).

Currently, a comprehensive full-length cDNAs are available for human and mouse and it greatly facilitated structural and functional studies of mammalian genome. A large-scale identification, sequencing and characterization of the full-length human cDNAs have greatly contributed to an integrative annotation of the human genome. As the result, a public database called H-Invitational (H-InvDB) (http://www.h-invitational.jp/) was developed and improved further (Yamasaki *et al*., 2008; Yamasaki *et al*., 2009) comprising wide range of useful information of the human genome function. Similarly, a comprehensive collection and functional annotation of mouse full-length enriched cDNAs were developed by the RIKEN Mouse Gene Encyclopedia project of the international annotation consortium (FANTOM) which comprehensively cloned, sequenced and annotated the full-length mouse cDNAs and continuously upgrading (Kawai *et al*., 2001; Maeda *et al*., 2006). Accordingly, they also developed an interactive database FANTOM mouse database (http://fantom.gsc.riken.jp/) which is pivotal resource in understanding the function of mammalian genome. In addition to developing the databases and functional annotations of both human and mouse genome, a large-scale collection and cloning of the full-length cDNA libraries provided a physical resource of cDNA clones for further research.

A review of the current resources of pig ESTs indicated that several research groups are increasingly generating ESTs from cDNA libraries representing various tissues and developmental stages of the pig. However, a profile of expressed sequences generated from full-length cDNA library is far from complete. The complete collection of ESTs is an essential resource for annotation, comparative genomics, assembly of the pig genome sequence, and further porcine transcription studies. Both genetic selection and biomedical research require an in-depth understanding of the genetic architecture of pig which is largely facilitated by the availability of the expression profile of the genome. To this end, there is a growing collaborative effort to deposit the comprehensive cDNA database of pig in the public nucleotide database and as part of that effort, the Animal Breeding and Genomics Centre (ABG) of Wageningen University is currently sequencing and characterizing a newly developed porcine normalized full-length cDNA library derived from a clone of the animal which was used in deriving the pig genome sequence. Therefore, the aim of this specific project is to sequence and analyze clones from the porcine normalized full-length cDNA library.

2. OBJECTIVES

1. To pick and sequence 10,000 normalized full-length cDNA clones of pig
2. To blast the cDNA sequences against pig, human and mouse nucleotide databases
3. To allocate the sequences on the pig genome
4. To build a resource of sequence clones

3. MATERIALS AND METHODS

3.1 RNA Extraction and cDNA Construction

Total RNA was extracted from 11 tissues (kidney, liver, lymph node, cerebellum, placenta, colon, hypothalamus, brain frontal lobe, spleen, small intestine and lung) of a fetal clone (113 days of gestation) from Duroc pig T.J. Tabasco. The total RNA was pooled and sent to a commercial company for construction of normalized full-length cDNA library (DNAFORM, 2009). Accordingly, the constructed library in phage stock and plasmid stock was acquired back and stored in the freeze (-80^0C). In addition, the detailed documentation of mRNA purification, normalization process, first and second strand cDNA synthesis, cloning of cDNA cloning into phage Lambda-FLC III, transformation of plasmid DNA into T1 bacteriophage resistant E. coli strain DH10B was obtained.

3.2 Culturing, Plasmid Isolation and Sequencing of Clones

The culturing of clones and sequencing was performed with established protocol provided in the Appendix (Appendix Table 1). The clones were cultured in a petri dishes of 145/20 mm with LB media supplemented with ampicillin (0.1 µg/ml), agar and grown overnight (18 hrs) in an incubator at 37^oC. Individual colonies were randomly picked from the petri dishes and transferred to 384-well plates (Microplates of 384-wells, 120 µl capacity with lid) with 1 % Lysogeny Broth (LB) media supplemented with ampicillin and freezing media and grown overnight. The original or master plates coded as POR_A were replicated using 384-pin replicator in two new plates labelled as POR_B and POR_C and grown overnight in the incubator. The first two plates (POR_A and POR_B) were stored in freeze (-80^0C) as recovery plate. The third plates (POR_C) were grown overnight in LB media supplemented with ampicillin and used for direct sequencing. A cell lysate was prepared from third plate by centrifuging and washing with deionised (MQ) water of 25 µL. The cell culture was denatured in a thermocycler at 95°C for 5 min followed by centrifuge. A 2µL cell culture from each of

the individual wells were transferred to a new 384-well PCR plates as a template for sequencing reaction and mixed with sequencing reaction comprising 0.5 µL of BigDye 3.1 (Applied Biosystems), 0.75 µL of 5X buffer, 1 µL of primer (0.8 µM) and MQ water, giving

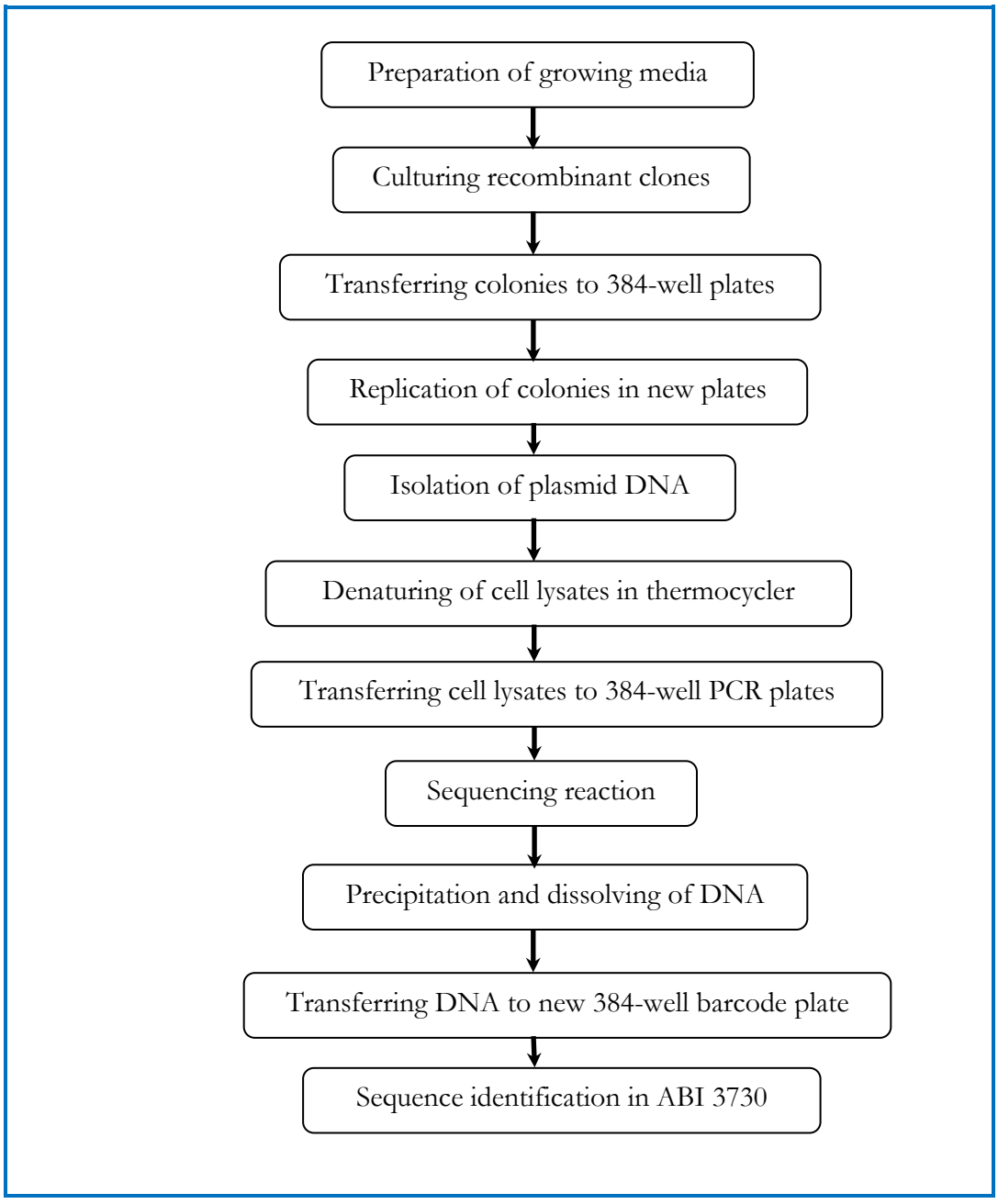


Figure 1. *Schematic flow of the cDNA clone sequencing*

a final volume of 5 µL per sample. Amplification was performed using cycle sequencing in BD50x50 program of the thermocycler (Appendix 1). Following amplification, the DNA was precipitated with 0.5 µL of NaAc-EDTA and 17 µL of ethanol and finally dissolved with 10 µL of formamide. The suspension was transferred to a new barcode 384-well PCR plates,

sealed and loaded into ABI3730 DNA Analyzer. The sequences were determined by capillary electrophoresis in ABI3730 DNA Analyzer (Applied Biosystems®).

3.3 Sequence Analyses

The raw sequences from the ABI 3730 DNA Analyzer were converted into FASTA files using the on-line DNA Baser Sequence Assembler (http://www.dnabaser.com/index.html). The cDNA sequences with the nucleotide length of a minimum of 50 base pairs (bp) were considered as a useful read sequence that could provide reliable sequence similarity when blasted against the reference nucleotide databases. The total numbers of useful reads from each plate were identified and success rates per plate were determined. Sequences from individual plates were pooled together in one file for blasting. The sequences were submitted for similarity search against nucleotide databases of pig genome (build9), pig cDNA, human cDNA, mouse cDNA, E. coli genome and FLCIII DNA available in Ensembl (www.ensembl.org/index.html) and NCBI (www.ncbi.nih.gov) public databases. Prior to blasting, the databases were downloaded and converted into blastable fasta format. The sequence similarity searches were performed using Basic Local Alignment Search Tool (BLAST) resource of nucleotide blastn (BLAST 2.2.23 release) in NCBI. The significant cut-off of the sequence similarity was 1e-10 and below. The clone sequences that has shown similarity to the pig cDNA and genome databases were further analyzed using bioinformatic scripts to map the sequences on the pig genome and identify the gene transcripts that cover the cDNA sequences. The script provided the start and end sites of the cDNA clone sequence on the reference databases, chromosomal information, ENSEMBL annotated gene transcripts and descriptions of the genes.

4. RESULTS

4.1 Sequencing Porcine Normalized Full-length cDNA Library

In total 12,288 clones have been picked and transferred to 32 384-well plates (Table 1). An overall of 8,520 one-pass 5′-end sequences were generated. The average length of the clone sequences was 314 bp, ranging from 51-811 bp.

Table 1: *Total number of sequenced cDNA clones*

SN	Plates	Sequences					
		Total number of sequences	*Length of sequences (bp)*	*Min (bp)*	*Max (bp)*	*Average*	*Efficiency* (%)*
1	POR-C024	137	46055	70	635	336	35.68
2	POR-C025	215	65887	51	575	307	55.99
3	POR-C026	351	118475	77	633	338	91.41
4	POR-C027	345	119046	111	697	345	89.84
5	POR-C028	49	15791	137	515	322	12.76
6	POR-C029	324	111800	71	609	345	84.38
7	POR-C030	332	132158	52	749	398	86.46
8	POR-C031	333	129743	85	661	390	86.72
9	POR-C032	366	145046	72	627	396	95.31
10	POR-C033	328	118025	87	571	360	85.42
11	POR-C034	313	117280	57	657	375	81.51
12	POR-C035	337	117020	69	652	347	87.76
13	POR-C036	330	108784	57	582	330	85.94
14	POR-C037	258	81241	62	685	315	67.19
15	POR-C038	323	115145	73	786	357	84.11
16	POR-C039	230	62020	53	653	270	59.90
17	POR-C040	337	111976	60	595	332	87.76
18	POR-C041	314	101741	85	560	324	81.77
19	POR-C042	219	57969	56	695	265	57.03
20	POR-C043	336	120007	69	706	357	87.50
21	POR-C044	298	88131	59	716	296	77.60
22	POR-C045	275	82692	56	611	301	71.61
23	POR-C046	131	42155	76	644	322	34.11
24	POR-C047	280	94367	82	678	337	72.92
25	POR-C048	195	49178	59	495	252	50.78
26	POR-C049	206	50355	57	561	244	53.65
27	POR-C050	232	64481	60	614	278	60.42
28	POR-C051	297	73873	65	811	249	77.34
29	POR-C052	335	93082	73	576	278	87.24
30	POR-C053	204	46255	66	578	227	53.13
31	POR-C054	175	39795	55	623	227	45.57
32	POR-C055	115	26791	53	718	233	29.95
Overall		**8520**	**2,746,364**	**51**	**811**	**314**	**69.34**

**Efficiency per plates is a ration of the total number of sequences with good reads to clones in total wells (384)*

The overall efficiency of the sequencing was 69.34% (Table 1), which is comparable with

similar previous study that reported an average efficiency of 71.21% (Table 2). The sequencing efficiency has shown wide variation among plates (12.76 to 95.31%). The lowest efficiency (12.76%) was obtained with plate POR-C029 which was mainly due to wrong PCR programming during cycle sequencing. Similarly, a low success rate of 35.68% from the first plate (POR_C024) was due to a wrong sealing of the plate during denaturing. The cell lysate in POR-C046 plate was stored in the freezer for couple of days before sequencing reaction and provided lower efficiency (34.11%). The last two plates (POR_C054 and POR_C055) were stored in the refrigerator for some time (hour) before they were loaded into the ABI sequencer because the sequencer was under minor calibration and provided low efficacy.

Table 2: *Total number of sequenced cDNA clones from previous study*

SN	Plates	Sequences					
		Total number of sequences	*Length of sequences (bp)*	*Min (bp)*	*Max (bp)*	*Average*	*Efficiency (%)*
1	POR_B001	253	85648	59	693	339	65.89
2	POR_B002	305	119788	125	672	393	79.43
3	POR_B006	334	119108	56	612	357	86.98
4	POR_B007	202	61785	64	581	306	52.60
5	POR_B008	347	130970	61	727	377	90.36
6	POR_B009	366	134370	63	685	367	95.31
7	POR_B010	294	86802	52	563	295	76.56
8	POR_B011	298	95403	80	549	320	77.60
9	POR_B012	305	106946	60	630	351	79.43
10	POR_B013	252	83541	58	714	332	65.63
11	POR_B014	295	94686	73	646	321	76.82
12	POR_B015	248	82466	52	594	333	64.58
13	POR_B016	261	78317	51	641	300	67.97
14	POR_B017	250	93369	56	633	374	65.10
15	POR_B018	265	114195	57	656	431	69.01
16	POR_B019	228	97330	74	653	427	59.38
17	POR_B020	212	89252	71	750	421	55.21
18	POR_B021	226	85400	75	677	378	58.85
19	POR_B022	246	104783	58	744	426	64.06
20	POR_B023	282	129475	51	693	459	73.44
Overall		**5469**	**1993634**	**51**	**750**	**365**	**71.21**

On the contrary, some plates provided very promising success rates of up to 95.31%. These could be due to well grown colonies, immediate processing without storage in the freezer and accurate processing of the plates. In general, direct sequencing on 384-well plate involves various manual processing and transfer of suspensions to several new plates including picking and transferring of clones on the plates, duplication of the plates, heat sealing, vortex, storage,

etc. Consequently, these steps possibly contributed to a reduced overall efficiency.

Part of the library was sequenced previously and a total of 5,469 useful clone sequences were generated from an overall of 7,680 clones processed in 20 384-well plates (Table 2). The average length of the clone sequences was 365 bp ranging from 51-750 bp. Combining the two datasets, a total of 13,989 useful pig cDNA clone sequences were generated from an overall of 19,968 cDNA clones processed.

4.2 Sequence Similarities against Reference Databases

The cDNA sequences from the entire plates were combined together in one file and blasted against individual databases of pig genome, pig cDNA, human cDNA, mouse cDNA, E. coli genome and FLCIII DNA sequence. Accordingly, a single hit per sequence with the lowest cut-off e-value was selected. The similarity search results from the cDNA and genome databases were similar except that the cDNA databases provide the gene transcript which encompasses the cDNA sequence while genomic database search provide genomic position of the clones including chromosome, start and end-sites (Table 3).

Blasting against FLCIII DNA sequence and E. coli genome were performed to check the similarity of sequence with their sequences in situations like short size inserts result in sequencing parts of the E. coli genome or FLCIII DNA sequence after the end of the insert. These hits were checked with the hits from the pig cDNA and pig genome databases to verify whether they also provided hit in the pig databases. A single transcript (POR_C037_L12) (Table 3e) gave hit against the E. coli genome database. The FLCIII DNA sequence provided several hits but cross-checking of the hits either to pig genome or pig cDNA revealed that only 29 hits were unique but the rest provided hit in pig databases. Accordingly, a total of thirty clone sequences gave hit against the E. coli genome and FLCIII DNA, representing only 0.21 percent of total useful sequences generated. The average length of the clones that provided hit against the FLCIII DNA was 375 bps ranging from 217-542 bps (Appendix Table 2) whereas the clone that gave hit to E.coli genome was 249 bps (Appendix Table 3). The sequence length of the clones that shown similarity with the FLCIII DNA were longer than that of the average of the total sequences. On the other hand, blasting against human and mouse cDNA databases were performed for comparative analyses of sequences with more evolutionary related mammals before speciation.

Table 3. *Sequence similarity results from all databases*

Table 3a. Example of hits with Pig Genome sequence

Query ID (Clones)	Chromosome number	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	E-value	Bit score
POR_B024_A21	9	98.28	116	2	0	221	336	62501709	62501594	2.00E-51	207
POR_B024_M14	15	99.2	125	1	0	116	240	131877222	131877346	1.00E-57	228
POR_B024_D12	7	94.52	219	10	2	81	297	92037407	92037189	5.00E-96	355
POR_B024_F02	2	94.59	314	16	1	37	349	119619441	119619754	4.00E-143	512

Table3b. *Example of hits with Pig cDNA sequences*

Query ID	Subject ID	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	e-value	Bit score
POR_B024_F12	ENSSSCT00000007438	93.98	133	4	2	72	203	247	376	5.00E-53	206
POR_B024_A21	ENSSSCT00000016652	95.67	277	11	1	139	414	161	437	2.00E-130	464
POR_B024_M14	ENSSSCT00000017825	98.44	192	3	0	116	307	81	272	2.00E-94	344
POR_B024_F02	ENSSSCT00000015579	93.45	351	21	2	37	385	1	351	2.00E-158	556

Table 3c. *Example of hits with Human cDNA sequences*

Query ID	Subject ID	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	e-value	Bit score
POR_B024_G20	ENST00000223136	86.55	290	31	7	38	324	3	287	1.00E-88	327
POR_B024_G01	ENST00000344359	83.97	287	42	4	82	366	9	293	1.00E-80	302
POR_B024_E01	ENST00000422500	87.45	271	30	4	85	352	70	339	2.00E-92	340

Table 3d. *Example of hits with mouse cDNA sequences*

Query ID	Subject ID	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	e-value	Bit score
POR_B024_E01	ENSMUST00000032775	82.11	246	38	6	111	352	64	307	8.00E-60	231
POR_B024_F03	ENSMUST00000025503	86.45	310	34	7	45	350	1	306	4.00E-97	355
POR_B024_B10	ENSMUST00000105964	85.23	352	46	6	40	387	89	438	5.00E-107	388

Table 3e. A *hit with E.coli genome sequence*

Query ID	Subject ID	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	e-value	Bit score
POR_C037_L12	DH10B_WithDup_FinalEdit	84.68	222	32	2	28	247	3504344	3504123	6.00E-76	279

Table 3f. *Examples of hits with FLCIII DNA seqquence*

Query ID	Subject ID	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	e-value	Bit score
POR_C027_H17	FLCIII	97.56	41	1	0	7	47	683	723	6.00E-17	73.1
POR_C030_N13	FLCIII	100	30	0	0	1	30	692	721	1.00E-11	56.5
POR_C031_E12	FLCIII	100	32	0	0	6	37	692	723	9.00E-13	60.2

From the overall useful clone sequences generated from a porcine cDNA library, a total of 12,220 clone sequences provided hit in one or more of the pig, human or mouse databases. Blasting against the porcine genome sequence (Build 9) provided the largest number of hits (10,857) (Table 4). On the other hand, the pig cDNA database has provided total hits of 6,597. Majority of the clones that gave a hit with pig cDNA sequences also provided hit in the pig genome sequence. This is partly due to the fact that there are limited expressed sequence tags of pig from normalized full-length cDNA library that represents the complete profile of the transcribed parts of the genome. Therefore, both improving the genome sequence and large-scale sequencing and characterization of the normalized full-length cDNA library enable to obtain more hits from a cDNA library.

The human and mouse cDNA provided a total hits of 4,786 and 2,801, respectively. Hits from human and mouse cDNA enable comparative analyses of the pig genome that could provide an opportunity to identify the homologous pig genes.

Table 4: *Blast hits with different Databases*

Databases	**Total Hit**	**Unique Hits**
Pig Genome	10,857	3,439
Pig cDNA	6,597	27
Human cDNA	4,786	813
Mouse cDNA	2,801	-
E.coli Genome	29	29
FLCIII DNA	1	1

In addition to the overall hits per database, clones that provided a hit only in a single database but not others were classified as a 'unique hit' and identified from all the sequence similarity search outputs (Table 5a,b,c). Accordingly, the pig genome provided highest unique hits (3,439) than the rest of other databases. In the contrary, pig cDNA provided only 27 unique hits. The human cDNA database also provided considerable unique hits from 813 clones. The unique hits especially from the pig genome provide an important insight in the discovery of novel genes due to the fact that the clone sequences are a transcribed region of the genome but may not identified through prediction methods. The clone sequence could be from either highly, moderately or rarely expressed genes but they were not identified at the current level of pig genome sequence and annotation. On the other hand, unique hits from cDNA databases

provided predicted gene transcripts and genes. The unique hits from human cDNA can be potential mapped in porcine using comparative mapping an maylead to the discovery of new genes or strengthen the structural and functional annotation of the existing predicted genes.

Table 5. *Examples of clones that provided unique hit from databases*

Table 5a. *Examples of unique hits from Pig genome*

Clones	Pig chromosome	Start-site	End-site
POR_B024_D12	7	92037407	92037189
POR_B024_O14	5	3537119	3537441
POR_B024_B03	6	31719348	31719233
POR_B024_D03	13	67539627	67539553
POR_B024_D06	X	99197081	99197331

Table 5b. *Examples of unique hits from pig cDNA*

Clones	Pig transcripts
POR_B024_J14	ENSSSCT00000000624
POR_C032_I12	ENSSSCT00000001158
POR_C040_D19	ENSSSCT00000013040
POR_C047_I15	ENSSSCT00000004837
POR_C055_A23	ENSSSCT00000013419

Table 5c. *Examples of unique hits from human cDNA*

Clones	Human transcripts
POR_B024_G20	ENST00000223136
POR_B024_K12	ENST00000370685
POR_B024_N02	ENST00000303646
POR_B024_E08	ENST00000376663
POR_B024_L03	ENST00000389134

4.3 Identification of Pig Genes

All the useful cDNA clone sequences were analyzed for identifying the pig gene transcripts and genes. All the cDNA clone sequences and genes were mapped on the pig genome (Table 6). The start and end-sites and directional orientation (strand) of both the cDNA clone sequences and the pig genes were determined on the pig genome. In addition, the lists of pig genes and gene transcripts and description of the genes were obtained. A preliminary result from hits against pig cDNA databases provided an overall of 6,597 genes from the sequenced library (Fig 2). However, several gene transcripts and genes were found redundant because different clones provided several transcripts that are part of a single gene or a redundant transcript. For example, a *F1SPG2_PIG* pig gene was obtained from 133 different cDNA clones. Similarly, the pig genes *F1RUN2_PIG* and *F1RUQ0_PIG* each were obtained from a total of 20 and 19 different cDNA clones, respectively. The ENSSSCT00000003148 transcript of the pig *GPT2* gene was obtained from six different clones (Table 7). The ENSSSCT00000007115 transcript of *LMNA* gene was obtained from 8 different clones. Further, the ENSSSCT00000008609 transcript of the *F1RPB0_PIG* gene was obtained from thirteen different clones.

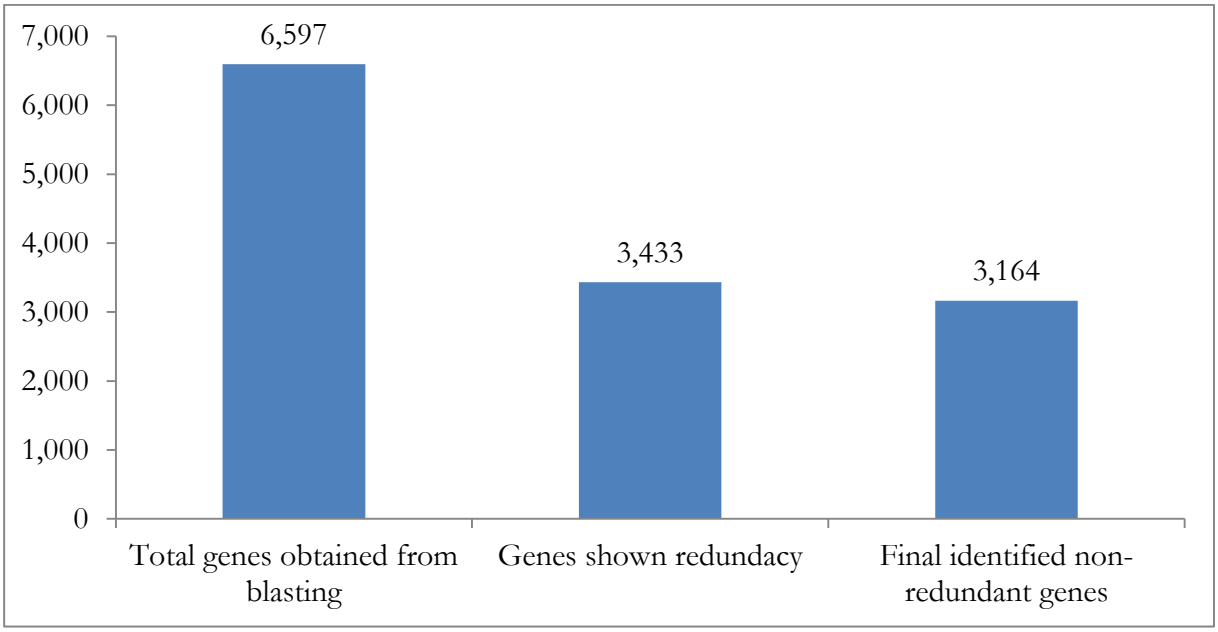


Figure 2. *Total number of genes identified from the library*

Thus, all the gene transcripts and genes were thoroughly checked for redundancy. Accordingly, a total of 3,164 non-redundant genes (Figure 2) were identified from the cDNA clone sequences from our library generated from 11 tissues of fetal pig clone.

Table 6. *Pig gene transcripts, genes, gene descriptions and genomic position*

Clones	Clone position	Transcript	Transcript position	Gene	Gene description
POR_B025_M06	2: 76243135-76243036	ENSSSCT00000015404	2:76258981-76286628:-1	ENSSSCG00000014102	*SCAMP1*
POR_B025_I17	10: 50118224-50118035	ENSSSCT00000012129	10:50114378-50118369:-1	ENSSSCG00000011082	*COMMD3*
POR_B025_I11	12: 40301482-40301600	ENSSSCT00000019300	12:40301423-40312721:1	ENSSSCG00000017733	*C17orf75*
POR_B025_H14	13: 18906627-18906280	ENSSSCT00000012317	13:18896001-18906667:-1	ENSSSCG00000011250	*ACAA1*
POR_B025_M23	X: 34084345-34084534	ENSSSCT00000013386	X:34238088-34255343:-1	ENSSSCG00000012240	*TSPAN7*
POR_B025_M17	2: 30907448-30907732	ENSSSCT00000014563	2:30907352-31012103:1	ENSSSCG00000013335	*LGR4*

Table 7. *Example of redundant pig gene transcripts* of *GPT2 gene*

Clones	Position of clones	Transcript	Gene	Transcript position
POR_B025_N09	6: 25390644-25390427	ENSSSCT00000003148	ENSSSCG00000002847	6:25348651-25390621:-1
POR_C036_A19	6: 25390644-25390440	ENSSSCT00000003148	ENSSSCG00000002847	6:25348651-25390621:-1
POR_C038_O20	6: 25390644-25390517	ENSSSCT00000003148	ENSSSCG00000002847	6:25348651-25390621:-1
POR_C042_I24	6: 25390640-25390379	ENSSSCT00000003148	ENSSSCG00000002847	6:25348651-25390621:-1
POR_C043_E01	6: 25390644-25390415	ENSSSCT00000003148	ENSSSCG00000002847	6:25348651-25390621:-1
POR_C048_M12	6: 25390644-25390444	ENSSSCT00000003148	ENSSSCG00000002847	6:25348651-25390621:-1

4.4 The Start-sites of Predicted Genes and cDNA Clone Sequences

The library was screening through one-pass 5′-end sequencing. The start-site of the cDNA clone sequences relative to the start-site of their corresponding predicted gene transcripts were determined to insight the fullness of the transcripts. The hit results from a pig cDNA databases were considered to identify the start-site of both the clones and transcripts. Accordingly, only 52 clone sequences (Table 9) have the same start-site with their corresponding predicted transcripts. In addition, clones with start-site varying from 1 to 60 nucleotides before or after the start-site of the predicted transcripts were also determined to show the proportion of clone sequences which show such variation (Table 8). Accordingly, a total of 999 clone sequences have a start-site varying from 1 to 60 nucleotides before or after the start-site of the predicted transcripts. For example, 17 clones have the start-site just one nucleotide before the predicted transcript while 19 clones shown that their start-site was one nucleotide after the start-site of the predicted transcript. Together, a total of 36 clones shown that their start-site is a single nucleotide before or after the start-site of their corresponding predicted transcripts. Further, a total number of clones with start-site varying from 1-5 before and after the start-site of the predicted transcripts were 93 and 69, respectively.

Table 8. *The start-site of clones relative to the predicted gene transcripts*

Variation in start-site (bps)	**Number of clones**		
	Before transcript start-site	*After transcript start-site*	*Total*
1	17	19	**36**
2	24	16	**40**
3	23	12	**35**
4	18	13	**31**
5	11	9	**20**
6–10	83	41	**124**
11–20	106	62	**168**
21–30	108	63	**171**
31–40	104	38	**142**
41–50	117	25	**142**
51–60	59	31	**90**
Total	**670**	**329**	**999**

Table 9. *Identical start-site of the cDNA clones and predicted gene transcripts*

SN	Clones	Pig chromosome	Clone start-site	Clone stop-site	Pig transcript	Transcript start-site	Transcript end-site	Pig gene
1	POR_C024_F03	2	119619441	119619754	ENSSSCT00000015579	119619441	119640124	ENSSSCG00000014258
2	POR_C026_D20	1	136415916	136416116	ENSSSCT00000005247	136415916	136421724	ENSSSCG00000004748
3	POR_C027_A09	15	102223292	102223684	ENSSSCT00000017560	102223292	102226921	ENSSSCG00000016128
4	POR_C027_P21	2	46464450	46464691	ENSSSCT00000014675	46464450	46466764	ENSSSCG00000013431
5	POR_C027_B11	10	49742656	49742880	ENSSSCT00000012125	49742656	49925053	ENSSSCG00000011079
6	POR_C036_P17	8	62038051	62038148	ENSSSCT00000009843	62038051	62129100	ENSSSCG00000008986
7	POR_C036_G07	4	104450611	104450783	ENSSSCT00000007341	104450611	104483333	ENSSSCG00000006702
8	POR_C037_K06	11	21507850	21508203	ENSSSCT00000010326	21507850	21510318	ENSSSCG00000009422
9	POR_C041_N17	8	62038051	62038148	ENSSSCT00000009843	62038051	62129100	ENSSSCG00000008986
10	POR_C042_C19	1	33720923	33721096	ENSSSCT00000004640	33720923	33830966	ENSSSCG00000004199
11	POR_C043_I09	8	101178233	101178427	ENSSSCT00000010036	101178233	101180799	ENSSSCG00000009164
12	POR_C043_B16	4	115237159	115237312	ENSSSCT00000007482	115237159	115296733	ENSSSCG00000006831
13	POR_C044_C06	5	15463454	15463639	ENSSSCT00000017099	15463454	15504357	ENSSSCG00000015699
14	POR_C045_J06	6	55114231	55114389	ENSSSCT00000003911	55114231	55131617	ENSSSCG00000003520
15	POR_C045_E17	2	291676	291881	ENSSSCT00000014049	291676	294411	ENSSSCG00000012853
16	POR_C047_E08	4	104450611	104450783	ENSSSCT00000007341	104450611	104483333	ENSSSCG00000006702
17	POR_C050_G06	2	13357967	13358106	ENSSSCT00000014462	13357967	13365882	ENSSSCG00000013242
18	POR_C051_C24	3	53192637	53192801	ENSSSCT00000009003	53192637	53244610	ENSSSCG00000008221
19	POR_C052_J23	8	101178233	101178321	ENSSSCT00000010036	101178233	101180799	ENSSSCG00000009164
20	POR_C052_N17	17	50933043	50933257	ENSSSCT00000008155	50933043	50939068	ENSSSCG00000007451
21	POR_C029_F06	17	59476535	59476758	ENSSSCT00000008201	59476535	59488099	ENSSSCG00000007492
22	POR_C029_I02	12	19392090	19392430	ENSSSCT00000019007	19392090	19401272	ENSSSCG00000017459
23	POR_C036_I16	4	115237159	115237316	ENSSSCT00000007482	115237159	115296733	ENSSSCG00000006831
24	POR_C038_G24	3	53192637	53192853	ENSSSCT00000009003	53192637	53244610	ENSSSCG00000008221
25	POR_C052_C10	4	64451450	64451579	ENSSSCT00000006773	64451450	64490934	ENSSSCG00000006179
26	POR_C029_F06	17	59476535	59476758	ENSSSCT00000008201	59476535	59488099	ENSSSCG00000007492
27	POR_C029_I02	12	19392090	19392430	ENSSSCT00000019007	19392090	19401272	ENSSSCG00000017459
28	POR_C036_I16	4	115237159	115237316	ENSSSCT00000007482	115237159	115296733	ENSSSCG00000006831

Table 9. *Continued*

SN	Clones	Pig chromosome	Clone start-site	Clone stop-site	Pig transcript	Transcript start-site	Transcript end-site	Pig gene
29	POR_B002_K22	3	5466046	5466124	ENSSSCT00000008350	5466046	5472201	ENSSSCG00000007610
30	POR_B006_D08	14	106165402	106165613	ENSSSCT00000011440	106165402	106166847	ENSSSCG00000010454
31	POR_B006_P05	14	106165402	106165690	ENSSSCT00000011440	106165402	106166847	ENSSSCG00000010454
32	POR_B008_C04	4	98064217	98064356	ENSSSCT00000007125	98064217	98071673	ENSSSCG00000006505
33	POR_B008_F05	3	7733351	7733460	ENSSSCT00000008422	7733351	8215757	ENSSSCG00000007681
34	POR_B008_I21	6	38624735	38624890	ENSSSCT00000003556	38624735	38627251	ENSSSCG00000003201
35	POR_B008_L16	9	69522303	69522440	ENSSSCT00000016700	69522303	69611149	ENSSSCG00000015327
36	POR_B008_O06	10	24925344	24925839	ENSSSCT00000011961	24925344	24930635	ENSSSCG00000010929
37	POR_B009_C15	7	20358750	20359191	ENSSSCT00000001186	20358750	20382925	ENSSSCG00000001090
38	POR_B009_D09	1	227339861	227340150	ENSSSCT00000005751	227339861	227361094	ENSSSCG00000005218
39	POR_B010_A16	13	14913077	14913247	ENSSSCT00000012294	14913077	14917024	ENSSSCG00000011227
40	POR_B011_J11	1	190796275	190796403	ENSSSCT00000005554	190796275	190822069	ENSSSCG00000005038
41	POR_B012_J06	1	254770032	254770413	ENSSSCT00000005941	254770032	254774282	ENSSSCG00000005398
42	POR_B013_M09	4	33020194	33020509	ENSSSCT00000006629	33020194	33107518	ENSSSCG00000006040
43	POR_B013_N18	13	112119404	112119556	ENSSSCT00000013064	112119404	112120668	ENSSSCG00000011939
44	POR_B013_P06	2	25555575	25555701	ENSSSCT00000014534	25555575	25592178	ENSSSCG00000013308
45	POR_B014_K07	10	49742656	49742893	ENSSSCT00000012125	49742656	49925053	ENSSSCG00000011079
46	POR_B015_J10	15	123918421	123918786	ENSSSCT00000017712	123918421	124009247	ENSSSCG00000016266
47	POR_B015_M07	8	62038051	62038148	ENSSSCT00000009843	62038051	62129100	ENSSSCG00000008986
48	POR_B016_H14	5	55056963	55057108	ENSSSCT00000000657	55056963	55063481	ENSSSCG00000000611
49	POR_B017_M06	1	226224429	226224546	ENSSSCT00000005739	226224429	226287788	ENSSSCG00000005207
50	POR_B018_L14	8	102084531	102084933	ENSSSCT00000010043	102084531	102175998	ENSSSCG00000009171
51	POR_B019_K05	14	106165402	106165797	ENSSSCT00000011440	106165402	106166847	ENSSSCG00000010454
52	POR_B019_O11	1	136415916	136416307	ENSSSCT00000005247	136415916	136421724	ENSSSCG00000004748

In addition, multiple cDNA clone sequences that provided the same gene transcript or genes were assessed to determine their genomic position in comparison with the transcripts or genes. For example, ENSSSCT00000003148 transcript of the *F1RP04_PIG* pig gene is located on pig chromosome 6:25,348,651-25,390,621:-1. The gene has a single transcript comprising 11 exons. A hit of this transcript was obtained from six different cDNA clones (Table 7). The locations of all the exons were compared with that of the cDNA clones. All the clone sequences are located on the pig chromosome 6: 25,390,640-25,390,379 (Figure 3). The 1st exon of the transcript is 243 bps long and located at 6:25,348,651-25,349,203:-1. The start-site of one of the clone sequences and the first exon is the same (25,390,379). However, other cDNA clones are also within the first exon but their start-site has slight variations.

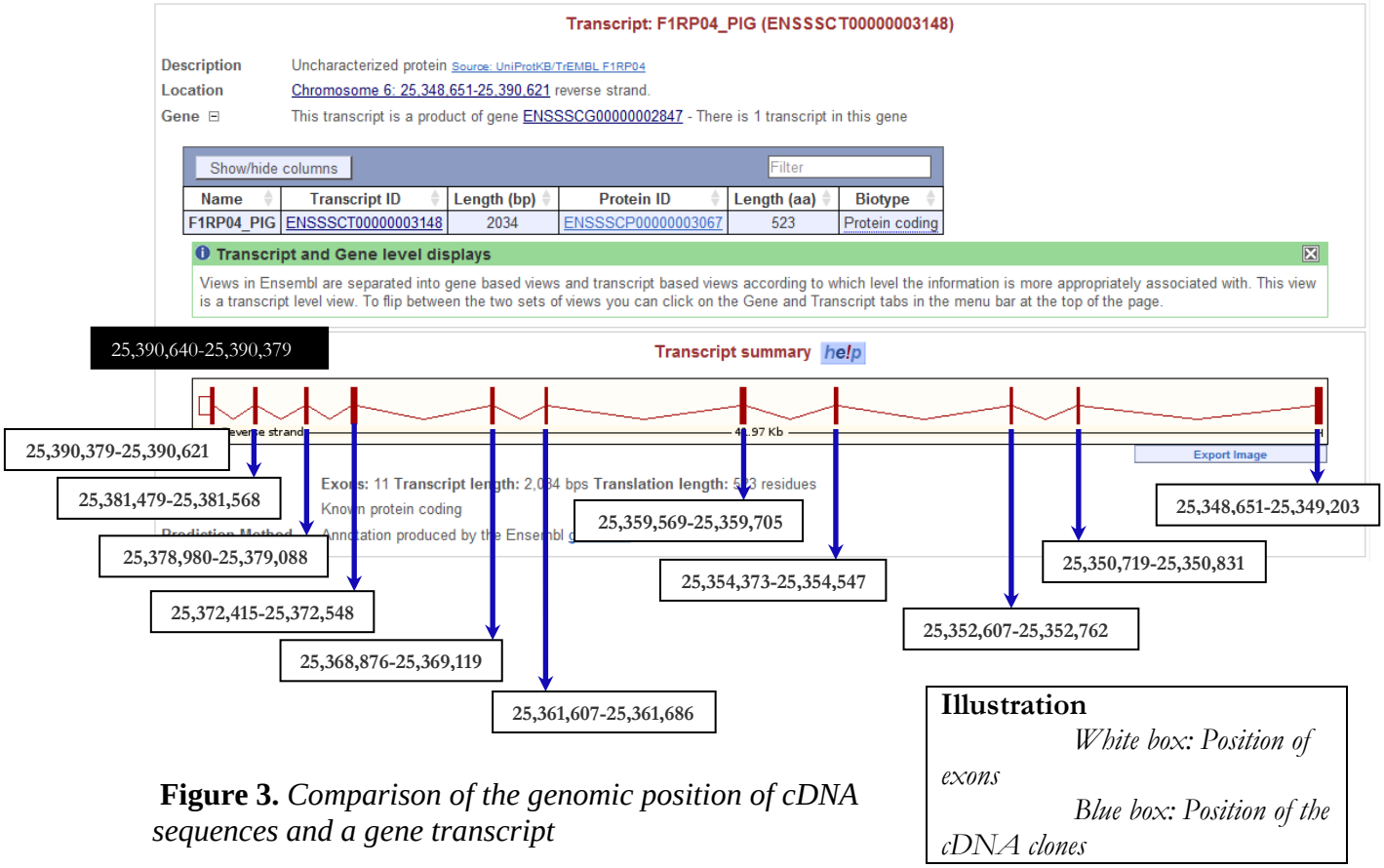


Figure 3. *Comparison of the genomic position of cDNA sequences and a gene transcript*

Similarly, the *F1SPG2_PIG pig* gene has a single ENSSSCT00000012709 transcript with 10 exons. The gene is located in the pig chromosome 13: 59,216,637-59,233,111:1. The first exon of the transcript is 313 bps long and located at 13: 59,216,637-59,216,949 and the last exon at 13:59,232,534-59,233,111. This gene was obtained from 133 different cDNA clone sequences. Among the total clone sequences, only POR_C039_E07 shown a minimum variation of start-site (13: 59,216,641) (Table 10) which is only 4 nucleotides after the start-

site of the first exon. The rest of the clones shown larger variation of start-site.

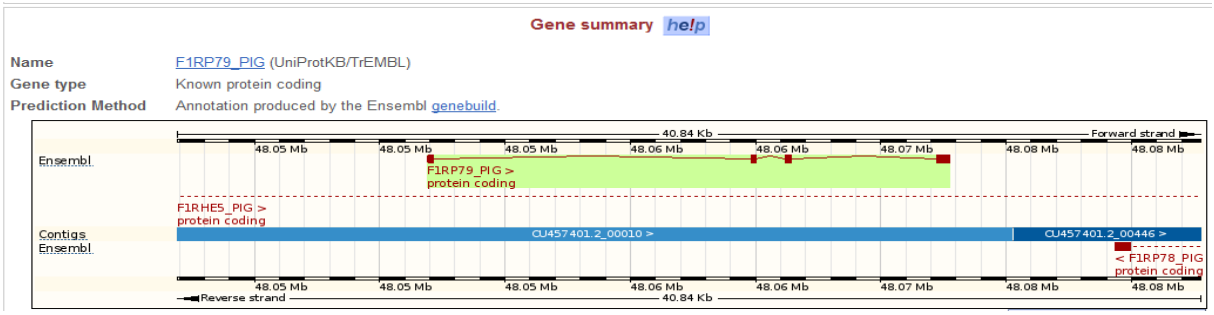
Table 10. *Start-site of the clone sequences*

cDNA Clones	Pig Chromosome	Clone Star-site	Clone Stop-site
POR_C027_P11	13	59,216,667	59,216,948
POR_C039_K22	13	59,216,643	59,216,901
POR_C039_E07	13	59,216,641	59,216,935
POR_B001_E04	13	59,216,688	59,216,950
POR_B001_M04	13	59,216,683	59,216,953
POR_B001_N04	13	59,216,723	59,216,889
POR_B001_O04	13	59,216,719	59,216,946
POR_B001_O06	13	59,216,718	59,216,946

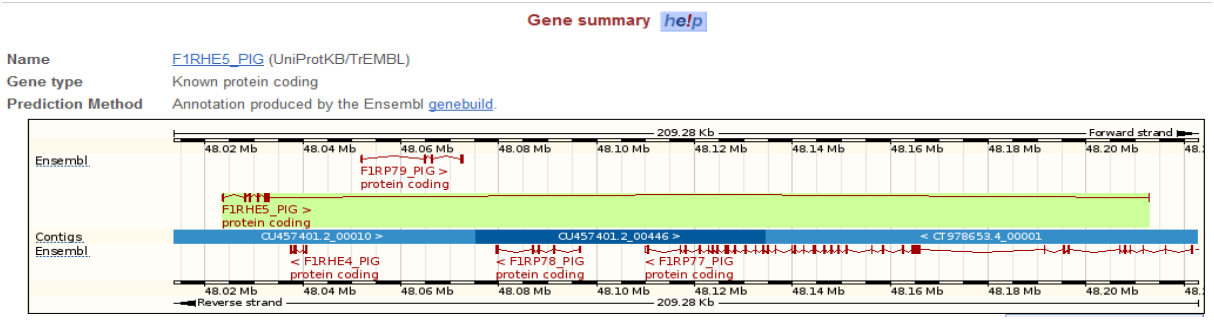
A sequence from POR_C030_B06 clone was located at 11: 48,051,963-48,051,903 provided ENSSSCT00000010384 transcript that belongs to the *F1RP79_PIG* pig gene located at 11: 48,051,903-48,072,745:1. The gene has a single transcript with four exons. The start-site of the clone was 60 nucleotides before the start-site of the predicted transcript. The genomic region that contain the above 60 nucleotides found annotated with another gene, *F1RHE5_PIG* located at 11: 48,023,540-48,212,816 forward strand. A 179,683 bps long 4-5 intron of the *F1RHE5_PIG* gene is located at 11: 48,033,114-48,212,796. The last (5^{th}) exon of this gene is located at 11: 48,212,797-48,212,816:1. This shows that the 60 nucleotides before the start-site of the *F1RP79_PIG* gene are contained within the intronic region of the *F1RHE5_PIG* gene. In addition, the *F1RP79_PIG* gene is contained in the region of the *F1RHE5_PIG* gene (Figure 4a, 4b). This implies that the region might cover sequences of a single gene that consist several transcripts.

Table 11. *The start-site of F1RP79_PIG and F1RHE5_PIG pig genes*

Genes	Start-site	End-site	Number of exons	Length (bps)
F1RP79_PIG	48,051,903	48,072,745	4	20,842
F1RHE5_PIG	48,023,540	48,212,816	6	189,276



a) *F1RP79_PIG gene*



b) *F1RHE5_PIG genes*

Figure 4. *Genomic region of the F1RP79_PIG and F1RHE5_PIG genes*

On the other hand, a sequence from clone POR_C029_N04 located at 15: 15,947,998-15,948,052 provided ENSSSCT00000017102 transcript belonging to the *LOC100512195* pig gene located at 15: 15,948,052-15,948,435:1. The start-site of the clone sequence was found 54 nucleotides after the start-site of the transcript. The gene has a single transcript with one exon with 384 bps length. Further, a sequence from clone POR_B024_J03 was located at 14: 57750130-57750368:1 and provided a ENSSSCT00000011123 transcript that belongs to the *F1RGW7_PIG* pig gene located at 14: 57748954-57750845:1. The start-site of the clone was 1,176 nucleotides before the star-site of the transcript. The gene has one transcript with three exons. The gene is bound by two coding genes of *TOMM20* at 14: 57,317,608-57,329,790:1 and *F1RGW5_PIG* at 14: 57,868,453-57,946,346:1 (Figure 5). The region of the chromosome between the above two genes was not reported for transcribed sequences. The length of the non-coding region ranging from the stop-site of the *TOMM20* gene to the start-site of the *F1RGW7_PIG* gene is 419,164 bps including the 1,176 nucleotides before the start-site of the *TOMM20* gene. This indicates that the cDNA clone could be part of one of the two genes, either part of the first-exon of the *TOMM20* gene or the last exon-of the *F1RGW7_PIG* gene.

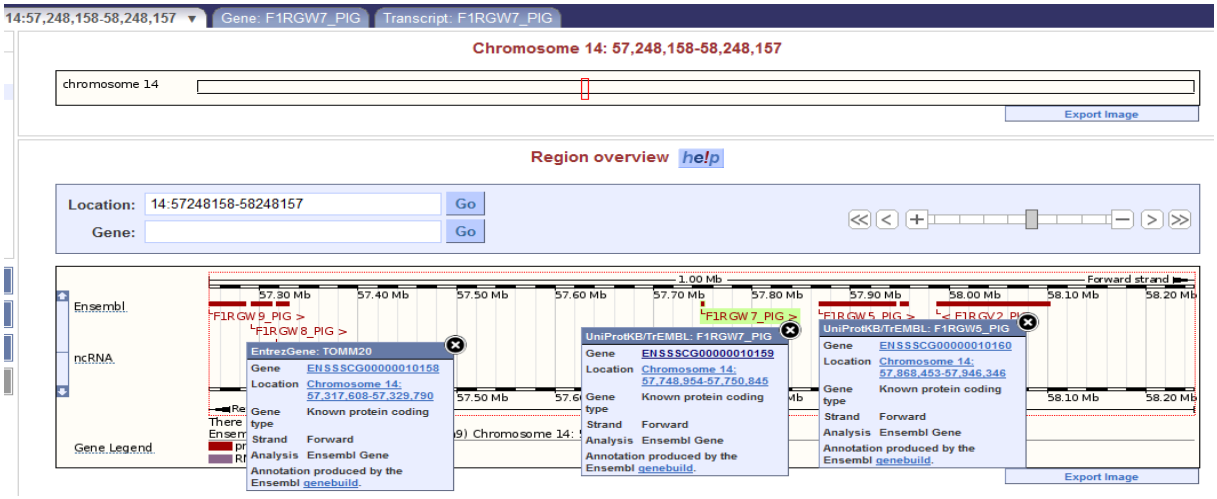


Figure 5. *Two genes that bound the F1RGW7_PIG* gene

Similarly, a sequence from POR_C038_J12 clone is located at 4: 114,202,948-114,203,083 provided a hit of ENSSSCT00000007456 transcript of the *F1S621_PIG* pig gene (ENSSSCG00000006808) at 4: 114,204,503-114,214,126:1. The start-site of the clone was 1,555 nucleotides before the start-site of the transcript. The gene is located between coding genes of *HBXIP_PIG* (ENSSSCG00000006807) at 4: 114,190,407-114,191,587:1 and *F1S619_PIG* (ENSSSCG00000006809) at 4: 114,230,643-114,237,676:-1 (Figure 6). The length between the stop-site of *HBXIP_PIG* gene and start-site of the *F1S621_PIG* gene is 12,916 bps. The genomic region that lies between the stop-site of the last exon of the previous gene and the star-site of the first exon of subsequent gene is 12,916 bps. This indicates that the start-site of the *F1S621_PIG* gene could be 1,555 bps before the predicted start-site of the first exon of the gene.

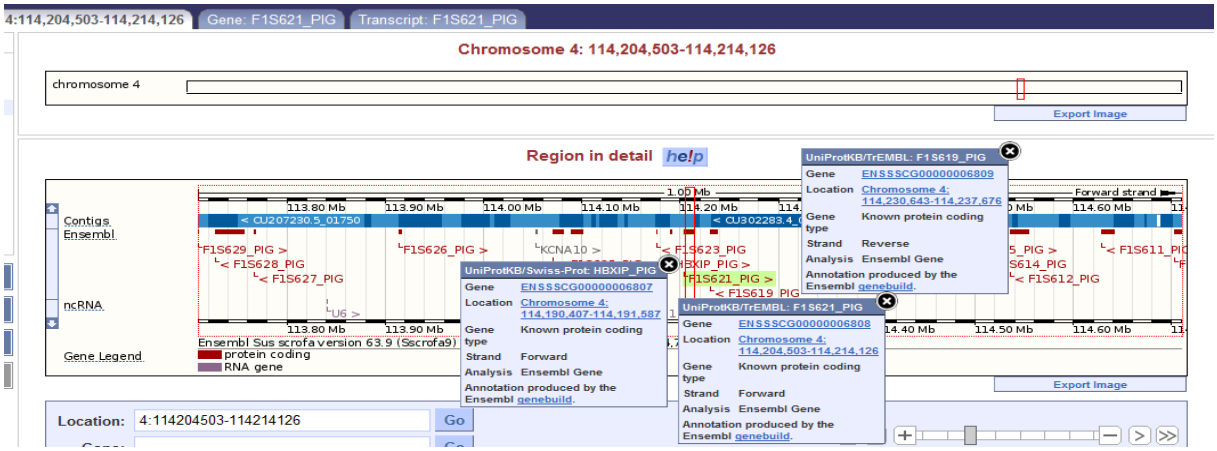


Figure 6. *F1S621_PIG gene surrounded by two other genes*

Again, a sequence from POR_C045_O04 clone at 11: 67,570,348-67,570,592:1 provided a hit

of ENSSSCT00000010423, a single transcript of the *LOC100155367* pig gene (ENSSSCG00000009506) located at 11: 67,570,524-67,570,841:1. The start-site of the clone is after 176 nucleotides of the start-site of the transcript. The transcript has a single 318 bps long exon and this shows that the start-site of the clone sequence is beyond the middle of the exon. Moreover, a sequence from POR_C050_L19 clone was located at 6: 114,714,014-114,714,281 and provided a ENSSSCT00000004286 transcript of the *LOC100518257* pig gene (ENSSSCG00000003874) at 6: 114,714,243-114,748,187:1. The star-site of the clone is 229 bps after the start-site of the transcript. A 391 bps long first exons of the transcript is located at 6: 114,714,243-114,714,633:1. Added, a sequence from POR_B024_B21 clone at 14: 140,701,050-140,701,281 provided a hit of ENSSSCT00000011746 transcript of the gene F1SDP3_PIG (ENSSSCG00000010734) at 14: 140,693,803-140,716,055:1. The start-site of the clone is found after 7,247 nucleotides of the start-site of the transcript. This indicates that the clone sequence lies between the end of the 2-3 intron at 14: 140,699,486-140,701,052 and the start of the 3-4 intron at 14: 140,701,281-140,705,705. The clone sequence matches with 4th exon 14:140,705,706-140,705,930 including few nucleotides from the bounding introns.

A sequence from POR_C024_C02 clone provided hit from pig genome at 10:19,404,958-19,405,418:1. However, the clone has no hit from pig cDNA database and thus the gene information not available. The sequences are located between two predicted genes located at 10:19,101,299-19,103,143 and 10:19,531,455-19,531,856 (Figure 7). Similarly, a sequence from POR_B024_O14 provided hit in pig genome at 5: 3,537,119-3,537,441. Again, it has no hit from pig cDNA and there is no gene information.

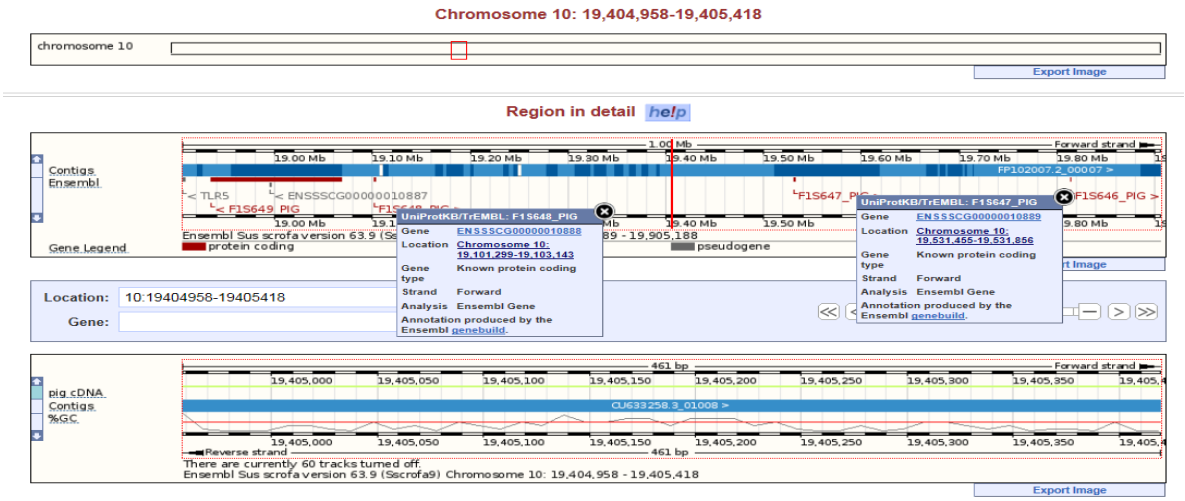


Figure 7. *The clone sequence located in non-coding region*

4.5 Identification of Homologous Pig Genes

One of the major goals of understanding the function of pig genome is to smooth path of human biomedical and evolutionary research. This is facilitated through identification of functionally conserved sequences that involved in genetic mechanisms of diseases and quantitative traits. Pig is evolutionary closely related mammal especially with human and also mouse. Compared to pig genome, the human and mouse genome are much comprehensively sequenced and annotated. Thus, understanding function of the pig genome could highly benefit from the vast information available from human and mouse genome. Our cDNA clone sequences provided considerable unique hits in human and mouse cDNA databases. Thus, the homologous pig genes were searched using the human and mouse genes obtained from their unique hits. For example, a clone sequence from POR_C034_D16 gave a unique ENST00000522709 human transcript. This transcript belongs to the protein-coding *EFR3A* human gene (ENSG00000132294) located at 8: 132,916,335-133,025,889:1 and the homologous pig gene is *F1RRT7_PIG* at 4:7,997,948-8,041,788:-1 (Figure 8).

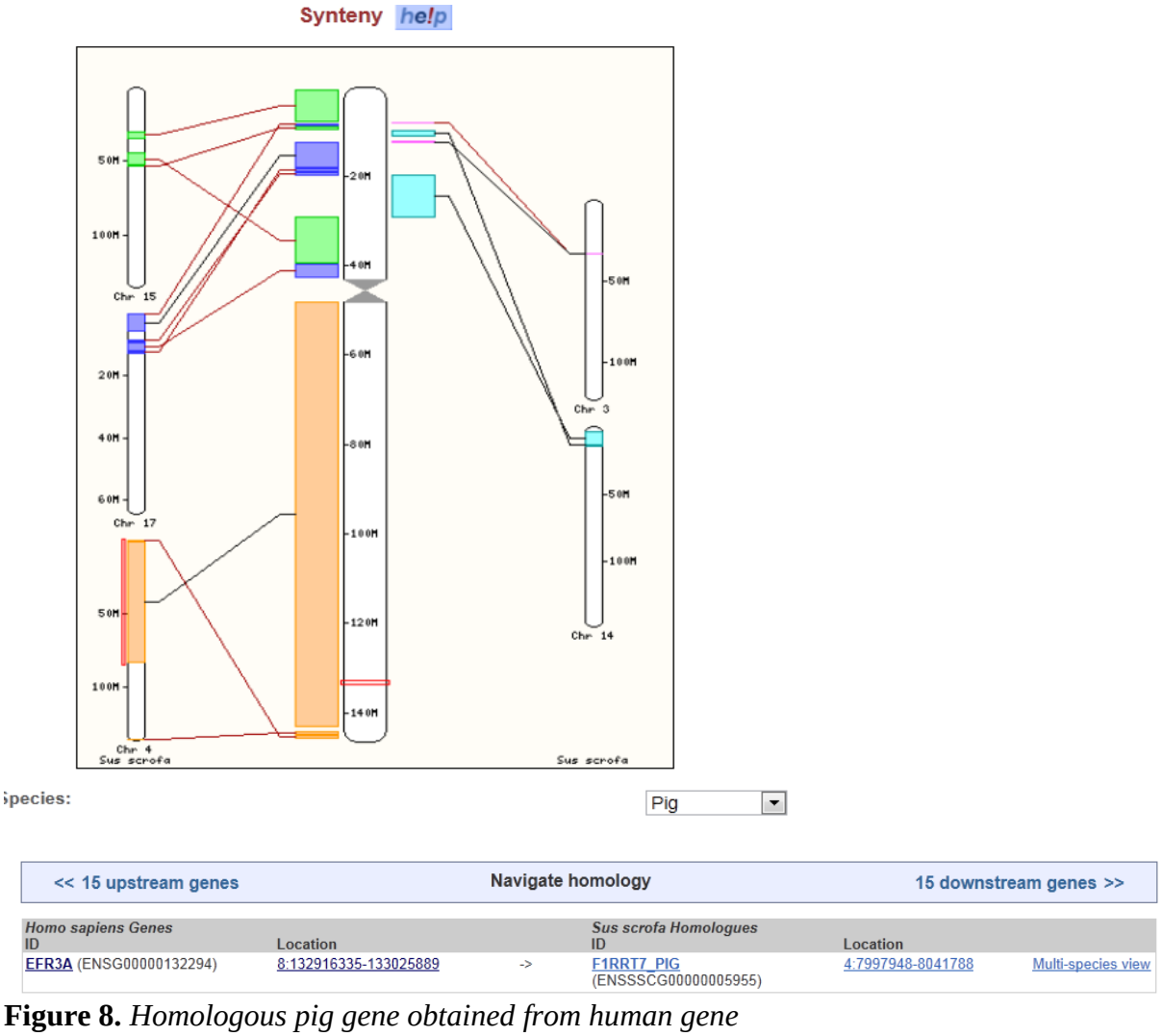


Figure 8. *Homologous pig gene obtained from human gene*

Likewise, a POR_C033_L12 clone provided unique hit of ENSMUST00000028087 mouse transcript located at 2: 39,049,874-39,081,971. It is the first transcript of the known protein-coding *Ppp6c* mouse gene (ENSMUSG00000026753) at 2:39,049,874-39,081,971:-1. A corresponding human homologous gene is *PPP6C* (ENSG00000119414) located on human chromosome 9: 127,908,852-127,952,218:-1. Since the mouse genes has no homologous pig gene in ENSEMBL tool and thus we first identified the human homologous gene and finally the pig homologous gene was determined from the human gene (Figure 9). Accordingly, the homology pig gene was found to be *PPP6C* located at 1:280,745,798-280,769,960:1 (Fig 10).

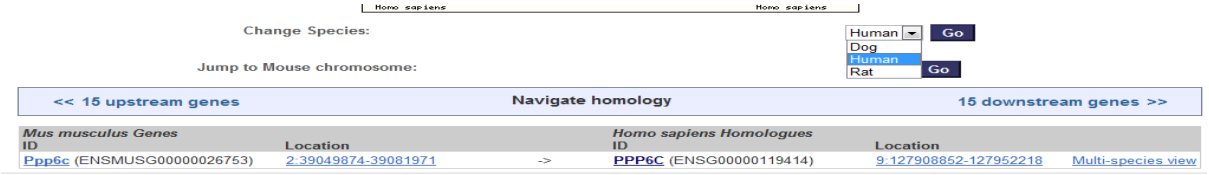


Figure 9. *Mouse homologous genes available only for human and dog*

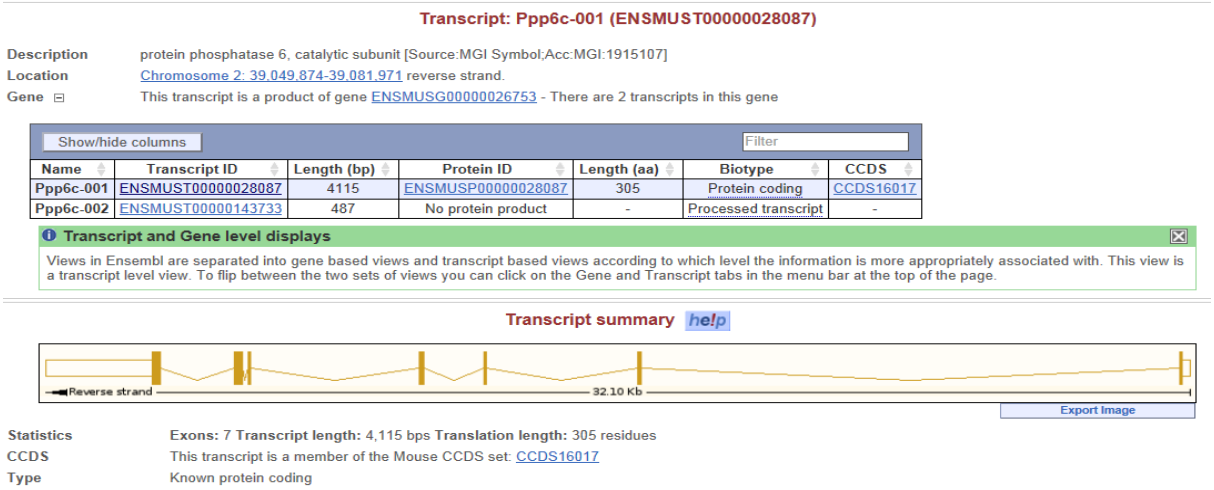


Figure 10a. *Mouse transcript*

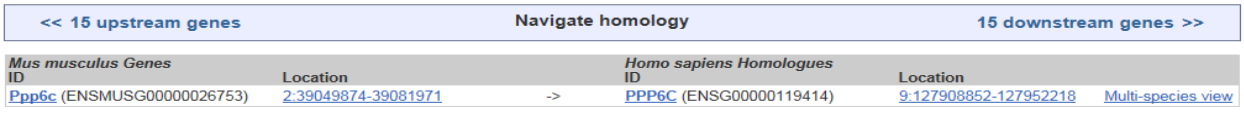


Figure 10b. *Homologous human gene*

Figure 10c. *Homologous pig gene*

Figure 10. *Pig homologous gene obtained from human and mouse gene*

Further, a clone sequence from POR_C040_H11 has given a unique ENST00000297029 human transcript. It belongs to the human *SCIN* gene located at 7: 12,610,203-12,693,228:1. The homologous pig gene is *F1SF64_PIG* located at 9:77,079,774-77,108,450:1. This transcript was also obtained from another clone, POR_C030_P19. Again, a clone sequence from POR_C044_E08 provided a unique hit of human ENST00000367362 transcript which is the first transcript of the human R5A2 (ENSG00000116833) gene at 1: 199,996,730-200,146,552:1. The pig homologous gene is *F1S5D5_PIG* located at 10:23,687,243-23,820,903:1.

On the contrary, some unique hits from both human and mouse cDNA databases have not provided homologous genes in pig thus the function of the pig homologous sequences could be drawn from the evolutionary related mammals. For example, a sequence from clone POR_C031_F08 provided a unique ENST00000481435 human transcript which belongs to the human SEC62 (ENSG00000008952) gene at 3: 169,684,423-169,716,161:1 consisting seven protein-coding and five noncoding transcripts. However, a homologous pig sequence was not found from the human sequence. The human SEC62 gene is speculated to perform post-translational protein translocation into the endoplasmic reticulum (ER) membrane, backward transport of ER proteins that are subject to the ubiquitin-proteasome-dependent degradation pathway and its encoded protein is an integral membrane protein located in the rough ER. Accordingly, the homologous pig gene could have a similar function. Similarly, a sequence from POR_C051_C11 provided a unique ENSMUST00000147232 mouse transcript of the mouse *Gripap1 (ENSMUSG00000031153)* gene at X: 7,366,891-7,397,693:1. The human homologous gene is *GRIPAP1* (ENSG00000068400) at X: 48,830,134-48,858,675:-1. However, a pig homologous gene was not found. The homologous human gene encodes a guanine nucleotide exchange factor for the Ras family of small G proteins (RasGEF). In brain studies, the encoded protein was found with the GRIP/AMPA receptor complex. Further, a sequence from POR_C049_G18 provided a unique hit of mouse ENSMUST00000078220 transcript which belongs to Eif4h mouse gene at 5: 135,095,747-135,115,218:-1. The human homologous gene is EIF4H (ENSG00000106682) at 7:73588575-73611431:1. But there is no

pig homologous gene. The human homologous gene encodes one of the translation initiation factors, which functions to stimulate the initiation of protein synthesis at the level of mRNA utilization. This gene is deleted in Williams's syndrome, a multisystem developmental disorder caused by the deletion of contiguous genes at 7q11.23. Again, some the unique human transcripts including ENST00000072516, ENST00000223136, ENST00000373816, ENST00000370685, ENST00000379805, ENST00000316377, ENST00000495607 and ENST00000455593 have not provided homologous pig gene.

5. DISCUSSION AND CONCLUSION

The objective of this study was to one-pass 5′-end sequencing and characterization of 10,000 cDNA clones from the normalized full-length library of the pig. The next generation sequencing technology can generate a large ESTs data but constructing and sequencing cDNA libraries has a unique role that it provides a physical clones for further study (Natarajan *et al*., 2010). Screening cDNA library with large number of bacterial colonies requires a rapid and cost-effective method that provides enhanced sequence quality. The currently available sequencing procedures, for example, direct sequencing using 96- and 384-well plates from Applied Biosystems (PE Biosystems, Foster City, CA, USA) (Poster 113199; Applied Biosystems, www.appliedbiosystems.com) is commonly applied in direct sequencing of lysed bacterial cells (Ganguly *et al*., 2005). A direct sequencing involving preparing template and sequencing reactions of normalized cDNA libraries in 384-well plates was reported to be rapid, inexpensive and reliable especially for generating large expressed sequence tag data (Smith *et al*., 2000). Similarly, a recent full-length cDNA library characterization in our laboratory using direct sequencing in 384-well plates also reported that the method is efficient, cost-effective and less laborious (Bernal *et al.*, 2011).

Together with the previous sequencing, a total of 13,989 useful reads were obtained with an overall success rate of 70.0%. The success rate of sequencing in a 384-well plate from previous study (71.21%) and the current data (69.34%) is not different from similar activity that reported 75%–80% (Smith *et al*., 2000). We have obtained lower success rates from 384-well plates that were stored in the freezer for longer time before sequencing reaction. Reduced success rate and poor quality or short read length are associated with poor growth or differential growth rates of bacterial clones during culturing (Das *et al*., 2001; Ganguly *et al*., 2005). There are recent development of replacing a manual cloning and sequencing procedure with automated and computer-aided method that involve automated transformation, inoculations, plasmid and PCR product purification and sequencing (Yehezkel *et al*., 2011). This could be an alternative option to be sought especially for large-scale collection and characterization of cDNA libraries.

A total of 13,989 clone sequences of at least 50 bps or more were generated from a normalized full-length cDNA library of the pig. The clone sequences provided a total of 12,220 hits in one or more of the pig, human or mouse databases. Blasting against the pig

genome provided larger hits of 10,857. In addition, the pig cDNA database also provided 6,597 hits. The human and mouse cDNA provided a total hits of 4,786 and 2,801, respectively, that enable comparative analyses to identify the homologous pig genes. A total of 12.6% clone sequences have not provided hit against the reference databases. In addition, considerable clone sequences provided unique hits of 3,439, 813 and 27, respectively in pig genome, human cDNA and pig cDNA databases. Both the unique hits obtained from specific databases or clones without hits implies that the expressed sequences of the pig genome were not yet fully identified and annotated.

The clone sequences represents the expressed parts of the pig genome but several of them have not provided hit at all or some unique hits in specific databases. The currently available genome assembly (Sscrofa9) covers only 88% of the pig genome sequence (Martien A.M. Groenen, personal communication). Thus, the absence of mismatches between the cDNA sequences and that of pig genome and cDNA databases is mainly due to the fact that the current genome built not represents the complete sequence of pig genome. Though many groups have reported characterization and deposition of the pig ESTs generated from cDNA libraries, it is still far from complete and thus several of our clones have not provided matches in the pig cDNA database. In addition, most of the ESTs available in the public database were generated from standard cDNA libraries that often encompass the vast majority of highly expressed genes and missing ESTs from moderately or rarely expressed genes. Thus, it is important to construct and sequence full-length cDNA library that provide a complete view of the transcribed pig genome. Furthermore, the current effort to upgrade the genome sequence from Suscrofa9 to Sscrofa10 could provide hits from our cDNA clones which currently lack matches.

A total of 3,164 pig genes were obtained from our library, representing only 18.2% of the total protein-coding pig genes. The current Ssscrofa9 assembly shows that the pig genome sequence consists of 17,463 known protein-coding genes, 22,050 gene transcripts and 159,909 gene exons (http://www.ensembl.org/Sus_scrofa/Info/Index?db=core) (July 2011). Our library was prepared from only 11 pig clone tissues at fetal stage. Thus, it is essential to construct and characterize more libraries from cells and tissues involved in biological systems of the pig and also representative of the various developmental and physiological states. This could help to obtain a complete profile of the expressed sequences of the genome sequence.

The sequencing of our library has shown that the start-site between the cDNA sequences and

their respective predicted genes was significantly different. It was found that only 52 cDNA sequences have the same start-site with their corresponding transcripts. Moreover, only 999 clones have shown a variation of 1 to 60 nucleotides either before or after the star-sites between the cDNA clones and their respective predicted transcripts. The proportion of cDNA clones that have the same start-site site and those varying from 1 to 60 nucleotides before or after the start-site of the transcripts represents only 8.6 percent of the total clones that provided hit in the references databases. In addition, the start-site of the cDNA sequences was also found located either before the start-site of the predicted gene, within the predicted exons, within the intron of the gene or in regions that were not reported for any transcribed genes. With extensive examples from our sequenced clones we have shown that most of the predicted sequences are not consistent with the actual information we generated through sequencing of the expressed sequences. Thus, a large-scale collection, sequencing and characterization of the cDNA clone could verify the structural and functional features of the predicted sequences and contributes toward understanding and annotation of the pig genome.

Fahrenkrug *et al* (2002) highlighted that the comparison of pig ESTs with sequences of other species provides an important tool for developing comparative map of the pig genome. The search for conserved sequences using the cDNA clone that provided hit only in the human and mouse databases provided homologous pig genes. These genes were not obtained directly from the sequence similarity search mainly because our library provided only 3,164 pig genes and the majority of the predicted genes were not represented. In addition, most of the predicted pig genes were not established through sequencing of the expressed sequences. This is particularly important in annotation of pig genome because majority of the pig gene were obtained through prediction programs. The blasting outputs have shown that most of the hits unique to pig genome lack similarity in the pig annotated transcripts. A comprehensive ESTs from normalized full-length cDNA library are available for human and mouse and thus, comparative analyses of the pig clone sequences to human and mouse cDNA databases highlight the homologous conserved regions (conserved segment or block) from their common ancestor before speciation. The identification of functionally conserved sequences has an unprecedented opportunity to understand the function of the in pig genome sequences thereby provide an insight into the genetic mechanism of quantitative traits and heritable diseases and also evolutionary footprint of the pig.

REFERENCES

Altschul S, Gish W, Miller W, Myers E, Lipman D: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403-410.

Archibald, A. L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M. A. M., Harlizius, B., et al. 2010. Pig genome sequence - analysis and publication strategy. *BMC Genomics, 11*(1).

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell, J and Sayers EW. 2011. GenBank. Nucleic Acids Res, 39: D32–D37.

Bernal S., Crooijmans R. and Groenen M. 2011. Characterization of a normalized full-length cDNA library from a cloned Pig. MSc Thesis. Animal Breeding and Genomics Centre, Wageningen University, The Netherlands.

Bonnet, A., Iannuccelli, E., Hugot, K., Benne, F., Bonaldo, M. F., Soares, M. B., et al. 2008. A pig multi-tissue normalised cDNA library: Large-scale sequencing, cluster analysis and 9K micro-array resource generation. *BMC Genomics, 9.*

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., et al. (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Research, 10*(10), 1617-1630.

Das, M., Harvey, I., Chu, L., Sinha, M. and Pelletier, J. 2001. Full-length cDNAs: more than just reaching the ends. *Physiol. Genomics* 6:57-80.

DNAFORM. 2009. Report on Construction of Normalized Full-length cDNA Library of Pig, Yokohama City, Kanagawa, Japan.

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

Ganguly, T., Chen, P., Teetsel, T., Zhang, L.P., Papaioannou, E. and Cianciarulo, J. 2005. High-throughput sequencing of high copy number plasmids from bacterial cultures by heat lysis. BioTechniques 39:304-308.

Gorodkin, J., Cirera, S., Hedegaard, J., Gilchrist, M. J., Panitz, F., Jørgensen, C., et al. 2007. Porcine transcriptome analysis based on 97 non-normalized cDNA libraries and assembly of 1,021,891 expressed sequence tags. *Genome Biology, 8*(4).

Harbers, M. (2008). The current status of cDNA cloning. *Genomics, 91*(3), 232-242.

Hayashizaki, Y., Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature, 420*(6915), 563-573.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., et al. (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biology, 2*(6).

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 Research 12, 53–62.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., et al. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature, 409*(6821), 685-689.

Kim, T., Kim, N., Lim, D., Lee, K., Oh, J., Park, H., et al. (2006). Generation and analysis for large-scale expressed sequence tags (ESTs) from a full-length enriched cDNA library of porcine backfat tissue. *BMC Genomics, 7*.

Kuroshu, R. M., Watanabe, J., Sugano, S., Morishita, S., Suzuki, Y., & Kasahara, M. (2010). Cost-effective sequencing of full-length cDNA clones powered by a de novo-reference hybrid assembly. *PLoS ONE, 5*(5).

Lee, K., Byun, M., Lim, D., Kang, K., Kim, N., Oh, J., et al. 2009. Full-length enriched cDNA library construction from tissues related to energy metabolism in pigs. *Molecules and Cells, 28*(6), 529-536.

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engström, P. G., et al. (2006). Transcript annotation in FANTOM3: Mouse gene catalog based on physical cDNAs. *PLoS Genetics, 2*(4).

Natarajan, P., Kanagasabapathy, D., Gunadayalan, G., Panchalingam, J., Shree, N., Sugantham, A.P., Singh, K.K. and Madasamy, P. 2010. Gene discovery from Jatropha curcas by sequencing of ESTs from normalized and full-length enriched cDNA library from developing seeds. BMC Genomics, 11:606.

Nguyen, D. T., Oh, Y., Dirisala, V. R., Choi, H., Park, K., Kim, J., et al. 2010. A simple, rapid, efficient and inexpensive strategy for sequencing clones from cDNA libraries. *Biotechnology and Bioprocess Engineering, 15*(5), 817-821.

Oshikawa, M., Sugai, Y., Usami, R., Ohtoko, K., Toyama, S., & Kato, S. (2008). Fine expression profiling of full-length transcripts using a size-unbiased cDNA library prepared with the vector-capping method. *DNA Research, 15*(3), 123-136.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., et al. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genetics, 36*(1), 40-45.

Sanger, F., Nicklen, S. and Coulson, A.R. 1977. DNA sequencing with chain terminating inhibitors. Proc. Nat. Acad. Sci. USA, 74, 5463-5467.

Shcheglov, A. S., Zhulidov, P. A., Bogdanova, E. A. and Shagin, D. A. 2007. Normalization of cDNA libraries. A. Buzdin and S. Lukyanov (eds.), *Nucleic Acids Hybridization,* 97–124.

Smith, T. P. L., Godtel, R. A., & Lee, R. T. 2000. PCR-based setup for high-throughput cDNA library sequencing on the ABI 3700^{TM} automated DNA Sequencer. *BioTechniques, 29*(4), 698-700.

Uenishi H, Eguchi T, Suzuki K, Sawazaki T, Toki D, Shinkai H, Okumura N, Hamasima N, 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:D484-8.

Uenishi, H., Eguchi-Ogawa, T. Shinkai, H, Okumura, N., Suzuki, K., Toki, D., Hamasima, N. and Awata, T. (2007) PEDE (Pig EST Data Explorer) has been expanded into Pig Expression Data Explorer, including 10 147 porcine full-length cDNA sequences. *Nucleic Acids Res.*, 35, D650-3.

Whitworth, K., Springer, G. K., Forrester, L. J., Spollen, W. G., Ries, J., Lamberson, W. R., et al. (2004). Developmental expression of 2489 gene clusters during pig embryogenesis: An expressed sequence tag project. *Biology of Reproduction, 71*(4), 1230-1243.

Yehezkel, T.B., Nagar, S., Mackrants, D., Marx, Z, Linshiz, G., Shabi, U. and Shapiro, E. 2011. Computer-aided high-throughput cloning of bacteria in liquid medium. BioTechniques, 50 (2): 124–127.

Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, et al. 2008. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*36:D793–799.

Yamasaki, C., Murakami, K., Takeda, J., Sato, Y., Noda, A., Sakate, R., et al. (2009). H-InvDB in 2009: Extended database and data mining resources for human genes and transcripts. *Nucleic Acids Research, 38*(SUPPL.1), D626-D632.

Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, et al. (2006) Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs. PLoS Genet 2(4): e62. doi:10.1371/journal.pgen.0020062.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., et al. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature, 409*(6821), 685-689.

APPENDIX

Appendix 1. Laboratory Protocols

1.Media Preparation:

1.1: Lysogeny Broth (LB) medium: to prepare 1 L

a. To 800 ml MQ water add:

10 g Bactotryptone

5 g yeast extract

10 g NaCl

b. Adjust pH to 7.5 with NaOH

c. Adjust volume to 1 L with MQ water

d. Sterilize by autoclaving

1.2: 1% Agar in Lysogeny Broth (LB): to prepare 1 L

a. 10 g agar (Bacto-agar) in 1 L of LB

b. Sterilize by autoclaving

c. When the medium reaches approximately 50^0C, add 1000L ampicillin (final concentration of 10mg/ml)

1.3: 10x Freezing Medium: to prepare 1 L

a. To prepare solution A:

360 mM k_2HPO4 (mw 174.18): 62.7 g

132 mM KH_2PO (mw 136.09): 17.96 g

Fill up to 160 ml with H_2O

Sterilize by autoclaving

b. To prepare solution B:

17 mM Na citrate (mM 294.11): 4.99g

4 mM $MgSO_4$ (mw 132.15): 0.99 g

68 mM (NH4)2 SO4 (mw 132.15) 8.99 g

Fill up to 400 ml and autoclave

c. Solution C:

Autoclave 440 ml of glycerol

d. Mix all the sterilized solutions in an horizontal laminar flow workstation

2: Plasmid Culturing

a. Four petridishes were filled with 50 ml of LB agar media (LB+ampicillin+agar)

b. 2-3µL library stock solution was diluted into 400 µL LB + ampicillin

c. 100 µL diluted libraries were spread on 140 mm petri dishes with glass beads

d. Bacteria were cultured in incubator at 37 ^{0}C overnight (18 hrs)

3: Preparation 384-well Plates and picking individual clones

a. 400 ml LB media was mixed with 400 µL ampicillin (10mg/ml)

b. 45 ml LB+Amp was mixed with 5 ml freezing media (FM). FM was for storing the plate in the -80^{0}C freezer.

c. The master plate (plate labelled A) POR_A25) was filled with the 100 µl of LB+Amp+FM

d. Individual colonies from the petridishes were transferred to the master plates using sterilized wooden sticks

e. The master plates were incubated at 37 ^{0}C overnight (18 hrs)

4: Making replica

a. For individual master plates, two replicas of 'B' and 'C' were prepared. B plates were filled with LB+AMP+FM whereas C plates were filled with only LB+AMP.

b. The sterilized replicator was placed into the master plate (A) in order to make a copy into the plate B and again into the final plate C. Both plates were grown overnight in incubator (18 hrs) at 37 °C.

c) Both A and B plates were cooled dawn to a room temperature for 30 minutes and finally stored at -80^{0}C

d) C plates were used for sequencing.

5: Preparing cell lysate

a. Cells were pelleted by spinning in the centrifuge for 20 min at 2000X g and 20°C

b. Invert the plate onto successive layers of paper towels

c. Add 25 µL MQ/well to wash the remaining medium

d. Centrifuge 5 min at 2000Xg at 20°C

e. Invert the plate onto successive layers of paper towels and remove the medium

f. Add 25 µL water per well

g. Centrifuge 5 min at 2000Xg at 20 °C

h. Invert the plate onto successive layers of paper towels and remove the medium

i. Suspend the pellet in 25 µL water/well

j. Vortex the plates (table vortex)

k. Transfer the cell suspensions to a new 384-well PCR plate and heat seal

l. Spin down shortly

m. Denaturing of the cell lysates was performed for 5 minutes at 95°C

n. Plates were transferred directly on ice

o. Centrifuge 10 min at 1000Xg at 4 °C

6: Sequencing reaction

a. Dilute the primer (T3 in this case): primer working solutions are around 40 pmol/µL, so each PCR primer should be diluted ±1:50

b. Add 2 µL from the lysate to a new 384-well plate

c. Prepare the master mix for sequencing reactions

Reagent	**Volume(µL)/1 sample**	**Volume(µL)/400 samples**
Cell lysate	2 (lysate in the new plates)	
5X sequencing buffer (5xB)	0.75	600
BD 3.1 (BigDye)	0.5	400
MQ	0.75	600
Primer (0.8 pmol/µL)	1	800
Total	1	2400

d. Add 3 µL of master mix to each well in a new 384-well PCR plate

e. Seal the new plates with heat sealing

f. Spin down shortly

g. Perform cycle sequencing as follows (Program: BD50x50).

Step	**Temperature (°C)**	**Time**	**Number of cycles**
1	95	5	1
2	96	30	50
	50	10	
	60	4	
3	4	∞	∞

7: Precipitation of DNA

a. 5 µl sequencing reaction

b. Add 0.5 µL NaAc-EDTA (1.5 M sodium acetate (pH > 8.0) and 250 mM EDTA)

c. Spin down shortly

d. Add 17 µL EtOH (-20 degrees) using 125 µL electronic pipette with 16 tips

e. Seal with aluminium (not heat sealing)

f. Mix by vortex (with Illumina vortex for 30 seconds

g. Incubate the plates for 30 minutes in the ice box

h. Centrifuge the plates for 30 minutes 3000g at 4 °C

i. Remove the aluminium seal

j. Centrifuge upside down for 1 minute 700g at 4 °C

k. Add 10 µl of formamide to each well of the plates

l. Transfer the solution into a barcode plate

j. Seal the plates with the sequencer devices and run the barcode plate with samples in the ABI3730

Appendix 2: Plates sequenced

Table 1. Batches of plates processed

Batches	Number of plates	Plates
1	2	POR-024, POR-025
2	2	POR-026, POR-027
3	2	POR-028, POR-029
4	4	POR-030, POR-031, POR-032, POR-033
5	4	POR-034, POR-035, POR-036, POR-037
6	4	POR-038, POR-039, POR-040, POR-041
7	4	POR-042, POR-043, POR-044, POR-045
8	4	POR-046, POR-047, POR-048, POR-049
9	4	POR-050, POR-051, POR-052, POR-053
10	2	POR-054, POR-055
Total	32	

Appendix Table 2. *The sequence length of clones that provided hit against FLCIII DNA sequence and search output*

Query ID	Clone length (bps)	Subject ID	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	e-value	Bit score
POR_C027_H17	217	FLCIII	97.56	41	1	0	7	47	683	723	6.00E-17	73.1
POR_C030_N13	400	FLCIII	100	30	0	0	1	30	692	721	1.00E-11	56.5
POR_C030_G02	444	FLCIII	100	30	0	0	1	30	692	721	1.00E-11	56.5
POR_C031_H03	363	FLCIII	100	32	0	0	1	32	692	723	8.00E-13	60.2
POR_C031_E12	414	FLCIII	100	32	0	0	6	37	692	723	9.00E-13	60.2
POR_C031_G14	373	FLCIII	100	30	0	0	6	35	692	721	1.00E-11	56.5
POR_C031_F16	506	FLCIII	97.3	37	1	0	1	37	687	723	2.00E-14	65.8
POR_C031_B12	392	FLCIII	100	31	0	0	6	36	692	722	3.00E-12	58.4
POR_C031_F11	417	FLCIII	96.88	32	1	0	1	32	687	718	1.00E-11	56.5
POR_C031_B14	347	FLCIII	96.97	33	1	0	1	33	691	723	3.00E-12	58.4
POR_C032_F11	502	FLCIII	100	34	0	0	1	34	687	720	8.00E-14	63.9
POR_C033_K14	336	FLCIII	100	32	0	0	1	32	687	718	7.00E-13	60.2
POR_C034_F11	382	FLCIII	100	29	0	0	1	29	692	720	4.00E-11	54.7
POR_C034_N14	374	FLCIII	100	37	0	0	1	37	687	723	1.00E-15	69.4
POR_C034_C06	319	FLCIII	100	32	0	0	1	32	692	723	7.00E-13	60.2
POR_C034_L12	329	FLCIII	97.3	37	1	0	1	37	687	723	2.00E-14	65.8
POR_C039_D08	333	FLCIII	100	32	0	0	6	37	692	723	7.00E-13	60.2
POR_C040_F07	332	FLCIII	100	32	0	0	1	32	692	723	7.00E-13	60.2
POR_C041_J05	471	FLCIII	100	32	0	0	1	32	692	723	1.00E-12	60.2
POR_C041_F07	327	FLCIII	100	29	0	0	6	34	692	720	3.00E-11	54.7
POR_C041_F18	268	FLCIII	100	28	0	0	1	28	692	719	1.00E-10	52.8
POR_C043_I03	321	FLCIII	100	32	0	0	1	32	692	723	7.00E-13	60.2
POR_C047_M03	542	FLCIII	100	29	0	0	1	29	692	720	5.00E-11	54.7
POR_C050_D06	287	FLCIII	100	32	0	0	1	32	692	723	6.00E-13	60.2

Appendix Table 3. *The sequence length of clones that provided hit against E. coli genome and search output*

Query ID	Clone length	Subject ID	% identity	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	e-value	Bit score
POR_C037_L12	249 bps	DH10B_WithDup_FinalEdit	84.68	222	32	2	28	247	3504344	3504123	6.00E-76	279