Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

# Sequence analysis and transcript length estimation of a normalized full-length porcine cDNA library

*Samuel Gebremedhn Etay*

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

390

Uppsala 2012

Master's Thesis, 30 HEC

Erasmus Mundus Programme
– European Master in Animal
Breeding and Genetics

SLU

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

# Sequence analysis and transcript length estimation of a normalized full-length porcine cDNA library

*Samuel Gebremedhn Etay*

**Supervisors:**
Richard Crooijmans, WU, The Netherlands
Göran Andersson, SLU, Department of Animal Breeding and Genetics
**Examiner:**
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 - EMABG
**Level:** Advanced, A2E

**Place of publication:** Uppsala
**Year of publication:** 2012
**Name of series:** Examensarbete / Swedish University of Agricultural Sciences,
Department of Animal Breeding and Genetics, 390
**On-line publication:** http://epsilon.slu.se

**Key words:** Pig (porcine), cDNA library, Direct sequencing

EUROPEAN MASTER IN ANIMAL BREEDING and GENETICS (EMABG)

# Sequence Analysis and Transcript Length Estimation of a Normalized Full-Length Porcine cDNA Library

Samuel Gebremedhn Etay

Registration Number-840712-231-060

MAJOR MSC THESIS ANIMAL BREEDING AND GENETICS (ABG-80436)

JUNE 2012

Animal Breeding and Genomics Centre (ABGC)

<u>SUPERVISORS</u>

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

# Sequence Analysis and Transcript Length Estimation of a Normalized Full-Length Porcine cDNA Library

**Samuel Gebremedhn Etay**

A Thesis submitted in partial fulfillment of the requirement for the degree

Of

## MASTER OF SCIENCE

In

**ANIMAL BREEDING and GENETICS**

| | |
|---|---|
| **Dr. Richard Crooijmans (WUR)**<br>**Supervisor** | **Prof. Dr. Martien Groenen (WUR)**<br>**Examiner** |
| **Prof. Dr. Göran Andersson (SLU)**<br>**Supervisor** | **Prof. Dr. Erling Strandberg (SLU)**<br>**Examiner** |

**Wageningen University**
**Wageningen, the Netherlands**

**JUNE 2012**

## TABLE OF CONTENT

## ACKNOWLEDGEMENT

I would like to thank many people and organizations who have helped me during my study and thesis work. I would like to forward my deepest gratitude to the European Master in Animal Breeding and Genetics (EMABG) consortium committee for creating the platform for young students who would like to build their career in the fields of Animal Breeding and Genomics. EMABG was indeed an eye-opening program. The program wouldn't have been successful without the financial grants from the European Union and it is highly appreciated.

My deepest gratitude goes to my thesis advisor Dr. Richard Crooijmans. Your support was immense and you‘ve coached me in a way to be better and qualified in the future. I also would like to thank Professor Martien Groenen for helping in the data analysis and for his suggestions. My second advisor; Professor Dr. Göran Andersson of the Swedish university of Agricultural Sciences is highly acknowledged for his valuable comments during the thesis work and being inspiring during my first year stay in Uppsala, Sweden.

Yogesh Paudel; you made the sequence data editing using LINUX faster than it would have taken me in windows. Ole Madsen; thank you very much for your help in the blast analysis. Bert Dibbits and Kimberley Laport your help in the laboratory work is highly appreciated.

Dieuwertje Lont and Britt De Klerk thank you very much for every help all the way from the introduction period to the final days of the program. Birgitta Malmfors and Helena Eken thank you very much for you care and help during the one year stay in Uppsala, Sweden.

I am highly indebted to my father Gebremedhn Etay and my mother Mebrat Atsbeha. Mebratey; this is what you always say to me "እዞም ደቀይ ኣብ ህይወትኩም እቱ ዘምልጥ ነገር ቀዲምኩም ሓዙ እሞ......" እቱ ዘምልጥ ንክሕዝ ይፍትን ኣለኹ:: Uncle Fikremariam Atsbeha, My brother and Sisters (Hailsh, Yonas, Zewdu, Seni, Yorda, Nigisti and Solome,) I always love you. Hewan Gebremariam and Tewodros Gebremariam; thank you for your encouragement. Long live to all of you.

Shikor ! You have been my strength all the time. Your love keeps me warm and delighted. Now is the time to start the long journey together. Du wisst, ich liebe dich sehr !

Friends in Wageningen University (Medhin Teklay, Mahlet, Desalegn, Daniel Kiros, Abraha Asmelash, Yihdego Tesfay and others) and in SLU ( Yonatan, kibebew, Alemayehu and others) thank you very much. Good friends make life viable.

Above all, I praise my Almighty God who gave me the courage and strength in my life.

## SUMMARY

*The Pig is one of the most important sources of animal protein and an essential animal model for human biomedical research due to its anatomical, physiological, biochemical and metabolic similarities with human being. Profound understanding of the pig genome helps to understand its biology in depth. To help this, availability of the pig genome sequence in the public databases facilitates research of wider spectrum. Generating Expressed sequence tags (EST) by partial sequencing normalized full-length cDNA libraries can improve the discovery of genes expressed in different tissues. cDNA libraries are also vital resources to elucidate alternative splicing pattern of genes and infer splice variants. The objective of this study was to pick and sequence the 5'end of 7,680 individual clones, identify clone sequences blasted against identical genes to determine the transcript size by Agarose-gel analysis and examining the presence of splice variants. Total RNA was isolated from eleven different tissues of an adult pregnant pig (113 days of gestation) and a normalized full-length cDNA library was constructed by a commercial company. Individual cDNA clones were cultured in 384-well plates and Sanger sequenced on an ABI3730 DNA analyser. The obtained sequence results were merged with results of two previous studies of the same cDNA library which resulted in 26,880 processed clones. In total 19,470 edited sequences were used for further analysis. Sequence similarity was searched using Basic Local Alignment Searching Tool (BLAST) against pig cDNA, human cDNA and mouse cDNA databases. A total of 12,461 sequences provided hit against the pig cDNA database and 8,300 sequences provided hit against the human cDNA database. 5,268 sequences provided hit against the mouse cDNA database. Moreover, the first 100 base pairs of the sequences were blasted against the newly released pig genome (build 10.2). In total of 12,222 sequences provided hit against the pig genome. Significant amount of clones provided database specific hits in either the Human or Mouse cDNA database which are important to retrieve homologous pig genes which are not mapped in the pig genome. A total of 7,074 non-redundant transcripts and 6,877 non-redundant genes were retrieved. PCR and Gel-electrophoresis protocols were optimized to determine insert the size of transcripts. Transcript length of 108 clones blasted against 10 different genes was determined and variation in size was obtained. The complete sequence of the 13 clones blasted to the same gene; the Swine Leucocyte Antigen (SLA-3) gene was obtained by sequencing with internal primers and variation in length and Exon-intron organization was confirmed. A comparison between the clone sequences and mRNA sequence stored in gene bank revealed higher degree of similarity and tissue specific transcripts were found. Large-scale screening and sequencing of cDNA libraries constructed from various tissues and development stages using the direct sequencing procedure can help to understand the pig genome. The procedure is also important to identify tissue specific splice variants.*

# 1. INTRODUCTION

The pig (*Sus scrofa*) is one of the most important sources of animal protein that provides about 36% of the global meat supply (FAO, 2009). Moreover, the pig is an essential animal model for human domestication (Archibald *et al*., 2010), evolutionary studies, biomedical and organ transplantation research (Fang *et al*., 2005). Its anatomical, physiological, biochemical and metabolic similarities with human makes the pig a useful target for human disease related research such as obesity and cardiovascular disease (Zhang *et al*., 2007). There are significant similarities in size and complexity of genomes of human and pig. Furthermore, comparative mapping revealed that there is similar genomic organization in the genome of pig and human than either of them compared to the mouse genome (Frönicke *et al*., 1996).

Detailed description and profound understanding of the pig genome brings significant benefits to both the economy and health sectors (Zhang *et al*., 2007). The pig genome has been sequenced using a combined approach of hierarchical shotgun and whole genome shotgun sequencing techniques by the swine genome sequencing consortium (Archibald *et al*., 2010). The latest version of the pig genome (Build 10.2) including the gene annotation is in the process of completion and being released to public databases. The availability of the pig genome sequence has an important role in widening our understanding of the pig biology in greater depth. Interpretation of a genome sequence in eukaryotes is enormously complicated and difficult. The coding regions of a genome; the exons are intermingled by non-coding intronic regions of the gene (Kawai *et al*., 2001). Furthermore, several gene products can be obtained from a single gene through alternative splicing (Zhang *et al*., 2007). This makes prediction of the real number of protein coding genes in an organism ambiguous. This leads to a conclusion that genomic sequences cannot decipher precisely the actual picture of the transcriptome in particular and the proteome in general (Kawai *et al*., 2001). However, scrutinizing the transcribed region of a gene in depth can provide more information. Therefore, sequencing the messenger RNA (mRNA) after converting into its complementary DNA (cDNA) is of paramount importance (Kato *et al*., 2005). The structure of a complete nuclear mRNA contains two distinct structural features; the Cap-structure at the 5'-end and the poly-A tail at the 3'-end of the mRNA. These structural features of mRNA play important roles in transport, intron splicing, stability and prevent mRNA degradation by endo-nuclease enzymes (Sachs, 2000). The region between the cap-structure and the poly-A tail is the coding region of matured mRNA. Kato *et al*., (2005) mentioned that sequencing the entire mRNA of a gene and mapping its sequence on the genome helps to easily identify structure of the coding and non-coding regions and other functional domains of a gene. The nature, distribution and expression of mRNA in eukaryotic cells vary spatiotemporally making full-length cDNA library construction challenging and complicated (Carninci, 2000). The mRNA constitutes only 1-5 % of the total RNA mass and it can further be classified according to the level of expression and abundance as super abundant, intermediate and rare (Shcheglov *et al*., 2007). Even though, there is a variation among cells derived from different tissues (Carninci, 2007), in a cellular transcriptome there can be 5-10 types of superabundant, 500-2000 types of intermediate and above 10,000 types of rare mRNA representing 20-30%, 30-50% and 30-40% of the total cellular mRNA mass respectively (Alberts *et al*., 1994). The presence of superabundant cellular mRNA can cause uneven distribution of expressed genes during cDNA library construction (Carninci, 2007). Eventually, this can cause difficulty in rare gene discovery and over representation of the highly expressed genes like housekeeping genes (Carninci, 2000). Despite the challenges, generating a normalized full-length cDNA library can improve the representativeness and discovery of genes (Natarajan *et al*., 2010, Nguyen *et al*., 2010). Therefore, normalizing the cDNA library can enrich the library and improve the sequencing efficiency by decreasing the abundance of superabundant genes in the library. This will help in the discovery of new genes and transcripts (Carninci, 2007, Natarajan *et al*., 2010).

cDNA libraries can be constructed using a PCR and solid matrix-based approaches (Shcheglov *et al*., 2007). Nevertheless, such cDNA libraries are characterized as non-normalized and of higher sequence redundancy. Carninci (2000) developed a cDNA library normalizing procedures based on physical separation of fractions. The normalization procedure includes biotinylation of the first cDNA strand and removal of the biotinylated RNA-DNA duplex using streptavidin magnetic beads. According to the length and entirety there are two types of cDNA libraries; the conventional and full-length cDNA libraries. The conventional cDNA libraries are truncated and contain about 20-30% of the full-length cDNA. The full-length cDNA libraries represent more than 90% of the full-length cDNA. The full-length enriched cDNA libraries have advantage over the conventional ones in providing a complete amino acid sequence of a particular protein and functional screening of genes (Shcheglov *et al*., 2007). There have been efforts to generate a full-length cDNA library sequences over the past decades. Nonetheless, the information obtained from the sequences was limited. The limitations arose due to technical complications; reduced efficiency of reverse transcriptase to produce full-length first strand cDNA and lack of efficient techniques to select full-length cDNA (Carninci *et al*., 1996). A cap-tapper method developed by Carninci *et al*. (1996) was reported to be an effective technique to construct a full-length and high-content cDNA library. Functional analysis of expressed genes in different tissues, cells and developmental stages can be effectively studied by analysing the expressed sequence tag (EST)

(Adams *et al*., 1992, Maeda *et al*., 2006). Partial sequencing of clones of cDNA libraries constructed from specific tissues can help to discover new and tissue specific genes (Yao *et al.*, 2002). Accumulating larger amount of expressed sequence tags on public repositories have been providing information on the transcriptome and enhancing the on-going mammalian genome sequencing projects (Al-Swailem *et al.*, 2010). In addition to the functional analysis of genes, cDNA libraries are vital resources to elucidate alternative splicing pattern of genes and infer novel splice variants.

Alternative splicing is a cellular activity in eukaryotes through which several gene products are obtained from a single gene (Zhang *et al*., 2007) which subsequently increases transcriptome and proteome complexity (Bonizzoni *et al.*, 2006). The alternatively splicing nature of immature-mRNA provides multiple transcripts of a single gene. This is the reason for the outnumbering transcripts of vertebrates to their protein coding genes. More than 70% of human genes undergo alternative splicing (Johnson *et al*., 2003). According to the splicing pattern of the coding regions of a gene, there are four different types of alternative splicing namely; exon skipping, alternative 5' splice site selection(5' SS), alternative 3' splice site selection (3' SS) and intron retention (Kim et al., 2008). The Exon skipping pattern of splicing can remove an entire exon from the transcript with its flanking introns and it is the most predominant type of splicing pattern in both vertebrate and invertebrates (Zhang *et al*., 2007). Alternative splicing of mRNA can be detected using both computational algorithms (Bonizzoni *et al.*, 2006) and experimental analysis of mRNA of either single or pooled tissue samples. However, computational approaches of splice variant detection are of lower specificity (Leparc and Mitra, 2007) and experimental analysis of mRNA can be an important resources to validate results of the computational approach of splice variant detection.

Over the past decades there have been remarkable efforts to generate EST by sequencing cDNA libraries derived from different porcine tissues. Smith *et al.* (2001) generated and sequenced 1,132 clones from a porcine early embryonic cDNA library. Fahrenkrug *et al*. (2002) also constructed two normalized porcine cDNA libraries derived from porcine embryonic and reproductive tissues and sequenced the 5'-end of 66,245 clones. Wang *et al.* (2006) constructed 131 randomly isolated clones from a longissimus Dorsi muscle tissue to study expression pattern of genes of skeletal muscle tissues of Chinese native pig. Lee *et al*. (2009) constructed five normalized and non-normalized cDNA libraries and generated a total of 71,000 high-quality ESTs. The cDNA library was constructed from porcine tissues related to energy metabolism; abdominal fat, induced fat cells, loin muscle, liver and pituitary gland.55,658 of the sequenced EST were stored in the database of expressed sequence tag (dbEST). Similarly, Kim *et al*. (2006) constructed a full-length enriched cDNA library from porcine back fat tissue to comprehend the expression of genes in backfat tissues. 16,110 sequences were deposited in the database of expressed sequence tags. Tan *et al.* (2006) also constructed and characterized a cDNA library from liver tissues of a Chinese Mini-pig inbred line. The cDNA library was constructed to investigate the genetic background of protein incompatibility after liver xenotransplantation and recommend an alternative strategy for further gene manipulation. Likewise, Yao *et al.* (2002) have constructed a normalized cDNA library from porcine skeletal muscles tissue to identify changes in the regulation of genes responsible for muscle growth. A total of 782 expressed sequence tags were generated. Chen *et al.* (2006) constructed a cDNA library of porcine adipose tissue to understand the functional expression of genes abundantly expressed in adipose tissue and sequenced 2880 individual clones.

Currently several research groups are highly involved in generating huge amount EST of cDNA libraries constructed from tissues derived from either multiple or specific organs and developmental stages of pigs. Useful information related to different pig metabolisms, and expression of specific genes affecting traits of economic importance is being generated. Nevertheless, in comparison to the human and mouse the profile of porcine EST generated from cDNA libraries is far from complete. Currently there are 1,669,337 pig ESTs stored in the database of expressed sequence tag (dbEST) (May 1, 2012). While for human and mouse there are 8,315,296 and 4,853,570 ESTs respectively (http://www.ncbi.nlm.nih.gov/dbEST). This indicates that the porcine transcriptome profile is not as well studied as the human and mouse ones.

The Animal Breeding and Genomics Centre (ABGC) of Wageningen University is one of the leading groups in the international swine genome consortium. It is leading the swine *Hapmap* consortium where over 4,000 pig samples from almost all breeds have been collected (either commercial, traditional as well as wild). Currently the ABGC is sequencing and characterizing a normalized full-length cDNA library constructed from pooled tissues samples of multiple organs of a clone of the pig from which the pig genome is derived.

Therefore, the aim of this project is to sequence and analyse cDNA clones from a normalized full-length cDNA library and blast all sequences generated by the project to the latest pig genome (build 10.2) to retrieve transcript ID, gene names and gene description. Furthermore, the project aims to identify clones sequences blasted to identical genes and examine the variation transcripts of these individual clones.

## 2. OBJECTIVES

a. To sequence the 5’-end of 7,680 individual clones in twenty 384 well-plates.
b. Combine all sequences (including the ones from previous experiments) and blast against the newly released pig genome (Build 10.2) and pig cDNA databases in order to retrieve transcript ID, gene names and gene descriptions.
c. To identify clones blasted to the same gene and elucidate presence of splice variants.

## 3. MATERIALS AND METHODS

### 3.1 RNA Extraction and cDNA Library Construction

Total RNA was isolated from individual 0.1 g of eleven different pig (Duroc Pig T.j. Tabasco; a clone of the pig from which the pig genome is derived) tissues (Kidney, Liver, lymph node, cerebellum, placenta, colon, hypothalamus, brain frontal lobe, spleen, small intestine and lung) of an adult pregnant pig (113 days of gestation) using Trizol extraction according to the manufacturer’s instruction (Appendix 1.1). Samples were loaded into RNeasy column and subjected for purification. Finally, 5-10 $\mu$g of the extracted total RNA of individual tissue samples were pooled together. The pooled total RNA samples were sent to a commercial company (K.K DNAFORM, JAPAN) for a normalized full-length cDNA library construction.

The normalized full-length cDNA library was constructed by a commercial company according the cap-tapper method (Carninci *et al.*, 2001). The full-length double stranded cDNA was digested, purified and cloned in **λ**-FLC III cloning vector (Detail information on cloning vector see Appendix 3**).** Finally, the plasmid DNA was transformed into an E.*coli* bacteria strain *DH10B*.

As part of the quality control procedures the insert size of 48 randomly isolated clones was determined. The average insert size was estimated to be about 2 Kb *(*Appendix 5. Table s1*)*. Sequences were also checked for redundancy, contamination, success rate and full-lengthiness. Finally, plasmid and phage stock were delivered and stored at -80 $^{o}$C.

### 3.2 Culturing and sequencing of Clones

Plasmid stock (1.75 *µl*) was mixed with 200 *µl* of LB media supplemented with Ampicillin (0.1 *µg/ml*). From the mix 100 *µl* was plated on petri dishes with 1% agar in LB medium which was supplemented with Ampicillin (0.1 *µg/ml*). Finally, Petri dishes were incubated upside down overnight (18 hours) at 37 $^{o}$C. (Details see: Appendix 1.2-1.5)

### 3.2.1 Master and Replica Plates Preparation

For the sequencing procedures one 384-well master plate (POR_A) and two 384-well replica plates (POR_B and POR_C) were prepared. Plates POR_A and POR_B were supplemented with LB medium, Ampicillin and 10X freezing medium (see: Appendix 1.6). The freezing media (Appendix 1.4) was added to plates because both plates (POR_A and POR_B) had to be stored in -80 $^{o}$C and serve as back-up plates. Plate POR_C was supplemented with only LB media and Ampicillin. (Freezing media was not added to POR_C in order to prevent negative effect of salt residues while sequencing). Individual bacterial colonies were picked from the agar petri dishes and transferred into 384 master plates (POR_A). Plates were incubated overnight (18 hours) at 37 $^{o}$C. The following day, replication of plates containing bacterial colonies was done using the 384-Pin replicator (Appendix 1.7). Similarly both replica plates (POR_B and POR_C) were incubated overnight at 37 $^{o}$C. Plates POR_A and POR_B were stored in -80 $^{o}$C to serve as back-up plates. Whereas, plates POR_C were used to prepare cell lysate for further sequencing procedures.

### 3.2.2 Cell lysate, Direct Sequencing Reaction and DNA Precipitation

Cell lysate was prepared by spinning down bacterial colonies at 2000 rcf for 20 minutes followed by series of washing the media using MQ water. Bacterial pellets were re-suspended in 25 $\mu l$ of MQ water. Plates were loaded into a thermo-cycler for cell denaturation at 95 $^{o}$c for 5 minutes. The cell lysate were stored directly on ice and centrifuged at 1000 rcf for 10 minutes to separate the cell debris from the plasmids. (Details *see*: Appendix 1.8). 2 $\mu l$ of the cell lysate was used for direct sequencing by mixing it with 3 $\mu l$ sequencing master mix (BigDye 3.1, 5X sequencing buffer, MQ water, Universal T3 forward Primer). Cycle sequencing was performed using BD50x50 program in PCR thermo-cycler (Appendix 1.9). DNA samples were precipitated with 0.5 $\mu l$ sodium acetate (NaAc-EDTA), 17 $\mu l$ of 100% ethanol. DNA samples were then dissolved in 10 $\mu l$ of formamide (Appendix 1.10). Finally, samples were transferred into a barcoded 384 well-plates and loaded into ABI 3730 DNA analyser (Applied Biosystems®). A summary of the whole sequencing procedure is represented schematically in figure 1.

## 3.3 Sequence Analysis

Raw sequences reads were retrieved from the ABI 3730 DNA analyser and converted to FASTA format (Appendix 2.1) using ABI-2-FASTA converter of the DNA baser sequence assembler package (http://www.dnabaser.com/download/Abi-to-Fasta-converter.html). MULTI-FASTA files of each 384 well-plates were generated using the Multi-Fasta builder of the DNA baser sequence assembler package (Appendix 2.2).

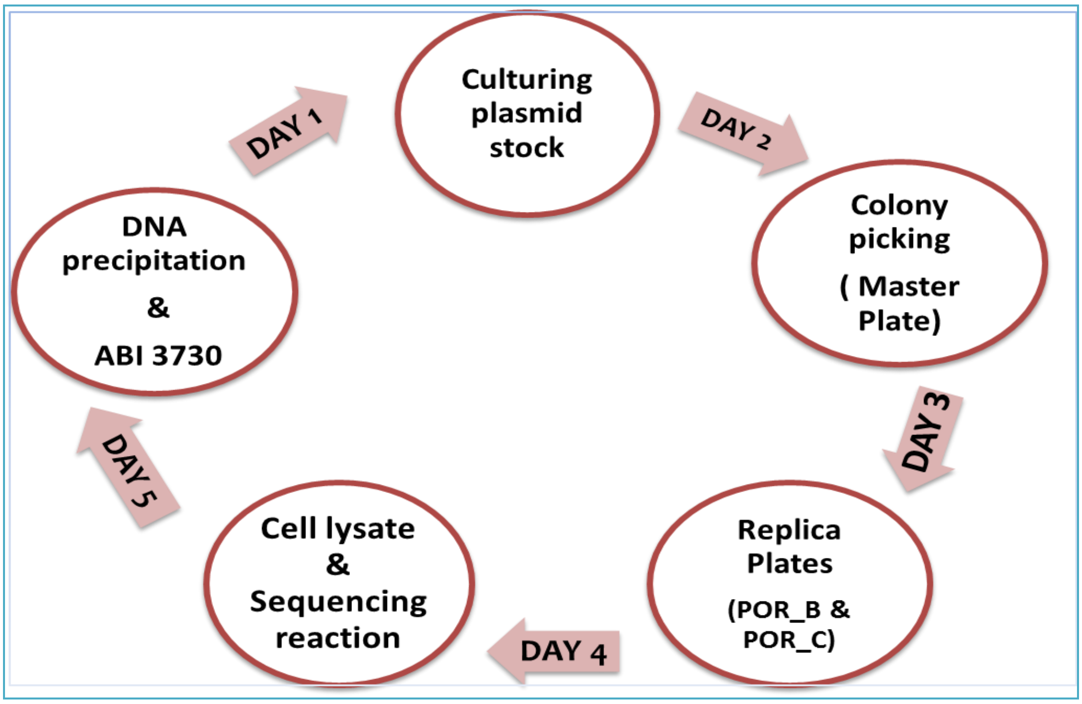

Culturing plasmid stock
DAY 2
Colony picking ( Master Plate)
DAY 3
Replica Plates (POR_B & POR_C)
DAY 4
Cell lysate & Sequencing reaction
DAY 5
DNA precipitation & ABI 3730
DAY 1

*Figure 1: Graphic representation of direct sequencing procedures*

Sequence lengths of 50 base pairs and above were considered as good reads for further analysis. Sequencing efficiency percentage, total number of sequences in base pairs, maximum sequence length, minimum sequence length and average sequence length of each 384 well-plate were calculated. All edited sequence files from this experiment were merged with another two sequence files from previous experiments of the same project for further analysis.

### 3.3.1 Sequence Quality Control and Open Reading Frame Detection

Presence of vector sequence and an open reading frame was searched for 10 randomly selected clone sequences of each plate sequenced. This was done to confirm the full-lengthiness of the cDNA sequences for blast search in the presence of the transcription start codon (ATG) and the 5' untranslated region (UTR). Moreover, the presence of open reading frame (ORF) signifies the presence of a gene and was searched using the ORF finder software hosted on the National Centre for Biotechnology information (NCBI) using the default settings.

### 3.3.2 Basic Local Alignment Search Tool Analysis

The pig genome (Build 10.2), Pig cDNA, Human cDNA, Mouse cDNA ,λFLC-III DNA and *E.Coli* genome databases were downloaded from Ensembl (www.ensembl.org/index.html) (Appendix 2.3) and converted into blastable database using specific command lines in MS DOS (Appendix 2.4). The Blast-2.2.26+ application for windows was also downloaded (Appendix 2.5) from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). Sequence similarity search was performed using the Basic Local Alignment Search Tool (BLAST) resource of nucleotide *Blastn* (Appendix 2.6). In order to reduce the number of redundant hits, only the first 100 bps of the clone sequences were used to blast against the pig genome (Build 10.2). However, the entire clone sequences were used to blast against the pig cDNA, Human cDNA, and Mouse cDNA databases. Sequence similarity with an e-value of less than $1E^{-10}$ was considered as a significant hit. The blast output was screened for redundancy and for clones blasted against identical genes.

## 3.4 PCR amplification and Gel-Electrophoresis Analysis

All sequences blasted against the pig cDNA database and blast output was merged with a separate gene annotation file that was used by Ensembl for gene annotation but not yet released when performing the analysis. Clones with hits to identical genes were selected. We randomly selected 108 clones blasted to 10 different genes (Appendix 5. Table 5s). The aim of the PCR reaction was to amplify the whole cDNA insert and verify presence of size variation among transcripts. Furthermore, sequence their 3'-end to infer presence of splice variants of the genes. All clones of one gene were selected for the whole transcript sequencing using internal primers.

### 3.4.1 Optimizing PCR protocol

PCR reaction was performed to amplify the cDNA insert of selected clones using both universal forward and universal reverse primers; T3 and T7 respectively. Optimized PCR protocol and PCR conditions were developed (Table 1).

Table 1: DNA template dilution rates and annealing temperatures for trial PCR protocol

| Sr.No | DNA template ( Cell lysate) Dilution rates | Annealing temperature |
|---|---|---|
| 1 | ± 1:7 | 50 and 55 |
| 2 | ± 1:12 | |
| 3 | Stock ( Undiluted cell-lysate) | |

### 3.4.2 Agarose Gel Electrophoresis

An Agarose gel of different percentages (Table 2) were prepared and dissolved in 300 ml of 0.5% TBE buffer. Etidium Bromide (EtBr) was added at a rate of 5 *µl*/100 *ml* of TBE buffer. (Detail see: Appendix 1.11) Per gel, 2 *µl* of PCR product of the three types of DNA dilution rates and two different annealing temperatures were mixed with 2 *µl* of loading dye and 6 *µl* of MQ water and loaded on the Agarose gel. Size standards of 100 bp and 500 bp were used to estimate PCR amplicon size.

Table 2: Agarose percentage and Gel electrophoresis running time

| Sr.No | Agarose percentage | Gel running time in minutes |
|---|---|---|
| 1 | 1.5 | 60 and 180 |
| 2 | 1.25 | |
| 3 | 1.00 | |

### 3.4.3 Purification of PCR products and Transcripts Length Estimation

The PCR products of the selected clones were purified using the Millipore PCR clean-up vacuum system to remove primers and dNTPs The quality and quantity of the purified PCR products were assessed on 1.5% Agarose gel by using the precision marker for the presence of single band without primer dimer(Appendix 1.12).

The purified PCR product was sequenced by adding BigDye 3.1, 5X dilution buffer, MQ water and universal T7 reverse primer on Biometra 384 well-plate thermo cycler with an annealing temperature of 55 °C (Appendix 1.13). The sequencing products were precipitated using (NaAc-EDTA) and Ethanol (Appendix 1.14). Samples were dissolved in formamide and transferred into barcoded 96 well-plate and analysed on the ABI3730 DNA analyser.

The poly-A tail at the 3’-end of the cDNA inserts made the sequencing procedure using the universal T7 reverse primer difficult and resulted bad quality sequences. To solve the problem further sequencing of the purified PCR product from the 5’-end by designing internal primers was considered as a solution. Sequence result from purified PCR products are longer than sequence product from direct sequencing. Additional sequencing on selected 13 clones blasted to one gene was performed by making internal primers in the program Primer3 (Appendix 2.7) (Appendix 5 table S4). Purified PCR amplicon were sequenced until the complete cDNA was sequenced. Sequences were aligned to the reference pig genome on UCSC genome browser to detect differences among transcripts.

# 4. RESULTS

## 4.1 Direct Sequencing of Clones of a Full-Length cDNA Library

In total 6,912 individual clones were picked and sequenced in eighteen 384 well-plates. The overall sequencing efficiency of the experiment was 79.4% ranging from 33.6 to 93.2%. The lowest sequencing efficiency was attained in plates POR_C062 and POR_C063 with overall efficiency of 33.6% and 49.5% respectively due to a failure in the sequencer. Therefore, only 5,481 clone sequences were considered as good sequence reads and used for further analysis. Sequences had an average length of 505 base pairs (bp) ranging from 50 bp to 857 bp (Table 3). On the contrary, seven plates (POR-C056, POR-C060, POR-C061, POR-C067, POR-C071, POR-C072 and POR-C073) had an efficiency of more than 90%.

Table 3: Total sequenced cDNA clones

| | | Sequences Statistics | | | | | |
|---|---|---|---|---|---|---|---|
| Sr.No | Plates Number | Total number of good sequences | Length of sequences (bp) | Min (bp) | Max (bp) | Average (bp) | Efficiency* (%) |
| 1 | POR-C056 | 358 | 210937 | 61 | 745 | 589 | 93.229 |
| 2 | POR-C057 | 339 | 184994 | 87 | 716 | 545 | 88.281 |
| 3 | POR-C058 | 299 | 161161 | 66 | 857 | 539 | 77.865 |
| 4 | POR-C059 | 290 | 150538 | 60 | 704 | 519 | 75.521 |
| 5 | POR-C060 | 347 | 195445 | 72 | 728 | 563 | 90.365 |
| 6 | POR-C061 | 357 | 189291 | 65 | 730 | 530 | 92.969 |
| 7 | POR-C062 | 129 | 27520 | 55 | 545 | 213 | 33.594 |
| 8 | POR-C063 | 190 | 45884 | 51 | 526 | 242 | 49.479 |
| 9 | POR-C064 | 254 | 95886 | 69 | 591 | 378 | 66.146 |
| 10 | POR-C065 | 226 | 78889 | 51 | 619 | 349 | 58.854 |
| 11 | POR-C066 | 318 | 139490 | 50 | 697 | 439 | 82.813 |
| 12 | POR-C067 | 352 | 171497 | 52 | 670 | 487 | 91.667 |
| 13 | POR-C068 | 292 | 94419 | 57 | 751 | 323 | 76.042 |
| 14 | POR-C069 | 332 | 186462 | 60 | 766 | 562 | 86.458 |
| 15 | POR-C070 | 339 | 196138 | 61 | 754 | 579 | 88.281 |
| 16 | POR-C071 | 353 | 204524 | 57 | 767 | 579 | 91.927 |
| 17 | POR-C072 | 358 | 221794 | 50 | 755 | 620 | 93.229 |
| 18 | POR-C073 | 348 | 213087 | 118 | 738 | 613 | 90.625 |
| **Overall** | | **5481** | **2,767,956** | **50** | **857** | **505** | **79.4** |

**Sequencing Efficiency of each plate is equals to the number of total good reads divided by 384 multiplied by 100*

The Sequences obtained in this project were combined with sequences obtained in two earlier performed experiments which generated 13,989 useful sequence reads from 19,968 processed clones from fifty two 384 well-plates. The average overall sequencing efficiency of the previous experiments was 71.21% and 69.34% respectively with an average sequence length of 365 and 314 bp respectively (Table 4). (Details see Appendix 5 Table s2 and Table s3).

## 4.2 Sequence Similarity Search against Reference Databases

Sequences from the three separate experiments were edited and merged together to create one file that included 19,470 individual cDNA clone sequences from seventy 384-well-plates (Table 4).

Table 4: Summery of the total number of sequences generated from the three experiments

| Experiment | Total No of plates | Number of clones Processed | Maximum length (bp) | Minimum length (bp) | Average length (bp) | Average efficiency (%) | Number of good sequences |
|---|---|---|---|---|---|---|---|
| **1** | 20 | 7,680 | 750 | 51 | 365 | 71.21 | 5,469 |
| **2** | 32 | 12,288 | 811 | 51 | 314 | 69.34 | 8,523 |
| **3** | 18 | 6,912 | 857 | 50 | 505 | 79.40 | 5,481 |
| **Total** | **70** | **26,880** | **857** | **50** | **394** | **72.46** | **19,470** |

Prior to the search for sequence similarity using the Blast analysis, the latest version of blast+ application (blast-2.2.26+) and fasta sequence files were downloaded. Blastable databases were created from the fasta sequences. Sequences were blasted against the pig genome (build 10.2), Pig cDNA, Human cDNA, Mouse cDNA, λFLC-III vector sequences and *E.Coli* genome. 80% of the sequences (15,388) of all useful sequence reads displayed hit in either of the pig genome, pig cDNA, Human cDNA and mouse cDNA databases ( Examples Table 6-10).

The blastn analysis output revealed that a total of 12,222 hits were obtained by blasting the first 100 bp of the sequences against the pig genome database. The reason for blasting only the first 100 base pairs of the sequences against the pig genome was to avoid redundant hits. The blastable part of the sequence mainly contains only the first exon of the cDNA sequences. In total 12,461 sequences gave a hit against the pig cDNA database (Table 5).

In addition to the blasting against the pig databases, blasting sequences against databases of species which are well analysed and evolutionarily related to pig (human and mouse) databases can help to identify homologous pig genes which are not mapped in the pig genome. As a result, 8,300 sequences showed hit against the human cDNA and 5,268 clone sequences showed hit against the mouse cDNA databases (Table 5).

Table 5: Summery of hits generated by blasting 19,470 cDNA sequences against different databases

| Sr. No | Database | Total Number of hits |
|---|---|---|
| 1 | Pig Genome (Build 10.2)$^{A}$ | 12,222 |
| 2 | Pig cDNA (Build 10.2) | 12,461 |
| 3 | Human cDNA | 8,300 |
| 4 | Mouse cDNA | 5,268 |
| 5 | *E.Coli* genome | 20 |

$^{A}$*: Only the first 100 bps of the sequences were blasted against the pig genome database*

Blasting against the *E.Coli* genome was also performed to check the inclusion of bacterial genome during the sequencing process. Only 20 clones displayed hits against the *E.coli* genome. All hits were thoroughly inspected if they can be found in the blast output of either of the pig databases. It turned out only 5 hits are obtained from the *E.coli* genome which represents only 0.04 per cent of the overall good sequence reads used for blastn analysis. The sequence reads were also blasted against the λFLC-III vector DNA sequence. There were 2,325 clones which gave hits against the vector sequence. It is quite normal to get significant amount of hit as the cDNA insert was cloned into the λFLC-III vector. Moreover, the vector sequences were not trimmed from the cDNA sequences. The average alignment length of the clones which displayed hit against the λFLC-III DNA sequence is 36 bps (Figure 2). This implies that on average 36 bp of the vector sequences before the start site of the inserted sequences are sequenced as the forward primer anneals into the vector sequence.

Blasting the cDNA clone sequence to the pig genome and comparing the alignment of the cDNA clone exons with the predicted positions of exons can indicate the full-lengthiness of the clone sequences. The clone sequence POR_C070_P17 blasted against the pig reference genome contains 5 exons. The positions of all exons of the clone perfectly match with the predicted positions of exons (Figure 3). The first exons of the predicted gene is not fully coding due to 5'-untranslated region (UTR).

The presence of both the vector sequence before the start of the cDNA insert can be an indicator of the quality of the sequencing product. For instance, the clone sequences *POR_C070_P17* indicated in figure 2 contains a vector sequence before the start of the cDNA insert. The first 5-10 bps of the cDNA sequence is of lower in quality which is the start site of the cDNA insert. This is due to the annealing of the T3 universal forward primer to the vector sequence

Table 6: Example of clones hits against the Pig genome database

| Clone ID | Chr. Number | % Identity | Alignment length | Mismatches | Gap Opens | Query start | Query End | Subject Start | Subject End | E-Value | Bit score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POR_C070_P17 | 8 | 100 | 57 | 0 | 0 | 44 | 100 | 130451297 | 130451353 | 2.00E-21 | 106 |
| POR_C073_P14 | 10 | 100 | 67 | 0 | 0 | 34 | 100 | 54244382 | 54244448 | 5.00E-27 | 124 |

Table 7: Example of clones hits against the Pig cDNA database

| Clone ID | Subject ID | % Identity | Alignment length | Mismatch | Gap Opens | Query start | Query End | Subject Start | Subject End | e-value | Bit Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POR_C070_P17 | ENSSSCT00000010058 | 99.81 | 526 | 0 | 1 | 44 | 569 | 38 | 562 | 0 | 965 |
| POR_C073_P14 | ENSSSCT00000022795 | 99.24 | 131 | 1 | 0 | 517 | 647 | 1 | 131 | 1.00E-61 | 237 |

Table 8: Example of clones hits against the Human cDNA database

| Clone ID | Subject ID | per cent Identity | Alignment length | Mismatch | Gap Opens | Query start | Query End | Subject Start | Subject End | e-value | Bit Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POR_C070_P17 | ENST00000508511 | 93.83 | 405 | 22 | 3 | 166 | 569 | 16 | 418 | 3.00E-172 | 606 |
| POR_C073_P14 | ENST00000376139 | 96.74 | 614 | 20 | 0 | 34 | 647 | 19 | 632 | 0 | 1024 |

Table 9: Example of clone hits against the Mouse cDNA database

| Clone ID | Subject ID | per cent Identity | Alignment length | Mismatch | Gap Opens | Query start | Query End | Subject Start | Subject End | e-value | Bit Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POR_C070_P17 | ENSMUST00000005964 | 90.35 | 404 | 36 | 3 | 167 | 569 | 89 | 490 | 2.00E-148 | 527 |
| POR_C072_E11 | ENSMUST00000147559 | 83.49 | 212 | 24 | 9 | 152 | 359 | 24 | 228 | 4.00E-46 | 187 |

Table 10: Example of clone hits against the E.Coli genome database

| Clone ID | Subject ID | per cent Identity | Alignment length | Mismatch | 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 |
| POR_C057_D17 | DH10B_ withdup_FinalEdit | 98.63 | 219 | 0 | 2 | 161 | 377 | 1384971 | 1385188 | 2.00E-107 | 385 |

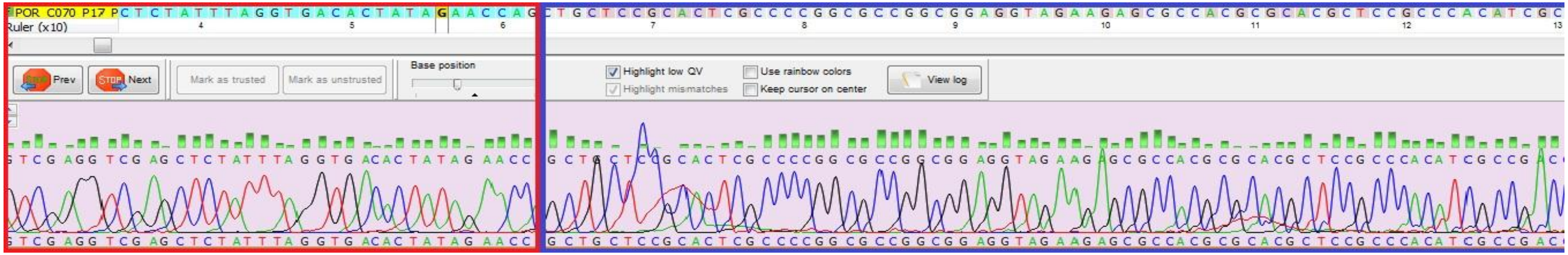


*Figure 2: A partial overview of a vector sequence and cDNA insert of the clone sequence POR_C070_P17. The vector sequence; highlighted in light blue is indicated in the red box and the cDNA insert sequence is indicated in dark blue box. The sequence quality of the first 10-15 base pairs of the cDNA insert clone is low. However, as we move further towards the 3'-end the quality gets better. The presence of the vector sequence before the 5'-end of the insert clone sequence signifies the full-lengthiness of the cDNA library.*

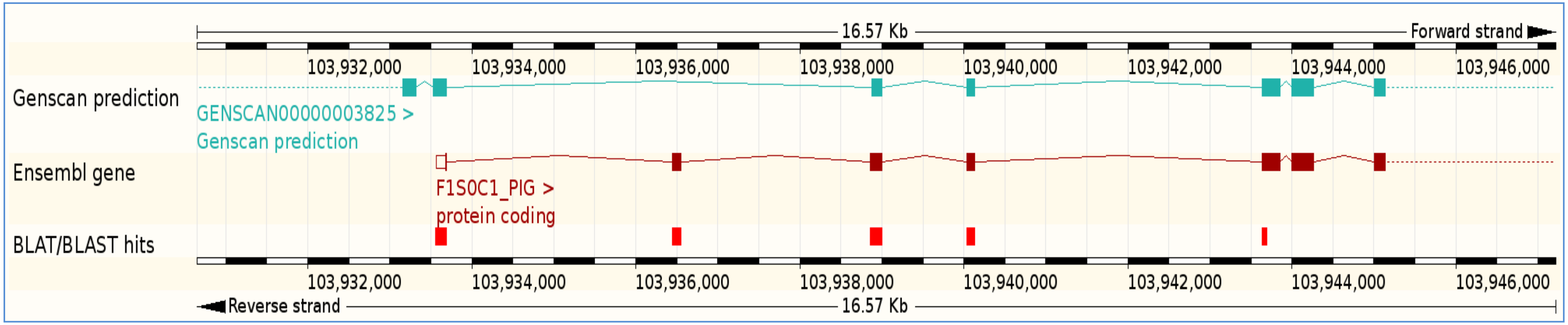
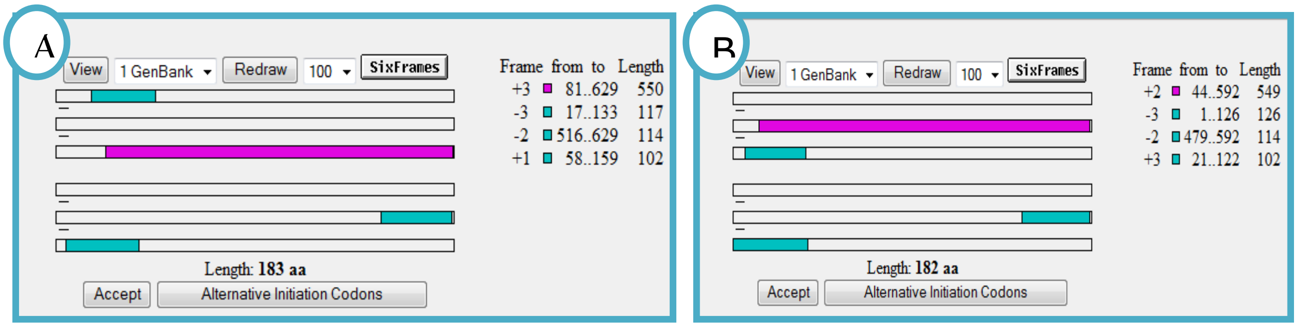


*Figure 3: BLAT/BLAST hit of the cDNA clone POR_C070_P17 against the pig genome using Ensembl genome browser. The clone sequence aligns with the gene F1SOC1_PIG . The Ensembl gene structure, gene scan prediction and the blast hit output of the cDNA clone are indicated. The cDNA Clone is a full-length as the first exon starts exactly at the 5'-end forward strand, which perfectly matches to the 5'-end forward strand of the Protein coding F1SOC1_PIG gene.*



| Frame | from to | Length |
|---|---|---|
| +3 | 81..629 | 550 |
| -3 | 17..133 | 117 |
| -2 | 516..629 | 114 |
| +1 | 58..159 | 102 |

Length: 183 aa

| Frame | from to | Length |
|---|---|---|
| +2 | 44..592 | 549 |
| -3 | 1..126 | 126 |
| -2 | 479..592 | 114 |
| +3 | 21..122 | 102 |

Length: 182 aa

*Figure 4: Open reading frame (ORF) search output of clone sequence POR_C070_P17. Figure A indicates the ORF search of a clone sequence including the vector sequence whereas; figure B indicates ORF without the vector sequence. Both ORF predictions showed the same frame which covers most of the query sequences. This signifies that the clone sequence is a full-length cDNA clone sequence. The 43 bps before the start of the open reading frame in figure B represents the 5'-untranslated region (5'-UTR).*

Finding the open reading frame of a sequence helps to identify the part of the gene which encodes for protein and assists gene prediction. The open reading frame of the clone POR_C070_P17 was searched both in the presence of the vector sequence and without the vector sequence. It was shown the second forward strand is the longest ORF and encode for 183 amino acids. Figure 4A indicates prediction of ORF including the vector sequence and the third frame is the longest one stretched from 81 to 629 base pair. Figure 4B indicates of ORF prediction without the vector sequence and covers 44 to 592 base pair. The first 43 base pairs are part of the gene but not part of the ORF and the position of the 5'-UTR region is presumably located in this region.

The blast outputs of each database were thoroughly inspected and there were significant number of sequences displayed hit to specific databases and were categorized as *database specific hits* (Examples Table 12 and 13). The pig genome showed to have highest number of database specific hits; 2,473 sequences provided hit only to the pig genome database. The pig cDNA database also provided 1,564 database specific hits. Whereas, Human cDNA, Mouse cDNA, and *E.coli* genome sequence provided 340, 109, and 5 specific hits respectively (Table 11).

Table 11: Summery of database specific blast hits

| Sr. No | Database | Number of Database specific blast hits* |
|---|---|---|
| 1 | Pig Genome(build 10.2) | 2,473 |
| 2 | Pig cDNA (build 10.2) | 1,564 |
| 3 | Human cDNA | 340 |
| 4 | Mouse cDNA | 109 |
| 5 | E.Coli genome | 5 |

**Database specific blast hits are clones that provided hit only in one of the databases but not in others.*

The human and mouse cDNA database specific blast hits are important for comparative mapping of pig genes. We can easily indicate genes that are not mapped on the pig genome by searching their homologous genes in either human or mouse genomes. This is due to the fact that both the human and the mouse genomes are studied comprehensively.

Table 12: Example of clone sequences provided hit only in Pig cDNA database

| Clone ID | Transcript ID | % Identity | Align. Length | Query Start | Query End | Subject Star | Subject End | E-Value | Bit Score |
|---|---|---|---|---|---|---|---|---|---|
| POR_C067_L01 | ENSSSCT00000019064 | 100 | 173 | 0 | 0 | 19 | 191 | 3.00E-87 | 320 |
| POR_C069_L16 | ENSSSCT00000007385 | 98.19 | 332 | 2 | 4 | 308 | 638 | 5.00E-164 | 577 |
| POR_C070_N16 | ENSSSCT00000014539 | 79.62 | 265 | 39 | 14 | 311 | 568 | 2.00E-43 | 176 |
| POR_C072_O24 | ENSSSCT00000010475 | 95.48 | 509 | 16 | 7 | 1 | 507 | 0 | 806 |

Table 13: Example of clone sequences provided hit only in Human cDNA database

| Clone ID | Transcript ID | % Identity | Align. Length | Query Start | Query End | Subject Star | Subject End | E-Value | Bit Score |
|---|---|---|---|---|---|---|---|---|---|
| POR_C063_H01 | ENST00000361264 | 85.09 | 161 | 18 | 174 | 132 | 287 | 4.00E-37 | 156 |
| POR_C064_B16 | ENST00000535674 | 85.71 | 168 | 3 | 165 | 1021 | 1185 | 1.00E-41 | 171 |
| POR_C070_K19 | ENST00000540734 | 83.69 | 325 | 45 | 8 | 41 | 361 | 555 | 875 |
| POR_C068_G01 | ENST00000263754 | 100 | 66 | 0 | 0 | 1 | 66 | 422 | 487 |

## 4.3 Identification of Pig Transcripts and Pig Genes

The blast output of all the useful sequence reads was analysed thoroughly for the number of transcripts and genes obtained. There were 12,461 clones which provided hit against the pig cDNA database. These clones were blasted against 70,023 transcripts. Nevertheless, most of the transcripts were redundant. The output was edited for redundancy and the total numbers of non-redundant transcripts were 7,074.

For example, the transcript ID ENSSSCT00000001005; is a novel transcript of the gene Plasma membrane calcium-transporting ATPase (ATP2B1) located on chromosome number **5:87,958,687-87,973,227** forward strand. It was obtained from five different clones (Detail see: Table 14).

Table 14: Example of redundant hits against the pig cDNA database

| Clone ID | Transcript ID | per cent Identity | Alignment Length | Query Start | Query End | Subject Star | Subject End | E-Value | Bit Score |
|---|---|---|---|---|---|---|---|---|---|
| POR_C033_K13 | ENSSSCT00000001005 | 93.85 | 260 | 41 | 298 | 1494 | 1753 | 5.00E-116 | 416 |
| POR_C037_J18 | ENSSSCT00000001005 | 96.76 | 278 | 39 | 315 | 1978 | 2253 | 1.00E-132 | 472 |
| POR_C049_N16 | ENSSSCT00000001005 | 94.02 | 234 | 36 | 264 | 1980 | 2213 | 5.00E-101 | 366 |
| POR_C070_C04 | ENSSSCT00000001005 | 99.1 | 555 | 28 | 582 | 1977 | 2529 | 0 | 996 |
| POR_C071_O02 | ENSSSCT00000001005 | 98.82 | 170 | 385 | 553 | 19 | 188 | 3.00E-81 | 302 |

The number of genes discovered from all useful sequence reads showed higher degree of redundancy like that of transcripts. For example, four different clones mentioned in table 15 provided hit against the Acetyl-CoA acyltransferase 1 gene (ACAA1); A gene encode for an enzyme operative in the beta-oxidation system of the peroxisomes located in chromosome number **13: 25,168,719-25,179,429** reverse strand.

Table 15: Example of redundant clone hits against the gene ACAA1

| Clone ID | Transcript ID | Sus_Gene | Ensembl_<br>GeneID | Chr | Start | End |
|---|---|---|---|---|---|---|
| POR_B006_E07 | ENSSSCT00000012317 | ENSSSCG00000011250 | ACAA1 | 13 | 25168719 | 25179429 |
| POR_B025_H14 | ENSSSCT00000012317 | ENSSSCG00000011250 | ACAA1 | 13 | 25168719 | 25179429 |
| POR_C040_B10 | ENSSSCT00000012317 | ENSSSCG00000011250 | ACAA1 | 13 | 25168719 | 25179429 |
| POR_C047_K12 | ENSSSCT00000012317 | ENSSSCG00000011250 | ACAA1 | 13 | 25168719 | 25179429 |

The blast output was merged with a gene annotation file used by ensembl which contains list of predicted genes with their transcripts to infer the number of genes obtained from the direct sequencing of the cDNA library. The file with cDNA database blast output and list of genes was edited for redundancy and a total number of 6,877 non-redundant genes were obtained. The numbers of non-redundant genes discovered from the first experiment were 3,028. Similarly, numbers of non-redundant genes discovered from the second and third experiments were 2,242 and 1,607 respectively. Figure below summarizes the number of non-redundant genes discovered from the three experiments

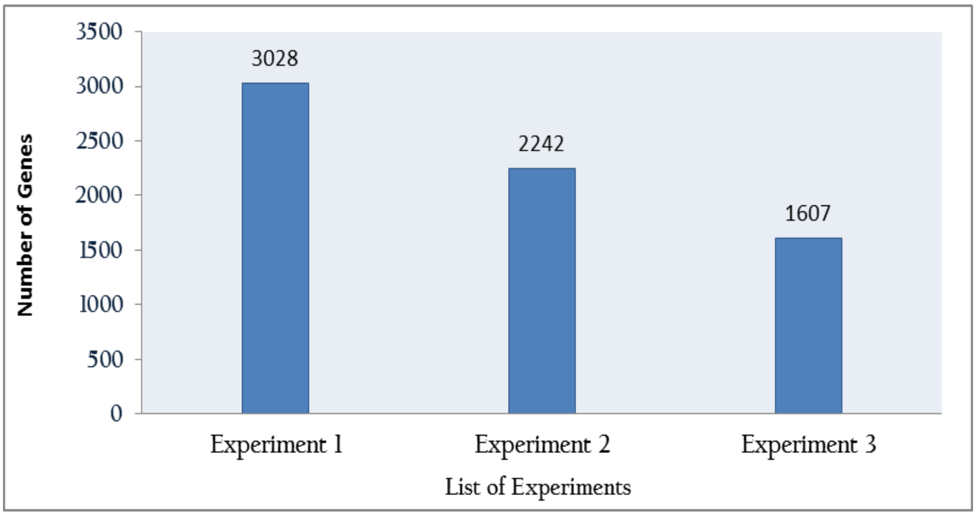


*Figure 5: Histogram of the number of non-redundant genes obtained from each experiment.*

The probability of finding new non-redundant genes from a cDNA library is higher in the first batch of sequenced plates than in last ones. This is the reason why the first experiment provided higher number of non-redundant genes than the other two experiments regardless of the number of 384 well-plates sequenced. The average number of genes discovered from each plate sequenced also showed variation among experiments. The first experiment has displayed higher number of non-redundant genes per plate. This is due to higher probability of each sequence being new and non-redundant. The number of non-redundant gene per plate discovered in the second experiment is relative lower than the two experiments. Table 16 illustrates the average number of non-redundant genes obtained per plate sequenced in the three experiments.

Table 16: Average number of non-redundant genes discovered per plate

| Experiment | Number of -plates sequenced | Average number of Non-redundant genes per plate |
|---|---|---|
| 1 | 20 | 151 |
| 2 | 32 | 70 |
| 3 | 18 | 89 |
| **Average** | **70** | **98** |

Sequencing more plates can minimize possibility of discovering non-redundant genes from the cDNA library. As shown in figure 6, there is still higher possibility of finding non-redundant genes by sequencing more new plates. On the other hand, the number of non-redundant genes obtained from plates POR_C062, POR_C063, POR_C064 and POR_C065 are comparatively lower. The overall sequencing efficiency of these plates was also much lower than the remaining plates; 33.59, 49.48, 66.15 and 58.85% respectively and the number of useful sequence reads generated from these plates was fewer (Detail see: Table 3).

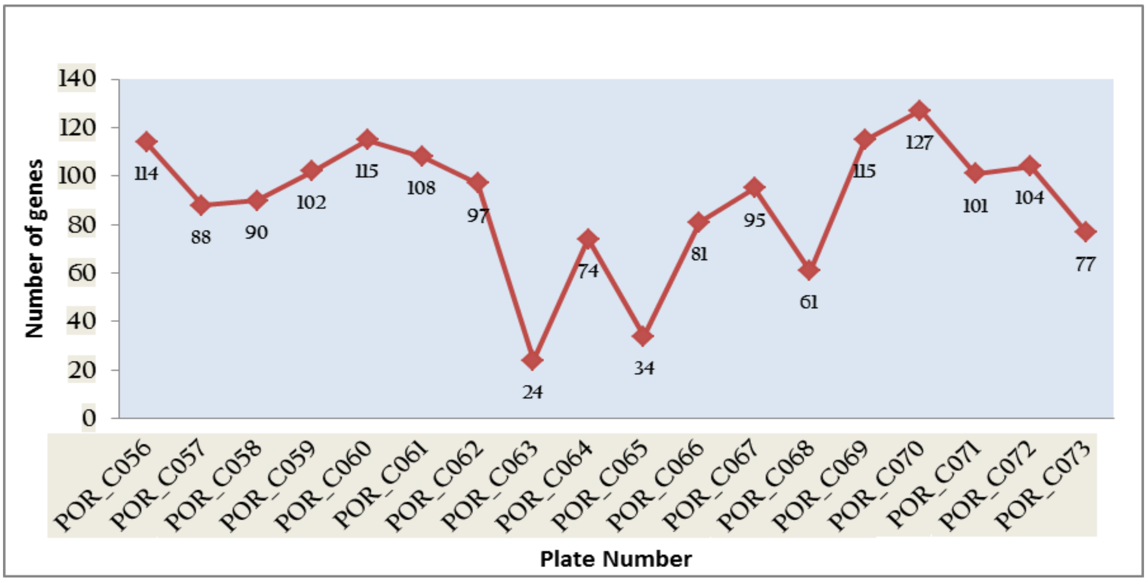
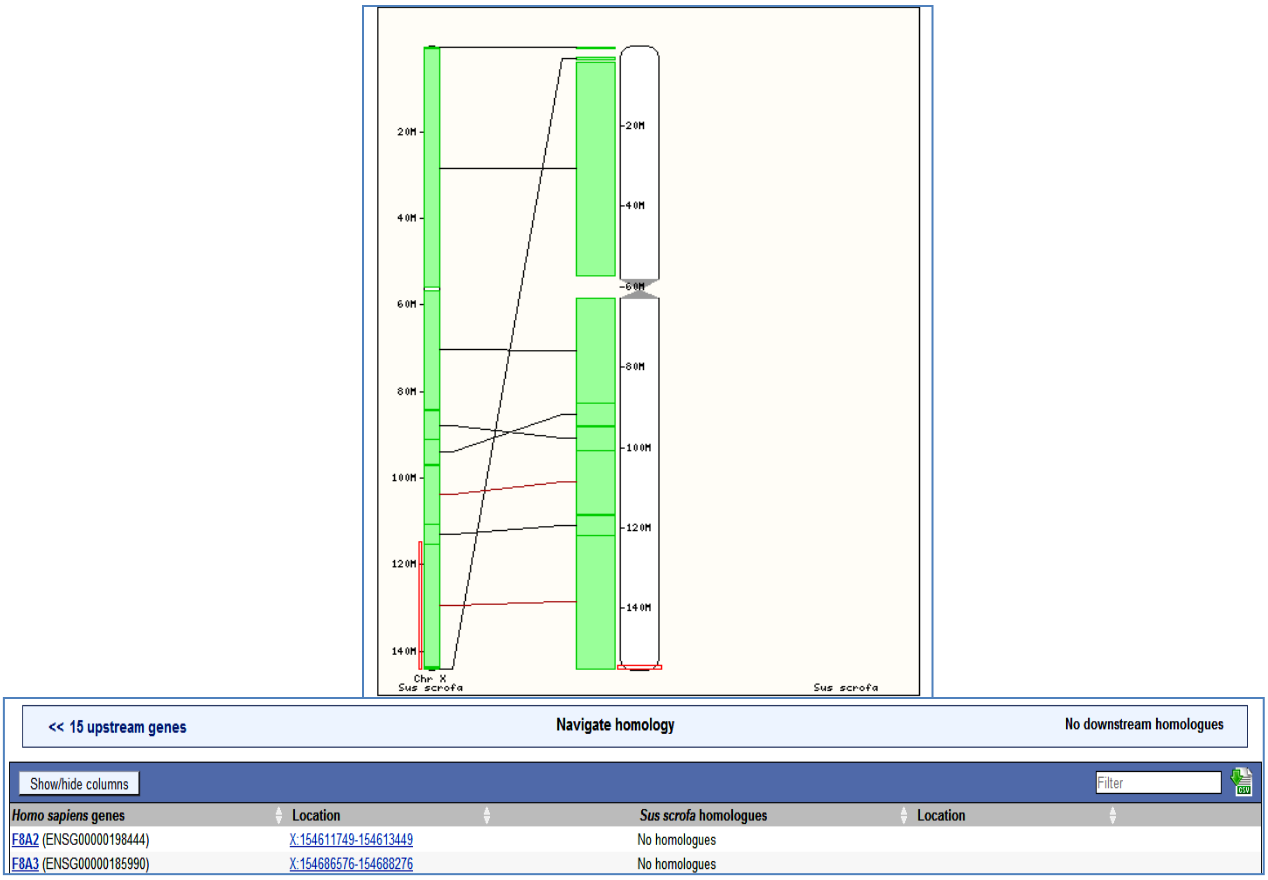


*Figure 6: number of non-redundant genes retrieved from each plates of experiment 3.*

### 4.3.1 Identification of Homologous Pig Genes in Human and Mouse genomes

The blast output of sequences showed significant number of database specific hits against both human and mouse genome. Thus, clone sequences which are specific to either of the human or mouse genome are vital sources for homologous pig genes identification. Genes that are not mapped onto the pig genome can be mapped by observing for the presence of identical flanking genes in both species. For instance, the clone sequence POR_C068_I17 provided hit only in human cDNA database to the human transcript ENST00000369505 located on chromosome **X: 154,609,763-154,614,139** forward strand. The transcript is one of the 9 gene products of the coagulation factor VIII-associated 2 (F8A2) gene. The homologous pig gene was searched by looking for a syntenic region shared by both human and pig. It was revealed that the gene F8A2 human gene does not have a homologous pig gene (Figure 7 and 8).

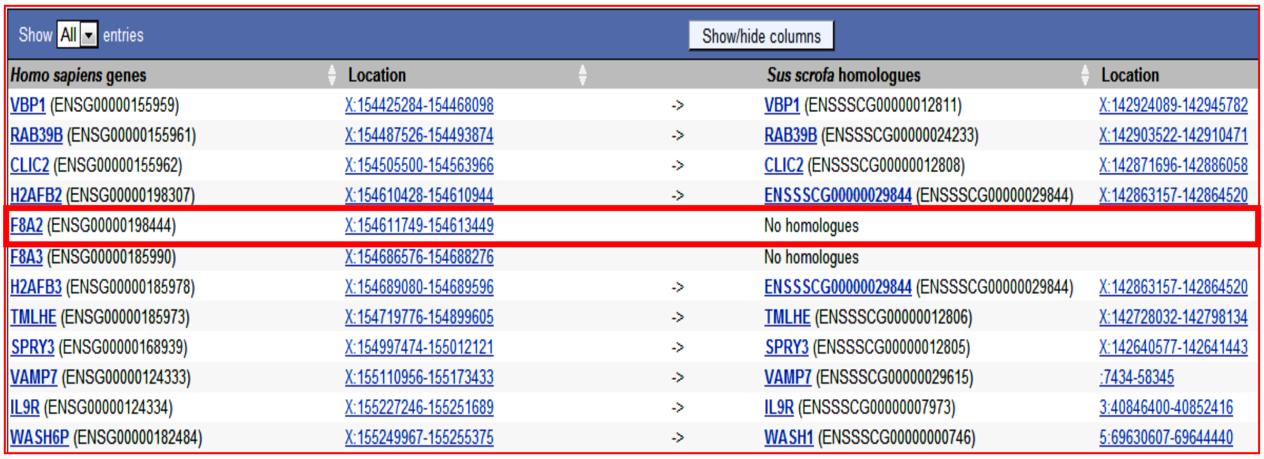


<< 15 upstream genes | Navigate homology | No downstream homologues

| *Homo sapiens* genes | Location | *Sus scrofa* homologues | Location |
|---|---|---|---|
| F8A2 (ENSG00000198444) | X:154611749-154613449 | No homologues | |
| F8A3 (ENSG00000185990) | X:154686576-154688276 | No homologues | |

*Figure 7: Homologous pig gene of the human gene* F8A2 (ENSG00000198444)*. There is not homologous pig gene displayed on the figure.*

The flanking genes around the F8A2 human gene were also navigated and compared with the flanking pig genes. The upstream and downstream genes are identical in both human and pig except the F8A2 and F8A3 genes which are not mapped on the pig genome. Therefore, we can deduce that the gene is a true homologous gene and not mapped on the pig genome. We can predict the position of both the F8A2 gene is on chromosome **X: 142,864,520** and **142,728,032** between the pig genes CLIC2 and TMLHE (Figure 8).



| *Homo sapiens* genes | Location | | *Sus scrofa* homologues | Location |
|---|---|---|---|---|
| VBP1 (ENSG00000155959) | X:154425284-154468098 | -> | VBP1 (ENSSSCG00000012811) | X:142924089-142945782 |
| RAB39B (ENSG00000155961) | X:154487526-154493874 | -> | RAB39B (ENSSSCG00000024233) | X:142903522-142910471 |
| CLIC2 (ENSG00000155962) | X:154505500-154563966 | -> | CLIC2 (ENSSSCG00000012808) | X:142871696-142886058 |
| H2AFB2 (ENSG00000198307) | X:154610428-154610944 | -> | ENSSSCG00000029844 (ENSSSCG00000029844) | X:142863157-142864520 |
| F8A2 (ENSG00000198444) | X:154611749-154613449 | | No homologues | |
| F8A3 (ENSG00000185990) | X:154686576-154688276 | | No homologues | |
| H2AFB3 (ENSG00000185978) | X:154689080-154689596 | -> | ENSSSCG00000029844 (ENSSSCG00000029844) | X:142863157-142864520 |
| TMLHE (ENSG00000185973) | X:154719776-154899605 | -> | TMLHE (ENSSSCG00000012806) | X:142728032-142798134 |
| SPRY3 (ENSG00000168939) | X:154997474-155012121 | -> | SPRY3 (ENSSSCG00000012805) | X:142640577-142641443 |
| VAMP7 (ENSG00000124333) | X:155110956-155173433 | -> | VAMP7 (ENSSSCG00000029615) | :7434-58345 |
| IL9R (ENSG00000124334) | X:155227246-155251689 | -> | IL9R (ENSSSCG00000007973) | 3:40846400-40852416 |
| WASH6P (ENSG00000182484) | X:155249967-155255375 | -> | WASH1 (ENSSSCG00000000746) | 5:69630607-69644440 |

*Figure 8: Upstream and downstream comparison of flanking genes of the F8A2 gene between human and pig. The pig genes CLIC2 and TMLHE are the flanking genes to the homologous gene in both human and pig.*

Similarly, the clone sequence POR_C070_K19 provided hit only against the Human cDNA database. It was blasted against human gene RPA interacting protein (RPAIN). The gene is located on chromosome **17:5,322,961-5,336,196** forward strand of human genome. The homologous pig gene is not located in the pig genome. The flanking genes of the RAPIN gene are MED31 and TXNDC17 in both human and pig genomes. Therefore, the position of the gene RPAIN in the pig genome is between the position of MED3 and TXNDC17 pig genes.

There are also 109 sequences displayed hit only to the mouse cDNA database. For example, the clone sequence POR_C058_A17 provided hit only in the mouse cDNA data base. It provided the gene ENSMUSG00000020719; a DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 (Ddx5) located on chromosome **11: 106,641,669-106,650,499** reverse strand of the mouse genome. The homologous pig gene cannot be found through the mouse genome. Nevertheless, it has a homologous human gene DDX5 (ENSG00000108654) located on chromosome **17:62,494,374-62,502,484** reverse strand of the human genome. The homologous pig gene can easily be navigated from the human DDX5 gene. However, the human gene DDX5 has no homologous pig gene. In addition to the DDX5 gene there are three other human genes (POLG2, LRR37A3 and RGS9) which are not mapped on the pig genome. Further inspection on both upstream and downstream of the DDX5 gene showed that identical genes are located in both human and pig genomes at the specific location. Therefore, we can deduce that the pig DDX5 gene is not mapped in the pig genome and its location is between **12: 13,602,445 and 12:13,030,855** on the pig genome.

### 4.3.2 Identification of Clone Sequences which did not provide hit to any of the data bases

Among the 19,470 obtained sequences which were blasted against the pig databases, Human cDNA and the Mouse cDNA databases only 80% of the sequences (15,388 sequences) provided hits in either of the databases. The remaining 20% of the sequences (4,082) did not provide hit in any of the data bases. To analysis these sequences further we took 10 sample sequences blasted the entire sequence against the nucleotide collection. It was revealed that 8 of the sequences provided hit against the pig genome but the start of the query sequences are beyond the first 100 base pairs or the alignment length is too short to be considered as a significant hit within the range of the given e-value. Meanwhile the sequence quality of the sequences was very low with several unidentified nucleotides (N) which might cause shorter alignment length.

## 4.4 PCR and Gel-Electrophoresis Protocol Optimization

An appropriate PCR and gel-electrophoresis protocols were established for transcript length estimation of clones blasted to identical genes. A DNA samples with different dilution rates were used to run PCR reactions with an annealing temperature of 50 and 55 °C. Similarly, size of PCR products was examined in agarose gel with different agarose percentage and electrophoresis running time. Pictures of the agarose gel analysis were examined for clarity of bands and presence of primer dimer. It was shown that PCR reactions with DNA dilution rate of both ± 1:7 and ± 1:12 and annealing temperature of 55 °C are best visualized in 1% agarose when running in the electrophoresis for three hours. Thus, it was optimum protocol to determine insert size of the clones effectively (figure 9).

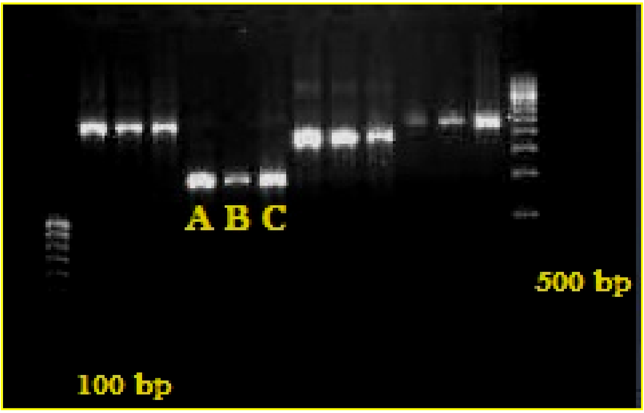


*Figure 9: Optimized PCR and gel-electrophoresis protocols with different DNA dilution rate, annealing temperature and gel running time. The picture represents 1% agarose, 3 hours of running time and annealing temperature of 55 °C. Letters A, B and C represent DNA dilution rates of ± 1:7 and ± 1:12 and stock undiluted DNA respectively. Size standards of 100 and 500 bp are indicated on the picture. The picture showed better result; single bands with no primer dimer.*

## 4.5 Transcript Length Estimation

The blast output of the pig cDNA database was thoroughly inspected for the presence of clone sequences with multiple hit against identical gene. There were several sequences blasted to identical genes. This might be due to either redundancy in the cDNA library or clones are from different transcripts of a single gene. However, the cDNA library is normalized and checked for redundancy by the commercial company. Thus, elucidating further for variation in the insert size among individual clones can be insightful. We selected 108 clones which provided hit to 10 different genes (Appendix 5 table s5). Clones were amplified by the optimized PCR protocol using both universal T3 forward and universal reverse T7 primers. Insert size was estimated using agarose gel electrophoresis (Appendix 4 Figure F1).

The result showed that there is variation in insert size of clones of the same gene (Appendix 4 Figure F1). It was also confirmed that the size of most of the selected clone was longer than the mean insert size of the cDNA library (i.e 2 kb). To decipher the variation in insert size among clones of the same gene, sequencing the clones from their 3'-end using universal T7 reverse primer was considered. Nonetheless, the sequencing procedure was not efficient and sequence reads were of bad quality. The presence of poly-A tail at the 3'-end of the clone sequences prevented to give a good sequence. Therefore, further sequencing of clones from their 5'-end by using internal primers was considered.

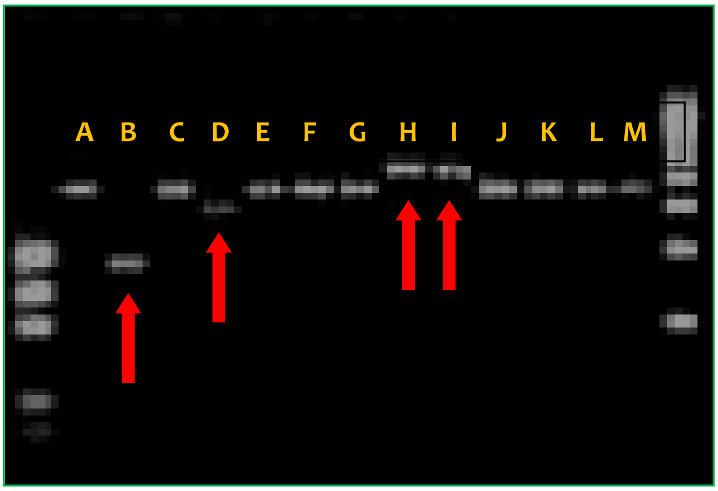


*Figure 10: cDNA insert size variation of SLA-3 gene is represented by 13 clones. 4 of the 13 clone sequences indicated by red arrows have different size than the remaining. Letters B, D, H and I represents clone sequences POR_B011_O02, POR_C058_G07, POR_C050_H02 and POR_C054_O20 respectively. The size standards of 500 bp and precision marker are located at the right and left of the PCR amplicon.*

The Swine leucocyte Antigen-3 gene (SLA-3) (ENSSSCG00000001227) was selected for further analysis. The SLA-3 gene is a classical major histocompatibility complex type I antigen family (MHC Class I) located on chromosome **7:24,641,613-24,645,323** of the pig genome. According to Ensembl, the gene has two transcripts ENSSSCT00000001325 and ENSSSCT00000001325 which are 1,733 and 1,730 bp long respectively. The gene transcripts have 9 exons encoding for 363 and 349 amino acids respectively (figure 11).

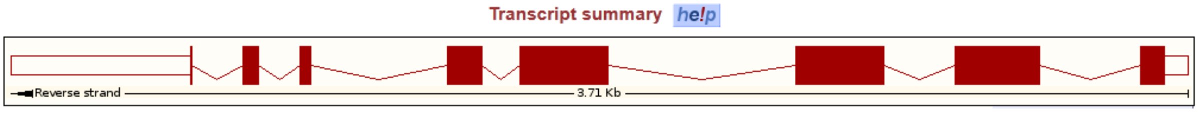


*Figure 11: Transcript summary of ENSSSCT00000001325. The transcript has 9 exons with reverse strand orientation. The line between each exon is position of the introns and the light boxes at both ends are 5' UTR and 3' UTR regions.*

Fragment size of the 13 clone sequences balsted to SLA-3 gene on agarose-gel showed variation (figure 10). 9 of the 13 clones (represented by letters A, C, E, F, G, J, K, L and M in figure 10) have fragment size between 1500 and 2000 bps. 2 clones (Letters H and I) have a fragment length between 2000 and 2500 bps. The remaining 2 clones (Letters B and D) have shorter fragment; around 800 bps and 1500 bps respectively. To elucidate further the variation in insert size among clones, they were sequenced using the universal T3 forward primers and aligned with the reference pig genome. Significant variation was shown on the exon-intron organization of clones which is in agreement with the Agarose gel analysis (Detail see Figure 14). The second exon of the clone POR_C054_O20 (*letter I in figure 10*) was longer than the remaining clone sequences which also showed to have longer fragment size on the agarose gel analysis. Sequencing the complete transcripts using internal primers revealed better picture of the exon-intron organization of all clone sequences. It was also proven that clone sequences which showed to have longer fragment size have longer exon sizes in one of their exons (Detail see: Figure 15).

The dot plot of the complete sequence of clones against the reference sequence of SLA-3 gene revealed that significant variation among sequences exists. The dot plot of each clone sequence is in consistent with both the Agarose gel analysis result and BLAT results indicated in (figure 10, 13 and 14). 4 of the clone sequences; POR_C054_O20, POR_c039_G07, POR_C050_H02 and POR_B011_O02 are significantly different from the remaining 9 clone sequences (Figure 14).

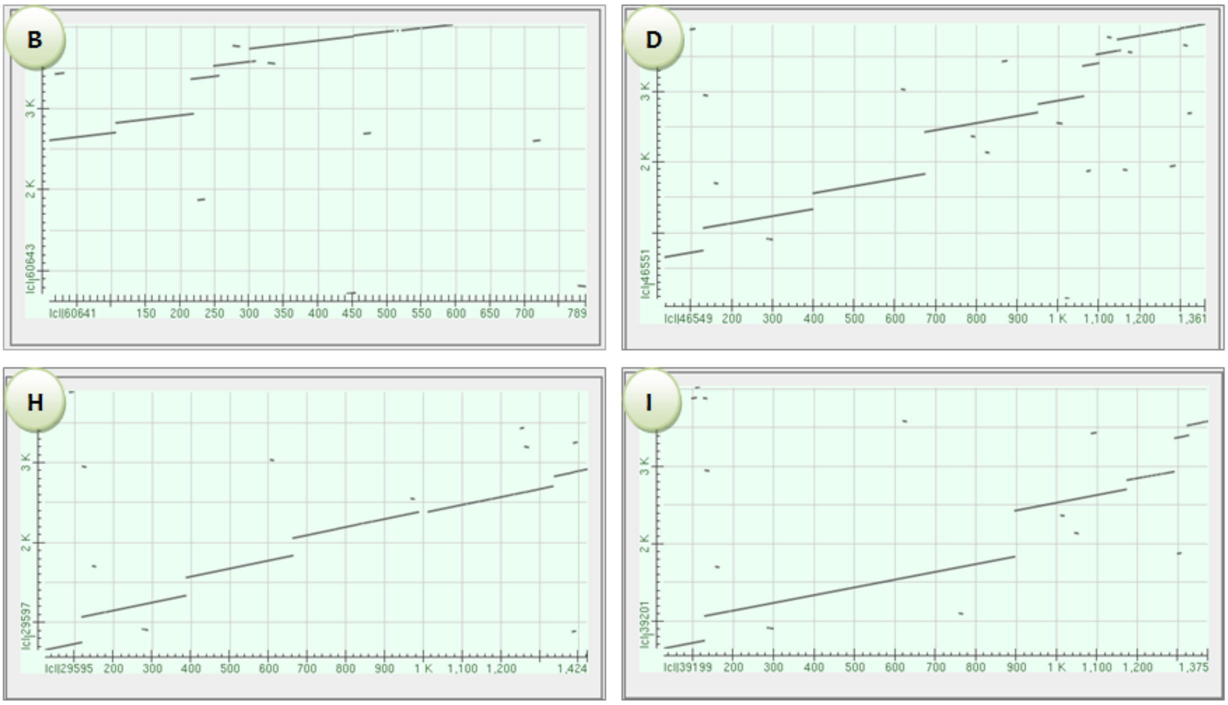
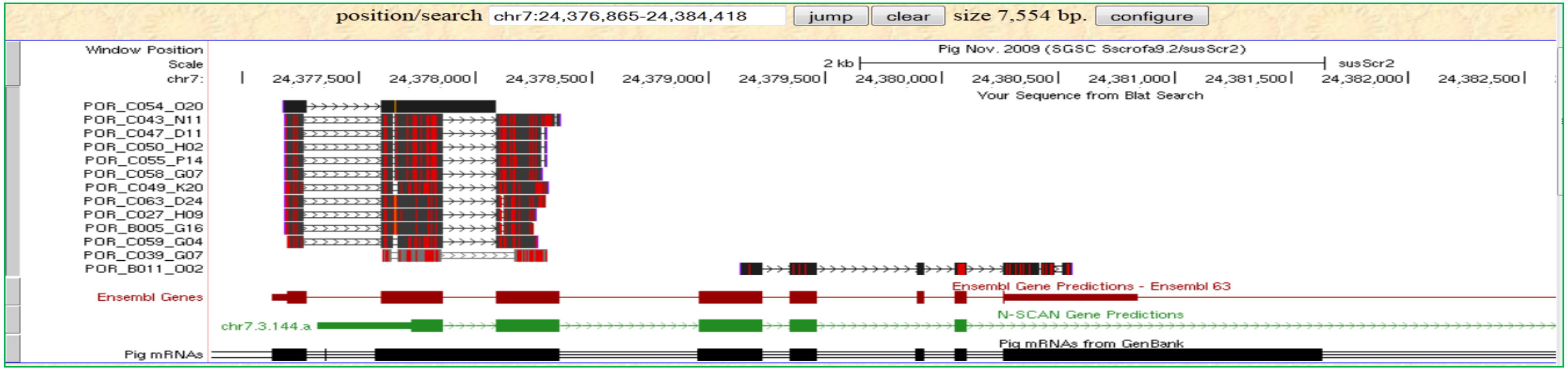


*Figure 12: Dot plot analysis for comparison of the 4 clone sequence which showed differences in insert size against the reference SLA-3 gene sequence. Letters B, D, H and I represent clone sequences POR_B011_O02, POR_C058_G07, POR_C050_H02 and POR_C054_O20 respectively.*

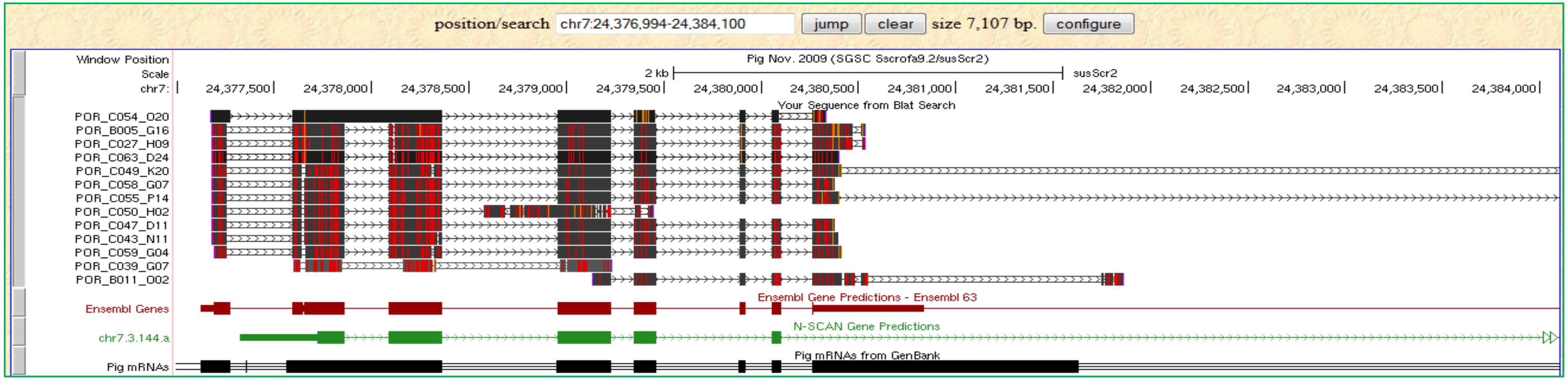
The figure above illustrates the variation among clone sequences. For instance, clone POR_C054_O20 indicated by letter I has a second exon that is longer than other clone sequences. This could be either due to segmental deletion in the other clone sequences or insertion into this particular clone sequence. This is in agreement with the agarose gel analysis result where insert size of this clone sequence is shown to be longer than other clones (Figure 10).The clone sequence POR_B011_O02 represented by letter B also revealed that the first 3 exons are not included in the transcript and the exon sizes are shorter than the others. Similarly, the clone POR_C058_G07 represented in letter D its first exon is not included which turned out to be shorter in size.

The presence of both the vector sequence and Open Reading Frame was checked for all clone sequences to confirm their full-lengthiness. All clone sequences except clone sequence POR_C039_G07 represented by letter D contain the vector sequences. It was separately blasted to the pig genome and it was not aligned to the first exon of the predicted gene. The four clone sequences which showed significant variation both in size and exon-intron organization were inspected for the presence of an open reading frame to confirm their full-lengthiness and presence of a gene. It was shown 4 of them have an open reading frame and are full-length.



position/search chr7:24,376,865-24,384,418 size 7,554 bp.

*Figure 13: Graphic representation of clone sequences obtained using the universal T3 primer aligned to the pig reference genome. The figure illustrates the exon-intron structure of clone sequences. 10 clone sequences showed higher degree of similarity in their organization except minor gaps in some of them. Meanwhile, 3 clones (POR_C054_O20, POR_C039_G07 and POR_B011_O02) Showed significant variation from the others.*



position/search chr7:24,376,994-24,384,100 size 7,107 bp.

*Figure 14: Graphic representation of the complete sequences of transcripts obtained using internal primers aligned to the reference pig genome. The figure shows the exon-intron arrangements of clone sequences in more depth than Figure 13. 9 of the 13 clone sequences showed higher degree of similarity whereas, 4 clone sequences showed significant variation from the remaining clones giving an insight of being splice variants. This figure is in agreement with the Agarose gel analysis.*

## 5. DISCUSSION

The objective of this study was to build a resource of porcine full-length cDNA clones with known gene annotation for further studies. For this reason we picked and sequenced the 5'-end of another 6,912 individual clones of a full-length normalized cDNA library constructed from 11 different porcine tissue samples. The study also intended to merge sequences results obtained from two previous experiments and blast to the newly released pig genome (Build 10.2), pig cDNA, Human cDNA and Mouse cDNA databases to retrieve the gene names, transcript name and their description. Additionally, to identify clone sequences blasted against identical genes and elucidate the variation in fragment size further.

Recent advancement in sequencing technologies like RNA-sequencing can give better understand in both expression of genes and relative abundance of transcripts (Wang *et al.*, 2009). However, sequencing cDNA library has an advantage over the RNA sequencing in a way the cloned cDNA can be used as back-up resources for further study on specific genes of interest (Natarajan *et al.*, 2010). Large scale screening and sequencing of cDNA library needs preparation of templates and cellular growth of bacterial colonies followed by plasmid purification. The plasmid purification steps remains expensive raising the cost of the whole sequencing procedure (Elkin *et al.*, 2001). Bypassing the plasmid DNA isolation procedure reduces the cost of sequencing by minimizing the amounts of reagents (Jennifer *et al.*, 2000). Previous experiment on the same cDNA library using direct sequencing on bacterial colonies was also proven to be cost effective way of large scale screening of cDNA libraries (Bernal *et al.*, 2011).

A total of 19, 470 individual sequences were obtained from the current and previous two studies with an average overall success rate of 72.46%. The sequencing success rate in 384 well-plate of the previous experiments was 71.21% (Bernal *et al.*, 2011) and 69. 34% (Ketema *et al.*, 2011) whereas, the success rate of this study is 79.4%. The overall sequence efficiency and average sequence length of this experiment is higher than the previous two experiments (Table 4). This is because during the first experiment the sequencing protocols were not fully optimized and technical failures in the second experiment. The cumulative effect of efficient sequencing, properly grown bacterial colonies, proper replicating procedures of plates, immediate processing of cell lysate, better pipetting skill and sample handling procedures resulted in better sequencing efficiency in this experiment. Moreover, there was no media contamination and all laboratory chemicals were available during the course of the experiment. The overall success rate of the cDNA library sequencing is higher than what Jennifer *et al.* (2000) obtained (66%). However, it is in consistence with the efficiency range of 75-80% obtained by Smith *et al*. (2000). The sequencing efficiency of this experiment could have been improved to up to 84% if the two plates with lower efficiency were not considered and technical inaccuracies were avoided. The computer aided bacterial colonies in liquid media can be a useful input in replacing manual and laborious procedure of bacterial colonies picking and transforming into well-plates throughout the sequencing procedure (Yehezkel *et al*., 2011).

Sequence reads were blasted against the pig genome, pig cDNA, Human cDNA and Mouse cDNA databases providing 12,222, 12,461, 8,300 and 5,268 hits respectively. The first 100 bp of the sequences were used for blasting against the pig genome database in order to avoid redundancy in the output file. The number of genes that could have been obtained is undermined as some sequences with no hit in all of the data bases displayed hit in the pig genome after the 100 bp. Significant amount of database specific hits were obtained; the pig genome database provided 2,473 database specific hits. These blast hits cannot be found in both the pig cDNA database and the database of expressed sequence tag (dbEST). This is due to the fact that ESTs are generated by cDNA library sequencing constructed from various tissues and developmental stages. Therefore, the 2,473 database specific sequences are possible candidates of novel EST obtained from this experiment.

Additionally, the human and mouse cDNA databases provided 340 and 109 database specific hits respectively. These database specific hits can be vital sources to map homologous pig genes which are not mapped on the pig genome. The human and mouse genomes are comprehensively studied than the pig genome and can be used to identify homologous pig genes which are not mapped on the pig genome. Fahrenkrug *et al.* (2002) underlined the importance of pig EST comparison with species of close evolutionary relation and comprehensively studied genome for mapping the pig genome comparatively. The newly released pig genome contains 21,640 protein-coding genes and 26,487 gene transcripts and its coverage is about 95%. It is expected to find nearly 1,000 unmapped genes on the genome (Martien A.M. Groenen, Personal communication). The number of hits specific to both human and mouse cDNA databases are in the range of the expectation. A total of 6,877 non-redundant pig genes and 7,074 non-redundant pig transcripts were obtained from the cDNA library sequenced. This represents 31.8% of the total protein-coding pig genes. The coverage of gene discovery can be improved by sequencing and characterization cDNA libraries constructed from various tissues and developmental stages. The cDNA library sequenced was constructed from 11 different tissues of an adult pregnant cloned pig. Sequence output of this study can only discover genes expressed in the tissues and developmental stage of the pig when the cDNA library is constructed.

The study also aimed to estimate transcript length of clones blasted to identical genes by Agarose gel analysis and further investigate the variation in insert size. The insert size variation was confirmed among 108 clones blasted to 10 genes. Gupta *et al.* (2004) proposed that the computational approach of alternative splice variants prediction should be accompanied by experimental validation for accurate delineation of tissue specific transcripts. Splice variants can be effectively revealed using a combined cDNA library screening and RT-PCR (Angelotti and Hofmann, 1996)

The SLA-3 gene is one of the three classical Major histocompatibility class-I genes. It was represented by 13 clone sequences of the blast output and insert size variation was confirmed by Agarose-gel analysis. Moreover, further sequence analysis through the 5'-end using the universal T3 forward primer and internal primers also confirmed the variation in insert size. The Exon-Intron organization of all the 13 transcripts was inspected using the BLAT tool of UCSC genome browser against the pig reference genome. As expected the 4 clone sequences showed significant variation than the remaining. Expression of mRNA is spatiotemporal dependent. Thus, different transcripts of identical genes can be expressed in different tissues and developmental stages (Gupta *et al.*, 2004). These tissues specific transcripts can have different arrangement and exon sizes. For instance, the second exon of clone POR_C054_O20 as shown in figure 13 and figure 11 is longer than the remaining clone sequences could be the reason for having longer fragment size. On the other hand, the first three exons are missing from clone sequence POR_B011_O02. Considering the alternative splicing nature of mRNA, the variation in both the insert size and genomic organization between transcripts of identical gene could be an indication for the presence of splice variants. However, the clone sequence POR_C039_G07 is represented only by second, third and fourth exons.

Thorough search in Gene-Bank was made for mRNAs with similar exon-intron organisation like the clone sequences which showed variation than the remaining clone sequences. The mRNA with accession number AK237682 located on Chromosome 7:24,377,188–24,397,078 of the pig genome has similar exon-intron organization with the clone sequence POR_C054_O20. Like the clone sequence the mRNA have longer second exon than the remaining mRNAs. Further information described that the mRNA (AK237682) is expressed in spleen (Uenishi *et al.*, 2004). It is long know that spleen is important in body immunity system of almost all vertebrates. This could be an indirect confirmation for the specific expression of the clone sequence in spleen for several reasons; first the gene SLA-3 is MHC class-I Antigen which plays a vital role in body immunity system. Secondly, expression of MHC genes including SLA-3 in spleen is expected due to the fact that expression of genes is time and space and spleen is immunologically important organ (Gupta *et al.*, 2004). Furthermore, the porcine cDNA library is constructed from 11 different tissues an adult pregnant cloned pig and tissue samples from spleen were included in the cDNA library construction.

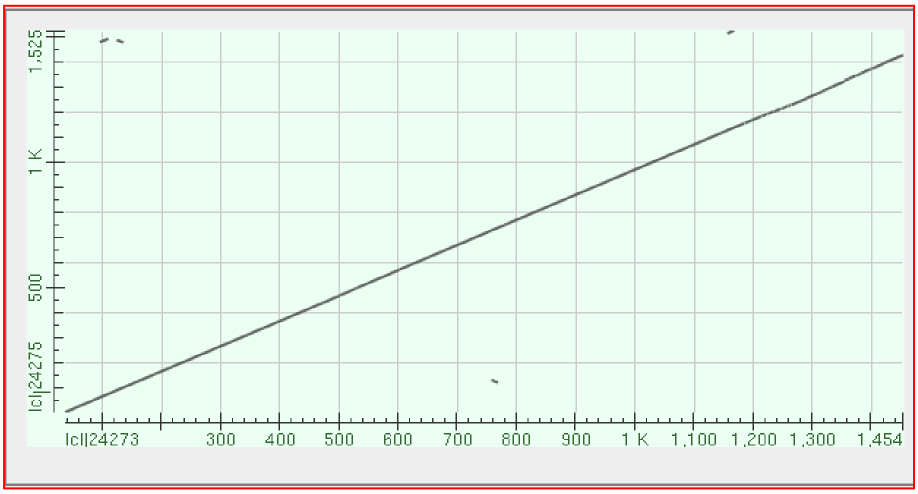


*Figure 15: Dot plot analysis for comparison of the clone sequence POR_C054_O20 with mRNA sequence expressed in spleen (AK237682). The clone sequence and the mRNA sequence are represented on X and Y-Axis respectively.*

Comparison of the clone sequence POR_C054_O20 with the mRNA sequence (AK237682) revealed that there is higher degree of similarity. The dot plot shows that the two sequences are nearly identical (Figure 15). We can deduce that the clone POR_C054_O20 is tissues specific splice variant expressed in spleen of adult pregnant pig.

There are also several gaps between exons of clones sequences and signify polymorphic nature of the sequences; small deletion and insertion. Rothschild and Ruvinsky (2011) describe that higher degree of within loci polymorphism is a remarkable feature of the MHC genes which increase the range of foreign antigen recognition. Smith *et al.* (2005) also mentioned that the SLA-3 gene is highly polymorphic and among 32 published DNA sequences 20 are unique. There is a distinct insertion of nine bp as a result of duplication creating additional insertion of three amino acids at the SLA-3*6 allele (Smith *et al.*, 2005). The human leucocyte antigen (HLA) is also the most polymorphic region of the human genome which signifies the polymorphic nature of the MHC gene family across species (Horton *et al.*, 2008).

# 6. CONCLUSION AND RECOMMENDATION

The direct sequencing technique of a normalized full-length cDNA library is an efficient, cost effective but laborious procedure. It has several advantages over high throughput sequencing techniques like the RNA Sequencing in a way it provides physical access to clones in the quest for further study on specific genes. Finding the functional domains of genes, reporter gene assay can be performed in the presence of backup clones of specific genes. Besides the functional screening of genes, the procedure can also be a vital resource to infer tissue specific splice variants. Blasting clone sequences against both the human and mouse genome helps for comparative mapping of homologous pig genes which are not mapped on the pig genome.

The rate of non-redundant gene discovery from the cDNA library is still high. Thus, sequencing more 384 well-plates is highly recommended to fully exploit the genes present in the cDNA library. The blast output contains significant number of clones blasted against identical genes. Further validation of transcripts by Agarose gel analysis, Sequencing the complete transcripts and decipher the exon-Intron organization is required.  It is also recommended to perform EST clustering analysis using gene ontology tools to functionally categorize the expressed genes according the biological process, cellular component and molecular function. The 2, 473 candidate novel ESTs found on this experiment should be validated further and submitted to the dbEST or the Pig Expression Data Explorer (PEDE). The 4, 082 clone sequences which didn't provide hit in any of the databases should be blasted against the available RNA-seq data or vice versa and identify the gene they blast against

## REFERENCES

ADAMS, M. D., DUBNICK, M., KERLAVAGE, A. R., MORENO, R., KELLEY, J. M., UTTERBACK, T. R., NAGLE, J. W., FIELDS, C. & VENTER, J. C. 1992. Sequence identification of 2, 375 human brain genes. *Nature,* 355**,** 632-634.

AL-SWAILEM, A. M., SHEHATA, M. M., ABU-DUHIER, F. M., AL-YAMANI, E. J., AL-BUSADAH, K. A., AL-ARAWI, M. S., AL-KHIDER, A. Y., AL-MUHAIMEED, A. N., AL-QAHTANI, F. H., MANEE, M. M., AL-SHOMRANI, B. M., AL-QHTANI, S. M., AL-HARTHI, A. S., AKDEMIR, K. C., INAN, M. S. & OTU, H. H. 2010. Sequencing, analysis, and annotation of expressed sequence tags for Camelus dromedarius. *PLoS One,* 5**,** e10720.

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K. & WATSON, J. 1994. Molecular biology of the cell Garland Publishing. *New York***,** 3-11.

ANGELOTTI, T. & HOFMANN, F. 1996. Tissue-specific expression of splice variants of the mouse voltage-gated calcium channel [alpha] 2/[delta] subunit. *FEBS letters,* 397**,** 331-337.

ARCHIBALD, A. L., BOLUND, L., CHURCHER, C., FREDHOLM, M., GROENEN, M. A., HARLIZIUS, B., LEE, K. T., MILAN, D., ROGERS, J., ROTHSCHILD, M. F., UENISHI, H., WANG, J., SCHOOK, L. B. & SWINE GENOME SEQUENCING, C. 2010. Pig genome sequence--analysis and publication strategy. *BMC Genomics,* 11**,** 438.

BERNAL, S., CROOJIMANS, R. & GROENEN, M. A. 2011. Characterization of a normalized full-length cDNA library from a cloned pig. *Animal Breedning and Genomics Centre, Wageningen University, The Netherlands.*

BONIZZONI, P., RIZZI, R. & PESOLE, G. 2006. Computational methods for alternative splicing prediction. *Brief Funct Genomic Proteomic,* 5**,** 46-51.

CARNINCI, P. 2000. Normalization and Subtraction of Cap-Trapper-Selected cDNAs to Prepare Full-Length cDNA Libraries for Rapid Discovery of New Genes. *Genome Research,* 10**,** 1617-1630.

CARNINCI, P. 2007. Constructing the landscape of the mammalian transcriptome. *J Exp Biol,* 210**,** 1497-506.

CARNINCI, P., KVAM, C., KITAMURA, A., OHSUMI, T., OKAZAKI, Y., ITOH, M., KAMIYA, M., SHIBATA, K., SASAKI, N., IZAWA, M., MURAMATSU, M., HAYASHIZAKI, Y. & SCHNEIDER, C. 1996. High-Efficiency Full-Length cDNA Cloning by Biotinylated CAP Trapper. *Genomics,* 37**,** 327-336.

CARNINCI, P., SHIBATA, Y., HAYATSU, N., ITOH, M., SHIRAKI, T., HIROZANE, T., WATAHIKI, A., SHIBATA, K., KONNO, H. & MURAMATSU, M. 2001. Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel [lambda]-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics,* 77**,** 79-90.

CHEN, C. H., LIN, E. C., CHENG, W. T., SUN, H. S., MERSMANN, H. J. & DING, S. T. 2006. Abundantly expressed genes in pig adipose tissue: an expressed sequence tag approach. *J Anim Sci,* 84**,** 2673-83.

ELKIN, C. J., RICHARDSON, P. M., FOURCADE, H. M., HAMMON, N. M., POLLARD, M. J., PREDKI, P. F., GLAVINA, T. & HAWKINS, T. L. 2001. High-throughput plasmid purification for capillary sequencing. *Genome Research,* 11**,** 1269-1274.

FAHRENKRUG, S. C., SMITH, T. P. L., FREKING, B. A., CHO, J., WHITE, J., VALLET, J., WISE, T., ROHRER, G., PERTEA, G., SULTANA, R., QUACKENBUSH, J. & KEELE, J. W. 2002. Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. *Mammalian Genome,* 13**,** 475-478.

FANG, M., HU, X., JIANG, T., BRAUNSCHWEIG, M., HU, L., DU, Z., FENG, J., ZHANG, Q., WU, C. & LI, N. 2005. The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Anim Genet,* 36**,** 7-13.

FAO 2009. food outlook- Global Market analysis. *FAO trade and market division*.

FRÖNICKE, L., CHOWDHARY, B., SCHERTHAN, H. & GUSTAVSSON, I. 1996. A comparative map of the porcine and human genomes demonstrates ZOO-FISH and gene mapping-based chromosomal homologies. *Mammalian Genome,* 7**,** 285-290.

GUPTA, S., ZINK, D., KORN, B., VINGRON, M. & HAAS, S. 2004. Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics,* 5**,** 72.

HORTON, R., GIBSON, R., COGGILL, P., MIRETTI, M., ALLCOCK, R. J., ALMEIDA, J., FORBES, S., GILBERT, J. G. R., HALLS, K. & HARROW, J. L. 2008. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics,* 60**,** 1-18.

JENNIFER, G., GLADDEN, B., RAY, R., GIETZ, R. D. & MOWAT, M. R. A. 2000. Rapid Screening of Plasmid DNA by Direct Sequencing from Bacterial Colonies. *BioTechniques,* 29**,** 436-437.

JOHNSON, J. M., CASTLE, J., GARRETT-ENGELE, P., KAN, Z., LOERCH, P. M., ARMOUR, C. D., SANTOS, R., SCHADT, E. E., STOUGHTON, R. & SHOEMAKER, D. D. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science,* 302**,** 2141-4.

KATO, S., OHTOKO, K., OHTAKE, H. & 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., ARAKAWA, T., HARA, A., FUKUNISHI, Y., KONNO, H., ADACHI, J., FUKUDA, S., AIZAWA, K., IZAWA, M., NISHI, K., KIYOSAWA, H., KONDO, S., YAMANAKA, I., SAITO, T., OKAZAKI, Y., GOJOBORI, T., BONO, H., KASUKAWA, T., SAITO, R., KADOTA, K., MATSUDA, H., ASHBURNER, M., BATALOV, S., CASAVANT, T., FLEISCHMANN, W., GAASTERLAND, T., GISSI, C., KING, B., KOCHIWA, H., KUEHL, P., LEWIS, S., MATSUO, Y., NIKAIDO, I., PESOLE, G., QUACKENBUSH, J., SCHRIML, L. M., STAUBLI, F., SUZUKI, R., TOMITA, M., WAGNER, L., WASHIO, T., SAKAI, K., OKIDO, T., FURUNO, M., AONO, H., BALDARELLI, R., BARSH, G., BLAKE, J., BOFFELLI, D., BOJUNGA, N., CARNINCI, P., DE BONALDO, M. F., BROWNSTEIN, M. J., BULT, C., FLETCHER, C., FUJITA, M., GARIBOLDI, M., GUSTINCICH, S., HILL, D., HOFMANN, M., HUME, D. A., KAMIYA, M., LEE, N. H., LYONS, P., MARCHIONNI, L., MASHIMA, J., MAZZARELLI, J., MOMBAERTS, P., NORDONE, P., RING, B., RINGWALD, M., RODRIGUEZ, I., SAKAMOTO, N., SASAKI, H., SATO, K., SCHONBACH, C., SEYA, T., SHIBATA, Y., STORCH, K. F., SUZUKI, H., TOYO-OKA, K., WANG, K. H., WEITZ, C., WHITTAKER, C., WILMING, L., WYNSHAW-BORIS, A., YOSHIDA, K., HASEGAWA, Y., KAWAJI, H., KOHTSUKI, S. & HAYASHIZAKI, Y. 2001. Functional annotation of a full-length mouse cDNA collection. 409, 685-690.

KETEMA, T. K., CROOJIMANS, R. & GROENEN, M. A. 2011. Sequence Analaysis of a Porcine Normalized Full-Length cDNA library. *Animal Breedning and Genomics Centre, Wageningen University, The Netherlands.*

KIM, E., GOREN, A. & AST, G. 2008. Alternative splicing: current perspectives. *Bioessays,* 30**,** 38-47.

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. & 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.

LEE, K. T., BYUN, M. J., LIM, D., KANG, K. S., KIM, N. S., OH, J. H., CHUNG, C. S., PARK, H. S., SHIN, Y. & KIM, T. H. 2009. Full-length enriched cDNA library construction from tissues related to energy metabolism in pigs. *Mol Cells,* 28**,** 529-36.

LEPARC, G. G. & MITRA, R. D. 2007. A sensitive procedure to detect alternatively spliced mRNA in pooled-tissue samples. *Nucleic Acids Res,* 35**,** e146.

MAEDA, N., KASUKAWA, T., OYAMA, R., GOUGH, J., FRITH, M., ENGSTRÖM, P. G., LENHARD, B., ATURALIYA, R. N., BATALOV, S. & BEISEL, K. W. 2006. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genetics,* 2**,** e62.

NATARAJAN, P., KANAGASABAPATHY, D., GUNADAYALAN, G., PANCHALINGAM, J., SHREE, N., SUGANTHAM, P. A., SINGH, K. K. & 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., OH, Y., DIRISALA, V., CHOI, H., PARK, K.-K., KIM, J.-H. & PARK, C. 2010. A simple, rapid, efficient and inexpensive strategy for sequencing clones from cDNA libraries. *Biotechnology and Bioprocess Engineering,* 15**,** 817-821.

ROTHSCHILD, M. F. & RUVINSKY, A. 2011. *The genetics of the pig*, CABI Publishing.

SACHS, A. 2000. Physical and functional interactions between the mRNA cap structure and the poly (A) tail. *Translational control of gene expression***,** 447-465.

SHCHEGLOV, A., ZHULIDOV, P., BOGDANOVA, E. & SHAGIN, D. 2007. Normalization of cDNA libraries. *Nucleic Acids Hybridization Modern Applications***,** 97-124.

SMITH, D., LUNNEY, J., MARTENS, G., ANDO, A., LEE, J. H., HO, C. S., SCHOOK, L., RENARD, C. & CHARDON, P. 2005. Nomenclature for factors of the SLA class-I system, 2004. *Tissue Antigens,* 65**,** 136-149.

SMITH, T. P., GODTEL, R. A. & LEE, R. T. 2000. PCR-Based Setup for High-Throughput cDNA Library Sequencing on the ABI 3700™ Automated DNA Sequencer *BioTechniques,* 29**,** 628-700.

SMITH, T. P. L., FAHRENKRUG, S. C., ROHRER, G. A., SIMMEN, F. A., REXROAD, C. E. & KEELE, J. W. 2001. Mapping of expressed sequence tags from a porcine early embryonic cDNA library. *Anim Genet,* 32**,** 66-72.

TAN, W., CHEN, Y., ZHANG, L., LU, Y., LI, S., ZENG, R., ZENG, Y., LI, Y. & CHENG, J. 2006. Construction and Characterization of a cDNA Library from Liver Tissue of Chinese Banna Minipig Inbred Line. *Transplantation Proceedings,* 38**,** 2264-2266.

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-D488.

WANG, X.-L., WU, K.-L., LI, N., LI, C.-L., QIU, X.-M., WANG, A.-H. & WU, C.-X. 2006. Analysis of Expressed Sequence Tags from Skeletal Muscle-specific cDNA Library of Chinese Native Xiang Pig. *Acta Genetica Sinica,* 33**,** 984-991.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics,* 10**,** 57-63.

YAO, J., COUSSENS, P. M., SAAMA, P., SUCHYTA, S. & ERNST, C. W. 2002. Generation of expressed sequence tags from a normalized porcine skeletal muscle cDNA library. *Anim Biotechnol,* 13**,** 211-22.

YEHEZKEL, T. B., NAGAR, S., MACKRANTS, D., GREGORY LINSHIZ, Z. M., SHABI, U. & SHAPIRO, E. 2011. Computer-aided high-throughput cloning of bacteria in liquid medium. *BioTechniques,* 50**,** 124-127.

ZHANG, L., TAO, L., YE, L., HE, L., ZHU, Y.-Z., ZHU, Y.-D. & ZHOU, Y. 2007. Alternative Splicing and Expression Profile Analysis of Expressed Sequence Tags in Domestic Pig. *Genomics, Proteomics & Bioinformatics,* 5**,** 25-34.

# APPENDICES

## 1. LABORATORY PROTOCOLS

### 1.1 RNA extraction using RNeasy mini protocol (QIAGEN, Mini handbook)

1. Determine the correct amount of starting material, as maximum 100 µg of RNA
    - A. Adjust the sample to a volume of 100 µl with RNase-free water. Add 350 µl Buffer RLT, and mix well.
    - B. Add 250 µl of ethanol (96–100%) to the diluted RNA, and mix well by pipetting. Do not centrifuge. Proceed immediately to step "c".
    - C. Transfer the sample (700 µl) to an RNeasy Mini spin column placed in a 2 ml collection tube (supplied). Close the lid gently, and centrifuge for 15 s at ≥8000 x g (≥10,000 rpm). Discard the flow-through.
    - D. Re-use the collection tube in step "d".
    - E. Note: After centrifugation, carefully remove the RNeasy spin column from the collection tube so that the column does not contact the flow-through. Be sure to empty the collection tube completely.
    - F. Add 500 µl Buffer RPE to the RNeasy spin column. Close the lid gently, and centrifuge for 15 s at ≥8000 x g (≥10,000 rpm) to wash the spin column membrane. Discard the flow-through.
    - G. Reuse the collection tube in step "e".
    - H. Note: Buffer RPE is supplied as a concentrate. Ensure that ethanol is added to Buffer RPE before use (see "Things to do before starting").
    - I. Add 500 µl Buffer RPE to the RNeasy spin column. Close the lid gently, and centrifuge for 2 min at ≥8000 x g (≥10,000 rpm) to wash the spin column membrane. The long centrifugation dries the spin column membrane, ensuring that no ethanol is carried over during RNA elution. Residual ethanol may interfere with downstream reactions.
    - J. Note: After centrifugation, carefully remove the RNeasy spin column from the collection tube so that the column does not contact the flow-through. Otherwise, carryover of ethanol will occur.
    - K. Optional: Place the RNeasy spin column in a new 2 ml collection tube (supplied), and discard the old collection tube with the flow-through. Close the lid gently, and centrifuge at full speed for 1 min. Perform this step to eliminate any possible carryover of Buffer RPE, or if residual flow-through remains on the outside of the RNeasy spin column after step "e".

L. Place the RNeasy spin column in a new 1.5 ml collection tube (supplied). Add 30–50 µl RNase-free water directly to the spin column membrane. Close the lid gently, and centrifuge for 1 min at ≥8000 x g (≥10,000 rpm) to elute the RNA.

M. If the expected RNA yield is >30 µg, repeat step "g" using another 30–50 µl RNase free water, or using the eluate from step "g" (if high RNA concentration is required). Reuse the collection tube from step "g". If using the eluate from step "g", the RNA yield will be 15–30 per cent less than that obtained using a second volume of RNase-free water, but the final RNA concentration will be higher.

1.2 Lysogeny broth (LB) medium: to prepare 1 Liter

A. To 800 ml MQ water add :
   - 10 grams of Bactotryptone
   - 5 grams of yeast extract
   - 10 grams of NaCl

B. Adjust pH to 7.5 with NaOH

C. Adjust volume to 1 L with MQ water

D. Sterilize by autoclaving

**1.3 1 per cent agar in Lysogeny broth (LB): to prepare 1 Liter**

A. To 1 L of LB Add: 10 grams of agarose

B. Sterilize by autoclaving

C. When the medium reaches approximately 50 °c, add Ampicillin to a final concentration of 0.1 µg/ml

### 1.4 10x Freezing Medium: to prepare 1 L

A. To prepare solution A:

- 360 mM k2HPO4 (mw 174.18): 62.7 g
- 132 mM KH2PO (mw 136.09): 17.96 g
- Fill up to 160 ml with H2O
- Sterilize by autoclaving

B. To prepare solution B:

- 17 mM Na citrate (mM 294.11): 4.99g
- 4 mM Mg SO4 (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 a horizontal laminar flow workstation.

### 1.5 Plasmid stock culturing

A. Prepare two 145/20 mm petri dishes and label them

B. Autoclave 50 ml of LB + 5 gm of agar and cool it down to 55 °C

C. Add Ampicillin at 1000 g/litre

D. Pour 50 ml of 1 per cent agar LB media (LM+Agar+Ampicillin)

E. Wait until the petri dishes cool down

F. Dilute 1.75 ul of library stock solution in 200 ul of LB + Ampicillin

G. Transfer 100 ul of diluted library stock solution on the 145/20 mm petri dishes

H. Spread the diluted library stock solutions using glass bids

I. Incubate bacteria upside down at 37 °c overnight (18 hrs)

### 1.6 Preparation of Master plate and individual colony picking

A. Mix 300 ml of LB media with 300 ul of Ampicillin

B. Mix 45 ml LB+Ampicillin with 5 ml of 10x freezing media (FM)

C. Add 100 ul of the LB+Ampicillin+FM using matrix pipette into every well of 384 well plate

D. Label the plate as POR_A followed by the number of plate prepared

E. Pick individual colonies and transfer them into using sterilized cocktail sticks

F. Incubate the master plates without shaking upside down at 37 °C overnight (18 hrs)

### 1.7 Making replica plates

A. Prepare two replica plates 'B' and 'C' for every individual master plate. Fill plates 'B' with LB+AMP+FM and plates 'C' with only LB+AMP.

B. Place a sterilized 384 pin replicator into the master plate (A) in order to make a copy into the plate B and again into the final plate

C. Rinse the 384-pin replicator every time by dipping in 1% bleach solution, distilled water and 100 % Ethanol

D. Place the device in flame to evaporate the ethanol

E. Cool down the device for making a new copy

F. Incubate both plates over night at 37 °C overnight (18hrs)

G. Cool down both A and B plates for about 30 minutes and finally store them at -80 °C

H. Use plate C for sequencing

### 1.8 Cell lysate preparation

A. Pellet cells by spinning plates 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. Re-suspend the pellet in 25 µl MQ water/well

J. Seal the plate with aluminum foil tape

K. Vortex the plates (table vortex)

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

M. Spin it down shortly

N. Place the plates into a thermo cycler for 5 minutes at 95°C to denature the cell lysates

O. store plates directly on ice

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

## 1.9 Sequencing reaction

A. Dilute the primer (T3 in this case is the forward primer): primer working solutions are 40 pmol/µl, so each PCR primer should be diluted ±1:50 to attain a concentration of 0.8 pmol/µl

B. Add 2 µl from the lysate to a new 384 well-plate and spin it down shortly

C. Prepare a master mix for sequencing reactions as follows ( for 410 samples)

| Reagent | Volume(µl / 1 sample | Volume(µl)/ 410 samples |
|---|---|---|
| 5X sequencing buffer | 0.75 | 307.5 |
| BD 3.1 (Big dye) | 0.5 | 205 |
| MQ | 0.75 | 307.5 |
| primer (0.8 pmol/µl) | 1 | 410 |
| **Total Volume** | **5** | **2050** |

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) on the Biometra 384-PCR machine

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

### 1.10 Precipitation of DNA samples

A. 5 µl sequencing reaction

B. Add 0.5 µl of NaAc-EDTA (1.5 M sodium acetate (pH > 8.0) and 250 mM EDTA)

C. Spin it down shortly

D. Add 17 µl of 70 % EtOH (-20 degrees) using 125 µl electronic pipette with 16 tips

E. Seal with aluminum (not heat sealing)

F. Mix by vortex (with Illumina vortex for 1 minute at 2000 rpm )

G. Incubate 30 minutes on ice box

H. Centrifuge 30 minutes 3000g at 4 °C

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

J. Add 10 µl of formamide to each well and dissolve pellet by pipetting 20X up and down

K. Transfer the solution into a barcoded plate

L. Seal the plates with sequencer devices and run the barcoded plates with samples in the ABI3730 DNA analyzer

### 1.11 Agarose gel-electrophoresis ( 1 per cent agarose gel preparation)

A. Add 3 grams of agarose powder

B. Add 300 ml of 0.5 % of TBE buffer

C. Dissolve the agarose by heating it in microwave

D. Cool down the dissolved agarose to 50 °c

E. Meanwhile, Prepare a Gel tray warped with tape and combs placed between

F. Add 15 µl of Etidium bromide (EtBr) i.e. 5 µl EtBr / 100 ml of TBE buffer

G. Mix thoroughly and pour into the gel tray with combs

H. Let the gel cool down

I. Meanwhile, prepare the DNA samples to be loaded into the gel as follows

| Sr.No | Ingredients | Amount per 1 well |
|---|---|---|
| 1 | PCR product | 2 µl |
| 2 | Loading dye | 2 µl |
| 3 | MQ water | 6 µl |

B. Add 10 µl of the above mix (PCR product, Loading dye and MQ water).

**C.** Add 3 µl of either 100 bp or 500 bp DNA markers or precision markers flanking the samples

### 1.12 Purification of PCR products

- (The PCR product should be cleaned from primers, to prevent the sequencing reaction to start at both ends)

A. Use Millipore PCR cleanup vacuum system (Multiscreen_ PCR vacu 030). Load the complete PCR products into the Multiscreen_PCR plate.

B. Place the Multiscreen_ PCR plate on top of the Vacuum manifold.

C. Apply vacuum at 24 inches Hg for 5 minutes or until the wells have emptied. Allow 30 extra seconds under vacuum after the well appears empty to be sure all liquid has been filtered. The filter appears shiny even after they are dry.

D. Load filter with 35 µl MQ water and apply the vacuum again for 5 minutes or until the wells have emptied.

E. Repeat procedure “D” once

F. After vacuum filtration is complete, remove the plate from the manifold, blot from underneath with paper towels and add 12 µl MQ water (equal to start volume of the PCR-reaction) to each well with a stepper-pipette.

G. Mix samples vigorously on a plate shaker for 5 minutes.

H. Retrieve purified PCR product from each well by pipetting and putting them in a new plate.

I. Check quality and quantity on agarose gel. For quantification, use Gene-ruler or EZ load precision as a marker. Load 1 µl on Agarose gel

J. The bands should be single, clean and without primer-dimer

## 1.13 Perform sequencing reaction of purified PCR product

A. Put a volume corresponding to 10-20 ng PCR product ( Below example is 1 µl of PCR product , maximum is 5.5 µl when no MQ is added) from each DNA sample in a Perkin Elmer 96-well PCR system plate

B. Dilute the primer (T7 and internal primers in this case):

NB: primer working solutions are around 40 pmol/µl, so each PCR primer should be diluted ± 1:50

C. Prepare the master mix for sequencing reactions

| Reagent | Volume(µl)/ 1 sample |
|---|---|
| PCR product | 1 |
| 5X sequencing buffer | 1.5 |
| BD 3.1 | 1 |
| MQ | 4.5 |
| Primer (0.8 pmol/µl) | 2 |
| **Total** | **10** |

D. add 9 µl of mix to each well in a new 96 well-plate PCR plate using a matrix pipette (in the example above 1 µl of PCR is added but it can vary according to the concentration of the product)

E. Heat seal the plate and mix by vortex

F. Spin all ingredients down shortly

G. Perform cycle sequencing as follows (Program: BD50x50. NB: Annealing temperature is correlated with the annealing temperature in the PCR reaction and may vary!):

## 1.14 Precipitation of DNA of sequencing reaction

A. 10 µl sequencing reaction

B. Add 1 µl NaAc-EDTA (1.5 M sodium acetate (pH > 8.0) and 250 mM EDTA)

C. Add 34 µl of 70 % EtOH (-20 degrees)

D. Mix by vortex (with Illumina vortex for 1 minute at 2000 rpm)

E. Incubate 30 minutes on ice

F. Centrifuge for 30 minutes 3000g at 4 °C

G. Centrifuge upside down for 1 minute 700g at 4 °C (NB: the plates cannot be frozen, then is better just to process one plate)

H. Add 10 µl of formamide into a barcode plate to each well and add 2 µl of sample.

**I.** Seal and run the barcode plate with samples in the ABI3730

## 2. COMPUTATIONAL PROTOCOLS

### 2.1 Convert the .abi files from ABI3730 sequencer into FASTA files

A. Download the free version of ABI to FASTA converter software downloaded from (http://www.dnabaser.com/download/Abi-to-Fasta-converter/abi-to-fasta-converter.html)

B. Open the program and go to settings. Deselect all the trim options in order to get the complete files

C. Go to open files and select the folder containing the .ab1 files

D. Select all the sequences to convert.

E. Press the CONVERT button

**F.** Open the converted files and check if it has been changed accordingly

### 2.2 Making Multi FASTA file with selected sequences ( in this case sequences from each dish)

A. Download the free version of DNA Baser V3.5.0 from (http://www.dnabaser.com/help/tools-converters/MultiFASTA%20Builder/index.html )

B. In the DNA BASER, start the MultiFasta Builder tool from the 'Tools -> MultiFasta Builder' menu.

C. Locate the folder that contains individual FASTA files ( in this case from each sequenced plates)

D. Select all individual FASTA files

E. Choose a name for the output file (the default name is "Result 1")

F. Deselect the option " Add empty line"

G. Press the 'Start' button.

H. check roughly if the multifasta file is the correct format and name

### 2.3 Download the databases

A. Go to the ensembl web site http://www.ensembl.org/info/data/ftp/index.html

B. download the appropriate databases. NB: For this project the databases downloaded were DNA and cDNA FASTA from *Sus Scrofa ( build 10.2)*

**C.** Select the location and save them. NB: the databases must be in the same folder with the downloaded BLAST files.

### 2.4 Make the databases BLASTable

A. Go to start and select Run

B. Type cmd and click OK

C. The window for DOS is opened

D. Type cd and the address of the folder in which you have all the files you already downloaded (BLAST, databases) and the .FASTA files with the query sequences

*E.* Type the name of the database (db) using the following commands in the same folder mention above: makeblastdb -in *the database.FA* -dbtype *nucl* -parse_seqids -out *your_blastable_ db*

F. Press enter

(NB: after pressing enter six different files will be created. This confirms that the *blastable* database was created)

### 2.5 Downloading BLAST Application

A. Go to the NCBI web site ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ and download the latest version for windows (win.64).NB: For this project the BLAST version blast-2.2.26+ for windows was used.

B. Select the install location and click install and accept the conditions.

### 2.6 Running Blastn

A. Go to start and select Run

B. Type cmd and click OK

C. The window for DOS is opened

D. Type cd and the address of the folder in which you have all the files you already downloaded (BLAST, databases) and the .FASTA files with the query sequences

E. Type the name of the database (db) and your .FASTA file using the following commands in the same folder mention above

F. In the address mention above run blastn typing the following commands: **blastn –query *your_combined_fasta_files.FASTA* -db *your_db* -outfmt 6 -out *your_query.blast* –evalue *1e-10***

H. For the –db option the name of the database should be the name of the file generated while making blastable database

## 2.7 Primer3: Internal primer development

A. Go the primer3 web site and run  the latest version of primer3Plus (http://primer3.wi.mit.edu/)

B. Upload or Paste your sequence into the  webpage

C. Name your sequence

D. Use the default setting

**E.** Pick your primers and check for the complementarity

## 3. CLONING VECTOR INFORMATION

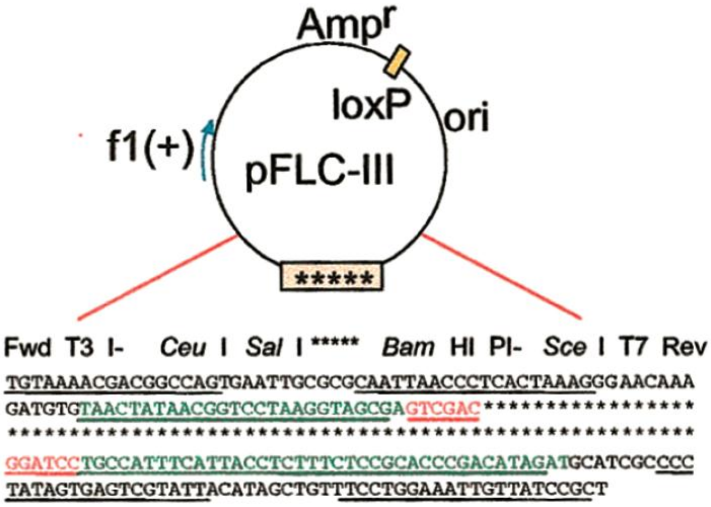

Amp^r
loxP
ori
f1(+)
pFLC-III
*****
Fwd T3 I- Ceu I Sal I ***** Bam HI PI- Sce I T7 Rev
TGTAAAACGACGGCCAGTGAATTGCGCGCAATTAACCCTCACTAAAGGGAACAAA
GATGTGTAACTATAACGGTCCTAAGGTAGCGAGTCGAC*****************
*******************************************************
GGATCCTGCCATTTCATTACCTCTTTCTCCGCACCCGACATAGATGCATCGCCCC
TATAGTGAGTCGTATTACATAGCTGTTTCCTGGAAATTGTTATCCGCT

Figure 1. Plasmid sequence after excision. The sequence of elements as they appear, abbreviated under the plasmid schematic structure (primer sequences, RNA polymerases, promoters, restriction sites and recombination sites), is underlined. (Carninci *et al.*, 2001)

pFLCIII-cDNA sequence

```
CTAAATTGTAAGCGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCAGCTC
ATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCTTATAAATCAAAAGAATAGACCGA
GATAGGGTTGAGTGTTGTTCCAGTTTGGAACAAGAGTCCACTATTAAAGAACGTGGACTC
CAACGTCAAAGGGCGAAAAACCGTCTATCAGGGCGATGGCCCACTACGTGAACCATCACC
CTAATCAAGTTTTTTGGGGTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAG
CCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAA
AGCGAAAGGAGCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAACCAC
CACACCCGCCGCGCTTAATGCGCCGCTACAGGGCGCGTCCCATTCGCCATTCAGGCTGCG
CAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGG
GGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTG
TAAAACGACGGCCAGTGAATTGCGCGCAATTAACCCTCACTAAAGGGAACAAAGATGTGT
AACTATAACGGTCCTAAGGTAGCGAGTCGAGGTCGAGCTCTATTTAGGTGACACTATAGA
ACCA********************************************************
************************************************************
************************************AAAAAAAAAAAAAAAACTCTTGTT
GGATCCTGCCATTTCATTACCTCTTTCTCCGCACCCGACATAGATGCATCGCCCCTATAG
TGAGTCGTATTACATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACA
ACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCA
CATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGC
ATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTT
CCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACT
CAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAG
CAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATA
GGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACC
CGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTG
TTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGC
TTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGG
GCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTC
TTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGA
TTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACG
GCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAA
```

AAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTG
TTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTT
CTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGAT
TATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTT**ATAACTTCGT**
**ATAGCATACATTATACGAAGTTAT**AAATCAATCTAAAGTATATATGAGTAAACTTGGTCT
GACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCA
TCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCT
GGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCA
ATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCC
ATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTG
CGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCT
TCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAA
AAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTA
TCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGC
TTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCG
AGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAA
GTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTG
AGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTC
ACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGG
GCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTAT
CAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATA
GGGGTTCCGCGCACATTTCCCCGAAAAGTGCCAC

Cloning site: *Xho* I/*Sal* I and *Bam*H I;

**LoxP** is inserted between Amp$^{r}$ and Ori.

Outside of Forward and Reverse primer sequences in the vector are same as pBluescript except a LoxP site.

## 4. FIGURES

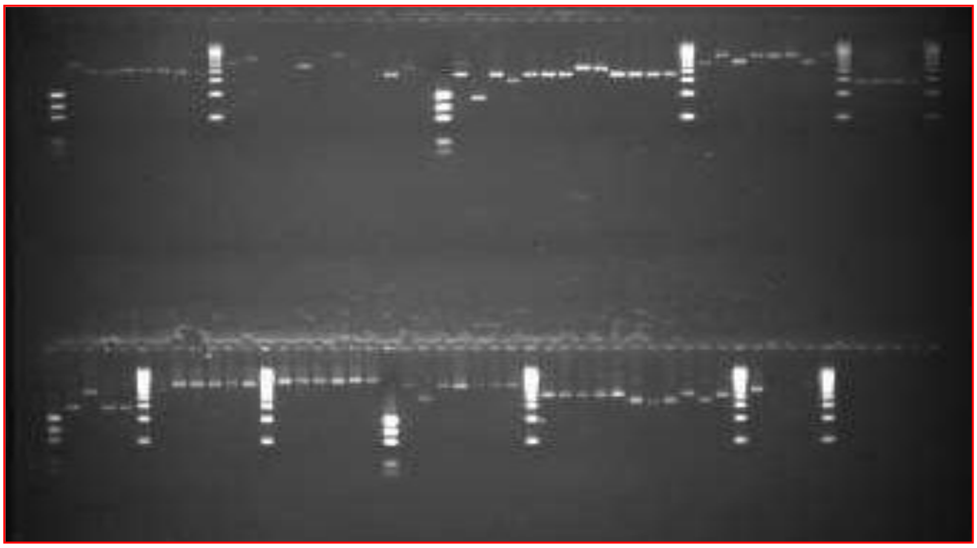


*Figure 1: fragment size variation among 108 clones blasted against 10 different genes. Each gene is indicated between precision markers*

## 5. TABLES

Table s1. Insert size obtained from 48 tested clones

| **Number of clones** | **Size (bp)** |
|---|---|
| 5 | 500 |
| 6 | 1000 |
| 7 | 1500 |
| 9 | 2000 |
| 7 | 2500 |
| 3 | 3000 |
| 4 | 3500 |
| 0 | 4000 |
| 1 | 4500 |
| 3 | 5000 |
| 1 | 5500 |
| 1 | 6000 |
| **Mean** | **2000** |

Table s2: Sequencing summery of Experiment 1

| Sr.No | **Plates** | *Total number of sequences* | *Length of sequences (bp)* | Min | Max | 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 |
| **Total** | | **5469** | **1993634** | **51** | 750 | **365** | **71.21** |

Table s3: *Sequencing summery of Experiment 2*

| **SN** | **Plates** | *Total number of sequences* | *Length of sequences (bp)* | *Min* | *Max* | *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 | 59.9 |
| 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 | 87.5 |
| 21 | POR-C044 | 298 | 88131 | 59 | 716 | 296 | 77.6 |
| 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** |

Table s4. List of primers and primer chemistry

| A. Universal T3 forward and T7 reverse primers | | | | |
|---|---|---|---|---|
| Primer | Sequence | Bases | TM | % GC |
| Universal T3 forward | ATTAACCCTCACTAAAGGGA | 20 | 56 | 40 |
| Universal T7 forward | GTAATACGACTCACTATAGGG | 21 | 45 | 42.9 |
| **B. Second round sequencing primers** | | | | |
| Internal Primers | Sequence | Bases | TM | % GC |
| A, C, E, F, G, H, J, K, L, M | ATTACATCGCCCTGAACGAG | 20 | 60.10 | 50 |
| I | CCCGGTTTCGTTTTCAGTT | 19 | 59.96 | 47.37 |
| D | GGACAAGTCCCGTGCTCAT | 19 | 61.10 | 57.89 |
| B | CTGGACACCATCTCCATCCT | 20 | 59.92 | 55 |
| **C. Third round sequencing primer** | | | | |
| Internal Primers | Sequence | Bases | TM | % GC |
| A, C, E, F, G, J, K, L, M | CTCCGATGTGTCCCTTACCA | 20 | 60.91 | 55 |
| H | TCAGAGCCTCCAAAGACACA | 20 | 59.55 | 50 |
| I | CCTGGAGAGGAGCAGAGCTA | 20 | 59.84 | 60 |
| D | ATGTCCAAGCCACTTTCCTG | 20 | 60.11 | 50 |

Table s5: List if clone sequences selected for transcript size estimation

| Clone name | Gene Name | Gene description |
|---|---|---|
| POR_C026_J04 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 2 |
| POR_C037_J05 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 2 |
| POR_C037_P07 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 2 |
| POR_C040_C13 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 2 |
| POR_C040_D13 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 2 |
| POR_C040_E13 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 2 |
| POR_C049_G01 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 3 |
| POR_C052_M24 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 4 |
| POR_C065_G03 | CAPZA2 | capping protein (actin filament) muscle Z-line, alpha 5 |
| POR_B007_B12 | CPE | carboxypeptidase E |
| POR_B010_B07 | CPE | carboxypeptidase E |
| POR_B011_M23 | CPE | carboxypeptidase E |
| POR_B011_N23 | CPE | carboxypeptidase E |
| POR_B011_O23 | CPE | carboxypeptidase E |
| POR_C031_M21 | CPE | carboxypeptidase E |
| POR_C033_G08 | CPE | carboxypeptidase E |
| POR_C034_G18 | CPE | carboxypeptidase E |
| POR_C034_J17 | CPE | carboxypeptidase E |
| POR_C036_D10 | CPE | carboxypeptidase E |
| POR_C038_P24 | CPE | carboxypeptidase E |
| POR_C043_C12 | CPE | carboxypeptidase E |
| POR_C043_P09 | CPE | carboxypeptidase E |

| | | |
|---|---|---|
| POR_C044_N06 | CPE | carboxypeptidase E |
| POR_C050_E16 | CPE | carboxypeptidase E |
| POR_C060_D12 | CPE | carboxypeptidase E |
| POR_C067_M07 | CPE | carboxypeptidase E |
| POR_C069_O22 | CPE | carboxypeptidase E |
| POR_C073_D10 | CPE | carboxypeptidase E |
| POR_B005_G16 | SLA-3 | MHC class 1 anti gene family |
| POR_B011_O02 | SLA-3 | MHC class 1 anti gene family |
| POR_C027_H09 | SLA-3 | MHC class 1 anti gene family |
| POR_C039_G07 | SLA-3 | MHC class 1 anti gene family |
| POR_C043_N11 | SLA-3 | MHC class 1 anti gene family |
| POR_C047_D11 | SLA-3 | MHC class 1 anti gene family |
| POR_C049_K20 | SLA-3 | MHC class 1 anti gene family |
| POR_C050_H02 | SLA-3 | MHC class 1 anti gene family |
| POR_C054_O20 | SLA-3 | MHC class 1 anti gene family |
| POR_C055_P14 | SLA-3 | MHC class 1 anti gene family |
| POR_C058_G07 | SLA-3 | MHC class 1 anti gene family |
| POR_C059_G04 | SLA-3 | MHC class 1 anti-gene family |
| POR_C063_D24 | SLA-3 | MHC class 1 anti gene family |
| POR_C066_B10 | SLA-3 | MHC class 1 anti gene family |
| POR_B006_H03 | SFXN1 | sideroflexin 1 |
| POR_C026_J11 | SFXN1 | sideroflexin 2 |
| POR_C040_L24 | SFXN1 | sideroflexin 3 |
| POR_C052_M09 | SFXN1 | sideroflexin 4 |
| POR_C058_H07 | SFXN1 | sideroflexin 5 |
| POR_C061_C24 | SFXN1 | sideroflexin 6 |
| POR_C070_C15 | SFXN1 | sideroflexin 7 |
| POR_C073_A16 | SFXN1 | sideroflexin 8 |
| POR_C029_O12 | OAZ1 | ornithine decarboxylase antizyme 1 |
| POR_C029_P11 | OAZ1 | ornithine decarboxylase antizyme 2 |
| POR_C033_B14 | OAZ1 | ornithine decarboxylase antizyme 3 |
| POR_C040_K01 | OAZ1 | ornithine decarboxylase antizyme 4 |
| POR_C040_M01 | OAZ1 | ornithine decarboxylase antizyme 5 |
| POR_C045_F08 | OAZ1 | ornithine decarboxylase antizyme 6 |
| POR_C056_M10 | OAZ1 | ornithine decarboxylase antizyme 7 |
| POR_C060_G05 | OAZ1 | ornithine decarboxylase antizyme 8 |
| POR_C073_O15 | OAZ1 | ornithine decarboxylase antizyme 9 |
| POR_B009_N16 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |
| POR_C029_B09 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |
| POR_C029_C17 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |

| | | |
|---|---|---|
| POR_C031_I18 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |
| POR_C056_J20 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |
| POR_C063_B11 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |
| POR_C071_K01 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |
| POR_C072_C18 | PRKAR2A | protein kinase, cAMP-dependent, regulatory, type II, alpha |
| POR_B005_L10 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_B007_A22 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C026_C23 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C031_C19 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C039_E22 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C044_P14 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C058_D11 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C061_N08 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C064_D01 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_C065_H18 | ELAVL1 | ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 |
| POR_B025_I11 | C17orf75 | protein Njmu-R1 |
| POR_C027_D05 | C17orf75 | protein Njmu-R2 |
| POR_C029_C08 | C17orf75 | protein Njmu-R3 |
| POR_C033_A06 | C17orf75 | protein Njmu-R4 |
| POR_C035_E12 | C17orf75 | protein Njmu-R5 |
| POR_C056_P09 | C17orf75 | protein Njmu-R6 |
| POR_C026_N14 | PSAP | prosaposin |
| POR_C032_M06 | PSAP | prosaposin |
| POR_C037_L11 | PSAP | prosaposin |
| POR_C037_N10 | PSAP | prosaposin |
| POR_C037_O08 | PSAP | prosaposin |
| POR_C040_F14 | PSAP | prosaposin |
| POR_C041_I07 | PSAP | prosaposin |
| POR_C052_F15 | PSAP | prosaposin |
| POR_C056_A13 | PSAP | prosaposin |
| POR_C060_G14 | PSAP | prosaposin |
| POR_C061_C17 | PSAP | prosaposin |
| POR_C062_C06 | PSAP | prosaposin |
| POR_C071_F04 | PSAP | prosaposin |
| POR_C073_L17 | PSAP | prosaposin |
| POR_B002_F19 | YWHAQ | tyrosine 3-monooxygenase |
| POR_B011_L02 | YWHAQ | tyrosine 3-monooxygenase |
| POR_B024_P04 | YWHAQ | tyrosine 3-monooxygenase |
| POR_C027_H16 | YWHAQ | tyrosine 3-monooxygenase |
| POR_C032_A14 | YWHAQ | tyrosine 3-monooxygenase |

|  |  |  |
|---|---|---|
| POR_C033_H01 | YWHAQ | tyrosine 3-monooxygenase |
| POR_C058_A06 | YWHAQ | tyrosine 3-monooxygenase |
| POR_C059_G18 | YWHAQ | tyrosine 3-monooxygenase |
| POR_C067_F22 | YWHAQ | tyrosine 3-monooxygenase |
| POR_C067_H10 | YWHAQ | tyrosine 3-monooxygenase |
| POR_C071_O06 | YWHAQ | tyrosine 3-monooxygenase |