Analysis of ovine uterus Expressed Sequence Tags (ESTs)

Suhaimi, A.H.M.S [1*], I.A. Azlina Azma[2], A.R. Roziatul Erin[1], W. M. Z. Somarny[1], M. N. Mat Isa[3], Z. A. Rabiatul Adawiah[4] , S. Md Tasol[1] and K. Musaddin[1]

[1]Strategic Livestock Research Centre,
MARDI Headquarters, 43400 Serdang, Selangor.
[2] MARDI Research Station Muadzam Shah, Muadzam Shah, Pahang.
[3]Malaysian Genome Institute, Bangi, Selangor.
[45] Biotechnology Research Centre, MARDI Headquarters,
43400 Serdang, Selangor
*Corresponding author: shusna@ mardi.gov.my

Abstract

The aim of this paper was to explore the different analyses available to examine Expressed Sequence Tags (ESTs). ESTs were generated from the uterus of indigenous sheep, Malin. A total of 675 unigenes were identified of which 603 were singletons and 72 were consensus sequences. Seventy nine of the unigenes had significant match to genes in SwissProt database while the remaining unigenes had to be converted to coding regions (CDS). Fifty seven of the CDS had significant matches to the SwissProt database. Upon functional annotation to both unigenes and CDS, binding and catalytic activity were the two most common molecular function while the most common biological processes were related to cellular processes. Enzyme commission pathway related to DNA, purine and primidine metabolism pathway had the most abundant number of genes Subjected to KEGG pathway mapping. Pfam protein family related to electron transport process towards the production of ATP family, was additionally identified among the ESTs. Analysis of ESTs with Pfam revealed the importance of ATP generation in the uterus. Albeit minimally, the different analyses performed in this project provide some light in the function of ovine uterus and can be applied to other indigenous species that have no or limited genome data availability to discover novel and unknown genes and proteins.

Keywords: Expressed Sequence TGS (ESTs), ovine, cDNA library, bioinformatics analyses

Introduction

Expressed sequence tags (ESTs) are short DNA sequences of 200-800 nucleotide bases in length corresponding to gene transcripts of specific tissues. ESTs have many applications including gene discovery, complementing genome annotation, aiding gene structure identification and discovery of single nucleotide polymorphisms (SNP) characterization. ESTs are obtained by sequencing a complementary DNA (cDNA) library in which fragments of the tissue of interest have been cloned. ESTs projects will yield a large volume of information and thus, is a source of valuable knowledge. In animal research, ESTs have been utilized to understand expression mechanisms in domestic animals such as cattle (Huang et al 2012) and pig (Chen et al. 2006) as well as ticks and parasites (Kanduma et al., 2012). AS of 1 July 2012, there are at least 73,360,923 ESTs submissions in the ESTs database at NCBI website (www.ncbi.nlm.nih.gov/dbEST/) of which 338,483 are of ovine ESTs. Analyzing

information obtained in a ESTs project can be an enormous task due to its large amount of data. Bioinformatics approach is commonly used to obtain valuable information from ESTs such as known gene, hypothetical and putative protein function, clustering, pathway analysis and species specificity.

In this study, ESTs is generated from the uterus of Malaysian indigenous sheep, Malin which is reputable for its high prolificacy and is known to give birth to high frequency of twins. ESTs are generated to gain insights on functional expression of genes and proteins in ovine uterus for further research. The information obtained will enable a better understanding of our local indigenous species and can then be used in a breeding program to improve the production traits of these animals.

Materials and Methods

Tissue Collection

Uterus tissues were obtained from a female Malin sheep with a known history of high twinning rate. Samples were sliced to pieces of about 0.1g. Tissues were quick-frozen in liquid nitrogen and stored at -80^{O}C until further use.

Total RNA and mRNA Isolation

The total RNA was isolated using Qiagen RNeasy Midi kit (Qiagen, Germany) according to manufacturer's protocol. The isolated total RNA (200µg/µl) was then cleaned using Qiagen RNeasy Clean-up (Qiagen, Germany) procedure. About 200 µg of cleaned total RNA was used to isolate it's poly A^{+} mRNA using Miltenyi MACS mRNA Isolation Kit (Miltenyi Biotech, Germany).

cDNA Library Construction

A cDNA library constructed from 6 ug of purified poly A+ mRNA using cDNA Synthesis Kit, ZAP-cDNA Synthesis Kit ZAP-cDNA Gigapack III Gold Cloning kit (Stratagene, USA). First and second-strand cDNA was synthesized and ligated into pBluescript SK (+/-) vector (Stratagene), packaged *in vitro* and amplified according to the manufacturer's protocol with minor modifications.

Mass- Excision of Phage Library

The phage clones were excised to plasmid form in *E. coli* strain *SOLR* using mass excision protocol according to the procedure described by Stratagene, (USA). Clones were plated at low density onto Luria-Bertani agar containing ampicillin (50µg/ml) at 37^{O}C overnight. The colonies were then selected randomly for DNA plasmid isolation.

Sequencing of cDNA Llibrary Clones

Each colony was then grown in a deep well plate containing 1.5 ml LB broth with 50 µg/ml ampicillin and grown at 200rpm, 37^{O}C overnight The glycerol stock (15% w/v) was then prepared and 200µl of glycerol stock was sent to The Malaysian Genome Institute for plasmid extraction and sequencing using SK primers. Plasmids were extracted using Montage Plasmid Miniprep Kit (Milipore, USA) and sequenced using ABI Automated Sequencer.

ESTs Processing, Assembly and Functional Annotation

An analysis pipeline was created to process the raw data of ovine uterus ESTs. Various Bioinformatics tools were utilized

to perform the ESTs analysis of ovine uterus. Phred software (Green and Ewing, 2002) was used for base-calling. The ESTs were further trimmed from vector sequences, adaptors and low quality bases using Lucy software (Chou and Holmes, 2001). The high quality ESTs sequences were then assembled using StackPack 2.2. The functional annotation of all the unigenes were carried out using BLASTX against the SwissProt database. The ESTs sequences with no significant matches to SwissProt database were translated into protein sequences using ESTscan and the translated proteins were then annotated to gene ontology terms using BLAST2GO. The gene ontology analysis was performed to understand the functional classification of known, putative and hypothetical genes in ovine uterus.

Results and Discussion

Generation of Expressed Sequence Tags

Complementary DNA (cDNA) libraries were constructed to reduce the number of genomic DNA sequences to only expressible genes. This would maximize the chances of finding new genes through random sequencing (Soares *et al.*, 1994). Initial data cleaning of ovine ESTs raw data which included vector trimming, removal of adapters and primers of was performed as a first step. This was the crucial stage to obtain the high quality sequences. A total of 746 raw sequences were obtained.

Expressed Sequence Tags Data Processing and Assembly

The assembly of 746 raw data produced 675 unigenes comprising of 603 sequences of singletones and 72 sequences of consensus. The assembled unigenes had an average length of 703 bp. The large number of singletons demonstrated the high complexity of the tissue's gene expression with many low copy number clones present. This would suggest that sequencing of additional uterus ESTs would identify other novel sequences. According to Wolfsberg and Landsman (1997) the singletons might also be caused by alternative splicing, whereby certain genes were derived from spliced or partially spliced transcripts either containing intron sequences or were spliced at previously unreported sites.

Functional Annotation of Expressed Sequence Tags

Similarity searching analysis was performed on the 675 unigenes using BLASTX against SwissProt database. BLASTX is a similarity searching tool to identify similar regions between an unknown nucleotide sequence and sequences from the protein database (Altschul *et al*., 1997). A $1x10^{-06}$ cut off e-value was used to identify potential candidate genes related to ovine uterus. SwissProt database was utilized for this purpose as the annotation was established, manually curated and highly cross referenced to other databases which would help to choose potentially abundant and functional genes in livestock (http://www.uniprot.org/). The analysis indicated that 79 unigenes in ovine ESTs had significant matches to SwissProt database (Table 1). The lowest similarity percentage obtained was 40%. The most represented ovine unigenes were retrovirus-related Pol polyprotein LINE-1 (POL2) with 26 redundancies followed by LINE-1 reverse transcriptase homolog (LINE1) with 12 redundancies (Table 2). All unigenes are listed in Appendix 1.

Since a large number of unigenes did not show any significant matches to the SwissProt database, the unigenes were

translated into protein coding sequences (CDS) using ESTcan. ESTcan is a program that can detect coding region in DNA/RNA sequences even if they are low quality sequences. Subsequently, unknown gene or hypothetical protein that may be expressed in ovine uterus can be discovered. Among the 596 unigenes that did not retrieve any BLASTX result, 121 unigenes had been successfully translated to CDS (Table 1). The average length of CDS was 198bp. The CDS were then subjected to GO annotations using BLAST2GO to identify potential genes. From this dataset, 57 CDS had significant matches with SwissProt database. All CDS are listed in Appendix 2. This low number was probably due to the lack of established ovine information in SwissProt database. The combined methods described above were able to identify more expressed genes that were enriched in cDNA library utilized.

Table 1. Functional annotation of ovine uterus unigenes

Unigenes	No. of unigenes
Unigenes annotated to SwissProt database	675
Unigenes with significant matches to SwissProt database	79
Unigenes without significant matches to SwissProt database	596
Unigenes translated to coding sequence (CDS)	121
Unigenes untranslated to CDS	475

Table 2. The highest represented unigenes and protein coding sequences (CDS) in ovine uterus

Accesion number	Protein name	No. of redundant unigenes
UNIGENES		
P11369.2	Retrovirus-related Pol polyprotein LINE-1 (POL2)	26
P08548.1	LINE-1 reverse transcriptase homolog (LINE1)	12
Q588U8	craniofacial development protein 2	6
P31625	bifunctional protease dutpase	2
CDS		
CAA10770.1	Reverse transcriptase-like protein	17
ADY76802.1	PP287	6
NP_001040099.1	Bitter taste receptor Bota-T2R65A	5
AAI26683.1	Catenin (cadherin-associated protein), alpha 3	3

Gene Ontology of Expressed Sequence Tags

The 79 unigenes were subjected to gene ontology (GO) analysis which will predict possible functions. The unigenes were assigned to GO Slims, the highest GO term level, providing a broad overview of the GO content. GO Slims are divided into biological processes, cellular component and molecular function categories and has a dynamic, controlled vocabulary and hierarchical relationship for the representation of those three categories (Conesa *et al*., 2005). At least one or more GO terms had been assigned to the 79 ovine unigenes. Based on the GO analysis, 37 unigenes had putative gene functions assigned to biological process, 22 to molecular function and 19 to cellular component. The most abundant GO Biological Process was cellular processes (GO: 0009987) and metabolic processes (GO: 0008512) while the most abundant molecular function was catalytic activity (GO: 0003824) and binding (GO: 0005488). The most abundant cellular component GO was cell (GO: 0005623) (Figure 1).

With regard to the CDS, a total of 57 CDS were then annotated with Gene Ontology (GO) terms. At least one or more GO terms had been assigned to the 57 ovine CDS. Based on the GO analysis, 23 CDS had putative gene functions assigned to biological process, 12 to molecular function and 10 to cellular component. The most abundant ontology terms in biological process were biological regulation (GO: 0065007) and cellular process (GO: 0009987). In molecular function, binding (GO: 0005488) and catalytic activity (GO: 0003824) were the most represented terms. Regarding cellular component, cell (GO: 0005623) and organelle (GO: 0031090) were the most abundant terms (Figure 2). A common feature of the functional annotation between the unigenes and CDS was catalytic activity which was the common molecular function. Wherelse, the common biological function among the unigenes and the CDS was related to cellular and metabolic processes. Annotations to similar biological and molecular functions for both unigenes and CDS confirmed the same origin of the two sets of data and analyzing the two sets of data together enabled more annotation information to be obtained. Highly represented categories indicated the occurrence of genes related to the categorized functions in the tissue studied. Thus, binding and catalytic activity as well as cellular and metabolic processes could potentially be important in the uterus and could play a major role in uterus function. Further analysis will be conducted to further characterize the function of known and hypothetical genes that were identified in this study.

Pathway Analysis of Expressed Sequence Tags

The unigenes and CDS were combined and subjected to KEGG pathway mapping to identify biological pathways that were prevalent to this set of data. KEGG pathway mapping was based on enzyme commission (EC) numbers, a system of enzyme nomenclature which was a numerical classification scheme for enzymes based on the chemical reactions they catalyze. KEGG pathway mapping based on EC numbers is an alternative approach to categorize gene functions with the emphasis on biochemical pathways. In total, 16 biochemical pathway based on the EC numbers obtained from GO annotations were identified. Among the pathways, EC number EC: 2.7.7.0 which corresponded to DNA, purine and primidine metabolism had the highest hit with 12 unigenes and CDS associated with it.

Protein Domain Family Identification in Sequences

The ESTs sequences were also further analyzed against Pfam, a comprehensive collection of protein domains and families (http://www. sanger.ac.uk/ resources/ databases/pfam.html). The current release of Pfam (22.0) contained 9318 protein families (Finn *et al.*, 2008). Alignment to the Pfam databases is a good approach as it can predict the function of gene transcripts represented by ESTs in species whose genomic information is still scarce. This approach can be utilized in other high-scale projects in various animal species to identify gene functions that may be involved in causing diseases or other physiological disorders.

About 161 sequences were aligned to known protein families or domains. In total, 16 protein families were identified with maximum of 12 hits and minimum of 1 hit (Figure 3). The three highest Pfam protein family identified were NADH-Ubiquinone oxidoreductase (complex I), chain 5 C-terminus (n=12), ATP synthase A chain (n=11) and Cytochrome C oxidase subunit II, transmembrane domain (n=9). The Pfam families identified above were involved in the electron transport process towards the production of ATP. Analysis of ESTs with Pfam revealed the importance of ATP generation in the uterus. This could be due to the role of uterus in implantation but further studies are needed to confirm this notion.

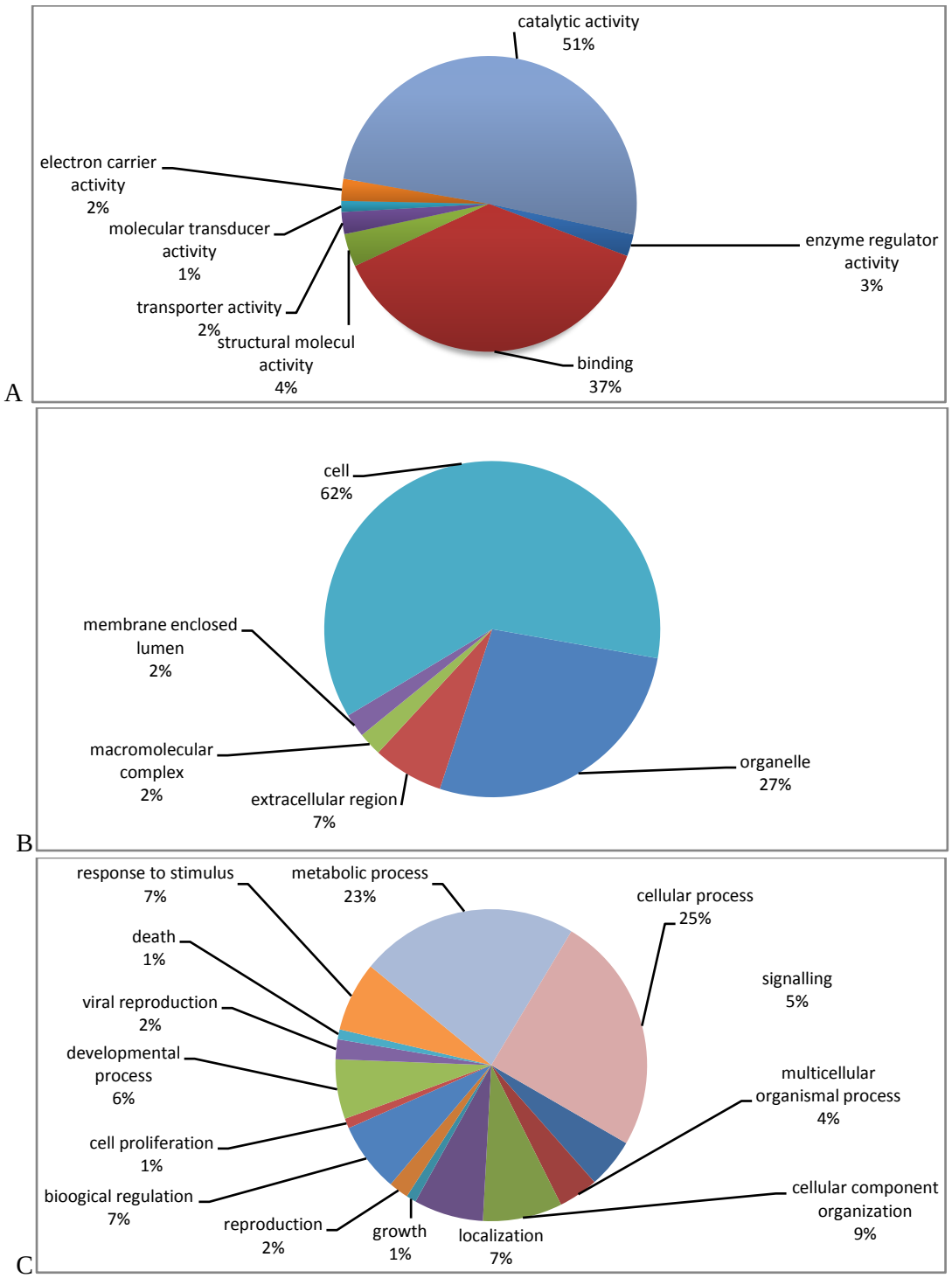

Figure 1: Gene ontology (GO) terms highly represented in unigenes generated from ovine uterus. A: Biological Process B: Cellular Component C: Molecular Function

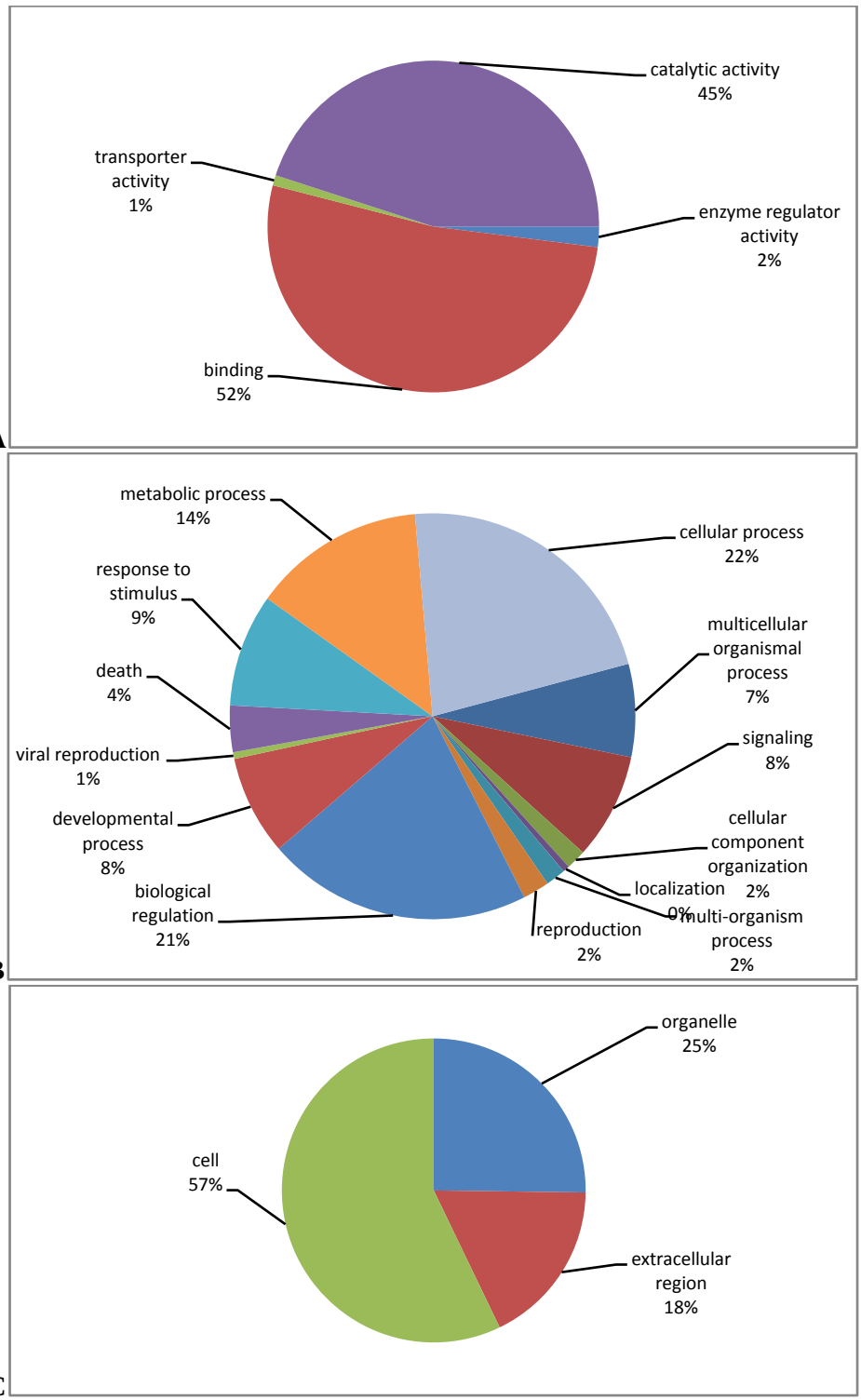

Figure 2: Gene ontology (GO) terms highly represented in Coding Sequences (CDS) generated from ovine uterus. A: Biological Process B: Molecular Function C: Cellular Component

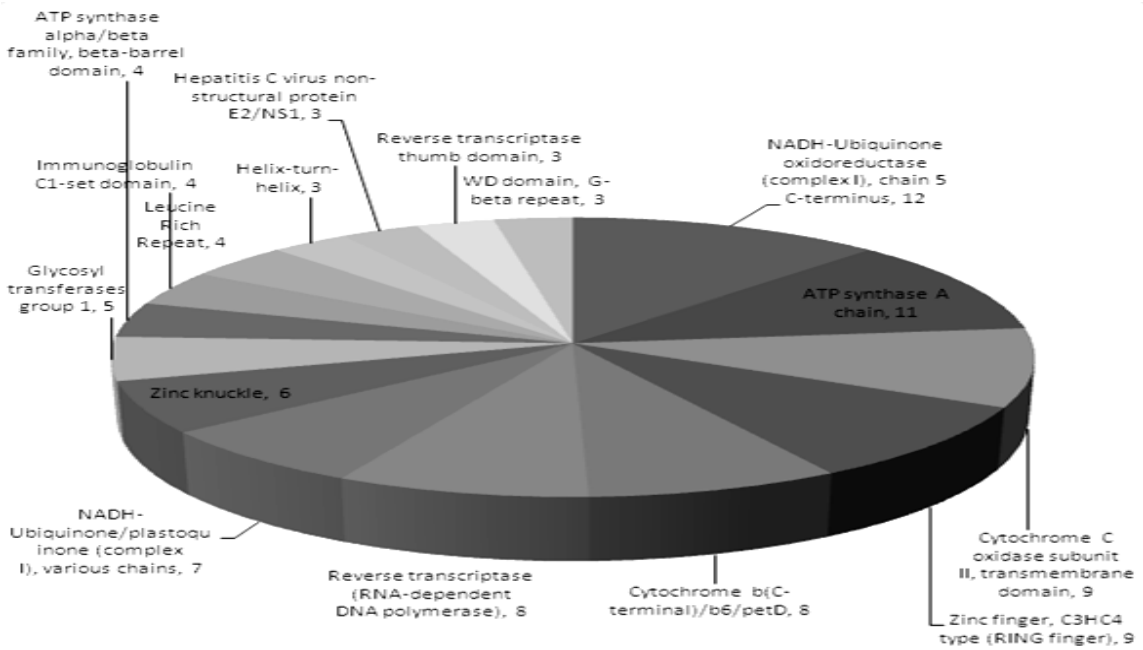

Figure 3: Protein domains and families identified in Malin ovine ESTs

Conclusion

Several genes, enzymes, biological pathways and protein domains that could be important in the function of sheep uterus were identified from ovine uterine ESTs sequences. This discovery was made using several bioinformatics approaches such as functional annotation and clustering, pathway analysis and protein domain identification. The number of ESTs generated in this project is fairly small, thus limiting the amount of information revealed from data analysis using bioinformatics. However, as not much is known regarding the local Malin sheep, any information is important to further understand our indigenous treasures. The analyses performed in this study can also be applied to other indigenous species that have no or limited genome data availability to discover novel and unknown genes and proteins in these species.

Acknowledgement

This project was funded by Ministry of Science, Technology and Innovation IRPA Grant 01-03-03-006 EA001. The authors also wish to thank staff of MARDI Strategic Livestock Research Centre's Animal Molecular Biology Laboratory, Muadzam Shah Research Station and Malaysian Genome Institute. Appreciation also goes to Dr. Tan Siang Hee, Lim Kean Jin, Habsah Bidin, Razean Haireen Mohd Razali, Umikalsum Mohd Bahari and other individuals who had contributed directly or indirectly to this paper.

References

Chen, C.H., Lin, E.C., Cheng, W.T., Sun, H.S., Mersmann, H.J. and Ding, S.T. 2006. Abundantly expressed genes in pig adipose tissue: an expressed sequence tag approach. J. Anim. Sci. 84(10): 2673-83.

Chou, H. and Holmes, M.H. 2001. DNA sequence quality trimming and vector removal. Bioinformatics. 17: 1093-1104.

Esther, G., Kanduma, E.G., Joram, M, Mwacharo, J.M., Jack,D., Sunter, J.D., Inosters, I., Nzuki, Mwaura, S.S., Peter, W., Kinyanjui, P.W., Kibe, M.M., Heyne, H.H., Hanotte, O.O., Skilton, R.R.A. and Bishop, R.P.R.P. 2012. Micro- and minisatellite-expressed sequence tag (EST) markers discriminate between populations of Rhipicephalus appendiculatus. Ticks and Tick-borne Diseases. 3(3): 128-136.

Finn, R. D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L. and Bateman, A. 2008. The Pfam protein families database. Nucleic Acids Research. 36: 281-288.

Green, P. and Ewing, B. 2002. *Phred*, version 0.020425c. http://phrap.org

Huang, W., Long, N. and Khatib, H. 2012. Genome-wide identification and initial characterization of bovine long non-coding RNAs from EST data. Animal Genetics. 43(6): 674-682.

Lu, Y., Huggins,P and Bar-Joseph, Z. 2009. Cross species analysis of microarray expression data. 25(12): 1476–1483.

Nagaraj, S. H. and Gasser, R. B. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bionformatics 8(1): 6-21.

Soares, M.B., Bonaldo, M.D.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. USA 91: 9228–9232.

Wolfsberg, T.G. and Landsman, D. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res. 25: 1626–1632.

Appendix 1: Ovine Uterine Expressed Sequence Tags (ESTS) unigenes

Accession number	Name of genes	Redundancies
Q86XU0	zinc finger protein 677	1
P43319	uncharacterized fimbrial-like protein yrak flags precursor	1
P30205	antigen flags precursor	1
P31625	bifunctional protease dutpase includes ame	2
P12378	udp-glucose 6-dehydrogenase short	1
Q3SYU7	transportin-1 ame	1
Q96Q15	serine threonine-protein kinase smg1 short	1
Q53H47	histone-lysine n-methyltransferase setmar ame	1
Q95SX7	probable rna-directed dna polymerase from transposon bs ame	1
Q9XSI3	60s ribosomal protein l10 ame	1
Q0T6C8	ribosomal protein s12 methylthiotransferase short	1
Q8N392	rho gtpase-activating protein 18 ame	1
O97764	zeta-crystallin	1
Q9TTC6	peptidyl-prolyl cis-trans isomerase a short	1
P31134	putrescine transport atp-binding protein	1
P11283	gag-pro-pol polyprotein contains ame	1
Q9TTC1	pro-pol polyprotein ame	1
P11369	retrovirus-related pol polyprotein line-1 ame	26
P10266	herv- provirus ancestral pol protein ame	1
O00443	phosphatidylinositol-4-phosphate 3-kinase c2 domain-containing subunit alpha short	1
Q0P5C2	bifunctional methylenetetrahydrofolate dehydrogenase mitochondrial includes ame	1
Q9BZ81	melanoma-associated antigen b5 ame	1
P08548	line-1 reverse transcriptase homolog	12
Q96RT1	protein lap2 ame	1
Q61768	kinesin-1 heavy chain ame	1
P58764	tyrosine-protein kinase etk	1
P0AEJ1	multidrug resistance protein b	1
P10443	dna polymerase iii subunit alpha	1
Q9UPY3	endoribonuclease dicer ame	1
P07014	succinate dehydrogenase iron-sulfur subunit	1
Q58DC0	calcineurin-like phosphoesterase domain-containing protein 1	1
O02751	craniofacial development protein 2 ame	6
Q5PPN4	carbonic anhydrase-related protein short	1
A3KQV2	bro1 domain-containing protein brox ame	1
Q9C0F0	polycomb group protein asxl3 ame	1
B7NG10	cation acetate symporter ame	1

Appendix 2: Ovine Uterine Expressed Sequence Tags (ESTS) Protein Coding Sequences (CDS)

Accesion number	Name of genes	Redundancies
NP_001193721.1	B7 homolog 6 precursor	1
NP_001040099.1	bitter taste receptor Bota-T2R65A	5
AAI26683.1	Catenin (cadherin-associated protein), alpha 3	3
ADI61825.1	endonuclease-reverse transcriptase	2
EAW91626.1	hCG2041411	1
ACH79982.1	hypothetical protein	1
ADY76802.1	PP287	6
ADY76801.1	PP288	1
XP_002764145.1	PREDICTED: hypothetical protein LOC100406107	1
XP_003584939.1	PREDICTED: Ig heavy chain V region MC101	1
XP_003581808.1	PREDICTED: LOW QUALITY PROTEIN: centrosomal protein of 70 kDa	1
XP_003279061.1	PREDICTED: LOW QUALITY PROTEIN: SH3 domain-binding protein 2-like	1
XP_002707826.2	PREDICTED: LOW QUALITY PROTEIN: SWI/SNF complex subunit SMARCC1	1
XP_003586095.1	PREDICTED: uncharacterized protein LOC100847767	1
XP_003586327.1	PREDICTED: uncharacterized protein LOC100848083	1
XP_003587971.1	PREDICTED: uncharacterized protein LOC100848498	1
XP_003587971.1	PREDICTED: uncharacterized protein LOC100848498	1
XP_003587971.1	PREDICTED: uncharacterized protein LOC100848498	1
XP_003587971.1	PREDICTED: uncharacterized protein LOC100848498	1
XP_003582233.1	PREDICTED: uncharacterized protein LOC100849819	1
CAA10770.1	reverse transcriptase-like	17
DAA13310.1	transmembrane 9 superfamily member 2	2
AAY53483.1	transposase	1
NP_001186599.1	zinc finger protein 33B	2