Expressed sequenced tags profiling of resistant and susceptible Gyr x Holstein cattle infested with the tick *Rhipicephalus* (*Boophilus*) *microplus*

C.S. Nascimento[1], M.A. Machado[2], S.E.F. Guimarães[1], M.F. Martins[2], J.O. Peixoto[3], J. Furlong[2], M.C.A. Prata[2], R.S. Verneque[2], R.L. Teodoro[2] and P.S. Lopes[1]

[1]Departamento de Zootecnia, Universidade Federal de Viçosa,
Viçosa, MG, Brasil
[2]Embrapa Gado de Leite, Juiz de Fora, MG, Brasil
[3]Embrapa Suínos e Aves, Concórdia, SC, Brasil

Corresponding author: M.A. Machado
E-mail: machado@cnpgl.embrapa.br

ABSTRACT. Tick resistance in cattle is mainly found in zebu (*Bos indicus*) animals, although it is also present in some taurine (*B. taurus*) breeds. In order to characterize functional genes involved in tick resistance/susceptibility in cattle, two cDNA libraries were generated using skin tissues of selected Holstein x Gyr animals. A total of 2700 high-quality reads from both resistant and susceptible cDNA were assembled into 458 sequences (contigs) and 834 singletons, with a mean size of 447.7 nucleotides. Assignment of homologous proteins by BLASTX revealed 790 (61.1%) and 300 (23.2%) hits in resistant and susceptible cDNA, respectively; 121 of these hits matched bovine proteins. A total of 502 (38.9%) unique sequences were found to have no significant homology with known sequences and were classified as novel sequences. In general, the most abundant sequences consisted of those coding for hypothetical proteins whose function had not yet been determined, in addition to ribosomal proteins, binding proteins

and structural proteins, such as keratin and collagen. The most abundant protein found was collagen type III alpha, although ribosomal proteins accounted for half of the 40 most frequent hits. In addition, five matches within the top 40 best hits corresponded to immune response proteins. These sequences could be used for future studies on functional genomics of cattle tick resistance as well as for genomic sequencing projects.

Key words: Expressed sequence tags; Skin; cDNA library; Bovine; Crossbreeds

INTRODUCTION

Riphicephalus (*Boophilus*) *microplus* is a cattle ectoparasite found in tropical and subtropical regions worldwide (Willadsen and Jongejan, 1999). In Brazil, ticks are a big problem for the cattle industry nationwide and the most common strategy used against this ectoparasite is through the wide use of acaricides. Due to its serious impact on cattle production, there is an urgent need to understand and to unveil the underlying mechanisms involved in tick resistance/susceptibility in cattle.

Genetic resistance to the bovine tick is one of the most important factors in reducing costs to control this parasite in tropical cattle systems. Differences in domestication and selection processes have contributed to considerable phenotypic and genotypic differences between *Bos taurus* and *Bos indicus* breeds. In general, *B. indicus* animals are more resistant to parasitic diseases than *B. taurus*. The high susceptibility of cattle breeds of European origin to *R. microplus* in Brazil, including Holstein and Brown Swiss dairy breeds, was reported by Villares (1941).

In bovine, the epithelium serves as the first barrier against parasites and the innate immune response constitutes the first line of defense against pathogen invasion (Gumbiner, 1993). Disturbances in this barrier can lead to the invasion of microorganisms and pathogenic agents, causing subsequent disease. So far, few genes related to the nonspecific immunity have been isolated and characterized in cattle. The innate immune response plays an important role in protecting cattle against foreign invasion.

Little is known about the genetic mechanisms involved in the resistance of *Bos indicus* animals to ticks. Genetic resistance can contribute to reducing the tick population, costs with medications, mortality and production losses as well as contributing to lesser use of acaricides, which in turn decreases animal intoxication and environment pollution. To identify genomic regions involved in tick resistance, we developed a F_2 population originated from a *B. taurus* x *B. indicus* cross (Martinez et al., 2006). A total of 360 F_2 animals were evaluated for tick resistance for two seasons (rainy and dry) and a genome wide scan with 180 microsatellite markers was performed (Machado et al., 2010).

Expressed sequence tag (EST) analysis, which surveys sequences of cDNA libraries, is a powerful approach for identifying new genes and profiling gene expression in tissues or cells. Genomic information is becoming more abundant for many species, which could be used to develop new control strategies. One way to understand the function of the skin tissue under infestation with ticks would be through the analysis of differentially expressed genes in the host skin tissue. The identification of new genes and host antigens involved in the mechanism of resistance/susceptibility to tick is a promising approach (Douglas et al., 1999). Therefore, in this report we present a comparative and functional genomics approach to analyze ESTs from

two cDNA libraries that could be useful for the whole genome annotation as well as for the understanding of the genetic and molecular interactions between *R. microplus* and its cattle host.

MATERIAL AND METHODS

Animals, artificial tick infestation and assessment of resistance

Animals used in this study were derived from a F_2 population developed at Embrapa, Brazil, from the crossing of F_1 females (50% Gyr: 50% Holstein) with F_1 sires of the same genetic composition (for more details, see Machado et al., 2010).

To determine tick resistance, each F_2 animal was artificially infested with tick larvae. Prior to experimental infestations with ticks, animals were bathed with acaricides to assure they were free of natural infestation on pastures. The animals were kept in a picket free of ticks until full clearance of the chemical acaricides was achieved and then submitted to artificial infestation with 10,000 tick larvae in the "dorsal-lumbar" region of the animals. The absolute number of ticks determines the level of resistance of each animal. The tick count was based on the procedure adopted by Wilkinson (1962) and modified by Wharton and Utech (1970), in which all engorged female ticks of 4.5 to 8.0 mm in diameter are counted on the right side of each animal. The counts of adult female ticks were done on the 21st day after infestation. Tick counts ranged from zero to 792 ticks per animal, showing extreme genetic variability in the F_2 population.

The tick count data were transformed to logarithmic scale using the following procedure: transformed count (Ct) = Log_{10} (tick count+1) (Gill et al., 1896; Smothers et al., 1999). Some independent traits, such as coat color, coat thickness, coat length, hair density, sex, age, infestation order, infestation year and animal group, can interfere in the number of ticks per animal. For this reason, the genetic value (u), which represents the average genetic difference of Ct of the animal in relation to the other animals, was calculated for each evaluated animal. The genetic value was obtained from a model that considers, besides the animal effect, the effect of other independent traits. The genetic value was used to remove the effect of these traits on the tick count (Ct). It was obtained by the solution of the system of equations, represented by:

$$\begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_e^2}{\sigma_a^2} \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

where β is the vector of fixed effects in the model (coat color, coat thickness, coat length, hair density, sex, age, infestation order, infestation year and animal group); u is a vector containing the genetic values of Ct for each animal evaluated; X is a matrix of the incidence of the fixed effects; Z is the matrix of the incidence of the random effects; σ_e^2 is the residual variance; σ_a^2 is the genetic additive variance and y is the transformed tick count vector. The solution of the equations was obtained by the system MTDFREML (multiple trait derivative free restricted maximum likelihood) (Boldman et al., 1995). Based on results of genetic values for tick resistance, animals with extreme breeding values for tick resistance/susceptibility were selected for the experimental groups. A total of six tick-resistant and six tick-susceptible F_2 animals were selected for tissue collection to generate the Res and Sus cDNA libraries.

Construction of skin cDNA libraries and single pass sequencing

Skin samples were collected from all animals on day 5 and 12 after artificial infestation. Skin biopsies of 6 mm in diameter were obtained near the animal tail insertion region where ticks most attached. After tissue collection, 800 mg of biopsies from each animal, collected on day 5 and 12 post infestation were gathered into two pools (Res and Sus) to perform RNA extraction. Total RNA was extracted using the RNeasy® Maxi kit (Qiagen, Valencia, CA) according to manufacturer instructions. The quality of the RNA was monitored by examination of the 18S and 28S ribosomal RNA bands after electrophoresis and quantified by spectrophotometry at 260 nm. Poly(A) RNA was isolated using the Oligotex® kit (Qiagen, Hilden, Germany), according to manufacturer instructions. Superscript™ Plasmid System with Gateway TM technology for cDNA Synthesis and Cloning kit (Invitrogen, Carlsbad, CA, USA) was used for construction of the two cDNA libraries. Plasmid DNA purification followed an alkaline lysis procedure (Birnboim and Doly, 1979; Sambrook and Russel, 2001)

Sequencing reactions were performed with 200 ng plasmid DNA template prepared in 96-well microplates using DYEnamic ET dye terminator kit (GE Healthcare, Piscataway, NJ) with M13 reverse or forward primer, followed by sequencing reaction clean up to remove residual dye and enzyme. Unidirectional single-pass sequencing was submitted to a MegaBACE 1000 capillary DNA sequencer (GE Healthcare, Piscataway, NJ).

Bioinformatics analyses

EST sequences were extracted from the raw data then rigorously screened with default parameters to remove contaminating vectors (minimatch 12, penalty -2, minscore 20), poly-A/T tails (8), primers and adaptor sequences. Contaminating bacterial sequences were masked using Cross-Match software and sequences of low-quality regions at both ends were eliminated using the trim option in Phred software with cutoff of 20 (Ewing et al., 1998). Alignment of the transcripts against *B. Microplus* EST database [http://compbio.dfci.harvard.edu/tgi/tgipage.html] was performed to ensure that the sequences from the animals with ticks were free of sequences of the parasite, due to the presence of their oral apparatus in tissue samples. After trimming, fragments above 150 bp and 80% identity were subjected to further analysis.

ESTs were clustered and assembled into contigs using CAP3 software (Huang and Madan, 1999) set to default parameters. ESTs that did not link to contigs were designated singletons. Collectively, the resultant contigs and singletons are referred to as unique sequences.

The unique sequences were then submitted to a stand-alone BLAST (Altschul et al., 1997) against non-redundant protein databases from NCBI GenBank and Swiss-Prot using an *E*-value of $1e^{-05}$. The coding region was determined by TargetIdentifier program [http://proteomics.ysu.edu/tools/]. Assignments to the gene ontology terms were performed using Blast2GO software suite v2 (Conesa et al., 2005) using default parameters. The BLAST hits were automatically assigned against the curate InterPro database to find functional motifs (InterProScan, EBI). Goslim and metabolic pathways (KEGG, Kyoto Encyclopedia of Genes and Genomes). A Venn diagram was used to compare the two dataset BLASTX results against Nr and Swiss-Prot protein database and display the relative similarity relationships.

Peptides were also compared, using BLASTX, with a protein sequence database from

Ensembl [http://ncbi36.ensembl.org/]. Protein databases of *Homo sapiens* (GRCh37), *Mus musculus* (NCBIM37) and *Rattus novergicus* (RGSC3.4) were built in-house for similarity searches.

Nucleotide sequence accession numbers

The nucleotide sequences that were identified were deposited in the dbEST database (http://www.ncbi.nlm.nih.gov/Genbank/dbEST/) under accession numbers GO578626 to GO581239 and dbEST-Id 64942695 to 64945308.

RESULTS

Sequencing and EST library characterization

Two independent bovine skin cDNA libraries from tick-resistant and tick-susceptible animals were sequenced to generate EST datasets. Single-pass sequencing was performed from 5'-end of 3563 cDNA and 3'-end of 507 randomly picked clones from the two non-normalized cDNA libraries. After sequence trimming, a total of 4070 reads were used for quality check. Clean-up sequences resulted in 2700 high quality ESTs (Table 1). Assembling the high quality sequences into contigs gave a total of 1292 unique sequences, consisting of 458 contigs and 834 singletons with an average size of 471.3 and 424.1 nucleotides to Res and Sus cDNA libraries, respectively.

Table 1. Statistical analyses of cattle ESTs derived from tick resistant and susceptible cDNA libraries.

Sequencing analysis	Number of sequences and percentage (%)		Total
	Resistance	Susceptible	
Category			
ESTs sequenced	3'-279 (14.8)	3'-228 (10.4)	507 (12.5)
	5'-1.606 (85.2)	5'-1.957 (89.6)	3.563 (87.5)
Reads	1.885 (46.3)	2.185 (53.7)	4.070 (100)
Clusterization and assembly			
High quality ESTs	1.235 (45.7)	1.465 (54.3)	2.700
Unique sequences	700 (54.2)	592 (45.8)	1.292
Contigs	225	233	458
Singletons	475	359	834
Mean sequence length (nt)	471.3	424.1	447.7

Annotation and functional classification of unique sequences

Using BLASTX to annotate unique sequences with an expectation value of 1E-05 or better, we were able to estimate the number of proteins identified by the 1292 unique sequences. Analysis of the positive homologs on NCBI non-redundant protein database revealed 790 hits to known proteins, indicating that 61.1% of them contained inserts of known genes (Table 2). Five hundred and two (38.9%) unique sequences showed no significant homology to known sequences and were classified as novel sequences. From the 790 hits with known proteins, 226 sequences contain inserts with full-length protein sequences, 184 unique sequences were derived from partial length sequences 5' and 128 were derived from partial length se-

quences 3'. In addition to that, 29 sequences showed a 5' stop codon without a start codon and were classified as ambiguous sequences. A total of 156 full-length and ORF completed sequences were found.

Table 2. Summary of homologous analysis with BLASTN and BLASTX algorithms against non-redundant nucleotide (Nr) and Nr and Swiss-Prot protein databases (%).

Libraries	Resistance	Susceptible	Total
BLASTX (Nr)			
No hits	255 (50.8)	247 (49.2)	502 (38.9)
Hits with known protein	495 (50.8)	295 (62.6)	790 (61.1)
Full length	139	87	226
Partial length 5'	120	64	184
Partial length 3'	102	26	128
Ambiguous	14	15	29
Full-length and ORF	74	82	156
BLASTX (Swiss-Prot)			
No hits	561 (56.6)	431 (43.4)	992 (76.9)
Hits with known protein	139 (46.3)	161 (53.7)	300 (23.1)
Hits with known bovine protein	52	69	121

Significant hits on the Nr database were followed up with protein function searches in the Swiss-Prot protein database, which provides value-added information reports for protein functions (Table 2). Roughly, we found 300 (23.1%) hits to homologous proteins with similarity >80%. A total of 121 of these hits matched bovine proteins, 52 from Res and 69 from Sus datasets. The remaining 992 (76.9%) sequences showed no putative identification or similarity <80%. The size distribution of the similarity of transcripts with matches to bovine homologous proteins ranged from 150 to 1600 nucleotides.

Gene ontology analysis

To get a better overview of the annotated unique sequences, we used the Blast2GO program to map and annotate unique EST unique sequences into functional categories. GO hits distribution for unique sequences assigned to the functional categories Biological Process (P), Cellular Component (C) and Molecular Function (M) are presented in Figure 1. We examined the assigned Biological Process categorization to determine differences in the distribution of these processes between the two libraries. From the sequences mapped on Biological Process, two major annotations on Res library were involved in macromolecule metabolic process (20%), followed by biosynthetic process (17%) (Figure 1a). In the Sus library, 86 unique sequences were mapped to the macromolecule metabolic process (26%) and biosynthetic process (23%) on this dataset (Figure 1d).

According to Cellular Component categorization, we attributed GO terms into cell, cell part, organelle, extracellular region, extracellular region part and extracellular matrix from Res and Sus datasets (Figure 1b and 1e). The major categories were intracellular (35 and 33%) and intracellular (35 and 32%) from Res and Sus cDNA datasets, respectively.

In addition, to categorize by Molecular Function, genes were distributed into the following categories: binding, structural molecule activity, catalytic activity, transport activity, enzyme regulatory activity and transcription regulatory activity, motor activity and molecular transducer activity (Figure 1c and 1f). Functional class protein binding (57 and 37%) and nucleic acid binding (35 and 32%) were the main categories found on Res and Sus cDNA datasets, respectively.

(a) Biological Process - Res dataset

- GO:0044238 primary metabolic process biological_process
- GO:0044237 cellular metabolic process biological_process
- GO:0007154 cell communication biological_process
- GO:0050789 regulation of biological process biological_process
- GO:0051674 localization of cell biological_process
- GO:0007275 multicellular organismal development biological_process
- GO:0006928 cell motility biological_process
- GO:0009058 biosynthetic process biological_process
- GO:0043170 macromolecule metabolic process biological_process
- GO:0048869 cellular developmental process biological_process
- GO:0006810 transport biological_process

(b) Biological Process - Sus dataset

- GO:0006928 cell motility biological_process
- GO:0051674 localization of cell biological_process
- GO:0007275 multicellular organismal development biological_process
- GO:0009058 biosynthetic process biological_process
- GO:0044237 cellular metabolic process biological_process
- GO:0006810 transport biological_process
- GO:0043170 macromolecule metabolic process biological_process
- GO:0044238 primary metabolic process biological_process
- GO:0048869 cellular developmental process biological_process
- GO:0007154 cell communication biological_process
- GO:0050789 regulation of biological process biological_process

(c) Cellular Component - Res dataset

- GO:0016020 membrane cellular_component
- GO:0044424 intracellular part cellular_component
- GO:0005622 intracellular cellular_component
- GO:0043229 intracellular organelle cellular_component
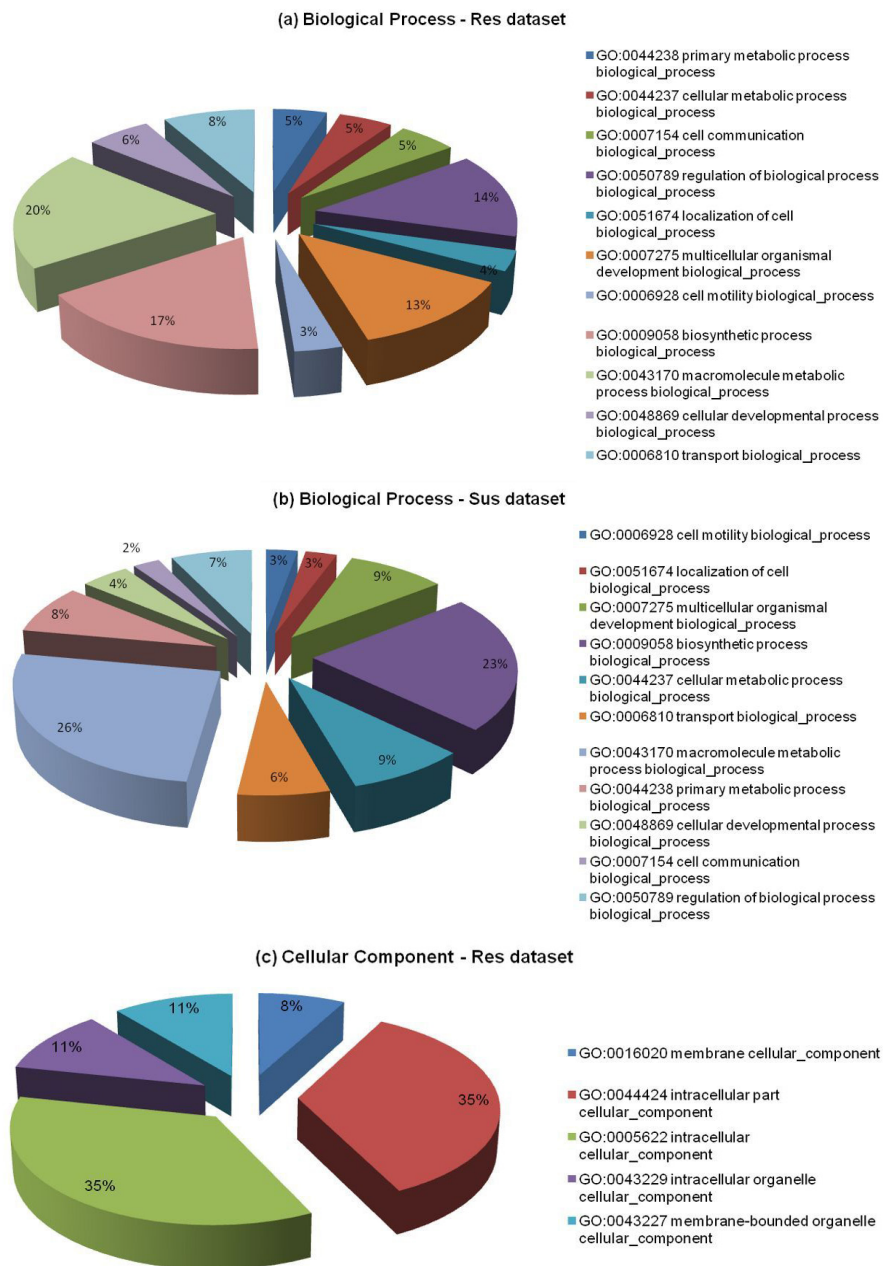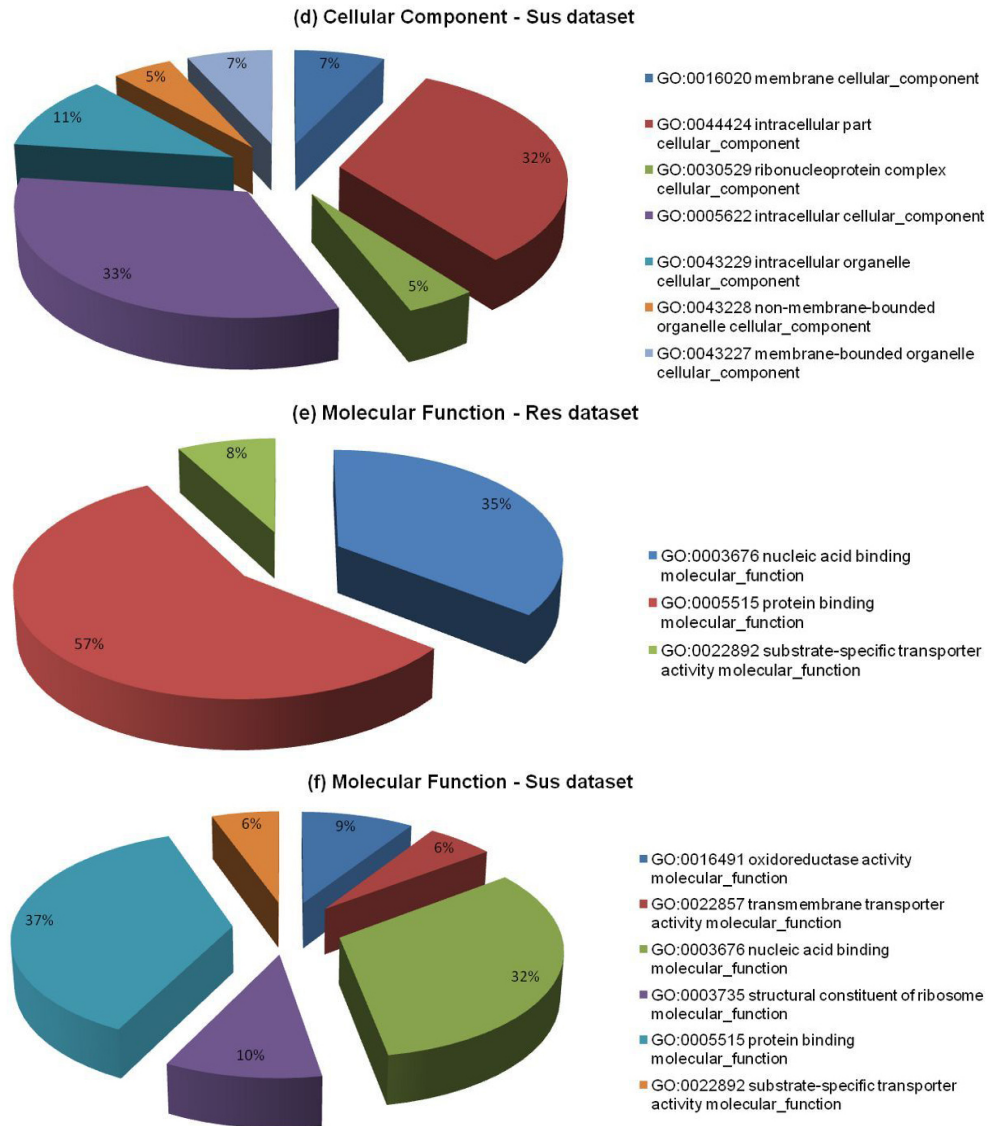- GO:0043227 membrane-bounded organelle cellular_component

Figure 1. Functional annotation of consensus sequences based on Gene Ontology (GO) categorization. Sequences were evaluated for their predicted involvement in Biological Processes (a and d), Cellular Components (b and e) and Molecular Functions (c and f) in Res and Sus datasets, respectively. Data are presented at level three GO categorization. Piers chart indicate the relative percentage of sequence consensus in each group.

Figure 1. Continued.

(d) Cellular Component - Sus dataset

- GO:0016020 membrane cellular_component
- GO:0044424 intracellular part cellular_component
- GO:0030529 ribonucleoprotein complex cellular_component
- GO:0005622 intracellular cellular_component
- GO:0043229 intracellular organelle cellular_component
- GO:0043228 non-membrane-bounded organelle cellular_component
- GO:0043227 membrane-bounded organelle cellular_component

(e) Molecular Function - Res dataset

- GO:0003676 nucleic acid binding molecular_function
- GO:0005515 protein binding molecular_function
- GO:0022892 substrate-specific transporter activity molecular_function

(f) Molecular Function - Sus dataset

- GO:0016491 oxidoreductase activity molecular_function
- GO:0022857 transmembrane transporter activity molecular_function
- GO:0003676 nucleic acid binding molecular_function
- GO:0003735 structural constituent of ribosome molecular_function
- GO:0005515 protein binding molecular_function
- GO:0022892 substrate-specific transporter activity molecular_function

Comparison of frequencies with Bovine (using whole *Bos taurus* genome annotation) within the Biological Processes, Cellular Component and Molecular Function GO categories showed overall consistency in terms of representation within the functional classes (Figure 2). We successfully annotated 1175 GO terms at a mean level, 621 from Res and 554 from Sus datasets. The two most representative GO terms were biological process and molecular function to both libraries.
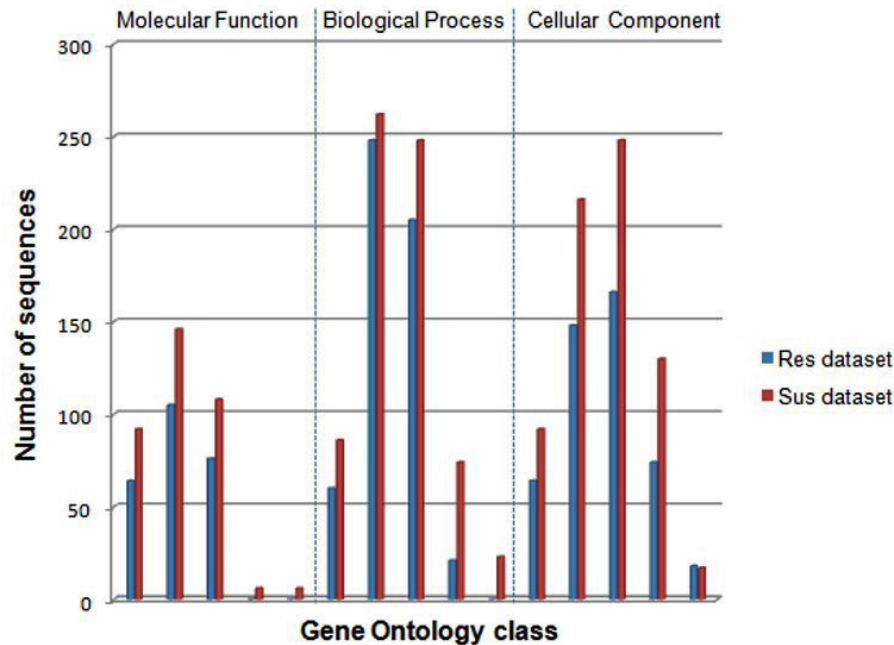
Figure 2. GO level distribution bar chart for skin cDNA libraries. The GO hits for Res and Sus unigenes were assigned to (a and d) biological process, (b and e) cellular components and (c and d) molecular functions, respectively. GO terms allowed assignment of 621 annotations to the Res library, mean level = 4.47, standard deviation = 1372 and 554 annotations to the Sus library, mean level = 4.13, standard deviation = 1302.

Comparison between cDNA libraries

A comparative analysis was done using the BLASTX hits information from each library to identify unique sequences in each library and also the common genes shared with them. The distribution of unique sequences with known function is presented as a Venn diagram in Figure 3. The two circles indicate the total number of sequences predicted in each library and the overlap area indicates the number of shared sequences. A total of 790 (Nr) and 221 (Swiss-Prot) unique hits were included in this comparison. The Venn diagram indicated that 43 (5.4%) and 22 (10.0 %) hits were shared by the two cDNA libraries. A total of 451 (57.1%) and 136 (61.5%) unique sequences show homologous protein sequences exclusively originated from the Res dataset on Nr and Swiss-Prot proteins databases, respectively. Likewise, 245 (31.0%) and 34 (15.4%) hits were exclusively originated from the Sus dataset.

Comparative genomics was used to characterize the conserved protein in several species. Data were compared with protein sequences available for *Homo sapiens*, *Mus musculus* and *Rattus novergicus* from the Ensembl protein database [http://www.ensembl.org/]. Three hundred and forty seven (84.6%) unique putative proteins, identified by BLASTX matches, had homologs in the Human protein dataset; 338 (84.9%) sequences had significant matches with the Mouse protein sequence and 329 (83.9%) were considered to be of Rat origin (Table 3). The table results show that, given the current database contents, the sampled skin transcriptome from *Bos taurus* x *Bos indicus* is equally close to available Human, Mouse and Rat protein sequences.
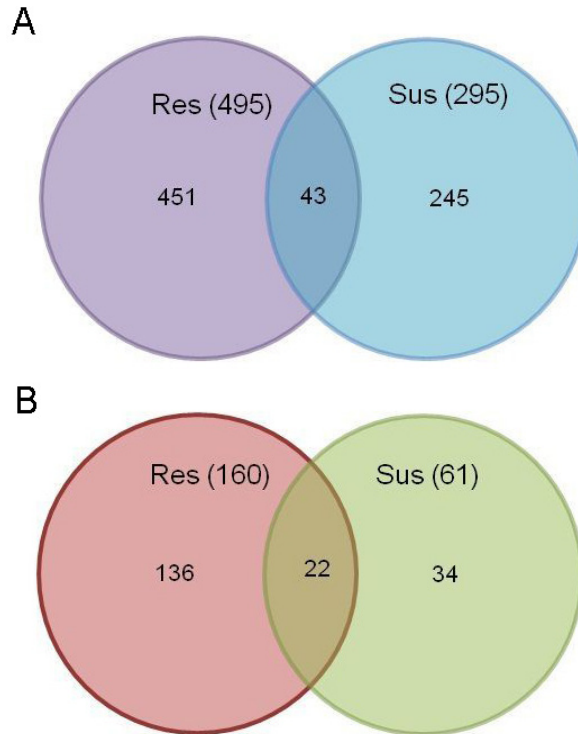
A

B

Figure 3. Expressed genes identified from blast searches against protein databases. Diagram Venn classification of the 790 (Nr) and 221 (Swiss-Prot) hits with BLASTX homology (Evalue <1e-5) from skin cDNA datasets where the diagram in Figure 2a represents protein hits with similarity on the Nr protein database and in Figure 2b shows hits identified on the Swiss-Prot protein database. The total number of hits in each class is given, including the number of the total matches from each cDNA library against Nr and Swiss-Prot protein databases.

Table 3. Comparative analysis of proteins from *Homo sapiens*, *Mus musculus* and *Rattus novergicus* with hits on Ensembl protein database (%).

	H. sapiens	*M. Musculus*	*R. novergicus*
Res	185	177	176
Sus	225	221	216
Unique proteins	347 (84.6)	338 (84.9)	329 (83.9)
Total	410	398	392

Identification of the most abundant and significant transcripts

The bovine sequences were clustered into more abundant categories and classes based on BLASTX results with associated predicted or known functions. Assignment of homologous proteins by BLASTX revealed 60 sequences with hits to hypothetical proteins on Res (N = 50) and Sus (N = 10) libraries. The most frequently found genes were ribosomal protein matches on Res (N = 35) and Sus (N = 55) libraries, followed by binding proteins matches on Res (N = 22) and Sus (N = 14), and structure cell proteins like keratin and collagen matches on Res (N = 13) and Sus (N = 12).

The top 40 most significant (Expect value = 1E-78 to 0) BLASTX hits are shown in Table 4. Highly expressed housekeeping proteins such as ribosomal proteins and collagen were abundantly found. The most significant was collagen type III alpha (Contig67_Res, E-value = 0) although ribosomal proteins account for half of the 40 best hits. Importantly, five matches within the top 40 correspond to immune response proteins. In fact, we detected two unique sequences that showed high significance to immune response proteins such as the cathepsin L2 precursor (Bs_Res_03G10_rc, E-value = 1E-100) and MHC class antigen I (Contig200_Res, E-Value = 1E-144) on Res dataset. In Sus, unique sequences matched to the CD44 antigen (Contig7_Sus, E-value = 1E-131), CD63 antigen ((Bs_Sus_18E06; E-value = 1.00E-98) and ADAM metallopeptidase (Bs_Sus_25D08, E-Value = 1E-88).

Table 4. The 40 unique sequences with the most significant sequence similarity (E-value) to sequences in the National Center for Biotechnology Information non-redundant protein database

Contig or singlet ID	Evalue	Accession No.	Protein Description
Contig67_Res	0	NP_000081.1	collagen type III alpha
Contig11_Res	1,00E-160	NP_001029211.1	collagen, type I, alpha
Contig34_Res	1,00E-149	NP_000997.1	ribosomal protein S3a
Contig200_Res	1,00E-144	gbAAO91983.1	MHC class I antigen
Contig7_Sus	1,00E-131	NP_776438.1	CD44 antigen
Contig185_Res	1,00E-125	NP_001038008.1	ribosomal protein L10
Contig164_Sus	1,00E-122	NP_001035610.1	ribosomal protein L7a
Bs_Res_87C06	1,00E-118	NP_777140.1	ribosomal protein L3
Bs_Sus_09A01	1,00E-115	NP_776993.1	poly(A) binding protein, cytoplasmic 1
Bs_Res_87D06	1,00E-115	YP_258382.1	hypothetical protein PFL_1251
Bs_Res_87C09	1,00E-114	NP_886542.1	putative extracellular solute-binding
Contig160_Res	1,00E-113	YP_258417.1	hypothetical protein PFL_1288
Contig157_Sus	1,00E-112	NP_001070466.1	ribosomal protein L13a
Contig47_Res	1,00E-108	NP_776802.1	keratin 10
Contig96_Sus	1,00E-106	NP_001038008.1	ribosomal protein L10
Bs_Res_89F07	1,00E-106	gbAAX29348.1	ribosomal protein S9
Bs_Res_89D02	1,00E-106	NP_001032906.1	RAN binding protein 6
Contig82_Sus	1,00E-104	gbAAI02075.1	Ribosomal protein, large, P0
Contig65_Res	1,00E-104	gbAAP36646.1	Homo sapiens keratin 14
Contig181_Sus	1,00E-103	gbAAX29111.1	ribosomal protein S7
Contig124_Res	1,00E-102	XP_543647.2	similar to keratin
Contig141_Res	1,00E-101	gbAAX43400.1	ribosomal protein S5
Bs_Res_03G10_rc	1,00E-100	spQ5E998CATL2	Cathepsin L2 precursor
Contig86_Sus	2,00E-99	NP_001030383.1	ribosomal protein L5
Contig186_Sus	1,00E-98	NP_001019640.2	ribosomal protein L9
Bs_Sus_18E06	1,00E-98	spQ9XSK2	CD63 antigen
Bs_Res_98D12	6,00E-98	gbAAX29834.1	ribosomal protein L11
Contig69_Sus	3,00E-97	XP_581277.2	similar to odorant binding
Bs_Res_92A02	4,00E-97	XP_231785.2	similar to 60S
Contig18_Res	1,00E-93	NP_001008663.1	keratin 5 (epidermolysis bullosa)
Bs_Res_02B12	8,00E-91	XP_523648.1	keratin 35
Bs_Sus_24E07	6,00E-90	XP_876960.1	hypothetical protein
Bs_Sus_25D08	1,00E-88	NP_777056.1	ADAM metallopeptidase
Contig43_Sus	5,00E-87	gbAAX29372.1	ribosomal protein S11
Contig35_Sus	3,00E-84	NP_001019642.1	ribosomal protein L27a
Contig115_Sus	2,00E-82	XP_868296.1	similar to ribosomal
Contig148_Sus	4,00E-82	XP_533657.2	similar to ribosomal
Contig156_Sus	5,00E-82	NP_035426.1	ribosomal protein S18
Contig65_Sus	5,00E-81	NP_001014862.1	ribosomal protein L29
Contig1_Sus	3,00E-78	NP_001019712.1	ribosomal protein S15

DISCUSSION

Identification and sequencing of ESTs have been successfully used in functional genomics and contribute to research in microarray construction and interspecies comparison. Furthermore, EST analysis is an efficient and fast method for gene discovery. In the present study, two skin cDNA libraries from F_2 (Holstein x Gyr) cattle were constructed and 2700 high quality ESTs were analyzed.

We used a BLASTX and GO annotation to reveal the putative functions of genes expressed in the skin tissue for both groups - resistant and susceptible animals. Both cDNA libraries were comparable in their GO annotation and GO terms likely to be related to the function and structure of the skin were discovered. However, it should be noted that the functional assignments are not from direct assessment but derived from computational annotation, and it is likely that all potential functions of a gene are not completely described (Dean et al., 2009).

In the analysis of ESTs, a high percentage showed no significant similarity to entries in the public nucleotide and protein databases. This was expected due to the small number of genetic studies in *Bos indicus* and *Bos indicus* x *Bos taurus* crossbred animals. Thus, a significant achievement of this study was to increase the sequence information for these breeds through the 502 putative novel sequences found (see Table 2).

Comparison of human, mouse and rat muscle proteins was used to demonstrate that more than 80% of cattle skin proteins were detected in at least one of these proteomes (see Table 3). The finding that most of these proteins showed high similarities with human is consistent with many previous studies on similarity between human and bovine genomes (Band et al., 2000; Wind et al., 2005; Itoh et al., 2005). Comparative genomics provides a powerful new tool for studying evolutionary changes among organisms, helping to identify the genes that are conserved among species along with genes that give each organism its own unique characteristics. The results obtained in this work could be a valuable resource with high sequence complexity for studying gene expression and the regulatory mechanism in skin tissue.

The most abundant group of protein hits in this study was related to ribosomal proteins followed by binding proteins. The high number of proteins associated with housekeeping processes was somehow expected due to the energy production and protein synthesis requirements in both normal and regenerating tissues. In addition, ribosomal protein genes are expressed ubiquitously at all development stages. Moreover, the ribosomal protein family is generally well conserved and contains 55 proteins in prokaryotes and 88 in eukaryotes (Doudna and Rath, 2002). An increasing number of studies have reported that numerous ribosomal proteins show extra-ribosomal functions, such as involvement with several human genetic disorders (Wool, 1996). Studies in gene expression profiling on cattle skin have found an enormous number of ribosomal proteins up regulated in response to a challenge with ticks in both high and low resistance animals (Wang et al., 2007). Furthermore, it is known that global changes in the expression of ribosomal proteins can occur as a response to stress (Causton et al., 2001). However, changes in the expression of individual ribosomal proteins could be explained if the protein has an additional extra ribosomal function (Wool, 1996).

A total of 40 unique sequences were also identified and showed the most significant alignments with protein of known function in the NCBI non-redundant protein database (Table 4). Some of them were related to molecules of the immune system. In the Res dataset, homologs to Cathepsin L2 and ADAM metallopeptidase were identified. Cathepsin L2, also

known as cathepsin V, encoded by the CTL2 gene, is a member of the peptidase C1 family (Adachi et al., 1998); ADAMTS2 is a procollagen N-proteinase that cleaves procollagen I, II, and III to the corresponding collagens and is required to generate collagen monomers that are able to assemble into elongated and cylindrical collagen fibrils (Colige et al., 2005). One transcript matched to MHC class I antigens although this sequence is not restricted to immune cells. In the Sus dataset, CD44 and CD63 antigens were identified. CD44 encodes a cell surface glycoprotein involved in cell/cell and cell/matrix interactions (Hale et al., 1995). The CD63 antigen was originally designated as a lysosomal membrane protein characterized as an activation dependent platelet surface antigen (Metzelaar et al., 1991).

CONCLUSIONS

We successfully generated and analyzed two skin cDNA libraries derived from a Holstein x Gyr (F_2) crossbred population artificially infested with ticks. The results reported here are a prerequisite for further studies on gene expression by cDNA microarray and for the identification of candidate genes for milk performance in cattle. The diversity of the EST collection developed in this study will facilitate genome annotation as well as identification and characterization of genes involved in host infection in bovine through a comparative and functional genomics approach.

ACKNOWLEDGMENTS

REFERENCES

Adachi W, Kawamoto S, Ohno I, Nishida K, et al. (1998). Isolation and characterization of human cathepsin V: a major proteinase in corneal epithelium. *Invest. Ophthalmol. Vis. Sci.* 39: 1789-1796.

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Band MR, Larson JH, Rebeiz M, Green CA, et al. (2000). An ordered comparative map of the cattle and human genomes. *Genome Res.* 10: 1359-1368.

Birnboim HC and Doly J (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* 7: 1513-1523.

Boldman KG, Kriese LA and Van Vleck LD (1995). A Manual for Use of MTDFREML. A Set of Programs to Obtain Estimates of Variances and Covariances. Department of Agriculture/Agricultural Research Service U.S. USDA-ARS, Lincoln.

Causton HC, Ren B, Koh SS, Harbison CT, et al. (2001). Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* 12: 323-337.

Colige A, Ruggiero F, Vandenberghe I, Dubail J, et al. (2005). Domains and maturation processes that regulate the activity of ADAMTS-2, a metalloproteinase cleaving the aminopropeptide of fibrillar procollagens types I-III and V. *J. Biol. Chem.* 280: 34397-34408.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.

Dean MD, Clark NL, Findlay GD, Karn RC, et al. (2009). Proteomics and comparative genomic investigations reveal

heterogeneity in evolutionary rate of male reproductive proteins in mice (*Mus domesticus*). *Mol. Biol. Evol.* 26: 1733-1743.

Doudna JA and Rath VL (2002). Structure and function of the eukaryotic ribosome: the next frontier. *Cell* 109: 153-156.

Douglas SE, Gallant JW, Bullerwell CE, Wolff C, et al. (1999). Winter flounder expressed sequence tags: Establishment of an EST database and identification of novel fish genes. *Mar. Biotechnol.* 1: 458-0464.

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

Gill JL, Ericsson GF and Helland IS (1986). Precision of assessing anthelmintic efficacy. *Biometrics* 42: 981-987.

Gumbiner BM (1993). Breaking through the tight junction barrier. *J. Cell Biol.* 123: 1631-1633.

Hale LP, Patel DD, Clark RE and Haynes BF (1995). Distribution of CD44 variant isoforms in human skin: differential expression in components of benign and malignant epithelia. *J. Cutan. Pathol.* 22: 536-545.

Huang X and Madan A (1999). CAP3: A DNA sequence assembly program. *Genome Res.* 9: 868-877.

Itoh T, Watanabe T, Ihara N, Mariani P, et al. (2005). A comprehensive radiation hybrid map of the bovine genome comprising 5593 loci. *Genomics* 85: 413-424.

Machado MA, Azevedo AL, Teodoro RL, Pires MA, et al. (2010). Genome wide scan for quantitative trait loci affecting tick resistance in cattle (*Bos taurus* x *Bos indicus*). *BMC Genom.* 11: 280.

Martinez ML, Machado MA, Nascimento CS, Silva MV, et al. (2006). Association of BoLA-DRB3.2 alleles with tick (*Boophilus microplus*) resistance in cattle. *Genet. Mol. Res.* 5: 513-524.

Metzelaar MJ, Wijngaard PL, Peters PJ, Sixma JJ, et al. (1991). CD63 antigen. A novel lysosomal membrane glycoprotein, cloned by a screening procedure for intracellular antigens in eukaryotic cells. *J. Biol. Chem.* 266: 3239-3245.

Sambrook J and Russel DW (2001). Molecular Cloning. A Laboratory Manual. 3rd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor.

Smothers CD, Sun F and Dayton AD (1999). Comparison of arithmetic and geometric means as measures of a central tendency in cattle nematode populations. *Vet. Parasitol.* 81: 211-224.

Villares JB (1941). Climatologia Zootecnica III: Contribuição ao Estudo de Resistência e Susceptibilidade Genética dos Bovinos ao *Boophilus microplus*. In: Boletim da Indústria Animal: 8-9 July 1941, Edited by Nova Serie, São Paulo.

Wang YH, Reverter A, Kemp D and McWilliam SM (2007). Gene expression profiling of Hereford Shorthorn cattle following challenge with *Boophilus microplus* tick larvae. *Aust. J. Exp. Agr.* 47: 1397-1407.

Westfall PH and Young SS (1993). Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. John Wiley & Sons, Inc., New York.

Wharton RH and Utech KBW (1970). The relation between engorgement and dropping of *Boophilus microplus* (Canestrini) (Ixodidae) to the assessment of tick numbers on cattle. 171.

Wilkinson PR (1962). Selection of cattle for tick resistance, and the effect of herds of different susceptibility on *Boophilus* populations. *Aust. J. Agr. Res.* 13: 974-983.

Willadsen P and Jongejan F (1999). Immunology of the tick-host interaction and the control of ticks and tick-borne diseases. *Parasitol. Today* 15: 258-262.

Wind AE, Larkin DM, Green CA and Elliott JS (2005). A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proc. Nat. Acad. Sci. U. S. A.* 102: 18526-18531.

Wool IG (1996). Extraribosomal functions of ribosomal proteins. *Trends Biochem. Sci.* 21: 164-165.