Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Generation of a preliminary bovine gene atlas, using expression clustering to annotate gene function

O.M. Keane[1,3], N. Maqbool[2], A.F. McCulloch[2], J.C. McEwan[2] and K.G. Dodds[2]

[1]Department of Biochemistry, University of Otago, AgResearch, Molecular Biology Unit, Dunedin, New Zealand
[2]AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand
[3]Animal Bioscience Centre, Teagasc, Grange, Dunsany, Co. Meath, Ireland

Corresponding author: O.M. Keane
E-mail: orla.keane@teagasc.ie

Genet. Mol. Res. 8 (3): 1013-1027 (2009)
Received February 3, 2009
Accepted June 10, 2009
Published August 25, 2009

ABSTRACT. Genes whose products function in a common biological process are often co-regulated. When regulation occurs at the transcriptional level, co-expressed genes can be detected globally by expression arrays or by sequencing non-normalized cDNA libraries. We examined bovine gene expression in 27 tissues using non-normalized cDNA library sequencing. Contigs were generated from expressed sequence tags whose sequences overlapped. Contigs containing a minimum of five expressed sequence tags were ordered via a hierarchical clustering process, where the distance between the contigs represents their expression pattern similarity across tissues. Gene ontology terms associated with the genes in each cluster showed that co-clustered genes encoded proteins involved in a common biological process. This process can be used to annotate genes of unknown function in the cluster. Gene expression was compared between bovine and human tissues; there were significant correlations between species for each tissue, with the exception of thyroid and placenta. Tissues were also clustered based on the genes they express; tissues with similar physiological functions clustered

closely. Based on this information, we generated the first preliminary gene atlas of the bovine genome. Genes with similar expression patterns were clustered, and genes with a common function co-clustered. This method can be used to annotate genes of unknown function in the bovine genome.

Key words: Bovine; Expressed sequence tag; Contig; Cluster; Transcriptional profiling; Gene atlas

INTRODUCTION

Expressed sequence tags (ESTs) are short single-pass DNA sequences obtained from either end of cDNA clones (Liang et al., 2000). Large-scale EST projects are common in many species and provide valuable sequence information for those species with a paucity of genomic sequence data. EST datasets also provide a vast resource for gene discovery and identification of important genetic variations such as alternative transcripts and single nucleotide polymorphisms (Ewing and Green, 2000; Irizarry et al., 2000; Gupta et al., 2004; Lee et al., 2006). The function of the genes represented by ESTs can be inferred by sequence homology between the EST and genes of known function. In addition, transcript abundance and tissue specificity may be predicted from EST data (Audic and Claverie, 1997).

Currently, there are over 1.5 million bovine ESTs deposited in NCBI dbEST (Boguski et al., 1993). The only mammalian species with more EST sequence data available are human and mouse. Despite this abundance of bovine EST data, there is a scarcity of information on bovine gene expression and variation.

Identifying the genes expressed in cells of a tissue is an important step towards providing essential information about gene function and tissue physiology. With the aim of cataloging a large proportion of the genes expressed in cattle, non-normalized cDNA libraries were created from a variety of tissues from animals of different ages and breeds. These cDNA libraries were single-pass sequenced from the 5' end, generating ESTs. These AgResearch ESTs, along with all bovine ESTs publicly available, were assembled into contigs in order to reduce the redundancy inherent in the cDNA libraries and to extend the length of known sequence for an individual transcript. The depth of AgResearch ESTs in a single contig was used to provide a measure of transcript abundance, and the variety of libraries in which the ESTs were detected were used to determine the profile of expression of that transcript across different tissues. Contigs were subsequently ordered via a hierarchical clustering process (Ward, 1963) with the distance between the contigs representing their expression pattern similarity in different tissues.

Some contigs either do not share significant sequence homology with sequences in public databases or are homologous to genes of unknown function, providing very little information on the function of the gene represented by the contig. Such contigs may have expression patterns similar to known genes, providing a level of annotation for the gene. This method of gene annotation has been successfully demonstrated in yeast, *Caenorhabditis elegans* and mouse (Kim et al., 2001; Wu et al., 2002; Zhang et al., 2004). In species of agricultural importance, contigs with expression patterns similar to genes, which confer desirable traits, are of particular interest, as they may represent economically important "novel genes". Additionally, the promoter region from contigs with similar expression patterns can be sequenced and common transcription fac-

tor binding site motifs detected in order to identify likely promoter regions contributing to the observed co-expression pattern (Leung and Chin, 2006; Zadissa et al., 2007).

MATERIAL AND METHODS

Generation of ESTs

Tissue samples were collected from a number of animals from research and commercial farms. These animals were of different ages and stages of development and the tissues that were collected are listed in Supplementary file 1. Tissue was sent to Genesis Biotechnology (Auckland, New Zealand) where cDNA was generated from isolated mRNA via a polyA primer and was directionally inserted into a pBKCMV vector. Clones were then single-pass sequenced from the 5' end, giving ESTs. All cDNA libraries were non-normalized and are listed in Table 1. The total library depth of each library is listed in Supplementary file 1 and the adjusted library depth (the number of ESTs from the library in contigs ≥5 ESTs) is listed in Table 1. All animal handling and procedures were approved by the AgResearch Animal Ethics Committee.

Table 1. List of bovine cDNA libraries and their human counterpart tissues and correlation of gene expression in human and bovine tissues.

Library	Bovine tissue	Human tissue	Pearson correlation coefficient (R)	No. of bovine ESTs*	P
BABA	Anterior brain stem	Pons	0.17	3088	$<1 \times 10^{-6}$
BAGA	Adrenal gland	Adrenal gland	0.05	1560	0.001
BAPA	Anterior pituitary	Pituitary gland	0.30	850	$<1 \times 10^{-6}$
BCEA	Cerebellum - post-natal	Cerebellum	0.27	3094	$<1 \times 10^{-6}$
BCEB	Cerebellum - 1 year				
BCMA	Cardiac muscle	Heart	0.29	2770	$<1 \times 10^{-6}$
BCNA	Contralateral ovary - non-ovulated	Ovary	0.08	7345	$<1 \times 10^{-6}$
BCOA	Contralateral ovary - ovulated				
BINA	Ipsilateral ovary - non-ovulated				
BIOA	Ipsilateral ovary - ovulated				
BOVA	Ovary				
BOVB	Ovary				
BCXB	Cortex	Prefrontal cortex	0.34	2793	$<1 \times 10^{-6}$
BHTA	Hypothalamus	Hypothalamus	0.30	1219	$<1 \times 10^{-6}$
BLIB	Liver	Liver	0.35	3359	$<1 \times 10^{-6}$
BPBA	Posterior brain stem	Medulla oblongata	0.27	3302	$<1 \times 10^{-6}$
BPLA	Placenta	Placenta	0.01	392	0.60
BPMA	Ficoll-purified peripheral blood mononuclear cells	Peripheral blood $CD4^+$ T cells	0.31	4977	$<1 \times 10^{-6}$
BMNA	Mesenteric lymph node	Lymph node	0.23	4347	$<1 \times 10^{-6}$
BPNA	Prescapular lymph node				
BPPA	Peyer's patch				
BPSA	Paratoid salivary	Salivary gland	0.12	552	$<1 \times 10^{-6}$
BSMA	Skeletal muscle	Skeletal muscle	0.47	3577	$<1 \times 10^{-6}$
BTMA	Thymus	Thymus	0.28	1335	$<1 \times 10^{-6}$
BTNA	Tonsil	Tonsil	0.30	1893	$<1 \times 10^{-6}$
BTSA	Testes	Testes	0.07	1071	3.2×10^{-5}
BTYA	Thyroid	Thyroid	0.01	420	0.41
Mean			0.22	2523	

*Number of bovine expressed sequence tags (ESTs) from the tissue used in the correlation analysis (adjusted library depth).

Contig assembly

Single-pass sequences were vector trimmed, quality clipped and masked by Genesis Biotechnology. Polymerase chain reaction analysis indicated the mean insert length of the ESTs to be 1265 bp and the median insert length to be 1100 bp, although this is likely to be an underestimate due to preferential amplification of small products. For the contig assembly, all EST sequences were cleaned by re-checking for possible vector contamination; trimming polyA tails and removing short and low complexity sequences. This was carried out using the TIGR SeqClean utility (available at http://www.tigr.org/tdb/tgi/software/). A total of 690,672 public and AgResearch bovine sequences were then assembled into contigs. The contig assembly used a divide-and-conquer strategy where an initial coarse partitioning of masked ESTs was performed to generate correct and feasible assembly sets of no more than 50,000 ESTs (Otu and Sayood, 2003). A simple graph-based clustering algorithm was used to generate the assembly sets in which each EST is a node and ESTs are joined by an edge if a BLAST alignment exists between them with an *e* value $\leq 1 \times 10^{-10}$. This value was chosen to ensure that pairs of ESTs with a 40-bp overlap of 80% identity (the relevant CAP3 parameters used for the assembly) would always be connected. Assembly sets correspond to connected components of the graph. Our initial partitioning generated one infeasible assembly set. This set was partitioned into smaller sets, using the MCL graph clustering algorithm (Enright et al., 2002). The parameters used for the MCL re-partitioning were -I 2 -adapt. For each assembly set, the unmasked sequences were assembled into contigs using CAP3 and a non-default value for the -y parameter of 40 (Huang and Madan, 1999). In order to ascertain the quality of the contigs, the coverage of protein coding sequence was examined by looking at those contigs with identity alignment ($e < 1 \times 10^{-130}$) to validated bovine RefSeqs. It was found that 46% of the contigs contained the entire protein sequence, while 70% of the contigs contained 50% or more of the protein sequence. Only 7% of the contigs contained untranslated sequence alone. The contigs clustered were also mapped to the bovine genome build Btau 4 using GMAP. Less than 1% of these contigs had an unexpected long alignment to the genome (>1000 bp), indicating they may contain some genomic DNA contamination. This contamination was not considered to be significant. The contig generation process and quality assessment is summarized in Table 2. All AgResearch bovine sequences generated are publicly available at NCBI and have accession numbers DY037420 - DY223196 and DY588300 - DY588367.

Table 2. Summary of contig generation process and quality control.

Total number of ESTs	740,181
Number of AgResearch ESTs	202,577
Number of AgResearch ESTs from non-normalized libraries	161,503
ESTs remaining after quality control	690,672
Assembly sets after BLAST	37,570
Singletons after BLAST	57,810
Contigs generated by CAP3	25,765
Total number of singletons	101,155
Number of contigs with ≥5 AgResearch ESTs	5812
Contigs containing 5' UTR alone	0.2%
Contigs containing 3' UTR alone	6.8%
Contigs covering up to 25% of CDS	8.4%
Contigs covering 25-50% of CDS	14.6%
Contigs covering 50-75% of CDS	12.4%
Contigs covering 75-100% of CDS	11.7%
Contigs covering 100% of CDS	45.8%

ESTs = expressed sequence tags; UTR = untranslated region; CDS = conding sequence.

Clustering of contigs

Contigs were clustered using a procedure based on the method described in Ewing et al. (1999). Only contigs containing at least 5 AgResearch ESTs were clustered, as the expression pattern of the ESTs belonging to these contigs was known in a variety of tissues and as they were derived from non-normalized libraries. The sequence of all 5812 contigs clustered is available in Supplementary file 2. Expression profiles were generated for each of these contigs by calculating the proportion of ESTs in a contig that were derived from each of the libraries (using the counts given in Supplementary file 3). Contigs were hierarchically clustered using Ward's minimum distance method (Ward, 1963). With this method, the distance between two clusters is the increase in the within cluster sum of squares (summed over contigs and libraries). This was expressed relative to the total sum of squares, giving the change in R^2 for joining the two clusters concerned. The hierarchical process was "cut" into groups at an arbitrary change in R^2 of 0.05 to investigate whether the clustering could give a reasonable description of known biological processes. The clusterings within each of these groupings were displayed as dendrograms, and as heat plots of correlations between the library proportions, while the data used (proportion of the contig's expression in each library) were also expressed as a heat plot.

Clustering of libraries

The distance between two libraries was taken to be one minus the correlation between the contig counts for those two libraries. These were used to hierarchically cluster the libraries using Ward's minimum distance method. The clustering was displayed as a dendrogram and as a heat plot of the correlations.

Annotation of contigs

Contigs were annotated with their top human RefSeq hit using BLASTN and an *e* value cut-off of 1×10^{-6}. In total 5188 (89%) of the bovine contigs available for clustering could be annotated with a human RefSeq. For contigs that matched more than one RefSeq with equivalent *e* value and bitscore (usually transcript variants of a gene), then the variant for which human expression data were available in UCSC Gene Sorter GNF Atlas 2 (Su et al., 2004) was chosen.

Comparison of bovine and human expression

Human expression data was available for 4325 of the 5188 human RefSeqs, which corresponded to the bovine contigs, i.e., 74% of the original 5812 bovine contigs. Relative gene expression levels in 79 human tissues were extracted from the GNF Atlas 2 array dataset (Su et al., 2004) in the UCSC Gene Sorter. In total, there were 19 tissues in common between the human and the bovine dataset (Table 1). The bovine peripheral blood mononuclear cell library (BPMA) was initially compared to both the human peripheral blood $CD4^+$ and $CD8^+$ T cell tissues. Gene expression in the bovine library was significantly correlated with gene expression in both $CD4^+$ and $CD8^+$ human tissues ($R = 0.31$ and 0.30, respectively); however, for all further analysis the human $CD4^+$ data alone was used. In order to compare gene expression in human and bovine ovary, lymph node

and cerebellum the BCNA, BCOA, BINA, BIOA, BOVA, and BOVB libraries were combined into a single bovine ovary dataset, the BMNA, BPNA, and BPPA libraries were combined into a single bovine lymph node dataset and the BCEA, and BCEB libraries were combined into a single bovine cerebellum dataset. As contigs that contained more ESTs had more accurate expression measurements, correlations were weighted by the number of ESTs in the contig in the 19 tissues compared. Contig counts were normalized by dividing by the number of contigs in that tissue, and then converting to a proportion (dividing by the sum of normalized values). Logged values ($\log_2(x + 0.001)$) were correlated against the Atlas GNF human expression values.

Gene Ontology

Gene Ontology Biological Process (GO-BP) terms significantly associated with genes were found using the Expression Analysis Systematic Explorer (EASE) (Hosack et al., 2003).

RESULTS

Over 200,000 bovine ESTs were generated from a variety of tissues. These data were used to generate clusters of contigs, where members of the same cluster had similar expression patterns across many tissues. In total, 5812 unique bovine contigs were clustered. This is approximately 23% of all contigs containing two or more ESTs, and thus represents a subset of highly expressed genes. The contigs used in the expression clustering varied in length from 179 bp to 13,430 bp with a mean contig length of 1725 bp and a median contig length of 1558 bp. The contig depth varied from 5 to 1873 with a mean contig depth of 18 and a median of 9. Figure 1 shows the number of ESTs per contig, for the contigs clustered. The contigs were grouped into 33 common expression clusters (Figure 2 and Supplementary file 3) with a median number of contigs per cluster of 148. The largest cluster, CL33, contained 973 contigs and these contigs were expressed in a variety of tissues and most likely represent genes that are widely expressed, i.e., housekeeping genes. The smallest cluster, CL292, contained only 17 contigs, and the expression of these contigs was primarily restricted to the seminal vesicle. Expression heat plots for each cluster are available in Supplementary file 4, and correlation heat plots and dendrograms of the clusters are available in Supplementary file 5.

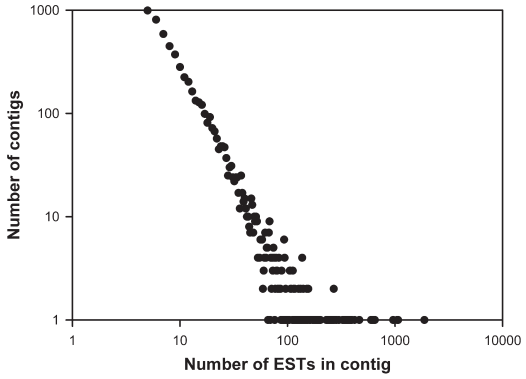

Figure 1. Scatter plot showing the number of contigs plotted against the number of expressed sequence tags (ESTs) in that contig for each of the contigs clustered. Both axes are plotted on a $\log_{10}$ scale.

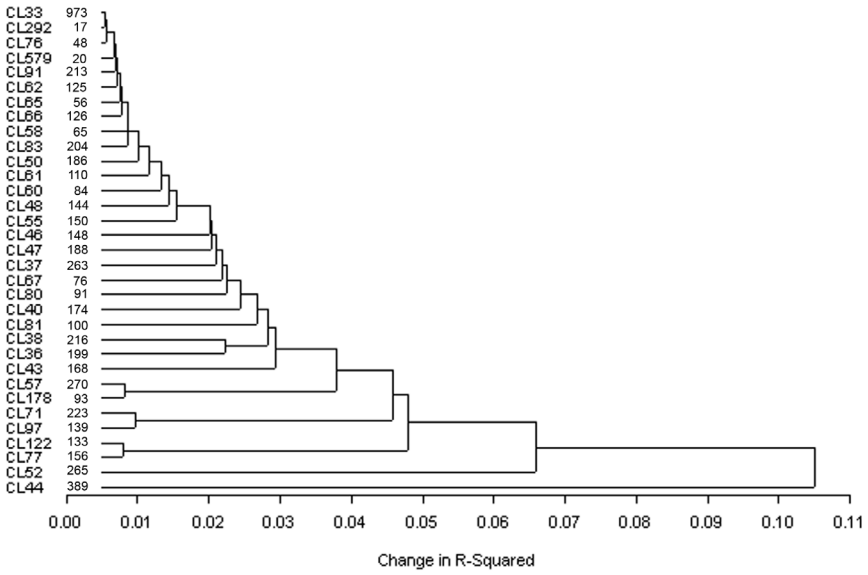

Figure 2. Hierarchial cluster analysis of the bovine contigs. The dendrogram shows the relationships between the clusters. The number after each cluster name represents the number of contigs in the cluster. The position of each contig within the cluster can be found in Supplementary file 3.

Clustering of ribosomal proteins

In order to demonstrate that the clustering was successful at grouping genes involved in a common biological process, we examined the clustering of genes encoding ribosomal proteins (RP). Ribosomal proteins are required in all cell types in equimolar amounts (Perry, 2005), and as such are expected to have similar expression patterns. Our dataset contained 89 bovine contigs (1.7% of the annotated contigs) corresponding to 77 unique RP-encoding genes. There was a significant association ($P = 1.4 \times 10^{-21}$) between cluster and RP-encoding annotation with many of the RP-encoding genes (39%) co-clustered in CL91. This indicates that the clustering is successful at grouping genes with similar expression pattern and function. However, not all RP-encoding genes were found in CL91. This may be due to inherent noise in the system or to the fact that much of the control of these proteins occurs at the translational and not transcriptional level (Meyuhas, 2000). Additionally, it has been reported that certain RPs are differentially expressed between different tissues (Bortoluzzi et al., 2001). Such RPs would not cluster in a single group.

Cluster function and annotation of genes of unknown function

With the exception of contigs in CL33, contigs that co-clustered were predominately expressed in a single library or in libraries from related tissues. This is illustrated for CL122 (Figure 3A) whose members are almost exclusively expressed in liver, CL77 (Figure 3B) whose members are primarily expressed in liver but are also expressed in other tissues, and

CL47 (Figure 3C) whose members are primarily expressed in longissimus dorsi skeletal muscle but are also expressed in the related tissues such as biceps femoris muscle and fetal muscle.

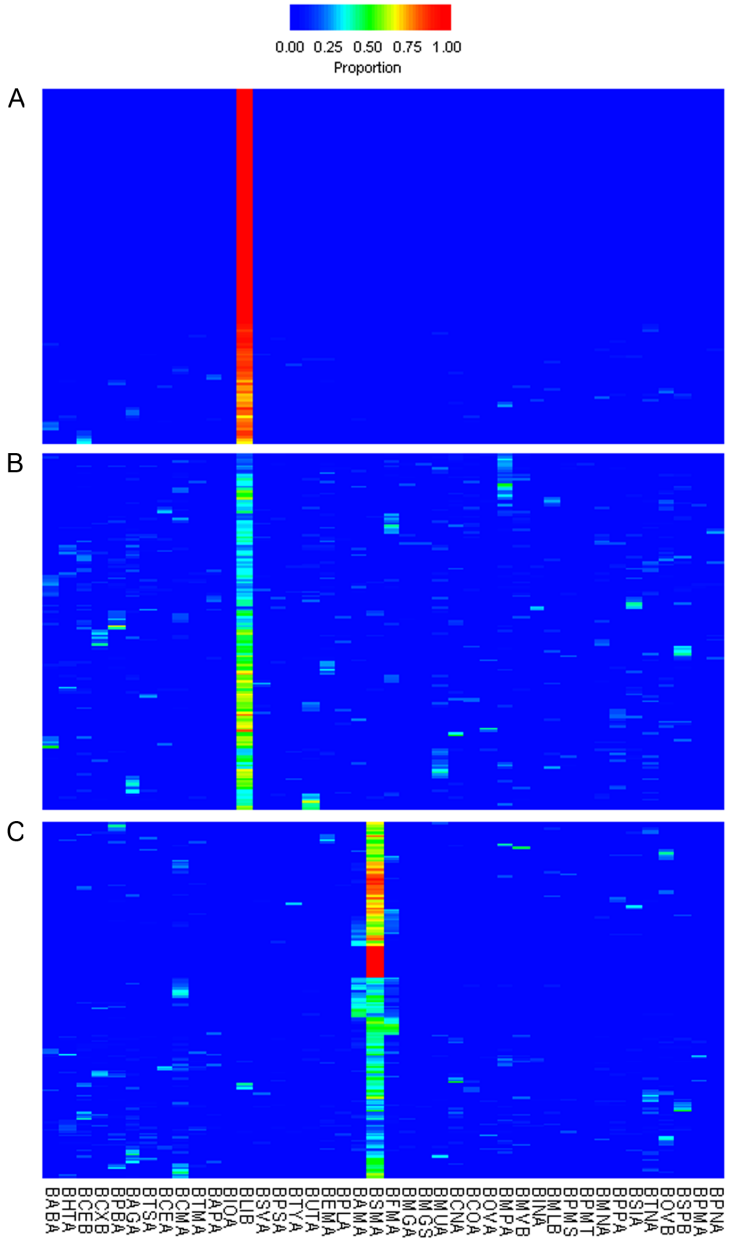

Figure 3. Expression heat plots of the CL122 (A), CL77 (B), and CL47 (C) clusters. The bovine libraries are plotted on the x-axis and contigs are plotted on the y-axis. For each contig the proportion of ESTs in each of the libraries is plotted. The contigs within each cluster are in the same order from top to bottom as they appear in Supplementary file 3.

The human ortholog of each bovine contig was found using BLAST (Altschul et al., 1990). The GO-BP terms associated with these orthologs were determined using EASE (Hosack et al., 2003). Table 3 shows the GO-BP terms significantly associated with genes from the three clusters shown in Figure 3. GO-BP terms associated with genes from the CL122 cluster (liver; Figure 3A) refer to the secretion of blood coagulation factors and maintenance of homeostasis, while GO-BP terms associated with the genes from the CL77 cluster (liver; Figure 3B) relate to many different types of metabolism. GO-BP terms associated with the genes from the CL47 cluster (muscle; Figure 3C) relate to striated muscle contraction and development and metabolism. Clustering based on gene expression therefore grouped genes with similar biological functions and successfully separated the liver-specific function of release of blood coagulation factors from the more general functions of the liver such as metabolism. The GO-BP terms significantly associated with the genes in each cluster are available in Supplementary file 6.

Table 3. Gene Ontology Biological Process (GO-BP) terms associated with genes from clusters CL122, CL77 and CL47.

Cluster	GO-BP term	No. of genes	P*
CL122			
	Regulation of body fluids	10	9.1×10^{-11}
	Blood coagulation	9	1.1×10^{-9}
	Hemostasis	9	1.7×10^{-9}
	Organismal physiological process	21	2.1×10^{-7}
	Regulation of blood pressure	4	2.0×10^{-6}
	Homeostasis	6	1.7×10^{-5}
	Response to external stimulus	17	9.3×10^{-5}
	Regulation of physiological process	4	1.2×10^{-4}
CL77			
	Nucleotide-sugar metabolism	3	5.5×10^{-4}
	Organic acid metabolism	11	5.7×10^{-4}
	Carboxylic acid metabolism	11	5.7×10^{-4}
	Lipid metabolism	12	3.0×10^{-3}
	Complement activation	4	0.0044
	Steroid metabolism	5	0.0046
	Protein secretion	3	0.0057
	Alcohol metabolism	7	0.0065
CL47			
	Striated muscle contraction	8	6×10^{-8}
	Muscle contraction	11	3.6×10^{-7}
	Muscle development	11	9.4×10^{-7}
	Hexose metabolism	8	3.3×10^{-5}
	Monosaccharide metabolism	8	3.3×10^{-5}
	Glucose metabolism	7	9.3×10^{-5}
	Regulation of striated muscle contraction	3	2.1×10^{-4}
	Glycolysis	5	3.1×10^{-4}

The most significant GO-BP terms are listed with limits of $P < 0.05$ and eight terms per cluster. *Fisher exact probability to test for over-represented GO terms in the cluster compared to all genes clustered.

In order to independently validate the results obtained using GO-BP terms, biochemical pathways containing the genes from the CL122, CL77, and CL47 clusters were found using Biorag (Pandey et al., 2004). The pathway containing the most genes from CL122 was for complement and coagulation cascade (12 RefSeqs, 11 genes), while the pathways containing

the most genes from CL77 and CL47 were those for ribosome (9 RefSeqs, 9 genes) and striated muscle contraction (19 RefSeqs, 11 genes), respectively. This confirms the results obtained using GO terms that proteins encoded by genes in the CL122 cluster are likely to be involved in blood coagulation, and that proteins encoded by genes in the CL77 cluster are likely to be involved in metabolism, while proteins encoded by genes in the CL47 cluster are likely to be involved in muscle contraction. The usefulness of this method of annotation can be demonstrated using contig CS3400474000001 from CL122 as an example. This contig is a 968-bp contig made up of 27 constituent ESTs. The majority of the ESTs are expressed in liver tissue with some found in heart and brain tissue. The contig corresponds to the bovine brain protein 44-like gene using BLAST and represents a near full length BRP44L transcript including all coding sequences. BRP44L is highly conserved across the animal kingdom but is very poorly annotated. It does not share significant homology with any well-annotated gene in the Genbank database and despite its name appears to be most highly expressed in liver and heart tissue in both *Bos Taurus* and humans. The contig shows good homology with an uncharacterized pfam domain and very weak homology to the proprotein convertase P domain of the subtilisin-like proprotein convertases. However, as it appears to lack the subtilisin-like catalytic domain, it is unlikely to function as a proprotein convertase. Our expression analysis indicates that this transcript may represent a gene involved in the secretion of blood coagulation factors and maintenance of homeostasis.

Within CL47, four sub-clusters could be observed (Figure 4). Sub-cluster 1 could be further split into 1A and 1B based on gene function. The genes in each sub-cluster code for proteins involved in different biological processes. The majority of genes encoding proteins involved in muscle contraction are found in sub-cluster 1B, while the majority of genes encoding proteins involved in metabolism are in sub-cluster 4. Sub-cluster 1B contains 38 unique annotated genes and includes many genes whose products are known to play a role in muscle architecture and contraction, such as myosin, tropomyosin, actin, troponin, myomesin, α-actinin, and dystrobrevin (Laing and Nowak, 2005). This cluster also contains a small number of genes such as *ZAK* and *ART1*, not previously known to encode proteins with a role in muscle contraction, indicating that they may play a previously unrecognized role in muscle tissue. The genes of unknown function in sub-cluster 1B of CL47 could be annotated as genes encoding proteins potentially involved in striated muscle contraction. Similarly, the genes of unknown function in sub-cluster 4 of CL47 could be annotated as genes encoding proteins putatively involved in muscle tissue metabolism.

Correlation of gene expression in human and bovine tissues

The correlation of gene expression between cattle and humans was calculated for each of 19 tissues in common (Table 1). Gene expression in 17 of the 19 tissues was significantly correlated. The correlation was highest in skeletal muscle ($R = 0.47$; $P < 1 \times 10^{-6}$). Gene expression was not correlated between human and bovine placenta and human and bovine thyroid.

It has previously been reported that there is a high correlation in expression of orthologous genes between mice and humans (Su et al., 2002). The expression pattern of each bovine contig and its corresponding human RefSeq was correlated across the 19 tissues in common between the two species (Supplementary file 7). A total of 267 of the genes examined were not expressed in the bovine tissues that were in common with human tissues, and so could not be

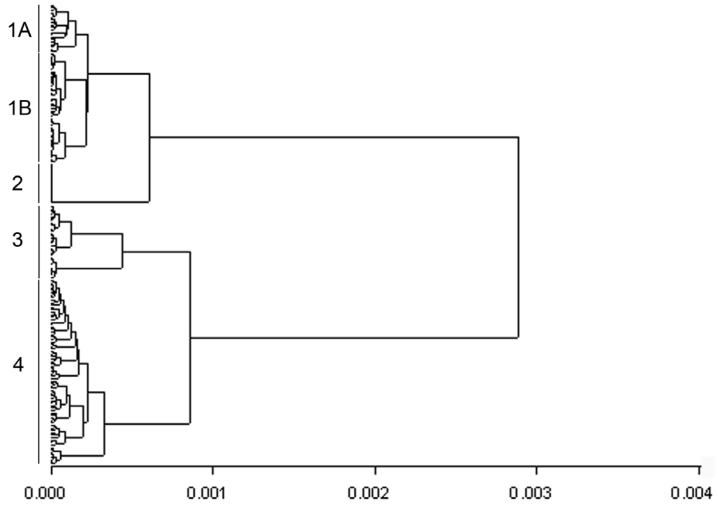

Figure 4. Dendrogram of the CL47 cluster. The cluster can be broken into four sub-clusters. The order of the contigs in the dendrogram, from top to bottom, is given in Supplementary file 3.

correlated, giving 4058 genes with both human and bovine expression measurements. Some bovine libraries were sequenced to a greater depth than others, providing a more comprehensive catalog of the genes expressed in this tissue. In order to adjust for this fact, the correlations were calculated by the adjusted library depth. Overall, the expression of orthologous genes was significantly correlated between humans and cattle (mean R = 0.09; test for zero mean gives $P = 9.8 \times 10^{-53}$). There was also a tendency for the contigs with more ESTs to have higher correlations (correlations increased by an average of 0.2 for each additional 100 ESTs; $P < 0.001$). In total, 747 (18.4%) individual genes showed significant ($P < 0.05$) positive correlation between their human and bovine expression patterns across the 19 tissues. A total of 199 (4.9%) genes have a significant ($P < 0.05$) negative correlation between their expression in human tissue and their expression in bovine tissues.

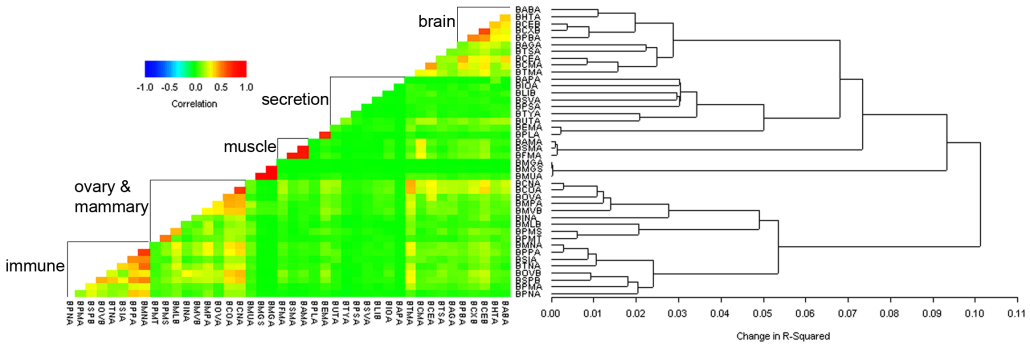

Figure 5. Correlation heat plot and dendrogram of the bovine EST libraries. Tissues with similar physiological function and which cluster together are noted.

Clustering of bovine libraries

In addition to clustering genes based on their tissue-specific expression pattern, tissues can also be clustered based on their gene expression pattern, revealing which tissues express similar genes. A previous study has shown that tissues with similar physiological functions or anatomical locations express similar genes (Shyamsundar et al., 2005). A correlation heat plot and dendrogram of the bovine libraries, based on the contigs they express, is shown in Figure 5. As can be seen in this figure, certain libraries group closely together, indicating similar gene expression in these libraries. Libraries from brain tissue clustered closely (anterior brain stem, hypothalamus, cerebellum, cortex, and posterior brain stem), as did libraries from secretory tissue (pituitary gland, liver, seminal vesicle, salivary gland, thyroid, and teat). Immune tissue (mesenteric lymph node, Peyer's patch, tonsil, spleen, peripheral blood and pre-scapular lymph node) co-clustered, while many of the ovary libraries grouped with mammary gland libraries. The three muscle libraries also grouped together (foetal and both skeletal muscle libraries), while the placenta library grouped with the embryonic library. As tissues with similar physiological functions express similar sets of genes, examining gene expression in a set of cells or in a tissue can help determine the function of the cells or tissue and their relationship to other tissues.

DISCUSSION

DNA sequences that are highly conserved across species are assumed to be constrained from mutating and to be functionally important. Genes with known functions are therefore commonly used to provide annotation for their sequence homologs, both in the same species and in other species. It has recently been shown that gene expression is largely conserved across species (Su et al., 2002; Khaitovich et al., 2005), indicating its functional importance. Consequently, annotation can be provided for genes of unknown function by comparing their expression pattern with the expression patterns of genes of known function. This method of gene annotation has the advantage that it can be used for both genes with known sequence homologs and those which lack homologs or are homologous to genes of unknown function.

Bovine EST data were used to determine gene expression profiles of almost 6000 transcripts across many tissues. ESTs were initially assembled into contigs and contigs clustered based on their expression pattern across all tissues. Human orthologs of 89% of the bovine contigs were found, but the annotation rate varied widely between clusters. The annotation rates of contigs in the seminal vesicle, embryonic tissue and thyroid clusters were particularly low (< 60%). These bovine tissues appear to express large numbers of novel genes that lack human orthologs or have diverged significantly at the sequence level from their human orthologs. The expression pattern of the novel genes can be used to provide some annotation for these genes.

GO-BP analysis was carried out on the human orthologs of the contigs in each cluster in order to assign biological functions to each cluster. The cross-species comparison may have introduced some noise into the data; however, the clustering process still successfully grouped genes whose products were involved in a coordinated biological process. There were two distinct clusters of genes primarily expressed in liver tissue. The GO-BP analysis showed that the genes in each cluster had distinct biological functions, thus showing that the clustering process can separate biological processes carried out in the same tissue based on their gene expression pattern. In some cases, clusters contained genes whose

products were involved in more than one biological process. This was shown for CL47 whose members encoded proteins involved in striated muscle contraction and metabolism. Sub-clustering CL47 successfully separated genes whose products were involved in striated muscle contraction from those whose products were involved in metabolism.

Human and bovine gene expression was compared across 19 tissues in common between the two species. Gene expression was significantly correlated in all tissues with the exception of placenta and thyroid. Bovine placenta differs markedly from human placenta (Liu et al., 2001). Human placenta is discoid or round, while bovine placenta is cotyledonary. Additionally, human placentas have a single large area of contact between maternal and fetal vascular systems, while in ruminants placental transfer occurs at specific predetermined uterine sites known as placentomes (Enders and Carter, 2004). These same sites are used for each pregnancy in cattle, while in humans a new placental attachment forms for each pregnancy. Therefore, it is unsurprising that gene expression in human and bovine placenta is not correlated. The reason gene expression in human and bovine thyroid is not correlated is unknown. It may reflect the metabolic differences between the species or the age or metabolic status of the thyroid tissues used to generate the data. Additionally, a human ortholog could only be found for 42% of the bovine contigs in the thyroid cluster (CL67) and only 24% of the orthologs had expression data. This may also contribute to the lack of correlation in gene expression between human and bovine thyroid tissue.

The expression of orthologous genes was also significantly correlated between humans and cattle overall, although only 747 individual genes were significantly ($P < 0.05$) positively correlated. There were also a small number of orthologous genes with a significant negative correlation in their expression pattern between humans and cattle. Interestingly, many of these genes were from ovary specific clusters (>20%). There are many possible reasons for this, such as the stage of the reproduction cycle at which the tissues were collected or the relative proportion of differentiated tissue types within the samples (e.g., presence of a prominent corpus luteum). Negative correlation in the expression of orthologous genes has been reported previously (Su et al., 2002).

In this study, we showed a highly significant correlation in gene expression between bovine and human tissue. However, the correlation coefficient *per se* is not particularly high. This may be due to many factors besides differences in gene expression between the two species. Different methods of measuring transcripts were used in the two species, and this can lead to considerable variability in the gene expression measurement. The precise amount of variability introduced into the system by this is impossible to quantify, but it is likely to be considerable, as gene expression in identical RNA preparations measured on different microarray platforms was found to have a correlation coefficient of only 0.53 (Tan et al., 2003). Other factors such as differences in the age or metabolic state of the tissues utilized will also result in variability in the set of expressed genes (Bahar et al., 2006; Li et al., 2006). Additionally, there is significant variation in gene expression between individuals of the same species within a tissue (Whitehead and Crawford, 2005). The fact that the expression levels were correlated to some degree, despite all these factors, indicates that the bovine EST counts provide useful information for expression.

In the present study, we were only able to examine gene expression in a subset of highly expressed genes as the limit of detection of a transcript using cDNA library sequencing is low. The limit of detection is also dependent on the depth of sequencing of the library.

The clusters generated are also in a sense arbitrary due to the choices of clustering method, cut-off for the minimum number of ESTs (with tissue expression information) in a contig and the point at which the dendrogram was divided into clusters (0.05 change in R^2). For these reasons, the robustness of an annotation should be checked by examining results using other choices of these methods and by examining the sub-structure of the clusters. It is also of note that there was a lack of data from a number of important organs such as the rumen, skin, lung, adipose tissue, kidney, and pancreas. Therefore, we examined here the expression of only a small number of highly expressed genes in a selection of tissues. Despite these caveats, the study created the first gene atlas of almost 6000 bovine transcripts in 27 different tissues. Expression clustering of these genes successfully classified the genes according to biological function. We also successfully demonstrated that cross-species comparisons can be used to provide annotation for contigs. This will allow researchers to study the transcriptomes of species for which few publicly available annotation resources are available.

ACKNOWLEDGMENTS

We thank AgResearch staff for collecting the tissues used to generate the EST libraries and Genesis Biotechnology for creating and sequencing the cDNA clones. Research supported by New Zealand Foundation for Research, Science and Technology.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Audic S and Claverie JM (1997). The significance of digital gene expression profiles. *Genome Res.* 7: 986-995.

Bahar R, Hartmann CH, Rodriguez KA, Denny AD, et al. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* 441: 1011-1014.

Boguski MS, Lowe TM and Tolstoshev CM (1993). dbEST - database for "expressed sequence tags". *Nat. Genet.* 4: 332-333.

Bortoluzzi S, d'Alessi F, Romualdi C and Danieli GA (2001). Differential expression of genes coding for ribosomal proteins in different human tissues. *Bioinformatics* 17: 1152-1157.

Enders AC and Carter AM (2004). What can comparative studies of placental structure tell us? - A review. *Placenta* 25 (Suppl A): S3-S9.

Enright AJ, Van Dongen S and Ouzounis CA (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30: 1575-1584.

Ewing B and Green P (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* 25: 232-234.

Ewing RM, Ben KA, Poirot O, Lopez F, et al. (1999). Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* 9: 950-959.

Gupta S, Zink D, Korn B, Vingron M, et al. (2004). Genome wide identification and classification of alternative splicing based on EST data. *Bioinformatics* 20: 2579-2585.

Hosack DA, Dennis G Jr, Sherman BT, Lane HC, et al. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4: R70.

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

Irizarry K, Kustanovich V, Li C, Brown N, et al. (2000). Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* 26: 233-236.

Khaitovich P, Hellmann I, Enard W, Nowick K, et al. (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309: 1850-1854.

Kim SK, Lund J, Kiraly M, Duke K, et al. (2001). A gene expression map for *Caenorhabditis elegans*. *Science* 293: 2087-2092.

Laing NG and Nowak KJ (2005). When contractile proteins go bad: the sarcomere and skeletal muscle disease. *Bioessays* 27: 809-822.

Lee MA, Keane OM, Glass BC, Manley TR, et al. (2006). Establishment of a pipeline to analyse non-synonymous SNPs in *Bos taurus*. *BMC Genomics* 7: 298.

Leung HC and Chin FY (2006). Finding motifs from all sequences with and without binding sites. *Bioinformatics* 22: 2217-2223.

Li RY, Zhang QH, Liu Z, Qiao J, et al. (2006). Effect of short-term and long-term fasting on transcriptional regulation of metabolic genes in rat tissues. *Biochem. Biophys. Res. Commun.* 344: 562-570.

Liang F, Holt I, Pertea G, Karamycheva S, et al. (2000). An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.* 28: 3657-3665.

Liu FG, Miyamoto MM, Freire NP, Ong PQ, et al. (2001). Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291: 1786-1789.

Meyuhas O (2000). Synthesis of the translational apparatus is regulated at the translational level. *Eur. J. Biochem.* 267: 6321-6330.

Otu HH and Sayood K (2003). A divide-and-conquer approach to fragment assembly. *Bioinformatics* 19: 22-29.

Pandey R, Guru RK and Mount DW (2004). Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics* 20: 2156-2158.

Perry RP (2005). The architecture of mammalian ribosomal protein promoters. *BMC Evol. Biol.* 5: 15.

Shyamsundar R, Kim YH, Higgins JP, Montgomery K, et al. (2005). A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.* 6: R22.

Su AI, Cooke MP, Ching KA, Hakak Y, et al. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 99: 4465-4470.

Su AI, Wiltshire T, Batalov S, Lapp H, et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101: 6062-6067.

Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, et al. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 31: 5676-5684.

Ward JH (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58: 236-244.

Whitehead A and Crawford DL (2005). Variation in tissue-specific gene expression among natural populations. *Genome Biol.* 6: R13.

Wu LF, Hughes TR, Davierwala AP, Robinson MD, et al. (2002). Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31: 255-265.

Zadissa A, McEwan JC and Brown CM (2007). Inference of transcriptional regulation using gene expression data from the bovine and human genomes. *BMC Genomics* 8: 265.

Zhang W, Morris QD, Chang R, Shai O, et al. (2004). The functional landscape of mouse gene expression. *J. Biol.* 3: 21.