Mammalian Genome 13, 373–379 (2002).
DOI: 10.1007/s00335-001-2145-4

Mammalian Genome
Incorporating Mouse Genome

Analysis of bovine mammary gland EST and functional annotation of the *Bos taurus* gene index

Tad S. Sonstegard,[1] Anthony V. Capuco,[1] Joseph White,[2] Curtis P. Van Tassell,[1] Erin E. Connor,[1] Jennifer Cho,[2] Razvan Sultana,[2] Larry Shade,[1] James E. Wray,[3] Kevin D. Wells,[1] John Quackenbush[2]

[1]USDA, ARS, Beltsville Agricultural Research Center, Beltsville, Maryland 20705, USA
[2]The Institute for Genomic Research, Rockville, Maryland 20850, USA
[3]USDA, ARS, U.S. Meat Animal Research Center, Clay Center, Nebraska 68933, USA

Received: 12 November 2001 / Accepted: 13 March 2002

Abstract. Functional genomic studies of the mammary gland require an appropriate collection of cDNA sequences to assess gene expression patterns from the different developmental and operational states of underlying cell types. To better capture the range of gene expression, a normalized cDNA library was constructed from pooled bovine mammary tissues, and 23,202 expressed sequence tags (EST) were produced and deposited into GenBank. Assembly of these EST with sequences in the *Bos taurus* Gene Index (BtGI) helped to form 5751 of the current 23,883 tentative consensus (TC) sequences. The majority (87%) of these 5751 assemblies contained only one to three mammary-derived EST. In contrast, 18% of the mammary EST assembled with TC sequences corresponding to 12 genes. These results suggest library normalization was only partially effective, because the reduction in EST for genes abundantly transcribed during lactation could be attributed to pooling. For better assessment of novel content in the mammary library and to add to existing annotation of all bovine sequence elements, gene ontology assignments, and comparative sequence analyses against human genome sequence, human and rodent gene indices, and an index of orthologous alignments of genes across eukaryotes (TOGA) were performed, and results were added to existing BtGI annotation. Over 35,000 of the bovine elements significantly matched human genome sequence, and the positions of some alignments (3%) were unique relative to those using human expressed sequences. Because 3445 TC sequences had no significant match with any data set, mammary-derived cDNA clones representing 23 of these elements were analyzed further for expression and novelty. Only one clone met criteria suggesting the corresponding gene was a divergent ortholog or expressed sequence unique to cattle. These results demonstrate that bovine sequence expression data serve as a resource for characterizing mammalian transcriptomes and identifying those genes potentially unique to ruminants.

Introduction

One goal of genomics research is to characterize expressed portions of a genome. Attempts to identify and annotate the human transcriptome have included performing various *in silico* analyses on the genome draft sequences (IHGSC 2001; Venter et al. 2001; Wright et al. 2001), assemblies of expressed sequence tags (EST; Liang et al. 2000), and comparative alignments of genomic and expressed sequences (Zhuo et al. 2001). The total number of genes identified has varied considerably among methods, thus impeding accurate identification and annotation of the complete transcriptome. This variation is due in part to incomplete data sets, software limitations, and contaminating sequence. Further sequencing of animal genomes, EST, and full-length (open reading frame) cDNA may provide additional evidence to better characterize the transcriptomes of humans and other biologically important species. For example, alignment of 61,227 mouse cDNAs to the human genome draft sequences identified 1141 aligned sequences that did not match any known protein, human gene, or EST (Kondo et al. 2001)

Even though livestock EST will facilitate functional genomic studies in animal science and increase the resolution of comparative maps that denote the general conservation of synteny between mammalian species, this sequence information may also be useful in identification of expressed genes from certain tissues. For example, ethical considerations have limited the feasibility of constructing cDNA libraries from high-quality human mammary tissues that represent the different developmental and operational states of this gland. Only 2500 human EST have been generated from libraries constructed from *normal* mammary tissue (http://www.ncbi.nlm.nih.gov/UniGene/). Although more than 224,000 mammary-derived EST have been generated from rodent cDNA libraries (http://www.ncbi.nlm.nih.gov/UniGene/), use of this sequence information for comparative purposes is limited by obvious physiological differences in mammary development between humans and rodents (Capuco et al. 2002). In contrast, a catalog of mammary-derived sequences from cattle could provide researchers with a more comparable sequence resource to aid annotation and functional genomic studies of genes expressed in this gland.

Previous efforts to generate bovine EST from mammary gland cDNA libraries were limited, in part, by the amplified expression of a few genes expressed in support of lactation. For example, a subtracted cDNA library made only from lactating gland produced EST corresponding to casein at a rate of 34% (J. Byatt, personal communication). A more recent effort, which produced the majority of bovine EST in the public domain, maximized gene sequence discovery by constructing normalized cDNA libraries from pooled tissues (Smith et al. 2001). However, none of the tissues used to construct these libraries was derived from mammary gland,

Correspondence to: T.S. Sonstegard; *e-mail:* tads@anri.barc.usda.gov

The nucleotide sequence data reported in this paper are in GenBank with accession numbers: BE476018-488146, BE588269-590261, BE845406-846390, BF230108-231509, BF868284-868340, BG687694-694329, BM878931-879001.

Table 1. Construction of a pooled tissue, bovine mammary gland cDNA library. The mRNA extracted from each of eight indicated tissue stages was pooled (12.5% mRNA pool/tissue stage). The 18 mammary tissue samples were recovered from Holstein heifers approximately 3 months of age (1–6), pregnant heifers 18–20 months in age (7–8), or mature Holstein cows (9–18). Success of coliform challenge by somatic cell count (SCC).

Tissue type	Physiological stage of mammary of gland	Specific treatment or phenotype	Collection animal	% of tissue stage for mRNA extraction
Mammary fat pad	Pre-pubertal	Estradiol	1	8
		Estradiol	2	10
		Control	3	14
		Estradiol	4	11
		Estradiol	5	11
		Control	6	46
Mammary epithelium	Pre-pubertal	Estradiol	1	67
		Control	3	33
Mammary gland	Mid-gestation (1st parity)	200 days gestation	7	48
		268 days gestation	8	52
Mammary gland	Late gestation	5 days prepartum	9	61
		5 days prepartum	10	39
Mammary gland	Peak lactation (w/coliform mastitis)	82 days (41 kg milk/d & SCC = 3.5×10^6/ml)	11	64
		65 days (38 kg milk/d & SCC = 3.2×10^6/ml)	12	36
Mammary gland	Peak lactation	89 days (43 kg milk/d)	13	51
		65 days (44 kg milk/d)	14	49
Mammary gland	Non-lactating (pregnant)	8 days post-lactation	15	34
		8 days post-lactation	16	66
Mammary gland	Non-lactating (pregnant)	29 days post-lactation	17	42
		30 days post-lactation	18	58

even though production from this organ has arguably the largest economic impact on beef and dairy production.

On the basis of these observations, we constructed and sequenced a normalized cDNA library, using mRNA extracted from eight different mammary tissue types or stages. The pooling and normalization were performed to more thoroughly sample gene expression during critical periods of mammary development and function, while decreasing redundancy for EST associated with lactoproteins. After sequencing, EST were integrated with the existing tentative consensus (TC) sequence assemblies of the *Bos taurus* Gene Index (BtGI; http://www.tigr.org/tdb/btgi/) developed by The Institute for Genomic Research (TIGR) as an interactive web site. Functional classification was performed on matching sequence elements by gene ontology assignments. To provide additional annotation, extensive BLAST analyses (Altshul et al. 1990) of all BtGI sequence elements were performed against assembled sequence data from the human genome and other animal gene indices. Sequence elements with no similarity were investigated further to provide support of expression and novelty relative to other species.

Materials and methods

Library construction and normalization. Mammary tissues were sampled from 18 purebred Holstein females raised at USDA, ARS, Beltsville Agricultural Research Center (BARC; Table 1). To increase pre-pubertal mammary growth, calves were injected subcutaneously with estradiol-17β (Sigma, St. Louis, Mo.) in corn oil (4 mg/ml) at a dosage of 0.1 mg/kg of body weight. After three daily injections (approximately 72 h after the first treatment), calves were killed at the BARC abattoir, and mammary tissue was recovered. For pregnant heifers and mature cows, mammary biopsies were taken from the rear quarters as described (Farr et al. 1996). Coliform mastitis was induced by infusion of pathogenic *E. coli* as described by Long et al. (2001), and mammary tissue was obtained by biopsy 24 h after bacterial infusion. All tissues were immediately frozen in liquid nitrogen and weighed. Owing to the relatively small mass of the mammary biopsies, tissues from individual animals were pooled within the physiological stage (Table 1). Total RNA extraction, mRNA purification, primary library construction, and normalization were purchased as a service from the Gene Discovery Services division of Invitrogen (Rockville, Md.), which used commercially available protocols and materials. After extraction and purification, 5 μg of poly-A-selected mRNA from each stage was pooled in equimolar amounts (40 μg total mRNA) prior to cDNA synthesis. Unidirectional cloning of cDNA into pCMV-SPORT6, library amplification, and normalization to C_{ot500} were performed as described by Smith and colleagues (2001).

At Invitrogen, normalization was estimated by independent hybridizations to colony lifts from a plated aliquot of library by using either a labeled probe for *elongation factor 1 alpha* (EF1α, previously determined to be among the highest copy number in other bovine libraries prior to normalization) or *alpha-S1-casein* probe (CSN1S1, determined to be the highest copy number by sequencing 89 clones from this library prior to normalization). Reduction in redundancy was calculated by using the ratio of the percentage of clones detected by hybridization before and after normalization. Average insert length of 1.55 kb was estimated by sizing PCR-generated products from 48 clones with standard M13 vector primers.

EST sequencing. The normalized library (BARC 5BOV) contained approximately 3.2×10^7 total transformants. After plating aliquots of the library, 55,296 individual colonies were picked and arrayed into 144 384-well plates by BACPAC resources (Oakland, Calif.). BACPAC maintains a master copy of the library for distribution of clones and library copies, which are freely available upon request. In total, 94 384-well plates were processed for PCR-based 5′-end single-pass sequencing as described by Smith et al. (2000). Sequencing of individual cDNA clones was performed on a CEQ2000XL from Beckman Coulter (Fullerton, Calif.) with plasmid DNA and DTCS Quick Start chemistry (Beckman Coulter). A relational database and automated dataflow were established to process and store the raw sequence data (MARCDB; J. Keele, unpublished). Sequences with 100 or more bases after trimming and vector removal were submitted to the GenBank dbEST database. Sequence quality assessment and trimming were performed using alt_trim option with phred v0.980904.e. Vector was identified and trimmed by using cross_match with the -minscore 18 and -minmatch 12 options.

EST assembly and functional annotation. Assembly of BARC EST and re-assembly of the BtGI were performed as described (Quackenbush et al. 2000). Assembly of component sequences for each cluster with CAP3 (Huang and Madan 1999) produced the TC sequence assemblies that comprise the BtGI. For annotation, bovine TC sequences containing a known gene were assigned the function of that gene, while TC sequences with significant similarity to orthologous sequence were assigned a putative function. Functional classification based on gene ontology assignments was performed by merging the GO.pep data set with TIGR's non-redundant amino acid (nraa) database and by conducting searches with DPS. Thresholds for BLAST analyses of BtGI vs. the human (HGI), mouse (MGI) and rat gene indices (RGI) at TIGR (http://www.tigr.org/tdb/tgi.shtml) were 70% sequence identity match (ID) with 20-bp gap and 70% ID with 30% EST coverage for comparisons with the "Golden Path" human genome sequence assembly (http://genome.ucsc.edu/).

Evaluation of mRNA representation. High-density grids of the BARC 5BOV library, prepared by spotting 18,432 arrayed clones in duplicate onto three 21 × 21-cm membranes, were purchased from BACPAC resources. Duplicate sets of these membranes were hybridized with labeled cDNA probes. These probes were synthesized by using an Ambion Strip-EZ kit (Austin, Tex) and aliquots of mRNA left over from library construction. Hybridization patterns were visualized on a STORM 860 phosphorimager (Amersham Biosciences, Sunnyvale, Calif.) and assessed with Array Vision v6.0 software developed by Imaging Research Inc. (Ontario, Canada).

Northern analysis. Total RNA samples from mammary tissues of collection animals 1, 3, 7, and 8 (Table 1) were extracted with Trizol according to the protocol supplied by Invitrogen. RNA concentration was determined spectrophotometrically. Gel electrophoresis, blotting, hybridization probe synthesis, and membrane washing were essentially performed as described by Kevil and colleagues (1997). Membranes were cut between sample lanes before hybridization. Approximately 10 µg of RNA sample was loaded per gel lane. The cDNA inserts of selected clones were released from the plasmid backbone by restriction enzyme digestion with *Not*I and *Sal*I. Inserts were gel-purified with Qiaquick spin-columns (Qiagen, Valencia, Calif.). Hybridization was performed overnight at 65°C in Perfect-hyb Plus from Sigma. Hybridization patterns were visualized on a STORM 860 phosphorimager.

Results and discussion

The main objectives of generating EST from the BARC 5BOV cDNA library were to maximize discovery of mammary expressed genes and to evaluate this sequence information to develop a resource for functional genomic studies. Before this effort, only 87 mammary EST derived from cattle had been generated and deposited into GenBank. Messenger RNAs extracted from mammary tissues representing eight distinct physiological states were pooled for normalized library construction (Table 1). The tissue stages represent periods of extensive postnatal growth and development, intensive synthesis and secretion of protein and lipid, tissue remodeling, and continuous immunological surveillance and activity. Pooling with mRNA from non-lactating tissues was expected to reduce the rate of sequencing EST associated with genes highly expressed in support of milk production and mammary growth (e.g., caseins and collagen). Pooling was also anticipated to enhance normalization, but not to the extent observed by Smith et al. (2001), where only mRNA extracted from unrelated tissues was pooled. In addition, EST generated from this library could be a resource for future studies to identify nucleotide sequence variation, because the tissues were sampled from 18 Holstein animals.

Initially, 17,664 clones from the BARC 5BOV library were processed for sequencing to generate over 12,000 EST. To assess rate of sequence redundancy, an intra-library overlap analysis was performed by using thresholds set at minimum overlap of 50 bp with no gaps. Because the results of this analysis indicated that 61% of the clones represented non-redundant sequence information, another one-third of the library was processed for sequencing. Overlap analysis of this larger data set indicated that non-redundant sequence information diminished to 31% and suggested that further sequencing for the purpose of unique sequence discovery within this library was limited. A total of 23,202 sequences met minimum criteria for automated GenBank submission to dbEST after quality assessment. Rejected sequences were found to correspond to clones with no insert (4%), no culture growth (5%), or low Phred quality scores (27%). In all, 9,476,952 bp of bovine sequence were generated with an average length of 408 bp per EST. BLASTN analysis of this sequence revealed that approximately 61% of the BARC EST had significant similarity (match score >300) to sequences in the GenBank nt database.

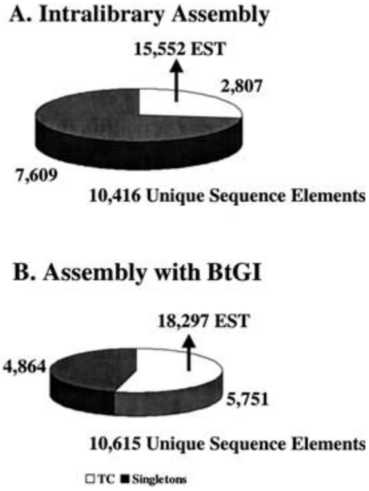


Fig. 1. The 23,161 sequences from the BARC 5BOV library were assembled independently **(A)** and with the existing BtGI **(B)**. Each pie represents the total number of unique sequence elements. Unique sequences are defined as sequence elements with no bovine equivalent under the constraints of sequence assembly. The number of BARC EST that represent TC sequences is denoted above the arrow originating from the corresponding pie slice.

To better evaluate the sequencing results from the 5BOV library and annotate EST for functional genomic studies, BARC EST were clustered and assembled both independently and with all the bovine expressed sequences (from GenBank) present in the BtGI (Release 6.0). An additional 41 EST (0.2%) were removed from the BARC data set after a second, more stringent screening for vector, poly-A/T tails, adaptor sequences, and contaminating bacterial sequences. Independent assembly of BARC EST identified 10,416 potentially unique sequence elements that could be used to develop mammary-specific microarrays (Fig. 1A). Assembly of BARC EST with BtGI produced a similar number of 10,615 unique sequence elements (Fig. 1B; 16% of total in BtGI); however, the percentage of BARC EST belonging to assemblies was increased from 67% to 80%. BARC EST populated 24% of the 23,883 TC sequences in BtGI, and approximately 2% of all TC sequences contained only BARC EST. Eleven percent of the 43,361 singletons in BtGI were BARC EST. For annotation, 47% of the cattle TC sequences and 66% of the mammary-related TC sequences could be assigned a gene function. Functional classification was performed by gene ontology assignments (http://www.tigr.org/tdb/btgi/GO/GO.html). Assignments could be made for approximately 55% of the TC sequences in BtGI.

The annotation and assembly results were used to evaluate the success of normalization for the 5BOV library. First, distribution of BARC EST among TC sequences was evaluated (Fig. 2). The majority (87%) of these assemblies contain one, two, or three BARC EST, indicating that overall sequence redundancy within this library is low. However, 13 TC corresponding to 12 genes contained 18% of these EST. These EST were sequenced at a higher frequency (> 0.3%) than any other EST sequenced from the four normalized bovine cDNA

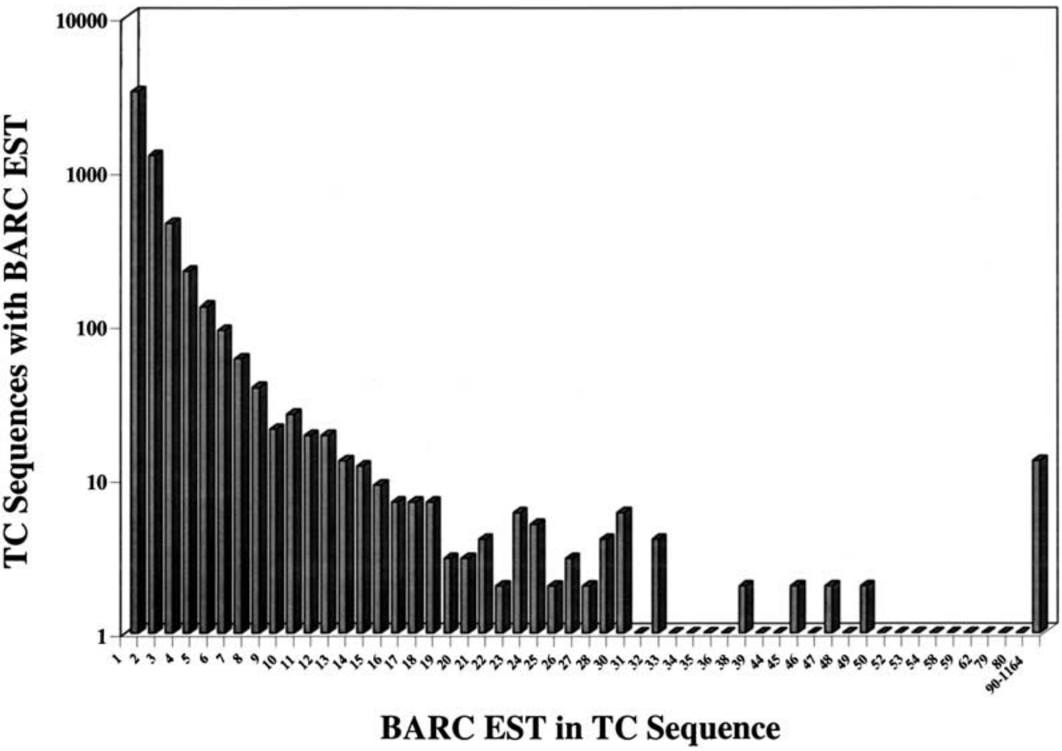


Fig. 2. The distribution of BARC EST among TC assemblies in the BtGI was determined. TC sequence assemblies were grouped into categories based on the number of BARC EST contained in each assembly (x-axis). The number of TC sequences in each group was plotted in log scale (y-axis).

Table 2. Assembly statistics and annotation for TC sequence assemblies containing BARC EST with a high rate of redundancy (>0.3%) in 5BOV library.

BARC EST in TC assembly	Sequence elements in TC assembly	BARC EST (%)	Percentage of total BARC EST (%)	Putative ID in BtGI
1164	1164	100	5.0	*alpha-s1-casein*
842	843	100	3.6	*beta casein*
524	524	100	2.3	*beta lactoglobulin*
430	433	99	1.9	*Lactotransferrin*
273	273	100	1.2	*kappa-casein precursor*
202	202	100	0.9	*alpha-s2-like casein precursor*
180	189	95	0.8	*Immunoglobulin variable region*
132	297	44	0.6	*not assigned*[a]
112	243	46	0.5	*alpha2(I) collagen*
100	332	30	0.4	*Osteonectin precursor*
98	165	59	0.4	*invariant chain*
98	126	78	0.4	*IgG1 heavy chain constant region*
90	121	74	0.4	*Osteopontin-k*

[a] One of two TC sequence assemblies matching *collagen alpha 1* (III).

libraries characterized by Smith and colleagues (2001). Together, these results suggest normalization was effective, but only for genes not typically associated with amplified expression during lactation. For example, some of the most common BARC EST assembled with TC sequences annotated as genes encode for five of the six major protein components in milk (Table 2).

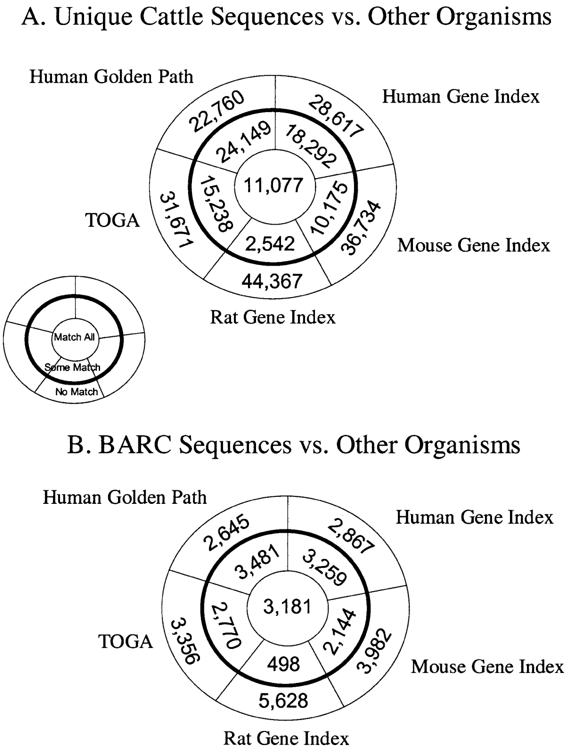
BARC EST (5%) that populated assembly of CSN1S1 were most prevalent, and the rate of redundancy was slightly lower than the 13.5% observed by sequencing random clones (N = 89) from the library prior to normalization. The rate of generating EST associated with EF1α was similar to that found in the other bovine normalized libraries produced at Invitrogen. Matches to EF1α for BARC EST occurred at a rate 0.16% compared with 1.1% prior to normalization. During library construction, the reduction in the relative abundance of hybridizing colonies for EF1α and CSN1S1 was calculated as 19- and 5-fold, respectively. These results correlate with the sequencing results in that normalization of EF1α was more effective than for CSN1S1. The difference in magnitude of normalization between sequencing and hybridization results may be attributed to variation in the handling of growth cultures prior to plating. Even though normalization had little effect on reducing the redundancy for some genes (Table 2), the number of EST for these genes was lower than that expected, considering that an average Holstein cow produces enough transcripts to synthesize approximately 1 kg of lactoprotein per day. This reduction of amplified transcripts was primarily the result of pooling mRNA extracted from different physiological conditions.

For confirmation of this observation, hybridization probes (cDNA) that represent the individual mRNA samples used to create the pools were generated. Labeled cDNA representing lactating, mastitic, mid-gestation, and pre-pubertal parenchyma mRNA samples produced positive signals from 12, 12, 18, and 16% of the library clones arrayed on filters, respectively (data not shown). As expected, the majority of the positive signals detected with the lactation probe were cDNA corresponding to milk protein genes. The overlap in hybridization signals between lactating and mastitic probes was nearly 100%, suggesting probe synthesis from these mRNA fractions was saturated by messages corresponding to genes encoding lactoproteins. Even though there was some overlap of hybridization signals among the membranes representing the four mRNA samples, hybridization signals for cDNA corresponding to lactoproteins were not detectable with prepubertal parenchyma probe. No conclusions could be drawn about the proportions of hybridization signals relative to each tissue stage, because less abundant mRNA probably produced an undetectable amount of probe. These results provide support that tissue pooling was more effective than library normalization at reducing the redundancy of cDNA for milk protein genes, and indicate that production of hybridization probes from lactating mammary gland for functional genomic studies may require subtraction of overly abundant transcripts to allow detection of less abundant messages.

The number of EST comprising the mouse and human gene indices is nearly an order of magnitude larger than the BtGI. Owing to this disparity, the number of novel sequences in the BtGI would be expected to be relatively small. To test this premise, extensive comparative sequence analysis was performed with the 57,986 unique sequence elements (TC sequences and singletons) of BtGI (Release 5.0) against assembled sequence data from the human genome (Golden Path release August, 2001), human and rodent gene indices, and orthologous alignments of genes across eukaryotes (TOGA). The results of this analysis are shown in Fig. 3, and alignments to Golden Path and TOGA have been made accessible in BtGI.

The highest number of matches (61%) was to the assembly of the human genome. Of the 35,226 alignments, 1102 mapped where no human TC sequence, transcript, or singleton EST was found, suggesting the potential locations of genes or exons not yet sampled in humans. Surprisingly, 17,535 (26%) of the sequence elements in BtGI did not have a significant match with these four sequence databases. Only 16% of the sequence elements comprised of BARC EST fell into the "no hit" category, suggesting a higher proportion of EST from the 5BOV library relative to BtGI represented expressed genes that had been sampled in humans and rodents.

These no-hit sequences were investigated further to determine whether they represented genes potentially unique to cattle, divergent orthologs, sequences for which expressed orthologs have not yet been sampled in other species, or sequence artifacts. Only the 3445 no-hit TC sequences, which represent expressed sequences sampled more than once, were considered for these analyses. The majority of these TC sequences were also assembled with EST from different libraries. First, BLASTN was performed against BtGI to determine matches to bovine repetitive elements (Table 3). Only a small fraction of the no-hit TC sequences matched these SINE sequences, suggesting the majority of no-hit TC sequences were not assemblies of sequence artifacts derived from immature transcripts or genomic DNA. BLASTN was performed against the GenBank nt databases for these TC sequences with a threshold set at 10^{-10}, and significant matches (excluding matches to other cattle sequences) were observed for 40 sequences. Together, these results suggest that over 3000 of the TC sequence assemblies for cattle correspond to potentially unique or divergent transcripts. However, these results


A. Unique Cattle Sequences vs. Other Organisms

B. BARC Sequences vs. Other Organisms

Fig. 3. Results of comparisons between assembled BtGI Release 5.0 **(A)** or BARC sequences **(B)** and sequences in the Human, Mouse and Rat Gene Indexes, TOGA and the Human Golden Path data sets. Each diagram contains five pie slices surrounding a central circle. The number in the central circle represents those bovine query sequences with a significant match to all five data sets (match all). The remaining query sequences relative to each data set are shown in the surrounding pie slices (name of data set denoted outside pie slice). The thick circle superimposed onto each pie diagram, divides those sequences with a match within a data set (some match) from those that have no match (no match). Sequences represented in the inner ring may match more than one data set, but not all data sets. In **A**, query of the sequence IDs common among the outer ring of the five pie slices yields 17,535 sequence elements, and these had no match to sequences in the five data sets.

Table 3. Results of BLASTN analysis of BtGI sequence elements against common bovine repetitive elements. GenBank Accessions for some of the short interspersed nuclear element (SINE) sequences present in the bovine genome were compared with sequences in BtGI. The significance threshold for BLAST analysis was set at 10^{-5}.

			Percentage of sequences		
Sequence ID	GenBank accession	Length (bp)	BtGI sequence elements (N = 67,870)	BARC EST (N = 23,161)	No hit TC sequences (N = 3445)
SINE sequence Bov-B	X64125	560	1.9	0.6	1.1
SINE sequence Bov-A2	AF327250	260	3.8	1.1	2.9
Retrotransposon conserved among Cetartiodactyla	U97688[a]	110	2.2	0.2	1.2

[a] Accession for a portion of bovine *alpha-mannosidase* exon 7.

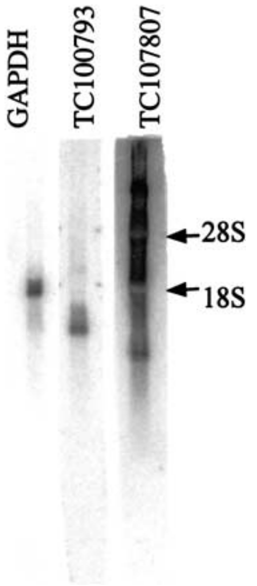


Fig. 4. Northern analysis of TC sequence assemblies with no comparative sequence matches. Total RNA from prepubertal mammary epithelium was electrophoresed and blotted onto nylon membrane. Shown are three strips from this membrane, and each has been hybridized to a different cDNA probe representing a no-hit TC sequence (sequence ID above membrane strip). A portion of the bovine *glyceraldehyde 3-phosphate dehydrogenase* (*GAPDH*) gene was used as a probe for the positive control (membrane strip on left).

could be caused by a lack of bovine sequence information in BtGI.

To test this suggestion, we selected cDNA clones representative of no-hit TCs for further sequence and expression analysis. The previous results obtained by hybridizing probes from prepubertal and mid-gestation mRNA to arrayed clones on membranes were used to compare the library plate addresses of detectable signal to BARC EST assembling with no-hit TC sequences. With membranes corresponding to 5BOV library plates 96–142, this comparison yielded clones corresponding to 23 TC sequences. The EST composition of these TC assemblies was analyzed, and all but two of the assemblies were constructed with sequences generated from different libraries. These two assemblies were eliminated from further analysis, because the EST sequence and trace files were of irregular quality. Furthermore, the EST for each assembly originated from a single plate within the 5BOV library, suggesting a contaminant was introduced during the processing of these two plates. BLASTX was performed for the other 21 TC sequences against the GenBank nr databases with a threshold set at 10^{-10}, and a significant match was observed for one sequence. Next, single-pass sequencing of the 5′ and 3′ ends of the 20 remaining clones followed by BLASTN analysis against GenBank nt and htgs databases revealed significant matches for eight clones. The other 12 clones did not significantly match with sequences from any other species in GenBank other than repetitive DNA elements. Probes for the cDNA inserts of these 12 clones were hybridized to Northern blots. Eight of the probes produced smear patterns similar to that shown for TC 107807 (Fig. 4). The smear was probably caused by hybridization to RNA containing repetitive element. The TC sequences for four of the probes were identified to contain repetitive sequence by BLAST analysis, and 3′ end sequencing revealed the same for a clone of TC107807. In the former case, TC sequences assembled with EST from different libraries, and the overlap of these sequences was across both flanks of the repetitive element. Regardless, the eight clones yielding smear patterns may be a product of cDNA synthesis from immature or improperly spliced transcripts, and full-length sequence analysis might help identify mammalian orthologs for these TC sequences.

For the remaining four candidates, discrete bands were detected by Northern analysis similar to that shown for TC100793 (Fig. 4). After Northern analysis, the annotation of these four TC assemblies was searched again in BtGI, because a scheduled update (Release 6.0) had added over 40,000 additional EST to the assemblies. Subsequently, this information improved the alignment of some TC sequences to a revised Golden Path, and as such three of the no-hit TC sequences with interpretable expression data aligned with the human genome. Only TC100793 remained potentially unique to cattle. The transcript corresponding to this TC sequence was approximately 800 nt in length and contained two small upstream open reading frames (ORF) and two ORF encoding longer peptides. Comparison of actual full-length size based on Northern analysis agreed with the predicted size from assembly of EST from eight different cDNA libraries. Determining map location and genomic sequence flanking this gene may provide evidence as to the origin or function of this gene relative to other species. Based on the results of this extended analysis, only a small percentage of the no-hit TC sequences potentially represent genes that have rapidly diverged or evolved in ruminants. Overall, our results suggest that additional high-quality bovine EST will provide the sequence information needed to annotate and align bovine genes onto the human genome sequence, and the additional resources added to BtGI provide powerful tools for identifying potentially novel genes.

Conclusions

This report provides the first publicly available assessment of EST generated from a pooled-tissue, normalized cDNA library of the mammary gland. This effort to better catalog gene expression in the mammary gland has aided a larger USDA effort aimed at developing high-quality genomic resources for livestock. Functional genomic studies hold great promise for improving basic knowledge of mammary gland biology. The mammary gland provides a research model for scientists investigating metabolism, secretion, tissue remodeling, and complex diseases like mastitis. Mastitis control is one of the most difficult management challenges facing the dairy industry worldwide, as this disease annually costs U.S. producers an estimated two billion dollars. Interpretation of gene expression data from these types of studies will rely on annotated sequence resources to help understand the changing dynamic among cell types within the mammary gland. Our results underscore the importance of comparative sequence analysis to aid characterization of mammalian transcriptomes and provide a valuable sequence resource that will facilitate functional genomic studies of the mammary gland regardless of species.

Acknowledgments. We thank D. Larry Wood for superior effort and technical assistance, Adilson Mota and Drs. Kurt Zuelke and Tim Smith for critical review of this manuscript. The BARC Animal Care and Use Committee approved protocols with experimental animals. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215, 403–410

Capuco AV, Ellis S, Wood DL, Akers RM, Garrett W (2002) Postnatal mammary ductal growth: three-dimensional imaging of cell proliferation, effects of estrogen treatment and expression of steroid receptors in prepubertal calves. Tissue Cell, in press

Farr VC, Stelwagen K, Cate LR, Molenaar AJ, McFadden TB et al. (1996) An improved method for the routine biopsy of bovine mammary tissue. J Dairy Sci 79, 543–549

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9, 868–877

Kevil CG, Walsh L, Laroux S, Kalogeris T, Grisham MB et al. (1997) An improved, rapid Northern protocol. Biochem Biophys Res Commun 238, 277–279

Kondo S, Shinagawa A, Saito T, Kiyosawa H, Yamanaka I et al. (2001) Computational analysis of full-length mouse cDNAs compared with human genomic sequences. Mamm Genome 12, 673–677

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL et al. (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. Nat Genet 25, 239–240

Long E, Capuco. AV, Wood DL, Sonstegard T, Tomita G et al. (2001) Apoptosis and cell proliferation are induced in *Escherichia Coli* infected bovine lactating mammary glands. Cell Death Differ 8, 808–816

Quackenbush J, Liang F, Holt I, Pertea G, Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res 28, 141–145

Smith TP, Godtel RA, Lee RT (2000) PCR-based reaction setup for high-throughput cDNA library sequencing on the ABI 3700 automated DNA sequencer. Biotechniques 29, 626–630

Smith TPL, Grosse WM, Freking BA, Roberts AJ, Stone RT et al. (2001) Sequence evaluation of four pooled-tissue normalized bovine cDNA libraries and construction of a gene index for cattle. Genome Res 11, 698–700

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al. (2001) The sequence of the human genome. Science 291, 1304–1351

Wright FA, Lemon WJ, Zhao WD, Sears R, Degen Z et al. (2001) A draft annotation and overview of the human genome. Gen Biol 2, 1–18

Zhuo D, Zhao WD, Wright FA, Yang HY, Wang JP et al. (2001) Assembly, annotation and integration of UNIGENE clusters into the human genome draft. Genome Res 11, 904–918