Sequence Evaluation of Four Pooled-Tissue Normalized Bovine cDNA Libraries and Construction of a Gene Index for Cattle

Timothy P.L. Smith,[1,3] William M. Grosse,[1] Brad A. Freking,[1] Andrew J. Roberts,[1] Roger T. Stone,[1] Eduardo Casas,[1] James E. Wray,[1] Joseph White,[2] Jennifer Cho,[2] Scott C. Fahrenkrug,[1] Gary L. Bennett,[1] Michael P. Heaton,[1] William W. Laegreid,[1] Gary A. Rohrer,[1] Carol G. Chitko-McKown,[1] Geo Pertea,[2] Ingeborg Holt,[2] Svetlana Karamycheva,[2] Feng Liang,[2] John Quackenbush,[2] and John W. Keele[1]

[1]United States Department of Agriculture, Agricultural Research Service, United States Meat Animal Research Center, Clay Center, Nebraska 68933, USA; [2]The Institute for Genomic Research, Rockville, Maryland 20850, USA

An essential component of functional genomics studies is the sequence of DNA expressed in tissues of interest. To provide a resource of bovine-specific expressed sequence data and facilitate this powerful approach in cattle research, four normalized cDNA libraries were produced and arrayed for high-throughput sequencing. The libraries were made with RNA pooled from multiple tissues to increase efficiency of normalization and maximize the number of independent genes for which sequence data were obtained. Target tissues included those with highest likelihood to have impact on production parameters of animal health, growth, reproductive efficiency, and carcass merit. Success of normalization and inter- and intralibrary redundancy were assessed by collecting 6000–23,000 sequences from each of the libraries (68,520 total sequences deposited in GenBank). Sequence comparison and assembly of these sequences was performed in combination with 56,500 other bovine EST sequences present in the GenBank dbEST database to construct a cattle Gene Index (available from The Institute for Genomic Research at http://www.tigr.org/tdb/tgi.shtml). The 124,381 bovine ESTs present in GenBank at the time of the analysis form 16,740 assemblies that are listed and annotated on the Web site. Analysis of individual library sequence data indicates that the pooled-tissue approach was highly effective in preparing libraries for efficient deep sequencing.

The rapid progress of genomic research in diverse organisms such as yeast, fruit flies, nematodes, mice, and humans has been driven by a combination of mapping, sequencing, and identification of expressed portions of each genome. Progress in these areas has lagged in the livestock species, limiting the use of functional genomics approaches to current problems in production-animal agriculture. Resources for a public effort to sequence the genomes of livestock species are not currently available. However, more modest sequencing efforts aimed at cDNA libraries have substantial value to the research community, especially when combined with mapping efforts to produce comparative maps with other mammalian species. Comparative maps make use of the general conservation of synteny between mammals and allow the livestock community to tap into the wealth of information generated in the human genome effort.

[3]Corresponding author.
E-MAIL smith@email.marc.usda.gov; FAX 402-762-4390.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.170101.

To provide a resource of livestock-specific expressed sequence data to facilitate proteomics and functional genomics approaches in animal science, an EST-sequencing program was initiated with the aim to maximize the efficiency of obtaining sequence from the highest possible number of independent genes. Four normalized bovine cDNA libraries specifically designed for this task were produced and arrayed for high-throughput sequencing. To make the data more accessible and useful, assembly and annotation analyses were performed and an interactive Web site constructed by The Institute for Genomic Research (TIGR) to view and analyze bovine genes. We report the construction of this Cattle Gene Index and assessment of the specialized libraries after collection of single-pass EST sequence from 74,890 clones.

RESULTS

The U.S. Meat Animal Research Center (MARC) libraries were made with RNA pooled from multiple tissues to increase efficiency of normalization and maximize the number of independent genes for which sequence data were obtained. Animals were selected based on

11:626–630 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

availability and breed diversity, and the tissues were selected as those most likely to express genes affecting important production traits of growth, carcass quality, reproduction, and animal health in cattle. Specific tissues included in each of the first three libraries (MARC 1BOV, MARC 2BOV, and MARC 3BOV) are shown in Table 1. The pooling strategy included dissimilar tissues within a library because the high and intermediate expression classes have the greatest likelihood of involving different genes in these different tissues. This approach was expected to improve overall normalization because some reduction in the relative frequency of higher abundance classes of dissimilar tissues is possible by pooling RNA in this fashion. Five different animals were used as sources for the first three libraries as listed in Table 1, to support detection of nucleotide sequence variation in silico from the resulting data. The fourth library (MARC 4BOV) was made from the RNA of whole bovine embryos at day 20 and day 40 of gestation (280-d gestation period).

The first objective was to obtain sufficient sequence from each library to evaluate its potential as a resource for deep sequencing. Sequences from 68,520 clones from the four libraries (91% of sequences) were suitable for submission to GenBank after removal of bacterial genomic artifacts and based on length and quality of sequence. A breakdown of the number of sequences for each library is shown in Table 1. The average length of sequence deposited was 398 bp, and a poly-A tract corresponding (in most cases) to the 3′ terminal end of the transcript was observed in 29% of the clones. A relational database and automated dataflow were established to process and store the raw sequence data and evaluate results of BLAST analysis for annotation (MARCDB; J. Keele, unpubl.). Analysis of BLASTN output for bovine EST data in MARCDB indicated that ~20% of sequences had significant (score >300) similarity to sequences in the GenBank nr database (Table 1), and ~30% had significant (score >200) matches to nonbovine sequences in the dbEST database. (The lower threshold was set in the latter case because of the relatively short lengths of EST sequences.)

To determine if the normalization process had significantly altered the representation of genes from the source tissue mRNAs, the representation of two tissues in the 1BOV library was assessed by hybridization of radiolabeled probes, prepared from mRNA of ovary and

Table 1. Production and Sequencing of Bovine EST Libraries

cDNA Library	Tissue	Collection animal	% of RNA	Fold normal	Avg insert length (bp)	No. of seq	% GenBank nr match	% Unique
MARC 1BOV	Mesenteric lymph node	3	18	18.3	1300 ± 632	23,470	18	63
TIGR library ID: 2819	Hilar lymph node	3	18		(n = 500)			
	Post-pubertal ovary	4	18					
	Kidney-associated fat	3	7					
	Hypothalamus	3	18					
	Pituitary	2	18					
	Subcutaneous fat	3	3					
MARC 2BOV	Testis	1	15	16.8	1143 ± 594	16,494	23	65
TIGR library ID: 2820	Thymus	2	16		(n = 139)			
	Semitendenosous muscle	2	15					
	Longissimus muscle	2	16					
	Pancreas	2	6					
	Adrenal gland	2	16					
	Endometrium	4	16					
MARC 3BOV	Bone marrow	2	21	14.6	1188 ± 564	6,196	23	82
TIGR library ID: 2821	Alveolar macrophage	3	20		(n = 310)			
	Pre-pubertal ovary	3	19					
	Fetal semitendonosous muscle	5	20					
	Fetal longissimus muscle	5	20					
MARC 4BOV	whole embryos (pool of 4)	d20 embryos	50	12.7	1008 ± 879	22,360	20	60
TIGR library ID: 2822	whole embryos (pool of 2)	d40 embryos	50		(n = 232)			
Total sequences						68,520		

RNA from the indicated tissue was collected from one of five animals, with the exception of the d20 and d40 embryos which each represented a pool of embryos. The five animals used were (1) a 48-h-old Piedmontese-Angus cross calf, (2) a 20-d-old Simmental-MARCIII calf, (3) an 800-lb MARCIII yearling heifer, (4) a pregnant six-year-old Limousin cow, or (5) a d100 fetus from a Gelbvieh-Limousin cross. A pool of 5 ug of mRNA was used to construct each library, with percentage of the total amount of RNA from each tissue as indicated (% of RNA). Effective normalization of each library (fold normal) is expressed as the fold decrease in abundance of EF1a clones. The number of EST sequences deposited in GenBank from each library is indicated. Sequences were compared to GenBank nr database via BLAST analysis, and percentage of clones showing significant match (score >300) is shown. The percentage of sequences with no significant overlap with other sequences from the same library is indicated (% unique).

pituitary, to high-density grids of the clones. This approach was used in place of 3′ end tags that potentially could have been included in the cDNA synthesis step because sequencing was performed from the 5′ end of the insert and only 15%–40% of picked clones were sequenced. Hybridization of labeled mRNA from bovine pituitary gland or ovary produced positive signals from 18% and 19% of the library clones, respectively, in agreement with the percentage of RNA from each tissue used in library construction (Table 1). Because the probes were generated from the mRNA population of the tissues, the sensitivity of the assay is biased toward genes with high steady-state levels of expression. A significant proportion of these highly expressed genes are housekeeping genes common between all cell types, such that one would predict that a significant number of positive clones would be in common between tissues. Indeed, 76% of the positive signals produced by pituitary probe and 73% of those from ovary probe were common to both tissues. However, the results of this experiment suggest that there is no large change in representation of tissues as a result of the pooling and normalization strategy.

Success of the four normalized libraries was evaluated by BLASTN analysis of sequences from each library to all others from the same library. A minimum overlap of 50 bp with no gaps was used to indicate redundancy at the sequence level. This approach can overestimate the level of redundancy, because the minimum overlap still allows significant unique sequence information to be obtained from each of the overlapping sequences. In addition, EST sequences with homology to bovine repetitive elements will share homology among large numbers of other ESTs, creating artifactual clusters. To minimize the impact of this artifact, overlaps based on known repetitive elements were eliminated. The overlap analysis indicated that 63% of clones represent nonredundant sequence information within the 1BOV library. Similar within-library redundancy rates were observed for the other three libraries (Table 1). We conclude that the pooled-tissue approach was successful in generating a resource for efficient collection of nonredundant EST sequence.

The overlap analysis suggested that continued random sequencing would efficiently produce sequence that would be unique within all four of the libraries. However, the level of redundancy between libraries, and with the full data set of bovine EST sequence in GenBank, must be considered to determine the overall value of continued deep sequencing. Individual EST sequences, therefore, were evaluated for overlap with each other and with all other bovine sequences in GenBank to form clusters. The number of clusters that results allows a first approximation of the number of genes for which sequence has been collected. However, this number will be overestimated, because multiple independent clusters may actually represent distinct portions of a single gene.

The TIGR Cattle Gene Index database was assembled from 125,485 EST sequences present in GenBank dbEST, of which 67,881 were from the four libraries described here. Screening for vector, poly-A/T tails, adaptor sequences, and contaminating bacterial sequences eliminated 1104 (0.9%) sequences and trimmed 933,355 bases of contaminating sequence from the ends of 29,422 ESTs. The procedure identified 15,993 clusters, leaving 30,544 unclustered singleton EST sequences. Assembly of component sequences for each cluster using CAP3 (Huang and Madan 1999) produced the 16,740 Tentative Consensus assemblies (TCs) that comprise the TIGR Cattle Gene Index. Consensus sequences are formed from an average of 3.4 ± 1.8 ESTs (range 1–88). TCs containing a known gene were assigned the function of that gene, and the remainder were searched using DPS (Huang et al. 1997) against a nonredundant protein database. TCs with significant matches were assigned a putative function. Statistics describing the results of cluster analysis for each of the four normalized libraries (see http://www.tigr.org/tdb/btgi) are shown in Table 2. There were 1138 assemblies having sequence homology with the 2898 Expressed Transcripts (ETs) present in GenBank.

DISCUSSION

At the inception of this project, there were 484 bovine EST sequences in the GenBank database. In the past year, the number of sequences has grown over 200-fold, ~50% of which were generated from the four libraries described here. A fifth library (BARC 5BOV) constructed from mammary gland has been produced and sequenced in an identical fashion at the Beltsville Agricultural Research Center, contributing an additional 14,122 sequences (T. Sonstegard, pers. comm.).

Table 2. Statistics of Assembly in the Cattle Gene Index

	In TCs	Singletons	Total
Sequences			
ESTs	93,837	30,544	124,381
ETs	2,291	607	2,898
Totals	96,128	31,151	127,279
Unique Sequences			
TCs			16,740
Singleton ETs			607
Singleton ESTs			30,544
Totals			47,891

The number of ESTs and expressed transcripts (ETs) that fit in tentative concensus sequences (TCs) or are singletons is indicated. The total numbers of unique bovine sequences present in GenBank is the sum of the number of TCs and singleton sequences.

The sequences from all five of these bovine libraries were generated from pools of tissues of multiple individual animals with different breed backgrounds, representing a resource for identification of nucleotide sequence diversity within and between cattle breeds.

The libraries described here were specifically designed to maximize the number of unique sequences collected from a variety of tissues of interest, with minimum investment in normalization and total number of sequence reactions. Recent improvements in library production and normalization technology may make the approach even more efficient than the methodology applied for these libraries, potentially improving average insert size and further reducing redundancy. Nevertheless, the approach appears to have been highly successful, based on the redundancy within and between the libraries. Because formation of clusters is a reflection of library redundancy, random sequencing of non-normalized libraries rapidly results in the formation of clusters representing portions of genes most highly expressed in source tissues. If library normalization was a significant problem, one would predict a relatively small number of assemblies, having a disproportionately large number of ESTs per assembly, skewing the distribution of number of ESTs per assembly versus the number of TCs. The distribution for the sequences in the four libraries described here is shown in Figure 1 and illustrates the effectiveness of the approach of pooling RNA from multiple, unlike tissues prior to normalization.

We used a hybridization approach to show that all tissues in the pool are represented in the library in proportion to the percentage of RNA in the pool. This hybridization procedure provides additional information relative to tagging the individual cDNA pools from each tissue and sequencing because it identifies clones expressed in more than one tissue, regardless of the origin of the clone itself. This information can be of value in construction of tissue-specific arrays for expression analysis, for example. The drawback to this approach is that rare messages may not generate positive hybridization signals due to the low level of specific probe generated by labeling total mRNA.

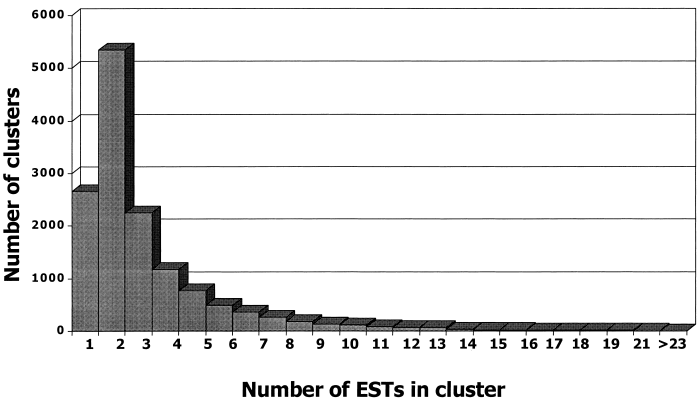

Figure 1 Overall sequence redundancy of the four normalized libraries. For each of the assemblies in The Institute for Genomic Research (TIGR) Cattle Gene Index, the number of MARC ESTs from the four normalized libraries that fit in each cluster was determined. This is plotted against the total number of clusters containing the same number of MARC ESTs. The majority (73%) of clusters contain one, two, or three MARC ESTs per cluster, indicating that overall sequence redundancy of the four libraries is low.

The large amount of sequence data produced in this type of project is difficult to use without appropriate bioinformatic support. In addition to allowing evaluation of the success of EST sequencing, the construction of the TIGR Cattle Gene Index facilitates comparative and molecular genetic studies, including construction of complete bovine cDNA sequences without the need for traditional cloning steps, and design of primers for gene-specific amplification reactions.

We conclude that pooling of tissues prior to normalization, as well as inclusion of unlike tissues within a library, is a very effective approach to developing deep-sequence ready libraries that maximize unique gene discovery. The sequence data developed during this project is underpinning efforts to create comparative maps using both radiation hybrid panel and genetic-linkage mapping approaches, creating an important resource to support genome research in livestock.

METHODS

Library Construction and Normalization

Tissues were collected into cryotubes, submerged in liquid nitrogen, and shipped to Life Technologies, Inc. Total RNA extraction, mRNA purification, primary library construction, and normalization were purchased as a service from the Gene Discovery Services division of Life Technologies, which used commercially available protocols and materials and provided the following pertinent data. Total RNA extraction and mRNA purification were performed separately for each tissue. The poly-A selected mRNAs were pooled prior to cDNA synthesis, with the proportion of each RNA in the pool indicated in Table 1. The first-strand-synthesis primer was 5′-GACTAGTTCTAGATCGC GAGCGGCCGCCC (T_{15})-3′ (no tissue-specific tag was included). Following second-strand synthesis, a *Sal*I adaptor primer was ligated to the cDNA (5′-TCG ACCCACGCGTCCG-3′) and the product was digested with *Not*I and ligated into vector pCMVSPORT6 (LTI) predigested with *Not*I and *Sal*I.

The primary library was amplified in semisolid agar as described by Hanahan et al. (1991) prior to normalization. One portion of double-stranded plasmid DNA representing the library was linearized by *Not*I. This *Not*I-digested library was used as a template for biotinylated RNA synthesis using SP6 RNA polymerase. Another portion of the double-stranded plasmid library was converted to single-stranded circles in vitro using Gene II and Exonuclease III (Life Technologies). Three µg of single-stranded DNA was hybridized (C0t 500) with 255 µg of Bio-RNA and vector-blocking oligonucleotides.

The hybridized Bio-RNA/ss-circles were removed by streptavidin:phenol extraction. Normalized libraries were processed as described by Swaroop et al. (1991)and Li et al. (1994).

Normalization was estimated by hybridization of an elongation factor 1 alpha (EF1a) probe (previously determined to be among the highest copy number in the primary libraries) to colony lifts of a plated aliquot of each library. Reduction in the relative abundance of hybridizing colonies was between 12- and 19-fold (Table 1), calculated as the ratio of the percentage of clones that were EF1a-hybridization positive before and after normalization. Average insert length was estimated (Table 1) based on sizing of PCR-generated products from 100–500 clones per library using standard m13 vector primers.

EST Sequencing

Approximately 2×10^6 total transformants per normalized library were obtained from Life Technologies. Individual colonies were picked into 384-well plates with a robot from plated aliquots of each library by BACPAC resources. Single-pass sequencing was performed as described by Smith et al. (2000) using PCR amplification of inserts for template preparation and SP6 sequencing primer. Sequences of sufficient length and quality for submission to GenBank have been deposited in the dbEST database. The number of sequences generated per library is shown in Table 1.

Cattle Gene Index Assembly

The TIGR Cattle Gene Index database was assembled using techniques developed at TIGR for the analysis of EST and gene sequences to construct transcript databases for a variety of species (Quackenbush et al. 2000). Cattle EST sequences were downloaded from dbEST records and then rigorously screened to remove contaminating vector, poly-A/T tails, adaptor sequences, and contaminating bacterial sequences. Also included were 2898 *Bos taurus* gene sequences (NP sequences) parsed through Entrez from CDS and CDS-join features in GenBank records. `FLAST`, a rapid sequence comparison program based on DDS (Huang et al. 1997), in which query sequences are concatenated and searched against a nucleotide database, was used to compare all sequences pairwise. Clusters were formed with sequences having >95% identity over at least 40 base pairs, with unmatched overhangs <20 base pairs. Component sequences of each cluster were assembled using CAP3 (Huang and Madan 1999) to produce the Tentative Consensus (TC) sequences that comprise the TIGR Cattle Gene Index.

Evaluation of Tissue Representation

High-density grids of each library were purchased from BACPAC resources. The 1BOV library was represented by three 21×21-cm membranes, each containing 18,432 arrayed clones spotted in duplicate. Two sets of these membranes were used for replicate hybridization with probes corresponding to two of the source tissues for the library. Bovine pituitary and ovarian tissues were collected from cows at random stages of the estrous cycle and various stages of pregnancy (pituitary only) and frozen. Frozen tissues were pulverized and then homogenized in 4 M guanidinium thiocyanate, 25 mM sodium citrate at pH 7, 0.5% sarcosyl, 0.1 M 2-mercaptoethanol. Total cellular RNA was purified by ultracentrifugation through 5.7 M cesium chloride. Poly-A-selected mRNA was prepared from each total cellular RNA sample by oligo dT-cellulose (Collaborative Biomedical Products) column chromatography. Two pools of poly-A-selected mRNA from each tissue type were prepared from three to six animals. Labeled cDNA probe was generated from each pool by random primer extension with MMLV reverse transcriptase and hybridized to the membranes. Hybridization patterns were visualized on a STORM 860 phosphoimager (Molecular Dynamics) and evaluated using `Array Vision Version 4.0 Rev. 1.3` software (Imaging Research, Inc.). Criteria for considering a clone positive for hybridization was that each duplicate have signal 0.5 S.D. above the mean frequency distribution of pixel intensities for the individual primary element (4×4 vector representing eight clones spotted in duplicate) in which the clone was spotted.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Hanahan, D., Jessee, J., and Bloom, F.R. 1991. Plasmid transformation of Escherichia coli and other bacteria. *Meth. Enzymol.* **204:** 63–113.

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

Huang, X., Adams, M.D., Zhou, H., and Kerlavage, A.R. 1997. A Tool for Analyzing and Annotating Genomic Sequence. *Genomics* **46:** 37–45.

Li, W.-B., Gruber, C.E., Lin, J.-J., Lim, R., D'Alessio, J.M., and Jessee, J.A. 1994. The isolation of differentially expressed genes in fibroblast growth factor stimulated BC3H1 cells by subtractive hybridization. *BioTechniques* **16:** 722–729.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. 2000. The TIGR Gene Indices: Reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28:** 141–145.

Smith, T.P.L., Godtel, R.A., and Lee, R.T. 2000. PCR-based reaction setup for high-throughput cDNA library sequencing on the ABI 3700 Automated DNA Sequencer. *Biotechniques* **29:** 698–700.

Swaroop, A., Xu, J., Agarwal, N., and Weissman, S.M. 1991. A simple and efficient cDNA library subtraction procedure: Isolation of human retina–specific cDNA clones. *Nucleic Acids Res.* **19:** 1954.

Received November 10, 2000; accepted in revised form January 24, 2001.