A Cattle–Human Comparative Map Built with Cattle BAC-Ends and Human Genome Sequence

Denis M. Larkin, Annelie Everts-van der Wind, Mark Rebeiz, et al.

Genome Res. 2003 13: 1966-1972
Access the most recent version at doi:10.1101/gr.1560203

Resource

A Cattle–Human Comparative Map Built with Cattle BAC-Ends and Human Genome Sequence

Denis M. Larkin,[1] Annelie Everts-van der Wind,[1] Mark Rebeiz,[1] Peter A. Schweitzer,[2] Sharon Bachman,[2] Cheryl Green,[1] Chris L. Wright,[2] Edhilvia J. Campos,[2] Leslie D. Benson,[2], Jennifer Edwards,[2] Lei Liu,[2] Kazutoyo Osoegawa,[3] James E. Womack,[4] Pieter J. de Jong,[3] and Harris A. Lewin[1,2,5]

[1]Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801 USA; [2]The W.M. Keck Center for Comparative and Functional Genomics, University of Illinois at Urbana-Champaign, Illinois 61801, USA; [3]Children's Hospital Oakland Research Institute, Oakland, California 94609, USA; [4]Department of Veterinary Pathobiology, Texas A & M University, College Station, Texas 77843, USA

As a step toward the goal of adding the cattle genome to those available for multispecies comparative genome analysis, 40,224 cattle BAC clones were end-sequenced, yielding 60,547 sequences (BAC end sequences, BESs) after trimming with an average read length of 515 bp. Cattle BACs were anchored to the human and mouse genome sequences by BLASTN search, revealing 29.4% and 10.1% significant hits ($E < e^{-5}$), respectively. More than 60% of all cattle BES hits in both the human and mouse genomes are located within known genes. In order to confirm in silico predictions of orthology and their relative position on cattle chromosomes, 84 cattle BESs with similarity to sequences on HSA11 were mapped using a cattle–hamster radiation hybrid (RH) panel. Resulting RH maps of BTA15 and BTA29 cover ~85% of HSA11 sequence, revealing a complex patchwork shuffling of segments not explained by a simple translocation followed by internal rearrangements. Overlay of the mouse conserved syntenies onto HSA11 revealed that segmental boundaries appear to be conserved in all three species. The BAC clone-based comparative map provides a foundation for the evolutionary analysis of mammalian karyotypes and for sequencing of the cattle genome.

[Supplemental material is available online at www.genome.org.]

Sequencing of the human and mouse genomes has created a watershed of opportunities in biology (Lander et al. 2001; Waterston et al. 2002). Among the most anticipated scientific advances from genome sequencing is the elucidation of genome evolution and function, with comparative analysis tools being the primary investigative strategy. It is now well accepted that progress in mammalian comparative genomics is dependent upon the position on the phylogenetic tree of species selected for genome mapping and sequencing (O'Brien et al. 2001). Thus, the mammalian order Cetartiodactyla has become a focus of great attention in comparative genomics, because it comprises a phylogenetically distant clade of eutherian mammals relative to primates, having diverged from a common ancestor ~85 million years ago (Kumar and Hedges 1998). On the basis of a limited amount of sequence information for orthologous regions in a number of mammals (Thomas et al. 2002), it is clear that a Cetartiodactyl genome will play an essential role in informing the human genome for conserved noncoding structural and regulatory elements, properly annotating exon/intron boundaries, and the identification of novel genes. For these reasons, the U.S. National Institutes of Health (NIH) has given high priority to the complete genome sequencing of two Cetartiodactyl species, *Bos taurus* (cattle) and *Sus scrofa domestica* (pig; http://www.genome.gov/page.cfm?pageID=10002154).

An important first step for efficiently sequencing a new mammalian genome is to have a high-quality, comparatively anchored physical map. Various strategies for creating such a map have been employed, including the use of BAC fingerprinting and BAC-end sequencing. For example, Fujiyama et al. (2002) produced a comparative clone-based map of the human and chimpanzee genomes using paired chimpanzee BAC-end sequences (BESs) aligned by BLAST with the human genome sequence. Approximately 98% of chimpanzee BESs had BLAST hits in the human genome that identified putative orthologs, thus yielding a powerful resource for comparative genome analysis. Gregory et al. (2002) produced a detailed comparative physical map of the mouse and human genomes by combining BAC-end sequencing with a whole-genome BAC contig created by BAC fingerprinting, revealing remarkable colinearity of the mouse and human genomes. Approximately 11% of mouse BESs had BLASTN hits ($E < e^{-5}$) in the human genome, permitting a framework for the assembly of whole-genome shotgun sequence and a minimum tile path of clones useful for generating and finishing the reference sequence. Although other methods have been used to generate contigs for targeted comparative sequencing (Thomas et al. 2002), large-scale BAC-end sequencing is currently the most efficient strategy for building whole-genome comparatively anchored physical maps in map-poor species. In this report we describe creation of the first BAC clone-based comparative map of the cattle and human genomes that has 32% coverage of the cattle genome and one comparatively anchored BAC-end every 217 Kbp of human genome sequence on average. Furthermore, we confirm in silico predictions of cattle chromosome location of BAC-ends using radiation hybrid (RH) mapping, simultaneously demonstrating that the collection of sequenced BAC-ends is a powerful resource for generating high-resolution comparative

[5]Corresponding author.
E-MAIL h-lewin@uiuc.edu; FAX (217) 244-5617.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1560203.

13:1966–1972 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

Table 1. Cattle BAC-End Sequencing Statistics

Project statistic	
BAC end sequencing reads	80,448
BESs[a]	63,541 (79%)
BESs after trimming[b]	60,547 (95.3%)
Average read length, bp	515
BESs mate-pairs[c]	24,855 (82%)
Total bases sequenced	31,250,134
Repeat masked bases	14,499,057 (46.4%)
Redundant BESs[d]	6.3%

[a]Number of sequences with edited read length ≥200 bp. [b]Number of sequences free of *E. coli* and vector DNA, with edited read length ≥200 bp. [c]Number of clones with both ends successfully sequenced. [d]Repeat-masked BESs that have >90% identity with other BESs and alignments >100 bp.

chromosome maps. These maps will facilitate assembly of the cattle genome sequence and the identification of genes affecting phenotypes of biological and economic importance.

RESULTS

BAC-End Sequencing Statistics

A total of 40,224 BAC inserts (~1× coverage of the cattle genome) were sequenced at both ends, resulting in 63,541 high-quality BESs >200 bp in length (79% overall success rate). Sequencing failures were mainly due to empty wells and low BAC DNA yield (Table 1). Trimming for *E. coli* and vector DNA yielded 60,547 (95.3%) BESs with an average read length of 515 bp, comparing favorably to mouse BAC-end sequencing results (Zhao et al. 2001). Of these, 49,710 (82.1%) were mate-pairs. Within the set of 60,547 BESs, 67% contained repeats. Repeat bases represent 46.4% of the 31,250,134 bp cumulative length of BAC-end sequences, with satellite DNA comprising 6.2% of that total. Most repetitive elements are LINE (28.7%), SINE (7.5%), and long terminal repeats (2.4%). Novel repeats (data not shown) accounted for the 6.3% of BESs that were redundant (identity cutoff >90% and hit length >100 bp using BLASTN). Assuming a genome size of 3 Gbp for cattle, and random sampling from the cattle genome, 60,547 BESs correspond to one BES every 50 Kbp of the cattle genome.

Anchoring of Cattle BESs to the Human and Mouse Draft Genome Sequences

After repeat masking, 46,621 BESs had ≥100 bp of contiguous nonrepetitive sequence. These cattle BESs were tested for similarity to human genome sequence (NCBI Build 30) using BLASTN. Search results produced 13,716 BLAST hits ($E < e^{-5}$), of which 10,732 matched a single position in the human genome (Table 2 and Supplementary Table 1, available online at www.genome.org). In comparison, there were only 4713 significant BLAST hits in the mouse genome that yielded 3704 single hits, of which 3239 are common with the human genome. Thus, random cattle BESs had 3.3-fold higher frequency of similarity hits to the human genome than the mouse genome, and 2.9-fold higher frequency of single hits. Analysis of the distribution of these single BLAST hits demonstrated that 6665 (62%) of the human hits and 2228 (60%) of the mouse hits were within known genes, including coding and noncoding regions (Table 2). Among 24,855 BAC clones with BES mate-pairs, 1242 had both ends matching human genome sequence, of which 1011 had ends <300 kb apart on the same human chromosome. The median distance of 184.3 Kbp between BES mate-pairs with human chromosome hits compares favorably to the average insert size of 180 Kbp for segment 1 of the CHORI 240 library. Many of the remaining 95 clones with paired-end hits ≥300 Kbp apart on the same chromosome most likely span internal rearrangements in cattle chromosomes relative to the human genome. This is supported by the predicted RH map location in cattle for these clones (Suppl. Table 1). There were also 136 mate-pairs (10%) with ends that hit different human chromosomes. At least three of these clones appear to cross evolutionary breakpoints, with the remainder likely to represent inaccurate BLAST predictions, colligation of BACs, and other sources of experimental error.

The distribution of significant cattle BES hits on the human chromosomes was generally as expected on the basis of chromosome size and gene density (Table 3). If we assume a human genome size of 2.973 Gbp (HSAY excluded in estimate), then the 13,716 cattle BES hits are distributed ~1/217 Kbp of human genome sequence on average (not accounting for paired ends). This estimate includes human centromeric and telomeric regions, which are comprised of repetitive sequences that are generally not suitable targets for BLAST similarity searches. A general estimate of the nonoverlapping fraction of the human genome covered by the 13,716 cattle BACs is 32.3%, ranging from 20.0% for HSA22 to 36.6% for HSA17 (Table 3).

Mapping of Cattle BESs In Silico

We used a new version of COMPASS software (see Methods) to predict the cattle chromosome location of the 13,716 cattle BESs with significant BLAST hits in the human genome (Table 3). The basis for predicting cattle chromosome and bin location was the first-generation human–cattle RH comparative map (Band et al. 2000). The positions of 51 BLAST hits could not be predicted because they were located in "unlinked" human contigs, that is, contigs not assigned to specific coordinates on any human chromosome. For the remaining 13,665 BLAST hits, chromosome and comparative bin on the cattle RH map were predicted for 9290 (68%), with the rest falling into gaps in coverage on the cattle–human comparative map. The efficiency of predictive mapping of BESs on the cattle genome ranged from 43% for BLASTN hits on HSA8 to 100% for BLASTN hits on HSA13, HSA17, HSA18, and

Table 2. BLASTN Comparison of Cattle BESs With Human and Mouse Genome Sequences

BLAST results	Human genome	Mouse genome
BESs used for comparative analysis[a]	46,621	46,621
BESs with significant BLAST hits[b]	11,877	4,103
Significant BLAST hits[c]	13,716 (29.4%)	4,713 (10.1%)
Single BLAST hits	10,732 (90.4%)	3,704 (90.2%)
Gene hits[d]	8,621 (62.9%)	2,864 (60.8%)
Gene hits in single BLAST hit set	6,665 (62.1%)	2,228 (60.2%)
Paired-ends, both with BLAST hits	1,242	183
Consistent paired-end BLAST hits[e]	1,106	172
Consistent paired-end BLAST hits with <300 Kbp distance in reference genome	1,011	162

[a]Number of BESs with ≥100 bp of contiguous nonrepetitive cattle sequence. [b]Number of BESs with BLAST hits having $E < e^{-5}$. [c]Total number of BLAST hits generated from 11,877 BESs. [d]Number of BLAST hits annotated as genes in the human or mouse genome. These regions include coding and noncoding sequences. [e]Number of paired-end hits where both ends matched the same human/mouse chromosome.

Table 3. Distribution of BES BLAST Hits in the Human Genome, Their Predicted Cattle Chromosome Locations, and Genome Coverage

HSA	Chr. size (Mb)	No. hits	Predicted cattle chromosomes (no. hits)	No. in gaps (%)	% Coverage[a]
1	246	1137	BTA2 (95), BTA3 (581), BTA16 (230), BTA28 (19)	212 (18.6)	29.6
2	240	1279	BTA2 (294), BTA11 (476)	509 (39.8)	35.5
3	194	999	BTA1 (428), BTA22 (51)	520 (52.0)	34.2
4	192	827	BTA6 (458), BTA17 (104)	265 (32.0)	31.3
5	181	936	BTA7 (282), BTA20 (236)	418 (44.6)	34.6
6	170	795	BTA9 (180), BTA23 (234)	381 (47.9)	34.9
7	157	716	BTA4 (470), BTA25 (16)	230 (32.1)	33.8
8	143	635	BTA8 (10), BTA14 (208), BTA27 (55)	362 (57.0)	32.8
9	132	552	BTA8 (294), BTA11 (37)	221 (40.0)	32.1
10	134	652	BTA13 (84), BTA26 (221), BTA28 (122)	225 (34.5)	34.9
11	137	673	BTA15 (365), BTA29 (17)	291 (43.2)	34.5
12	131	611	BTA5 (491), BTA17 (71)	49 (8.0)	34.3
13	113	397	BTA12 (397)	0	28.0
14	104	529	BTA10 (229), BTA21 (51)	249 (47)	32.2
15	99	449	BTA10 (227), BTA21 (108)	114 (25.3)	28.5
16	81	348	BTA18 (27), BTA25 (110)	166 (47.7)	29.4
17	80	372	BTA19 (372)	0	36.6
18	77	342	BTA24 (342)	0	32.9
19	60	259	BTA7 (91), BTA18 (64)	104 (40.0)	30.2
20	62	284	BTA13 (284)	0	31.8
21	44	153	BTA1 (153)	0	23.0
22	47	155	BTA5 (43), BTA17 (53)	59 (38.10	20.0
X	149	565	BTAX (565)	0	29.7
Total	2,973	13,665		4,375 (32.0)	32.3

[a]Coverage of human genome by cattle BACs was estimated by summing the lengths of human genome sequences between cattle BESs for which both ends matched the same human chromosome with a distance <300 Kbp and the number of BESs having only one high-confidence hit in the human genome at a distance ≥180 Kbp multiplied by 180 Kbp. The coverage within the cattle genome should be approximately the same, given that 91% of all paired-end cattle BESs with BLAST hits on the same chromosome were separated by <300 Kbp.

HSAX; chromosomes that appear to have complete conserved synteny with their cattle homologs.

Confirmation of In Silico Map Predictions by RH Mapping of Cattle BESs

Radiation hybrid mapping was used to confirm COMPASS predictions of cattle chromosome positions of BESs, thus testing the validity of the COMPASS approach for detecting true orthologous relationships. In addition, the usefulness of BESs for creating human–cattle comparative maps with enhanced detail was investigated. As an exemplar chromosome, the 673 cattle BESs with similarity to HSA11 were used as a resource pool of markers for mapping on the cattle–hamster RH_{5000} panel. The 673 conserved cattle BESs are evenly distributed across the entire 137 Mbp length of HSA11, with the exception of the centromere region, yielding an average hit spacing of 1 cattle BES/203.5 Kbp of HSA11 sequence (Fig. 1). A total of 109 BESs were chosen for mapping on the RH panel, with a spacing of ~1 Mbp on HSA11. Of the 109 BESs selected, 97 (89%) gave a distinct PCR product that was acceptable for RH mapping. Out of a total of 176 comparatively anchored markers distributed on HSA11, 84 BESs were mapped on either BTA15 or BTA29; the rest are genes mapped previously (Band et al. 2000; A. Everts-van der Wind, S. Kata, J.E. Womack, and H.A. Lewin, unpubl.). Eleven of the remaining 13 BESs were linked at LOD <8.0, and two were linked to other chromosomes. Analysis of a more current build of the human genome revealed that the markers mapping to other chromosomes are assembly errors in Build 29. Therefore, a conservative estimate of the accuracy of the BLAST-COMPASS approach is 86.6% (84/97), but finishing of the human genome sequence and improvements to the cattle RH map will likely result in further gains in chromosome/bin prediction accuracy. We term cattle BESs that have significant homology to human genome sequence, and with RH mapping information, "comparatively anchored sequence tagged sites" (CASTS). The CASTS were incorporated into RH maps of BTA15 and BTA29, and detailed comparative maps were constructed (Fig. 1). The distribution of CASTS on the cattle RH map was as expected based on their order in the human genome, with two exceptions on BTA29 (BZ942629 and BZ932447). These exceptions are changes of order within conserved segments, not between conserved segments, thus suggesting the possibility of additional internal rearrangements or genotyping errors for these markers. The comparative map of HSA11 covers 113.56 Mbp (85%) in 10 distinct conserved segments, with an average comparative spacing of 822 Kbp between markers. The smallest gap size between adjacent conserved segments is 180 Kbp, and the largest is 5.96 Mbp. Visual inspection of the comparative organization of HSA11, BTA15, and BTA29 reveals a patchwork shuffling of segments that is not explained by a simple translocation of HSA11 into two syntenic segments followed by multiple internal rearrangements.

The high degree of accuracy of the COMPASS predictions permits the construction of a virtual map of BESs on the cattle chromosomes. By identifying the coordinates of the BESs in the human genome and using COMPASS to identify the closest linked marker in the RH maps of BTA15 and BTA29, most (89%) of the 673 BESs with hits on HSA11 could be ordered on these cattle chromosomes (Fig. 1). This virtual map will be useful for selecting a tiling path for sequencing the cattle genome and for comparative proofing of BAC contigs produced by fingerprinting (Gregory et al. 2002).

The Borders of Conserved Synteny

A three-way comparative map of mouse and cattle chromosomes overlaid on HSA11 was constructed using a combination of results from BLAST of cattle BAC ends against the human and mouse genomes, and COMPASS (Fig. 1). The mouse chromosome segments were defined from BLASTN hits of cattle BAC-ends in the mouse genome and then converting these hits to coordi-

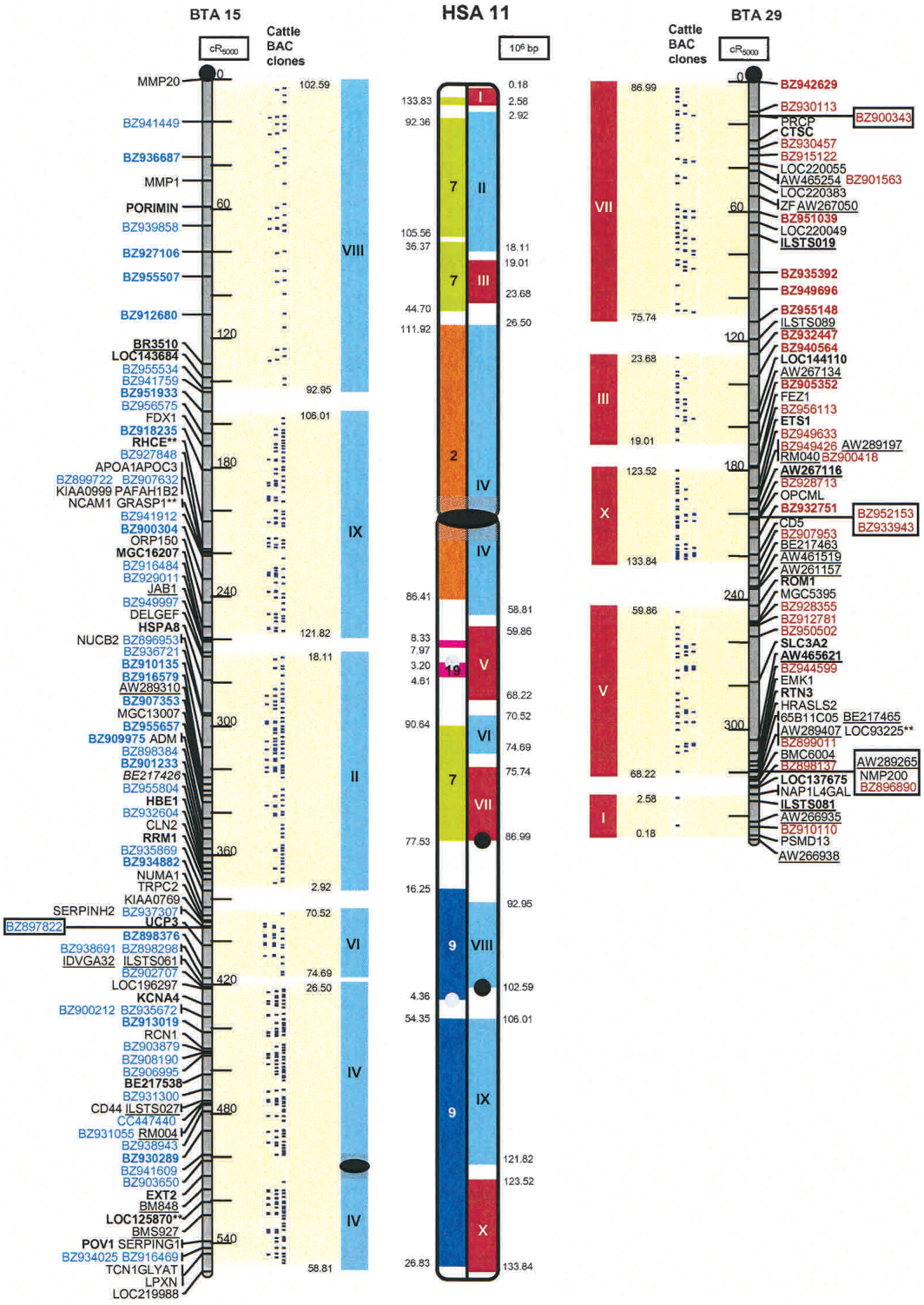

Figure 1 High-resolution cattle- and mouse-on-human comparative map of HSA11 created with current RH maps of BTA15 and BTA29 and 84 new cattle BESs. Comparative coverage on the current cattle RH_{5000} map is shown in solid red or blue blocks on the *right half* of the HSA11 drawing, with comparative segments indicated in Roman numerals. Mouse-on-human segments and mouse chromosome number are shown within the *left half* of the HSA11 drawing, with the sequence coordinates of the corresponding positions in the mouse genome indicated to the *left*. The distribution of BACs predicted to conserved segments of BTA15 and BTA29 is shown between the RH maps and human chromosome segments. Mapped conserved BAC ends (CASTS) increased the total comparative coverage and enabled merging of linkage groups for both cattle chromosomes. The CASTS are identified by GenBank accession number and are colored blue or red on chromosome maps. The human centromere is indicated by black ovals, and the 3.37-Mbp pericentric region on HSA11 with no BLAST hits is shaded. Black and gray circles indicate the location of centromeres on cattle and mouse chromosomes, respectively. Placement of markers appearing in boxes is defined by two-point linkage, because their order could not be determined with certainty. Framework markers are in bold text. Underlined markers are those with no hit against the human genome (microsatellites, and novel ESTs). Markers with two asterisks (**) are not predicted by COMPASS to be in the observed segment on the basis of the available comparative map. Comparative maps were drawn using NCBI Build 29 (April 2002) of the human genome.

nates in the human genome on the basis of existing comparative mapping information. Therefore, the displayed comparative mouse and cattle segment lengths may differ somewhat because of the incomplete overlap of BLAST hits against the two reference genomes (Table 2). In addition, the representation of the mouse-on-human map is partially complete because the segment boundaries are defined by hits with cattle BESs and not by the actual segment boundaries defined by mouse genome-to-human genome comparison. The map was constructed this way in order to properly represent the comparative maps that can be built solely with the cattle BES data. Even so, the boundaries of the 10 conserved segments in the cattle genome appear to correspond well to the boundaries of nine conserved syntenic segments in the mouse genome that correspond to HSA11 (Fig. 1). Segment I is located at the telomere of HSA11, BTA29, and MMU7. Segment IV shows nearly complete overlap in all three species, with the boundaries of the segment appearing identical on BTA15 and MMU7. The boundaries of segment VIII appear nearly identical on BTA15 and MMU9, with the additional interesting feature that the centromeres of BTA15 and MMU9 are located in a position that appears conserved in mouse and cattle (BTA15 and MMU9 are acrocentric). One boundary of segment VII on BTA29 has a centromere in cattle, but the same boundary on MMU7 is acentric. Other conserved segments in cattle appear as subsegments of larger conserved segments in the mouse (e.g., segments VI and VII, IX and X). These first multi-ordinal sequence-based comparative map alignments provide firm evidence for segmental evolution of mammalian chromosomes across a time span of ~85 million years.

DISCUSSION

Providing functional and evolutionary context to DNA sequences is a major goal of comparative genomics. Among mammals, it is expected that comparative genomics will lead to greater understanding of phenotypic, physiological, and metabolic diversity at the molecular and systems levels. Thus, sequenced mammalian genomes are the reference standards that will enable comparative genomics to be used as a tool to understand the molecular basis for the vast phenotypic diversity within this vertebrate class (O'Brien et al. 2001). Each mammalian genome sequenced, particularly those that are chosen from phylogenetically distant clades, adds another piece of the evolutionary puzzle that is necessary to unveil the coding potential of the human and other mammalian genomes (Thomas and Touchman 2002). In addition, the great importance of cattle to the world's supply of meat and dairy products and the global ecosystem makes this organism a prime target for genome studies.

Approximately 60,500 high-quality cattle BESs with an average read length of 515 bp were generated from end-sequencing of 40,224 random BAC clones. These BESs comprise more than 14 Mbp of nonrepetitive cattle DNA, thus providing a resource for anchoring cattle genomic sequences to the human and mouse genomes. A somewhat surprising result was the frequency of significant BLASTN hits in the human genome (29%) compared with the mouse genome (10%) among filtered and repeat-masked cattle BESs. These data are corroborated by the results of Gregory et al. (2002) who found ~11% BLAST hits of mouse BESs against human genome sequence using the same BLAST parameters (Gregory et al. 2002). The obvious explanation for this observation is that the cattle genome has diverged more from the mouse genome than it has from the human genome, perhaps because of the shorter generation interval in the rodent lineage. In the mouse and human genomes, approximately 60% of the cattle BES BLAST hits are in genes, and roughly 90% have single-match hits (Table 2). Clearly then, a greater number of both conserved coding and noncoding sequences are detected in the human genome compared to the mouse, supporting the idea that the human and cattle genomes have diverged to a lesser extent. It remains to be determined whether this represents greater similarity in the biology of these two species or differences in the rate of sequence divergence in evolutionary time. Regardless of which theory is correct, the large number of conserved noncoding cattle BESs will be useful for identification of noncoding regulatory elements and structural motifs in the human and cattle genomes, as suggested by more limited comparative sequencing results (Thomas et al. 2002).

The fraction of repeat bases in our set of cattle BESs is 47%, which is higher than the 37% found in mouse BESs (Zhao et al. 2001). This difference might be due to the restriction enzyme used to make the BAC libraries; *Mbo*I was used for CHORI-240 (cattle) and RPCI-24 (mouse), and *Eco*RI was used to prepare RPCI-23 (mouse). However, the differences observed are more likely to be due to the greater number of SINEs and LINEs in the cattle genome. The large number and wide dispersal of repeats in the cattle genome will undoubtedly complicate the assembly of a cattle whole-genome shotgun sequence, thus increasing the level of importance of obtaining high-quality BAC-based physical and comparative maps. Unfortunately, repetitive elements are not only a nuisance for assembling genome sequences, they also complicate comparative analyses of BESs. The problem of repetitive sequences reduced the effective amount of total DNA available for similarity search by nearly half, and the number of BAC-ends by 23%.

Despite the complications caused by repetitive elements, BAC-end sequencing was still found to be a robust and efficient means of anchoring mammalian genomes to each other using the combined BLAST-COMPASS approach. A total of 13,716 significant BLAST hits in the human genome were identified, of which 10,732 are single hits, yielding one cattle BES for every 217 Kbp of human genome sequence. Of these 2484 (1242 BAC clones) are mate-pairs, of which 2022 (81%) are separated by <300 Kbp on the same chromosome. These results give high confidence to the BLAST analysis. Furthermore, these clones cover ~185 Mbp of cattle DNA and will thus provide an excellent resource for assembly of BAC fingerprint contigs and in silico prediction of gene content based on comparative mapping information. The remaining 95 mate-pairs with BLAST hits on the same human chromosome, but ≥300 Kbp apart, may be useful for defining rearrangements within conserved syntenies. Three additional BAC clones were identified that likely cross evolutionary translocation breakpoints (data not shown) which, when sequenced, will facilitate comparative analysis of these important chromosome domains. The coverage was less for the mouse (~1 cattle BES/527 Kbp mouse sequence) because there were about threefold fewer BLAST hits against the mouse genome. Nevertheless, reasonably good coverage of the mouse genome was obtained, as demonstrated by the overlay of MMU2, MMU7, MMU9, and MMU19 segments on HSA11 (Fig. 1).

The COMPASS approach (Band et al. 1998; Ma et al. 1998; Rebeiz and Lewin 2000) was used to predict the cattle chromosome location (Table 3) and comparative bin of all BESs with a single significant BLAST hit in the human genome (see Suppl. Table 1). The regions not covered in the comparative map include those surrounding human centromeres and the human Y chromosome. The centromeric regions represent a problem for comparative analysis because assembled human sequence is not yet available for most of them, and even when it is, such regions are comprised mostly of repeats that will be masked in a typical BLAST analysis. For example, HSA1 has a gap in sequence coverage of ~ 21.5 Mb in the centromeric region. Such gaps reduce the apparent comparative coverage because they are counted in the

human chromosome size estimates. With respect to the Y chromosome, most of the significant BLAST hits on HSAY had hits on other human chromosomes with more significant E-values. Some of these hits were in the pseudoautosomal region. Because of these problems the Y chromosome was excluded from the present analysis. The proportional representation of BLAST hits on all human autosomes and HSAX, and the moderate resolution of the first-generation cattle–human comparative RH map (Band et al. 2000), allowed prediction of the map location of 68% of all cattle BESs (Table 3). The COMPASS strategy thus permits in silico mapping and ordered alignment of BACs, as shown for BTA15 and BTA29 (Fig. 1). The coverage of these BACs in the cattle genome is 32.3%, thus providing a solid framework for scaffolding BAC contigs in the cattle genome as done for the mouse genome (Gregory et al. 2002) and for cattle chromosome-based BAC skim sequencing.

The accuracy of COMPASS and the utility of the BESs for creating high-resolution comparative maps of cattle, human, and mouse chromosomes was demonstrated by RH mapping of cattle BESs with BLAST hits on HSA11 (Fig. 1). The accuracy of cattle chromosome prediction was found to be ~90%, and a cattle–human comparative map with greater than 1-Mbp resolution was created by adding 84 BAC ends to an existing cattle RH map. No gap was greater than 6 Mbp, and comparative coverage is 85% on the basis of human genome coordinates. These results demonstrate that BESs combined with COMPASS and RH mapping, yielding large numbers of CASTS, represent an exceptionally powerful approach for generating detailed comparative maps that should be extensible to other cattle chromosomes and to whole-genome comparative mapping in map-poor species.

The cattle–human comparative map generated with CASTS revealed 10 distinct conserved segments that can be aggregated into two conserved syntenies on BTA15 and BTA29. The conserved segments range in size from 2.4 Mbp (segment I) to 32.31 Mbp (segment IV, including the centromere) with a mean of 11.41 Mbp. These results are in general agreement with genome averages for the mouse–human comparison of 6.9 Mbp based on direct sequence comparison (Waterston et al. 2002). A startling observation was made when the mouse, human, and cattle segments were aligned (Fig. 1), clearly showing conservation of several segmental boundaries. Thus, it appears from the conserved segmental boundaries that breakages occur within defined chromosomal regions. The apparent stability of segment boundaries over an evolutionary time scale exceeding 85 million years can be due to insufficient time for additional breakages within conserved segments to occur and/or purifying selection acting against breakages within gene-dense regions. Chromosomes appear to evolve as mosaics of these segments, with insertion of centromeres occurring at or near these boundaries. Our data clearly show that the gaps between segments have very low levels of sequence similarity or are masked by repetitive elements, because the gaps are devoid of BES hits in both the human and mouse genomes (Fig. 1). We have also noted that narrower regions of the segment gaps contain gene deserts in the human genome, which would be desirable with a segment-shuffling evolutionary strategy so as not to produce gene loss. The continued progress in sequenced-based mapping of the dog (Mellersh et al. 2000), cat (Murphy et al. 2000), pig (Rink et al. 2002), and other mammalian genomes will enable a more comprehensive evaluation of the conserved segment model of mammalian chromosomal evolution.

The long history of cattle genetics (Fries and Ruvinsky 1999) and a vibrant community of cattle genome scientists greatly facilitate the development of the maps and other genomic resources necessary for genome sequencing and genomic biology. The work we report here provides the first sequence-based physical and multispecies comparative maps for cattle. These maps and accompanying BAC resources will enable robust assembly of the cattle genome sequence. Furthermore, comparatively anchored BESs provide a powerful tool for evolutionary studies of vertebrate genome organization and an important new resource for positional cloning of genes affecting health and production traits in this agriculturally important species.

METHODS

BAC Culture and End Sequencing

Segment 1 of the CHORI 240 cattle BAC library (http://www.chori.org/bacpac) was used for BAC-end sequencing. The library, consisting of approximately 200,000 clones, was created by cloning of partially digested *Mbo*I genomic DNA isolated from a Hereford bull into the *Bam*HI cloning site of the pTARBAC1.3 vector. Preculture was used to increase BAC DNA yields for end sequencing. Briefly, BAC clones were inoculated into 2-mL 96-well culturing blocks (Millipore) containing 1 mL of 2× Luria Broth and 12.5 µL/mL chloramphenicol. Blocks were covered with plastic lids and incubated for 21 h at 37°C with shaking overnight at 320 rpm. New blocks with 1.5 mL of fresh media were inoculated with 3 µL of the preculture and incubated overnight for 17 h as described above. The blocks were centrifuged at 1500*g* for 7 min, the culture supernatant was decanted, and the blocks were inverted and tapped on absorbent paper to remove residual culture supernatant. DNA extractions were performed robotically (QIAGEN 9600 BioRobot) using a commercial kit (Montage BAC_{96} Miniprep Kit, Millipore) according to the manufacturer's specifications. BAC DNA was transferred to V-bottom plates and stored at 4°C.

Dye terminator sequencing reactions were set up robotically using the following ingredients: 10 µL BAC DNA solution, 2 µL standard T7 sequencing primer (20 µM solution) or custom-designed SP6 (GGCCGTCGACATTTAGGTGACA) primer (10 µM solution), 2 µL BigDye cycle sequencing-ready reaction mix with AmpliTaq DNA polymerase FS (Applied Biosystems), and 6 µL of dilution buffer (200 mM Tris-HCl, pH 9.0, 5 mM $MgCl_2$). Thermal cycling (MJ Research) was performed at 96°C for 4 min followed by 100 cycles of 96°C for 30 sec, 56°C for 10 sec, 60°C for 4 min. Reaction products were purified by precipitation in 75 µL of 0.3 mM $MgSO_4$ in 70% EtOH, followed by washing in 75 µL of 70% EtOH. The samples were then loaded on ABI 3700 automated capillary sequencers (Perkin Elmer-Applied Biosystems).

Sequence Processing and Bioinformatics

The BAC-end sequences were trimmed of vector sequences, stored in a local Oracle database, and uploaded to GenBank (accession nos. BZ896445–BZ956676 and CC447354–CC447937) after quality assessment using the Genome Project Management System (GPMS), a local laboratory information management system for large-scale DNA sequencing projects (Liu et al. 2000). Quality assessment was performed using Phred software (Ewing and Green 1998; Ewing et al. 1998) using Q ≥20 as a cutoff. Repeats were masked using RepeatMasker software (ftp://ftp.genome.washington.edu) and BESs tested for similarity using NCBI-BLASTN (Altschul et al. 1997) mounted on an SGI/Cray 2000 16-processor supercomputer. Target databases were Build 30 of the human genome draft sequence (June 2002 release) and the mouse genome database (MGSCv3). An expectation value (E) of e^{-5} was used as the significance threshold for comparison of cattle BESs with the human and mouse sequence contigs. This empirically chosen threshold was shown to identify orthologs with ~95% accuracy (Band et al. 2000).

The COMPASS strategy (comparative mapping by annotation and sequence similarity) permits the prediction of chromosome map location based upon sequence similarity of orthologous genes, if comparative map information is available for two species (Ma et al. 1998; Band et al. 2000; Rebeiz and Lewin 2000). A new version of the COMPASS Perl scripts (COMPASS III) was created specifically for predicting cattle chromosome and bin

location of BESs. The COMPASS III script utilizes BLASTN output to first identify the sequence coordinates in the human genome of each BAC end, and then predicts cattle chromosome and comparative bin using data from the first-generation cattle–human comparative RH map (Band et al. 2000). Up to three significant BLAST hits are stored for each BES. The Perl script parses BLAST output into a spreadsheet, and transforms hit positions within the human contigs to actual positions in the assembled human genome sequence on the basis of the NCBI file *seq_contig.md*. The COMPASS III predictions were integrated with the comprehensive information and annotation for each end-sequenced BAC clone.

Radiation Hybrid Mapping

The database of annotated BESs was used for selection of primers for RH mapping of those BESs with significant BLASTN hits on HSA11. A total of 109 BESs were selected at approximately 1-Mbp intervals on HSA11. Only those BESs with a single high-confidence human hit on HSA11 were selected for primer design. This is to avoid paralogs and duplicated segments that might complicate mapping and interpretation. Oligonucleotide primers were designed to discriminate cattle from rodent sequences in the RH panel using Vector NTI v7.0 software (InforMax). For the 109 primer pairs selected for the HSA11 pilot study, 89% gave distinct PCR products using the cattle–hamster RH_{5000} panel. Primers were purchased from a commercial source (Operon). All primer pairs were typed in duplicate against a cattle 5000 rad RH panel (Womack et al. 1997) as described (Band et al. 1998). Two-point linkage and multipoint map construction were carried out with RHMAPPER 1.22 (Slonim et al. 1997) using procedures similar to those described (Band et al. 2000). A threshold of LOD 8.0 was used for BESs with COMPASS-predicted assignments to one cattle chromosome. For BESs falling into gaps on the comparative map, a threshold for linkage of LOD 12.0 was used. Initial framework maps were created using the RHMAXLIK program of RHMAP 3.0 (Boehnke 1992) with an LOD threshold of 3.0. This order was further expanded using the grow frameworks option of RHMAPPER, and finally, a placement map incorporated all remaining markers in the most likely framework intervals within 40 cR. Two-point linkage analysis was used to confirm the position of markers that mapped outside terminal framework markers. Microsatellite markers incorporated in the maps were used to orient linkage groups within chromosomes according to the USDA-MARC linkage map (Kappes et al. 1997).

ACKNOWLEDGMENTS

This work was supported in part by grants from the USDA-National Research Initiative (AG99-35205-8534), the USDA Cooperative State Research Service (AG2002-34480-11828), and the USDA Agricultural Research Service (Agreement No. 58-5438-2-313).

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

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Band, M., Larson, J.H., Womack, J.E., and Lewin, H.A. 1998. A radiation hybrid map of BTA23: Identification of a chromosomal rearrangement leading to separation of the cattle MHC class II subregions. *Genomics* **53:** 269–275.

Band, M.R., Larson, J.H., Rebeiz, M., Green, C.A., Heyen, D.W., Donovan, J., Windish, R., Steining, C., Mahyuddin, P., Womack, J.E., et al. 2000. An ordered comparative map of the cattle and human genomes. *Genome Res.* **10:** 1359–1368.

Boehnke, M. 1992. Multipoint analysis for radiation hybrid mapping. *Ann. Med.* **24:** 383–386.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Fries, R. and Ruvinsky, A. 1999. *The genetics of cattle*, p. 710. CABI Publishing, Oxon, UK.

Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T.D., Itoh, T., Tsai, S.F., Park, H.S., Yaspo, M.L., Lehrach, H., Chen, Z. et al. 2002. Construction and analysis of a human–chimpanzee comparative clone map. *Science* **295:** 131–134.

Gregory, S.G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C.E., Evans, R.S., Burridge, P.W., Cox, T.V., Fox, C.A., et al. 2002. A physical map of the mouse genome. *Nature* **418:** 743–750.

Kappes, S.M., Keele, J.W., Stone, R.T., McGraw, R.A., Sonstegard, T.S., Smith, T.P., Lopez-Corrales, N.L., and Beattie, C.W. 1997. A second-generation linkage map of the bovine genome. *Genome Res.* **7:** 235–249.

Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Liu, L., Roinishvili, L., Pan, X., Liu, Z., and Kumar, C. 2000. GPMS: A web based genome project management system. *The Proceeding of 4th World Multiconference on Systematics, Cybernectics, and Informatics SCI2000* 62–67.

Ma, R.Z., van Eijk, M.J., Beever, J.E., Guérin, G., Mummery, C.L., and Lewin, H.A. 1998. Comparative analysis of 82 expressed sequence tags from a cattle ovary cDNA library. *Mamm. Genome* **9:** 545–549.

Mellersh, C.S., Hitte, C., Richman, M., Vignaux, F., Priat, C., Jouquand, S., Werner, P., André, C., DeRose, S., Patterson, D.F., et al. 2000. An integrated linkage-radiation hybrid map of the canine genome. *Mamm. Genome* **11:** 120–130.

Murphy, W.J., Sun, S., Chen, Z., Yuhki, N., Hirschmann, D., Menotti-Raymond, M., and O'Brien, S.J. 2000. A radiation hybrid map of the cat genome: Implications for comparative mapping. *Genome Res.* **10:** 691–702.

O'Brien, S.J., Eizirik, E., and Murphy, W.J. 2001. Genomics. On choosing mammalian genomes for sequencing. *Science* **292:** 2264–2266.

Rebeiz, M. and Lewin, H.A. 2000. COMPASS of 47,787 cattle ESTs. *Anim. Biotechnol.* **11:** 75–241.

Rink, A., Santschi, E.M., Eyer, K.M., Roelofs, B., Hess, M., Godfrey, M., Karajusuf, E.K., Yerle, M., Milan, D., and Beattie, C.W. 2002. A first-generation EST RH comparative map of the porcine and human genome. *Mamm. Genome* **13:** 578–587.

Slonim, D., Kruglyak, L., Stein, L., and Lander, E. 1997. Building human genome maps with radiation hybrids. *J. Comput. Biol.* **4:** 487–504.

Thomas, J.W. and Touchman, J.W. 2002. Vertebrate genome sequencing: Building a backbone for comparative genomics. *Trends Genet.* **18:** 104–108.

Thomas, J.W., Prasad, A.B., Summers, T.J., Lee-Lin, S.Q., Maduro, V.V., Idol, J.R., Ryan, J.F., Thomas, P.J., McDowell, J.C., and Green, E.D. 2002. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* **12:** 1277–1285.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Womack, J.E., Johnson, J.S., Owens, E.K., Rexroad III, C.E., Schläpfer, J., and Yang, Y.P. 1997. A whole-genome radiation hybrid panel for bovine gene mapping. *Mamm. Genome* **8:** 854–856.

Zhao, S., Shatsman, S., Ayodeji, B., Geer, K., Tsegaye, G., Krol, M., Gebregeorgis, E., Shvartsbeyn, A., Russell, D., Overton, L., et al. 2001. Mouse BAC ends quality assessment and sequence analyses. *Genome Res.* **11:** 1736–1745.

WEB SITE REFERENCES

http://www.genome.gov/page.cfm?pageID=10002154; National Human Genome Research Institute.

http://www.chori.org/bacpac; BACPAC Resources Center.

ftp://ftp.genome.washington.edu; RepeatMasker software.

Received May 19, 2003; accepted in revised form June 10, 2003.