The cattle genome reveals its secrets

Journal of Biology

BioMed Central

Minireview

The cattle genome reveals its secrets

David W Burt

Address: Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Midlothian EH25 9PS, UK.
Email: Dave.Burt@roslin.ed.ac.uk.

Published: 24 April 2009

Journal of Biology 2009, **8**:36 (doi:10.1186/jbiol137)

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/8/4/36

Abstract

The domesticated cow is the latest farm animal to have its genome sequenced and deciphered. The members of the Bovine Genome Consortium have published a series of papers on the assembly and what the sequence reveals so far about the biology of this ruminant and the consequences of its domestication.

Cattle belong to an ancient group of mammals, the Cetartiodactyla, that first appeared around 60 million years ago. Domesticated cattle (*Bos taurus* and *Bos taurus indicus*) diverged from a common ancestor 250,000 years ago, and have had a long and rich association with human civilization since Neolithic times 8,000-10,000 years ago. All modern cattle breeds originate from large populations of the ancestral aurochs (*Bos taurus primigenius*; Figure 1) through thousands of years of domestication. During this time, more than 800 cattle breeds have been established, representing an important resource for understanding the genetics of complex traits in ruminants. More than a billion cattle are raised annually worldwide for beef and dairy products, as well as for hides. Cattle therefore represent significant scientific opportunities, as well as an important economic resource.

Sequencing of the cattle genome began in December 2003, led by Richard Gibbs and George Weinstock at the Baylor College of Medicine's genome sequencing center in Houston, Texas, USA. The first draft sequence of the bovine genome was based on DNA taken from a Hereford dam, L1 Dominette 01449 (Figure 2), a cattle breed used in beef production. In parallel, a large number of single-nucleotide polymorphisms (SNPs) have also been generated from the partial sequence of six breeds (Holstein, Angus, Jersey, Limousin, Norwegian Red and Brahman). Taken together with the sequence of L1 Dominette 01449 (the reference bovine genome [1]) these represent a valuable resource for marker-assisted selection of genetic traits in commercial breeding programs.

The Bovine Genome Project represents a complex collaborative effort between multiple groups and funding from the United States, Canada, France, United Kingdom, New Zealand and Australia.

Undoubtedly the current bovine genome sequence will be improved in both its sequence coverage and its annotation, but this draft sequence will form the basis for cattle genetics and genomics for the next 20 years or more.

So what have we learned?

The genome assembly problem - still not solved?

The technology for generating raw sequence data has advanced rapidly over the past 35 years, starting with Sanger sequencing in the 1970s, automated fluorescent Sanger sequencing in the 1980s and, recently, ultra-high-

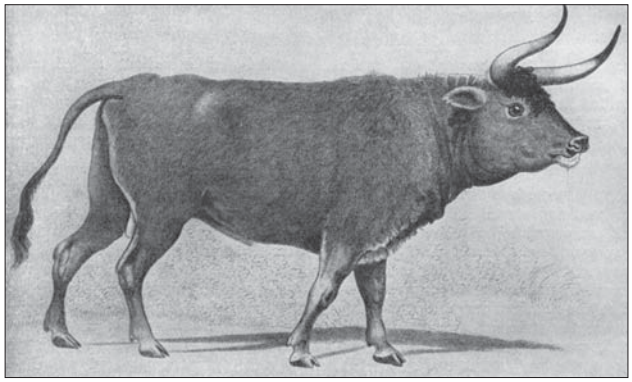

Figure 1
A picture of the ancestral aurochs (*Bos taurus primigenius*) taken from Brehms Tierleben (picture from Wikipedia).

Figure 2
This Hereford cow, known as L1 Dominette 01449, provided scientists with the first genome sequence for cattle.

throughput methods based on the parallel sequencing platforms produced by 454, Illumina, and ABI. However, the scale of these advances has not been matched by new algorithms and tools for sequence assembly, particularly for large genomes. Common problems associated with large genomes have been repetitive sequences (generally around 50% of a vertebrate genome), gene families and genetic polymorphisms, all of which can cause errors in assembly. Genome assembly is still a problem, requiring a combination of parallel computing and hard work from teams of manual annotators, and there is a need for a step change in the algorithms and approaches used to assemble a sequence. The bovine genome is the latest in a series of large-scale sequencing projects based on the conventional automated Sanger methods. It illustrates many of these problems and provides some solutions [2-3].

There are two bovine genome assemblies: BCM4 from Baylor College and UMD2 from the University of Maryland. Both assemblies are based on the sequence data generated by the Baylor genome sequencing center. How do they compare? Which is the more accurate?

BCM4 is the latest assembly from a series - BCM1 (2004), BCM2 (2005), and BCM3.1 (2006) - which claims to be more accurate, with greater coverage and fewer misassemblies than before. The earlier inaccuracies were due to the assemblies having been largely based on whole-genome shotgun (WGS) data alone: because of the sizes of the fragments generated in WGS sequencing, this is highly prone to errors caused by the repeated sequences that pose a significant problem in genome assemblies. BCM4 by contrast was assembled by combining WGS reads (sequencing of 30 million reads) with the reads and fingerprinted contig (FPC) maps of large genomic inserts cloned into bacterial artificial chromosomes (BACs). The large inserts allow the smaller fragments to be correctly assembled with fewer mistakes due to repetitive sequences [2]. The WGS reads ensure coverage of the whole genome. In addition, through the development of a new assembler (Atlas) the Baylor team was able to integrate these sequences with other data, from FPC BAC maps, genetic maps and chromosome assignments. The sequence data themselves were based on a sire and daughter, mostly on the daughter's DNA. Therefore, the coverage of the sex chromosomes X and Y is not as good as that of the autosomes, especially in the case of the Y chromosome, of which only a small amount of DNA was available from the single Y chromosome of the sire, whereas the two animals together provided three X chromosomes (and of course four of each of the autosomes) [2,3].

For BCM4, more than 90% of sequences have been assigned to a specific chromosome and total sequence assembled is 2.54 Giga base-pairs (Gbp). On the basis of overlaps with 1.04 million expressed sequenced tags (ESTs), the gene coverage is estimated at 95%. Comparisons between 73 fully sequenced BAC clones showed few misassemblies and more than 92% coverage. Finally, 99.2% of 17,482 SNPs have been mapped correctly onto the BCM4 assembly. The sequence of the bovine MHC (BoLA) provides a critical test of accuracy [4], as it contains many polymorphic gene families densely clustered on chromosome 23 and automated genome assembly software is prone to errors of deletion and duplication in such regions. The paper by Brinkmeyer-Langford *et al.* [4] shows extremely good agreement between the radiation hybrid (RH) map derived by mapping DNA markers from this region on RH panels and the BCM4 sequence assembly.

Figure 3
Bovinae have diverged into **(a)** cattle **(b)** antelope and **(c)** buffalo over a relatively short time period. (a) shows a domesticated cow (*Bos taurus*) (photograph by Daniel Schwen, Wikipedia), (b) is the Common Eland (*Taurotragus oryx*) (Ablestock) and (c) is a Cape Buffalo (*Syncerus caffer*) (Ablestock).

The University of Maryland's assembly, UMD2, is based on the same raw data as BCM4 and integrates a wider range of external data to improve and validate the final sequence assembly [3]. In particular, it uses comparison between the cattle and human genome sequences to orientate or place cattle contigs when the data from the cattle genome alone cannot. It has therefore been able to assemble more sequence (2.86 Gbp, with 91% of sequences assigned to a specific chromosome and some of the Y), with fewer gaps (for example UMD2 assigned 136 Mb to the bovine X chromosome and BCM4 only 83 Mb), fewer misassemblies and with SNP errors corrected (BCM4 may have threefold more errors than UMD2).

Accuracy was also improved in the UMD2 assembly by paired-end reads for regions containing segmental duplications, gene families and gene polymorphisms, where assembly is particularly error-prone. In a paired-end read, about 500 bp are sequenced at each end of a large BAC insert to place the insert on the genome map. If the length of the BAC insert fails to correspond to the distance between the sequences matching the two ends of the insert on the genome assembly, then a duplication or a deletion must have been introduced in the assembly. As a result of this analysis, the UMD2 group report only 662 segmental duplications compared with 3,098 for BCM4. Duplications can be due to copy-number variation, a focus of much current interest because of its association, in different cases, with genetic disease and with disease resistance. However, quantification of WGS reads in these regions did not suggest any over-representation that might indicate increased copy number. WGS should be over- or under-represented in the corresponding BCM4 sequences where the two assemblies disagree, and this should clearly be checked.

The use by the UMD2 assembly of comparative maps between cattle and human allowed more sequence to be assembled, but somewhat undermines conclusions based on human-bovine sequence comparisons. The data can, however, now be used to highlight potential problem areas or predict specific arrangements and guide more sequencing to generate bovine data to confirm these predictions. These studies will presumably go ahead in the coming months at Maryland, Baylor and elsewhere.

What these assemblies also illustrate is the benefit of and need for community support for the final success of a genome project. The cattle community provided DNA samples of breeds, chromosome assignments of specific contigs, genetic linkage maps, BAC and FPC BAC maps, EST libraries for gene prediction and genome annotations [1] for gene and protein predictions. However, the integration of datasets from multiple sources posed a substantial challenge for the bioinformaticians at Baylor College and Maryland in the absence of the genome sequence as a reference point.

Finally, we should ask what we can expect in the future. The availability of ultra-high-throughput sequence technologies will provide more raw sequence data, which could be used to fill in gaps, for example in regions not cloned in the current assembly. The extra reads would also increase the quality and number of SNPs detected by comparing several breeds, and increase the accuracy of sequence divergence and diversity estimates by providing some assurance that apparent SNPs are really SNPs and not sequencing errors.

Genome evolution

The availability of a cattle genome sequence with more than 95% coverage is an excellent resource for comparative and evolutionary biologists. In addition, physiologists and biochemists will be interested in the unique biology of ruminants specialized for converting low-grade forage into energy-rich fat, milk and muscle.

Figure 4
A phylogeny using unambiguous sites in the Bovinae results in three main groups: cattle, bison **(a)** and sister group, yak **(b)** and banteng **(c)**. (a) shows North American Bison (*Bison bison*), (b) Yak (*Bos grunniens*) and (c) Banteng (*Bos javanicus*). All photographs are from Ablestock.

Elsik and colleagues [1] have led the way to annotate the genome, to give it meaning in terms of genomic structure, genes and proteins. This was achieved using a combination of automated pipelines and 4,000 manual annotations, which were made as part of a 'Bovine Annotation Jamboree' as well as by dedicated teams of annotators. Analysis predicted 26,835 genes, of which 82% were validated from external data sources. This suggests that the bovine genome encodes at least 22,000 genes, which is broadly in line with gene counts in all other mammals. In addition, 496 microRNAs were detected, including 135 novel sequences.

Multiple species comparisons between the cow and other mammals define a core set of 14,345 orthologous genes, 1,217 of which are specific to placental mammals and missing in marsupials and monotremes. Comparative mapping with other mammalian genomes defines 124 evolutionary breakpoints, mostly associated with repetitive sequences and segmental duplications. Interestingly, genes associated with lactation and immune responses are also associated with these breakpoints. Does this suggest a selective advantage or simply a mechanism for expanding these gene families?

Comparisons between human and bovine coding regions aimed at identifying genes under strong selection define 2,210 genes with elevated dN/dS ratios (a measure of selective constraint on proteins). Seventy-one genes have dN/dS >1, and among these, not surprisingly, genes with roles in reproduction, lactation and fat metabolism are over-represented [1,5-6]. More surprisingly, they include genes encoding proteins of the immune system. These are the genes that distinguish the ruminants from other mammals, and may reflect special needs of ruminants, which retain the low-grade food they ingest, along with any associated pathogens, for up to a day in the rumen before releasing it into the intestines from which infectious organisms are readily expelled.

One of the novel features of the Bovine Genome Project has been to use the sequence to examine the evolution and process of domestication of cattle. The aims of these studies were to uncover more about phylogenetic relationships amongst the Bovinae and the importance of natural and artificial selection, and to identify genes or genomic regions that have been critical in the domestication process - the so called 'signatures of selection'.

The divergence of the Bovinae (antelope, buffalo and cattle; Figure 3) over a relatively short period makes it difficult to determine a robust phylogeny for this group. MacEachern *et al*. [8] have exploited cattle genomic sequences to design primers to amplify across a wide range of species, 16 in total. Sequence comparison of 30,000 sites from all species identify 1,800 variable sites. However, 111 sites are ambiguous in all trees because of apparently multiple substitutions whose ancestry cannot readily be traced. Fifty-three of these ambiguous, or aberrant, sites are segregating within the Bovina (cattle, bison and yak) and Bubalina (Asian and African buffaloes) lineages, which diverged from their common ancestor 5-8 million years ago (Mya). Further investigation has suggested that these are ancient polymorphisms, because they are associated with very small haplotypes. The other possible explanation for aberrant sites is hybridization between species, but this would be characterized by more extensive haplotypes, reflecting exchanges during meiotic recombination. This in turn would suggest that ancestral populations were very large, probably with effective breeding sizes of 90,000 or more [9], because large numbers of aberrant sites would not be expected to survive in a small population (this is consistent with the extremely abundant fossil record). The distribution of these ancient polymorphisms into species-specific lineages would then be a matter of chance. The other aberrant sites probably arose independently in the ancestors of the Bovina, 2-3 Mya, again from large breeding populations. These findings are novel and show that genetic

polymorphisms present 2-8 Mya are still segregating in many present-day lineages.

The large number of aberrant sites in the Bovinae probably explain how the yak came to be reported, erroneously, as a close phylogenetic relative of cattle: many of these sites are shared by the two species. However, when only unambiguous sites are examined, the resulting phylogeny has three main groups: domestic cattle, bison/yak and banteng (Figure 4). The phylogeny is star-like, suggesting rapid evolution in a relatively short time of 1-3 million years [8], a period too short for reliable identification of points of divergence.

Genome biology and domestication

From the analysis of ancestral mutations [10], it appears that domesticated cattle populations are able to maintain a high load of unfavorable mutations. This is probably a consequence of the domestication process itself. The selection of specific cattle breeds has been through many small populations, and thus bottlenecks, which may favor the chance survival of unfavorable alleles. Survival of potentially deleterious alleles will of course be further favored by strong artificial selection: for example, the double-muscling genes favored for beef production would almost certainly be lost in the wild through natural selection.

Like other genome projects, the cattle project also has a parallel SNP discovery pipeline [7]. The reference Hereford genome has been compared with six other breeds, with the identification of 37,470 SNPs polymorphic in all breeds. An immediate practical outcome of this SNP project is the definition of a set of 50 SNPs that could be used for unique parentage assignment and proof of identity.

Recently (in the last 10,000 years), population sizes have fallen sharply to small numbers, with many bottlenecks due to domestication and artificial selection for milk and beef. The decline in diversity seen in some breeds is a matter for concern. But even in these contracted populations, the pattern of linkage disequilibrium suggests that cattle started from a very large base 1-2 Mya with ancestral populations of 90,000 or more [9].

Various measures of genomic selection have been used (iHS, FST and CLR) to map regions of selective sweep on chromosomes 2, 6 and 14 [7]. Selective sweep is the term used for the presence of genes on either side of a selected gene that are unusually conserved by virtue of their linkage to the selected gene. These regions in the bovine genome are, not surprisingly, associated with genes with a function in muscling (*MSTN*), milk yield and composition (*ABCG2*) and energy homeostasis (*R3HDM1*, *LCT*). The evidence of selection in these regions correlates with genes associated with efficiency of food utilization, immunity and behavior. It is possible that under domestication, mutations at these genes have been selected to produce animals more able to resist the infectious diseases prevalent in herds and showing the docile behavior suited to human husbandry [7].

Acknowledgements

DWB is supported by the Biotechnology and Biological Sciences Research Council and the University of Edinburgh.

References

1. The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC: **The genome sequence of taurine cattle: a window to ruminant biology and evolution.** *Science* 2009, **324:**522-528.
2. Liu Y, Qin X, Song X-ZH, Jiang H, Shen Y, Durbin KJ, Lien S, Kent MP, Sodeland M, Ren Y, Zhang L, Sodergren E, Havlak P, Worley KC, Weinstock GM, Gibbs RA: ***Bos taurus* genome assembly.** *BMC Genomics* 2009, **10:**180.
3. Zimin AV, Delcher AL, Florea L, Kelley DA, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL: **A whole-genome assembly of the domestic cow, Bos taurus.** *Genome Biol* 2009, **10:**r42.
4. Brinkmeyer-Langford CL, Childers CP, Fritz KL, Gustafson-Seabury AL: **A high resolution RH map of the bovine major histo-compatibility complex.** *BMC Genomics* 2009, **10:**182.
5. Walker AM, Roberts RM: **Characterization of the bovine Type I IFN locus: rearrangements, expansions, and novel subfamilies.** *BMC Genomics* 2009, **10:**187.
6. Seo S, Lewin HA: R**econstruction of metabolic pathways for the cattle genome.** *BMC Systems Biol* 2009, **3:**33.
7. The Bovine HapMap Consortium: **Genome wide survey of SNP variation uncovers the genetic structure of cattle breeds.** *Science* 2009, **324:**528-532.
8. MacEachern S, John McEwan J, Goddard M: P**hylogenetic reconstruction and the identification of ancient polymorphism in the Bovini tribe (Bovidae, Bovinae).** *BMC Genomics* 2009, **10:**177.
9. MacEachern S, Hayes B, John McEwan J, Goddard M: **An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in domestic cattle.** *BMC Genomics* 2009, **10:**181.
10. MacEachern S, McEwan J, McCulloch A, Mather A, Savin K, Goddard M: **Molecular evolution of the Bovini tribe (Bovidae, Bovinae): is there evidence of rapid evolution or reduced selective constraint in domestic cattle?** *BMC Genomics* 2009, **10:**179.

Bovine genome coverage in BioMed Central:

- Burt DW: **The cattle genome reveals its secrets.** *J Biol* 2009, **8:**36.
- Capuco AV, Akers RM: **The origin and evolution of lactation.** *J Biol* 2009, **8:**37.
- Church DM, Hillier LW: **Back to Bermuda: how is science best served?** *Genome Biol* 2009, **10:**105.