Multi-dimensional functional annotation of bovine genome for the BovReg project.

G.C.M. Moreira[1*], S. Dupont[1], D. Becker[2], M. Salavati[3], R. Clark[4], E.L. Clark[3], G. Plastow[5], C. Kühn[2,6] and C. Charlier[1] on behalf of the BovReg consortium

[1] Unit of Animal Genomics, GIGA Institute, University of Liège, 4000, Liège, Belgium; [2] Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), 18196, Dummerstorf, Germany; [3] The Roslin Institute, University of Edinburgh, EH25 9RG, Edinburgh, UK; [4] Genetics Core, Edinburgh Clinical Research Facility, The University of Edinburgh, EH4 2XU, Edinburgh, UK; [5] Livestock Gentec, Department of Agricultural, Food and Nutritional Science, University of Alberta, T6G 2R3, Edmonton, Canada; [6] Faculty of Agricultural and Environmental Sciences, University Rostock, 18059, Rostock, Germany; *gcosta@uliege.be

Abstract

A multi-dimensional transcriptome map was generated using a diverse catalogue of tissue samples collected from six individuals of both sexes, different ages, kept in different environments and from three divergent breeds/crosses. More than 15k genes exhibit at least one potentially novel transcript in comparison with existing annotations. Tissue-specific miRNAs were detected contributing to the understanding of vital body functions in bovine. PIWI-interacting RNAs (piRNAs) were detected in ovary and testis only and, the repertoire of piRNA clusters in the male germline was expanded. A regulatory landscape was also generated including open chromatin regions, putative active promoters and transcription start sites, active enhancers, repressive states and insulators. This constitutes a key repository dataset for biology-driven genomic prediction that will help the scientific community and industry stakeholders to apply genomics to address the challenges facing cattle production.

Introduction

In humans, a catalogue of functional elements in the genome was generated by the ENCODE consortium (ENCODE Project Consortium, 2012; Davis *et al.,* 2018) to link genotype to phenotype. In bovine, a comprehensive catalogue of functional elements still does not exist despite the huge contribution of bovids to global food production (Adesogan *et al.,* 2020). Recent efforts from the FAANG consortium have contributed to the functional annotation of the bovine genome but, with a limited number of tissues, individuals and populations analyzed (Foissac et al., 2019; Kern et al*.,* 2021). Multi-dimensional functional annotation of the bovine genome built from a diverse catalogue of tissue samples collected from individuals of both sexes, different ages, kept in different environments and from divergent breeds/crosses is critical in linking genotype to phenotype in cattle, to further improve biology-driven genomic prediction models.

Materials & Methods

Samples and tissues. A diverse catalogue of 129 tissue samples were collected from six individuals of both sexes, different ages, kept in different environments and from three divergent breeds/crosses: Belgian dairy (Holstein), Canadian beef composite (Kinsella) and German beef/dairy cross (Charolais x Holstein).

RNA isolation and library preparation. For transcriptome map generation, Total RNA was extracted using miRNeasy kit (QIAGEN) from the 129 snap-frozen tissues samples. After quality control (quantity, purity and integrity assessment), Total RNA libraries were generated

using TruSeq Stranded Total RNA Library Prep Gold (Illumina); mRNA libraries using TruSeq Stranded mRNA Library Prep (Illumina); and small RNA libraries using QIAseq miRNA Library (QIAGEN).
For regulatory landscape generation, ATACseq libraries were generated using the ATACseq kit (Active Motif). The chromatin immunoprecipitations (ChIP) were performed for five antibodies (H3K4me3, H3K4me1, H3K27me3, H3K27ac and CTCF - all from Diagenode) using the iDeal ChIP-seq kit for TF (Diagenode); and ChIPseq libraries were prepared using the NEBNext Ultra DNA library prep kit for Illumina libraries (New England Biolabs).

Sequencing and analyses. Libraries were sequenced on the Illumina NovaSeq 6000 (150nt 25M paired-end – mRNA; 150nt 40M paired-end – Total RNA; 100nt/50nt 25M single-end - small-RNA; 150nt 50M paired-end – ATACseq; 100nt paired-end – ChIPseq 30M reads for narrow marks and CTCF; 35M reads for broad marks).
Data analyses for mRNA and Total RNA assays were performed using the BovReg pipeline (https://github.com/BovReg/rnaseq) and for small RNA, ATACseq and ChIPseq assays, using Nextflow/nf-core pipelines (smrnaseq-1.0.0, atacseq-1.2.1, chipseq-1.2.2). The ARS-UCD1.2_Btau5Y reference file from 1000bulls project was utilized. The GTF file from Ensembl v.102 with the coordinates lifted over to the 1000bulls reference genome was used.
For piRNA analyses, small-RNA were mapped to the genome and after the selection of reads ranging between 18-35nt, reads that mapped to rRNAs, miRNAs, snoRNAs, snRNAs, and tRNAs were removed. PiRNA clusters were detected using proTRAC software (Rosenkranz and Zischler, 2012) with default parameters.

Results

Transcriptome map. A *de novo* transcriptome assembly from mRNA and Total RNA assays was created. 43,117 genes were assembled and from those, more than 15K genes exhibit at least one potentially novel transcript in comparison with existing annotations (Table 1).

Table 1. Results from the comparison between BovReg transcriptome assembly and existing annotations.

BovReg annotation features	Compared with Ensembl v.104*	Compared with NCBI v.106
Genes with at least one potentially novel transcript assembled	15,954	18,636
Complete transcripts with exact match of intron chain	37,257	34,430
Transcripts with partial match, encompassing or within reference transcripts	211,616	223,926
Novel transcripts – no match with reference (unknown, intergenic)	28,925	19,442

*To compare the *de novo* transcriptome assembly from BovReg with existing annotations, Ensembl v.104 annotation was utilized instead of v.102.

Based on mRNA assays, tissues originating from the central nervous system exhibit a greater number of genes with elevated expression compared to the other tissues (Figure 1A). Charolais x Holstein/adult testis exhibited four-fold more genes with elevated expression than Holstein/neonate testis and most of those are tissue enriched (elevated expression in testis compared to any other tissue), with GO terms enriched for spermatogenesis related processes known to occur in post-puberal testis (Figure 1A). When considering only tissues shared between the six animals, Holstein/neonate animals exhibited a greater number of genes expressed/assembled and also a greater number of genes with at least one potentially novel transcript compared to Kinsella crossbred/juvenile and Charolais x Holstein.

In the small RNA dataset, miRNAs were the predominant class detected (with 81.1% of the reads corresponding to miRNAs, 0.4% to rRNA, 0.3% to tRNA, 3.02% to snoRNA, 0.93% to snRNA and 14.25% to either mitosRNA, unknown - including piRNAs - or artifacts). Taking the top five known miRNAs with the highest expression in each tissue, tissue-specific miRNAs were confirmed (Figure 1B) and most of these miRNAs are already known to play important roles in key biological processes. As an example, *Bta-mir-206* is highly expressed in skeletal muscle and is involved in muscular hypertrophy in Texel sheep (Clop et al., 2006).
PIWI-interacting RNAs (piRNAs) were only detected in ovary and testis. PiRNAs are known to silence transposons and regulate gene expression (Czech and Hannon, 2016). In testis they represent up to 95% of small-RNAs in adult and around 30% in neonate. A huge proportion of piRNAs, especially those expressed in late spermatogenesis, are produced from large genomic regions known as piRNA clusters (Aravin, Hannon & Brennecke, 2007). A greater number of piRNA clusters were detected on neonate testis compared to adult testis and also, from what is known from the database (https://www.smallrnagroup.uni-mainz.de/piRNAclusterDB/), expanding the repertoire of piRNA clusters in bovine.
The data presented here will be combined with a new map of transcription start sites (TSS) (CAGE), lncRNA and circRNA, which will be generated in collaboration with other partners in the BovReg consortium, to create a multi-dimensional transcriptome map.

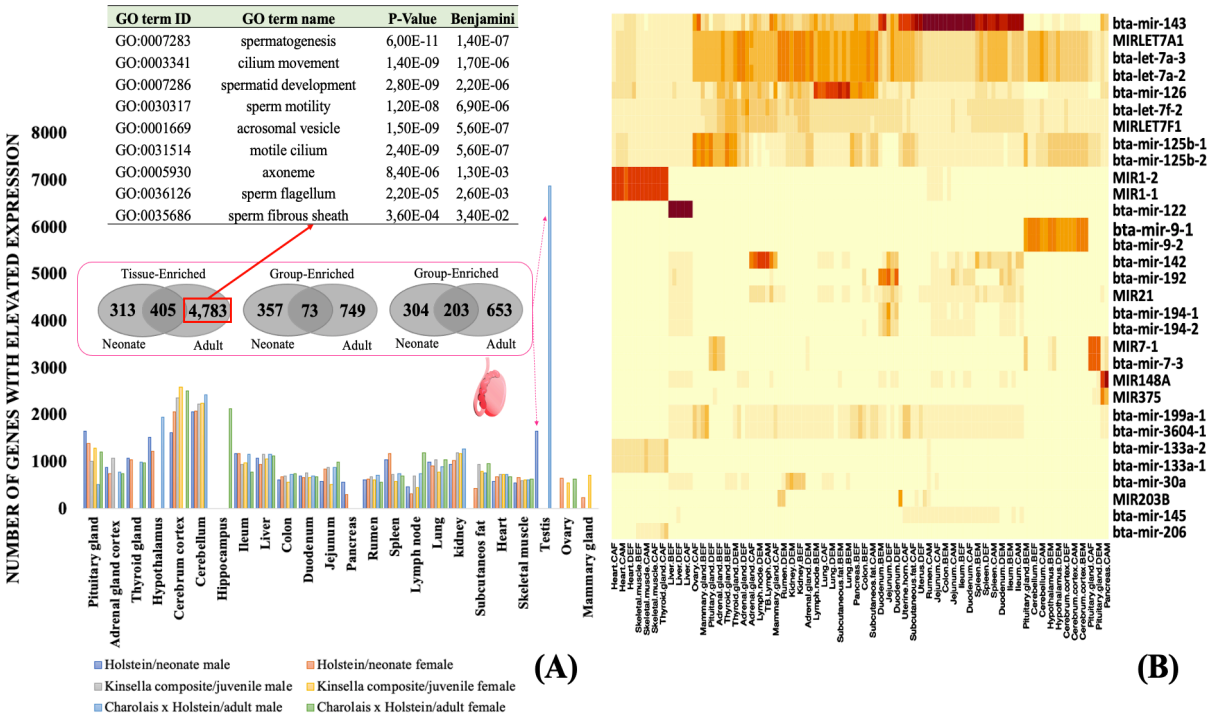

Figure 1 – Overview of mRNA and miRNA specificity profiles in the bovine transcriptome
(A) Genes with elevated expression (normalized by TPM) in each tissue; classification and GO enrichment of genes with elevated expression in testis. Elevated expression encompass three subcategory types (Jain and Tuteja, 2019): tissue enriched (at least 4x higher mRNA level compared to any other tissue); tissue enhanced (at least 4x higher average mRNA level in a group of 2-5 tissues compared to any other tissue) and; group enriched (at least 4x higher mRNA level compared to the average level in all other tissues). (B) Heatmap of the top 5 known miRNAs with the highest expression (normalized by CPM) in each tissue. MiRNAs in y-axis and tissues on x-axis; BEM: Belgian animal male; BEF: Belgian animal female; CAM: Canadian animal male; CAF: Canadian animal female; DEM: German animal male; DEF: German animal female.

Regulatory regions. Preliminary ATACseq data revealed approximately 666K open chromatin regions in the bovine genome. From peak annotation relative to gene features, using the *de novo* transcriptome assembly ~4-fold more peaks were annotated in exons, ~3-fold more peaks annotated in promoter-TSS and in transcription termination sites (TTS) and ~2-fold less annotated in intergenic regions, compared with the annotation obtained using Ensembl v.102 assembly.
From preliminary ChIPseq data, 230,923 putative active promoters and transcription start sites (TSS) states (H3K4me3), 627,454 active enhancer states (H3K27ac and H3K4me1), 659,150 repressive states (H3K27me3) and 409,062 insulators bound by CTCF were detected.

Discussion
Despite a greater number of genes harboring potentially novel transcripts that are expressed in purebred/neonate compared to composite breeds/juvenile and adult animals, it is difficult to claim if the greater number is driven by age and/or breed. However, our findings demonstrate the importance of having different ages, breeds/crosses and environments to build a comprehensive functional annotation of bovine genome. Our investigation of gene expression regulation by miRNAs, especially the novel ones, provides important biological knowledge to better understand bovine physiology.
The repertoire of miRNAs, piRNAs and piRNA clusters in the bovine germline was expanded. There is a lack of information about piRNAs in the bovine germline and their potential impact on gene expression regulation, which emphasises the importance of the new annotation information provided by BovReg.
The data presented here already represents a substantial improvement on the functional annotation of the bovine genome. An integrative approach with the *de novo* transcriptome map, chromatin states as well as TSS catalogue will provide additional evidence to support the novel genes, miRNAs and piRNA clusters detected.
The BovReg project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668.

References
Adesogan A.T., Havelaar A.H., McKune S.L., Eilittä M. and Dahl G.E. (2020) Glob Food Secur 25:100325. https://doi.org/10.1016/j.gfs.2019.100325
Aravin A.A., Hannon G.J. and Brennecke J. (2007) Science 318(5851):761-764. https://doi.org/10.1126/science.1146484
Clop A., Marcq F., Takeda H., Pirottin, D., Tordoir, X. *et al*. (2006). Nat Genet 38(7):813-818. https://doi.org/10.1038/ng1810
Czech B. and Hannon G.J. (2016) Trends Biochem Sci 41(4):324-337. https://doi.org/10.1016/j.tibs.2015.12.008
Davis C.A., Hitz B.C., Sloan C.A., Chan E.T., Davidson J.M. *et al.* (2018) Nucleic Acids Res 46(D1):D794-D801. https://doi.org/10.1093/nar/gkx1081
ENCODE Project Consortium (2012) Nature 489(7414):57-74. https://doi.org/10.1038/nature11247
Foissac S., Djebali S., Munyard K., Vialaneix N., Rau A. *et al.* (2019) BMC Biol 17(1):108. https://doi.org/10.1186/s12915-019-0726-5
Jain A. and Tuteja G. (2019) Bioinformatics 35(11):1966-1967. https://doi.org/10.1093/bioinformatics/bty890
Kern C., Wang Y., Xu X., Pan Z., Halstead M. *et al.* (2021) Nat Commun 1821. https://doi.org/10.1038/s41467-021-22100-8
Rosenkranz D. and Zischler H. (2012) BMC Bioinform 13 (5). https://doi.org/10.1186/1471-2105-13-5