GigaScience

Facilitating Functional genomics of cattle through integration of multi-omics data

--Manuscript Draft--

Manuscript Number:	GIGA-D-23-00037
Full Title:	Facilitating Functional genomics of cattle through integration of multi-omics data
Article Type:	Research
Funding Information:	National Institute of Food and Agriculture (2018-67015-27500) — Dr Huaijun Zhou National Institute of Food and Agriculture (2015-67015-22940) — Dr Huaijun Zhou
Abstract:	Background The accurate identification of the functional elements in the bovine genome is a fundamental requirement for high quality analysis of data informing both genome biology and genomic selection. Functional annotation of the bovine genome was performed to identify a more complete catalogue of transcript isoforms across bovine tissues. Results A total number of 171,985 unique transcripts (50% protein-coding) representing 35,150 unique genes (64% protein-coding) were identified across tissues. Among them, 118,563 transcripts (70% of the total) were structurally validated by independent datasets (PacBio Iso-seq data, ONT-seq data, de novo assembled transcripts from RNA-seq data) and comparison with Ensembl and NCBI gene sets. In addition, all transcripts were supported by extensive data from different technologies such as WTTS-seq, RAMPAGE, ChIP-seq, and ATAC-seq. A large proportion of identified transcripts (69%) were un-annotated, of which 87% were produced by annotated genes and 13% by un-annotated genes. A median of two 5’ untranslated regions were expressed per gene. Around 50% of protein-coding genes in each tissue were bifunctional and transcribed both coding and noncoding isoforms. Furthermore, we identified 3,744 genes that functioned as non-coding genes in fetal tissues, but as protein coding genes in adult tissues. Our new bovine genome annotation extended more than 11,000 annotated gene borders compared to Ensembl or NCBI annotations. The resulting bovine transcriptome was integrated with publicly available QTL data to study tissue-tissue interconnection involved in different traits and construct the first bovine trait similarity network. Conclusions These validated results show significant improvement over current bovine genome annotations.
Corresponding Author:	James Reecy Iowa State University Ames, IA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Iowa State University
Corresponding Author's Secondary Institution:	
First Author:	Hamid Beiki
First Author Secondary Information:	
Order of Authors:	Hamid Beiki
	Brenda M. Murdoch

	Carissa A. Park
	Chandlar Kern
	Denise Kontechy
	Gabrielle Becker
	Gonzalo Rincon
	Honglin Jiang
	Huaijun Zhou
	Jacob Thorne
	James E. Koltes
	Jennifer J. Michal
	Kimberly Davenport
	Monique Rijnkels
	Pablo J. Ross
	Rui Hu
	Sarah Corum
	Stephanie McKay
	Timothy P.L. Smith
	Wansheng Liu
	Wenzhi Ma
	Xiaohui Zhang
	Xiaoqing Xu
	Xuelei Han
	Zhihua Jiang
	Zhi-Liang Hu
	James Reecy
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.	Yes

Have you included all the information requested in your manuscript?	
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Facilitating Functional genomics of cattle through integration of multi-omics data

Hamid Beiki[1], Brenda M. Murdoch[2], Carissa A. Park[1], Chandlar Kern[3], Denise Kontechy[2], Gabrielle Becker[2], Gonzalo Rincon[4], Honglin Jiang[5], Huaijun Zhou[6], Jacob Thorne[2], James E. Koltes[1], Jennifer J. Michal[7], Kimberly Davenport[2], Monique Rijnkels[8], Pablo J. Ross[6], Rui Hu[5], Sarah Corum[4], Stephanie McKay[9], Timothy P.L. Smith[10], Wansheng Liu[3], Wenzhi Ma[3], Xiaohui Zhang[7], Xiaoqing Xu[6], Xuelei Han[7], Zhihua Jiang[7], Zhi-Liang Hu[1], James M. Reecy[1]

[1]Department of Animal Science, Iowa State University; [2]Department of Animal and Veterinary and Food Science, University of Idaho; [3]Department of Animal Science, Pennsylvania State University; [4]Zoetis; [5]Department of Animal and Poultry Sciences, Virginia Tech; [6]Department of Animal Science, University of California, Davis; [7]Department of Animal Science, Washington State University; [8]Department of Veterinary Integrative Biosciences, Texas A&M University; [9]University of Vermont; [10]USDA, ARS, USMARC.

Corresponding author:

James M. Reecy

Professor of Animal Breeding and Genetics, Department of Animal Science, Ames, IA, USA

jreecy@iastate.edu

Abstract

Background

The accurate identification of the functional elements in the bovine genome is a fundamental requirement for high quality analysis of data informing both genome biology and genomic selection. Functional annotation of the bovine genome was performed to identify a more complete catalogue of transcript isoforms across bovine tissues.

Results

A total number of 171,985 unique transcripts (50% protein-coding) representing 35,150 unique genes (64% protein-coding) were identified across tissues. Among them, 118,563 transcripts (70% of the total) were structurally validated by independent datasets (PacBio Iso-seq data, ONT-seq data, *de novo* assembled transcripts from RNA-seq data) and comparison with Ensembl and NCBI gene sets. In addition, all transcripts were supported by extensive data from different technologies such as WTTS-seq, RAMPAGE, ChIP-seq, and ATAC-seq. A large proportion of identified transcripts (69%) were un-annotated, of which 87% were produced by annotated genes and 13% by un-annotated genes. A median of two 5' untranslated regions were expressed per gene. Around 50% of protein-coding genes in each tissue were bifunctional and transcribed both coding and noncoding isoforms. Furthermore, we identified 3,744 genes that functioned as non-coding genes in fetal tissues, but as protein coding genes in adult tissues. Our new bovine genome annotation extended more than 11,000 annotated gene borders compared to Ensembl or NCBI annotations. The resulting bovine transcriptome was integrated with publicly available QTL data to study tissue-tissue interconnection involved in different traits and construct the first bovine trait similarity network.

Conclusions

These validated results show significant improvement over current bovine genome annotations.

Introduction

Domestic bovine (*Bos taurus*) provides a valuable source of nutrition and an important disease model for humans [1]. Furthermore, cattle have the greatest number of genotype associations and genetic correlations of the domesticated livestock species, which means they provide an excellent model to close the genotype-to-phenotype gap. Therefore, the accurate identification of the functional elements in the bovine genome is a fundamental requirement for high quality analysis of data informing both genome biology and genomic selection.

Current annotations of farm animal genomes largely focus on the protein-coding regions and fall short of explaining the biology of many important traits that are controlled at the transcriptional level [2]. In humans, 88% of trait-associated single nucleotide polymorphisms (SNP) identified by genome-wide association studies (GWAS) are found in non-coding regions

[3]. Therefore, elucidating non-coding functional elements of the genome is essential for understanding the mechanisms that control complex biological processes.

Untranslated regions play critical roles in the regulation of mRNA stability, translation, and localization [4], but these regions have been poorly annotated in farm animals [2, 5]. A recent study of the pig transcriptome using single-molecule long-read isoform sequencing technology resulted in the extension of more than 6000 annotated gene borders compared to Ensembl or National Center for Biotechnology Information (NCBI) annotations [2].

Small non-coding RNAs, such as microRNAs (miRNA), are known to be involved in gene regulation through post-transcriptional regulation of expression via silencing, degradation, or sequestering to inhibit translation [6-8]. The number of annotated miRNAs in the current bovine genome annotation (Ensembl release 2018-11; 951 miRNAs) is much lower than the number reported in the highly annotated human genome (Ensembl release 2021-03; 1,877 miRNAs).

This study applied a comprehensive set of transcriptome and chromatin state data from 47 cattle tissues and cell types to identify previously unannotated genes and improve the annotation of thousands of protein-coding and non-coding genes. Predicted un-annotated genes and transcripts were highly supported by independent Pacific Biosciences single-molecule long-read isoform sequencing (PacBio Iso-Seq), Oxford Nanopore Technologies sequencing (ONT-seq), Illumina high-throughput RNA sequencing (RNA-seq), Whole Transcriptome Termini Site Sequencing (WTTS-seq), RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression (RAMPAGE), chromatin immunoprecipitation sequencing (ChIP-seq), and Assay for Transposase-Accessible Chromatin using sequencing

(ATAC-seq) data. The transcriptome data was integrated with publicly available Quantitative Trait Loci (QTL) and gene association data to construct the first bovine trait similarity network that recapitulates published genetic correlations. Thus, it may be possible to begin to examine the genetic mechanisms underlying genetic correlations.

Results

The diversity of RNA and miRNA transcript diversity among 47 different bovine tissues and cell types was assessed using miRNA-seq and polyadenylation (poly(A)) selected RNA-seq data. Most of the tissues studied were from Hereford cattle closely related to L1 Dominette 01449, the individual from which the bovine reference genome (ARS-UCD1.2) was sequenced. The 47 tissues and cell samples included follicular cells, myoblasts, five mammary gland samples from various stages of mammary gland development and lactation, eight fetal tissues (78-days of gestation), eight tissues from adult digestive tract, and 16 other adult organs. A total of approximately 4.1 trillion RNA-seq reads and 1.2 billion miRNA-seq reads were collected, with a minimum of 27.5 million RNA-seq and 9.3 million miRNA-seq reads from each tissue/cell type (average 87.8 ± 49.7 million and 27.6 ± 12.9 million, respectively) (Supplemental file 1: Fig. S1 and Supplemental file 2).

Transcript level analyses

A total of 171,985 unique transcripts (76% spliced) were identified (Table 1) with a median of 51,231 transcripts per tissue. There was a median of 9.1 exons per spliced transcript, and all of the predicted acceptor and donor splice sites conformed to the canonical consensus sequences. All of the predicted splice junctions across tissues were supported by RNA-seq reads that spanned the splice junction, substantiating the accuracy of the transcript definition from RNA-seq reads.

A total of 31,476 transcripts appeared tissue-specific by virtue of being assembled from RNA-seq reads in just a single tissue, but 20,100 of those transcripts (64%) were actually expressed in multiple tissues. Thus, reliance solely on assembled transcripts in a given tissue to predict a tissue transcript atlas may overestimate tissue specificity due to a high false-negative rate for transcript detection. To solve this problem of over-prediction of tissue specificity, we marked a transcript as "expressed" in a given tissue only if (1) it had been assembled from RNA-seq data in that tissue; or (2) its expression and all of its splice junctions has been quantified using RNA-seq reads in the tissue of interest with an expression level more than 1 reads per kilobase of transcript per Million reads mapped (RPKM) (see Methods section). This resulted in 11,375 apparently tissue-specific transcripts (7%) and 156,423 transcripts (91%) expressed in more than one tissue (Fig. 1), among which 9,125 transcripts (5%) were found in all 47 tissues examined.

The unique transcripts identified were equally distributed between 85,658 (50%) protein-coding transcripts and 86,327 (50%) non-coding transcripts (ncRNAs) (Fig. 2). Non-coding transcripts were further classified as long non-coding RNAs (lncRNAs) (56%), nonsense-mediated decay (NMD) transcripts (38%), non-stop decay (NSD) transcripts (5%), and small non-coding RNAs (sncRNAs) (1%). While the majority of expressed transcripts in each tissue were protein coding (median of 62% of tissue transcripts), NMD transcripts (median of 14.58% of tissue transcripts) and antisense lncRNAs (median of 12% of tissue transcripts) each made up more than 10% of the transcripts (Supplemental file 1: Fig. S2A and B, Supplemental file 3 and

4). Fetal muscle and fetal gonad tissues showed the highest proportion of antisense lncRNAs compared to that observed in other tissues (Supplemental file 1: Fig. S2B) and around 60% of antisense lncRNAs (17,982 transcripts) were expressed from these two tissues. Compared to non-coding transcripts, protein-coding transcripts were more likely to have spliced exons (p-value < 2.2e-16) and were expressed in a higher number of tissues (median of 11 tissues for protein-coding transcripts versus six tissues for non-coding transcripts; p-value < 2.2e-16) (Additional file1: Fig. S2C). The lncRNAs had a significantly lower splice rate (36%) compared to other non-coding transcripts (p-value < 2.2e-16). Splice rate was highest (70%) in sncRNAs (p-value < 2.2e-16; NMD transcripts were not included in this analysis, as they were all spliced transcripts by definition).

There were no significant correlations between the number of RNA-seq reads for a given tissue and the number of unique transcripts identified, except for a modest correlation for the antisense lncRNA class (Supplemental file 1: Fig. S3A). There was a significant positive correlation (p-value 1.3e-04) between the number of unique NMD transcripts in a tissue and the number of protein-coding transcripts, and the NMD transcript class showed the lowest median expression level across tissues, followed by antisense-lncRNAs and sense intronic-lncRNAs (Supplemental file 1: Fig. S2D and Fig. S3B). In addition, there was a significant positive correlation (p-value 3.4e-03) between the number of NMD transcripts and the number of protein-coding transcripts across tissues (Supplemental file 1: Fig. S3A). The expression levels of sncRNAs and protein-coding transcripts were higher (p-values: 1.1e-02 and 2.6e-06, respectively) than that observed for other transcript biotypes (Supplemental file 1: Fig. S2D and Fig. S3B).

Transcript similarity to other species

Protein/peptide homology analysis of transcripts with an open reading frame (protein-coding transcripts, lncRNAs, and sncRNAs) revealed a higher conservation of protein-coding transcripts (86%) compared to lncRNA and sncRNA transcripts (8%; p-value < 2.2e-16) (Table 2). Bovine non-coding transcripts had significantly (p-value < 2.2e-16) less similarity to other species than protein-coding transcripts (Table 2 and Table 3). Within non-coding transcripts, NSD transcripts showed the lowest conservation rate (35%), followed by sncRNAs (37%), lncRNAs (49%), and NMD transcripts (55%), while sense intronic lncRNAs had the highest conservation rate (60%) compared to other non-coding transcripts (Table 4).

Transcript expression diversity across tissues

A median of 70% of protein-coding transcripts were shared between pairs of tissues (Supplemental file 1: Fig. S4A), significantly higher than that was observed for non-coding transcripts (53%; p-value < 2.2e-16; Supplemental file 1: Fig. S5). Clustering of tissues based on protein-coding transcripts was different than that observed based on non-coding transcripts (Supplemental file 1: Fig. S4B and Fig. S5B, Fig. S35F). The fetal tissues clustered together and were generally more similar to one another than to the corresponding adult tissue in both dendrograms, but thymus was closely related to fetal tissues for protein-coding transcript content, while it appeared more similar to lymph nodes, myoblasts, and pregnant/lactating mammary tissue using non-coding transcript profiles. The digestive tract tissues clustered together in the non-coding dendrogram with ileum as a slight outlier, while both jejunum and ileum were distant from the other digestive tissues in the protein-coding transcript profile. The "adult mammary gland" (78 day pregnant) and "virgin mammary gland" samples did not cluster with the three other pregnant/lactating mammary samples nor with each other in either dendrogram. This is mostly likely because: 1) these are from different physiological stages, 2) these were whole tissue samples while the other three pregnant/lactating samples are enriched for mammary gland epithelial cells, 3) the virgin and 78-day pregnant samples are from Hereford background while other pregnant/lactating samples are from Holstein-Frisian breed. Fetal tissues had significantly higher proportions than adult tissues of unique non-coding transcripts (specifically NSDs, antisense lncRNAs, and intragenic lncRNAs) compared to protein-coding transcripts (p-value < 2.2e-16; Supplemental file 5).

Transcript validation

Prediction of transcripts and isoforms from RNA-seq data may produce erroneous predicted isoforms. The validity of transcripts was therefore examined by comparison to a library of isoforms taken from Ensembl (release 2021-03) and NCBI gene sets (Release 106), as well as isoforms identified through complete isoform sequencing with Pacific Biosciences, a de novo assembly produced from its matched RNA-seq reads, and isoforms identified from Oxford Nanopore platforms (see Methods section). A total of 118,563 transcripts (70% of predicted transcripts) were structurally validated by independent datasets (PacBio Iso-seq data, ONT-seq data, *de novo* assembled transcripts from RNA-seq data) and comparison with Ensembl and NCBI gene sets. A total of 160,610 transcripts were expressed in multiple tissues (93% of predicted transcripts), providing further support for their validity (Fig. 3). All transcripts were also extensively supported by data from different technologies such as WTTS-seq, RAMPAGE, histone modification (H3K4me3, H3K4me1, H3K27ac), CTCF-DNA binding, and ATAC-seq (Fig. 3).

Comparison of predicted transcript structures with annotated transcripts in the current bovine genome annotations (Ensembl release 2021-03 and NCBI Release 106) resulted in a total of 52,645 annotated transcripts that exactly matched previously annotated transcripts (31% of all transcripts), including 47,054 annotated NCBI transcripts, 31,740 annotated Ensembl transcripts, and 26,149 transcripts that were common to both annotated gene sets (Fig. 3). The median expression level of annotated transcripts in their expressed tissues (1.8 RPKM) was similar to that observed for un-annotated transcripts (1.4 RPKM) (Supplemental file 1: Fig. S6). Annotated transcripts were expressed in a median of 17 tissues, which was higher (p-value 7.4e-03) than that observed for un-annotated transcripts (median of seven tissues) (Supplemental file 1: Fig. S6). In addition, compared to un-annotated transcripts, annotated transcripts were enriched with protein-coding (p-value 1.37e-02) and spliced transcripts (p-value 3.76e-02).

The median length of coding sequence (CDS) of annotated transcripts was 1,014 nt, significantly longer than that observed in un-annotated transcripts (510 nt; p-value 0.0) (Additional file1: Fig. S7A). In addition, un-annotated transcripts had longer 5’ untranslated regions (UTR) (400 bp) compared to that was observed in annotated transcripts (300 nt, p-value 2.631E-06; Additional file1: Fig. S7A). Un-annotated transcripts encoding proteins with homology to proteins annotated in other species had longer CDS (687 bp) compared to transcripts without such homology (192 nt; p-value 0.0). Annotated protein-coding transcripts showed a higher GC content in their 5’ UTRs (61%) than un-annotated transcripts (53%; p-value 5.562E-18), but both classes of transcripts showed similar GC content within their CDS (Supplemental file 1: Fig. S7B).

Gene level analyses

The transcripts correspond to a total of 35,150 genes, which were classified into protein coding (21,193), non-coding (10,928), and pseudogenes (3,029) (Supplemental file 3 and 4, and Fig. 4). Genes transcribed at least a single "expressed" transcript (see Transcript level analysis section) in a given tissue, were marked as "expressed gene" in that tissue. Most genes expressed in each tissue were protein coding (median of 83% of tissue genes), followed by non-coding (median of 14% of tissue genes) and pseudogenes (median of 3% of tissue genes) (Supplemental file 1: Fig. S8). Testis showed the highest number of expressed genes with observed transcripts compared to other tissues (Supplemental file 1: Fig. S8). Fetal brain and fetal muscle tissues showed the highest number and percentage of non-coding genes compared to that observed in other tissues (Supplemental file 1: Fig. S8). In addition, more than 40% of transcripts corresponded to non-coding genes (1,271 genes) in fetal brain and fetal muscle. The proportion (6%) and number (1,271) of transcript-producing pseudogenes was higher in testis than in other tissues. There was no significant correlation between the number of input reads and the number of expressed genes across tissues, but the numbers of genes from different coding potential classes were significantly correlated across tissues (Supplemental file 1: Fig. S9).

Transcripts corresponding to the predicted genes that had at least one exon overlapping an Ensembl- or NCBI-annotated gene were considered to belong to an annotated gene. This supported an intersection analysis of predicted and previously annotated genes that indicated 22,452 (64%) of our predicted genes correspond to previously annotated genes. Approximately 87% of un-annotated transcripts (103,387) were associated with this set of annotated genes. The remaining 12,698 genes (36% of predicted genes) represent un-annotated genes, i.e., genes not found on Ensembl (release 2021-03) or NCBI (release 106), with which 15% of un-annotated transcripts (22,364 transcripts) were associated. The median number of unique transcripts per annotated gene (tpg) was four, which was higher than that observed in either the Ensembl (1.5 tpg) or NCBI (2.3 tpg) annotated gene sets, while the median number of transcripts per un-annotated gene was one, with an average of 1.31 and standard deviation of 1.36. Most of the transcripts identified were transcribed from annotated genes, including 96% of protein-coding transcripts (82,060), 79% of lncRNA transcripts (38,662), 78% of sncRNA transcripts (413), and more than 95% of NMD transcripts (31,422). Annotated genes were enriched with protein-coding genes (p-value < 2.2e-16). The median transcript abundance from annotated genes in their expressed tissues (6.59 RPKM) was significantly higher than that observed for un-annotated genes (median of 1.68 RPKM; p-value < 2.2e-16; Supplemental file 1: Fig. S10A). The median number of tissues in which annotated genes were expressed (42 tissues) was also significantly higher than that observed for un-annotated genes (median of four tissues; p-value < 2.2e-16; Supplemental file 1: Fig. S10B).

More than a third (37%) of genes with at least one predicted protein-coding transcript displayed either multiple 5’ UTRs or multiple 3’ UTRs (median of three 5’ UTRs and three 3’ UTRs per gene) among associated transcript isoforms (Fig. 5). The 496 genes with the highest number of UTRs (the top 5% in this metric) were highly enriched (q-value 1.7E-7) for the “response to protozoan” Biological Process (BP) Gene Ontology (GO) term (Supplemental file 1: Fig. S11 and Supplemental file 6).

A median of 51% of the expressed protein-coding genes in each tissue transcribed both protein-coding and non-coding transcripts and were denoted as bifunctional genes. These genes were mostly previously annotated (95%) and had both coding and non-coding transcripts in a median of 21 tissues, representing 57% of their expressed tissues (Fig. 6A and B). Protein-coding transcripts and NMD transcripts covered more than 90% of the exonic length in bifunctional genes (Fig. 6C). This percentage was significantly lower for other types of non-coding transcripts transcribed from bifunctional genes (77%, 81%, and 62% for NSD transcripts, sncRNAs, and intragenic lncRNAs, respectively) (Fig. 6C). Although transcript terminal sites (TTS) of transcripts encoded by bifunctional genes were centralized around these genes' 3' ends, transcript start sites (TSS) varied greatly among transcript biotypes (Fig. 6C). The TTSs of NSD transcripts, sncRNAs, and intragenic lncRNAs were shifted from their protein-coding genes' start sites (Fig. 6C). Genes that transcribed both protein-coding and non-coding transcripts in all of their expressed tissues (1,661 genes) were highly enriched for "mRNA processing" (q-value 6.08E-16) and "RNA splicing" (q-value 1.35E-14) BP GO terms that were mostly (65%) related to different aspects of transcription and translation (Fig. 6D and Supplemental file 7).

A total of 3,744 protein-coding genes (17% of all predicted protein-coding genes) only transcribed non-coding transcripts in a median of two tissues (equivalent to 15% of their expressed tissues). Detailed investigation of these genes in tissues from both adult and fetal samples (brain, kidney, muscle, and spleen) revealed the total of 106 non-coding genes (90% annotated) in fetal tissues that were switched to protein-coding genes with only protein-coding transcripts in their matched adult tissues (Supplemental file 1: Fig. S12). Functional enrichment analysis of these genes resulted in the identification of enriched BP GO terms related to "humoral immune response", "sphingolipid biosynthetic process", "negative regulation of wound healing”, “cellular senescence”, “symporter activity”, “regulation of lipid biosynthetic process”, and “filopodium assembly” (Supplemental file 1: Fig. S12, Supplemental file 8).

A median of 32% of protein-coding genes in each tissue expressed at least a single potentially aberrant transcript (PAT), i.e., NMDs and NSDs. In this group of genes, the number of PATs was strongly correlated with the total number of transcripts (median correlation of 0.61 across all tissues). The median expression level of these genes in their expressed tissues (11.52 RPKM) was significantly higher (p-value < 2.2e-16) than for protein-coding genes with no PATs (4.48 RPKM). In each tissue, protein-coding genes with PATs showed a significantly higher number of introns (p-value < 2.2e-16; median of 65 introns per gene) than that observed in the remainder of protein-coding genes (median of 15 introns per gene). In addition, genes from this group were expressed in a median of 47 tissues, significantly higher (p-value < 2.2e-16) than that observed for the other coding genes (median of 24 tissues), non-coding genes (median of five tissues), and pseudogenes (median of four tissues) (Supplemental file 1: Fig. S13A and B). These genes transcribed a median of two PATs in half (median 54%) of their expressed tissues, equivalent to a median of 22% of all their transcripts in each tissue. Protein-coding genes that transcribed PATs as their main transcripts (PATs comprised >50% of their transcripts) in all of their expressed tissues were highly enriched with RNA splicing–related BP GO terms (Supplemental file 9).

Gene similarity to other species

Eighty-five percent of protein-coding genes (18,087) encoded either homologous proteins (17,150 genes or 80% of protein-coding genes) or homologous ncRNAs (7,347 genes or 35% of protein-coding genes) (Supplemental file 1: Fig. S14A). Nineteen percent of protein-coding genes (4,043) encoded cattle-specific proteins (Supplemental file 1: Fig. S14A). Most of these genes (2,750 or 68%) were either annotated genes or genes with homology to another cattle gene(s) that has established homology to genes in other species (Supplemental file 1: Fig. S14C). The remaining 32% of cattle-specific, protein-coding genes (1,293 genes or six percent of protein-coding genes) were denoted as protein-coding orphan genes (Supplemental file 1: Fig. S14C). A median of 70 protein-coding orphan genes were expressed in each tissue. The expression level of these genes was significantly lower than other types of protein-coding genes (Additional file1: Fig. S15A and B). The median number of expressed tissues for protein-coding orphan genes (one tissue) was lower than for other types of protein-coding genes (46 tissues) (Supplemental file 1: Fig. S15C). In addition, protein-coding orphan genes only transcribed protein-coding transcripts in their expressed tissue(s).

Fifty percent of non-coding genes (5,559) encoded either homologous short peptides (9-43 amino acids; 5.8% of non-coding genes) or homologous ncRNAs (49% of non-coding genes) (Supplemental file 1: Fig. S14B). There were 5,546 non-coding genes (51% of non-coding genes) that encoded cattle-specific ncRNAs (Supplemental file 1: Fig. S14B). Ninety-nine percent of these genes (5,537 genes) were either annotated genes or genes with homology to another cattle gene(s) that has established homology to genes in other species (Supplemental file 1: Fig. S14C). The remaining 1% (nine non-coding genes) were denoted as non-coding orphan genes (Supplemental file 1: Fig. S14C). The median number of expressed tissues for non-coding orphan genes was 17 tissues, which was higher (p-value < 2.2e-16) than for homologous non-coding genes (six tissues) and protein-coding orphan genes (one tissue) (Supplemental file 1: Fig. S15C).

A total of 3,029 pseudogenes were expressed. The median expression level of these genes in their expressed tissues was 2.15 RPKM, which was lower than that observed for protein-coding genes (7.08 RPKM) and similar to that observed for non-coding genes (1.7 RPKM) (Supplemental file 1: Fig. S16A). Pseudogenes were expressed in a median of four tissues (Supplemental file 1: Fig. S16B). The median number of expressed tissues for protein-coding and non-coding genes was 44 tissues and five tissues, respectively (Supplemental file 1: Fig. S16B). In addition, a total of 1,038 pseudogene-derived lncRNAs were expressed. The median expression of pseudogene-derived lncRNAs was 1.8 RPKM, similar to that observed for other lncRNAs (1.6 RPKM) (Supplemental file 1: Fig. S17A). In addition, pseudogene-derived lncRNAs were expressed in a median of four different tissues, which was lower than observed for other lncRNAs (seven tissues) (Supplemental file 1: Fig. S17B).

Testis had the highest number of expressed pseudogene-derived lncRNAs (427), followed by fetal brain (315) (Supplemental file 1: Fig. S8A and B). The correlation between the number of input reads and the number of pseudogene-derived lncRNAs was not significant (0.25, p-value 0.09).

Gene expression diversity across tissues

Tissue similarities increased dramatically from transcript level to gene level (Supplemental file 1: Fig. S4A, Fig. S5A, Fig. S18A, Fig. S19A). The median percentage of shared genes between pairs of tissues was significantly higher in protein-coding genes compared to non-coding genes

(90% and 57%, respectively; p-value < 2.2e-16; Supplemental file 1: Fig. S18A, Fig. S19A). Clustering of tissues based on protein-coding genes was similar to that observed based on protein-coding transcripts (Supplemental file 1: Fig. S18B, Fig. S19B). The same result was observed in non-coding genes and transcripts. In addition, clustering of tissues based on protein-coding genes was different than that of non-coding genes (Supplemental file 1: Fig. S4B, Fig. S5B, Fig. S18B, Fig. S19B, Fig. S35F).

Tissues with both fetal and adult samples (brain, kidney, muscle, and spleen) were used to investigate gene biotype differences between these developmental stages. Similar to what was observed at transcript level, fetal tissues were significantly enriched for non-coding genes and pseudogenes and were depleted for protein-coding genes (p-value < 2.2e-16; Supplemental file 10). These results were consistent across all tissues with both adult and fetal samples (Supplemental file 10).

Gene validation

A total of 32,460 genes (92% of predicted genes) were structurally validated by independent datasets (PacBio Iso-seq data, ONT-seq data, *de novo* assembled transcripts from RNA-seq data) and comparison with Ensembl and NCBI gene sets (see Method section). In addition, a total of 31,635 genes (90% of predicted genes) were expressed in multiple tissues (31,635 genes or 90%) (Fig. 7). All genes were extensively supported by data from different technologies such as WTTS-seq, RAMPAGE, histone modification (H3K4me3, H3K4me1, H3K27ac) and CTCF-DNA binding, and ATAC-seq data generated from the samples (Fig. 7).

Identification and validation of annotated gene border extensions

This new bovine gene set annotation extended (5′ end extension, 3′ end extension, or both) more than 11,000 annotated Ensembl or NCBI gene borders. Extensions were longer on the 3′ side, but the median increase was 104 nt for the 5′ end (Table 5). To validate gene border extensions, independent WTTS-seq (24 tissues) and RAMPAGE datasets (30 tissues) were utilized. More than 80% of annotated gene border extensions were validated by independent data (Fig. 8). The extension of annotated gene borders on both ends resulted in an approximate nine-fold expression increase of these genes in the new bovine gene set annotation compared to their matched Ensembl and NCBI genes (Table 6). This effect was smaller in annotated genes extended only on 5′ or 3′ ends (Table 6).

Alternative splicing events

Alternative splicing (AS) events (Supplemental file 1: Fig. S20A) are commonly distinguished in terms of whether RNA transcripts differ by inclusion or exclusion of an exon, in which case the exon involved is referred to as a “skipped exon” (SE) or “cassette exon”, “alternative first exon”, or “alternative last exon”. Alternatively, spliced transcripts may also differ in the usage of a 5' splice site or 3' splice site, giving rise to alternative 5' splice site exons (A5Es) or alternative 3' splice site exons (A3Es), respectively. A sixth type of alternative splicing is referred to as “mutually exclusive exons” (MXEs), in which one of two exons is retained in RNA but not both. However, these types are not necessarily mutually exclusive; for example, an exon can have both an alternative 5' splice site and an alternative 3' splice site, or have an alternative 5' splice site or 3' splice site, but be skipped in other transcripts. A seventh type of alternative splicing is "intron retention", in which two transcripts differ by the presence of an unspliced intron in one transcript that is absent in the other. An eighth type of alternative splicing is "unique splice site exons" (USEs), in which two exons overlap with no shared splice junction. A total of 102,502 transcripts (85% of spliced transcripts) were involved in different types of AS events, a large increase over Ensembl (63% of spliced transcripts) and NCBI (75% of spliced transcripts) annotations (Additional file1: FigureS20B). Skipped exons were observed in a greater number of transcripts compared to other types of AS events (Supplemental file 1: Fig. S21).

A median of 60% of tissue transcripts showed at least one type of AS event (Supplemental file 1: Fig. S22A). There was no significant correlation between the number of input reads and the number of AS event transcripts across tissues (Supplemental file 1: Fig. S22B).

The median expression level of AS transcripts (111,366 transcripts or 65% of transcripts) was 1.38 RPKM, which was similar to that observed for other types of transcripts (1.58RPKM) (Supplemental file 1: Fig. S23A). In addition, AS transcripts were expressed in a median of 10 tissues (Supplemental file 1: Fig. S23B), which was higher than for the other transcript types (median of nine tissues). Alternatively spliced transcripts were enriched with protein-coding transcripts (p-value < 2.2e-16). Meanwhile, transcripts that did not show AS events, i.e., unspliced transcripts and spliced transcripts from single transcript genes, were enriched for non-coding transcripts (p-value < 2.2e-16). A median of 67% of protein-coding genes showed at least one type of AS event. In contrast, this was only 3% in non-coding genes. In most cases, AS events did not change transcript biotypes (Supplemental file 1: Fig. S24). In addition, a switch from protein-coding to ncRNAs was the main biotype change resulting from AS events (Supplemental file 1: Fig. S24).

A median of four AS events were expressed in alternatively spliced genes (14,260 genes or 40% of genes) (Supplemental file 1: Fig. S25). The top five percent of genes with the highest number of AS events (2,734 genes, Fig. 35A) were highly enriched for several BP GO terms related to different aspects of RNA splicing (Supplemental file 1: Fig. S26B, Supplemental file 11).

Comparison of tissues with both fetal and adult samples (brain, kidney, Longissimus Dorsi (LD) muscle, and spleen) revealed a significantly higher rate of AS events in fetal tissues (only genes expressed in both fetal and adult samples were included in this analysis) (Supplemental file 1: Fig. S27).

Tissue specificity

Nine percent of all genes (3,174) and transcripts (15,562) were only expressed in a single tissue and were denoted as tissue-specific (Supplemental file 1: Fig. S28A). Most tissue-specific genes (75%) and transcripts (84%) were un-annotated. Forty-nine percent of tissue-specific transcripts (11,748) were produced by annotated genes. Most tissue-specific genes (61%) and transcripts (57%) were protein-coding (Supplemental file 1: Fig. S28A and B). In addition, more than 70% of tissue-specific transcripts (11,222) were transcribed from non-tissue-specific genes. Compared to other tissues, testis and thymus had the highest number of tissue-specific genes and transcripts (Supplemental file 1: Fig. S28C, Supplemental file 12). The expression level of tissue-specific genes and transcripts was significantly lower than that of their non-tissue-specific counterparts (p-value < 2.2e-16; Supplemental file 1: Fig. S28D). A median of 71% of tissue-specific transcripts showed any type of AS event in their expressed tissues (Supplemental file 1: Fig. S29). This was only 3.9% for tissue-specific genes (Supplemental file 1: Fig. S29). Testis, myoblasts, mammary gland, and thymus had the highest proportion of tissue-specific genes displaying any type of AS event (Supplemental file 1: Fig. S29).

A total of 16,806 multi-tissue expressed genes (53% of all multi-tissue expressed genes) and 74,487 multi-tissue expressed transcripts (51% of all multi-tissue expressed transcripts) showed Tissue Specificity Index (TSI) scores (Supplemental file 13) greater than 0.9 and were expressed in a tissue-specific manner. These genes and transcripts were expressed in a median of six tissues and four tissues, respectively (Supplemental file 1: Fig. S30A and B). Functional enrichment analysis of the top five percent of genes with the highest TSI score (3,171 genes) resulted in the identification of “sexual reproduction” (p-value 3.06e-24) and “fertilization” (p-value 1.04e-8) as their top enriched BP GO terms (Supplemental file 1: Fig. S30C-E, Supplemental file 14).

Tying genes to phenotypes

There were 9,800 predicted genes identified as the closest expressed gene to an existing QTL (QTL-associated genes) in their expressed tissues (Supplemental file 15). These genes had either QTLs located inside (6,511 genes) or outside (5,306 genes) their genomic borders (either from their 5’ end or 3’ end) with a median distance of 51.9 kilobases (KB) and a maximum distance of 2.6 million bases (MB) (Supplemental file 1: Fig. S31). Most QTL-associated genes were annotated genes (8,130 genes or 83%). In addition, the median number of AS events in these genes (eight) was significantly higher than that observed in other genes (median of seven AS events; p-value 5.69e-09).

Potential testis-pituitary axis

Testis tissue was not clustered with any other tissues and had the highest number of tissue-specific genes (1,195 genes) compared to the rest of the tissues (Supplemental file 1: Fig. S4, Fig. S5, Fig. S18, and Fig. S19). Testis-specific genes were highly enriched with different traits related to fertility (e.g., percentage of normal sperm and scrotal circumference), body weight (e.g., body weight gain and carcass weight), and feed efficiency (e.g., residual feed intake) (Supplemental file 16). The extent of testis-pituitary axis involvement in the "percentage of normal sperm" was investigated using animals with both testis and pituitary samples (three samples per tissue). The *SPACA5* gene was the only testis-specific gene encoded protein with a signal peptide (SP) that was close to the "percentage of normal sperm" QTLs. The expression of this gene in testis samples showed significant positive correlation with 70 pituitary expressed genes that were closest to the "percentage of normal sperm" QTLs (Supplemental file 1: Fig. S32, Supplemental file 17). These pituitary genes were enriched with the "signal transduction in response to DNA damage" BP GO term (Supplemental file 1: Fig. S32). In addition, the expression of testis genes that encoded protein with a signal peptide that were close to the "percentage of normal sperm" QTLs was significantly correlated with expression of pituitary genes close to this trait (Fig. 9, Supplemental file 18). The same result was observed for the pituitary-testis tissue axis (Supplemental file 1: Fig. S33, Supplemental file 19).

Trait similarity network

The extent of genetic similarity between different bovine traits was investigated using their associated QTLs. A total of 1,857 significantly similar trait pairs (184 different traits) were identified and used to create a bovine trait similarity network (https://www.animalgenome.org/host/reecylab/a; Supplemental file 20).

miRNAs

A total of 2,007 miRNAs (at least ten mapped reads in each tissue) comprised of 973 annotated and 1,034 un-annotated miRNAs were expressed (Supplemental file 21). In each tissue, a median of 704 annotated miRNAs and 549 un-annotated miRNAs were expressed (Fig. 10A). The median expression of un-annotated miRNAs was 0.10 Reads Per Million (RPM), which was significantly lower than that observed for annotated miRNAs (0.41 RPM; p-value 3.25e-25; Fig. 10B). In addition, un-annotated miRNAs were expressed in a median of 23 tissues, significantly lower than for annotated miRNAs (43 tissues; p-value 1.00e-45; Fig. 10C). A median of 84.53% of miRNAs were shared between pairs of tissues (Supplemental file 1: Fig. S34). Clustering of tissues based on miRNAs was similar to what was observed based on non-coding genes (Supplemental file 1: Fig. S35).

A total of 113 miRNAs (5.6%) were expressed in a single tissue and were denoted as tissue-specific (Supplemental file 1: Fig. S36A). The proportion of tissue-specific miRNAs was higher for un-annotated miRNAs, such that 75% of the tissue-specific miRNAs (85) were un-annotated. The number of un-annotated miRNAs was higher in pre-adipocytes compared to other tissues, followed by fetal gonad and testis (Supplemental file 1: Fig. S36B). Un-annotated miRNAs showed a significantly lower expression level compared to annotated miRNAs (p-value 1.4e-19; Supplemental file 1: FigureS36 C). In addition, a total of 1,047 multi-tissue expressed miRNAs (55% of all multi-tissue expressed miRNAs) had a TSI score greater than 0.9 and were expressed in a tissue-specific manner (Additional file1: Fig. S36D). These miRNAs were expressed in a median of 19 tissues (Supplemental file 1: Fig. S36E).

Chromatin features across 500-base pair (bp) windows surrounding upstream of miRNA precursors' start sites or downstream of miRNA precursors' terminal sites from independent cattle experiments were used to investigate the relationship between miRNAs and chromatin accessibility. More than 99% of un-annotated miRNAs (1,027) and 94% of annotated miRNAs (923) were supported by at least one of the H3K4me3, H3K4me1, H3K27ac, CTCF-DNA binding, or ATAC-seq peaks (Fig. 11).

Summary of expressed transcripts, genes, and miRNAs

The numbers of expressed transcripts, genes, and miRNAs in different tissues are summarized in Supplemental file 1: Fig. S37. In addition, the number of annotated and un-annotated genes, transcripts, and miRNAs in different tissues are summarized in Supplemental file 1: Fig. S38.

Discussion

Despite many improvements in the current bovine genome annotation ARS-UCD1.2 assembly (Ensembl release 2021-03 and NCBI release 106) compared to the previous genome assembly (UMD3.1), these annotations are still far from complete [9, 10]. In this study, using RNA-seq and miRNA-seq data from 47 different bovine tissues/cell types, 12,698 un-annotated genes and 1,034 un-annotated miRNAs were identified that have not been reported in current bovine genome annotations (Ensembl release 2021-03, NCBI release 106 and miRbase [11]). In addition, we identified protein-coding transcripts with a median ORF length of 270 nt for 822 annotated bovine genes that have been annotated as non-coding in current bovine genome annotations (Supplemental file 1: Fig. S14C). The high frequency of validation of these un-annotated genes and un-annotated miRNAs using multiple independent datasets from different technologies verifies the improvement in terms of the number of genes and miRNAs using our methods.

Five prime and 3‘untranslated region length plays a critical role in regulation of mRNA stability, translation, and localization [4]. However, only a single 5’ UTR and 3’ UTR per gene is annotated in current bovine genome annotations (Ensembl release 2021-03 and NCBI release 106), and variations in UTR length are not available. In this study, 7,909 genes (22% of predicted genes) with multiple UTRs were identified. Genes with multiple 5’ UTRs are common, primarily due to the presence of multiple promoters [12] or alternative splicing mechanisms within 5’ UTRs [12]. Fifty-four percent of human genes have multiple transcription start sites [12]. In addition, the length of 3’ UTRs often varies within a given gene, due to the use of different poly(A) sites [4, 13].

In this study, around 50% of expressed protein-coding genes in each tissue transcribed both coding and non-coding transcript isoforms. Several studies have shown evidence of the existence of bifunctional genes with coding and non-coding potential using RNA-seq and ribosome footprinting followed by sequencing (Ribo-seq) [14-16]. More than 20% of human protein-coding genes have been reported to transcribe non-coding isoforms, often generated by alternative splicing [17] and recurrently expressed across tissues and cell lines [16]. A considerable number of non-coding isoform variants of protein-coding genes appear to be sufficiently stable to have functional roles in cells [18]. It has been shown that the proportion of non-coding isoforms from protein-coding genes dramatically increases during myogenic differentiation of primary human satellite cells and decreases in myotonic dystrophy muscles [19]. In this study, 106 non-coding genes were identified in fetal tissues that switched to protein-coding genes in their matched adult tissues. Taken together this supports the notion that protein-coding/non-coding transcript switching plays an important role in tissue development in cattle as well.

Nonsense-mediated RNA decay is an evolutionarily conserved process involved in RNA quality control and gene regulatory mechanisms [20]. For instance, the RNA-binding protein polypyrimidine tract binding protein 1 (*PTBP1*) can promote the transcription of NMD transcripts via alternative splicing, which negatively regulates its own expression [21]. In this study, NMD transcripts comprised 19% of bovine transcripts that were transcribed from 30% of bovine genes (10,498). In humans, NMD-mediated degradation can affect up to 25% of transcripts [22] and 53% of genes [23]. As expected, in this study, most genes that transcribed NMD transcripts were protein coding (83% or 8,687 genes), while a considerable portion (17%) were pseudogenes. Many pseudogenes are annotated to give rise to NMD transcripts [24, 25]. Bioinformatic study of the human transcriptome revealed that 78% of NMD transcript–producing genes were protein coding, followed by pseudogenes (nine percent), long intergenic noncoding RNAs (six percent), and antisense transcripts (four percent) [25].

Despite the important regulatory function of lncRNAs and miRNAs, very low numbers of these elements have been annotated in the current bovine genome annotations (Table 7). In this study, a total of 10,789 lncRNA genes and 2,007 miRNA genes were expressed in the bovine transcriptome, which is similar to what has been reported for the human transcriptome (Table

7). While, a total of 3,770 human miRNAs and 1,203 cattle miRNAs have been reported in miRbase [11].

In this study, 1,038 pseudogene-derived lncRNAs were identified that were recurrently expressed across tissues and cell types. Ever-increasing evidence from different studies suggests pseudogene derived RNAs are key components of lncRNAs [26-28]. lncRNAs expressed from pseudogenes have been shown to regulate genes with which they have sequence homology [26, 27] or to coordinate development and disease in metazoan systems [26].

Correct annotation of gene borders has an important role in defining promoter and regulatory regions. Our novel transcriptome analysis extended (5′-end extension, 3′-end extension, or both) more than 11,000 annotated Ensembl or NCBI gene borders. Extensions were longer on the 3′ side, which was relatively similar to that we observed in the pig transcriptome using PacBio Iso-Seq data [2].

A growing body of evidence indicates that a considerably large portion of lncRNAs encode microproteins that are less conserved than canonical open reading frames [29-33]. In this study, a vast majority (98%) of predicted lncRNAs had short ORFs (<44 amino acids) that were less conserved than canonical ORFs (Table 2).

Alternative splicing is the key mechanism to increase the diversity of the mRNA expressed from the genome and is therefore essential for response to diverse environments. In this study, skipped exons and retained introns were the most prevalent AS events identified in the bovine transcriptome, similar to what has been observed in other vertebrates and invertebrates [34]. A higher rate of AS events was observed in fetal tissues compared to their adult tissue counterparts. The same result has been observed in a recently published study in humans [35].

We hypothesized that the integration of the gene/transcript data with previously published QTL/gene association data would allow for the identification of potential molecular mechanisms responsible for a) tissue-tissue communication as well as b) genetic correlations between traits. To test the first hypothesis, we developed a novel approach to study the involvement of tissue-tissue interconnection in different traits based on the integration of the transcriptome with publicly available QTL data. In particular, the interconnection between testis and pituitary tissues with respect to the “percentage of normal sperm” trait was investigated in more detail. This resulted in the identification of the regulation of ubiquitin-dependent protein catabolic process, the regulation of Nuclear factor-κB (NF-κB) transcription factor activity, and Rab protein signal transduction as key components of this tissue-tissue interaction (Supplemental file 18 and 19). Interestingly, expressed genes that were closest to “percentage of normal sperm” QTLs, and also encoded protein with a signal peptide (short peptide present at the N-terminus of proteins that are destined toward the secretory pathway[36]) in both testis and pituitary tissues, were highly enriched for the BP GO term “regulation of ubiquitin-dependent protein catabolic process” (Supplemental file 18 and 19). The expression of these genes in testis tissue was significantly correlated with expression levels of pituitary expressed genes closest to “percentage of normal sperm” QTLs that were highly enriched for the “positive regulation of NF-kappaB transcription factor activity” BP GO term (Supplemental file 1: Fig. S32 and Supplemental file 18). Activation of NF-κB requires ubiquitination, and this modification is highly conserved across different species [37]. NF-κB induces secretion of adrenocorticotropic hormone from the pituitary [38], which directly stimulates testosterone production by the testis [39]. In addition, ubiquitinated proteins in testis cells are required for the progression of mature spermatozoa [40]. The expression levels of pituitary expressed genes closest to “percentage of normal sperm” QTLs that also encoded signal peptides were significantly correlated with expression levels of testis expressed genes closest to “percentage of normal sperm” QTLs (Supplemental file 1: Fig. S33). These testis genes were highly enriched for the “Rab protein signal transduction” BP GO term (Supplemental file 19). Rab proteins have been reported to be involved in male germ cell development [41]. Thus, it appears that integration of gene data with QTL/association data can be used to identify putative molecular pathways underlying tissue-tissue communication mechanisms.

To test the second hypothesis, we also developed a novel approach to study trait similarities based on the integration of the transcriptome with publicly available QTL data. Using this approach, we could identify significant similarity between 184 different bovine traits. For example, clinical mastitis showed significant similarity with 23 different cattle traits that were greatly supported by published studies, such as milk yield [42], milk composition traits [43], somatic cell score [44], foot traits [45], udder traits [46], daughter pregnancy rate [47], length of productive life [48] and net merit [49]. Similar results were observed for residual feed intake, which showed significant similarity with 14 different traits such as average daily feed intake [50], average daily gain [51], carcass weight [52], feed conversion ratio [53], metabolic body weight [54], subcutaneous fat [55], and dry matter intake [56].

Taken together, these results identify a list of candidate genes that might harbor genetic variation responsible for the genetic mechanisms underlying genetic correlations

(Supplemental file 18 and 1. If this is the case, in the future, these novel methods should be able to predict the impact of a given set of genetic variants that are associated with a trait of interest on other traits that were not measured in a given study. This might then lead to the optimization of variants used (or not used) in genomic selection to minimize any non-beneficial effect of selection on selected traits. However, it is important to acknowledge that the nearest neighbor gene to a genotype association is not necessarily the causal gene. None the less, these results are intriguing in that meaningful genetic correlation can be recapitulated.

Conclusions

In-depth analysis of multi-omics data from 47 different bovine tissues/cell types provided evidence to improve the annotation of thousands of protein-coding, lncRNA, and miRNA genes. These validated results increase the complexity of the bovine transcriptome (number of transcripts per gene, number of UTRs per gene, lncRNA transcripts, AS events, and miRNAs), comparable to that reported for the highly annotated human genome. We provided direct evidence that the predicted un-annotated transcripts extend existing annotated gene models, by verifying such extensions using independent WTTS-seq and RAMPAGE data. We utilized a novel approach to integrate the transcriptome with publicly available QTL data and showed its application in a study of tissue axis involvement in different traits and genetic similarity between different traits. This approach is particularly important in the selection of indicator traits for breeding purposes, study of artificial selection side effects in livestock species, and functional annotation of poorly annotated livestock genomes.

Methods

Tissue and cell collection, total RNA extraction and construction of RNA-seq, miRNA-seq, WTTS-seq, ATAC-seq, and ChIP-seq libraries

Cell sample collections. Skeletal muscle and subcutaneous fat samples were collected from Angus-crossbred steers slaughtered at the Virginia Tech Meat Center. Satellite cells were isolated from skeletal muscle by pronase digestion as described previously (Leng et al. 2019). The isolated satellite cells were activated to proliferate as myoblasts by culturing in growth medium composed of Dulbecco's Modified Eagle Medium (DMEM), 10% fetal bovine serum (FBS), and 1% antibiotics-antimycotics. To induce myoblasts to differentiate into myocytes, myoblasts cultured in growth medium were switched to differentiation medium composed of DMEM and 2% horse serum for 2 days. Preadipocytes from subcutaneous fat were isolated by collagenase digestion as previously described (Hausman et al. 2008). To induce preadipocytes to differentiate into adipocytes, preadipocytes were initially cultured in growth medium (DMEM/F12, 10% FBS, 1% antibiotics-antimycotics) to reach confluency, then in induction medium (DMEM/F12, 10% FBS, 1% antibiotics-antimycotics, 10 μg/mL insulin, 1 μM dexamethasone, 0.5 mM isobutyl methylxanthine, and 200 μM indomethacin) for 2 days, and lastly in maintenance medium (DMEM/F12, 10% FBS, 1% antibiotics-antimycotics, 1 μg/mL insulin) for 10 days.

Adult tissue collections. Procedures for tissue collection followed the Animal Care and Use protocol (#18464) approved by the Institutional Animal Care and Use Committee (IACUC), University of California, Davis (UCD). Four cattle (2 males and 2 females) were slaughtered at UCD using captive bolt under USDA inspection at 14 months old and were intact male and female Line 1 Herefords that had the same sire, provided by Fort Keogh Livestock and Range Research Lab [57]. Tissue samples were flash frozen in liquid nitrogen then stored at –80 °C until further assay processing.

Fetal tissue collections. Fetal sample collection and tissue collection were approved by IACU), University of Idaho (2017-67). Four pregnant females at day 78 of gestation Line 1 Herefords were slaughtered at UI meats lab using captive bolt under USDA inspection. Animals were provided by Fort Keogh Livestock and Range Research Lab (Tixier-Boichard et al. 2021). Tissue samples were flash frozen in liquid nitrogen then stored at –80 °C until further assay processing.

RNA-seq library construction. Tissue samples (Supplemental file 22) were collected from storage at -80 °C and ground to a powder using a mortar and pestle and liquid nitrogen. The tissue was next homogenized in QIAzol Lysis Reagent (Qiagen Catalog No. 79306) using a QIAshredder spin column (Qiagen Catalog No. 79656). After centrifugation, the lysate was mixed with chloroform, shaken vigorously for 15 sec, incubated for 2 – 3 min at room temperature, and centrifuged for 15 min at 12,000 x g at 4°C. The upper, aqueous phase was transferred to a new collection tube and 1.5 vol of 100% ethanol was added and mixed thoroughly by pipetting up and down several times. Total RNA was then isolated from the sample using the RNeasy Mini Kit (Qiagen Catalog No. 74106) according to the manufacturer's instructions. Contaminating DNA was removed by treating total RNA with DNase (AM1906, Ambion). Total RNA quantity was measured with the Quant-It RiboGreen RNA Assay Kit (Life Technologies Corp., Carlsbad, CA) and quality assessed by fragment analysis (Advance Analytical Technologies, Inc., Ankeny IA).

Mammary gland tissue collection and RNA-seq library construction. The 16 animals used in this study were Holstein-Friesian heifers from a single herd managed at the AgResearch Research Station in Ruakura, NZ. All experimental protocols were approved by the AgResearch, NZ, ethics committee, and carried out according to their guidelines. Samples were collected from the same animals at 5 time points: virgin state before pregnancy between 13 and 15 months of age (virgin), mid-pregnant at day 100 of pregnancy, late pregnant ~2 weeks pre-calving, early lactation ~2 weeks post-calving, and at peak lactation, 34-38 days post-calving. Tissue samples were obtained by mammary biopsy using the Farr method [58]. Lactating cows were milked before biopsy and sampled within 5 hours of milking. Biopsy sites were clipped and given aseptic skin preparation (povidone iodine base scrub and iodine tincture) and subcutaneous local anesthetic (4 ml per biopsy site). Core biopsies were taken using a powered sampling cannula (4.5 mm internal diameter) inserted into a 2 cm incision. The resulting samples of mammary gland parenchyma measured 70 mm in length, with a 4 mm diameter. Small slices from each sample were preserved for histology before mammary epithelial organoids were separated from surrounding adipose and connective tissue to allow for secretory-specific signals in the RNA-seq analysis. In preparation for isolating organoids, tissue samples were digested in a freshly prepared collagenase solution containing 0.2% collagenase A (Roche), 0.05% trypsin (1:250 powder, 100U/ml Gibco), hyaluronidase (Sigma), 5% fetal calf serum (Hyclone), Pen/Strep/Fungizone solution (Hyclone) or 5 µg/ml Gentamycin (Sigma) in DMEM/F12 (Gibco) with 10 ng/ml insulin. Samples were minced to a fine slurry and incubated in this freshly prepared collagenase solution (10 ml solution/g tissue) for 3.5 hours at 37°C in a 50 ml conical tube with slow shaking (120 rpm). Digested tissue was centrifuged at 453 x g for 10 min at 4°C, after which the supernatant and fat layers were discarded, and the pellet was gently resuspended in 5 ml DMEM/F12 without serum. A further 5 ml DMEM/F12 without serum was added, and the sample was centrifuged at 453 x g for 10 min at 4°C. The media was discarded, and the pellet was gently resuspended in 10 ml DMEM/F12 and centrifuged for another 10 min at 453 x g and 4°C. The media was discarded, and pellet resuspended in 10 ml DMEM/F12 for a third time, and the sample centrifuged in a series of brief spins achieved by allowing the centrifuge to reach 453 x g for two seconds before applying the brake. These brief pulse spins were repeated at least 4 times, or until examining the sample under a microscope revealed primarily epithelial organoid clusters and very few single cells. At this point, the organoid pellet was resuspended in 1 ml TRIzol and stored at -80°C until RNA isolation. High-quality total RNA (RIN > 7) was extracted from frozen mammary epithelial organoid pellets using NucleoSpin® miRNA isolation kit (MACHEREY-NAGEL) according to the manufacturer's protocol, isolating large and small (<200 bp) fractions separately. The "large" RNA fraction was used to prepare strand-specific poly(A)+ RNA-seq libraries for sequencing. The "small" RNA fraction was used to make miRNA-seq libraries using NEXTflex^{TM} Small RNA-Seq Kit v3.

miRNA-seq library construction. Tissue samples (Supplemental file 22) were collected similarly to the method described in the previous section. QIAseq miRNA Library Kit (Qiagen, cat no. 331505) and QIAseq miRNA NGS 96 Index IL Kit (Qiagen, cat no. 331565) were used to isolate miRNAs from all tissues except mammary gland. miRNAs from mammary gland were isolated using NEXTflex^{TM} Small RNA-Seq Kit v3 (Illumina) according to the manufacturer's instructions. The isolated miRNA was subjected to 3' ligation to ligate a pre-adenylated DNA adaptor to the 3' ends of all miRNAs. An RNA adaptor was then ligated to the 5' end of the mature miRNA to complete 5' ligation. cDNA synthesis was completed using a reverse transcriptase (RT) primer containing integrated unique molecular identifiers (UMI). The RT primer bound to the 3' adaptor region and facilitated conversion of the 3'/5' ligated miRNAs into cDNA while a UMI was assigned to every miRNA molecule. After reverse transcription, a clean-up of the cDNA was performed using a streamlined magnetic bead-based method. Library amplification was accomplished by a universal forward primer from a plate being paired with 1 of 96 dried reverse primers in the same plate (Qiagen, cat no. 331565) to assign each sample a unique custom index. Following library amplification, a clean-up of the miRNA library was performed using a streamlined magnetic bead-based method. Libraries were then evaluated for quantity and quality measures before being normalized and pooled for llumina sequencing (1×50bp).

WTTS-seq library construction. Construction of the WTTS-seq libraries from tissue samples (Supplemental file 22) involved fragmentation, poly(A)+ RNA enrichment, first-strand cDNA synthesis by reverse transcription and second-strand cDNA synthesis by PCR as described previously [59]. The starting material was 2.5 µg of total RNA per library, which was fragmented with 1 µl of 10X RNA fragmentation buffer (Ambion, AM8740), followed by enrichment of poly(A)+ RNA using Dynabeads (Ambion 61002). The poly(A)+ RNA molecules were then used for the first-strand cDNA synthesis with both 5' adaptor (switching primer, 100 µM) and 3' adaptor (containing oligo (dT10), 100 µM) catalyzed by the SuperScript III reverse transcriptase (200 U/µl) (Invitrogen, 18080). The first-strand cDNA molecules were chemically enriched with RNases I and H and used to synthesize the second-strand cDNA using PCR. Base PCR conditions were as follow: initial denaturation at 98 °C for 30 s, PCR cycles of 98 °C for 10 s, 50°C for 30 s, and 72°C for 30 s, and final extension at 72°C for 10 min. The size-selected cDNA (200 – 500 bp) was purified with SPRI beads (Agencourt AMPure XP beads, Beckman Coulter, Brea, CA) and sequenced using an Ion PGM™ Sequencer at Washington State University.

ATAC-seq library construction. Frozen tissue samples (Supplemental file 22) were pulverized under liquid nitrogen using mortar and pestle. Permeabilized nuclei were obtained by resuspending pulverized tissue (5-15 mg) in 250 µL Nuclear Permeabilization Buffer (0.2% IGEPAL-CA630 [I8896, Sigma], 1 mM DTT [D9779, Sigma], Protease inhibitor [05056489001, Roche], and 5% BSA [A7906, Sigma] in PBS [10010-23, Thermo Fisher Scientific]), and incubating for 10 min on a rotator at 4°C. Nuclei were then pelleted by centrifugation for 5 min at 500 x g at 4°C. The pellet was resuspended in 25 µL ice-cold Tagmentation Buffer (33 mM Tris-acetate [pH = 7.8; BP-152, Thermo Fisher Scientific], 66 mM K-acetate [P5708, Sigma], 11 mM Mg-acetate [M2545, Sigma], 16% DMF [DX1730, EMD Millipore] in molecular biology grade water [46000-CM, Corning]). An aliquot was then taken and counted by hemocytometer to determine nuclei concentration. Approximately 50,000 nuclei were resuspended in 20 µL ice-cold Tagmentation Buffer and incubated with 1 µL Tagmentation enzyme (FC-121-1030, Illumina) at 37 °C for 30 min with shaking at 500 rpm. The tagmentated DNA was purified using MinElute

PCR purification kit (28004, Qiagen). The libraries were amplified using NEBNext High-Fidelity 2X PCR Master Mix (M0541, NEB) with primer extension at 72°C for 5 min, denaturation at 98°C for 30 s, followed by 8 cycles of denaturation at 98°C for 10 s, annealing at 63°C for 30 s and extension at 72°C for 60 s. Amplified libraries were then purified using MinElute PCR purification kit (28004, Qiagen), and two size selection steps were performed using SPRIselect bead (B23317, Beckman Coulter) at 0.55X and 1.5X bead-to-sample volume ratios, respectively. ATAC-seq libraries were sequenced on an Illumina Nextseq 500 platform using Nextra V2 sequencing chemistry to generate 2 × 75 paired-end reads.

Sequencing the transcriptomes of seven bovine tissues by using the PacBio Iso-Seq and Illumina RNA-Seq technologies

Publicly available PacBio Iso-seq reads and matched RNA-seq reads (PRJNA386670) were used in this experiment. Sequence reads were generated using the following procedure. Frozen tissue samples (Supplemental file 22) were pulverized by grinding with disposable mortar and pestle in liquid nitrogen. RNA was extracted using TRIzol reagent as directed by the manufacturer (Invitrogen) with integrity examined using a BioAnalyzer (Agilent). Only samples with RIN values >8 were used for cDNA synthesis. Libraries for RNA-seq short-read sequencing were prepared using the TruSeq RNA Kit following the “TruSeq RNA Sample Preparation v2 Guide” as recommended by the manufacturer (Illumina). RNA-seq libraries were sequenced on a NextSeq500 instrument. IsoSeq libraries for long-read sequencing were prepared using the SMRTbell Template Prep Kit 1.0. First strand cDNA synthesis was performed with approximately 1 µg of extracted RNA from each tissue using the Clontech SMARTer PCR cDNA Synthesis Kit (Clontech) as directed by the manufacturer. cDNA was then converted to SMRTbell template library following the “Iso-Seq using Clontech cDNA Synthesis and BluePippin Size Selection” protocol as directed by the manufacturer (Pacific Biosciences). Three size fraction pools for each tissue were prepared using the BluePippin instrument (Sage Science), representing insert sizes of 1-2 kb, 2-3 kb, and 3-6 kb. The two smaller fractions were sequenced in three to five SMRT cells on an RSII instrument (Pacific Biosciences), and the largest fraction sequenced in five or six cells, using P6/C4 chemistry. The sequences were processed into HQ isoforms using SMRT Analysis v6.0 for each tissue independently but with all size fractions within tissue included in the analysis.

RNA-seq data analysis and transcriptome assembly

Single-end Illumina RNA-Seq reads (75 bp) from each tissue sample were trimmed to remove the adaptor sequences and low-quality bases using Trim Galore (version 0.6.4) [60] with --quality 20 and --length 20 option settings. The resulting reads were aligned against ARS-UCD1.2 bovine genome using STAR (version 020201) [61] with a cut-off of 95% identity and 90% coverage. FeatureCounts (version 2.0.2) [62] was used to quantify genes reported in the NCBI gene build (version 1.21) with -Q 255 -s 2 --ignoreDup --minOverlap 5 option settings. The resulting gene counts were adjusted for library size and converted to Counts Per Million (CPM) values using SVA R package (version 3.30.0) [63]. In each tissue, sample similarities were checked using hierarchical clustering and regression analysis of gene expression values (log2 based CPM), and outlier samples were expressed and removed from downstream analysis. Samples from each tissue were combined to get the most comprehensive set of data in each tissue. To reduce the processing time due to huge sequencing depth, the trimmed reads were in silico normalized using insilico_read_normalization.pl from Trinity package (version 2.6.6)

[64] with --JM 350G and --max_cov 50 option settings. Normalized RNA-seq reads were aligned against ARS-UCD1.2 bovine genome using STAR (version 020201) [61] with a cut-off of 95% identity and 90% coverage. The normalized reads were assembled using *de novo* Trinity software (version 2.6.6) [64] combined with massively parallelized computing using HPCgridRunner (v1.0.1) [65] and GNU parallel software [66]. The resulting transcript reads were collapsed and grouped into putative gene models (clustering transcripts that had at least a one-nucleotide overlap) by the pbtranscript-ToFU from SMRT Analysis software (v2.3.0) [67] with min-identity = 95%, min-coverage = 90% and max_fuzzy_junction = 15 nt, whereas the 5'-end and 3'-end difference were not considered when collapsing the reads. Base coverage of the resulting transcripts was calculated using mosdepth (version 0.2.5) [68]. Predicted transcripts were required to have a minimum of three times base coverage in their assembled tissues. The predicted acceptor and donor splice sites were required to be canonical and supported by Illumina-seq reads that spanned the splice junction with 5-nt overhang. Spliced transcripts with the exact same splice junctions as their reference transcripts but that contained retained introns were removed from analysis, as they were likely pre-RNA sequences. Unspliced transcripts with a stretch of at least 20 A's (allowing one mismatch) in a genomic window covering 30 bp downstream of their putative terminal site were removed from analysis, as they were likely genomic-DNA contaminations. To decrease the false positive rate, unspliced transcripts that were only expressed in a single tissue were removed from downstream analysis. In addition, single-exon genes without histone mark (H3K4me3, H3K4me1, H3K27ac) or ATAC-seq peaks mapped to their promoter (see Relating transcripts and genes to epigenetic data section) were removed from downstream analysis as they were likely transcriptional noise.

The resulting transcripts from each tissue were re-grouped into gene models using an in-house Python script. Structurally similar transcripts from the different tissues (see Comparison of transcript structures across datasets/tissues section) were collapsed using an in-house Python script to create the RNA-seq based bovine transcriptome.

The resulting transcripts and genes were quantified using align_and_estimate_abundance.pl from the Trinity package (version 2.6.6) [64] with --aln_method bowtie --est_method RSEM --SS_lib_type R option settings.

“Isoform” and “transcript” terms are used interchangeably throughout the manuscript.

PacBio Iso-Seq data analysis

PacBio Iso-seq data has been processed as described for the pig transcriptome [2] with the following exceptions. Errors in the full-length, non-chimeric (FLNC) cDNA reads were corrected with the preprocessed RNA-Seq reads from the same tissue samples using the combination of proovread (v2.12) [69] and FMLRC (v1.0.0) [70] software packages. Error rates were computed as the sum of the numbers of bases of insertions, deletions, and substitutions in the aligned FLCN error-corrected reads divided by the length of aligned regions for each read (Table 8).

The RNA-seq-based transcriptome was assembled as described in the previous section.

Oxford Nanopore data analysis

Assembled isoforms from a previously published Oxford Nanopore experiment were used in this study (Halstead et al. 2021).

Comparison of transcript structures across datasets/tissues

When comparing transcripts across datasets/tissues, transcripts whose 5' and 3' borders were supported by RAMPAGE and/or WTTS data (see Transcript and gene border validation section) and whose splice junctions were identical (maximum fuzzy junction was set to 15 bp) were considered "structurally equivalent transcripts". The maximum of 100 nt fuzzy 5' and 3' transcript borders were applied when comparing transcripts were not supported by RAMPAGE and/or WTTS data. Other transcripts that did not met these criteria were considered "structurally different transcripts".

A pair of genes was considered as structurally equivalent across datasets if they transcribed at least single "structurally equivalent transcript".

Prediction of transcript and gene biotypes

Transcripts' open reading frames (ORFs) were predicted using the stand-alone version of ORFfinder [71] with "ATG and alternative initiation codons" as ORF start codon. The longest three ORFs were matched to the NCBI non-redundant vertebrate database and Uniprot vertebrate database using Blastp [71] with E-value cutoff of 10^{-6}, min coverage 60%, and min identity 95%. The ORFs with the lowest E-value to a protein were used as the representative, or if no matches were found, the longest ORF was used. Putative transcripts that had representative ORFs longer than 44 amino acids were labelled as protein-coding transcripts. If the representative ORF had a stop codon that was more than 50 bp upstream of the final splice junction, it was labelled as a nonsense-mediated decay transcript [72]. Transcripts with start codon but no stop codon before their poly(A) site were labelled non-stop decay RNAs. Putative non-coding transcripts (ORFs shorter than 44 amino acids and lack of coding potential predicted by CPC2 [73]) with lengths less than 200 bp that did not overlap with annotated or un-annotated miRNA precursors (see miRNA-seq data analysis section) were labelled as small non-coding RNAs [72]. Putative non-coding transcripts with lengths greater than 200 bp were labelled as long non-coding RNAs [72]. Long non-coding RNAs overlapping one or more coding loci on the opposite strand were labelled as antisense lncRNAs. Long non-coding RNAs located in introns of coding genes on the same strand were labelled as sense-intronic lncRNAs. Long non-coding RNAs that had an exon(s) that overlapped with a protein-coding gene were labeled as Intragenic lncRNAs. Long non-coding RNAs located in intergenic regions of the genome were labeled as Intergenic lncRNAs.

Putative genes that transcribed at least a single protein-coding transcript were labelled as protein-coding genes. Putative genes with homology to existing vertebrate protein-coding genes (Blastx [71], E-value cut-off 10^{-6}, min coverage 90%, and min identity 95%) but containing a disrupted coding sequence, i.e., transcribe only nonsense-mediated decay or non-stop decay transcripts in all of their expressed tissues, were labelled as pseudogenes. The rest of the putative genes were labeled as non-coding.

ncRNAs homology analysis

Putative non-coding transcripts were matched to NCBI and Ensembl vertebrate ncRNA databases using Blastn [71] with E-value cutoff of 10^{-6}, min coverage 90%, and min identity 95%. Transcripts with at least one hit were considered as homologous ncRNAs.

Transcriptome termini site sequencing data analysis

T-rich stretches located at the 5′ end of each WTTS-seq raw read were removed using an in-house Perl script, as described previously [59]. T-trimmed reads were error-corrected using Coral (version 1.4.1) [74] with -v -Y -u -a 3 option settings. The resulting reads were quality trimmed using FASTX Toolkit (version 0.0.14) [75] with -q 20 and -p 50 option settings. High-quality, error-corrected WTTS-seq reads were aligned against the ARS-UCD1.2 bovine genome using STAR (version 020201) [61] with a cut-of of 95% identity and 90% coverage.

ChIP-seq data analysis

Regions of signal enrichment (“peaks”) from a previously published ChIP-seq experiment were used in this study [76].

ATAC-seq data analysis

The UC Davis FAANG Functional Annotation Pipeline was applied to process the ATAC-seq data, as previously described [76]. Briefly, the ARS-UCD1.2 genome assembly and Ensembl genome annotation (v100) were used as references for cattle. Sequencing reads were trimmed with Trim Galore! (Krueger et al. 2015) (v.0.6.5) and aligned with either STAR (Dobin et al. 2012) (v.2.5.4a) or BWA (Li et al. 2013) (v0.7.17) to the respective genome assemblies. Alignments with MAPQ scores <30 were filtered using Samtools (Li et la. 2009) (v.1.9). Duplicate reads were marked and removed using Picard (v.2.18.7). Regions of signal enrichment were called by MACS2 (Zhang et al. 2008) (v.2.1.1).

Relating transcripts and genes to epigenetic data

The promoter was defined as the genomic region that spans from 500 bp 5′ to 100 bp 3′ of the gene/transcript start site. Histone mark (H3K4me3, H3K4me1, H3K27ac), CTCF-DNA binding or ATAC-seq peaks mapped to the promoter of a given gene/transcript were related to that gene/transcript.

Transcript and gene border validation

RAMPAGE peaks from a previously published experiment [10] were used to validate gene/transcript start site. Peaks within the genomic region that spans from 30 bp 5′ to 10 bp 3′ of a gene/transcript start site were assigned to that gene/transcript. WTTS-seq reads (median length of 161 bp) within the genomic region that spans from 10 bp 5′ to 165 bp 3′ of a gene/transcript terminal site were assigned to that gene/transcript.

Functional enrichment analysis

The potential mechanism of action of a group of genes was deciphered using ClueGO [77]. The latest update (May 2021) of the Gene Ontology Annotation database (GOA) [78] was used in the analysis. The list of genes with at least one transcript expressed in a given tissue was used as background for that tissue. The GO tree interval ranged from 3 to 20, with the minimum number of genes per cluster set to three. Term enrichment was tested with a right-sided hypergeometric test that was corrected for multiple testing using the Benjamini-Hochberg procedure [79]. The adjusted p-value threshold of 0.05 was used to filter enriched GO terms.

Alternative splicing analysis

Alternative splicing events, except Unique Splice Site Exons, were detected using generateEvents from SUPPA (version 2.3) [80] with default settings. Unique Splice Site Exons were detected using an in-house Python script.

miRNA-seq data analysis

Single-end Qiagen miRNA-seq reads (50 bp) from each tissue sample were trimmed to remove the adaptor sequences and low-quality bases using Trim Galore (version 0.6.4) [60] with --quality 20, --length 16, --max_length 30 -a AACTGTAGGCACCATCAAT option settings. miRNA reads were aligned against the ARS-UCD1.2 bovine genome using mapper.pl from mirDeep2 (version 0.1.3) [81] with -e -h -q -j -l 16 -o 40 -r 1 -m -v -n option settings. miRNA mature sequences along with their hairpin sequences for Bos taurus species were downloaded from miRbase [11]. These sequences, along with the aligned miRNA reads, were used to quantify annotated miRNAs in each sample using miRDeep2.pl from mirDeep2 (version 0.1.3) [81] with -t bta -c -v 2 setting options. miRNA normalized Reads Per Million (RPM) were used to check sample similarities using hierarchical clustering and regression analysis of gene expression values (log2 based CPM), and outlier samples were detected and removed from downstream analysis. In order to predict the most comprehensive set of un-annotated miRNAs, samples from different tissues were concatenated into a single file that were aligned against the ARS-UCD1.2 bovine genome using mapper.pl from mirDeep2 (version 0.1.3) [81] with the aforementioned settings. Aligned reads from the previous step were used, along with annotated miRNAs' mature sequences and their hairpins, to predict un-annotated miRNAs using miRDeep2.pl from mirDeep2 (version 0.1.3) [81] with the aforementioned settings. Samples from each tissue were combined to get the most comprehensive set of data for that tissue. Mature miRNA sequences and their hairpins for both annotated and predicted un-annotated miRNAs' sequences along with the aligned miRNA reads from each tissue were used to quantify annotated and un-annotated miRNAs in each tissue using mirDeep2 (version 0.1.3) [81] with the aforementioned settings.

Tissue-specificity index

Tissue Specificity Index (TSI) calculations were utilized to present more comprehensive information on transcript/gene/miRNA expression patterns across tissues. This index has a range of zero to one with a score of zero corresponding to ubiquitously expressed transcripts/genes/miRNAs (i.e., "housekeepers") and a score of one for transcripts/genes/miRNAs that are expressed in a single tissue (i.e., "tissue-specific") [82]. The TSI for a transcript/gene/miRNA j was calculated as [82]:

$$TSI_j = \frac{\sum_{i=1}^{N}\left(1 - x_{j,i}\right)}{N - 1}$$

where N corresponds to the total number of tissues measured, and $x_{j,i}$ is the expression intensity of tissue i normalized by the maximal expression of any tissue for transcript/gene/miRNA j.

QTL enrichment analysis

Publicly available bovine QTLs were retrieved from Animal QTLdb [83]. Closest expressed gene to a given trait's QTLs were denoted as QTL-associated genes for that trait. The median distance of QTLs located outside gene borders to the closest expressed gene was 51.9 kilobases and the maximum distance was 2.6 million bases. QTL enrichment was tested with a right-sided Fisher Exact test using an in-house Python script. The resulting p-values were corrected for multiple testing by the Benjamini-Hochberg procedure [79]. The adjusted p-value threshold of 0.05 was used to filter QTLs.

Trait similarity network

For a given pair of traits, trait A was denoted as “similar” to trait B if a significant portion of trait A's QTL-associated genes were also the closest expressed genes to trait B QTLs based on 1000 permutation tests. The resulting p-values were corrected for multiple testing using the Benjamini-Hochberg procedure [79]. The same procedure was used to test trait B's similarity to trait A. The adjusted p-value threshold of 0.05 was used to filter significant trait similarities. A graphical presentation of the method used to construct the tissue similarity network is presented in Supplemental file 1: Fig. S39. The resulting network was visualized using Cystoscape software [84].

Testis-pituitary axis correlation significance test

The presence of signal peptides on representative ORFs of protein-coding transcripts was predicted using SignalP-5.0 [85]. Spearman correlation coefficients were used to study expression similarity between testis genes encoding signal peptides that were closest to the "percentage of normal sperm" QTLs (62 genes) and pituitary expressed genes closest to the "percentage of normal sperm" QTLs (246 genes). To test the statistical difference between these correlation coefficients (reference correlations) and random chance, 1000 random sets of 246 pituitary genes were selected, and their correlation coefficients with 62 previously described testis genes were calculated (random correlations). The reference correlations were compared with 1000 sets of random correlations using a right-sided t-test. The resulting p-values were corrected for multiple testing by the Benjamini-Hochberg procedure [79]. The distribution-adjusted p-values were used to determine the significance level of expression similarities for genes involved in the testis-pituitary axis related to "percentage of normal sperm". The same analysis was conducted to determine the significance of pituitary-testis axis involvement in this trait.

Tissue dendrogram comparison across different transcript and gene biotypes

Tissues were clustered based on the percentage of their transcripts/genes that were shared between tissue pairs using the hclust function in R. Cophenetic distances for tissue dendrograms were calculated using the cophenetic R function. The degree of similarity between dendrograms constructed based on different gene/transcript biotypes was obtained using the Spearman correlation coefficient between the dendrograms' Cophenetic distances.

Figure legends

Figure 1. Distribution of the number of expressed transcripts (A) and genes (B) across tissues.

Figure 2. Classification of the predicted transcripts into different biotypes.

Figure 3. Support of predicted transcripts using data from different technologies and datasets.

Figure 4. Classification of the predicted genes into different biotypes.

Figure 5. Distribution of the number of 5’ UTRs and 3’ UTRs per gene in genes with multiple UTRs.

Figure 6. (A) Classification of protein-coding genes based on their novelty and types of encoded transcripts. (B) Number of expressed tissues for bifunctional genes. Dots have been color coded based on their density. (C) Location of different transcript biotypes on bifunctional genes. (D) Functional enrichment analysis of genes that remained bifunctional in all of their expressed tissues.

Figure 7. Support of predicted genes using data from different technologies and datasets

Figure 8. Functional enrichment analysis of non-coding genes in fetal tissues that were switched to protein coding with only coding transcripts in their matched adult tissue.

Figure 9- (A) Correlation between testis genes encoded protein with a signal peptide that were close to the “percentage of normal sperm” QTL and pituitary expressed genes closest to this trait (reference correlations). (B) Distribution of p-values resulting from a right-sided t-test between reference correlation coefficients and correlation coefficients derived from random chance (see methods for details).

Figure 10- (A) Distribution of the number of expressed annotated and un-annotated miRNAs across tissues. (B) Expression of annotated and un-annotated miRNAs across their expressed tissues. (C) Number of expressed tissues for annotated and un-annotated miRNAs.

Figure 11- Support of annotated (A) and un-annotated (B) miRNAs using different histone marks and CTCF-DNA binding data.

Tables

Table 1. Summary of expressed transcripts/genes

	Annotation[1]		
Feature	Current project	Ensembl	NCBI
		(Release 2021-03)	(Release 106)
Number of genes	35,150 (21,193)	27,607 (21,880)	35,143 (21,355)
Number of transcripts	171,985 (85,658)	43,984 (37,538)	83,195 (47,280)
Number of spliced transcripts	130,531	37,299	73,423
Number of transcripts per gene	4.9	1.5	2.3
Median number of 5' UTRs per gene	2	1	1
Median number of 3' UTRs per gene	1	1	1

[1] Numbers in parentheses indicate the number of protein-coding genes/transcripts.

Table 2. Protein/peptide homology of transcripts with coding potential

Transcript biotype	Number of transcripts	Transcripts with protein/peptide homology to other species[1]
Protein-coding transcripts	85,658	73,268 (86%)
sncRNAs and lncRNAs that encode short peptides[2]	48,425	4,054 (8%)

[1] Number in parentheses indicates the percentage of each transcript biotype.

[2] Open reading frame of 9 to 43 amino acids

Table 3. Sequence homology of non-coding transcripts

Transcript biotype	Number of transcripts	Transcripts with sequence homology to ncRNAs in other species[1]
Long non-coding RNAs	48,661	23,707 (49%)
Small non-coding RNAs	526	194 (37%)
Non-stop decay RNAs	4,359	1,551 (35%)
Nonsense-mediated decay RNAs	32,781	18,195 (55%)

[1] Number in parentheses indicates the percentage of each transcript biotype.

Table 4. Sequence homology of different types of lncRNAs

lncRNA biotype	Number of transcripts	Transcripts with sequence homology to ncRNAs in other species[1]
antisense lncRNAs	29,987	13,793 (46%)
sense-intronic lncRNAs	1,694	1,029 (60%)
intragenic lncRNAs	5,569	2,314 (41%)
intergenic lncRNAs	11,841	5,820 (49%)

[1] Number in parentheses indicates the percentage of each transcript biotype.

Table 5. Gene border extensions in current ARS-UCD1.2 genome annotations by *de novo* assembled transcriptome from short-read RNA-seq data

Annotation	Type of gene extension	Number of genes	Median extension (nucleotides)
Ensembl	5' extension only	1,848	128
(Release 2021-03)	3' extension only	5,701	422
	Both ends extended	4,874	122, 5' 439, 3'
NCBI	5' extension only	2,214	80
(Release 106)	3' extension only	5,496	126
	Both ends extended	3,613	66, 5' 210, 3'

Table 6. Median number of reads mapped to the extended region of annotated genes[1]

Annotation	5’ end extension	3’ end extension	Both ends extension
Ensembl (release 2021-03)	92 (1.10)	220 (1.24)	1,766 (8.90)
NCBI (release 106)	72 (1.05)	95 (1.10)	2,009 (9.05)

[1] Numbers in parentheses indicate the median fold change in expression level resulting from gene extensions.

Table 7. Comparison of different gene builds based on gene biotypes

Species	Gene build	Protein-coding genes	lncRNA genes	miRNA genes	Other types of small non-coding genes[1]	Pseudo-genes
Bovine (ARS-UCD1.2)	Ensembl (Release 2021-03)	21,880	1,480	951	2,209	492
	NCBI (Release 106)	21,039	5,179	797	3,249	4,569
	Current project[2]	21,193 (18,096)	10,789 (2,847)	2,007 (973)	139 (0)	3,029 (1,509)
Human (GRCh38.104)	Ensembl (release 2021-03)	20,442	16,876	1,877	2,930	15,266

[1] Small nucleolar RNAs, small non-coding RNAs, small Cajal body specific RNAs, small conditional RNAs, and tRNAs

[2] Numbers in parentheses indicate the number of un-annotated RNAs in each biotype.

Table 8. Summary of error-corrected, FLNC Iso-Seq reads and their matched RNA-seq reads

Tissue	Error-corrected FLNC Iso-Seq reads[1]	Median error rate in error-corrected FLNC Iso-Seq reads	Normalized RNA-seq reads used for error correction[2]
Thalamus	664,900 (90%)	0.21%	32,452,612
Testes	711,821 (86%)	1.43%	31,939,024
Liver	1,064,146 (84%)	1.84%	13,657,156
Medulla	380,531 (86%)	0.43%	48,256,918
Subcutaneous fat	215,759 (93%)	0.45%	42,043,313
Cerebral cortex	440,797 (87%)	1.01%	21,285,864
Jejunum	604,436 (90%)	2.331%	34,457,447

[1] Number in parentheses indicates mapping rate (90% coverage and 95% identity).

[2] In silico normalized using insilico_read_normalization.pl from Trinity (version 2.6.6) with the following settings: --max_cov 50 --max_pct_stdev 100 --single

Supplemental files

Supplemental file 1: Fig. S1 Distribution of the number of RNA-seq reads across tissues. **Fig. S2** (A) Comparison of tissues based on number of transcript biotypes and (B) percentage of transcript biotypes. (C) Comparison of transcript biotypes based on their number of expressed tissues and (D) their expression level across expressed tissues. **Fig. S3** (A) Relation between the number of input reads and the number of transcript biotypes (B) Comparison of expression level between different transcript biotypes. **Fig. S4** Tissue similarities (A) and clustering (B) based on the percentage of protein-coding transcripts shared between pairs of tissues. **Fig. S5** Tissue similarities (A) and clustering (B) based on the percentage of non-coding transcripts shared between pairs of tissues. **Fig. S6** Comparison of annotated and un-annotated transcripts based on their expression (A) and number of expressed tissues (B). **Fig. S7** Comparison of annotated and un-annotated protein-coding transcripts based on the length (A) and GC content (B) of their 5' UTR, CDS, and 3' UTR. **Fig. S8** (A) Comparison of tissues based on number of gene biotypes and (B) percentage of gene biotypes. **Fig. S9** Relation between the number of input reads and the number of gene biotypes**. Fig. S10** Comparison of annotated and un-annotated genes based on their expression (A) and number of expressed tissues (B). **Fig. S11** Functional enrichment analysis of the top five percent of genes with the highest number of UTRs. **Fig. S12** Similarity of tissues based on the number of non-coding genes in their fetal samples that switched to protein-coding genes with only coding transcripts in their adult samples. **Fig. S13** (A) Distribution of genes that transcribed PATs, based on their number of expressed tissues, percentage of genes' transcripts that are PATs and percentage of genes' expressed tissues in which PATs were transcribed. (B) Comparison of genes that transcribed PATs with other gene biotypes. **Fig. S14** (A) Homology analysis of protein-coding genes. (B) Homology analysis of non-coding genes. (C) Detection of orphan genes based on homology classification of cattle-specific protein-coding genes and non-coding genes. **Fig. S15** Comparison of the expression level of homologous and orphan genes across (A) and within (B) their expressed tissues. (C) Comparison of homologous and orphan genes based on the number of expressed tissues. **Fig. S16** Comparison of different gene biotypes based on the expression (A) and the number of expressed tissues (B). **Fig. S17** Comparison of different pseudogene-derived lncRNAs and non-pseudogene derived lncRNAs based on the expression level (A) and the number of expressed tissues (B)**. Fig. S18** Tissue similarities (A) and clustering (B) based on the percentage of protein-coding genes shared between pairs of tissues. **Fig. S19** Tissue similarities (A) and clustering (B) based on the percentage of non-coding genes shared between pairs of tissues. **Fig. S20** (A) Different types of alternative splicing events. (B) Comparison of bovine genome builds based on the number of transcripts that showed any type of alternative splicing (AS) events**. Fig. S21** Comparison of tissues based on the number (A) and the percentage (B) of transcripts that showed different types of alternative splicing events. Comparison of tissues based on the number (C) and the percentage (D) of alternative splicing events**. Fig. S22** (A) Comparison of tissues based on the percentage of transcripts that showed any type of alternative splicing events, spliced transcripts from single-transcript genes, and unspliced transcripts and (B) the relation between the number of input reads and the number of these transcripts across tissues. **Fig. S23** Comparison of transcripts that showed different types of alternative splicing events based on (A) the expression level in the expressed tissues and (B) the number of expressed tissues. **Fig. S24** Transcript biotype switching due to alternative splicing events**. Fig. S25**

Comparison of tissues based on the number of alternative splicing events per alternatively spliced gene**. Fig. S26** (A) Distribution of the number of alternative splicing events per alternatively spliced gene. The 5% quantile is shown using a dashed red line. (B) Functional enrichment analysis of the top five percent of genes with the highest number of alternative splicing events. **Fig. S27** Comparison of the alternative splicing rate between adult and fetal tissues. **Fig. S28** (A) Distribution of gene's number of expressed tissues. Tissue-specific gene biotypes are shown in the pie chart. (B) Distribution of transcript's number of expressed tissues. Tissue-specific transcript biotypes are shown in the pie chart. (C) Comparison of tissues based on the number of tissue-specific genes and transcripts. (D) Comparison of the expression level of tissue-specific genes and transcripts versus their non-tissue-specific counterparts. **Fig. S29** Relationship between tissue specificity and alternative splicing events**. Fig. S30** Relationship between tissue specificity index and the number of multi-tissue expressed genes (A) and transcripts (B). Distribution of tissue specificity indexes in multi-tissue expressed genes (C) and transcripts (D). The 5% quantile is shown using dashed red lines. (E) Functional enrichment analysis of the top five percent of multi-tissue expressed genes with the highest tissue specificity indexes. **Fig. S31** Distribution of QTLs located outside gene borders in relation to the closest expressed gene**. Fig. S32** (A) Distribution of correlation coefficients between *SPACA5* gene expression and pituitary expressed genes closest to "percentage of normal sperm" QTLs. Dashed lines show the minimum significant positive and negative correlation (p-value <0.05). (B) Expression atlas of *SPACA5* gene in human tissues from The Human Protein Atlas [86]. **Fig. S33** (A) Correlation between pituitary genes with signal peptides that were close to the "percentage of normal sperm" QTL and testis expressed genes closest to this trait's QTL (reference correlations). (B) Distribution of p-values resulting from right-sided t-test between reference correlation coefficients and correlation coefficients derived from random chance (see methods for details)**. Fig. S34** Tissue similarities (A) and clustering (B) based on the percentage of miRNAs shared between pairs of tissues. **Fig. S35** Clustering of tissues based on protein-coding genes (A), protein-coding transcripts (B), non-coding genes (C), non-coding transcripts (D), and miRNAs (E). (F) Comparison of tissue dendrograms based on the correlation between their Cophenetic distances. **Fig. S36** (A) Distribution of the number of expressed tissues for annotated and un-annotated miRNAs. Classification of miRNAs as annotated, or un-annotated is presented in the pie chart. (B) Comparison of tissues based on their number of tissue-specific miRNAs. (C) Expression of annotated and un-annotated miRNAs in their expressed tissues. (D) Distribution of multi-tissue expressed miRNAs' tissue specificity indexes. (E) Relationship between tissue specificity index and number of expressed tissues in multi-tissue expressed miRNAs. Dots have been color coded based on their density. **Fig. S37** Distribution of the number of expressed genes (A), transcripts (B), and miRNAs (C) across tissues. **Fig. S38** Distribution of the number of annotated and un-annotated genes (A), transcripts (B), and miRNAs (C) across tissues. **Fig. S39** Graphical representation of the method used to construct the tissue similarity network.

Supplemental file 2: Summary of RNA-seq and miRNA-seq reads

Supplemental file 3: Detailed description of the number of transcripts, genes, and miRNAs expressed in each tissue

Supplemental file 4: List of transcripts and genes expressed in each tissue and their expression values (RPKM)

Supplemental file 5: Transcript biotype enrichment analysis in adult and fetal tissues

Supplemental file 6: Functional enrichment analysis of the top five percent of genes with the highest number of UTRs

Additional file7: Functional enrichment analysis of genes that remained bifunctional in all their expressed tissues

Additional file8: Functional enrichment analysis of non-coding genes in fetal tissues that were switched to protein coding with only coding transcripts in their matched adult tissue

Additional file9: Functional enrichment analysis of protein-coding genes that transcribed PATs as their main transcripts (PATs comprised >50% of their transcripts) in all their expressed tissues

Supplemental file 10: Gene biotype enrichment analysis in adult and fetal tissues

Supplemental file 11: Functional enrichment analysis of the top five percent of genes with the highest number of alternative splicing events

Additional file12: List of tissue-specific genes and transcripts

Additional file13: Genes and transcripts tissue specificity indexes

Additional file14: Functional enrichment analysis of the top five percent of multi-tissue expressed genes with the highest tissue specificity indexes

Additional file15: List of QTL's closest expressed genes in each tissue

Additional file16: Trait enrichment analysis of testis-specific genes

Additional file16: Pituitary expressed genes closest to "percentage of normal sperm" QTLs that showed positive significant correlation with SPACA5 gene in testis

Additional file18: List of expressed genes closest to "percentage of normal sperm" QTLs that were involved in testis-pituitary tissue axis and their functional enrichment analysis results

Additional file19: List of genes expressed closest to "percentage of normal sperm" QTLs that were involved in pituitary-testis tissue axis and their functional enrichment analysis results

Additional file20: Similarity of traits based on the integration of the assembled bovine transcriptome with publicly available QTLs

Additional file21: List of miRNAs expressed in each tissue and their expression values

Additional file22: List of tissues related to different omics datasets used in the experiment

Abbreviations

A3Es: Alternative 3' splice site Exons; A5Es: Alternative 5' splice site Exons; AFEs: Alternative First Exon; ALEs: Alternative Last Exon; AS: Alternative Splicing; ATAC-seq: Assay for Transposase-Accessible Chromatin using sequencing; bp: base pair; BP: Biological Process; CDS: coding sequence; ChIP-seq: Chromatin Immunoprecipitation Sequencing; CPM: Counts Per

Million; CTCF: CCCTC-binding factor; DMEM: Dulbecco's Modified Eagle Medium; FLNC: Full-Length, Non-Chimeric; GO: Gene Ontology; GOA: Gene Ontology Annotation database; GWAS: Genome-Wide Association Studies; H3K27ac: N-terminal acetylation of lysine 27 on histone H3; H3K4me1: tri-methylation of lysine 4 on histone H1; H3K4me3: tri-methylation of lysine 4 on histone H3; IACUC: Institutional Animal Care and Use Committee; LD: Longissimus Dorsi; lncRNAs: long non-coding RNAs; miRNA: microRNAs; MXEs: Mutually Exclusive Exons; NCBI: National Center for Biotechnology Information; ncRNAs: non-coding RNAs; NMD: Nonsense-Mediated Decay; NSD: Non-Stop Decay; ONT-seq: Oxford Nanopore Technologies sequencing; ORFs: Open Reading Frames; PacBio Iso-Seq: Pacific Biosciences single-molecule long-read isoform sequencing; PAT: Potentially Aberrant Transcript; poly(A): Polyadenylation; PTBP1: polypyrimidine tract binding protein 1; QTL: Quantitative Trait Loci; RAMPAGE: RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression; Ribo-seq: Ribosome footprinting followed by Sequencing; RIEs: Retained Intron Exons; RNA-seq: Illumina high-throughput RNA sequencing; RPKM: Reads Per Kilobase of Transcript per Million reads mapped; RPM: Reads Per Million; SEs: Skipped Exons; sncRNAs: small non-coding RNAs; SNP: Single Nucleotide Polymorphism; tpg: transcripts per annotated gene; TSI: Tissue Specificity Index; TSS: Transcript Start Sites; TTS: Transcript Terminal Sites; UCD: University of California, Davis; USEs: Unique Splice Site Exons; UTR: untranslated region; WTTS-seq: Whole Transcriptome Termini Site Sequencing.

Data availability

RNA-seq and miRNA-seq, ATAC-seq, and WTTS-seq datasets generated in this study are submitted to the ArrayExpress database (https://www.ebi.ac.uk/biostudies/arrayexpress) under accession numbers E-MTAB-11699, E-MTAB-11815, and E-MTAB-12052, respectively. The constructed bovine trait similarity network is publicly available through the Animal Genome database (https://www.animalgenome.org/host/reecylab/a). The constructed cattle transcriptome and related sequences are publicly available in the Open Science Framework database (https://osf.io/jze72/?view_only=d2dd1badf37e4bafae1e12731a0cc40d). Custom code used is available at https://github.com/hamidbeiki/Cattle-Genome.

Ethics approval and consent to participate

Procedures for tissue collection followed the Animal Care and Use protocol (#18464) approved by the Institutional Animal Care and Use Committee (IACUC), University of California, Davis (UCD).

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

Funding

This study was supported by Agriculture and Food Research Initiative Competitive Grant no. 2018-67015-27500 (H.Z., P.R. etc.) and sample collection was supported by no. 2015-67015-22940 (H.Z. and P.R.) from the USDA National Institute of Food and Agriculture.

Acknowledgments

We are grateful to Nathan Weeks for helping with massive parallel computing of transcriptome assembly.

Authors' contributions

H.B., B.M.M., H.J., H.Z., M.R., P.J.R., S.M., T.P.L.S., W.L., Z.J., and J.M.R. conceived and designed the project; C.K., W.M., and W.L. generated RNA-seq and miRNA-seq data; D.K., G.B., J.T., and K.D. participated in tissue collection; R.H and H.J prepared cells; J.J.M., X.Z., X.H., and Z.J. generated W.T.T.S-seq data, X.X., P.J.R. and H.J generated ChIP-seq data; M.R.J. generated ATAC-seq data; T.P.L.S. generated PacBio Iso-seq data; G.R. and S.C. conducted sequencing of RNA-seg, miRNA-seq, ChIP-seq, and ATAC-seq data; H.B. conducted bioinformatics data analysis and drafted the manuscript, which was edited by C.A.P., B.M.M., H.J., H.Z., J.E.K., M.R., P.J.R., S.M., T.P.L.S., W.L., Z.J. and J.M.R.; Z.H. created the web-based database for the trait similarity network; all authors read and approved the final manuscript.

Endnotes

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

The results reported here were made possible with resources provided by the USDA shared computing cluster (Ceres) as part of the ARS SCINet initiative.

References

1. Roth JA and Tuggle CK. Livestock models in translational medicine. ILAR J. 2015;56 1:1-6. doi:10.1093/ilar/ilv011.
2. Beiki H, Liu H, Huang J, Manchanda N, Nonneman D, Smith TPL, et al. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. BMC Genomics. 2019;20 1:344. doi:10.1186/s12864-019-5709-y.
3. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106 23:9362-7. doi:10.1073/pnas.0903103106.
4. Jereb S, Hwang HW, Van Otterloo E, Govek EE, Fak JJ, Yuan Y, et al. Differential 3' Processing of Specific Transcripts Expands Regulatory and Protein Diversity Across Neuronal Cell Types. Elife. 2018;7 doi:10.7554/eLife.34042.
5. Schurch NJ, Cole C, Sherstnev A, Song J, Duc C, Storey KG, et al. Improved annotation of 3' untranslated regions and complex loci by combination of strand-specific direct RNA sequencing, RNA-Seq and ESTs. PLoS One. 2014;9 4:e94270. doi:10.1371/journal.pone.0094270.
6. Ambros V. The functions of animal microRNAs. Nature. 2004;431 7006:350-5. doi:10.1038/nature02871.
7. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116 2:281-97. doi:10.1016/s0092-8674(04)00045-5.
8. Yates LA, Norbury CJ and Gilbert RJ. The long and short of microRNA. Cell. 2013;153 3:516-9. doi:10.1016/j.cell.2013.04.003.
9. Halstead MM, Islas-Trejo A, Goszczynski DE, Medrano JF, Zhou H and Ross PJ. Large-Scale Multiplexing Permits Full-Length Transcriptome Annotation of 32 Bovine Tissues From a Single Nanopore Flow Cell. Front Genet. 2021;12:664260. doi:10.3389/fgene.2021.664260.
10. Goszczynski DE, Halstead MM, Islas-Trejo AD, Zhou H and Ross PJ. Transcription initiation mapping in 31 bovine tissues reveals complex promoter activity, pervasive transcription, and tissue-specific promoter usage. Genome Res. 2021;31 4:732-44. doi:10.1101/gr.267336.120.
11. Kozomara A, Birgaoanu M and Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47 D1:D155-D62. doi:10.1093/nar/gky1141.

12. Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, Suresh U, et al. Before It Gets Started: Regulating Translation at the 5' UTR. Comp Funct Genomics. 2012;2012:475731. doi:10.1155/2012/475731.

13. Gerber S, Schratt G and Germain PL. Streamlining differential exon and 3' UTR usage with diffUTR. BMC Bioinformatics. 2021;22 1:189. doi:10.1186/s12859-021-04114-7.

14. Andrews SJ and Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 2014;15 3:193-204. doi:10.1038/nrg3520.

15. Kumari P and Sampath K. cncRNAs: Bi-functional RNAs with protein coding and non-coding functions. Semin Cell Dev Biol. 2015;47-48:40-51. doi:10.1016/j.semcdb.2015.10.024.

16. Nam JW, Choi SW and You BH. Incredible RNA: Dual Functions of Coding and Noncoding. Mol Cells. 2016;39 5:367-74. doi:10.14348/molcells.2016.0039.

17. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J and Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013;14 7:R70. doi:10.1186/gb-2013-14-7-r70.

18. Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, et al. MBASED: allele-specific expression detection in cancer tissues and cell lines. Genome Biol. 2014;15 8:405. doi:10.1186/s13059-014-0405-3.

19. Hubé F, Velasco G, Rollin J, Furling D and Francastel C. Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation. Nucleic Acids Res. 2011;39 2:513-25. doi:10.1093/nar/gkq833.

20. Kurosaki T, Popp MW and Maquat LE. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nat Rev Mol Cell Biol. 2019;20 7:406-20. doi:10.1038/s41580-019-0126-2.

21. Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA and Smith CW. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell. 2004;13 1:91-100. doi:10.1016/s1097-2765(03)00502-1.

22. Nickless A, Bailis JM and You Z. Control of gene expression through the nonsense-mediated RNA decay pathway. Cell Biosci. 2017;7:26. doi:10.1186/s13578-017-0153-7.

23. Supek F, Lehner B and Lindeboom RGH. To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases. Trends Genet. 2021;37 7:657-68. doi:10.1016/j.tig.2020.11.002.

24. Mitrovich QM and Anderson P. mRNA surveillance of expressed pseudogenes in C. elegans. Curr Biol. 2005;15 10:963-7. doi:10.1016/j.cub.2005.04.055.

25. Colombo M, Karousis ED, Bourquin J, Bruggmann R and Mühlemann O. Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. RNA. 2017;23 2:189-201. doi:10.1261/rna.059055.116.

26. Milligan MJ and Lipovich L. Pseudogene-derived lncRNAs: emerging regulators of gene expression. Front Genet. 2014;5:476. doi:10.3389/fgene.2014.00476.

27. Stewart GL, Enfield KSS, Sage AP, Martinez VD, Minatel BC, Pewarchuk ME, et al. Aberrant Expression of Pseudogene-Derived lncRNAs as an Alternative Mechanism of Cancer Gene Regulation in Lung Adenocarcinoma. Front Genet. 2019;10:138. doi:10.3389/fgene.2019.00138.

28. Lou W, Ding B and Fu P. Pseudogene-Derived lncRNAs and Their miRNA Sponging Mechanism in Human Cancer. Front Cell Dev Biol. 2020;8:85. doi:10.3389/fcell.2020.00085.

29. Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell. 2015;160 4:595-606. doi:10.1016/j.cell.2015.01.009.

30. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, et al. Extensive identification and analysis of conserved small ORFs in animals. Genome Biol. 2015;16:179. doi:10.1186/s13059-015-0742-x.

31. Olexiouk V, Crappé J, Verbruggen S, Verhegen K, Martens L and Menschaert G. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2016;44 D1:D324-9. doi:10.1093/nar/gkv1175.

32. Li J and Liu C. Coding or Noncoding, the Converging Concepts of RNAs. Front Genet. 2019;10:496. doi:10.3389/fgene.2019.00496.

33. Wei L-H and Guo JU. Coding functions of “noncoding” RNAs. Science. 2020;367 6482:1074-5. doi:10.1126/science.aba6117.

34. Sammeth M, Foissac S and Guigó R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol. 2008;4 8:e1000147. doi:10.1371/journal.pcbi.1000147.

35. Mazin PV, Khaitovich P, Cardoso-Moreira M and Kaessmann H. Alternative splicing during mammalian organ development. Nature Genetics. 2021;53 6:925-34. doi:10.1038/s41588-021-00851-w.

36. Wu Z, Yang KK, Liszka MJ, Lee A, Batzilla A, Wernick D, et al. Signal Peptides Generated by Attention-Based Neural Networks. ACS Synth Biol. 2020;9 8:2154-61. doi:10.1021/acssynbio.0c00219.

37. Chen J and Chen ZJ. Regulation of NF-κB by ubiquitination. Curr Opin Immunol. 2013;25 1:4-12. doi:10.1016/j.coi.2012.12.005.

38. Karalis KP, Venihaki M, Zhao J, van Vlerken LE and Chandras C. NF-kappaB participates in the corticotropin-releasing, hormone-induced regulation of the pituitary proopiomelanocortin gene. J Biol Chem. 2004;279 12:10837-40. doi:10.1074/jbc.M313063200.

39. O'Shaughnessy PJ, Fleming LM, Jackson G, Hochgeschwender U, Reed P and Baker PJ. Adrenocorticotropic hormone directly stimulates testosterone production by the fetal and neonatal mouse testis. Endocrinology. 2003;144 8:3279-84. doi:10.1210/en.2003-0277.

40. Richburg JH, Myers JL and Bratton SB. The role of E3 ligases in the ubiquitin-dependent regulation of spermatogenesis. Semin Cell Dev Biol. 2014;30:27-35. doi:10.1016/j.semcdb.2014.03.001.

41. Kumar S, Lee HJ, Park HS and Lee K. Testis-Specific GTPase (TSG): An oligomeric protein. BMC Genomics. 2016;17 1:792. doi:10.1186/s12864-016-3145-9.

42. Rajala-Schultz PJ, Gröhn YT, McCulloch CE and Guard CL. Effects of clinical mastitis on milk yield in dairy cows. J Dairy Sci. 1999;82 6:1213-20. doi:10.3168/jds.S0022-0302(99)75344-0.

43. Martí De Olives A, Díaz JR, Molina MP and Peris C. Quantification of milk yield and composition changes as affected by subclinical mastitis during the current lactation in sheep. J Dairy Sci. 2013;96 12:7698-708. doi:10.3168/jds.2013-6998.

44. Halasa T and Kirkeby C. Differential Somatic Cell Count: Value for Udder Health Management. Front Vet Sci. 2020;7:609055. doi:10.3389/fvets.2020.609055.

45. Remnant J, Green MJ, Huxley J, Hirst-Beecham J, Jones R, Roberts G, et al. Association of lameness and mastitis with return-to-service oestrus detection in the dairy cow. Vet Rec. 2019;185 14:442. doi:10.1136/vr.105535.

46. Miles AM, McArt JAA, Leal Yepes FA, Stambuk CR, Virkler PD and Huson HJ. Udder and teat conformational risk factors for elevated somatic cell count and clinical mastitis in New York Holsteins. Prev Vet Med. 2019;163:7-13. doi:10.1016/j.prevetmed.2018.12.010.

47. Lima FS, Silvestre FT, Peñagaricano F and Thatcher WW. Early genomic prediction of daughter pregnancy rate is associated with improved reproductive performance in Holstein dairy cows. J Dairy Sci. 2020;103 4:3312-24. doi:10.3168/jds.2019-17488.

48. Hertl JA, Schukken YH, Tauer LW, Welcome FL and Gröhn YT. Does clinical mastitis in the first 100 days of lactation 1 predict increased mastitis occurrence and shorter herd life in dairy cows? J Dairy Sci. 2018;101 3:2309-23. doi:10.3168/jds.2017-12615.

49. Kaniyamattam K, De Vries A, Tauer LW and Gröhn YT. Economics of reducing antibiotic usage for clinical mastitis and metritis through genomic selection. J Dairy Sci. 2020;103 1:473-91. doi:10.3168/jds.2018-15817.

50. Green TC, Jago JG, Macdonald KA and Waghorn GC. Relationships between residual feed intake, average daily gain, and feeding behavior in growing dairy heifers. J Dairy Sci. 2013;96 5:3098-107. doi:10.3168/jds.2012-6087.

51. Elolimy AA, Abdelmegeid MK, McCann JC, Shike DW and Loor JJ. Residual feed intake in beef cattle and its association with carcass traits, ruminal solid-fraction bacteria, and epithelium gene expression. J Anim Sci Biotechnol. 2018;9:67. doi:10.1186/s40104-018-0283-8.

52. Weber C, Hametner C, Tuchscherer A, Losand B, Kanitz E, Otten W, et al. Variation in fat mobilization during early lactation differently affects feed intake, body condition, and lipid and glucose metabolism in high-yielding dairy cows. J Dairy Sci. 2013;96 1:165-80. doi:10.3168/jds.2012-5574.

53. Yi Z, Li X, Luo W, Xu Z, Ji C, Zhang Y, et al. Feed conversion ratio, residual feed intake and cholecystokinin type A receptor gene polymorphisms are associated with feed intake and average daily gain in a Chinese local chicken population. J Anim Sci Biotechnol. 2018;9:50. doi:10.1186/s40104-018-0261-1.

54. Liu E and VandeHaar MJ. Relationship of residual feed intake and protein efficiency in lactating cows fed high- or low-protein diets. J Dairy Sci. 2020;103 4:3177-90. doi:10.3168/jds.2019-17567.

55. Clare M, Richard P, Kate K, Sinead W, Mark M and David K. Residual feed intake phenotype and gender affect the expression of key genes of the lipogenesis pathway in subcutaneous adipose tissue of beef cattle. J Anim Sci Biotechnol. 2018;9:68. doi:10.1186/s40104-018-0282-9.

56. Houlahan K, Schenkel FS, Hailemariam D, Lassen J, Kargo M, Cole JB, et al. Effects of Incorporating Dry Matter Intake and Residual Feed Intake into a Selection Index for Dairy Cattle Using Deterministic Modeling. Animals (Basel). 2021;11 4 doi:10.3390/ani11041157.

57. Tixier-Boichard M, Fabre S, Dhorne-Pollet S, Goubil A, Acloque H, Vincent-Naulleau S, et al. Tissue Resources for the Functional Annotation of Animal Genomes. Front Genet. 2021;12:666265. doi:10.3389/fgene.2021.666265.

58. Farr VC, Stelwagen K, Cate LR, Molenaar AJ, McFadden TB and Davis SR. An improved method for the routine biopsy of bovine mammary tissue. J Dairy Sci. 1996;79 4:543-9. doi:10.3168/jds.S0022-0302(96)76398-1.

59. Zhou X, Li R, Michal JJ, Wu XL, Liu Z, Zhao H, et al. Accurate Profiling of Gene Expression and Alternative Polyadenylation with Whole Transcriptome Termini Site Sequencing (WTTS-Seq). Genetics. 2016;203 2:683-97. doi:10.1534/genetics.116.188508.

60. Krueger F: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (2019).

61. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29 1:15-21. doi:10.1093/bioinformatics/bts635.

62. Liao Y, Smyth GK and Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30 7:923-30. doi:10.1093/bioinformatics/btt656.

63. Leek J, Johnson W, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, et al. *sva: Surrogate Variable Analysis* . R package version 3.30.0. 2021.

64. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29 7:644-52. doi:10.1038/nbt.1883.

65. Hass B: https://hpcgridrunner.github.io/. (2015).

66. Tange O: GNU Parallel. https://doi.org/10.5281/zenodo.1146014. (2018).

67. PacificBiosciences: https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/. (2018).

68. Pedersen BS and Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34 5:867-8. doi:10.1093/bioinformatics/btx699.

69. Hackl T, Hedrich R, Schultz J and Förster F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics. 2014;30 21:3004-11. doi:10.1093/bioinformatics/btu392.

70. Wang JR, Holt J, McMillan L and Jones CD. FMLRC: Hybrid long read error correction using an FM-index. BMC Bioinformatics. 2018;19 1:50. doi:10.1186/s12859-018-2051-3.

71. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003;31 1:28-33. doi:10.1093/nar/gkg033.

72. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. Database (Oxford). 2016;2016 doi:10.1093/database/baw093.

73. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45 W1:W12-W6. doi:10.1093/nar/gkx428.

74. Salmela L and Schröder J. Correcting errors in short reads by multiple alignments. Bioinformatics. 2011;27 11:1455-61. doi:10.1093/bioinformatics/btr170.

75. Hannon GJ: FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit. (2010).

76. Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun. 2021;12 1:1821. doi:10.1038/s41467-021-22100-8.

77. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25 8:1091-3. doi:10.1093/bioinformatics/btp101.

78. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, et al. The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res. 2015;43 Database issue:D1057-63. doi:10.1093/nar/gku1113.

79. Kim KI and van de Wiel MA. Effects of dependence in high-dimensional multiple testing problems. BMC Bioinformatics. 2008;9 1:114. doi:10.1186/1471-2105-9-114.

80. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19 1:40. doi:10.1186/s13059-018-1417-1.

81. Friedländer MR, Mackowiak SD, Li N, Chen W and Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40 1:37-52. doi:10.1093/nar/gkr688.

82. Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, et al. Distribution of miRNA expression across human tissues. Nucleic Acids Res. 2016;44 8:3865-77. doi:10.1093/nar/gkw116.

83. Hu ZL, Park CA and Reecy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic Acids Res. 2019;47 D1:D701-D10. doi:10.1093/nar/gky1084.

84. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13 11:2498-504. doi:10.1101/gr.1239303.

85. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology. 2019;37 4:420-3. doi:10.1038/s41587-019-0036-z.

86. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347 6220:1260419. doi:10.1126/science.1260419.

Figure 1

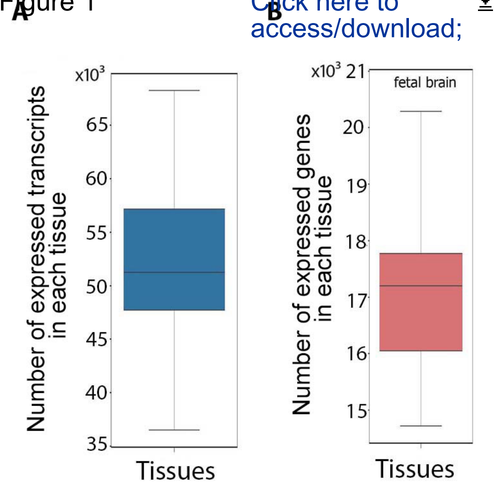

A
x10³
65
60
55
50
45
40
35
Number of expressed transcripts in each tissue
Tissues
B
x10³
21
20
19
18
17
16
15
fetal brain
Number of expressed genes in each tissue
Tissues

Figure 2

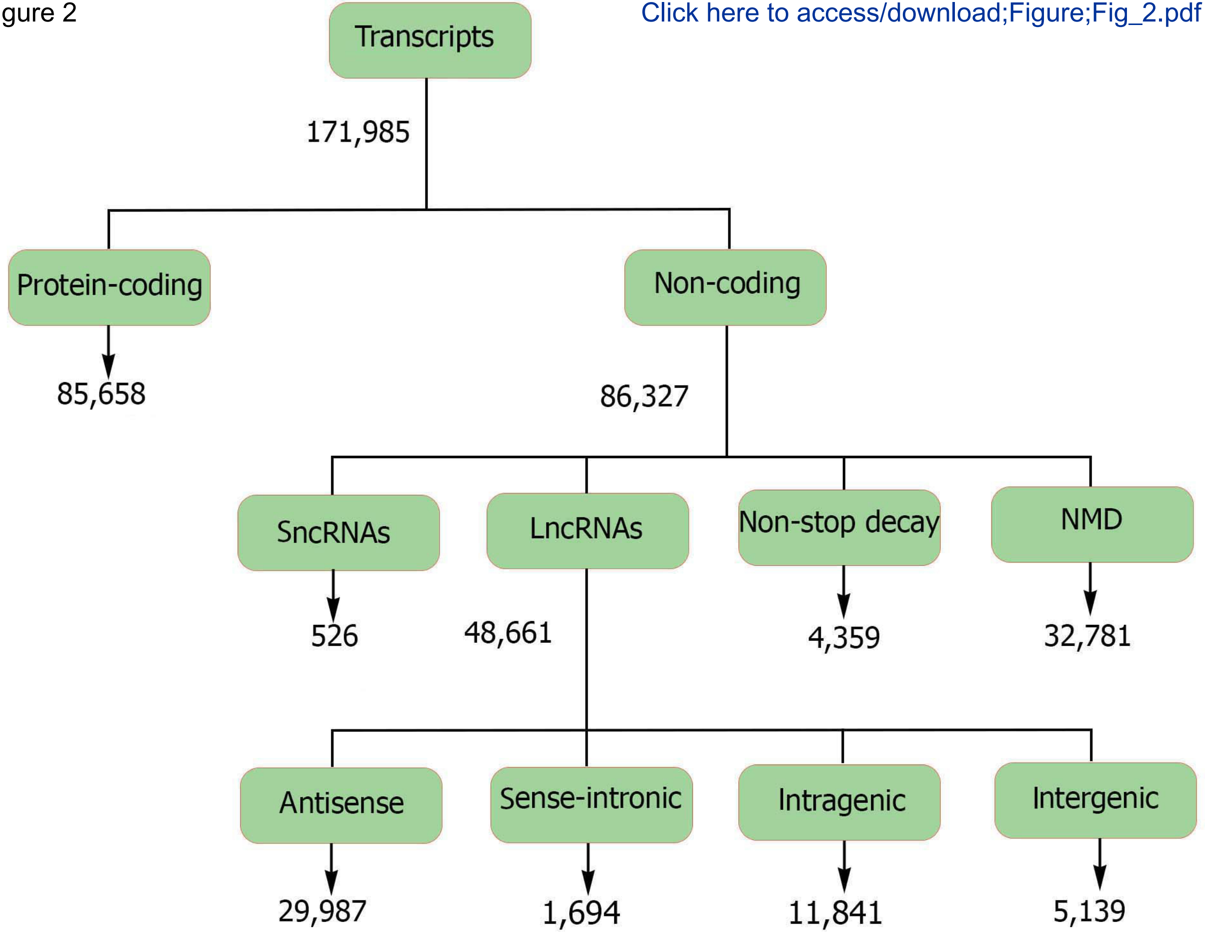

Transcripts
171,985
Protein-coding
85,658
Non-coding
86,327
SncRNAs
526
LncRNAs
48,661
Non-stop decay
4,359
NMD
32,781
Antisense
29,987
Sense-intronic
1,694
Intragenic
11,841
Intergenic
5,139

Figure 3

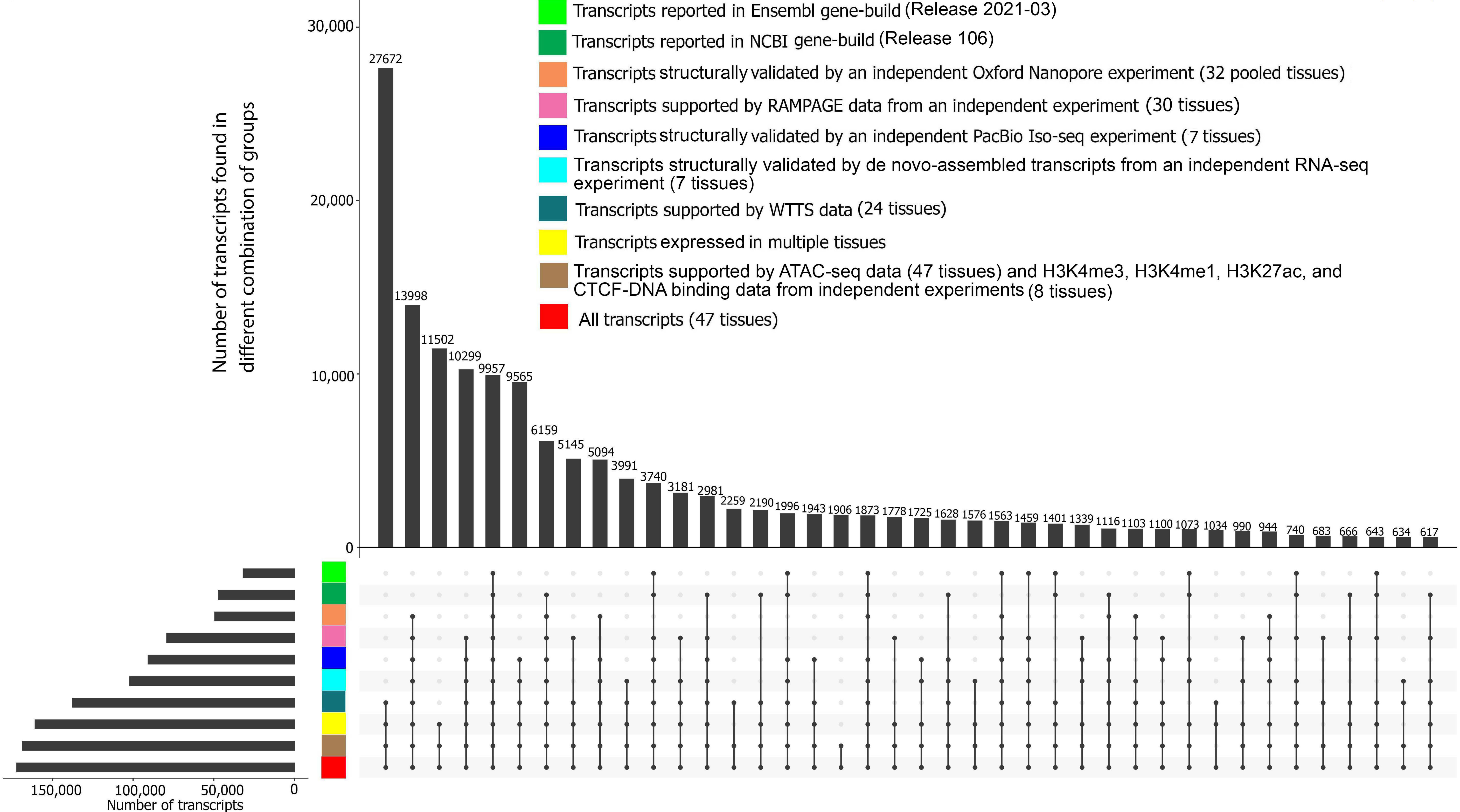

Transcripts reported in Ensembl gene-build (Release 2021-03)
Transcripts reported in NCBI gene-build (Release 106)
Transcripts structurally validated by an independent Oxford Nanopore experiment (32 pooled tissues)
Transcripts supported by RAMPAGE data from an independent experiment (30 tissues)
Transcripts structurally validated by an independent PacBio Iso-seq experiment (7 tissues)
Transcripts structurally validated by de novo-assembled transcripts from an independent RNA-seq experiment (7 tissues)
Transcripts supported by WTTS data (24 tissues)
Transcripts expressed in multiple tissues
Transcripts supported by ATAC-seq data (47 tissues) and H3K4me3, H3K4me1, H3K27ac, and CTCF-DNA binding data from independent experiments (8 tissues)
All transcripts (47 tissues)
Number of transcripts found in different combination of groups
30,000
20,000
10,000
0
27672
13998
11502
10299
9957
9565
6159
5145
5094
3991
3740
3181
2981
2259
2190
1996
1943
1906
1873
1778
1725
1628
1576
1563
1459
1401
1339
1116
1103
1100
1073
1034
990
944
740
683
666
643
634
617
150,000
100,000
50,000
0
Number of transcripts

Figure 4

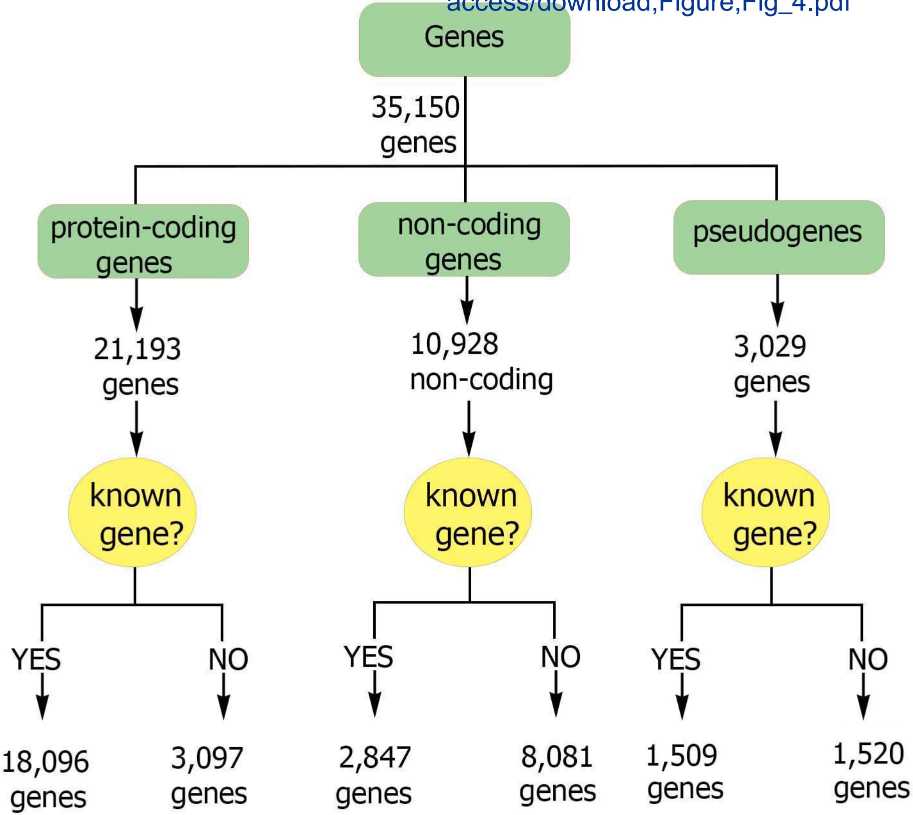

Genes
35,150 genes
protein-coding genes
non-coding genes
pseudogenes
21,193 genes
10,928 non-coding
3,029 genes
known gene?
known gene?
known gene?
YES
NO
YES
NO
YES
NO
18,096 genes
3,097 genes
2,847 genes
8,081 genes
1,509 genes
1,520 genes

Figure 5

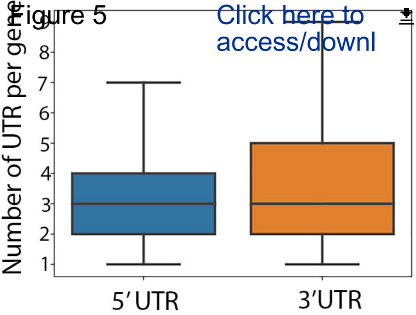


Figure 6

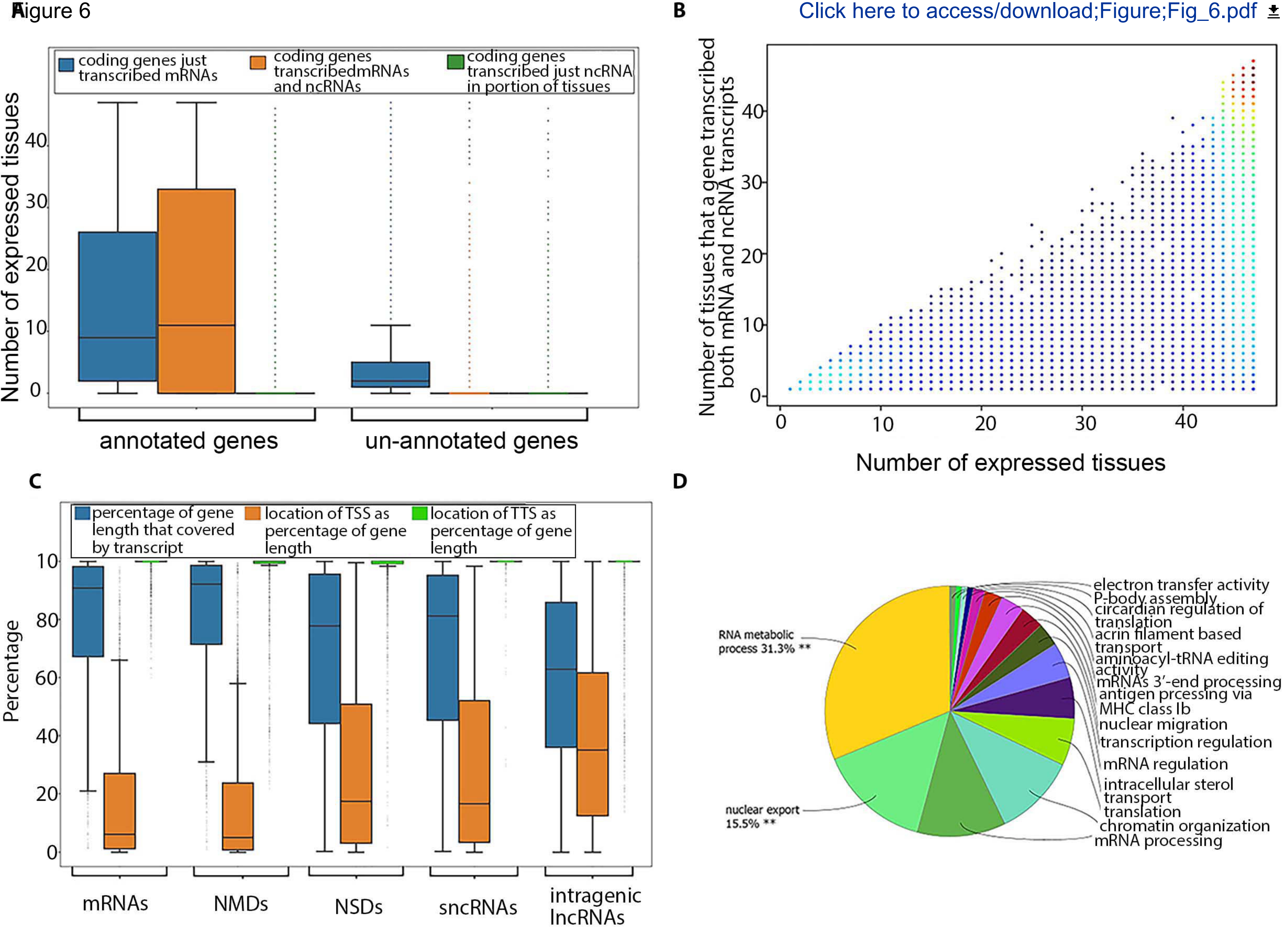

A
coding genes just transcribed mRNAs
coding genes transcribedmRNAs and ncRNAs
coding genes transcribed just ncRNA in portion of tissues
Number of expressed tissues
0
10
20
30
40
annotated genes
un-annotated genes
B
Number of tissues that a gene transcribed both mRNA and ncRNA transcripts
0
10
20
30
40
Number of expressed tissues
C
percentage of gene length that covered by transcript
location of TSS as percentage of gene length
location of TTS as percentage of gene length
Percentage
0
20
40
60
80
10
mRNAs
NMDs
NSDs
sncRNAs
intragenic lncRNAs
D
RNA metabolic process 31.3% **
nuclear export 15.5% **
electron transfer activity
P-body assembly
circardian regulation of translation
acrin filament based transport
aminoacyl-tRNA editing activity
mRNAs 3'-end processing
antigen prcessing via MHC class Ib
nuclear migration
transcription regulation
mRNA regulation
intracellular sterol transport
translation
chromatin organization
mRNA processing

Figure 7

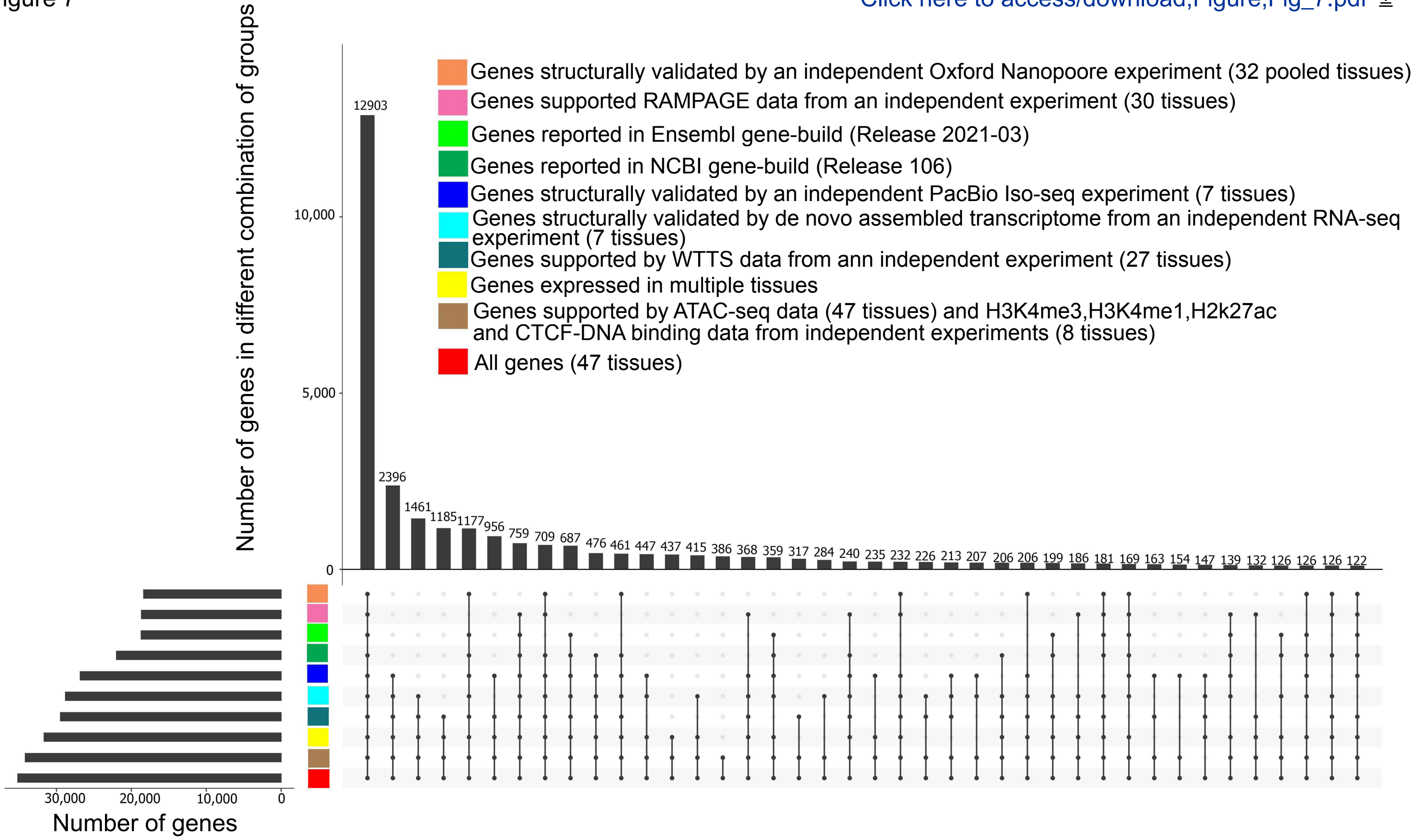

Genes structurally validated by an independent Oxford Nanopoore experiment (32 pooled tissues)
Genes supported RAMPAGE data from an independent experiment (30 tissues)
Genes reported in Ensembl gene-build (Release 2021-03)
Genes reported in NCBI gene-build (Release 106)
Genes structurally validated by an independent PacBio Iso-seq experiment (7 tissues)
Genes structurally validated by de novo assembled transcriptome from an independent RNA-seq experiment (7 tissues)
Genes supported by WTTS data from ann independent experiment (27 tissues)
Genes expressed in multiple tissues
Genes supported by ATAC-seq data (47 tissues) and H3K4me3,H3K4me1,H2k27ac and CTCF-DNA binding data from independent experiments (8 tissues)
All genes (47 tissues)
Number of genes in different combination of groups
10,000
5,000
0
12903
2396
1461
1185
1177
956
759
709
687
476
461
447
437
415
386
368
359
317
284
240
235
232
226
213
207
206
206
199
186
181
169
163
154
147
139
132
126
126
126
122
30,000
20,000
10,000
0
Number of genes

Figure 8

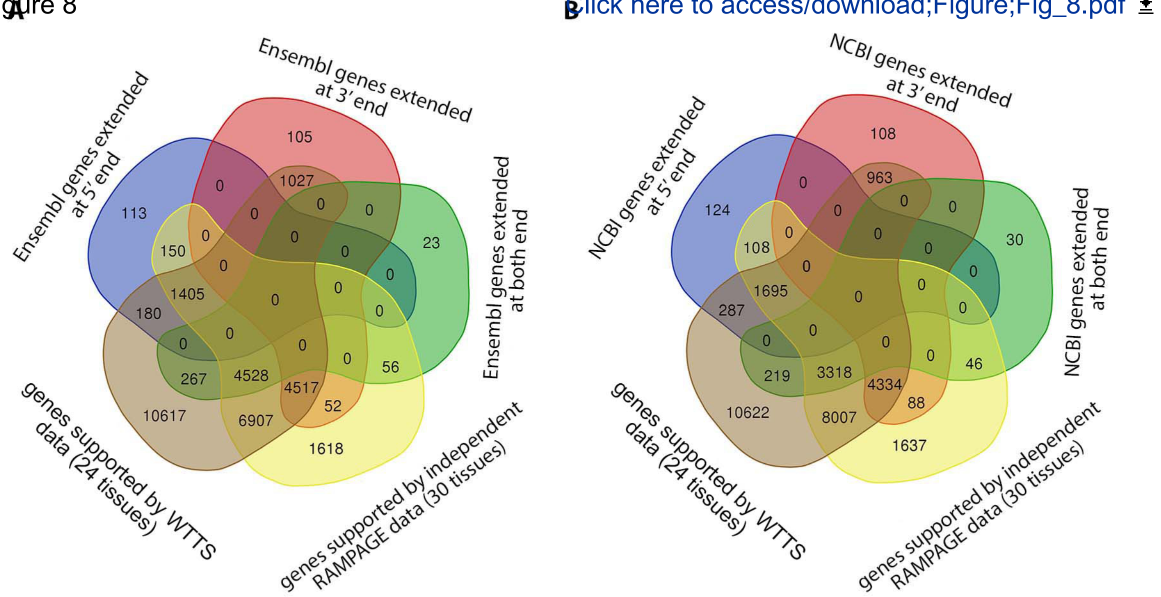

A
Ensembl genes extended at 5′ end
Ensembl genes extended at 3′ end
Ensembl genes extended at both end
genes supported by independent RAMPAGE data (30 tissues)
genes supported by WTTS data (24 tissues)
105
113
0
1027
0
0
0
23
150
0
0
0
0
0
0
1405
0
0
0
180
0
0
0
0
0
267
4528
4517
52
56
10617
6907
1618
B
NCBI genes extended at 5′ end
NCBI genes extended at 3′ end
NCBI genes extend at both end
genes supported by independent RAMPAGE data (30 tissues)
genes supported by WTTS data (24 tissues)
108
124
0
963
0
0
0
30
108
0
0
0
0
0
0
1695
0
0
0
287
0
0
0
0
0
219
3318
4334
88
46
10622
8007
1637

Figure 9

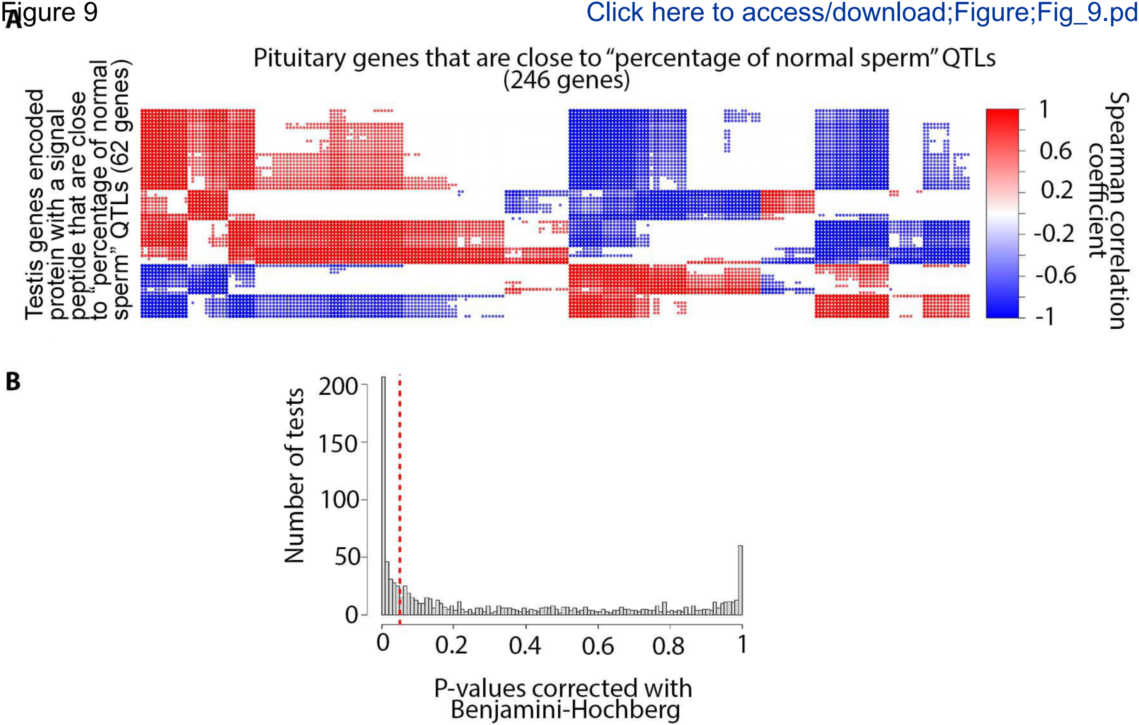
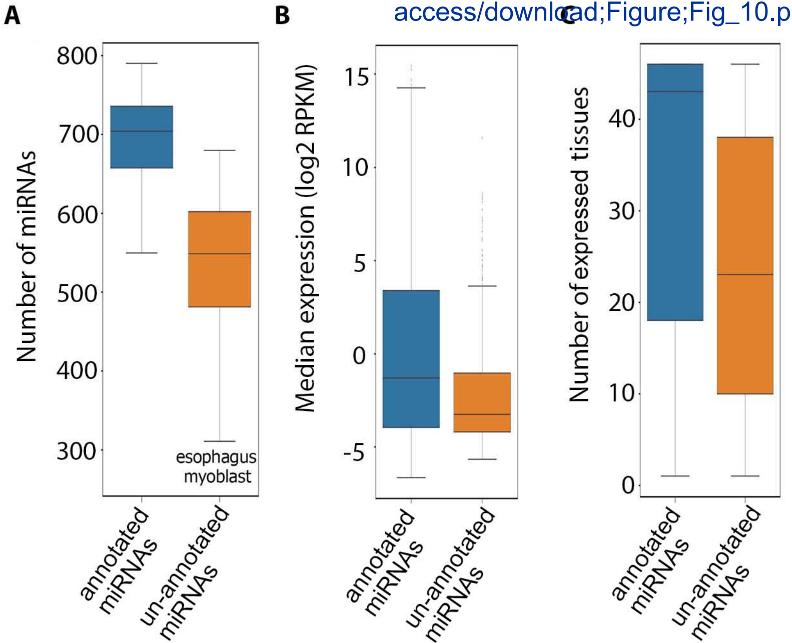
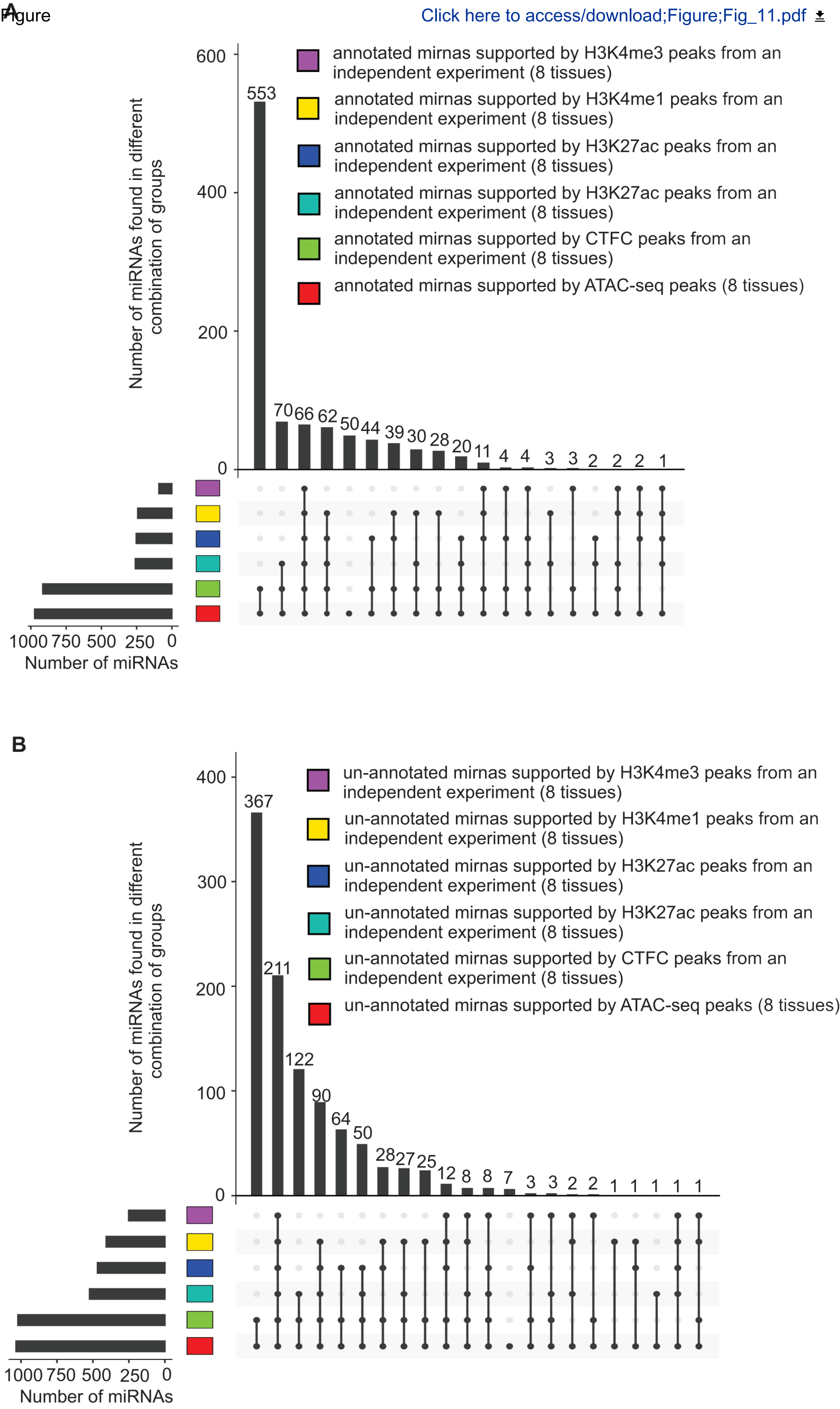

A
Pituitary genes that are close to "percentage of normal sperm" QTLs (246 genes)
Testis genes encoded protein with a signal peptide that are close to "percentage of normal sperm" QTLs (62 genes)
1
0.6
0.2
-0.2
-0.6
-1
Spearman correlation coefficient
B
Number of tests
200
150
100
50
0
0
0.2
0.4
0.6
0.8
1
P-values corrected with Benjamini-Hochberg

Figure 10

Click here to access/download;Figure;Fig_10.p




A
Number of miRNAs found in different combination of groups
600
400
200
0
553
70
66
62
50
44
39
30
28
20
11
4
4
3
3
2
2
2
1
annotated mirnas supported by H3K4me3 peaks from an independent experiment (8 tissues)
annotated mirnas supported by H3K4me1 peaks from an independent experiment (8 tissues)
annotated mirnas supported by H3K27ac peaks from an independent experiment (8 tissues)
annotated mirnas supported by H3K27ac peaks from an independent experiment (8 tissues)
annotated mirnas supported by CTFC peaks from an independent experiment (8 tissues)
annotated mirnas supported by ATAC-seq peaks (8 tissues)
1000 750 500 250 0
Number of miRNAs
B
Number of miRNAs found in different combination of groups
400
300
200
100
0
367
211
122
90
64
50
28
27
25
12
8
8
7
3
3
2
2
1
1
1
1
1
un-annotated mirnas supported by H3K4me3 peaks from an independent experiment (8 tissues)
un-annotated mirnas supported by H3K4me1 peaks from an independent experiment (8 tissues)
un-annotated mirnas supported by H3K27ac peaks from an independent experiment (8 tissues)
un-annotated mirnas supported by H3K27ac peaks from an independent experiment (8 tissues)
un-annotated mirnas supported by CTFC peaks from an independent experiment (8 tissues)
un-annotated mirnas supported by ATAC-seq peaks (8 tissues)
1000 750 500 250 0
Number of miRNAs

Supplemental File 1

Supplemental File 2

Supplemental File 3

Supplemental File 4

Supplemental File 5

Supplemental File 6

Supplemental File 7

Supplemental File 8

Supplemental File 10

Supplemental File 11

Supplemental File 12

Supplemental File 13

Supplemental File 14

Supplemental File 15

Supplemental File 16

Supplemental File 17

Supplemental File 18

Supplemental File 19

Supplemental File 20

Click here to access/download

Supplementary Material

Supplemental_file20.xlsx

Supplemental File 21

Supplemental File 22

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Office of the Vice President for Research
2610 Beardshear Hall
Ames, Iowa 50011-2036
515 294-6344
FAX 515 294-6100

February 11, 2023

Dear GigaScience Editor,

I am very excited to share our new manuscript, entitled "Facilitating Functional genomics of cattle through integration of multi-omics data" for your consideration. Given the impact and novelty of our findings, we are confident that this work is ideally suited for publication in *Genome Research*. The comprehensive analyses presented herein crystallized into a decisive advance for the cattle transcriptomics field and beyond. This study applied a comprehensive set of transcriptome and chromatin state data from 47 cattle tissues and cell types to identify previously unannotated genes and improve the annotation of thousands of protein-coding and non-coding genes.

Predicted novel genes and transcripts were highly supported by independent Pacific Biosciences single-molecule long-read isoform sequencing (PacBio Iso-Seq), Oxford Nanopore Technologies sequencing (ONT-seq), Illumina high-throughput RNA sequencing (RNA-seq), Whole Transcriptome Termini Site Sequencing (WTTS-seq), RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression (RAMPAGE), chromatin immunoprecipitation sequencing (ChIP-seq), and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data.

Our key findings show that around half of protein-coding genes in each tissue are bifunctional and transcribe both in coding and noncoding isoforms. Critically, we identified 3,744 genes that functioned as non-coding genes in fetal tissues, but as protein coding genes in adult tissues. Most interestingly when the transcriptome was integrated with publicly available cattle QTL/association data using a novel bioinformatics approach, we were able to study tissue-tissue interconnection involved in different traits and construct the first bovine trait similarity network. These independent findings agree with published trait correlation data and move us closer to being able to identify the gene networks that underlie genetic correlations between traits.

Given these results, we strongly maintain that our work will be of general interest to the broad readership of *Genome Research*. All datasets generated in this study have been submitted to public databases, and all gene/protein names and symbols used in the paper adhere to approved nomenclature guidelines for specific species. We also confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere.

Thank you for your consideration and we look forward to hearing from you.

Sincerely,

James Reecy

James Reecy
Associate Vice President for Research