



## Original Article

## Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning



Yahui Gao<sup>a,b</sup>, Lingzhao Fang<sup>c</sup>, Ransom L. Baldwin VI<sup>a</sup>, Erin E. Connor<sup>d</sup>, John B. Cole<sup>a</sup>, Curtis P. Van Tassell<sup>a</sup>, Li Ma<sup>b</sup>, Cong-jun Li<sup>a,\*</sup>, George E. Liu<sup>a,\*</sup>

<sup>a</sup> Animal Genomics and Improvement Laboratory, BARC, USDA-ARS, Beltsville, MD 20705, USA

<sup>b</sup> Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA

<sup>c</sup> MRC Human Genetics Unit at the Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom

<sup>d</sup> Department of Animal and Food Sciences, University of Delaware, Newark, DE 19716, USA

## ARTICLE INFO

## Keywords:

Cattle  
Ruminal epithelial cell  
Single-cell RNA-seq

## ABSTRACT

Using the 10× Genomics Chromium Controller, we obtained scRNA-seq data of 5064 and 1372 individual cells from two Holstein calf ruminal epithelial tissues before and after weaning, respectively. We detected six distinct cell clusters, designated their cell types, and reported their marker genes. We then examined these clusters' underlying cell types and relationships by performing cell cycle, pseudotime trajectory, regulatory network, weighted gene co-expression network and gene ontology analyses. By integrating these cell marker genes with Holstein GWAS signals, we found they were enriched for animal production and body conformation traits. Finally, we confirmed their cell identities by comparing them with human and mouse stomach epithelial cells. This study presents an initial effort to implement single-cell transcriptomic analysis in cattle, and demonstrates ruminal tissue epithelial cell types and their developments during weaning, opening the door for new discoveries about tissue/cell type roles in complex traits at single-cell resolution.

## 1. Background

Rumen development is a critical process necessary for the digestion of solid feed (concentrates and roughage) and optimal growth performance in weaned cattle. Moreover, calf health during the weaning phase, specifically digestive health, has a central role in lifelong feed efficiency and methane emissions [1,2]. The neonatal rumen is undeveloped at birth, exhibiting rudimentary papillae without the high degree of keratinization, which is characteristic of the mature organ. The rumen's physical and metabolic development is incomplete at birth and largely remains so until further development is triggered by short-chain fatty acids (SCFA) resulting from bacterial fermentation of solid feed-stuffs. After establishing a viable ruminal fermentation, the maturation process proceeds [3] resulting in the epithelial layer increasing in

surface area to support absorption and metabolic differentiation to use SCFA as the primary energy substrate. Rumen epithelium serves as both a protective barrier from the digestive luminal environment and a metabolically important tissue for whole-animal energy metabolism [4–6]. A clear understanding of regulatory control of epithelial cell proliferation and differentiation and nutrient-gene interactions is crucial for the optimization of management strategies to support healthy ruminal development. The stratified squamous epithelium absorbs SCFA, which provides up to 70% of the energetic needs of mature animals, and serves as the primary producer of ketones in fed animals [7]. The ruminal epithelium consists of four strata: stratum basale, stratum spinosum, stratum granulosum, and stratum corneum [8]. The stratum basale is the layer of cells immediately adjacent to the basal lamina. These cells contain fully functional mitochondria and other organelles.

**Abbreviations:** AW, after weaning; BW, before weaning; FAANG, Functional Annotation of Animal Genome project; FDR, False Discovery Rate; GO, Gene Ontology; SCFA, short-chain fatty acids; TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site; UMI, unique molecular identifier.

\* Corresponding authors at: Animal Genomics and Improvement Laboratory, USDA-ARS, Building 306, Room 111, BARC-East, Beltsville, MD 20705, USA.

**E-mail addresses:** [gyhalvin@gmail.com](mailto:gyhalvin@gmail.com) (Y. Gao), [Lingzhao.fang@igmm.ed.ac.uk](mailto:Lingzhao.fang@igmm.ed.ac.uk) (L. Fang), [Ransom.Baldwin@usda.gov](mailto:Ransom.Baldwin@usda.gov) (R.L. Baldwin), [eeconnor@udel.edu](mailto:eeconnor@udel.edu) (E.E. Connor), [John.B.Cole@gmail.com](mailto:John.B.Cole@gmail.com) (J.B. Cole), [curt.vantassell@usda.gov](mailto:curt.vantassell@usda.gov) (C.P. Van Tassell), [lma@umd.edu](mailto:lma@umd.edu) (L. Ma), [Congjun.Li@usda.gov](mailto:Congjun.Li@usda.gov) (C.-j. Li), [George.Liu@usda.gov](mailto:George.Liu@usda.gov) (G.E. Liu).

<https://doi.org/10.1016/j.ygeno.2021.04.039>

Received 9 October 2020; Received in revised form 20 March 2021; Accepted 27 April 2021

Available online 29 April 2021

0888-7543/Published by Elsevier Inc.

The intermediate cell layers are the stratum spinosum and stratum granulosum, which are not distinctively separated [5]. RNA-sequencing has been used to identify the molecular mechanisms involved in rumen development, and several functional studies have been reported in the last decade [9–13]. However, those studies were performed using RNA isolated from whole tissues that include a composite of differentiated cell types. Therefore, they were limited to only measuring whole tissues and providing an average expression profile for all constituent cells [14].

Because of the unique and variable physical composition of the ruminal epithelium, it has been difficult to investigate the effects of differing ruminal environments on all aspects of rumen epithelial functions. The use of isolated ruminal epithelial cells provides several advantages in the study of ruminal metabolism and development [15]. As a crucial and high-value tool to study the development of rumen, we established a stable rumen epithelial primary cell (REPC) culture, explored its transcriptomic profile, and identified the direct effects of butyrate on gene expression in these cells. Correlated gene networks elucidated the putative roles and mechanisms of butyrate action in rumen epithelial development [2,16]. However, the transcriptome profiles are unique to individual cell type, developmental stage, health status, and biological function [17]. In our previous study, transcriptomic profiling of a total of 18 single cells and the clustering of the differentially expressed transcripts showed high divergence and variation in gene expression among the REPC [2]. Single-cell transcriptome complexity and single-cell transcriptome variation have also been reported in different cell types, such as gonadal and stem cell populations [18–20]. It is expected that individual cell phenotypes such as cell type, size, ultrastructure, and stage of the cell cycle could directly control cell-to-cell transcriptome variability. Therefore, large-scale sampling and single-cell transcriptome sequencing are necessary to identify cell type from tissues, evaluate dynamic cellular transitions with complex cell compositions, and illustrate impacts of cell-to-cell interactions among hundreds- to tens-of-thousands of cells.

Breakthroughs in the development of single-cell RNA-seq (scRNA-seq) technologies provide an avenue for dissecting tissue heterogeneity and understanding cell identity, fate, and function. High-throughput single-cell transcriptomes offer an unbiased approach to understanding gene expression variations between seemingly identical cells [21]. Han et al. used scRNA-seq to determine the cell-type composition of all major human organs and constructed a schematic representation of the human cell landscape (HCL) [22]. Their ‘single-cell HCL analysis’ pipeline helped to define cell identity. It was used to perform a single-cell comparative analysis of landscapes from humans and mice to identify conserved genetic networks. Several studies reported scRNA-seq analyses in the human small intestinal epithelium [23], in the human esophagus, stomach, and small and large intestines [24], as well as in the murine gastric organoid [25]. Despite all those developments, cell type profiles of cattle rumen epithelium at a single-cell resolution are lacking.

Many state-of-art analysis tools are available to process scRNA-seq data. For example, Seurat 3.0 [26] is an R package designed for quality control, integration, and scRNA-seq data analysis. Based on highly variable genes (HVG), Seurat performs clustering on cells using the Louvain algorithm [27–29]. Another analytical challenge is the interpretation of clusters and the assignment of cell types. SingleR 1.2.4 is an automatic annotation method that labels new cells from a test dataset based on similarity to the reference cell types [30]. Within SingleR, seven reference atlases are available, including the Human Primary Cell Atlas dataset [31] and the Blueprint and Encode dataset [32,33]. Gene regulatory network (GRN) inference can also reveal regulatory interactions and help identify the role of single cells. SCENIC (Single-Cell rEgulatory Network Inference and Clustering) can construct GRN from scRNA-seq data and infer transcription factor (TF) activity using a statistical method (AUCell 1.8.0) [34]. Trajectory inference methods interpret single-cell data as a snapshot of a continuous process. Monocle 2 [35] can order single cells in pseudotime to represent a biological process such as cell differentiation, according to an individual cell’s

asynchronous progression, using advanced machine learning techniques (such as Reversed Graph Embedding).

In this report, using the 10× Genomics Chromium Controller, we obtained transcriptomic profiling of 5064 and 1372 cells from ruminal epithelial cells of Holstein calves during weaning. We detected thousands of candidate marker genes among different cell clusters. We then examined these clusters’ underlying cell types and relationships by performing cell cycle, pseudotime trajectory, regulatory network, weighted gene co-expression network, and gene ontology (GO) analyses. This study provides an initial example for bovine single-cell analysis and opens the door for discoveries about tissue/cell type roles in complex traits at single-cell resolution.

## 2. Results

### 2.1. Data generation and quality assessment

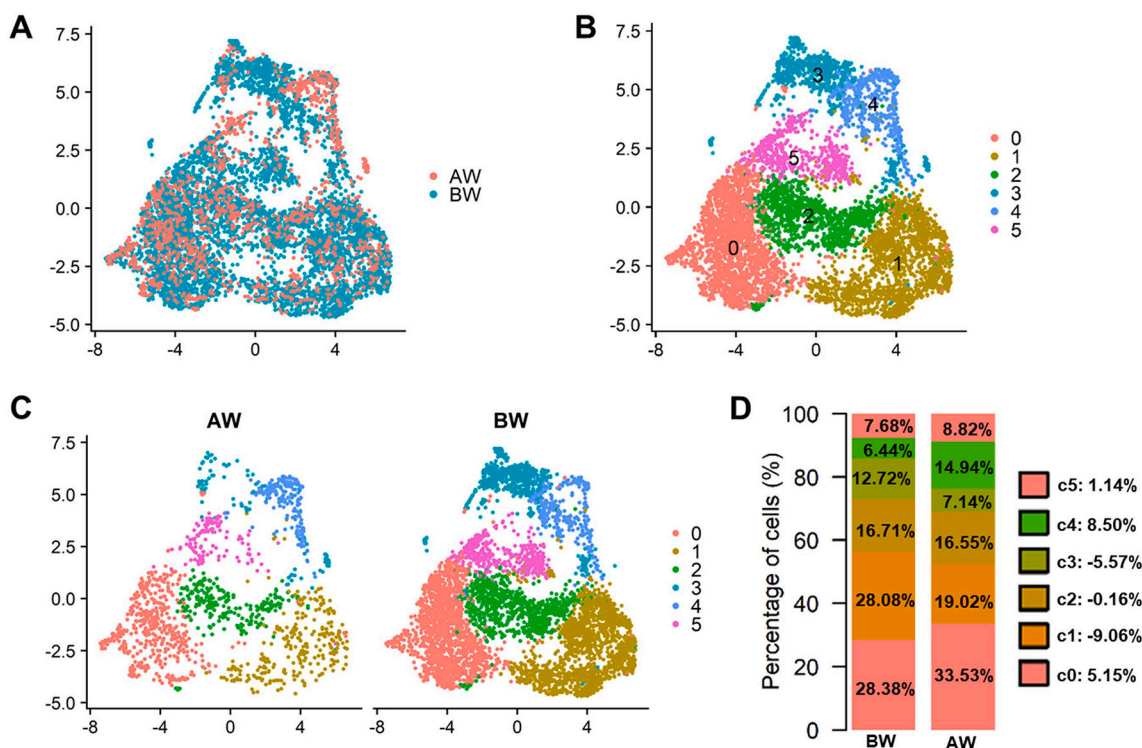
We used the 10× Genomics Chromium platform [36] to generate single-cell transcriptomes for two rumen tissues during weaning, one before weaning (BW) and another after weaning (AW) from two animals. In total, we sequenced 7479 single cells, with approximately 180,000 reads per cell (Table S1). After quality filtering and integration, we obtained 6436 single cells, which corresponded to a median of 37,000 unique molecular identifiers per cell, and more than 15,000 total genes detected in the whole population. Overall, 79% of all single cells (5064) belonged to BW, while the remaining 21% (1372) to AW.

### 2.2. Seurat cell cluster analyses

Using the Seurat v3.0 R package [26], we performed a community detection-based clustering to groups of cells according to their gene expression profiles. After visualizing the Uniform Manifold Approximation and Projection (UMAP) plots, we found the single-cell transcriptomes of two studied samples were largely similar (Fig. 1A), indicating a high degree of reproducibility. In total, we obtained 6 distinct clusters and named them Clusters C0, C1, C2, C3, C4, and C5 (Fig. 1B & C).

Additionally, we attempted to assign the cell types to the 6 Seurat clusters with SingleR [30], using the human cell reference datasets - Blueprint [37] and Encode [33]. In total, we obtained 12 cell types of the 6436 individual cell transcriptomes from the two rumen tissues (Fig. S1A). The cell count of different cell types ranged from 1 to 5634 (Table S2). After removing cell types with fewer than 12 cells, the three main cell types remaining were epithelial cells, keratinocytes, and mesangial cells (Fig. S1B). The majority of cells in C0, C1, C2, and C3 (Table 1) as epithelial cells. Compared to other clusters, the percentages of keratinocytes and mesangial cells were the highest in C4 and C5, respectively. When comparing the BW versus AW samples, based on the relative cell proportions, we noticed that C1 decreased 9.06%, C0 increased 5.15%, and C3 decreased 5.57%, while C4 increased 8.50% post-weaning. On the other hand, C2 decreased barely at 0.16% and C5 increased slightly at 1.14%, respectively (Fig. 1D).

To compare the performances of bulk and scRNA-seq, we calculated the gene expression correlations between bulk and scRNA-seq. We retrieved RNA-seq data of rumen bulk samples both before and after weaning, as we reported before [2]. As shown in Fig. S1C, the overall correlations were more than 0.65, no matter using all cells’ gene expression (Left panel) or the clustered cells’ gene expression within each cluster (Right panel), which suggested that scRNA-seq and bulk RNA-seq results were generally consistent. However, for a heterogeneous tissue, conventional bulk RNA-seq approaches have difficulties in accurately revealing cell-type-specific changes in gene expression, particularly for rare cell types. On the other hand, since scRNA-seq can capture the gene expression at both cell and bulk tissue levels, we used it to study individual cell types by analyzing alterations in gene transcription at the single-cell level.



**Fig. 1.** Cluster analysis of single-cell transcriptomes from two calf rumen tissues. (A) UMAP projection plot showing dimensional reduction of the distribution of 6436 individual cell transcriptomes from two rumen tissues (green = before weaning; red = after weaning); (B) UMAP projection plot showing six major clusters of the 6436 individual cell transcriptomes; (C) UMAP projection plot showing annotation by before and after weaning rumen tissues. (D) The percentage of cell types across the pre-and post-weaning rumen tissues. The cell types were annotated based on (C). The numbers in the legend indicate the differences between the pre-and post-weaning rumen tissues within each cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**

Cell components and cell-cycle index for each cluster.

| Cluster | Epithelial cells |       | Keratinocytes |       | Muscular cells |       | Others          |      | All  | Dividing cells |   |
|---------|------------------|-------|---------------|-------|----------------|-------|-----------------|------|------|----------------|---|
|         | Count            | %     | Count         | %     | Count          | %     | Count           | %    |      | Count          | % |
| 0       | 1740             | 91.72 | 27            | 1.42  | 130            | 6.85  | 0               | 0    | 1897 | 1.42           |   |
| 1       | 1593             | 94.65 | 17            | 1.01  | 71             | 4.22  | 2               | 0.12 | 1683 | 96.14          |   |
| 2       | 999              | 93.1  | 10            | 0.93  | 62             | 5.78  | 2               | 0.19 | 1073 | 20.32          |   |
| 3       | 614              | 82.75 | 27            | 3.64  | 87             | 11.73 | 14              | 1.89 | 742  | 21.29          |   |
| 4       | 337              | 63.47 | 113           | 21.28 | 45             | 8.47  | 36              | 6.78 | 531  | 41.81          |   |
| 5       | 351              | 68.82 | 43            | 8.43  | 102            | 20    | 14              | 2.75 | 510  | 32.75          |   |
| Total   | 5634             |       | 237           |       | 497            |       | 68 <sup>a</sup> |      | 6436 |                |   |

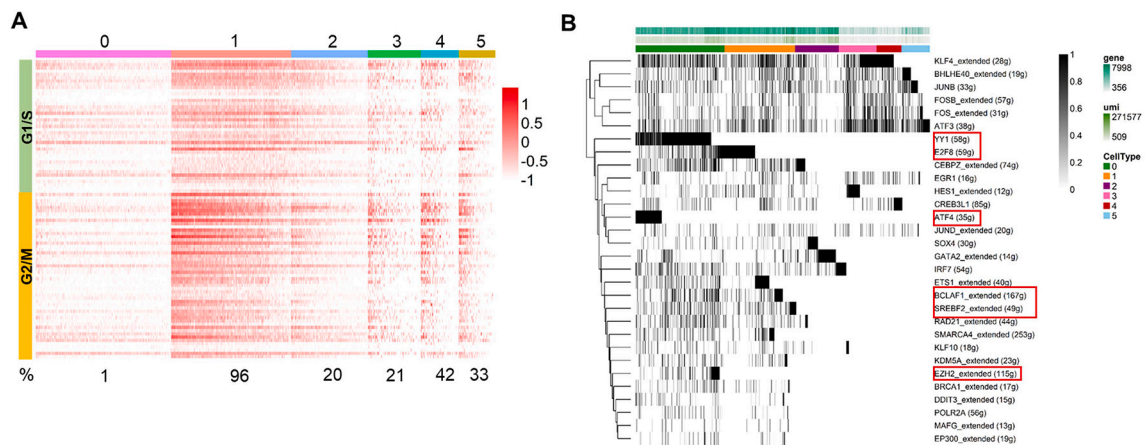
<sup>a</sup> Others included 35 Fibroblasts, 14 Myocytes, 7 Erythrocytes, 6 MEP and 1 Astrocyte, 1 Chondrocyte, 1 CLP, and 1 Endothelial cell. In each cluster, these other cell counts were fewer than 12.

### 2.3. Cell cycle analysis and SCENIC results for the rumen tissues

To explore the proliferation status of Seurat cell clusters, we performed the cell cycle analysis to calculate their cell cycle indices, using sets of 43 G1/S and 55 G2/M genes [38] (Table S3). The expression profiles of cell cycle-related genes revealed that the overall cell cycle indices were 1%, 96%, 20%, 21%, 42%, and 33% for Clusters C0 to C5, respectively (Fig. 2A and Table 1). These results suggested that C1 cells were actively dividing (96.14%), whereas C0 cells were not actively proliferating (1.42%). The cell cycle indices for C2 and C3 were around 20%, while those for C4 and C5 were 41.81% and 32.75%, respectively. Additionally, the average cell cycle indices were 75%, 59% for all BW or AW cells, respectively (Fig. S2A). Within each cell cluster, we also compared cell cycle indices between BW and AW cells. We found that cell cycle indices stayed similar for C0 and C1; increased for C3 (18% to 30%), but decreased for C2 (22% to 14%), C4 (46% to 12%), and C5 (55% to 18%) between BW and AW cells, respectively (Fig. S2B).

Furthermore, we performed another Seurat cell clustering after removing these cell cycle genes. When we compared the Seurat results with vs. without the cell cycle genes (Fig. S1D and Fig. S1E), we found that the global distribution patterns of BW and AW cells, as well as those of their corresponding cell clusters, were generally similar.

As important regulators of gene expression, transcription factors (TF) are very useful for identifying cell types. Thus, we performed the SCENIC analysis [34] to identify regulators and gene regulatory networks. Briefly, SCENIC infers co-expression modules between TF and candidate target genes using machine learning regression techniques (e.g., random forest or gradient boosting machines), which are pruned based on the enrichment of the TF motif around the TSS of the potential target genes, resulting in regulons. Based on the AUCell algorithm, SCENIC calculates the activity of each regulon in single-cell transcriptomes to obtain the corresponding area under the curve (AUC) scores, which are used to rank the cells for a given regulon and determine a threshold for active or inactive expression. Through this analysis, we identified 30 active



**Fig. 2.** Cell-cycle analysis and SCENIC results on the rumen tissues. (A) Heatmap showing expression levels of cell-cycle-related genes in each Seurat cluster. Cells were ordered according to the average expression level of cell-cycle-related genes within each cell. The color key from white to red indicated expression levels from low to high. The cell-cycle index of each cell type is shown at the bottom of the heatmap. (B) SCENIC binary regulon activity matrix shows all correlated regulons active in at least 1% of all regulons. Each column represents a single cell, and cluster labels correspond to those used in the UMAP plot. Representative transcription factors are highlighted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

regulons in the rumen (Fig. 2B). The count range of target genes of these regulons was between 12 and 252 (Table S4). SCENIC analysis revealed several important transcriptional regulators modulating cell type-specific gene regulatory networks. For Cluster C0, we identified its specific TF, including ATF4, EZH2, and YY1. For clusters combining C1 and C0, we detected E2F8, ETS1\_extended, BCLAF1\_extended, SREBP2\_extended, SMARCA4\_extended, KDM5A\_extended, BRCA1\_extended, DDIT3\_extended, POLR2A, MAFG\_extended, and EP300\_extended. For Clusters C0, C1, and C2, especially for C2, we discovered CEBPZ\_extended, SOX4, and GATA2\_extended.

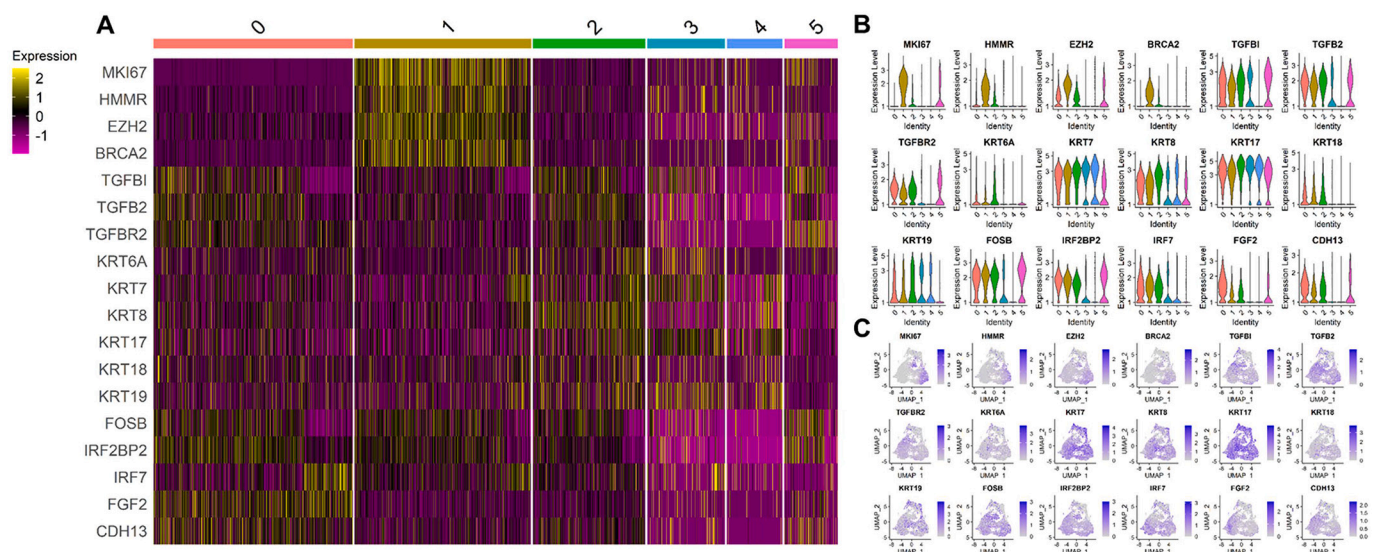
2.4. Marker gene expression for rumen cell clusters

To profile gene expression patterns of the different clusters identified by Seurat above (Table S5), we analyzed the expression of the top 10 marker genes in each cluster, as compared to all other clusters. Heatmap analysis revealed distinct signatures from each cluster (Fig. S3A). Of note, some of these top marker genes were highly expressed in only one

cluster, such as *UHRF1* (Ubiquitin Like With PHD And Ring Finger Domains 1) in C1 and *ACTA2* (Actin Alpha 2, Smooth Muscle) in C2, whereas other markers were conserved across two or more clusters, such as *FTH1* (Ferritin Heavy Chain 1), *MT-ND3* (Mitochondrially Encoded NADH: Ubiquinone Oxidoreductase Core Subunit 3), *MT-COX1* (official name *MT-CO1*: mitochondrially encoded cytochrome c oxidase I), and *RPLP1* (Large Ribosomal Subunit Protein P1 in Figs. S3B and S3C). Among these marker genes, there were some specific genes related to the cell cycle, such as *MKI67*, *HMMR*, *EZH2*, and *BRCA2*. We also detected epithelial cell marker genes, including *TGFβ1*, *TGFβ2*, and *TGFβR2*. Moreover, we obtained distinct sets of keratins for these cell clusters, some of which were up- or down-regulated in a specific cluster (Fig. 3).

2.5. Pseudotime analysis

In order to estimate the lineage relationships among the Seurat clusters and better understand the development states of all cells, we conducted a pseudotime analysis to infer the cell trajectories using



**Fig. 3.** Characterization of differential gene expressions for rumen tissues. (A) Gene expression heatmap of the 10 most differentially expressed genes in each cluster compared to all other clusters. Genes are represented in rows and cell clusters in columns. (B) Violin plots of gene expression. Expression in each cell is shown along with the probability density of gene expression, denoted by the shape of the plot. (C) UMAP projection plots showing transcript accumulation for cell marker genes in individual cells. Color intensity indicates the relative transcript level for the indicated gene in each cell.

Monocle 2 [35]. Following a “developmental/transitional” path according to their transcriptomic similarity, we identified one major and long-trajectory branch and one minor and short-trajectory branch, with cells ordered in an arrangement from proximal to distal distribution (Fig. 4A). Combining with the pseudotime values (Table S6), we observed that the long-trajectory tree rooted from C1, sprouted into C0, C3, and C4 (C1 → C0 → C3 → C4), while the short trajectory tree rooted from C2, sprouted into C5 (C2 → C5). The long path appeared to agree with our definitions on the Seurat clusters previously, i.e., from proliferating epithelial cells (C1) to resting epithelial cells (C0), to differentiated epithelial cells (C3), finally to keratinized epithelial cells (C4). The short path from C2 to C5 seemed to correspond to vascular smooth muscle development. C4 and C5 with the highest pseudotime scores might represent the terminal developmental states for either of these two paths, respectively.

## 2.6. Co-expression analyses

To systematically investigate the genetic program dynamics, we performed weighted gene co-expression network analysis (WGCNA) [39] using 2000 marker genes derived by Seurat. WGCNA identified 6 gene modules (Fig. 5A), each containing gene sets that tend to be co-expressed (Table S7). We then performed gene ontology (GO) analyses for genes in each module to investigate their biological functions (Fig. 5B, Table S8). To assign co-expressed gene functions to Seurat clusters, we generated a correlation heatmap in Fig. 5C. For example, the blue module genes were enriched for cell cycle and division, and the blue module was significantly represented in all cell clusters from C1 to C5, except for C0, likely corresponding to C0’s resting nature. The brown module genes were enriched for epithelial cell proliferation, and the negative regulation of the developmental process and metabolic process. The blue module is significantly associated with all other Clusters, except for C4, indicating C4’s terminal differentiation. The turquoise module genes were enriched for multiple GO terms, including epithelial cell mobility and differentiation, cell-cell junction and adhesion, cell division and death, and blood vessel development. This module was more correlated with all other clusters (correlation coefficients of 0.63–0.98) than with C1 (0.49), suggesting C1’s dividing but undifferentiated states. The yellow module genes were enriched for extracellular matrix organization and this module was strongly associated with Clusters C3 and C4, probably related to their absorption and protection

mechanisms. Especially for C4, the yellow module was most correlated, suggesting its keratinized epithelial cells could provide skin-like function inside rumen and further support our C4 assignment.

## 2.7. Trait-relevant cell clusters

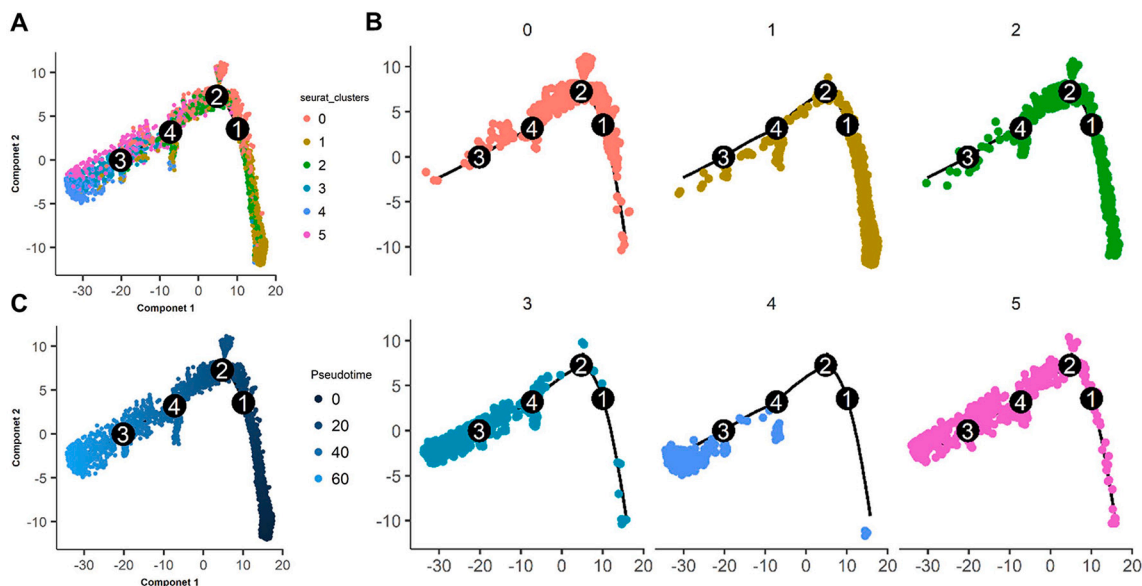
Using a permutation-based marker-set test approach (Methods), we tested the enrichment of 45 GWAS signals within marker genes of distinct clusters (Table S5) reported by Seurat (FDR < 0.05) (Fig. 6A). Production and body conformation traits were significantly associated with all clusters, especially C5, C0, and C4, reflecting the important functions of these cell types related to SCFA absorption and tissue development. In addition, health traits, such as SCS (somatic cell score) were associated with all clusters except C1 and C2, suggesting that the differentiated and terminal cell types have a role in tissue integrity and immunity. This might reflect that rumen plays a role in the regulation of immunity. Moreover, based on the marker genes reported by edgeR (Table S9) between cell clusters across the BW and AW rumen samples, we also detected similar results (Fig. 6B).

## 2.8. Cross-species comparison

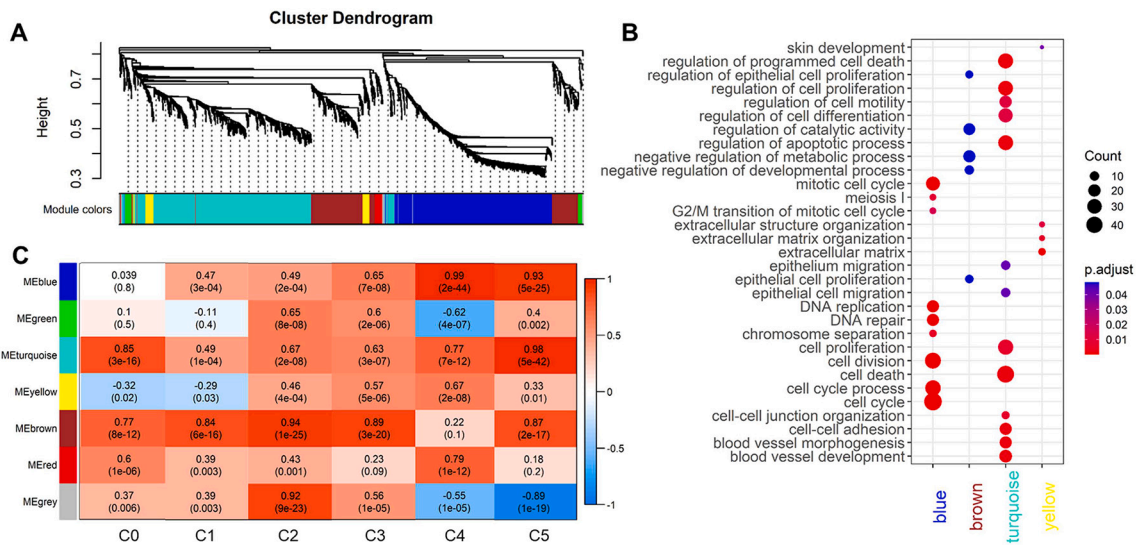
To support our cattle rumen results using human data, we downloaded the scRNA-seq dataset of the human stomach from GSE134355 [22] and performed Seurat clustering analysis [26]. Plotting the single-cell transcriptomes via UMAP projection yielded largely overlapping distributions of cells from 2 cattle and 7 human samples (Fig. S5A), validating our scRNA-seq data generation, processing, and cell type assignment. In total, we identified 13 distinct clusters (Fig. S5A). Using SingleR [30], we obtained the 29,926 individual cell transcriptomes of 42 cell types from the nine samples (Fig. S5B). The UMAP plot distribution reflected that the main cell types were adipocytes, epithelial cells, keratinocytes, mesangial cells, monocytes, plasma cells, and skeletal muscle (Fig. S5B). Within them, we could validate epithelial cells, keratinocytes, and mesangial cells identified in our cattle rumen samples.

## 3. Discussion

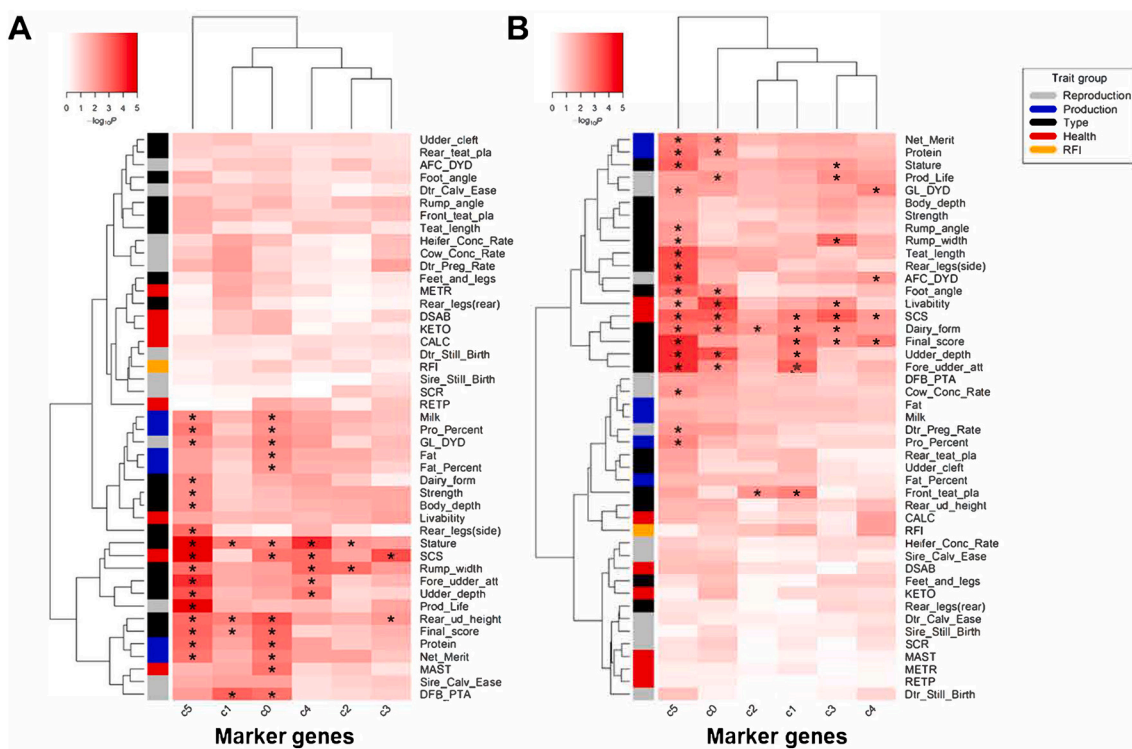
A previous human study showed that stratified squamous epithelia of internal organs are generally similar to skin, although they make



**Fig. 4.** Pseudotime analysis using Monocle 2 for cell transcriptomes. Solid black lines indicate the main diameter path of the minimum spanning tree (MST) and provide the backbone of Monocle’s pseudotime ordering of the cells. Each dot represents an individual cell colored by cluster (A, B), or pseudotime (C).



**Fig. 5.** WGCNA suggested genetic networks. (A) Dendrogram showing the gene co-expression network constructed using WGCNA. The color bar labeled as “Module colors” beneath the dendrogram represents the module assignment of each gene. (B) Significantly enriched GO terms based on genes within each module. (C) The relationship between Modules and Seurat clusters. The upper numbers within each grid are the correlation between each module and Seurat cluster. The numbers in brackets represent the *P*-values.



**Fig. 6.** Associations of cell clusters with complex traits based on GWAS signal enrichment analyses using marker genes among cell clusters (A) and between pre-and post-weaning (top 5%) (B). “\*” denotes *FDR* < 0.05.

different types of keratins [40]. An early study reported influences of extracellular matrix components on the growth and differentiation of ruminal epithelial cells in primary culture [41]. Xiang et al. studied the response to and regulation of the layers of the full rumen wall to different diets fed to sheep, using bulk RNA-s [11]. They identified clusters of genes characteristic of cell proliferation and differentiation, as well as metabolism-specific genes within the epithelium network. The expressions of cell-cycle and metabolic genes were positively correlated with dry matter intake, ruminal SCFA concentrations, and methane

production. They reported that the TF expression patterns and their targets, ruminal epithelium, mimic proliferating and differentiating skin, suggesting conservation of regulatory networks. From an evolutionary perspective, Pan et al. reported that a positively selected and ruminant-specific gene, *WDR66*, regulates the expression of occludin, which then tightens the intercellular space and controls epithelial permeability [42]. They speculated that when junction structure (desmosome) between keratinocytes of the ruminal epithelium becomes loose, the enlarged intercellular space with its copious blood supply

enables nutrient absorption across the ruminal epithelium. However, all the above studies were performed with bulk samples. Cell-type profiles for cattle rumen epithelial cells at a single-cell resolution are necessary to compare the undeveloped ruminal epithelium, consisting primarily of strata basal and spinosum cell types, and fully differentiated ruminal epithelium, comprising four strata with highly variable cell content [15].

Using the 10× Genomics Chromium Controller, we did scRNA-seq on Holstein ruminal epithelial cells during weaning. To our knowledge, this represents the first reported single-cell transcriptomic analysis in cattle. Our study was successful in generating rumen single-cell transcriptomes, revealing major and some novel cell types. In the current study, we identified 6 distinct cell clusters and tentatively assigned their cell types using Human Cell Atlas/Blueprint reference cell datasets (Fig. 1 and Table 1). For mesangial cell assignment, because we used human reference cell types to assign cattle cells, these designations may be biased towards human structure and function. Even in humans, mesangial cells are believed to share the same origin as vascular smooth muscle cells and are sometimes considered to be a type of specialized vascular smooth muscle cell [43]. Therefore, it might be more helpful to rename those cells as vascular smooth muscle cells despite their mesangial cell assignments. We also detected thousands of marker genes among cell clusters during weaning (Table S1). We then performed GO and KEGG-based gene enrichment and cell cycle analyses (Fig. S4 and Fig. 2A). In the co-expression analyses (Fig. 5), we obtained 6 distinct modules (Fig. 5A) and significantly enriched GO terms based on genes within each module (Fig. 5B). We then assigned co-expressed gene functions to specific cell clusters (Fig. 5C). When we integrated these marker genes with Holstein GWAS signals, we observed all clusters, especially C5 and C0, were enriched for animal production and body type traits. Additionally, we also substantiated the cattle cell identities by comparing them with the human and mouse stomach epithelial cells.

We found that Cluster C1 contained 94.65% epithelial cells and 96.14% cells were dividing. C1-specific genes were enriched for cell cycle, chromosome segregation, cytoskeleton, DNA replication, nuclear division, etc. (Fig. S4 C1). C0 contained 91.72% epithelial cells, but only 1.42% cells were dividing, and C0-specific genes were enriched for RNA binding, localization, and degradation, cytosolic ribosome, regulation of peptidase activity, and metabolic processes (Fig. S4 C0). C2 contained 93.10% epithelial cells, 0.93% keratinized epithelial cells, 5.79% vascular muscle cells, and 20.32% cells were dividing. C2-specific genes were enriched for cell differentiation processes, including myofibril, smooth muscle proliferation and contraction, extracellular matrix-receptor interaction, regeneration, and cell cycle (Fig. S4 C2). C3 contained 82.75% epithelial cells, 3.64% keratinized epithelial cells, 11.73% vascular muscle cells, and 21.29% cells were dividing. C3-specific genes were enriched for cell aging, positive regulation of establishment of protein localization to the telomere, insulin-like growth factor binding, response to cytokine, and interaction with symbiont (Fig. S4 C3). C4 contained 63.47% epithelial cells, 22.28% keratinized epithelial cells, 8.47% of vascular muscle cells, and 41.81% cells were dividing. Notably, C4-specific genes were enriched for skin development (keratinization), response to hypoxia, extracellular matrix organization, apoptotic process, cell cycle and death, as well as cell adhesion and migration (Fig. S4 C4). C5 contained 68.82% epithelial cells, 8.43% keratinized epithelial cells, 20.00% vascular muscle cells, and 32.75% cells were dividing. C5-specific genes were enriched for mitochondrial respiratory chain complex I assembly and peptide and protein metabolic process (Fig. S4 C5).

To estimate the effects of the cell cycle genes on cell clustering, we performed Seurat cell clustering with or without these cell cycle genes. Our cell clustering results showed that there were no significant distribution differences for BW or AW samples, or for six rumen cell types during the weaning process (Fig. S1 D and E). These results indicated that the cell clustering and type assignment mainly reflected physiological differences among these cell types (cell differentiation), rather

than the effects of the cell cycle statuses (cell division) under our analysis conditions. Based on these results and existing literature, we proposed the following model for cattle rumen epithelial development (Fig. S6). We designate cell types to Seurat clusters along 2 lineages in the following networks, which can better reflect the temporal and spatial distributions for the ruminal epithelium layer. Lineage 1 is C1 → C0 → C3 → C4, including proliferating epithelial cells (C1), resting poised epithelial cells (C0), differentiated epithelial cells (C3), and keratinized epithelial cells (C4). Lineage 2 is C2 → C5, including vascular muscle precursor cells (C2) and vascular muscle cells (C5). As C2 appeared at roughly the same time as C1, and both were earlier than C0, it is less likely that C2 was derived from C0. However, the relationship between C1 and C2 was not clear, even though the overwhelming majority of their cells were the same type (epithelial cells) at 94.65% and 93.10%, respectively. Thus, we speculated that C1 and C2 could be derived from the same epithelial stem cells, which linked Lineages 1 and 2 together.

With the above model in mind, we checked cell-type-specific marker genes and TF. For cell cycle-related TF, we could readily identify them from Cluster C1's marker genes and SCENIC results, including MKI67 (ranked as No. 8 by its *P*-value), HMMR (No. 54), and EZH2 (No. 70). These three TFs were the same as reported by Xiang et al. in sheep rumen feed efficiency research [11]. They also reported the cell cycle regulator, BRCA1, was present in sheep, while we found its close relative, BRCA2, in cattle.

For epithelial cell marker genes, we detected transforming growth factor-beta receptors or ligands, such as TGFβ1, TGFβ2, and TGFβR2. Their expressions were decreased in C4, while expressions of TGFβ1 and TGFβR2 were increased in C5. The TGFβ1 (Transforming growth factor, beta-induced) protein contains the common peptide motif (arginylglycylaspartic acid - RGD), which binds to type I, II, and IV collagens. Previous studies reported that the TGFβ1 protein is secreted, induced by TGFβ, and associated with normal skin and adhesion of dermal fibroblasts [44] or keratinocytes [45]. Bond et al. reported their first discovery of TGFβ1 in rumen epithelium, possibly modulating cell adhesion [46]. The TGFβ superfamily is critical in wound healing and repair. It must be activated by release from the extracellular matrix where it is bound by latent TGFβ-binding proteins and active proteases, such as the matrix metalloproteinase [47]. TGFβ has been shown to inhibit the proliferation of keratinocytes [48,49]. Additionally, in humans, ligands TGFβ1, TGFβ2, and TGFβ3 all function through the same receptor signaling systems. This pathway is involved in many cellular processes in both the adult organism and the developing embryo, including cell growth, cell differentiation, apoptosis, cellular homeostasis, and other cellular functions. We previously reported that TGFβ1 is an important transcriptional regulator of gene expression networks related to certain diets using the same calf rumen epithelium samples during weaning [50]. Our rediscovery of the same TGFβ pathways from the scRNA-seq assay, further confirmed that these cytokines and their related proteins are likely involved in regulating the growth and differentiation of the rumen epithelium. In C4 and C5, their expression repressions and inductions may correspond to the different cell specializations for keratinized epithelial cells and vascular smooth muscle cells, respectively. Further characterization of TGFβ pathway gene expression and distribution within the extracellular matrix and among the layers of the rumen epithelium during proliferation and differentiation is needed to better understand its function in the ruminal mucosa related to these physiological processes.

Interestingly, we also obtained distinct sets of keratins from the marker genes among these cell clusters, such as C1: down-regulation of KRT17; C2: up-regulation of KRT8, KRT17; C3: up-regulation of KRT17, but down-regulation of KRT6A; and C5: down-regulation of KRT7, KRT19, KRT8, KRT18, and KRT17. For example, KRT8 is well-known to be expressed in epithelial cells of the human gastrointestinal tract (including stomach, colon, small intestine, gall bladder, liver, and pancreas) and mammary gland ducts [51]. Additionally, we found *FOSB*

(FosB proto-oncogene, AP-1 transcription factor subunit) was down-regulated in C4 but up-regulated in C5, which was reported to play a role in epithelial proliferation and differentiation [11]. We also detected that interferon *IRF2BP2*, which is known to be involved in immunity, was decreased in C4, and *IRF7* was also decreased in C5. At the same time, we also obtained distinct sets of genes from Cluster marker genes, like *FGF2* (Fibroblast Growth Factor 2, which is related to *FGF7* – a potent epithelial cell-specific growth factor). *FGF2* gene expression was decreased in C1 but increased in C0. *FGF2*'s other close relatives include *FGF6*, which has been shown to be associated with cattle traits, like body depth, rump width, sire calving ease, stature, and others [52]. *CDH13* (cadherin 13, related to the cell adhesion protein *E-Cadherin*) was down-regulated in C4.

Within the marker genes reported among cell clusters, the *CENPF* gene encodes a protein that associates with the centromere-kinetochore complex and it may play a role in chromosome segregation during mitosis [53] and *NPM1* encoded a protein that is involved in several cellular processes, including centrosome duplication and cell proliferation [54]. We also discovered important TF modulating cell-type-specific gene regulatory networks. For Cluster C0, we identified its specific TF, including ATF4, EZH2, and YY1. *EZH2* encodes a member of the Polycomb-group (PcG) family, which maintains the transcriptionally repressive state. YY1 is a ubiquitously distributed transcription factor belonging to the GLI-Kruppel class of zinc finger proteins, which can activate or repress the promoter [55]. For clusters combining C1 and C0, we detected *BCLAF1*, *BRCA1*, *SMARCA4*, *EP300*, and other TF. *BRCA1* encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor [56]. *BCLAF1* encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. As a member of the *BRCA1*-associated genome surveillance complex (BASC), the gene product plays a role in transcription, DNA repair of double-stranded breaks, and recombination [57]. Both *SMARCA4* and *EP300* regulate transcription via chromatin remodeling and are important in cell proliferation and differentiation [58,59]. For Clusters C0, C1, and C2, especially for C2, we discovered *CEBPZ*, *SOX4*, and *GATA2*, all of which are important transcriptional regulators for cell growth and differentiation [60–62].

**Conclusions:** In summary, this study provides an initial example for bovine single-cell analysis and opens the door for new discoveries about tissue/cell type roles in complex traits at single-cell resolution. We provided the first cell type profiles for cattle rumen epithelial cells at a single-cell resolution. We characterized their cell cycle, component, relative timing, and regulatory networks, as well as co-expression and gene function patterns. With our proposed cell lineage development model, we reported 6 cell types identified across their temporal and spatial distributions, which appear to be correlated with the rumen epithelium's underlying layers, structures, and functions. This rumen cell development model will need to be further tested and improved by more replicates and functional validations. For example, spatial transcriptomics data will be needed to locate the relative position for each cell cluster over the development stages. More future experiments are warranted to investigate the mechanisms, which regulate the commitment of epithelial stem cells to differentiate in distinct lineages.

## 4. Methods

### 4.1. Sample collection

Animals and tissue collection were fully described in our previous report [8]. Briefly, two Holstein bull calves were chosen: one calf (pre-weaning) was fed with milk replacer only (MRO - Cornerstone 22:20, Purina Mills, St. Louis, MO, USA; 22.0% crude protein, 20.0% crude fat, 0.15% crude fiber, 0.75 to 1.25% Ca, 0.70% P, 66,000 IU/kg vitamin A, 11,000 IU/kg vitamin D3, and 220 IU/kg vitamin E) for two weeks; while the other (post-weaning) was fed with MRO for six weeks,

followed by a combination of milk replacer and grain-based commercial calf starter for four weeks. Calves were euthanized by captive bolt followed by exsanguination at day 14 or day 70 to represent development at two stages of weaning on a grain concentrate diet. Rumen epithelial tissue was collected from the anterior portion of the ventral sac of the rumen beneath the reticulum and below the rumen fluid layer at slaughter. The epithelial layer of the rumen tissue was separated manually from the muscular layer. After rinsing with tap water to remove residual feed particles, samples were further rinsed in ice-cold physiological saline, and subsamples of epithelial tissues (approximately 600 mg) were fixed in RNAlater (Life Technologies, Grand Island, NY, USA) RNA stabilization solution according to the manufacturer's instructions and stored at  $-80^{\circ}\text{C}$  until use.

### 4.2. Single-cell isolation and RNA-seq library preparation and sequencing

Rumen tissue samples from the one pre-weaned and one weaned calf were collected and processed by a commercial service provider, Singulomics (New York, NY, USA), for scRNA-seq analysis. Library preparation was performed according to instructions by using the 10× Genomics Chromium single-cell controller. The libraries were then pooled and sequenced on a HiSeq4000 (Illumina, San Diego, CA, USA).

### 4.3. Generation of single-cell transcriptomes

We first processed 10× Genomics raw data by the Cell Ranger Single-Cell Software Suite (release 3.1.0), including using Cell Ranger *mkfastq* to demultiplex raw base-call files into FASTQ files followed by the use of Cell Ranger *count* to perform alignment, filtering, barcode counting, and UMI counting. The raw reads were aligned to the ARS-USD1.2 cattle reference genome [63] by Cell Ranger *pipeline* using default parameters. The output summary of the two samples is shown in Supplemental Table 1. All downstream single-cell analyses were performed using the Seurat 3.0 [26] R package v3.6.3 unless explicitly mentioned.

### 4.4. Quality control, dimension reduction, and cell clustering

Overall, 5064 and 1372 cells passed the following quality control thresholds: all genes expressed in fewer than 3 cells were removed; the number of genes expressed per cell  $>200$  as low and  $< 8000$  as high cut-off; UMI counts less than 200; the percent of mitochondrial-DNA derived gene-expression  $<30\%$ . We used the LogNormalize method of the “Normalization” function to calculate the expression value of genes. We then restricted the corrected expression matrix to the subsets of highly variable genes (HVG), and centered and scaled values before performing dimension-reduction and clustering on them. We selected 2000 genes as HVG using the “FindVariableFeatures” function with default parameters. We then used the “RunPCA” function to perform the principal components analysis (PCA) on the single-cell expression matrix with genes restricted to HVG. The number of significant principal components was determined using a permutation test implemented by the “JackStraw” function. Within all the PC, the top 10 PC were used for clustering and Uniform Manifold Approximation and Projection (UMAP) analysis. To find clusters, we used the weighted Shared Nearest Neighbor (SNN) graph-based clustering method implemented by the “FindNeighbors” function and then utilized the “FindClusters” function to conduct the cell-clustering analysis by embedding cells into a graph structure in PCA space. Based on the number of cells in our study, we set the parameter resolution to 0.24. Visualization of the cells was performed using the UMAP algorithm as implemented by the Seurat “RunUMAP” function.

### 4.5. Assigning cell types to single-cell clusters

Two methods were utilized to assign the cell clusters identified by Seurat. First, canonical cell-type marker genes that are conserved across

conditions were identified using the “FindConservedMarkers” function with default parameters. Marker genes with significant specificity to each cluster were annotated with their known functions. Additionally, raw expression data for the filtered cells were used for cell type assignment using SingleR [30] with default parameters using the Blueprint [37] and Encode [33] human cell atlases. To compare scRNA-seq gene expression levels with bulk RNA-seq data, we calculated average gene expression values across all cells or cells within a cluster. Pearson correlation coefficients were calculated between snRNA-seq and bulk RNA-seq values for all genes expressed in both data sets.

#### 4.6. Pseudotime trajectory analysis

For trajectory analysis, we used Monocle 2 [35] to order cells in pseudotime based on their transcriptional similarities, with UMI counts modeled using a negative binomial distribution. First, we integrated the preprocessed Seurat objects into Monocle 2, utilizing the “new-CellDataSet” function. We then determined the differentially expressed genes or marker genes that were identified using the “differentialGeneTest” function. We next reduced the dimensionality of the data to two dimensions using the discriminative dimensionality reduction with trees (DDRTree) method implemented in the “reduceDimension” function. Finally, after pseudotime calculations were made for each cell, we projected clusters derived from the Seurat object onto the minimum spanning tree upon cell order using the “plot\_cell\_trajectory” function.

#### 4.7. Cell-cycle analysis

Sets of 43 G1/S and 55 G2/M genes [38] were used in the cell-cycle analysis. To calculate the ratio of actively proliferating cells of each feature, such as different clusters and different weaning stages, we first calculated the total expression levels of all 98 cell-cycle genes in every single cell, and only cells with mean expression levels higher than the average values of all clusters were regarded as actively proliferating.

#### 4.8. Single-cell regulatory network inference and clustering (SCENIC) analysis

We conducted SCENIC analysis on cells after filtering for each major cell type using the R package SCENIC v1.1.2 [34], which is a computational workflow that predicts TF activities from scRNA-seq data. Instead of interrogating predefined regulons, individual regulons are constructed from the scRNA-seq data. Regions for TF searching were restricted to a 10 kb distance centered on the transcriptional start site (TSS) or 500 bp upstream of the TSS. First, TF-gene co-expression modules are defined in a data-driven manner with GENIE3 v1.8.0. Subsequently, those modules are refined via RcisTarget by keeping only those genes that contain the respective transcription factor binding motif (TFBS). Once the regulons are constructed, the method AUCell scores individual cells by assessing for each TF separately whether target genes are enriched in the top quantile of the cell signature.

#### 4.9. Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) was performed with functions in the WGCNA v1.69 R package following the previously published study by Tosches and colleagues [64]. According to the methods, the analyses were performed on pseudocells, calculated as averages of 100 cells randomly chosen within each cluster. The top 2000 highly variably expressed genes determined in Seurat were used for analysis. Briefly, the topological overlap matrix (TOM) was constructed with softPower and was set to 2. The hub genes for each module were identified as module eigengene. The GO enrichment analysis was performed by ClusterProfiler [65] R package using hub gene data sets and the BH method was employed for multiple test correction. GO terms

with a *P*-value lower than 0.05 were considered significantly enriched.

#### 4.10. Gene function analysis

To get the lists of marker genes, we first extracted the genes' UMIs across cells within each cluster and then assigned cells to the BW or AW group. Based on the gene x cells matrix, we utilized edgeR to detect marker genes for each cluster between BW and AW (Table S9). We used lists of genes differentially expressed in each of the six clusters for GO and KEGG using Cytoscape 3.8.0 analyses with the ClueGO app [66]. Fisher exact test was used to measure gene enrichment in annotation terms. FDR corrected *P*-values were used to search for significantly enriched terms. GO terms and KEGG pathways with a *P*-value lower than 0.05 were considered significantly enriched (Table S10).

#### 4.11. GWAS signal enrichment analysis

We previously reported details of the single-marker GWAS and fine-mapping analyses for the body type, reproduction, and production traits from 27,214 U.S. Holstein bulls, for health traits from 11,880–24,699 bulls, and for feed efficiency (i.e., RFI) from 3947 Holstein cows [52,67–69]. Because the complex traits being studied here are highly polygenic, we applied the sum-based marker-set test approach shown in eq. 1, as implemented in QGG package v1.0 [70], to determine whether GWAS signals were enriched in marker genes of distinct cell clusters and marker genes of AW vs. BW. We added 20-kb windows around gene regions to include the potential *cis*-regulatory variants. Previous studies showed that this approach had at least equal power when compared to other commonly used GWAS signal enrichment methods in humans [71,72], *Drosophila melanogaster* [73], and livestock [74–76], especially for the highly polygenic traits.

$$T_{sum} = \sum_{i=1}^{m_f} b^2 \quad (1)$$

In this expression,  $m_f$  is the number of genomic markers within a list of genes (marker genes of each cell cluster or marker genes of AW vs. BW in each cell cluster), and  $b$  is the marker effect from single-marker GWAS. We controlled marker-set sizes and linkage disequilibrium patterns among markers through applying the following genotype cyclical permutation strategy [70]. Briefly, we first ordered marker effects (i.e.,  $b^2$ ) using their chromosome positions (i.e.,  $b_1^2, b_2^2, \dots, b_{m-1}^2, b_m^2$ ). We then randomly selected one marker (i.e.,  $b_k^2$ ) from this vector as the first place, and shifted the remaining ones to new positions, while retaining their original orders (i.e.,  $b_k^2, b_{k+1}^2, \dots, b_{m-1}^2, b_m^2, b_1^2, \dots, b_{k-1}^2$ ) to maintain LD patterns among markers. We calculated a new summary statistic for a given list of genes using their original chromosome locations. To obtain an empirical *P*-value for the list of genes, we repeated this permutation procedure 10,000 times. We employed a one-tailed test of the proportion of random summary statistics greater than that observed.

#### 4.12. Cross-species comparison

We downloaded a single-cell RNA-seq dataset of the human stomach from GSE134355. We first merged expression matrices of the two species (cattle and humans) based on the detected homologous genes' intersection. Next, we performed expression matrix preprocessing separately for the two species using the Seurat v3 R package, followed by integrating three datasets using functions in Seurat v3 [26]. The resolution was set to 0.4 to yield 13 cell clusters.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ygeno.2021.04.039>.

#### Ethics approval and consent to participate

All samples were collected with the approval of the US Department of

Agriculture (USDA) Agriculture Research Service (ARS) Institutional Animal Care and Use Committee under Protocol 07–025. Consent to participate: not applicable.

### Consent for publication

Not applicable.

### Availability of data and material

All cell and gene expression matrix data are available as supplemental data S1. The GWAS summary statistics for all complex traits have been submitted to Figshare, i.e., body type, production, and reproduction traits under <https://figshare.com/s/ea726fa95a5bac158ac1>, and the remaining ones under <https://figshare.com/s/94540148512ddd7ed32>. All scripts and source codes can be found in the Supplemental Material, as well as in <https://github.com/YahGao/Rumen-scRNA-seq>.

### Author statement

None.

### Funding

This work was supported in part by USDA National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) grant numbers 2013-67015-20951, 2016-67015-24886, 2019-67015-29321, and 2021-67015-33409 and the US-Israel Binational Agricultural Research and Development (BARD) Fund grant number US-4997-17. This work was also supported in part by USDA ARS appropriated projects 8042-31000-001-00-D, 8042-31000-002-00-D, and 8042-31310-078-00-D.

### Authors' contributions

GEL and CJL conceived and designed the experiments. EEC, RLB, and JBC collected samples and/or generated NGS data. YG, LF, CJL, CPVT, LM, and GEL performed *in silico* prediction and computational analyses. YG, CJL, and GEL wrote the paper. All authors read and approved the final manuscript.

### Declaration of Competing Interest

All authors declare no potential conflict of interest.

### Acknowledgements

We thank Reuben Anderson, Mary Bowman, Donald Carbaugh, Christina Clover, Sarah McQueeney, Mary Niland, Marsha Campbell, Dennis Hucht, and Research Animal Services staff at the Beltsville Dairy Unit for technical assistance. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture (USDA). The USDA is an equal opportunity provider and employer.

### References

- Lin, L., Xie, F., Sun, D., Liu, J., Liu, W., Zhu, S., Mao, S., Ruminant microbiome-host crosstalk stimulates the development of the ruminal epithelium in a lamb model, *Microbiome* 7 (1) (2019) 83.
- Lin, L., Fang, X., Kang, S., Liu, M., Liu, E.E., Connor, R.L., VI Baldwin, G., Liu, C.J., Li, Establishment and transcriptomic analyses of a cattle rumen epithelial primary cells (REPC) culture by bulk and single-cell RNA sequencing to elucidate interactions of butyrate and rumen development, *Heliyon* 6 (6) (2020), e04112.
- R.L. VI Baldwin, Use of isolated ruminal epithelial cells in the study of rumen metabolism, *J. Nutr.* 28 (2) (1998) 293S–296S.
- P. Gálfi, S. Neogrady, T. Sakata, 3 - Effects of volatile fatty acids on the epithelial cell proliferation of the digestive tract and its hormonal mediation, in: T. Tsuda, Y. Sasaki, R. Kawashima (Eds.), *Physiological Aspects of Digestion and Metabolism in Ruminants*, Academic Press, San Diego, 1991, pp. 49–59.
- C. Stevens, Fatty acid transport through the rumen epithelium, *Physiol. Digest. Metab. Ruminant* (1970) 101–112.
- R.L. VI Baldwin, E.E. Connor, Rumen function and development, *Vet. Clin. N. Am. Food Anim. Pract.* 33 (3) (2017) 427–439.
- E.N. Bergman, Energy contributions of volatile fatty acids from the gastrointestinal tract in various species, *Physiol. Rev.* 70 (2) (1990) 567–590.
- A.T. Phillipson, Physiology of digestion and metabolism in the ruminant. Proceedings of the Third International Symposium, Cambridge, August 1969, in: *Physiology of digestion and metabolism in the ruminant Proceedings of the Third International Symposium*, Cambridge, August 1969, Oriel Press Ltd., 32 Ridley Place, Newcastle upon Tyne, NE1 8LH, 1970.
- R.L. VI Baldwin, S. Wu, W. Li, C. Li, B.J. Bequette, R.W. Li, Quantification of transcriptome responses of the rumen epithelium to butyrate infusion using RNA-seq technology, *Gene Regul. Syst. Biol.* 6 (2012) 67–80.
- A. Naeem, J.K. Drackley, J.S. Lanier, R.E. Everts, S.L. Rodriguez-Zas, J.J. Loor, Ruminant epithelium transcriptome dynamics in response to plane of nutrition and age in young Holstein calves, *Funct. Integr. Genomics* 14 (1) (2014) 261–273.
- R. Xiang, J. McNally, S. Rowe, A. Jonker, C.S. Pinares-Patino, V.H. Oddy, P. E. Vercoe, J.C. McEwan, B.P. Dalrymple, Gene network analysis identifies rumen epithelial cell proliferation, differentiation and metabolic pathways perturbed by diet and correlated with methane production, *Sci. Rep.* 6 (2016) 39022.
- K. Zhao, Y.H. Chen, G.B. Penner, M. Oba, L.L. Guan, Transcriptome analysis of ruminal epithelia revealed potential regulatory mechanisms involved in host adaptation to gradual high fermentable dietary transition in beef cattle, *BMC Genomics* 18 (2017) 976.
- W. Li, S. Gelsinger, A. Edwards, C. Riehle, D. Koch, Changes in meta-transcriptome of rumen epimural microbial community and liver transcriptome in young calves with feed induced acidosis, *Sci. Rep.* 9 (1) (2019) 18967.
- I. Kanter, T. Kalisky, Single cell transcriptomics: methods and applications, *Front. Oncol.* 5 (2015) 53.
- R.L. VI Baldwin, Use of isolated ruminal epithelial cells in the study of rumen metabolism, *J. Nutr.* 128 (2) (1998) 293S–296S.
- L. Fang, S. Liu, M. Liu, X. Kang, S. Lin, B. Li, E.E. Connor, R.L. VI Baldwin, A. Tenesa, L. Ma, et al., Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations, *BMC Biol.* 17 (1) (2019) 68.
- M.J. Peters, R. Joeanes, L.C. Pilling, C. Schurmann, K.N. Conneely, J. Powell, E. Reinmaa, G.L. Sutphin, A. Zhernakova, K. Schramm, et al., The transcriptional landscape of age in human peripheral blood, *Nat. Commun.* 6 (2015) 8570.
- H. Dueck, M. Khaladkar, T.K. Kim, J.M. Spaethling, C. Francis, S. Suresh, S. A. Fisher, P. Seale, S.G. Beck, T. Bartfai, et al., Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation, *Genome Biol.* 16 (2015) 122.
- F. Zhou, X. Li, W. Wang, P. Zhu, J. Zhou, W. He, M. Ding, F. Xiong, X. Zheng, Z. Li, et al., Tracing haematopoietic stem cell formation at single-cell resolution, *Nature* 533 (7604) (2016) 487–492.
- L. Li, J. Dong, L. Yan, J. Yong, X. Liu, Y. Hu, X. Fan, X. Wu, H. Guo, X. Wang, et al., Single-cell RNA-seq analysis maps development of human Germline cells and gonadal niche interactions, *Cell Stem Cell* 20 (6) (2017) 858–873 (e854).
- A.K. Shalek, R. Satija, J. Shuga, J.J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J.T. Gaublotte, N. Yosef, et al., Single-cell RNA-seq reveals dynamic paracrine control of cellular variation, *Nature* 510 (7505) (2014) 363–369.
- X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge, et al., Construction of a human cell landscape at single-cell level, *Nature* 581 (7808) (2020) 303–309.
- A.L. Haber, M. Biton, N. Rogel, R.H. Herbst, K. Shekhar, C. Smillie, G. Burgin, T. M. Delorey, M.R. Howitt, Y. Katz, et al., A single-cell survey of the small intestinal epithelium, *Nature* 551 (7680) (2017) 333–339.
- S. Gao, L. Yan, R. Wang, J. Li, J. Yong, X. Zhou, Y. Wei, X. Wu, X. Wang, X. Fan, et al., Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing, *Nat. Cell Biol.* 20 (6) (2018) 721–734.
- J. Chen, B.T. Lau, N. Andor, S.M. Grimes, C. Handy, C. Wood-Bouwens, H.P. Ji, Single-cell transcriptome analysis identifies distinct cell types and niche signaling in a primary gastric organoid model, *Sci. Rep.* 9 (1) (2019) 4536.
- T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data, *Cell* 177 (7) (2019) 1888–1902 (e1821).
- V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.: Theory Exp.* (2008) 10.
- A. Duo, M.D. Robinson, C. Soneson, A systematic performance evaluation of clustering methods for single-cell RNA-seq data, *F1000Res* 7 (2018) 1141.
- S. Freytag, L. Tian, I. Lonnstedt, M. Ng, M. Bahlo, Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data, *F1000Res* 7 (2018) 1297.
- D. Aran, A.P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R.P. Naikawadi, P. J. Wolters, A.R. Abate, et al., Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage, *Nat. Immunol.* 20 (2) (2019) 163–172.
- L.P. Chung, S. Keshav, S. Gordon, Cloning the human lysozyme cDNA: inverted Alu repeat in the mRNA and *in situ* hybridization for macrophages and Paneth cells, *Proc. Natl. Acad. Sci. U. S. A.* 85 (17) (1988) 6227–6231.

- [32] J.H. Martens, H.G. Stunnenberg, BLUEPRINT: mapping human blood cell epigenomes, *Haematologica* 98 (10) (2013) 1487–1489.
- [33] E.P. Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature* 489 (7414) (2012) 57–74.
- [34] S. Aibar, C.B. Gonzalez-Blas, T. Moerman, V.A. Huynh-Thu, H. Imrichova, G. Hulsemans, F. Rambow, J.C. Marine, P. Geurts, J. Aerts, et al., SCENIC: single-cell regulatory network inference and clustering, *Nat. Methods* 14 (11) (2017) 1083–1086.
- [35] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell trajectories, *Nat. Methods* 14 (10) (2017) 979–982.
- [36] G.X. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, et al., Massively parallel digital transcriptional profiling of single cells, *Nat. Commun.* 8 (2017) 14049.
- [37] H.G. Stunnenberg, International Human Epigenome C, Hirst M: The International Human Epigenome Consortium, A Blueprint for scientific collaboration and discovery, *Cell* 167 (5) (2016) 1145–1149.
- [38] M.S. Jackson, M. Rocchi, G. Thompson, T. Hearn, M. Crosier, J. Guy, D. Kirk, L. Mulligan, A. Ricco, S. Piccininni, et al., Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications, and unstable sequences with homologies to telomeric and other centromeric locations, *Hum. Mol. Genet.* 8 (1999) 205–215.
- [39] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics* 9 (2008) 559.
- [40] G.P. Smith, Evolution of repeated DNA sequences by unequal crossover, *Science* 191 (1976) 528–535.
- [41] A.F. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes, *Curr. Opin. Genet. Dev.* 9 (6) (1999) 657–663.
- [42] X. Pan, Y. Wang, Z. Li, X. Chen, R. Heller, N. Wang, C. Zhao, Y. Cai, H. Xu, S. Li, et al., Tracing the origin of a new organ by inferring the genetic basis of rumen evolution, *bioRxiv* (2020), <https://doi.org/10.1101/2020.02.19.955872>.
- [43] C. Schell, N. Wanner, T.B. Huber, Glomerular development—shaping the multicellular filtration unit, *Semin. Cell Dev. Biol.* 36 (2014) 39–49.
- [44] A. Smit, A. Riggs, MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation, *Nucleic Acids Res.* 23 (1995) 98–102.
- [45] R.G. LeBaron, K.I. Bezverkov, M.P. Zimber, R. Pavelec, J. Skonier, A.F. Purchio, Beta IG-H3, a novel secretory protein inducible by transforming growth factor-beta, is present in normal skin and promotes the adhesion and spreading of dermal fibroblasts in vitro, *J. Invest. Dermatol.* 104 (5) (1995) 844–849.
- [46] J.J. Bond, A.J. Donaldson, J.V.F. Coumans, K. Austin, D. Ebert, D. Wheeler, V. H. Oddy, Protein profiles of enzymatically isolated rumen epithelium in sheep fed a fibrous diet, *J. Anim. Sci. Biotechnol.* 10 (2019) 5.
- [47] O. Tatti, P. Vehviläinen, K. Lehti, J. Keski-Oja, MT1-MMP releases latent TGF-beta1 from endothelial cell extracellular matrix via proteolytic processing of LTBP-1, *Exp. Cell Res.* 314 (13) (2008) 2501–2514.
- [48] M. Smidt, I. Kirsch, L. Ratner, Deletion of Alu sequences in the fifth c-sis intron in individuals with meningiomas, *J. Clin. Investig.* 86 (4) (1990) 1151–1157.
- [49] J. Kalucka, A. Ettinger, K. Franke, S. Mamlook, R.P. Singh, K. Farhat, A. Muschter, S. Olbrich, G. Breier, D.M. Katschinski, et al., Loss of epithelial hypoxia-inducible factor prolyl hydroxylase 2 accelerates skin wound healing in mice, *Mol. Cell. Biol.* 33 (17) (2013) 3426–3438.
- [50] E.E. Connor, R.L. VI Baldwin, M.P. Walker, S.E. Ellis, C. Li, S. Kahl, H. Chung, R. W. Li, Transcriptional regulators transforming growth factor-beta1 and estrogen-related receptor-alpha identified as putative mediators of calf rumen epithelial tissue development and function during weaning, *J. Dairy Sci.* 97 (7) (2014) 4193–4207.
- [51] N.O. Ku, P. Strnad, H. Bantel, M.B. Omary, Keratins: biomarkers and modulators of apoptotic and necrotic cell death in the liver, *Hepatology* 64 (3) (2016) 966–976.
- [52] J. Jiang, J.B. Cole, E. Freebern, Y. Da, P.M. VanRaden, L. Ma, Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls, *Commun. Biol.* 2 (1) (2019) 212.
- [53] C. Perez-Stable, C.K. Shen, Competitive and cooperative functioning of the anterior and posterior promoter elements of an Alu family repeat, *Mol. Cell. Biol.* 6 (6) (1986) 2041–2052.
- [54] C. Vascotto, D. Fantini, M. Romanello, L. Cesaratto, M. Deganuto, A. Leonardi, J. P. Radicella, M.R. Kelley, C. D'Ambrósio, A. Scaloni, et al., APE1/Ref-1 interacts with NPM1 within nucleoli and plays a role in the rRNA quality control process, *Mol. Cell. Biol.* 29 (7) (2009) 1834–1854.
- [55] L. Perez-Jurado, R. Peoples, P. Kaplan, B. Hamel, U. Francke, Molecular definition of the chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth, *Am. J. Hum. Genet.* 59 (1996) 781–791.
- [56] L.M. Perelygina, N.V. Tomilin, O.I. Podgornaia, Nekotorye kharakteristiki belkov iz kletok HeLa, spetsificheski svyazyvaiushchikh ALU-posledovatel'nost' cheloveka, *Mol. Biol.* 21 (6) (1987) 1610–1619.
- [57] A. Edwards, H.A. Hammond, L. Jin, C.T. Caskey, R. Chakraborty, Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups, *Genomics* 12 (1992) 241–253.
- [58] L. Edelmann, E. Spiteri, K. Koren, V. Puljajal, M.G. Bialer, A. Shanske, R. Goldberg, B.E. Morrow, AT-rich palindromes mediate the constitutional t(11;22) translocation, *Am. J. Hum. Genet.* 68 (1) (2001) 1–13.
- [59] L. Edelmann, P. Stankiewicz, E. Spiteri, R.K. Pandita, L. Shaffer, J.R. Lupski, B. E. Morrow, Two functional copies of the DGC6 gene are present on human chromosome 22q11 due to a duplication of an ancestral locus, *Genome Res.* 11 (2) (2001) 208–217.
- [60] L. Edelmann, R.K. Pandita, B.E. Morrow, Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome, *Am. J. Hum. Genet.* 64 (4) (1999) 1076–1086.
- [61] P.M. Bingham, T.B. Chou, I. Mims, Z. Zacher, On/off regulation of gene expression at the level of splicing, *Trends Genet.* 4 (1988) 134–138.
- [62] F. Bigoni, R. Stanyon, U. Koehler, A. Morescalchi, J. Wienberg, Mapping homology between human and black and white colobine monkey chromosomes by fluorescent in situ hybridization, *Am. J. Primatol.* 42 (1997) 289–298.
- [63] B.D. Rosen, D.M. Bickhart, R.D. Schnabel, S. Koren, C.G. Elsik, E. Tseng, T. N. Rowan, W.Y. Low, A. Zimin, C. Couldrey, et al., De novo assembly of the cattle reference genome with single-molecule sequencing, *Gigascience* (2020) 9(3).
- [64] M.A. Tosches, T.M. Yamawaki, R.K. Naumann, A.A. Jacobi, G. Tushev, G. Laurent, Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles, *Science* 360 (6391) (2018) 881.
- [65] G. Yu, L.G. Wang, Y. Han, Q.Y. He, ClusterProfiler: an R package for comparing biological themes among gene clusters, *OMICS* 16 (5) (2012) 284–287.
- [66] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W. H. Fridman, F. Pages, Z. Trajanoski, J. Galon, ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, *Bioinformatics* 25 (8) (2009) 1091–1093.
- [67] B. Li, L. Fang, D.J. Null, J.L. Hutchison, E.E. Connor, P.M. VanRaden, M. J. VandeHaar, R.J. Tempelman, K.A. Weigel, J.B. Cole, High-density genome-wide association study for residual feed intake in Holstein dairy cattle, *J. Dairy Sci.* 102 (12) (2019) 11067–11080.
- [68] L. Fang, W. Cai, S. Liu, O. Canela-Xandri, Y. Gao, J. Jiang, K. Rawlik, B. Li, S. G. Schroeder, B.D. Rosen, et al., Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle, *Genome Res.* 30 (5) (2020) 790–801.
- [69] E. Freebern, D.J.A. Santos, L. Fang, J. Jiang, K.L. Parker Gaddis, G.E. Liu, P. M. VanRaden, C. Maltecca, J.B. Cole, L. Ma, GWAS and fine-mapping of livability and six disease traits in Holstein cattle, *BMC Genomics* 21 (1) (2020) 41.
- [70] P.D. Rohde, I. Fourie Sørensen, P. Sørensen, qgg: an R package for large-scale quantitative genetic analyses, *Bioinformatics* 36 (8) (2019) 2614–2615.
- [71] S. Liu, Y. Yu, S. Zhang, J.B. Cole, A. Tenesa, T. Wang, T.G. McDanel, L. Ma, G. E. Liu, L. Fang, Epigenomics and genotype-phenotype association analyses reveal conserved genetic architecture of complex traits in cattle and human, *BMC Biol.* 18 (1) (2020) 80.
- [72] P.D. Rohde, D. Demontis, B.C.D. Cuyabano, A.D. Børglum, P. Sørensen, Covariance association test (CVAT) identifies genetic markers associated with schizophrenia in functionally associated biological processes, *Genetics* 203 (4) (2016) 1901–1913.
- [73] I.F. Sørensen, S.M. Edwards, P.D. Rohde, P. Sørensen, Multiple trait covariance association test identifies gene ontology categories associated with chill coma recovery time in *Drosophila melanogaster*, *Sci. Rep.* 7 (1) (2017) 2413.
- [74] P. Sarup, J. Jensen, T. Ostensen, M. Henryron, P. Sørensen, Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs, *BMC Genet.* 17 (1) (2016) 11.
- [75] L. Fang, G. Sahana, G. Su, Y. Yu, S. Zhang, M.S. Lund, P. Sørensen, Integrating sequence-based GWAS and RNA-Seq provides novel insights into the genetic basis of mastitis and milk production in dairy cattle, *Sci. Rep.* 7 (1) (2017) 45560.
- [76] L. Fang, G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M.S. Lund, P. Sørensen, Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection, *Genet. Sel. Evol.* 49 (1) (2017) 44.