Contributed paper

EFFECT OF BOVINE REFERENCE GENOME CHOICE IN RNA-SEQ ALIGNMENT AND DIFFERENTIAL GENE EXPRESSION ANALYSIS IN BRANGUS CATTLE

E. Mantilla Valdivieso[1], E. Ross[1], A. Raza[1], B. Hayes[1], N. Jonsson[2], P. James[1] and A. Tabor[1]

[1] Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, QLD, 4072 Australia
[2] Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, Glasgow, United Kingdom

SUMMARY

New and improved assemblies for bovine genomes have been released in the past two years, contributing to the growing field of livestock genomic information, but they still require to be more comprehensively evaluated in RNA-seq bioinformatic pipelines in terms of their reproducibility in mapping and differential gene expression analysis. The present study aimed to evaluate these parameters by mapping Brangus-derived leukocyte sequence data to three bovine reference genome assemblies (Hereford, Brahman, and Angus) in order to find differentially expressed genes related to ectoparasite host resistance. We observed similar mapping rates across the three genome assemblies and a similar number of differentially expressed genes (DEGs) detected with each genome (84-86 genes). However, using haplotype-resolved genomes (Angus and Brahman) was found to be important to discover an additional 45 DEGs that could not be identified with the non–haplotype-resolved Hereford reference genome.

INTRODUCTION

High-throughput RNA sequencing technology (RNA-Seq) is currently the most powerful approach for profiling transcriptomes and identifying differentially expressed genes (DEGs) between experimental conditions (Wang *et al*. 2009). This technology is now extensively applied in the field of animal research, particularly to better understand the mechanisms responsible for genetic variation in complex phenotypes in livestock (Georges *et al.* 2019). In cattle, for instance, genetic improvement to enhance traits such as host resistance against parasites is highly desirable since the reduction of parasitic burden can improve animal welfare and increase productivity (Tabor *et al.* 2017). Ectoparasites such as the cattle tick (*Rhipicephalus microplus* species complex) represent a major animal health challenge for the cattle industry; thus, finding effective ways to control tick infestations is a priority for producers.

One of the most feasible options to protect cattle herds from ticks is through the use of tick-resistant breeds which have *Bos indicus* genetics, as *Bos taurus* breeds are mostly susceptible (Utech *et al.* 1978). Crossbred cattle (*B. indicus x B. taurus*), such as Brangus, have more desirable meat quality than purebred *Bos indicus* but exhibit a range of tick-resistant and susceptible phenotypes. On top of this, targeting host resistance for genetic improvement is challenging because the underlying biological mechanisms are not yet fully understood (Tabor *et al.* 2017). Previous work suggests that variation in immune gene expression can contribute to the variation in the phenotype (Piper *et al*. 2010). Therefore, it is hypothesised that biomarker discovery by differential gene expression analysis could provide feasible opportunities for selecting for tick-resistant hosts in cattle with *Bos taurus* content.

The accurate quantification of gene expression heavily relies on the availability of high-quality genomes and their corresponding annotations (Oshlack *et al.* 2010). Currently, the *Bos taurus* ARS_UCD1.2 assembly (Rosen *et al.* 2020), which originated from an inbred Hereford animal, is widely accepted as the reference genome for taurine and indicine cattle. However, Low *et al*. (2020) released two novel reference-quality assemblies UOA_Angus_1 and UOA_Brahman_1 from Angus

and Brahman parental haplotypes of an F1 *B. taurus x B. indicus* hybrid (Brangus), which provides the opportunity to further study breed-specific gene expression patterns that could be related to the expression of host resistance. Therefore, this study aimed to investigate if the choice of bovine reference genome (Hereford, Angus, and Brahman) may affect the mapping rate of short-read sequencing data and produce substantial differences in downstream differential gene expression analysis in circulating leukocytes from Brangus cattle of high and low resistance to tick infestation.

MATERIALS AND METHODS

Animals. 30 Brangus steers (~9 months old) without previous exposure to ticks were recruited for this study conducted under animal ethics approval (QAAFI/469/18). The animals were exposed to artificial infestation with approximately 10,000 tick larvae (*R. australis*) over 12 weeks, during which animals were ranked for their resistance to infestation and blood samples were collected. The number of developing adult ticks after an infestation cycle (21 days) was estimated with a tick scoring scale from 1 (<50 ticks = Resistant) to 5 (>300 ticks = Susceptible). The animals subsequently classified as the most resistant (R, n=3), and most susceptible (S, n=5) hosts were selected for RNA sequencing of leukocytes isolated from blood collected immediately before primary infestation.

RNA extraction and sequencing. RNA was extracted from frozen leukocytes in Qiazol reagent with the miRNeasy mini kit (QIAGEN, USA) as per manufacturer's instructions. RNA samples were treated with DNAse and RNA was quantified using the Nanodrop 2000 (Thermofisher, USA). The RNA RIN quality analysis was evaluated with the 2100 Bioanalyzer Instrument (Agilent Technologies, USA). The cDNA libraries were prepared with the TruSeq Stranded mRNA kit and sequenced as 100 bp single-end reads in one flow cell lane on the Illumina NovaSeq 6000 sequencer (Illumina, USA) through the Australian Genome Research Facility.

Bioinformatics pipeline. The RNA-Seq pipeline for this study is shown in Figure 1. Briefly, read quality control was performed with FastQC v.0.11.4 (Andrews 2015) and adapters were removed with Trimmomatic v.0.35 (Bolger *et al.* 2014). The reads were mapped with STAR .2.5.2b (Dobin *et al.* 2012) to the ARS-UCD1.2 (Rosen *et al.* 2020), UOA_Angus_1, and UOA_Brahman_1 (Low *et al.* 2020) assemblies. Genomes and annotations were sourced from the Ensembl Release 102 (https://asia.ensembl.org). The gene count matrices were processed in RStudio with the edgeR Bioconductor package (Robinson e*t al.* 2009). A generalized linear model was fitted to test phenotype (S vs. R) as the main factor with sample RIN number as a covariate. Differentially expressed genes (DEGs) were considered significant based on a false discovery rate (FDR) < 0.05 and $|\log_2$ (fold change)| >1.

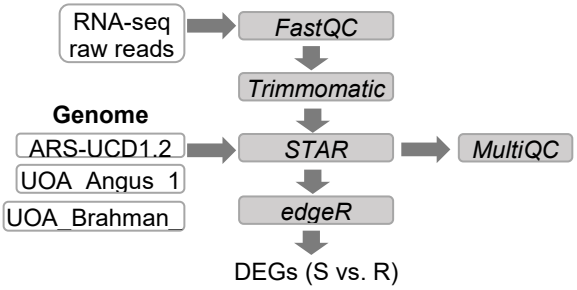

Figure 1. RNA-Seq pipeline for differential gene expression in leukocytes of tick-susceptible (S) compared to resistant (R) Brangus cattle pre-infestation

RESULTS AND DISCUSSION

RNA-Seq mapping. The sequencing produced an average of 36.2 million raw single-end reads

per sample. After the adapter trimming and QC steps, the average number of reads per sample was 35.5 million. The percentage of reads that uniquely mapped to the ARS_UCD1.2 genome was higher by approximately 4% and 6% compared to the UOA_Brahman_1 and UOA_Angus_1 genomes, respectively (Table 1). Additionally, the percentage of multi-mapped reads was between 5.1 and 7.1 across all three genomes, but a larger proportion (4%) of unmapped reads was obtained with the Angus genome. Therefore, for this Brangus-derived transcriptomic dataset, mapping rates were consistently high with all three genomes, but the performance of the STAR aligner improved slightly when using the Hereford assembly.

Differential gene expression. In total, 131 DEGs were identified in the circulatory leukocytes from tick-resistant compared to tick-susceptible Brangus with all three bovine reference genomes (Figure 2). Of these genes, 51 (38.9%) were commonly identified by all three genomes, 47 (35.9%) were common to the taurine genomes (ARS_UCD1.2 and UOA_Angus_1), and 19 (14.5%) were unique to the indicine genome (UOA_Brahman_1). Overall, mapping our sequencing data to the haplotype-resolved reference genomes was useful to identify an additional 45 DEGs that otherwise could not have been identified by the ARS_UCD1.2 genome alone; however, many of these genes did not have full annotations. This result further highlights the need for an improved gene annotation pipeline for both the UOA_Brahman_1 and UOA_Angus_1 assemblies, particularly to be able to characterise indicine-derived DEGs and their relevance in conferring host resistance against ticks.

Moreover, it was found that choice of reference genome did not significantly alter the total number of genes that were differentially expressed in the two phenotypes of host resistance (susceptible vs. resistant), but the number of up- and down-regulated genes varied slightly for each reference genome (Table 1).

Table 1. RNA-seq mapping results (%) for three bovine reference genomes and the resulting number of differentially expressed genes (DEGs) in leukocytes from tick-susceptible compared to tick-resistant Brangus cattle

	Hereford ARS_UCD1.2	Angus UOA_Angus_1	Brahman UOA_Brahman_1
Uniquely mapped reads	94.07	88.63	90.43
Multi-mapped reads	5.15	6.56	7.07
Unmapped reads	0.42	4.46	2.14
Total no. of DEGs	86	84	84
Up-regulated	33	26	20
Down-regulated	53	58	64

CONCLUSIONS

Continuous improvement to the cattle reference genome has led to the latest release of the *B. taurus* ARS_UCD1.2 assembly. Although this is generally considered a high-quality assembly, it is based on an inbred taurine animal and does not hold the potential to characterise all the variation that exists in other cattle subspecies, i.e. *B.t. indicus, B.t. africanus*, and crosses thereof (Low *et al.* 2020). The UOA_Brahman/Angus_1 haplotype-resolved genomes provide an opportunity to address these concerns, but they have not been extensively tested in RNA-Seq bioinformatic pipelines. This study explored how the choice of reference genome input can influence short-read mapping and differential gene expression in leukocyte transcriptomic data from crossbred Brangus cattle.

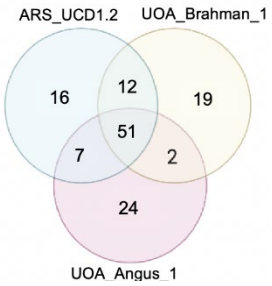

Figure 2. Venn diagram showing the number of unique and overlapping DEGs (tick-susceptible vs. -resistant Brangus) detected from three bovine reference genomes.

It was found that the choice of bovine reference genome had a mild effect on read mapping and different gene expression detection, likely reflecting on the very high-quality of all three genomes. Importantly, using haplotype-resolved genomes allowed the detection of additional DEGs that appeared to be specific to the indicine and taurine components of the Brangus breed (an Angus and Brahman cross). However, many of these genes are yet to be fully annotated, thus, further gene overlap could still be expected in addition to 51/131 DEGs discovered with all three genomes, once gene annotations pipelines improve. Further work on characterising which unannotated up- and down-regulated DEGs are homologous to other ARS_UCD1.2 sequences or orthologous to other species (human or rat) will be the first step towards elucidating these novel genes and potentially shed light on the biological mechanisms underlying tick host resistance. Ultimately, testing a variety of high-quality genome resources in well-established bioinformatic pipelines such as RNA-Seq can greatly improve interpretations from transcriptomic data, particularly if the end goal is discovering biomarkers that can assist for genetic improvement of a wider range of cattle breeds.

ACKNOWLEDGMENTS

Meat and livestock Australia (MLA) Donor Company project funding P.PSH.0798, and The University of Queensland RTP Scholarship for funding E. Mantilla's candidature.

REFERENCES

Andrews, S. (2015) *FastQC A quality control tool for high throughput sequence data.*; Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bolger, A.M., Lohse, M., Usadel, B. (2014) *Bioinformatics* **30**(15): 2114.

Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., and Gingeras T.R. (2012). *Bioinformatics* **29**:15.

Georges M., Charlier C., and Hayes B. (2019). *Nat Rev Genet* **20**: 135.

Low W.Y., Tearle R., Liu R., Koren S., Rhie A., Bickhart D.M., Rosen B.D., Kronenberg Z.N., Kingan S.B., Tseng E., *et al.* (2020) *Nat Commun* **11**: 2071.

Oshlack A., Robinson M.D., and Young M.D. (2010) *Genome Biol* **11**: 220.

Piper E.K., Jackson L.A., Bielefeldt-Ohmann H., Gondro C., Lew-Tabor A.E., and Jonsson N.N. (2010) *Int J Parasitol* **40**: 431.

Tabor A.E., Ali A., Rehman G., Garcia G., Zangirolamo A., Malardo T., and Jonsson N.N. (2017) *Front Cell Infect Microbiol* **7**: 1.

Robinson M.D., McCarthy D.J., and Smyth, G.K. (2009) *Bioinformatics* **26**: 139.

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G., Tseng E., Rowan T.N., Low W.Y., Zimin A., Couldrey C., *et al.* (2020) *Gigascience* **9**.

Utech K.B., Wharton R.H., and Kerr J.D. (1978) *Aus J Agric Res* **29**: 885.

Wang Z., Gerstein M., and Snyder M. (2009). *Nat Rev Genet* **10**: 57.