RNA sequencing (RNA-seq) reveals extremely low levels of reticulocyte-derived globin gene transcripts in peripheral blood from horses (*Equus caballus*) and cattle (*Bos taurus*)

Carolina N. Correia[1], Kirsten E. McLoughlin[1], Nicolas C. Nalpas[1, †], David A. Magee[1], John A. Browne[1], Kevin Rue-Albrecht[1, §], Stephen V. Gordon[2, 3], David E. MacHugh[1, 3]*

[1] Animal Genomics Laboratory, UCD School of Agriculture and Food Science, UCD College of Health and Agricultural Sciences, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[2] UCD School of Veterinary Medicine, UCD College of Health and Agricultural Sciences, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[3] UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[†] Current address: Quantitative Proteomics and Proteome Centre Tübingen, Interfaculty Institute for Cell Biology, University of Tübingen, Tübingen, 72076, Germany.

[§] Current address: Kennedy Institute of Rheumatology, University of Oxford, Oxford OX3 7FY, United Kingdom.

*** Correspondence:** David E. MacHugh, Animal Genomics Laboratory, UCD School of Agriculture and Food Science, UCD College of Health and Agricultural Sciences, University College Dublin, Belfield, Dublin D04 V1W8, Ireland. Email: david.machugh@ucd.ie

Keywords: biomarker, blood, cattle, globin, horses, pigs, reticulocyte, RNA-seq, transcriptome

Note: British English language style preferred for publication.

Abstract

RNA-seq has emerged as an important technology for measuring gene expression in peripheral blood samples collected from humans and other vertebrate species. In particular, transcriptomics analyses of whole blood can be used to study immunobiology and develop novel biomarkers of infectious disease. However, an obstacle to these methods in many mammalian species is the presence of reticulocyte-derived globin mRNAs in large quantities, which can complicate RNA-seq library sequencing and impede detection of other mRNA transcripts. A range of supplementary procedures for targeted depletion of globin transcripts have, therefore, been developed to alleviate this problem. Here, we use comparative analyses of RNA-seq data sets generated from human, porcine, equine and bovine peripheral blood to systematically assess the impact of globin mRNA on routine transcriptome profiling of whole blood in cattle and horses. The results of these analyses demonstrate that total RNA isolated from equine and bovine peripheral blood contains very low levels of globin mRNA transcripts, thereby negating the need for globin depletion and greatly simplifying blood-based transcriptomic studies in these two domestic species.

1. Introduction

It is increasingly recognised that new technological approaches are urgently required for infectious disease diagnosis, surveillance and management in burgeoning domestic animal populations as livestock production intensifies across the globe (Thornton, 2010; Nabarro and Wannous, 2014; Animal Task Force, 2016). In this regard, new strategies have emerged that leverage peripheral blood gene expression to study host immunobiology and to identify panels of RNA transcript biomarkers that can be used as specific biosignatures of infection by particular pathogens for both animal and human infectious disease (Ramilo and Mejias, 2009; Mejias and Ramilo, 2014; Chaussabel, 2015; Ko et al., 2015; Holcomb et al., 2017). For example, we and others have applied this approach to bovine tuberculosis (BTB) caused by infection with *Mycobacterium bovis* (Meade et al., 2007; Killick et al., 2011; Blanco et al., 2012; Churbanov and

Milligan, 2012; McLoughlin et al., 2014; Cheng et al., 2015). It is also important to note that peripheral blood transcriptomics using technologies such as microarrays or RNA-sequencing (RNA-seq) can be used to monitor changes in the physiological status of domestic animals due to reproductive status, diet and nutrition or stress (O'Loughlin et al., 2012; Takahashi et al., 2012; Song et al., 2013; Kolli et al., 2014; Shen et al., 2014; de Greeff et al., 2016; Elgendy et al., 2016; Jegou et al., 2016).

During the last 15 years, a major hindrance to whole blood transcriptomics studies has emerged, which is the presence of large quantities of globin mRNA transcripts in peripheral blood from many mammalian species (Wu et al., 2003; Fan and Hegde, 2005; Liu et al., 2006). This is a consequence of abundant α globin and β globin mRNA transcripts in circulating reticulocytes, which in humans, may account for more than 95% of the total cellular mRNA content in these immature erythrocytes (Debey et al., 2004). Reticulocytes, in turn, account for 1–4% of the erythrocytes in healthy adult humans, which corresponds to between 5×10^7 and 2×10^8 cells per ml compared to 7×10^6 cells per ml for leukocytes (Greer et al., 2013). Hence, globin transcripts can account for a substantial proportion of total detectable mRNAs in peripheral blood samples collected from humans and many other mammals (Bruder et al., 2010; Winn et al., 2010; Schwochow et al., 2012; Choi et al., 2014; Shin et al., 2014; Bowyer et al., 2015; Huang et al., 2016; Morey et al., 2016). In particular, for humans, more than 70% of peripheral blood mRNA transcripts are derived from the haemoglobin subunit alpha 1, subunit alpha 2 and subunit beta genes (*HBA1*, *HBA2* and *HBB*) (Wu et al., 2003; Field et al., 2007; Mastrokolias et al., 2012).

The emergence of massively parallel transcriptome profiling for clinical applications in human peripheral blood—initially with gene expression microarrays, but more recently using RNA-seq—has prompted development of methods for the systematic reduction of globin mRNAs in total RNA samples purified from peripheral blood samples, including: oligonucleotides that bind to globin mRNA molecules with subsequent digestion of the RNA strand of the RNA:DNA hybrid (Wu et al.,

2003); peptide nucleic acid (PNA) oligonucleotides that are complementary to globin mRNAs and block reverse transcription of these targets (Liu et al., 2006); the GLOBINclear™ system, which uses biotinylated oligonucleotides that hybridise with globin transcripts followed by capture and separation using streptavidin-coated magnetic beads (Field et al., 2007); and the recently introduced GlobinLock method that uses a pair of modified oligonucleotides complementary to the 3' portion of globin transcripts and that block enzymatic extension (Krjutskov et al., 2016).

In the present study we use RNA-seq data generated from globin-depleted and non-depleted total RNA purified from human and porcine peripheral blood, in conjunction with non-depleted total RNA isolated from equine and bovine peripheral blood, for a comparative investigation of the impact of reticulocyte-derived globin mRNA transcripts on routine transcriptome profiling of blood in domestic cattle and horses. The primary objective of the present study to test the hypothesis that both cattle and horses exhibit significantly lower quantities of haemoglobin gene transcripts compared to humans and pigs.

2. Materials and Methods

2.1. Data sources

RNA-seq data sets from human peripheral whole blood samples used for assessment of globin depletion and with parallel non-depleted controls (Shin et al., 2014) were obtained from the NCBI Gene Expression Omnibus (GEO) database (accession number GSE53655). A comparable RNA-seq data set from globin-depleted and non-depleted porcine peripheral whole blood was obtained directly from the study authors (Choi et al., 2014). A published RNA-seq data set (Ropka-Molik et al., 2017) from equine non-depleted peripheral whole blood was obtained from the NCBI GEO database (accession number GSE83404). Finally, bovine RNA-seq data from peripheral whole blood were generated by us as described below and can be obtained from the European Nucleotide Archive (ENA) database (accession number to be determined). A summary overview of the methodology used for the current study is shown in **Figure 1**.

2.2. Human, porcine and equine sample collection, globin depletion and RNA-seq libraries

Detailed information concerning ethics approval, sample collection, total RNA extraction, and RNA-seq library preparation and sequencing for the human, porcine, and equine data sets is provided in the original publications (Choi et al., 2014; Shin et al., 2014; Ropka-Molik et al., 2017). Supplementary Table 1 provides summary information on the human, porcine and equine samples and RNA-seq libraries.

In brief, for the human samples, peripheral blood from six healthy subjects (three females and three males) was collected into PAXgene blood RNA tubes (PreAnalytiX/Qiagen Ltd., Manchester, UK). Total RNA, including small RNAs, was purified from the collected blood samples using the PAXgene Blood miRNA Kit (PreAnalytiX/Qiagen Ltd.) as described by Shin et al. (Shin et al., 2014). Human *HBA1*, *HBA2* and *HBB* mRNA transcripts were depleted from a subset of the total RNA samples using the GLOBINclear kit (Invitrogen™/Thermo Fisher Scientific, Loughborough, UK). RNA-seq data was then generated using 24 paired-end (PE) RNA-seq libraries (12 undepleted and 12 globin-depleted) generated from the six biological replicates and six identical technical replicates created from pooled total RNA across all six donor samples. The multiplexing and sequencing was then performed such that data for the 12 samples in each treatment group (undepleted and globin depleted) was generated from two separate lanes of a single flow cell twice, for a total of four sequencing lanes (Shin et al., 2014).

Porcine peripheral blood samples were collected from 12 healthy crossbred pigs (Duroc × (Landrace × Yorkshire)) using Tempus™ blood RNA tubes (Applied Biosystems™/Thermo Fisher Scientific, Warrington, UK) and total RNA was purified using the MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit (Invitrogen™/Thermo Fisher Scientific) (Choi et al., 2014). Porcine *HBA* and *HBB* mRNA transcripts were subsequently depleted from a subset of the total RNA samples using a modified RNase H globin depletion method with custom porcine-specific antisense

oligonucleotides for *HBA* and *HBB*. RNA-seq data was then generated from 24 PE RNA-seq libraries (12 undepleted and 12 globin-depleted).

Equine peripheral blood samples were collected using Tempus™ blood RNA tubes from 12 healthy Arabian horses (five females and seven males) at three different time points during flat racing training (Ropka-Molik et al., 2017). In addition, peripheral blood samples were collected from six healthy untrained Arabian horses (two females and four males). Total RNA was purified using the MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit and 37 of the 42 total RNA samples were used to generate single-end (SE) libraries for RNA-seq data generation. Globin depletion for the equine samples was not performed prior to RNA-seq library preparation (Katarzyna Ropka-Molik, pers. comm.).

2.3. Bovine peripheral blood collection and RNA extraction

Approximately 3 ml of peripheral blood from ten age-matched healthy male Holstein-Friesian calves were collected into Tempus™ blood RNA tubes. The Tempus™ Spin RNA Isolation Kit (Applied Biosystems™/Thermo Fisher Scientific) was used to perform total RNA extraction and purification, following the manufacturer's instructions. RNA quantity and quality checking were performed using a NanoDrop™ 1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and an Agilent 2100 Bioanalyzer using an RNA 6000 Nano LabChip kit (Agilent Technologies Ltd., Cork, Ireland). The majority of samples displayed a 260/280 ratio greater than 1.8 and an RNA integrity number (RIN) greater than 8.0 (Supplementary Table 2). Globin mRNA depletion was not performed on the total RNA samples purified from bovine peripheral blood samples.

2.4. Bovine RNA-seq library generation and sequencing

Individually barcoded strand-specific RNA-seq libraries were prepared with 1 μg of total RNA from each sample. Two rounds of poly(A)$^+$ RNA purification were performed for all RNA samples

using the Dynabeads® mRNA DIRECT™ Micro Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. The purified poly(A)$^+$ RNA was then used to generate strand-specific RNA-seq libraries using the ScriptSeq™ v2 RNA-Seq Library Preparation Kit, the ScriptSeq™ Index PCR Primers (Sets 1 to 4) and the FailSafe™ PCR enzyme system (all sourced from Epicentre®/Illumina® Inc., Madison, WI, USA), according to the manufacturer's instructions.

RNA-seq libraries were purified using the Agencourt® AMPure® XP system (Beckman Coulter Genomics, Danvers, MA, USA) according to the manufacturer's instructions for double size selection (0.75× followed by 1.0× ratio). RNA-seq libraries were quantified using a Qubit® fluorometer and Qubit® dsDNA HS Assay Kit (Invitrogen™/Thermo Fisher Scientific), while library quality checks were performed using an Agilent 2100 Bioanalyzer and High Sensitivity DNA Kit (Agilent Technologies Ltd.). Individually barcoded RNA-seq libraries were pooled in equimolar quantities and the quantity and quality of the final pooled libraries (three pools in total) were assessed as described above. Cluster generation and high-throughput sequencing of three pooled RNA-seq libraries were performed using an Illumina® HiSeq™ 2000 Sequencing System at the MSU Research Technology Support Facility (RTSF) Genomics Core (https://rtsf.natsci.msu.edu/genomics; Michigan State University, MI, USA). Each of the three pooled libraries were sequenced independently on five lanes split across multiple Illumina® flow cells. The pooled libraries were sequenced as PE 2 × 100 nucleotide reads using Illumina® version 5.0 sequencing kits.

Deconvolution (filtering and segregation of sequence reads based on the unique RNA-seq library barcode index sequences; Supplementary Table 2) was performed by the MSU RTSF Genomics Core using a pipeline that simultaneously demultiplexed and converted pooled sequence reads into discrete FASTQ files for each RNA-seq sample with no barcode index mismatches permitted. The RNA-seq FASTQ sequence read data for the bovine samples were obtained from the MSU RTSF Genomics Core FTP server.

2.5. RNA-seq data quality control and filtering/trimming of reads

Bioinformatics procedures and analyses were performed as described below for the human, porcine, equine, and bovine samples, except were specifically indicated. All of the bioinformatics workflow scripts were developed using GNU bash (version 4.3.48) (Free Software Foundation, 2013), Python (version 3.5.2) (Python Software Foundation, 2017), and R (version 3.4.0) (R Core Team, 2017). The scripts and further information are available at a public GitHub repository (https://github.com/carolcorreia/Globin_RNA-sequencing). Computational analyses were performed on a 32-core Linux Compute Server (4× AMD Opteron™ 6220 processors at 3.0 GHz with 8 cores each), with 256 GB of RAM, 24 TB of hard disk drive storage, and with Ubuntu Linux OS (version 14.04.4 LTS). Deconvoluted FASTQ files (generated from SE equine RNA-seq libraries and PE RNA-seq libraries for the other species) were quality-checked with FastQC (version 0.11.5) (Andrews, 2016).

Using the ngsShoRT software package (version 2.2) (Chen et al., 2014), filtering/trimming consisted of: (1) removal of SE or PE reads with adapter sequences (with up to three mismatches); (2) removal of SE or PE reads of poor quality (i.e., at least one of the reads containing ≥ 25% bases with a Phred quality score below 20); (3) for porcine samples only, 10 bases were trimmed at the 3' end of all reads; (4) removal of SE or PE reads that did not meet the required minimum length (70 nucleotides for human and equine, 80 nucleotides for porcine and 100 nucleotides for bovine). Filtered/trimmed FASTQ files were then re-evaluated using FastQC. Filtered FASTQ files were transferred to a 36-core/64-thread Compute Server (2× Intel® Xeon® CPU E5-2697 v4 at 2.30 GHz with 18 cores each), with 512 GB of RAM, 96 TB SAS storage (12× 8 TB at 7200 rpm), 480 GB SSD storage, and with Ubuntu Linux OS (version 16.04.2 LTS).

2.6. Transcript quantification

The Salmon software package (version 0.8.2) (Patro et al., 2017) was used in quasi-mapping-mode for transcript quantification. Sequence-specific and fragment-level GC bias correction was enabled and transcript abundance was quantified in transcripts per million (TPM) for each filtered library (multiple lanes from the same library were processed together) was estimated after mapping of SE or PE reads to their respective reference transcriptomes. As summarised in **Table 1**, the NCBI RefSeq database is currently the only one to contain haemoglobin gene annotations for all species analysed. Hence, NCBI RefSeq reference transcript models were used for the human, porcine, equine, and bovine data sets. Detailed information about these reference transcriptomes is provided in Supplementary Table 3.

2.7. Gene annotations and summarisation of TPM estimates at the gene level

Using R (3.5.0) within the RStudio IDE (version 1.1.447) (RStudio Team, 2015) and Bioconductor (version 3.7 using BiocInstaller 1.30.0) (Gentleman et al., 2004), the GenomicFeatures (version 1.32.0) (Lawrence et al., 2013) and AnnotationDbi (version 1.42.1) (Pagès et al., 2017) packages were used to obtain corresponding gene and transcript identifiers from the NCBI RefSeq annotation releases pertinent to each species, as detailed in **Table 1**. Using these identifiers, the tximport (version 1.8.0) package (Soneson et al., 2015) was used to import into R and summarise at gene level the TPM estimates obtained from the Salmon tool. A threshold of greater than or equal to 1 TPM across at least half of the total number of samples (≥ 12 for human and porcine, ≥ 18 for equine, and ≥ 5 for bovine) was applied in order to remove lowly expressed genes.

2.8. Data exploration, plotting and summary statistics

Data wrangling and tidying from all species was performed using the following R packages: tidyverse (version 1.2.1) (Wickham, 2017b), dplyr (version 0.7.5) (Wickham et al., 2017), tidyr (version 0.8.1) (Wickham and Henry, 2017), reshape2 (version 1.4.3) (Wickham, 2017a), and magrittr (version 1.5) (Bache and Wickham, 2017). The ggplot2 (version 2.2.1) (Wickham and

Chang, 2017), and ggjoy (version 0.4.1) (Wilke, 2017), packages were used for figure generation. Finally, the mean and standard deviation were calculated for the undepleted and globin-depleted groups in each species using the skimr (version 1.0.2) R package (McNamara et al., 2017).

3. Results and Discussion

3.1. Status of human, porcine, equine and bovine haemoglobin gene annotations

Annotation of the haemoglobin subunit alpha 1 and 2 genes (*HBA1* and *HBA2*, respectively) is well established for the human genome; however, annotations for these genes in the porcine, equine and bovine genomes are inconsistent across databases. As shown in **Table 1**, the porcine *HBA* gene annotation is absent from Ensembl and the UCSC Table Browser. For the NCBI RefSeq database, this gene has been assigned to two loci (*LOC110259958* and *LOC100737768*) that have similar descriptions (haemoglobin subunit alpha and haemoglobin subunit alpha-like). Therefore, these NCBI LOC symbols were used.

Equine *HBA* (*HBA1*) and *HBA2* genes are absent from the current Ensembl annotation release. Similarly, bovine *HBA1* and *HBA* (*HBA2*) have been annotated as *GLNC1* in Ensembl, whereas *HBA1* is absent from the UCSC Table Browser annotation (**Table 1**). In the NCBI RefSeq database, equine *HBA* (*HBA1*) is described as haemoglobin subunit alpha 1; and bovine *HBA* (*HBA2*) is described as haemoglobin subunit alpha 2, thus their descriptions are shown in parenthesis herein. In contrast to these observations, haemoglobin subunit beta (*HBB*) genes for the four species are well annotated in Ensembl, NCBI RefSeq and UCSC Genome Browser databases (**Table 1**).

At the time of writing, NCBI RefSeq is the only database that contains annotations for all three haemoglobin genes in all species analysed. Additionally, equine and bovine gene annotations are based on the latest genome assemblies (**Table 1**). EquCab3 and ARS-UCD1.2 have incorporated major improvements compared to previous versions, including increased genome coverage (from

6.8× and 9×, to 80×, respectively), and incorporation of PacBio sequencing reads (Kalbfleisch et al., 2018; Rosen et al., 2018).

3.2. Basic RNA-seq data outputs

Unfiltered SE (equine libraries) or PE (human, porcine, and bovine libraries) RNA-seq FASTQ files were quality-checked, adapter- and quality-filtered prior to transcript quantification. As shown in **Table 2**, the human and porcine undepleted groups each had approximately 40 million (M) raw reads per library, whereas globin-depleted libraries showed a mean of approximately 37 M and 31 M, respectively. Equine and bovine libraries, which did not include a globin depletion step had an average of 24 M raw reads and 21 M raw read pairs, respectively.

After adapter- and quality-filtering of RNA-seq libraries, an average of 20% and 29% read pairs were removed from the human undepleted and globin-depleted libraries, respectively. Conversely, approximately 12% of read pairs were removed from each of the porcine undepleted and globin-depleted libraries. For the undepleted equine and bovine RNA-seq libraries, an average of 0.2% reads and 17% read pairs were removed, respectively. Detailed information on filtering/trimming of RNA-seq libraries from all species, including technical replicates from libraries sequenced over multiple lanes, is presented in Supplementary Table 4. All data sets exhibited a mean mapping rate greater than 70% (**Table 2**). Supplementary Tables 5 contain sample-specific RNA-seq mapping statistics.

3.3. Transcript quantification

Transcript-level TPM estimates generated using the Salmon tool were imported into the R environment and summarised at gene level with the package tximport (Soneson et al., 2015). Gene-

level TPM estimates represent the sum of corresponding transcript-level TPMs and provide results that are more accurate and comprehensible than transcript-level estimates (Soneson et al., 2015). In the current study, gene-level TPM estimates are referred as TPM.

Filtering of lowly expressed genes (see **Section 2.7**) resulted in 12,951 genes expressed across all human samples, and represented 24% of 54,644 total annotated genes and pseudogenes. Porcine samples showed a total of 9,396 expressed genes (31% of 30,334 annotated genes and pseudogenes); and equine and bovine samples exhibited 12,724 (38% of 33,146) and 14,044 (40% of 35,143) expressed genes, respectively.

The density distribution of TPM values for the human and porcine samples improved after globin depletion; this is evident by the shift of gene detection levels towards greater $\log_{10}$ TPM values for the globin-depleted samples in **Figure 2**. In this regard, it is noteworthy that the undepleted bovine and equine samples also exhibited similar TPM density distributions to the human and porcine globin-depleted samples.

3.4. Proportions of human and porcine haemoglobin gene transcripts in undepleted and depleted peripheral blood

In line with previous reports (Field et al., 2007; Mastrokolias et al., 2012), the proportion of haemoglobin gene transcripts (*HBA1*, *HBA2*, and *HBB*) detected in undepleted human peripheral blood samples for the current study averaged 70% (**Figure 3** and Supplementary Table 6), which is lower than the mean proportion of 81% reported by Shin et al. (2014). On the other hand, after depletion the human samples exhibited an identical reduction to a 17% proportion of globin sequence reads in both the present study and that of Shin et al. (2014) (**Figure 3** and Supplementary Table 6).

In the current study, for the undepleted porcine peripheral blood samples, the percentage of haemoglobin gene transcripts (*LOC110259958* [*HBA*], *LOC100737768* [*HBA*], and *HBB*) observed

as a proportion of the total expressed genes was 72% (**Figure 3** and Supplementary Table 6), which is considerably larger than the mean of 46.1% reported in the original study (Choi et al., 2014). Similarly, after depletion, the porcine samples in the present study contained a mean proportion of 22% globin transcripts (**Figure 3** and Supplementary Table 6) compared to a mean proportion of 8.9% reported by Choi et al. (2014). Additionally, **Table 3** shows the mean TPM for each haemoglobin gene across undepleted or globin-depleted samples.

A number of possible explanations, including the different approaches used for read mapping and transcript quantification, may account for the different proportions of haemoglobin gene transcript detected in human and porcine samples for the present study compared to the original studies (Choi et al., 2014; Shin et al., 2014). For the present study, a recently developed lightweight alignment method was adopted (Salmon and tximport), in contrast to the more traditional methodologies used in the original publications. Shin and colleagues (2014) used the TopHat and Cufflinks software tools (Trapnell et al., 2012), while Choi et al. (2014) implemented TopHat with Htseq-count (Anders et al., 2015). In addition to this, different gene annotations were used: NCBI *Homo sapiens* Annotation Release 109 and NCBI *Sus scrofa* Annotation Release 106 were used for the present study, while UCSC hg18 (*Homo sapiens*) and Ensembl release 71 (*Sus scrofa*) were used by Shin et al. (2014) and Choi et al. (2014), respectively.

3.5. Equine and bovine peripheral blood contains extremely low levels of haemoglobin gene transcripts

The equine and bovine peripheral blood samples, which did not undergo globin depletion, had extremely low proportions of haemoglobin gene transcripts to total expressed genes: 0.21% and 0.17%, respectively (**Figure 3** and Supplementary Table 6). Notably, similar results have been reported in a transcriptomics study of bovine peripheral blood in response to vaccination against neonatal pancytopenia. In that study, 12 cows were profiled before and after vaccination (24 peripheral blood samples in total), and a mean proportion of 1.0% of RNA-seq reads were observed

to map to the bovine α haemoglobin gene cluster on BTA25 or to the β haemoglobin gene cluster on BTA15 (Demasius et al., 2013). To the best of our knowledge, this is the first time that the average number of equine haemoglobin transcripts have been reported for RNA-seq data.

Finally, it is important to note that $\log_2$ TPM values for haemoglobin gene transcripts in the undepleted equine and bovine peripheral blood RNA samples are substantially lower than $\log_2$ TPM values for the globin-depleted human and porcine peripheral blood RNA samples (**Figure 4**). This is a direct consequence of extremely low levels of circulating reticulocytes in equine and bovine peripheral blood (Tablin and Weiss, 1985; Harper et al., 1994; Hossain et al., 2003; Cooper et al., 2005).

3.6. Conclusion

In light of our RNA-seq data analyses, we propose that globin mRNA transcript depletion is not a pre-requisite for transcriptome profiling of bovine and equine peripheral blood samples. This observation greatly simplifies the laboratory and bioinformatics workflows required for RNA-seq studies of whole blood collected from domestic cattle and horses. It will also be directly relevant to future work on blood-based biomarker and biosignature development in the context of infectious disease, reproduction, nutrition and animal welfare. For example, transcriptomics of peripheral blood has been used extensively in development of new diagnostic and prognostic modalities for human tuberculosis (HTB) disease caused by infection with *Mycobacterium tuberculosis* (for reviews see: Blankley et al., 2014; Haas et al., 2016; Weiner and Kaufmann, 2017; Goletti et al., 2018). Therefore, as a consequence of this HTB research, comparable transcriptomics studies in cattle (Meade et al., 2007; Killick et al., 2011; Blanco et al., 2012; Churbanov and Milligan, 2012; McLoughlin et al., 2014; Cheng et al., 2015), and the ease with which RNA-seq can be performed in bovine peripheral blood, it should be feasible to develop transcriptomics-based biomarkers and biosignatures for bovine tuberculosis caused by *M. bovis* infection.

4. Data Accessibility

The RNA-seq data generated for this study using peripheral blood from ten age-matched healthy male Holstein-Friesian calves can be obtained from the ENA database (accession number to be determined).

5. Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6. Ethics Statement

Animal experimental work for the present study (cattle samples) was carried out according to the UK Animal (Scientific Procedures) Act 1986. The study protocol was approved by the Animal Health and Veterinary Laboratories Agency (AHVLA–Weybridge, UK), now the Animal & Plant Health Agency (APHA), Animal Use Ethics Committee (UK Home Office PCD number 70/6905).

7. Author Contributions

DEM, SVG, CNC, and KEM conceived and designed the project and organised bovine sample collection; KEM, NCN, DAM, and JAB performed RNA extraction and RNA-seq library generation; CNC, KEM, NCN, KRA, and DEM performed the analyses; CNC and DEM wrote the manuscript and all authors reviewed and approved the final manuscript.

8. Funding

This work was supported by Investigator Grants from Science Foundation Ireland (Nos: SFI/08/IN.1/B2038 and SFI/15/IA/3154), a Research Stimulus Grant from the Department of Agriculture, Food and the Marine (No: RSF 06 405), a European Union Framework 7 Project Grant (No: KBBE-211602- MACROSYS), a Brazilian Science Without Borders – CAPES grant (No: BEX-13070-13-4) and the UCD Wellcome Trust funded Computational Infection Biology PhD Programme (Grant no: 097429/Z/11/Z).

9. Acknowledgements

The authors wish to express their gratitude to Prof Martin Vordermeier and Dr Bernardo Villarreal-Ramos (Animal and Plant Health Agency, UK) for provision of bovine peripheral blood samples, Prof Graham Plastow (University of Alberta, Canada) for provision of porcine peripheral blood RNA-seq data. We also thank Drs Gabriella Farries (University College Dublin) and Kerri Malone (EMBL-EBI, Cambridge, UK) for stimulating discussion and advice concerning genome annotations, equine genetics and data visualisation.

10. References

Anders, S, Pyl, PT, and Huber, W (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31. doi: 10.1093/bioinformatics/btu638.

Andrews, S. 2016. FastQC: a quality control tool for high throughput sequence data. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Animal Task Force. 2016. A strategic research and innovation agenda for a sustainable livestock sector in Europe. Available: http://www.animaltaskforce.eu.

Bache, SM, and Wickham, H. 2017. magrittr: a forward-pipe operator for R. Available: https://github.com/tidyverse/magrittr.

Blanco, FC, Soria, M, Bianco, MV, and Bigi, F (2012). Transcriptional response of peripheral blood mononuclear cells from cattle infected with *Mycobacterium bovis*. *PLoS ONE* 7**,** e41066. doi: 10.1371/journal.pone.0041066.

Blankley, S, Berry, MP, Graham, CM, Bloom, CI, Lipman, M, and O'Garra, A (2014). The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis. *Philos Trans R Soc Lond B Biol Sci* 369**,** 20130427. doi: 10.1098/rstb.2013.0427.

Bowyer, JF, Tranter, KM, Hanig, JP, Crabtree, NM, Schleimer, RP, and George, NI (2015). Evaluating the stability of RNA-seq transcriptome profiles and drug-induced immune-related expression changes in whole blood. *PLoS ONE* 10**,** e0133315. doi: 10.1371/journal.pone.0133315.

Bruder, CE, Yao, S, Larson, F, Camp, JV, Tapp, R, McBrayer, A, Powers, N, Granda, WV, and Jonsson, CB (2010). Transcriptome sequencing and development of an expression microarray platform for the domestic ferret. *BMC Genomics* 11**,** 251. doi: 10.1186/1471-2164-11-251.

Chaussabel, D (2015). Assessment of immune status using blood transcriptomics and potential implications for global health. *Semin Immunol* 27**,** 58-66. doi: 10.1016/j.smim.2015.03.002.

Chen, C, Khaleel, SS, Huang, H, and Wu, CH (2014). Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol Med* 9**,** 8. doi: 10.1186/1751-0473-9-8.

Cheng, Y, Chou, C-H, and Tsai, H-J (2015). In vitro gene expression profile of bovine peripheral blood mononuclear cells in early *Mycobacterium bovis* infection. *Exp Ther Med* 10**,** 2102-2118. doi: 10.3892/etm.2015.2814.

Choi, I, Bao, H, Kommadath, A, Hosseini, A, Sun, X, Meng, Y, Stothard, P, Plastow, GS, Tuggle, CK, Reecy, JM, Fritz-Waters, E, Abrams, SM, Lunney, JK, and Guan le, L (2014).

Increasing gene discovery and coverage using RNA-seq of globin RNA reduced porcine blood samples. *BMC Genomics* 15**,** 954. doi: 10.1186/1471-2164-15-954.

Churbanov, A, and Milligan, B (2012). Accurate diagnostics for bovine tuberculosis based on high-throughput sequencing. *PLoS ONE* 7**,** e50147. doi: 10.1371/journal.pone.0050147.

Cooper, C, Sears, W, and Bienzle, D (2005). Reticulocyte changes after experimental anemia and erythropoietin treatment of horses. *J Appl Physiol* 99**,** 915-921. doi: 10.1152/japplphysiol.00438.2005.

de Greeff, A, Bikker, P, Smit-Heinsbroek, A, Bruininx, E, Zwolschen, H, Fijten, H, Zetteler, P, Vastenhouw, S, Smits, M, and Rebel, J (2016). Increased fat and polyunsaturated fatty acid content in sow gestation diet has no effect on gene expression in progeny during the first 7 days of life. *J Anim Physiol Anim Nutr (Berl)* 100**,** 127-135. doi: 10.1111/jpn.12345.

Debey, S, Schoenbeck, U, Hellmich, M, Gathof, BS, Pillai, R, Zander, T, and Schultze, JL (2004). Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. *Pharmacogenomics J* 4**,** 193-207. doi: 10.1038/sj.tpj.6500240.

Demasius, W, Weikard, R, Hadlich, F, Muller, KE, and Kuhn, C (2013). Monitoring the immune response to vaccination with an inactivated vaccine associated to bovine neonatal pancytopenia by deep sequencing transcriptome analysis in cattle. *Vet Res* 44**,** 93. doi: 10.1186/1297-9716-44-93.

Elgendy, R, Giantin, M, Castellani, F, Grotta, L, Palazzo, F, Dacasto, M, and Martino, G (2016). Transcriptomic signature of high dietary organic selenium supplementation in sheep: A nutrigenomic insight using a custom microarray platform and gene set enrichment analysis. *J Anim Sci* 94**,** 3169-3184. doi: 10.2527/jas.2016-0363.

Fan, H, and Hegde, PS (2005). The transcriptome in blood: challenges and solutions for robust expression profiling. *Curr Mol Med* 5**,** 3-10. doi: 10.2174/1566524053152861.

Field, LA, Jordan, RM, Hadix, JA, Dunn, MA, Shriver, CD, Ellsworth, RE, and Ellsworth, DL (2007). Functional identity of genes detectable in expression profiling assays following globin mRNA reduction of peripheral blood samples. *Clin Biochem* 40**,** 499-502. doi: 10.1016/j.clinbiochem.2007.01.004.

Free Software Foundation. 2013. Bash (4.3.48) Unix shell program. Available: http://ftp.gnu.org/gnu/bash/.

Gentleman, RC, Carey, VJ, Bates, DM, Bolstad, B, Dettling, M, Dudoit, S, Ellis, B, Gautier, L, Ge, Y, Gentry, J, Hornik, K, Hothorn, T, Huber, W, Iacus, S, Irizarry, R, Leisch, F, Li, C, Maechler, M, Rossini, AJ, Sawitzki, G, Smith, C, Smyth, G, Tierney, L, Yang, JY, and

Zhang, J (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5**,** R80. doi: 10.1186/gb-2004-5-10-r80.

Goletti, D, Lee, MR, Wang, JY, Walter, N, and Ottenhoff, THM (2018). Update on tuberculosis biomarkers: From correlates of risk, to correlates of active disease and of cure from disease. *Respirology*. doi: 10.1111/resp.13272.

Greer, JP, Arber, DA, Glader, B, List, AF, Means, RT, Paraskevas, F, Rodgers, GM, and Foerster, J (2013). *Wintrobe's Clinical Hematology*. Philadelphia, PA, USA: Lippincott, Williams and Wilkins.

Haas, CT, Roe, JK, Pollara, G, Mehta, M, and Noursadeghi, M (2016). Diagnostic 'omics' for active tuberculosis. *BMC Med* 14**,** 37. doi: 10.1186/s12916-016-0583-9.

Harper, SB, Hurst, WJ, Ohlsson-Wilhelm, B, and Lang, CM (1994). The response of various hematologic parameters in the young bovine subjected to multiple phlebotomies. *ASAIO J* 40**,** M816-825.

Holcomb, ZE, Tsalik, EL, Woods, CW, and McClain, MT (2017). Host-based peripheral blood gene expression analysis for diagnosis of infectious diseases. *J Clin Microbiol* 55**,** 360-368. doi: 10.1128/jcm.01057-16.

Hossain, MA, Yamato, O, Yamasaki, M, Otsuka, Y, and Maede, Y (2003). Relation between reticulocyte count and characteristics of erythrocyte 5'-nucleotidase in dogs, cats, cattle and humans. *J Vet Med Sci* 65**,** 193-197. doi: 10.1292/jvms.65.193.

Huang, Z, Gallot, A, Lao, NT, Puechmaille, SJ, Foley, NM, Jebb, D, Bekaert, M, and Teeling, EC (2016). A nonlethal sampling method to obtain, generate and assemble whole blood transcriptomes from small, wild mammals. *Mol Ecol Resour* 16**,** 150-162. doi: 10.1111/1755-0998.12447.

Jegou, M, Gondret, F, Vincent, A, Trefeu, C, Gilbert, H, and Louveau, I (2016). Whole blood transcriptomics is relevant to identify molecular changes in response to genetic selection for feed efficiency and nutritional status in the pig. *PLoS ONE* 11**,** e0146550. doi: 10.1371/journal.pone.0146550.

Kalbfleisch, TS, Rice, E, DePriest, MS, Walenz, BP, Hestand, MS, Vermeesch, JR, O'Connell, BL, Fiddes, IT, Vershinina, AO, Petersen, JL, Finno, CJ, Bellone, RR, McCue, ME, Brooks, SA, Bailey, E, Orlando, L, Green, RE, Miller, DC, Antczak, DF, and MacLeod, JN (2018). EquCab3, an updated reference genome for the domestic horse. *bioRxiv 306928*. doi: 10.1101/306928.

Karolchik, D, Hinrichs, AS, Furey, TS, Roskin, KM, Sugnet, CW, Haussler, D, and Kent, WJ (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32**,** D493-496. doi: 10.1093/nar/gkh103.

Kent, WJ, Sugnet, CW, Furey, TS, Roskin, KM, Pringle, TH, Zahler, AM, and Haussler, D (2002). The human genome browser at UCSC. *Genome Res* 12**,** 996-1006. doi: 10.1101/gr.229102.

Killick, KE, Browne, JA, Park, SD, Magee, DA, Martin, I, Meade, KG, Gordon, SV, Gormley, E, O'Farrelly, C, Hokamp, K, and MacHugh, DE (2011). Genome-wide transcriptional profiling of peripheral blood leukocytes from cattle infected with *Mycobacterium bovis* reveals suppression of host immune genes. *BMC Genomics* 12**,** 611. doi: 10.1186/1471-2164-12-611.

Ko, ER, Yang, WE, McClain, MT, Woods, CW, Ginsburg, GS, and Tsalik, EL (2015). What was old is new again: using the host response to diagnose infectious disease. *Expert Rev Mol Diagn* 15**,** 1143-1158. doi: 10.1586/14737159.2015.1059278.

Kolli, V, Upadhyay, RC, and Singh, D (2014). Peripheral blood leukocytes transcriptomic signature highlights the altered metabolic pathways by heat stress in zebu cattle. *Res Vet Sci* 96**,** 102-110. doi: 10.1016/j.rvsc.2013.11.019.

Krjutskov, K, Koel, M, Roost, AM, Katayama, S, Einarsdottir, E, Jouhilahti, EM, Soderhall, C, Jaakma, U, Plaas, M, Vesterlund, L, Lohi, H, Salumets, A, and Kere, J (2016). Globin mRNA reduction for whole-blood transcriptome sequencing. *Sci Rep* 6**,** 31584. doi: 10.1038/srep31584.

Lawrence, M, Huber, W, Pages, H, Aboyoun, P, Carlson, M, Gentleman, R, Morgan, MT, and Carey, VJ (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol* 9**,** e1003118. doi: 10.1371/journal.pcbi.1003118.

Liu, J, Walter, E, Stenger, D, and Thach, D (2006). Effects of globin mRNA reduction methods on gene expression profiles from whole blood. *J Mol Diagn* 8**,** 551-558. doi: 10.2353/jmoldx.2006.060021.

Mastrokolias, A, den Dunnen, JT, van Ommen, GB, t Hoen, PA, and van Roon-Mom, WM (2012). Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood RNA. *BMC Genomics* 13**,** 28. doi: 10.1186/1471-2164-13-28.

McLoughlin, KE, Nalpas, NC, Rue-Albrecht, K, Browne, JA, Magee, DA, Killick, KE, Park, SD, Hokamp, K, Meade, KG, O'Farrelly, C, Gormley, E, Gordon, SV, and MacHugh, DE (2014). RNA-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with *Mycobacterium bovis*. *Front Immunol* 5**,** 396. doi: 10.3389/fimmu.2014.00396.

McNamara, A, Arino de la Rubia, E, Zhu, H, Lowndes, J, Ellis, S, Waring, E, Quinn, M, McLeod, H, Wickham, H, and Müller, K. 2017. skimr: compact and flexible summaries of data. Available: https://github.com/ropenscilabs/skimr.

Meade, KG, Gormley, E, Doyle, MB, Fitzsimons, T, O'Farrelly, C, Costello, E, Keane, J, Zhao, Y, and MacHugh, DE (2007). Innate gene repression associated with *Mycobacterium bovis* infection in cattle: toward a gene signature of disease. *BMC Genomics* 8**,** 400. doi: 10.1186/1471-2164-8-400.

Mejias, A, and Ramilo, O (2014). Transcriptional profiling in infectious diseases: ready for prime time? *J Infect* 68 Suppl 1**,** S94-99. doi: 10.1016/j.jinf.2013.09.018.

Morey, JS, Neely, MG, Lunardi, D, Anderson, PE, Schwacke, LH, Campbell, M, and Van Dolah, FM (2016). RNA-seq analysis of seasonal and individual variation in blood transcriptomes of healthy managed bottlenose dolphins. *BMC Genomics* 17**,** 720. doi: 10.1186/s12864-016-3020-8.

Nabarro, D, and Wannous, C (2014). The potential contribution of Iivestock to food and nutrition security: the application of the One Health approach in livestock policy and practice. *Rev Sci Tech* 33**,** 475-485. doi: 10.20506/rst.33.2.2292.

O'Leary, NA, Wright, MW, Brister, JR, Ciufo, S, Haddad, D, McVeigh, R, Rajput, B, Robbertse, B, Smith-White, B, Ako-Adjei, D, Astashyn, A, Badretdin, A, Bao, Y, Blinkova, O, Brover, V, Chetvernin, V, Choi, J, Cox, E, Ermolaeva, O, Farrell, CM, Goldfarb, T, Gupta, T, Haft, D, Hatcher, E, Hlavina, W, Joardar, VS, Kodali, VK, Li, W, Maglott, D, Masterson, P, McGarvey, KM, Murphy, MR, O'Neill, K, Pujar, S, Rangwala, SH, Rausch, D, Riddick, LD, Schoch, C, Shkeda, A, Storz, SS, Sun, H, Thibaud-Nissen, F, Tolstoy, I, Tully, RE, Vatsan, AR, Wallin, C, Webb, D, Wu, W, Landrum, MJ, Kimchi, A, Tatusova, T, DiCuccio, M, Kitts, P, Murphy, TD, and Pruitt, KD (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44**,** D733-745. doi: 10.1093/nar/gkv1189.

O'Loughlin, A, Lynn, DJ, McGee, M, Doyle, S, McCabe, M, and Earley, B (2012). Transcriptomic analysis of the stress response to weaning at housing in bovine leukocytes using RNA-seq technology. *BMC Genomics* 13**,** 250. doi: 10.1186/1471-2164-13-250.

Pagès, H, Carlson, M, Falcon, S, and Li, N. 2017. AnnotationDbi: Annotation Database Interface. R package version 1.38.0. Available: https://doi.org/doi:10.18129/B9.bioc.AnnotationDbi.

Patro, R, Duggal, G, Love, MI, Irizarry, RA, and Kingsford, C (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14**,** 417-419. doi: 10.1038/nmeth.4197.

Python Software Foundation. 2017. Python (version 3.5.2). Available: https://www.python.org/downloads/release/python-352/.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. Available: https://www.R-project.org.

Ramilo, O, and Mejias, A (2009). Shifting the paradigm: host gene signatures for diagnosis of infectious diseases. *Cell Host Microbe* 6**,** 199-200. doi: 10.1016/j.chom.2009.08.007.

Ropka-Molik, K, Stefaniuk-Szmukier, M, Zukowski, K, Piorkowska, K, Gurgul, A, and Bugno-Poniewierska, M (2017). Transcriptome profiling of Arabian horse blood during training regimens. *BMC Genet* 18**,** 31. doi: 10.1186/s12863-017-0499-1.

Rosen, BD, Bickhart, DM, Schnabel, RD, Koren, S, Elsik, CG, Zimin, A, Dreischer, C, Schultheiss, S, Hall, R, Schroeder, SG, Van Tassell, CP, Smith, TPL, and Medrano, JF (Year). "Modernizing the bovine reference genome assembly", in: *Proceedings of the World Congress on Genetics Applied to Livestock Production*), 802.

RStudio Team. 2015. RStudio: Integrated Development for R. Available: http://www.rstudio.com.

Schwochow, D, Serieys, LE, Wayne, RK, and Thalmann, O (2012). Efficient recovery of whole blood RNA--a comparison of commercial RNA extraction protocols for high-throughput applications in wildlife species. *BMC Biotechnol* 12**,** 33. doi: 10.1186/1472-6750-12-33.

Shen, J, Zhou, C, Zhu, S, Shi, W, Hu, M, Fu, X, Wang, C, Wang, Y, Zhang, Q, and Yu, Y (2014). Comparative transcriptome analysis reveals early pregnancy-specific genes expressed in peripheral blood of pregnant sows. *PLoS ONE* 9**,** e114036. doi: 10.1371/journal.pone.0114036.

Shin, H, Shannon, CP, Fishbane, N, Ruan, J, Zhou, M, Balshaw, R, Wilson-McManus, JE, Ng, RT, McManus, BM, and Tebbutt, SJ (2014). Variation in RNA-seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PLoS ONE* 9**,** e91041. doi: 10.1371/journal.pone.0091041.

Soneson, C, Love, MI, and Robinson, MD (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* 4**,** 1521. doi: 10.12688/f1000research.7563.2.

Song, KD, Dowd, SE, Lee, HK, and Kim, SW (2013). Long-term dietary supplementation of organic selenium modulates gene expression profiles in leukocytes of adult pigs. *Anim Sci J* 84**,** 238-246. doi: 10.1111/j.1740-0929.2012.01060.x.

Tablin, F, and Weiss, L (1985). Equine bone marrow: a quantitative analysis of erythroid maturation. *Anat Rec* 213**,** 202-206. doi: 10.1002/ar.1092130212.

Takahashi, J, Waki, S, Matsumoto, R, Odake, J, Miyaji, T, Tottori, J, Iwanaga, T, and Iwahashi, H (2012). Oligonucleotide microarray analysis of dietary-induced hyperlipidemia gene expression profiles in miniature pigs. *PLoS ONE* 7**,** e37581. doi: 10.1371/journal.pone.0037581.

Thornton, PK (2010). Livestock production: recent trends, future prospects. *Philos Trans R Soc Lond B Biol Sci* 365**,** 2853-2867. doi: 10.1098/rstb.2010.0134.

Trapnell, C, Roberts, A, Goff, L, Pertea, G, Kim, D, Kelley, DR, Pimentel, H, Salzberg, SL, Rinn, JL, and Pachter, L (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7**,** 562-578. doi: 10.1038/nprot.2012.016.

Tyner, C, Barber, GP, Casper, J, Clawson, H, Diekhans, M, Eisenhart, C, Fischer, CM, Gibson, D, Gonzalez, JN, Guruvadoo, L, Haeussler, M, Heitner, S, Hinrichs, AS, Karolchik, D, Lee, BT, Lee, CM, Nejad, P, Raney, BJ, Rosenbloom, KR, Speir, ML, Villarreal, C, Vivian, J, Zweig, AS, Haussler, D, Kuhn, RM, and Kent, WJ (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* 45**,** D626-d634. doi: 10.1093/nar/gkw1134.

Weiner, J, and Kaufmann, SH (2017). High-throughput and computational approaches for diagnostic and prognostic host tuberculosis biomarkers. *Int J Infect Dis* 56**,** 258-262. doi: 10.1016/j.ijid.2016.10.017.

Wickham, H. 2017a. reshape2: flexibly reshape data: a reboot of the reshape package. Available: https://github.com/hadley/reshape.

Wickham, H. 2017b. tidyverse. Available: http://tidyverse.tidyverse.org.

Wickham, H, and Chang, W. 2017. ggplot2: create elegant data visualisations using the grammar of graphics. Available: http://ggplot2.tidyverse.org.

Wickham, H, Francois, R, Henry, L, and Müller, K. 2017. dplyr: a grammar of data manipulation. Available: http://dplyr.tidyverse.org.

Wickham, H, and Henry, L. 2017. tidyr: easily tidy data with 'spread()' and 'gather()' functions. Available: http://tidyr.tidyverse.org.

Wilke, CO. 2017. ggjoy: joyplots in 'ggplot2'. Available: https://github.com/clauswilke/ggjoy.

Winn, ME, Zapala, MA, Hovatta, I, Risbrough, VB, Lillie, E, and Schork, NJ (2010). The effects of globin on microarray-based gene expression analysis of mouse blood. *Mamm Genome* 21**,** 268-275. doi: 10.1007/s00335-010-9261-y.

Wu, K, Miyada, G, Martin, J, and Finkelstein, D. 2003. Technical note: globin reduction protocol: a method for processing whole blood RNA samples for improved array results. Available: https://tools.thermofisher.com/content/sfs/brochures/blood2_technote.pdf.

Zerbino, DR, Achuthan, P, Akanni, W, Amode, MR, Barrell, D, Bhai, J, Billis, K, Cummins, C, Gall, A, Giron, CG, Gil, L, Gordon, L, Haggerty, L, Haskell, E, Hourlier, T, Izuogu, OG, Janacek, SH, Juettemann, T, To, JK, Laird, MR, Lavidas, I, Liu, Z, Loveland, JE, Maurel, T, McLaren, W, Moore, B, Mudge, J, Murphy, DN, Newman, V, Nuhn, M, Ogeh, D, Ong, CK, Parker, A, Patricio, M, Riat, HS, Schuilenburg, H, Sheppard, D, Sparrow, H, Taylor, K, Thormann, A, Vullo, A, Walts, B, Zadissa, A, Frankish, A, Hunt, SE, Kostadima, M, Langridge, N, Martin, FJ, Muffato, M, Perry, E, Ruffier, M, Staines, DM, Trevanion, SJ, Aken, BL, Cunningham, F, Yates, A, and Flicek, P (2018). Ensembl 2018. *Nucleic Acids Res* 46**,** D754-D761. doi: 10.1093/nar/gkx1098.

11. Figures

FIGURE 1 | Schematic of the bioinformatics workflow for RNA-seq data acquisition, quality control, analysis and interpretation.

FIGURE 2 | Ridge plots showing density of sample gene-level transcripts per million (TPM). Results are shown from undepleted (purple) or globin-depleted (green) treatments.

FIGURE 3 | Average proportions of haemoglobin genes to total expressed genes from peripheral blood RNA-seq data in humans, pigs, horses and cattle.

FIGURE 4 | Distribution of haemoglobin gene-level transcripts per million (TPM). Results are shown from undepleted (purple) or globin-depleted (green) treatments.

12. Tables

Table 1: Status of current human, porcine, equine and bovine haemoglobin gene annotations in the Ensembl, NCBI RefSeq, and UCSC databases.

Homo sapiens			
	Ensembl	**NCBI RefSeq**	**UCSC Table Browser**
Annotation release	Human release 92 (April 2018)[1]	NCBI *Homo sapiens* Annotation Release 109 (March 2018)[2]	hg38.refGene annotation track (last updated on May 2018)[3, 4, 5]
Genome assembly used to derive annotation	GRCh38.p12, GCA_000001405.27, December 2017	GRCh38.p12, GCF_000001405.38, December 2017	GRCh38, GCF_000001405.15, December 2013
HBA1	Annotated with gene ID ENSG00000206172	Annotated with Entrez Gene ID 3039	Annotated with Entrez Gene ID 3039
HBA2	Annotated with gene ID ENSG00000188536	Annotated with Entrez Gene ID 3040	Annotated with Entrez Gene ID 3040
HBB	Annotated with gene ID ENSG00000244734	Annotated with Entrez Gene ID 3043	Annotated with Entrez Gene ID 3043
Sus scrofa			
	Ensembl	**NCBI RefSeq**	**UCSC Table Browser**
Annotation release	Pig release 92 (April 2018)[1]	NCBI *Sus scrofa* Annotation Release 106 (May 2017)[2]	susScr3.refGene annotation track (last updated on May 2018)[4, 5]
Genome assembly used to derive annotation	Sscrofa11.1, GCA_000003025.6, February 2017	Sscrofa11.1, GCF_000003025.6, February 2017	Sscrofa11.1, GCF_000003025.6, February 2017
***LOC110259958* (*HBA*)**	Absent from current annotation release, accessible via online search with ID 110259958.1	Annotated with Entrez Gene ID 110259958	Absent
***LOC100737768* (*HBA*)**	Absent from current annotation release, accessible via online search with ID 100737768.1	Annotated with Entrez Gene ID 100737768	Absent

HBB	Annotated with gene ID ENSSSCG00000014725	Annotated with Entrez Gene ID 407066	Annotated with Entrez Gene ID 407066
Equus caballus			
	Ensembl	**NCBI RefSeq**	**UCSC Table Browser**
Annotation release	Horse release 92 (April 2018)[1]	NCBI *Equus caballus* Annotation Release 103 (January 2018)[2]	equCab2.refGene annotation track (last updated on May 2018)[4, 5]
Genome assembly used to derive annotation	EquCab 2, GCA_000002305.1, September 2007	EquCab3, GCF_002863925.1, May 2018	EquCab 2, GCA_000002305.1, September 2007
***HBA* (also known as *HBA1*)**	Absent	Annotated with Entrez Gene ID 100036557	Annotated with Entrez Gene ID 100036557
HBA2	Absent	Annotated with Entrez Gene ID 100036558	Annotated with Entrez Gene ID 100036558
HBB	Annotated with gene ID ENSECAG00000010020	Annotated with Entrez Gene ID 100054109	Annotated with Entrez Gene ID 100054109
Bos taurus			
	Ensembl	**NCBI RefSeq**	**UCSC Table Browser**
Annotation release	Cow release 92 (April 2018)[1]	NCBI *Bos taurus* Annotation Release 106 (May 2018)[2]	bosTau8.refGene annotation track (last updated on May 2018)[4, 5]
Genome assembly used to derive annotation	UMD3.1, GCA_000003055.3, November 2009	ARS-UCD1.2, GCF_002263795.1, April 2018	UMD3.1.1, GCA_000003055.4, June 2014
HBA1	Annotated as *GLNC1* with ID ENSBTAG00000026417	Annotated with Entrez gene ID 100140149	Absent
***HBA* (also known as *HBA2*)**	Annotated as *GLNC1* with ID ENSBTAG00000026418	Annotated with Entrez Gene ID 512439	Annotated with Entrez Gene ID 512439

HBB	Annotated with gene ID ENSBTAG00000038748	Annotated with Entrez Gene ID 280813	Annotated with Entrez Gene ID 280813

[1] (Zerbino et al., 2018); [2] (O'Leary et al., 2016); [3] (Kent et al., 2002); [4] (Karolchik et al., 2004); [5] (Tyner et al., 2017); [6].

Table 2: Summary of RNA-seq filtering/trimming and mapping statistics.

Species	Treatment	RNA-seq library type	Sequencing mode	Mean no. of reads (SE) or pairs (PE)	Mean no. of reads (SE) or pairs (PE) removed	Mean proportion of reads (SE) or pairs (PE) removed	Mean no. of observed fragments*	Mean no. of mapped fragments*	Average mapping rate	Reference source
Homo sapiens	Undepleted	Inward unstranded	PE	40,218,886	8,203,217	20.4%	32,015,669	25,593,239	80.9%	NCBI RefSeq
Homo sapiens	Globin depleted	Inward unstranded	PE	36,874,759	10,704,088	29.0%	26,170,671	20,371,624	75.8%	NCBI RefSeq
Sus scrofa	Undepleted	Inward unstranded	PE	39,036,515	4,613,991	11.8%	28,685,437	25,129,140	87.7%	NCBI RefSeq
Sus scrofa	Globin depleted	Inward unstranded	PE	31,339,886	3,899,427	12.4%	22,867,049	19,959,995	87.3%	NCBI RefSeq
Equus caballus	Undepleted	Unstranded	SE	24,271,141	38,892	0.2%	14,850,797	11,387,774	76.5%	NCBI RefSeq
Bos taurus	Undepleted	Inward stranded forward	PE	20,495,983	3,474,597	17.0%	17,021,386	12,353,147	72.6%	NCBI RefSeq

*The Salmon tool categorises fragments as single read (for SE RNA-seq libraries) or a read pair (for PE RNA-seq libraries).

Table 3: Summary statistics for haemoglobin gene-level transcripts per million (TPM).

Species	Gene symbol	Treatment	No. of samples	Mean TPM	Standard deviation
Homo sapiens	*HBA1*	Undepleted	12	191,209	16,601
Homo sapiens	*HBA1*	Globin depleted	12	66,718	23,557
Homo sapiens	*HBA2*	Undepleted	12	300,000	29,523
Homo sapiens	*HBA2*	Globin depleted	12	79,818	31,259
Homo sapiens	*HBB*	Undepleted	12	200,000	43,262
Homo sapiens	*HBB*	Globin depleted	12	20,770	6,706
Sus scrofa	*LOC110259958* (*HBA*)	Undepleted	12	86	30
Sus scrofa	*LOC110259958* (*HBA*)	Globin depleted	12	13	13
Sus scrofa	*LOC100737768* (*HBA*)	Undepleted	12	243,864	31,605
Sus scrofa	*LOC100737768* (*HBA*)	Globin depleted	12	84,021	86,095
Sus scrofa	*HBB*	Undepleted	12	476,284	52,939
Sus scrofa	*HBB*	Globin depleted	12	136,172	128,232
Equus caballus	*HBA* (*HBA1*)	Undepleted	37	443	560
Equus caballus	*HBA2*	Undepleted	37	653	789
Equus caballus	*HBB*	Undepleted	37	1,024	1,144
Bos taurus	*HBA1*	Undepleted	10	21	29
Bos taurus	*HBA* (*HBA2*)	Undepleted	10	1,101	1,102
Bos taurus	*HBB*	Undepleted	10	532	469

Figure 1

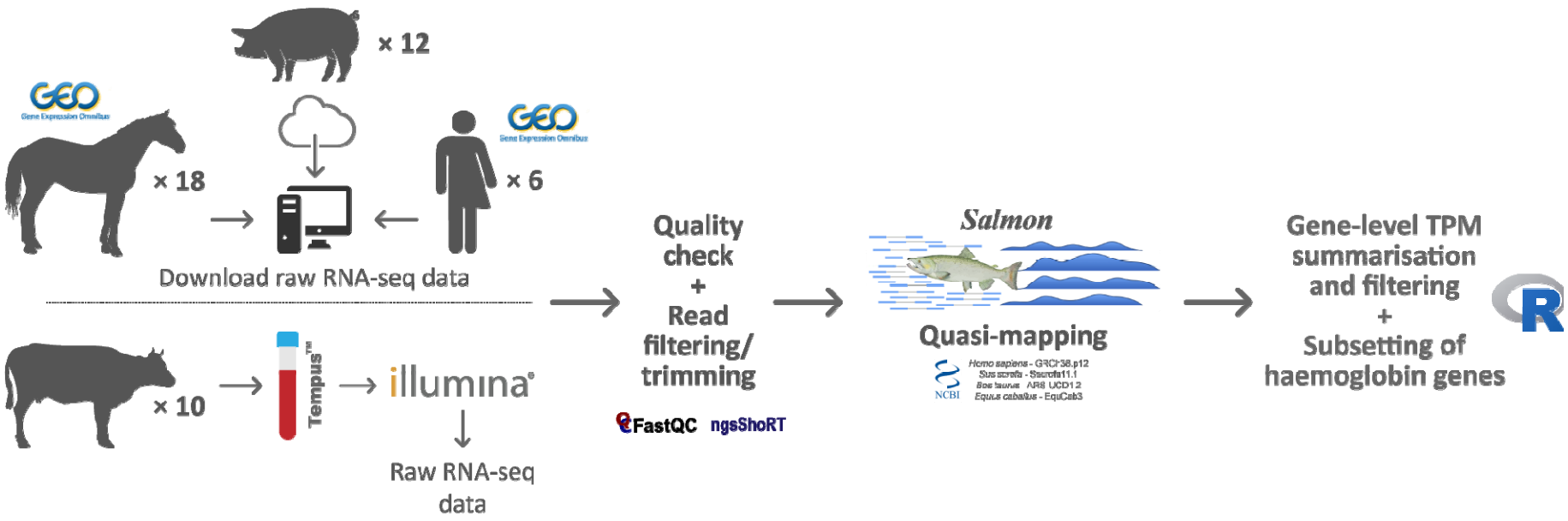

× 12
× 18
× 6
Download raw RNA-seq data
× 10
Tempus™
illumina
Raw RNA-seq data
Quality check + Read filtering/ trimming
FastQC
ngsShoRT
Salmon
Quasi-mapping
NCBI
Gene-level TPM summarisation and filtering + Subsetting of haemoglobin genes

Figure 2

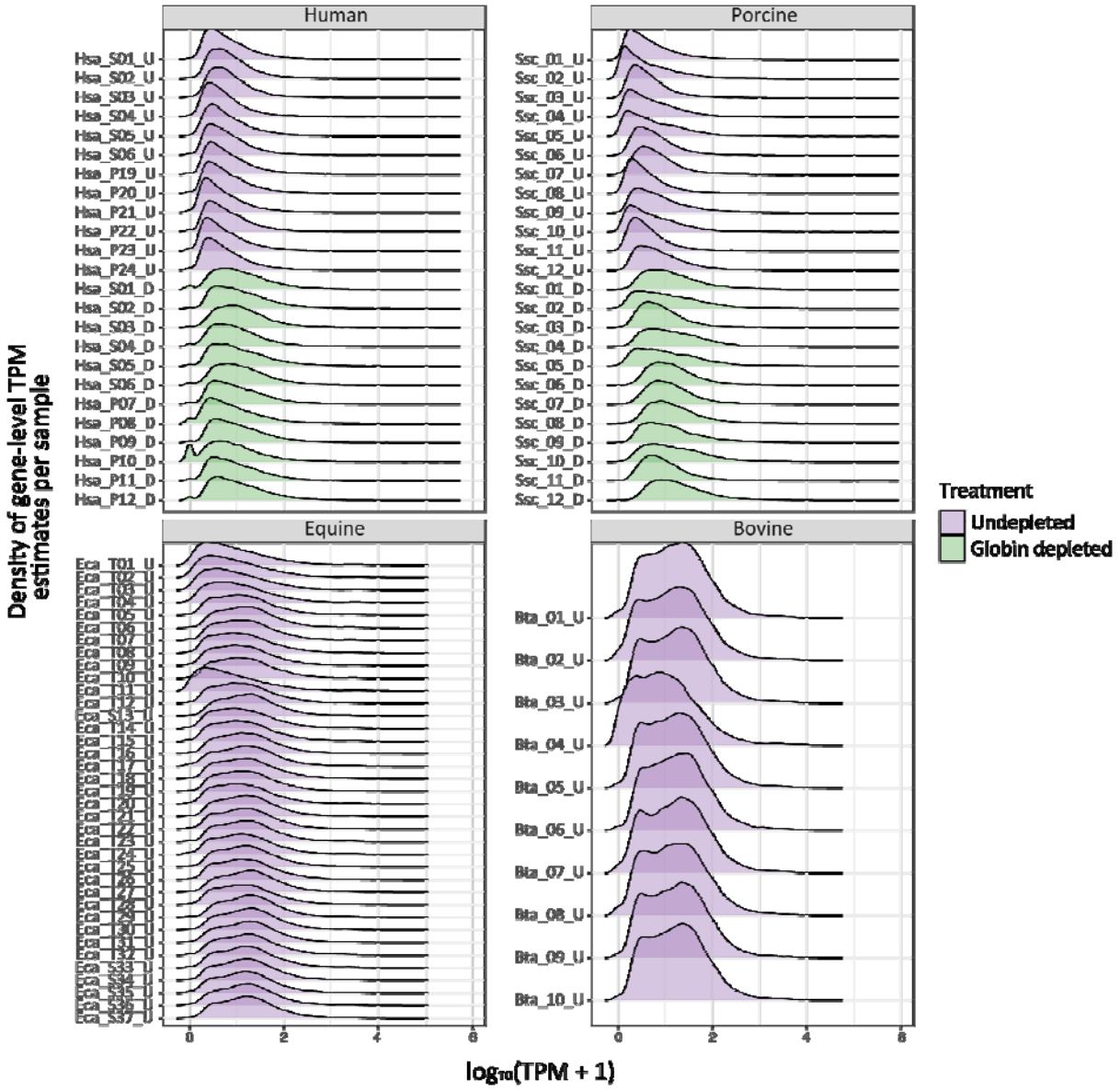

Human
Porcine
Equine
Bovine
Density of gene-level TPM estimates per sample
log10(TPM + 1)
Treatment
Undepleted
Globin depleted

Figure 3

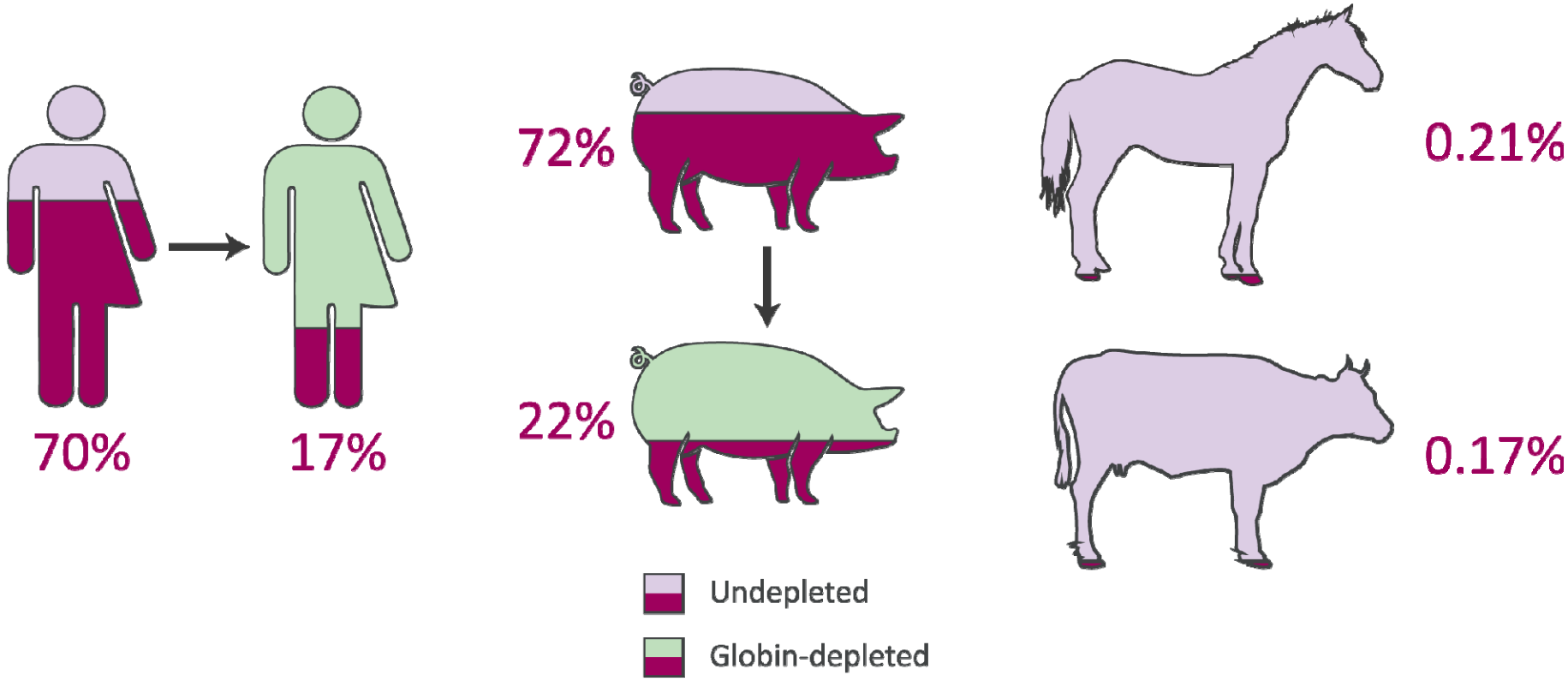
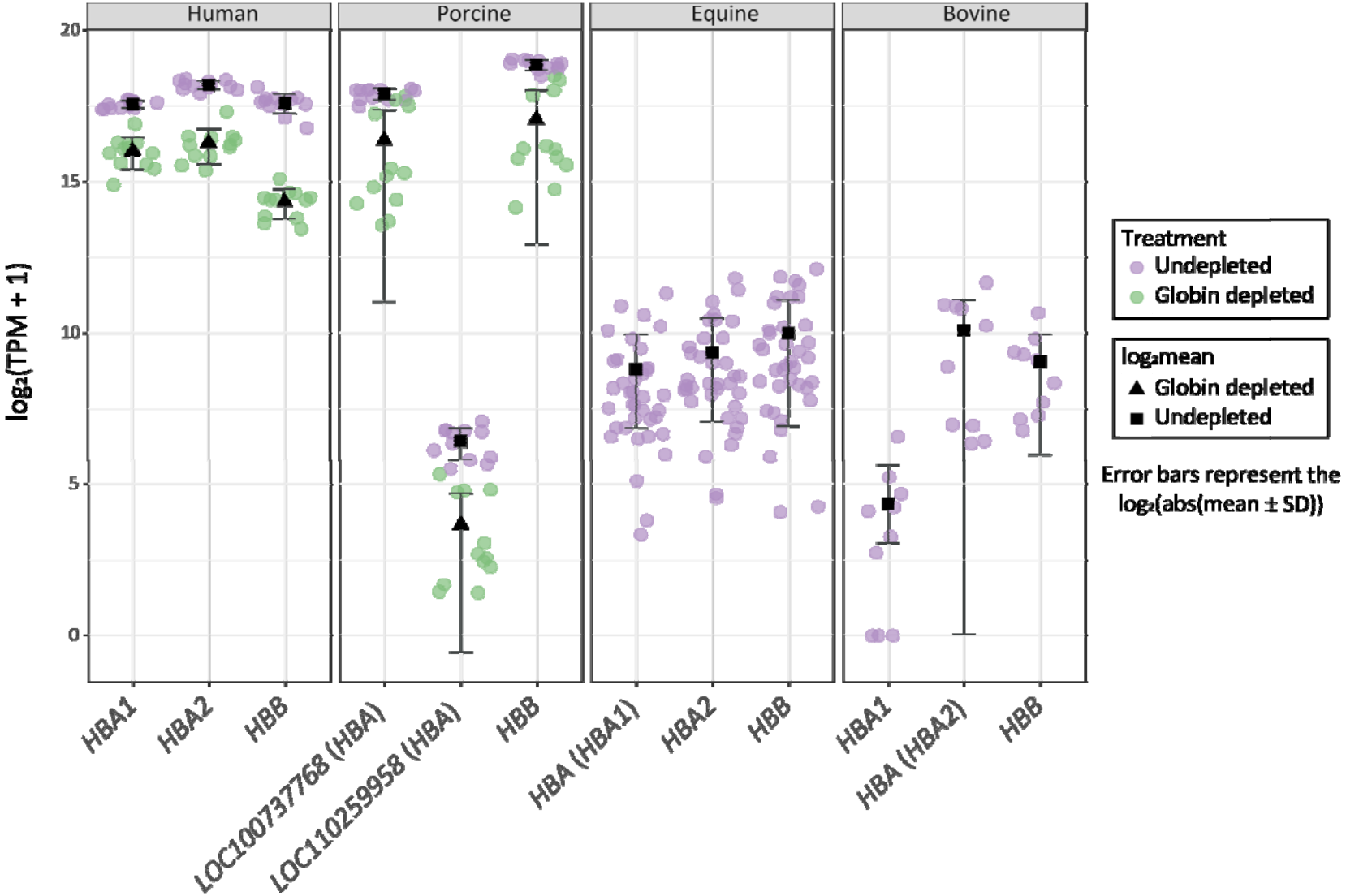


Figure 4


Human
Porcine
Equine
Bovine
log2(TPM + 1)
20
15
10
5
0
HBA1
HBA2
HBB
LOC100737768 (HBA)
LOC110259958 (HBA)
HBB
HBA (HBA1)
HBA2
HBB
HBA1
HBA (HBA2)
HBB
Treatment
Undepleted
Globin depleted
log2mean
Globin depleted
Undepleted
Error bars represent the
log2(abs(mean ± SD))