Title

Characterisation of the horse transcriptome from immunologically active tissues

Authors

J. Moreton[1, 2, 3], S. Malla[2], A. A. Aboobaker[4], R. E. Tarlinton[3] and R. D. Emes[1, 3]

1 Advanced Data Analysis Centre, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, LE12 5RD, UK
2 Deep Seq, Centre for Genetics and Genomics, University of Nottingham, Queen's Medical Centre, NG7 2UH, UK
3 School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, LE12 5RD, UK
4 Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK

Corresponding author:
Joanna Moreton, ADAC, School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, LE12 5RD, UK
0115 951 6359
Joanna.Moreton@nottingham.ac.uk

Abstract

The immune system of the horse has not been well studied, despite the fact that the horse displays several features such as sensitivity to bacterial lipopolysaccharide that make them in many ways a more suitable model of some human disorders than the current rodent models. The difficulty of working with large animal models has however limited characterisation of gene expression in the horse immune system with current annotations for the equine genome restricted to predictions from other mammals and the few described horse proteins. This paper outlines sequencing of 184 million transcriptome short reads from immunologically active tissues of three horses including the genome reference "Twilight". In a comparison with the Ensembl horse genome annotation, we found 8,763 potentially novel isoforms.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.286v1 | CC-BY 4.0 Open Access | received: 14 Mar 2014, published: 14 Mar 2014

Introduction

While no longer the principal means of transport in much of the world, the horse is still an economically important animal in agriculture, sport and gambling associated with horse racing. Individual stallions may be worth several millions of dollars and attract high stud fees creating considerable interest in the genetics of performance traits (Hill *et al.* 2010). The immune response of the horse has not been well characterised, largely due to the difficulties in working with large animals in experimental settings. There are however several components of the equine immune system that make them in many ways a better model of some human disorders than the current rodent models. These include, similarly to human, an exquisite sensitivity to the effects of lipopolysaccharide (LPS) with associated endotoxemia and sepsis (Bryant *et al.* 2007).

Due to a lack of expressed sequence tag (EST) data, the current annotation of the protein coding regions of the horse genome is largely derived from extrapolation from the genes of other species (Coleman *et al.* 2010). Several recent papers have outlined transcription profiles using digital gene analysis of a variety of horse tissues including muscle, leukocytes, cartilage, brain, reproductive tissue, embryos, sperm and blood (Coleman *et al.* 2010; McGivney *et al.* 2010; Serteyn *et al.* 2010; Park *et al.* 2012; Capomaccio *et al.* 2013; Das *et al.* 2013; Iqbal *et al.* 2014). Capomaccio *et al.* (2013) identified new putative non-coding sequences within intergenic and intronic regions whilst Das *et al.* (2013) suggested additions to the structural annotation of four sperm genes. Two of the other studies (Coleman *et al.* 2010; Park *et al.* 2012) detailed extensions to the annotated gene catalogue in the horse based on transcriptome analysis of quite differing tissue sets, methods and results to those used in this paper. They show that the actual expressed transcription profile only partially overlaps the annotated gene set. A direct comparison of our and these two studies is difficult due to the differing tissues, methodologies and the lack of available locations of the predicted novel genes from these studies.

This paper focuses on immunologically active tissues in the horse to further explore this issue. Uniquely we present data on the transcriptional profile from lymphocytes from Twilight, the animal that the published horse genome is derived from. Comparison of this animal with lymphocytes, core immunologically active tissues (lymph node and spleen) and other tissues (liver, kidney and jejunum) from two unrelated animals allows a unique insight into expression of genes with a functional role in the immune system.

Materials and methods

Samples, library preparation and sequencing

The methods are described fully in our previous work (Brown *et al.* 2012) but briefly, five tissue samples; kidney, jejunum, liver, spleen and mesenteric lymph node were collected from an aged gelding (castrated male horse) euthanised due to osteoarthritis. The tissue samples listed were collected from an animal euthanized for clinical reasons, by the veterinary surgeon, under the Veterinary Surgeons act of 1966. Full informed consent of the owner was obtained for use of the samples, taken from that animal post-mortem.

Lymphocytes isolated by Ficoll Paque (GE healthcare) from a healthy 11 year old welsh mountain pony gelding were kindly provided by Dr Julia Kydd (School of Veterinary Medicine and Science, University of Nottingham) under the Home Office and local Ethical Approval Committee (PPL 40/3354). RNA from lymphocytes isolated from a healthy Thoroughbred mare (DNA the horse genome is derived from) was kindly provided by Donald Miller (Baker Institute of Animal Health, Cornell University, USA). This horse was maintained at the Baker Institute for Animal Health, Cornell University, Ithaca, N.Y., USA. Animal care and research activities were performed in accordance with the guidelines set forth by the Institutional Animal Care and Use Committee of Cornell University, protocol #

1986-0216, approved until March 2013.

Sequencing was performed on a SOLiD 3 ABI sequencer generating 50bp reads according to the manufacturer's instructions. Read data are available at the EBI Sequence Read Archive (SRA) under the study accession number ERP001116.

Read trimming and alignment

The horse genome assembly EquCab2 (Wade *et al.* 2009) was downloaded from Ensembl v71 (www.ensembl.org) and contained 26,991 genes and 29,196 transcripts. CLC Genomics Workbench version 6 (CLC Bio, Aarhus, Denmark, www.clcbio.com) was used to apply quality, SOLiD adapter and Poly-N trimming to the read sequences (supplemental file 1). The limit for the removal of low quality sequences was set at 0.2 and a maximum of two ambiguous nucleotides were permitted in each sequence. Any reads less than 20bp were removed after trimming and the average read lengths were 47bp.

TopHat 2.0.9 (Trapnell *et al.* 2009) was used to align the reads to the repeat masked version of the horse genome (Ensembl v71) to enable non-redundant transcriptome analysis. TopHat first aligns non-spliced reads using Bowtie 1.0.0 (Langmead *et al.* 2009) then identifies splice junctions. Gapped alignments are then used by TopHat to map the reads not aligned by Bowtie. In order to utilise the splice sites in all samples, two iterations of TopHat alignments were carried out (Cabili *et al.* 2011). Firstly, the reads from each sample were aligned to the repeat-masked horse genome with default parameters. The splice sites ("junctions") were extracted from all of the output files and duplicates were removed leaving 216,007 sites. These splice sites were pooled together with the non-redundant sites extracted from the Ensembl annotation yielding 399,264 non-redundant splice sites. Each of the samples were then re-aligned with TopHat using the pooled non-redundant splice sites file (with 'raw-juncs' and 'no-novel-juncs' parameters) to the repeat-masked genome.

TopHat was used for the read alignment because it is part of the Tuxedo suite and is therefore a natural input for the Cufflinks assembler (Trapnell *et al.* 2010). It is also the preferred aligner for Scripture (Guttman *et al.* 2010). Cufflinks and Scripture are described in the transcriptome assembly section.

De novo transcriptome assembly

Each of the samples were assembled into separate transcriptomes using two different tools; Cufflinks v2.1.1 (Trapnell *et al.* 2010) and Scripture (Guttman *et al.* 2010) (beta2 version, December 2010). These tools use different approaches for transcript assembly. A minimal set of transcripts is assembled by Cufflinks using a probabilistic model. It performs a minimum cost maximum matching in bipartite graphs (Trapnell *et al.* 2010). Scripture however creates a connectivity graph which represents the adjacency that occurs in the RNA but that is broken in the genome by an intron sequence. A statistical segmentation strategy is used to determine paths with aligned read enrichment over background noise (Guttman *et al.* 2010).

Both Cufflinks and Scripture were run using default parameters, however due to computational time Scripture was run on the named chromosomes only (not on the unanchored contigs "chrUn"). The samples were assembled individually to reduce the complexity of isoforms and hence reduce the chance of incorrectly assembled transcripts (Trapnell *et al.* 2012). The Cufflinks and Scripture assembly files are provided as supplemental files 2 and 3.

The "Cuffmerge" program (included in the Cufflinks package) was used to merge the Cufflinks and Scripture assemblies separately. Stranded transcripts from the two assemblies were compared using the Cufflinks inclusive program "Cuffcompare" with the Cufflinks assembly as a mock reference. The class codes in the Cuffcompare output were used to generate a consensus assembly (University of Nottingham "UoN", supplemental file 4). This consensus assembly was compared to the Ensembl annotations using Cuffcompare (supplemental file 5).

Annotation

The UoN cDNA sequences (supplemental file 6) were extracted from the consensus assembly (*gtf) file and the longest open reading frames (ORFs) were determined. Gene annotation was conducted by prediction of Pfam domains (PfamA.hmm library downloaded June 2013) (Punta *et al.* 2012) using HMMER (Eddy 2011). Associated gene ontology (GO) terms (Ashburner *et al.* 2000) were determined using the Pfam2GO database (version compiled 15/6/2013) of Interpro (Hunter *et al.* 2009). The UoN transcripts were searched against the NCBI non-redundant (NR) database (downloaded 14th November 2013) using BLASTX (Altschul *et al.* 1997), a cutoff evalue of 1e-10 was used to infer homology.

Gene expression analyses

The TopHat BAM files were filtered for unique alignments (SAM flag NH:i:1) and the number of tags per Ensembl gene was calculated using htseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html). These counts were converted into Reads per Kb per million (RPKM) values (Mortazavi *et al.* 2008). A table of RPKM values for all Ensembl genes is provided as supplemental file 7.

As the number of replicates was limiting, identification of genes differentially expressed between samples was not attempted. However, genes enriched in each sample were identified using the following criteria; RPKM > 5 for a sample and RPKM > 10 x the mean of RPKMs for the other samples (supplemental file 8). The "hclust" command in R (R-Core-Team 2013) was used for the hierarchical clustering analysis of gene expression values (RPKMs). It was performed using the default complete linkage method and Euclidean distance. Probability values for each cluster were calculated using the "pvclust" R package (Suzuki & Shimodaira 2006) (bootstrap n = 1000).

Comparison of horse and human gene families

To identify orthologous and potential paralogous gene expansions in the horse evident in our transcriptome data, translations of all horse transcripts were compared to proteins encoded by known human genes (Ensembl build GRCh37.71). Both human and horse proteome sets were first clustered to collapse within-species identical protein sequences generated from alternative transcripts using CD-HIT (Li & Godzik 2006). This resulted in 64,231 human and 29,090 horse sequences. These were compared using Inparanoid (version 4.1, overlap cutoff = 0.5, group merging cutoff = 0.5, scoring matrix BLOSUM62) (Remm *et al.* 2001). Functional comparison of gene sets was conducted using Ingenuity Pathway Analysis (Ingenuity Systems).

Results

Transcriptome assemblies

Around 184 million reads were generated and 159 million remained after trimming; approximately 68.6 million of which were aligned to the reference genome EquCab2 (Table 1). Scripture assembled 102,270 stranded transcripts (27,610 with >1 exon, supplemental file 3) whereas Cufflinks reconstructed 58,182 (20,459 with >1 exon, supplemental file 2). There were 10,518 Cufflinks transcripts that completely matched the intron chain of the Scripture transcripts. In addition to this 18,152 Cufflinks transcripts contained or covered at least one Scripture transcript with the same compatible intron structure. The union of these two sets resulted in 28,230 transcripts, 14,762 of which contained more than one exon (supplemental file 4).

Comparison of consensus assembly to Ensembl

The similarities between the 28,230 consensus transcripts (henceforth referred to as "UoN", University Of Nottingham) and the 28,944 Ensembl transcripts on the named chromosomes were compared (supplemental file 5). There were only 507 UoN transcripts which completely matched the intron chain of an Ensembl transcript. The majority of transcripts (8763, 31%) were identified as potentially novel isoforms of a predicted Ensembl transcript with at least one splice junction shared.

The majority of Ensembl transcripts (18668, 65%) did not overlap with a UoN transcript (supplemental file 9). This could be due to the strict consensus approach used for the UoN assembly. Also, the specific tissues analysed would not be expected to reconstruct all the transcripts from Ensembl which are predicted from genomic DNA and hence all potential transcriptomes.

Around 9,500 (34%) of the 28,230 UoN transcripts were annotated with a Pfam protein domain, approximately 6,600 (23%) with at least one GO term and 16,166 (57%) had at least one significant BLASTX hit against NCBI-NR (supplemental file 10). In total there were 16,305 UoN transcripts with at least one annotation. The UoN annotated transcripts were split into Cuffcompare categories based on the comparison to the Ensembl annotations (supplemental file 10). As expected, the transcripts matching the intron chain ("=") or sharing at least one splice junction ("j") of the Ensembl annotations had the highest percentage of annotated transcripts (e.g. 97% and 99% with BLASTX hits respectively). There were 367 of the 16,166 UoN transcripts with a BLASTX hit that showed homology to only a single species and just under half of these (163) were to *Equus caballus*. The top hit was extracted for each transcript and as expected most of these hits were also to the *Equus caballus* genome. Other mammals with a high number of top hits were *Homo sapiens, Mus musculus, Ceratotherium simum simum, Tursiops truncatus and Sus scrofa.* The full list is shown in supplemental file 11.

Gene expression analyses

The number of Ensembl genes specific to each sample is shown in Table 2 and supplemental file 8 (see also materials and methods). No genes were enriched in more than one sample. The Lymphocyte A sample had many more specific genes than Lymphocyte B. This is possibly due to sample A being taken from the same horse that the published genome is derived from, however the read alignment rate between these two samples is similar suggesting this may not be the major factor. Alternatively this may reflect the immune states of individual horses at the time of sample collection.

The top ten gene ontology (GO) terms for the sample-enriched genes largely reflect the known function of the tissues sampled (supplemental file 12). Hierarchical clustering analysis of the RPKMs between tissues showed three clades (Figure 1). For each of the nodes, the approximately unbiased (au) bootstraps are over 80%. These are reported having superiority over the bootstrap probabilities (bp) (Suzuki & Shimodaira 2006). The lymphocyte samples cluster most closely with the spleen sample which likely reflects the high number of lymphocytes present in the spleen at the time of collection. Whilst the kidney and liver have general shared roles in waste excretion suggesting a possible overlap of transcription profile, determining a definitive reason for the separation of the clade containing lymph node, kidney and liver is not clear. The jejunum sample forms an outgroup and this separation from the other immune-like tissues likely reflects the relatively smaller proportion of lymphoid (Peyer's patch) tissue to non-lymphoid material in this organ. It is also important to consider that only a limited number of samples and animals are compared and so robustness of these relationships is not ensured.

Analysis of genes enriched in each sample identified related enriched canonical pathways. The kidney sample is enriched in genes involved in the “γ-glutamyl Cycle”, “Leukotriene Biosynthesis”, “Glycine Cleavage Complex”, “β-alanine Degradation I” and “4-hydroxyproline Degradation I” pathways. Amino-acid catabolism pathways, possibly reflecting high-energy consumption of the kidney, dominate these. The liver sample is enriched with genes involved in the degradation of chemical products (e.g. nicotine and melatonin). Enzymes including members of the CYP450 and UDP-Glucuronosyltransferase (UGT) gene families, which are known to be highly expressed in the liver, are enriched. The spleen shows enrichment of genes involved in the pathways “Autoimmune Thyroid Disease Signaling”, “Hematopoiesis from Pluripotent Stem Cells”, “Primary Immunodeficiency Signaling”, “Dendritic Cell Maturation”, and “Agranulocyte Adhesion and Diapedesis”. Largely these are due to the enrichment of genes encoding the immunoglobulin heavy chain and Fc fragment of IgG. Enrichment of these pathways reflects the role of the spleen as the primary site of white blood cell differentiation and storage. The lymph node sample is enriched in the pathways, “Primary Immunodeficiency Signaling”, “Hematopoiesis from Pluripotent Stem Cells”, “Autoimmune Thyroid Disease Signaling”, “Allograft Rejection Signaling” and “Communication between Innate and Adaptive Immune Cells”. As with the spleen these are predominantly due to the enrichment of genes encoding the immunoglobulin heavy chain proteins and result from the contained white blood cell content.

Identification of paralogous gene expansions in horse

Previously the horse genome was described as containing lineage specific expansions of olfactory and immune genes (Wade *et al.* 2009). The expansion of these families particularly immune related genes is often seen in mammalian genome comparisons (Emes *et al.* 2003). Wade *et al.* (2009) reported that there were 99 gene families expanded in the horse genome. Comparison of the proteins encoded by the transcripts identified here identified 4,605 groups of horse:human orthologs and 10,607 inparalogs. The majority of these represent expansions in human where a single horse protein was encoded by the transcriptome data generated here. 91 families were identified with a specific expansion in horses (many:1 relationship). Of these the large majority (83/91) represent simple duplications in the horse transcriptome compared to human. Three families have four non-identical encoded proteins orthologous to a single protein in humans. Annotation of these genes identifies them as T cell receptor alpha constant (TRAC), heparin sulfate proteoglycan 2 (HSPG2 and solute carrier family 23 (ascorbic acid transporter) member 1 (SLC23A1). An additional four gene families are identified with three encoded proteins in horse compared to a single protein in human. These are GTPase, IMAP family member 7 (GIMAP7), UDP glucuronosyltransferase 1 family polypeptide A6 (UGT1A6), solute carrier family 44 (SLC44A2), ATP-binding cassette, sub-family C member 8 (ABCC8 and sushi, nidogen and EGF-like domains 1 (SNED1).

An additional 99 families were found with expansions in both human and horse (many:many relationship). Reflecting the tissues used for RNA extraction, genes in this category are highly enriched for immune functions. The most significantly populated pathways are “Role of NFAT in regulation of the immune response”, “CD28 Signaling in T helper cells”, “iCOS-iCOSL signaling in T helper cells”, “Natural killer cell signaling” and “PKCθ signaling in T lymphocytes”.

Discussion

The analysis conducted here provided insight into the transcriptome of immune tissues from the horse and made these analyses freely available (supplemental files). Whilst it is unclear why the horse transcriptome should contain the specific expansions of gene families described, the analysis provided insight into potential areas of T-cell biology which may underlie equine specific immunobiology. The analysis conducted also allowed the identification of gene expansions such as UGT1A6, part of a putative paralogous gene expansion in horse relative to human. UGT1A6 is a member of the UDP-glucuronosyltransferases (UGTs), a gene family essential for metabolism of both xenobiotic and endobiotic substances. In contrast to humans and model organisms, there is currently little information regarding specific drug metabolism in animals of veterinary importance. This is particularly true in the horse, despite it being potentially exposed to extensive medical care. Due to the broad application of its mechanisms on xenobiotic substances, the UGT enzyme group has important implications in pharmacokinetics, the development of drugs and their associated elimination rates. Importantly, as many of the drugs used in equids are adopted from those designed from human UGT research, understanding the differences in genes encoding these proteins may provide a basis for investigation into the UGT group of enzymes in horses and will open up further opportunities for specific pharmacokinetic research into UGT related equine drug metabolism potentially reducing toxic drug interactions.

The data presented here demonstrated the utility of second generation sequencing in significantly advancing knowledge of gene transcription in a poorly characterised species. A large number of potential novel genes were identified alongside some extensions to existing genes. The completeness of these predictions remains to be confirmed by traditional mRNA isolation and sequencing but the data presented provides a starting point for the study of whole groups of genes.

Acknowledgements

Dr. Julia Kydd (School of Veterinary Medicine and Science, University of Nottingham) and Dr. Donald Miller (Baker Institute of Animal Health, Cornell University, USA) for their kind donation of lymphocyte samples and RNA. We would also like to thank Dr Martin Blythe, Damian Kao, Victoria Wright and Katharine Rangeley (University of Nottingham) for useful discussions.

Supplemental Information

Supplemental Files

- *S1_CLC_SOLiD_trim_adapter_list.xls*: The trim adapter list used in CLC with SOLiD adapter sequences and single-base-repeat sequences.
- *S2_Cufflinks_assembly.gtf.gz*: Annotation of the individual pre-consensus horse Cufflinks assembly.
- *S3_Scripture_assembly.gtf.gz*: Annotation of the individual pre-consensus horse Scripture assembly.
- *S4_UoN_horse_consensus_assembly.gtf.gz*: Annotation of the University of Nottingham (UoN) final consensus horse assembly.
- *S5_Ensembl-vs-UoN_Cuffcompare-results.xls*: Table of Cuffcompare results showing the similarities between the University of Nottingham (UoN) consensus assembly and the Ensembl annotations (using Ensembl as a Cuffcompare reference).
- *S6_UoN_horse_cDNA_sequences.fa.gz*: University of Nottingham (UoN) consensus horse cDNA sequences.
- *S7_Ensembl71_genes_RPKMs_UoN-reads.xls*: RPKM values for all Ensembl v71 genes.
- *S8_sample-enriched-genes.zip*: Tables containing the sample enriched Ensembl v71 gene IDs.
- *S9_Ensembl-vs-UoN_Cuffcompare-results_UoN-as-ref.xls*: Table of Cuffcompare results showing the similarities between the University of Nottingham (UoN) consensus assembly and the Ensembl annotations (using UoN as a Cuffcompare

reference).

- *S10_Ensembl-vs-UoN_Cuffcompare-results_PFAM-GO-BLASTX.xls*: Table of Cuffcompare results showing the similarities between the University of Nottingham (UoN) consensus assembly and Ensembl. The number of annotated UoN transcripts for each of the Cuffcompare categories is also shown.
- *S11_species_UoN_numTopHits_BLASTX.xls*: Table showing the number of top hit BLASTX hits for each species.
- *S12_top10-GO-names_sample-enriched-genes.xls*: Table showing the top ten gene ontology (GO) term names for the sample-enriched genes.

Figure Legends

Figure 1: Hierarchical clustering of gene expression profiles in 7 tissues

The R command "hclust" was used for the hierarchical clustering analysis. The branch values are the pvclust approximately unbiased (AU) p-values (left) and bootstrap (BP) probability values (right) where the p-values are expressed as percentages (95% is equivalent to p-value < 0.05) (Beliakova-Bethell *et al.* 2013).

Tables

Table 1: Read statistics for the seven samples

Sample	Horse	Raw reads	Trimmed reads	Reads aligned
Lymphocyte A	A	20,853,992	18,243,283	7,856,017
Lymphocyte B	B	32,050,093	27,315,182	11,659,787
Jejunum	C	19,902,170	17,241,772	7,659,938
Kidney	C	33,158,285	27,746,321	10,937,750
Liver	C	23,176,545	19,982,256	8,565,159
Lymph node	C	24,671,029	21,444,476	9,221,340
Spleen	C	30,421,675	26,828,834	12,708,499

(A) "Twilight", healthy Thoroughbred (B) healthy castrated male welsh mountain pony (C) aged gelding euthanised for arthritis.

Table 2: Sample-enriched genes

Sample	# Sample-enriched genes
Lymphocyte A	201
Lymphocyte B	23
Jejunum	228
Kidney	318
Liver	272
Lymph node	44
Spleen	79

References

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M. & Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.

Beliakova-Bethell N., Massanella M., White C., Lada S.M., Du P., Vaida F., Blanco J., Spina C.A. & Woelk C.H. (2013) The effect of cell subset isolation method on gene expression in leukocytes. *Cytometry* **85**, 94-104.

Brown K., Moreton J., Malla S., Aboobaker A.A., Emes R.D. & Tarlinton R.E. (2012) Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. *Virology* **433**, 55-63.

Bryant C.E., Ouellette A., Lohmann K., Vandenplas M., Moore J.N., Maskell D.J. & Farnfield B.A. (2007) The cellular Toll-like receptor 4 antagonist E5531 can act as an agonist in horse whole blood. *Vet Immunol Immunopathol* **116**, 182-9.

Cabili M.N., Trapnell C., Goff L., Koziol M., Tazon-Vega B., Regev A. & Rinn J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25**, 1915-27.

Capomaccio S., Vitulo N., Verini-Supplizi A., Barcaccia G., Albiero A., D'Angelo M., Campagna D., Valle G., Felicetti M. & Silvestrelli M. (2013) RNA Sequencing of the Exercise Transcriptome in Equine Athletes. *PloS one* **8**, e83504.

Coleman S.J., Zeng Z., Wang K., Luo S., Khrebtukova I., Mienaltowski M.J., Schroth G.P., Liu J. & MacLeod J.N. (2010) Structural annotation of equine protein-coding genes determined by mRNA sequencing. *Anim Genet* **41 Suppl 2**, 121-30.

Das P.J., McCarthy F., Vishnoi M., Paria N., Gresham C., Li G., Kachroo P., Sudderth A.K., Teague S. & Love C.C. (2013) Stallion sperm transcriptome comprises functionally coherent coding and regulatory RNAs as revealed by microarray analysis and RNA-seq. *PloS one* **8**, e56535.

Eddy S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195.

Emes R.D., Goodstadt L., Winter E.E. & Ponting C.P. (2003) Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* **12**, 701-9.

Guttman M., Garber M., Levin J.Z., Donaghey J., Robinson J., Adiconis X., Fan L., Koziol M.J., Gnirke A. & Nusbaum C. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology* **28**, 503-10.

Hill E.W., Gu J., McGivney B.A. & MacHugh D.E. (2010) Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance. *Anim Genet* **41 Suppl 2**, 56-63.

Hunter S., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., Bork P., Das U., Daugherty L., Duquenne L., Finn R.D., Gough J., Haft D., Hulo N., Kahn D., Kelly E., Laugraud A., Letunic I., Lonsdale D., Lopez R., Madera M., Maslen J., McAnulla C., McDowall J., Mistry J., Mitchell A., Mulder N., Natale D., Orengo C., Quinn A.F., Selengut J.D., Sigrist C.J., Thimma M., Thomas P.D., Valentin F., Wilson D., Wu C.H. & Yeats C. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, D211-5.

Iqbal K., Chitwood J.L., Meyers-Brown G.A., Roser J.F. & Ross P.J. (2014) RNA-Seq Transcriptome Profiling of Equine Inner Cell Mass and Trophectoderm. *Biology of Reproduction*.

Langmead B., Trapnell C., Pop M. & Salzberg S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

Li W. & Godzik A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-9.

McGivney B.A., McGettigan P.A., Browne J.A., Evans A.C., Fonseca R.G., Loftus B.J., Lohan A., MacHugh D.E., Murphy B.A., Katz L.M. & Hill E.W. (2010) Characterization of the equine skeletal muscle transcriptome identifies novel functional responses to exercise training. *BMC genomics* **11**, 398.

Mortazavi A., Williams B.A., McCue K., Schaeffer L. & Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-8.

Park K.D., Park J., Ko J., Kim B.C., Kim H.S., Ahn K., Do K.T., Choi H., Kim H.M. & Song S. (2012) Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq. *BMC genomics* **13**, 473.

Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G. & Clements J. (2012) The Pfam protein families database. *Nucleic acids research* **40**, D290-D301.

R-Core-Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

Remm M., Storm C.E. & Sonnhammer E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-52.

Serteyn D., Piquemal D., Vanderheyden L., Lejeune J.P., Verwilghen D. & Sandersen C. (2010) Gene expression profiling from leukocytes of horses affected by osteochondrosis. *J Orthop Res* **28**, 965-70.

Suzuki R. & Shimodaira H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540-2.

Trapnell C., Pachter L. & Salzberg S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11.

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D.R., Pimentel H., Salzberg S.L., Rinn J.L. & Pachter L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-78.

Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., van Baren M.J., Salzberg S.L., Wold B.J. & Pachter L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5.

Wade C.M., Giulotto E., Sigurdsson S., Zoli M., Gnerre S., Imsland F., Lear T.L., Adelson D.L., Bailey E., Bellone R.R., Blocker H., Distl O., Edgar R.C., Garber M., Leeb T., Mauceli E., MacLeod J.N., Penedo M.C., Raison J.M., Sharpe T., Vogel J., Andersson L., Antczak D.F., Biagi T., Binns M.M., Chowdhary B.P., Coleman S.J., Della Valle G., Fryc S., Guerin G., Hasegawa T., Hill E.W., Jurka J., Kiialainen A., Lindgren G., Liu J., Magnani E., Mickelson J.R., Murray J., Nergadze S.G., Onofrio R., Pedroni S., Piras M.F., Raudsepp T., Rocchi M., Roed K.H., Ryder O.A., Searle S., Skow L., Swinburne J.E., Syvanen A.C., Tozaki T., Valberg S.J., Vaudin M., White J.R., Zody M.C., Lander E.S. & Lindblad-Toh K. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865-7.